

西安交通大学

毕业设计（论文）

题 目 基于统计学习的骨关节炎风险预测模型构建

生命学院 理科试验班(化学生物 H) 专业 81 班

学生姓名 郭骐瑞

学 号 2186113661

指导教师 郭燕

设计所在单位 西安交通大学

2022 年 6 月

摘要

骨关节炎是一种由遗传因素与环境因素共同作用所产生的关节退行性病变，目前尚无有效的治疗方法。全球范围内骨关节患者群体庞大且规模逐渐扩大，使得骨关节炎成为导致残疾与疼痛的主要因素之一。研究显示通过风险预测模型的骨关节的早期预防与诊断能显著改善患者预后，然而现有的骨关节炎风险预测模型存在着性能差，可解释性弱的缺点，远不能达到临床使用需求。本文因此提出并设计了一种以图神经网络为核心的基于个体基因型与表型信息的骨关节炎风险预测模型。

本文主要完成以下三个方面工作：一、本文从 UKBiobank 数据库获取了 13706 名个体的基因型与表型数据，并根据现有的全基因组关联研究与相关指标对该数据进行质控与预处理，用以模型的训练与测试。二、本文设计了以切比雪夫图神经网络为核心的结合基因型与表型信息的风险预测模型，该模型能够通过图估计器将输入无结构数据转化为图数据并进行图神经网络处理，并同时具备预测结果解释模块。三、本文对该风险预测模型的功能与性能加以测试评估。本文选择了一系列指标对模型的预测性能加以评估，并同包括多基因风险评分模型、常见机器学习算法在内的传统疾病风险预测模型相比较。证明本模型（AUC 0.74）较传统疾病风险预测模型（AUC 0.5）而言有着十分明显的性能改善。本文还通过对图估计器工作过程分析展现了本模型优秀的可解释性。

在模型中，本文创新性使用了基于变分期望最大化的图估计器结合图神经网络来处理无结构的基因型数据。研究结果证实该方法不仅给出了较好的风险预测准确度，还能通过挖掘基因型数据中潜藏的信息对疾病病因、分型等因素加以推断。本文的研究成果一方面为无结构数据在图神经网络中的处理提供了新方法，另一方面也为疾病风险预测模型的构建提供了新思路，对图神经网络与疾病风险预测模型的广泛应用具有积极意义。

关键词：骨关节炎；风险预测模型；图神经网络

ABSTRACT

Osteoarthritis is a degenerative joint disease for which there is no effective treatment. The large and expanding population of osteoarthritis patients worldwide makes osteoarthritis one of the leading causes of disability and pain. Studies have shown that the early prevention and diagnosis of osteoarthritis through risk prediction models can significantly improve the prognosis of patients. Nevertheless, the existing osteoarthritis risk prediction models are of poor performance and weak interpretability, which are far from meeting the needs of clinical use. Therefore, this paper proposes a graph neural network(GNN)-based osteoarthritis risk prediction model based on individual genotype and phenotype information.

This paper mainly carries out three parts: First, this paper obtains the genotype and phenotype data of 13,706 individuals from the UKB database and conducts quality control and preprocessing on the data according to GWAS research and related indicators. Second, this paper designs a Chebyshev-GNN-based osteoarthritis risk prediction model, which can integrate phenotype and genotype information. Equipped with a graph estimator, this model is also capable of transforming plain data to a graph for GNN, the model interpreter qualifies the model to elaborate on the predicted result. Third, this paper introduces a series of indicators to evaluate the model's predictive performance and compares it with traditional disease risk prediction models. It is demonstrated that this model (AUC 0.74) has a very significant performance improvement compared with the traditional disease risk prediction model (AUC 0.5). This paper also demonstrates the model's excellent interpretability by analysing the graph estimator's working process.

This paper innovatively uses a graph estimator based on VEM combined with a graph neural network to process unstructured genotype data. The study's results confirmed that the model gives better risk prediction accuracy and can infer factors such as disease aetiology and classification by mining the hidden information in genotype data. The result of this study provides a new method for the processing of unstructured data in the graph neural network; it also provides a new idea for the construction of the disease risk prediction model.

KEY WORDS: Osteoarthritis; Risk Prediction Model; Graph Neural Network

目 录

1	绪论	1
1.1	研究背景与意义	1
1.2	现状研究综述	1
1.2.1	骨关节炎与风险预测模型	1
1.2.2	图神经网络	3
1.3	本文研究内容	3
2	数据预处理	5
2.1	数据来源	5
2.2	样本标注	5
2.3	基因型位点质控	6
2.4	特征筛选	7
2.4.1	基于卡方的特征筛选	8
2.4.2	基于支持向量机的特征筛选	8
2.5	数据集描述	9
2.6	本章小结	9
3	骨关节炎风险预测模型构建	11
3.1	图估计器	11
3.1.1	图神经网络中无结构数据的处理方法	11
3.1.2	问题描述	11
3.1.3	基于变分期望最大化的图估计	13
3.2	图神经网络	22
3.2.1	基本定义	22
3.2.2	切比雪夫层	23
3.2.3	结构设计	24
3.3	表型融合	25
3.4	图解释器	26
3.5	本章小结	27
4	结果讨论与分析	28
4.1	评估指标	28

4.2 模型预测效果评价	29
4.2.1 基准计算	29
4.2.2 特征筛选方法对比	29
4.2.3 图神经网络处理	31
4.2.4 表型融合	32
4.3 图估计效果	32
4.3.1 估计器在训练过程中的效果	32
4.3.2 图估计器工作过程分析	36
4.4 图解释器结果案例分析	45
4.5 本章小结	45
5 总结与展望	48
5.1 工作总结	48
5.2 展望	49
参考文献	50
附录 A 外文文献原文	56
附录 B 外文文献译文	66

1 绪论

1.1 研究背景与意义

骨关节炎是一种由遗传因素与环境因素共同作用所产生的关节退行性病变，目前尚无有效的治疗方法。世界范围内至少三亿人罹患骨关节炎且患者群体逐年扩大。而其较差的预后也使得骨关节炎成为了全球范围内导致残疾与疼痛的主要因素之一。临床研究显示，早期诊断与介入对骨关节炎患者病程控制有着积极的影响。因此，构建骨关节炎患病风险预测模型有助于通过早期筛查与预警的方式帮助潜在患者控制病程发展。目前已有诸多关于骨关节炎基因型的全基因组关联研究（GWAS）鉴定了若干骨关节炎的易感 SNP 位点，也有相关研究基于这些位点建立了骨关节炎风险预测模型。但是这些模型的预测效果较差，远不能达到临床预测的要求。此外，现有的模型也无法处理输入基因型数据之间的复杂网络关系。同时这些模型只能根据输入的位点信息给出判断，无法解释输入位点之间的关联与特定位点在预测过程中的贡献值，模型可解释性方面存在较大不足。以上存在的问题与不足极大限制了骨关节炎风险预测模型的实际应用，如何改进构建模型的方法以提高模型的预测效果，也成为相关研究亟待解决的关键问题。

图作为一种数据结构能通过对节点与边的描述同时反映节点信息与节点之间关系，目前已被广泛应用于分子生物学代谢网络、动物行为学社交网络等生物学领域的研究中。鉴于基因型位点相互关联乃至成网的特征，同时考虑到图在提取和处理处理节点信息（特征）及节点间关系过程中的优势，本文使用图来描述基因型位点特征与位点-位点关系。但是，图数据作为一种非欧数据，传统的诸如卷积神经网络在内的机器学习方法并不适用于直接对图数据进行处理和分析。近年来随着图数据在各个领域的广泛应用，为了解决这一问题，适用于处理图数据的图神经网络应运而生。图神经网络能够基于图信息完成诸如图分类、节点分类等任务并且具有效果好、可解释性强等优点，适合用来处理基因型数据网络。因此，本文将基于图形式的基因型位点，通过图神经网络，结合患者表型信息来构建一种高效的、可解释的骨关节炎风险预测模型。该模型能够基于患者基因型及表型信息，在给出高效患病风险预测结果的同时对患者基因型网络进行解释，因此对骨关节炎的早期诊断与分型具有重要的临床意义。

1.2 现状研究综述

1.2.1 骨关节炎与风险预测模型

骨关节炎是最常见的关节退行性病变之一，目前报道的案例显示骨关节炎主要影响着人体膝、髋、手等若干关节。骨关节炎的症状主要包括关节疼痛，僵硬、柔韧性丧失，甚至导致关节失能。更严重的是，目前对骨关节炎的治疗以缓解患者痛苦为主，尚

无对骨关节的有效临床治愈手段。以上原因也使得骨关节炎成为了导致失能与残疾的主要因素之一，并且严重影响到了患者的生活质量。^[1]同时，骨关节炎在世界范围内有着较庞大的患者群体：世界范围内至少有三亿人罹患骨关节炎^[2]，仅在英国范围内的70岁以上群体中就有约40%的个体受骨关节炎影响^[3]，而在全年龄个体中则至少有一千万患者，这导致了每年至少148亿英镑的直接医疗支出。^[4]而研究显示，对骨关节炎的早期诊断与介入能很大程度上延缓关节异常生长进程，对骨关节炎患者的预后改善有着积极的效果^[1]。因此许多研究也将注意力转到了基于风险预测模型的骨关节炎的早期诊断与预防上，目前已有的风险预测模型主要着眼于揭示与骨关节炎病程发展相关的影响因子，包括肥胖，关节错位，关节损伤，骨质增生与高强度的运动^[5-7]。同时有研究发现骨关节炎也受遗传因素调控。^[8]并试图通过全基因组关联分析研究骨关节炎的基因背景，以此鉴定并得到了丰富的具有统计学意义、能够作为疾病风险预测模型标志物的疾病易感单核苷酸多态性位点（Single Nucleotide Polymorphism, SNP）^[9, 10]。然而，基于这些疾病易感位点建立的风险预测模型性能却不尽如人意：例如^[9]等人建立的基于PRS算法的风险预测模型，其效果与随机预测相当，远不能达到临床需求，还存在着极大的改进空间。

目前基于基因信息的疾病风险预测模型主要分为两种思路，一种思路通过统计学分析计算个体基因型位点对目标性状的贡献，并根据总体值对样本患病风险进行评估，该方法以PRS为代表^[11]。该方法及其变体已被运用于精神分裂症、I型糖尿病与过敏性肠炎的诊断与筛查中^[12-14]；另一思路则基于目前具有广泛应用机器学习算法，通过对样本信息进行学习，进而输出模型预测的样本患病风险。构建疾病风险预测模型所常用的机器学习方法主要分为基于回归的机器学习算法与基于树的机器学习算法。前者主要包括决策树与随机森林算法，该算法主要通过构建决策分类规则来完成输入输出数据的建模。有研究便通过随机森林法构建了II型糖尿病的疾病预测模型。^[15]该研究采用的随机森林法相较于基于回归的支持向量机法有着较高的预测准确性。而后者主要有逻辑回归法、支持向量机、神经网络等算法。这类算法通过参数或非参数回归的方法构建损失函数并完成回归计算。这些算法已经被运用于癌症、老年痴呆症、心脏病以及糖尿病的风险预测^[16-20]。而近年来随着神经网络的广泛应用，基于其发展来的深度学习疾病风险预测模型也受到越来越多的关注。一项研究肥胖预测模型的研究展现了其发掘样本信息的能力^[21]。相较于传统机器学习算法，深度学习算法具有更好的预测准确性。

但是以上常见的风险预测模型构建方法仍存在着许多问题：首先，对于诸如基因型位点网络或代谢物网络等存在复杂结构的数据，上述方法都无法深层次挖掘数据的内在联系。目前的算法只将输入的位点作为独立的数据点处理，这便会导致输入阶段潜在信息的丧失。其次，这些基于机器学习或者深度学习的方法只能建立从输入数据到输出数据的映射关系，但无法结合输入数据对该映射关系的形成过程给出因果解释，即所谓的“黑箱化”。该问题使得基于以上算法构建的风险预测模型虽然能给出预测值，但是无法因此了解到使得模型做出该预测的决策过程，使得潜藏在输出数据内部的信

息被浪费。

综上，目前基于基因型信息的骨关节炎风险预测模型仍存在着预测准确率低，处理网络数据乏力，可解释性差等问题，相关风险预测模型建立的方法亟待改进。

1.2.2 图神经网络

图是由顶点与边组成的一种数据结构，该数据结构既描述了顶点的性质，也描述了顶点与顶点之间的相关关系。这种数据原理上同生物学领域的许多概念相契合，能够用来描述常见的例如代谢网络、SNP 网络的性质。本文主要以图的形式来整合并对输入的 SNP 数据进行处理。但是，同我们耳熟能详的图片、文本、序列等数据不同，图数据不满足平移不变性，不能投影到欧几里得空间中。而平移不变性又是目前常见的诸如卷积神经网络，递归神经网络等深度学习网络所依赖的关键假设。^[22] 因此这些神经网络不能被直接用来处理图数据。但是随着图数据的广泛应用，能够处理图数据的神经网络也在逐渐发展。A. Sperduti and A. Starita^[23] 率先提出了一种能够应用于有向图的神经网络。Gori^[24] 等人则定义了图神经网络这一概念。而随着卷积方法在传统的成功，图卷积方法也成为了图神经网络中一项热门的研究方向，并产生了诸多已被应用到实际生产生活中用来解决诸如节点分类、图分类等问题的神经网络框架。

图卷积神经网络根据卷积核功能分为两种类：基于空间的图卷积与基于谱的图卷积。基于空间的卷积借助了信息传播^[25] 的思想，认为图中的节点信息通过边进行扩散，卷积核作用于节点的空间邻域，继而通过该空间邻域计算节点信息。目前得到广泛应用的空间卷积算法主要有 GIN^[25], GAT^[26], DCNN^[27] 等；而基于谱的图卷积则借助图的拉普拉斯量将图结构于傅里叶空间展开，有助于识别图结构中的潜藏结构。基于这一方法的谱卷积方法主要有 GCN^[28], ChebyNet^[29], AGCN^[30] 等。目前图神经网络已被广泛应用于包括药物筛选^[31, 32], 计算机视觉^[33] 等图数据的分析。Ghosal^[34] 等人便将图神经网络用于阿尔兹海默症疾病风险的基因型预测图分类问题中，证明了使用图神经网络构建基于基因型的疾病风险预测模型的可行性。

此外，由于图的结构及其蕴含的信息特点，也有研究^[35] 开发出了对图神经网络的解释器。该解释器通过分析已训练好的图神经网络与预测结果，给出与预测结果相关的子图。利用该类型解释器，我们可以在生成风险预测结果时同时获取与该预测结果相关的子图信息，并为致病位点或疾病分型相关工作提供便利。而这类型工作是传统疾病预测模型所无法实现的。

1.3 本文研究内容

综上所述，本文将基于谱图卷积网络，试图构建一个可解释的基于患者基因型信息与表型信息的高效骨关节炎风险预测模型。本文工作主要分为三个方面：首先，本文根据目前已发表的 GWAS 研究及公共数据库 UK BioBank 获取患者表型与基因型数据并进行数据预处理；之后构建了包括特征筛选、邻接矩阵估计、图卷积神经网络、表型

信息融合四个模块在内的骨关节炎风险预测模型；最后本文对该模型的性能以及预测结果进行了进一步的分析和解读，继而对模型的预测准确性，可解释性等指标进行评估。具体研究内容分为以下五个章节

第一章，绪论。该章首先阐述论文的研究背景与意义，并对骨关节炎、风险预测模型以及图神经网络领域加以综述。最后介绍了本文的主要研究内容。

第二章，数据预处理。该章介绍了根据已有研究从公共数据库获取患者基因型与表信息以及根据一定评判标准对该数据进行筛选与预处理的过程。

第三章，风险预测模型搭建。该章介绍了从特征筛选、邻接矩阵估计、图卷积神经网络、表型信息融合四个模块出发构建骨关节炎风险预测模型的过程与理论原理。

第四章，模型性能评价。该章分析了本文搭建模型的预测准确率并将其同传统风险预测模型准确率进行比较。同时还从模型可解释性角度出发分析了模型预测结果。一方面证明本文建立模型相较于传统模型具有较高程度的提升，另一方面也展现了模型对骨关节炎性状之外信息的揭示能力。

第五章，总结与展望。总结本文对骨关节风险预测模型的研究工作，并对未来本研究还需要解决的问题进行展望。

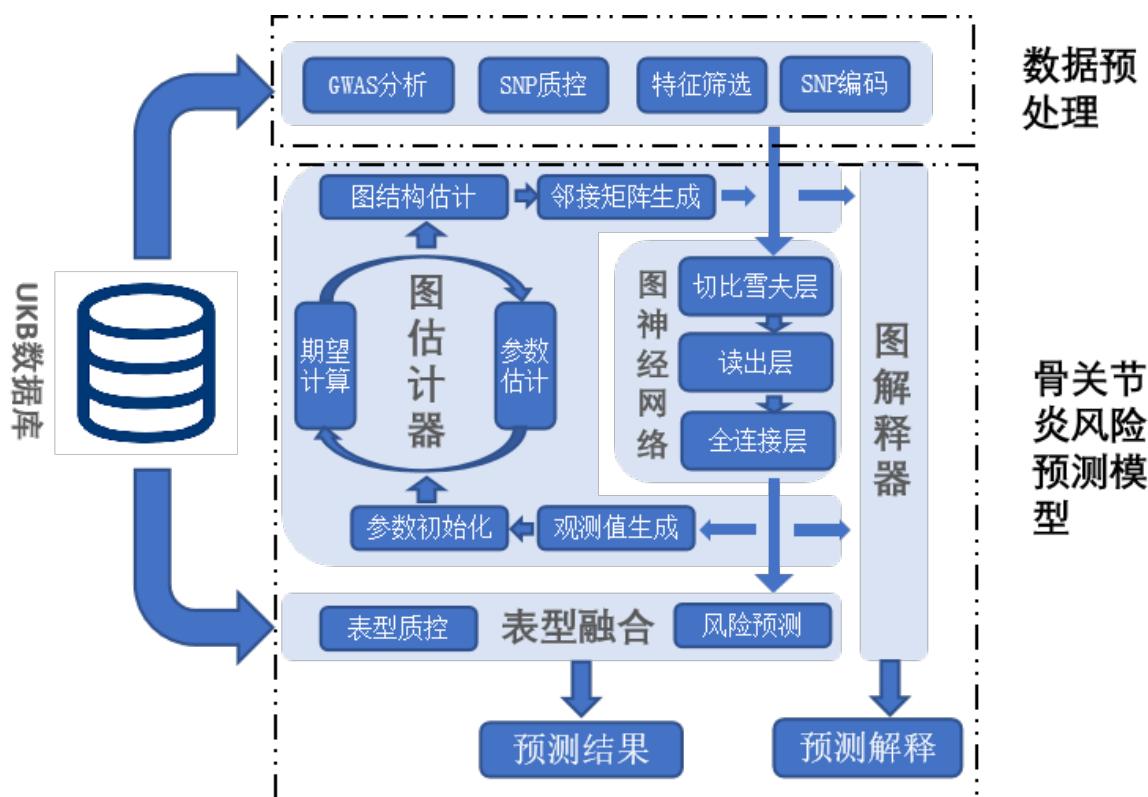


图 1-1 技术路线图

2 数据预处理

2.1 数据来源

本文所使用的基因型与表型数据主要来自于 UK Biobank[36] 数据库。该数据库收集并整合了 2006-2010 年间于英国招募的约 500000 名志愿者的生物样本、身体状况、功能评估、多种表型及遗传学数据。由于其所提供的丰富多样的遗传与表型信息，UK Biobank 已经成为了研究多种复杂表型和遗传病的重要数据来源。本文主要基于 UK Biobank 的第三批发布数据开展骨关节炎相关研究。

2.2 样本标注

我们从 UK Biobank 数据库中获取骨关节炎患者（阳性样本）与非骨关节炎患者（阴性样本），并使用个体疾病伤害分类标准编码来对 UK Biobank 所提供的样本信息进行区分。国际疾病伤害及死因分类标准（The International Statistical Classification of Diseases and Related Health Problems, ICD）是由世界卫生组织所制定的根据特定规则对人群可能出现疾病进行归类的一套编码系统。[37] UK Biobank 已经对数据库内样本进行了 ICD 标注，因此我们可通过这一方法筛选研究对象。

参考其他骨关节炎的相关研究，[10] 我们确定如下筛选规则。我们认定包含如表2-1所含 ICD 编码的个体为骨关节炎患者。

表 2-1 患者评判依据

ICD 编码	分类
M15-X	骨关节炎
M16-X	髋关节炎
M17-X	膝关节炎
M18-X	腕掌关节关节炎
M19-X	其他关节炎

在筛选非骨关节患者时，为防止其他关节症状的干扰，我们也根据文献报道对表2-2所含编码的个体予以排除。

依照以上规则，我们从 UK Biobank 数据库中筛选并获得 6706 名骨关节炎患者作为模型训练的阳性集。同时，为保证风险预测模型训练数据的均衡性，我们又挑选出了 7000 名正常个体作为阴性集。两部分数据共同构成本次研究所使用的数据集。

表 2-2 非患者评判依据

ICD 编码	分类
M11-X	软骨钙化症
M20-X、M21-X	获得性关节畸形
M22-X	髌骨功能异常
M23-X	膝关节异常
M24-X	其他关节异常
M25-X	关节疼痛
M42-X	脊柱软骨病

2.3 基因型位点质控

人类基因组具有复杂性与多样性，本研究中的每个样本个体基因型数据都由数十万乃至数百万个基因型位点组成。由于目前任何算法都无法短时间内对如此多的样本进行处理，因此需要从其中选取部分具有统计学意义的位点。本文根据目前已发表的骨关节炎的 GWAS 结果 [9, 10] 对样本的基因型位点进行关联显著值（P-Value）与等位基因频率（Allele Frequency, AF）质控。

关联显著值是用来衡量 GWAS 研究中单核苷酸突变与给定性状关联显著性的一个统计量。在 GWAS 研究中，我们设定零假设为数据中没有 SNP 位点与特定性状相关联；而备择假设为至少有一个 SNP 位点同特定性状相关联。同时我们定义统计量 p 为当零假设为真时观察到该关联的概率。显然， p 值越小，有越高把握在观察到关联时认为零假设为假。因此我们可以设定一个阈值，当统计量 p 低于该阈值时拒绝原假设，认为该位点同目标性状相关。[38] 目前 GWAS 研究中对位点统计显著值的描述可通过 Manhattan 图进行，Manhattan 图的横坐标为 SNP 位点在基因组中的坐标，纵坐标为对应 SNP 位点在 GWAS 研究中的统计显著值。本研究所选区的 SNP 位点 Manhattan 图如图2-1所示。

但是，由于遗传漂变的影响，随机的基因突变也有可能具有类似相关关系。因此我们并不能认为 p 值较小的位点同目标位点直接相关，在指定阈值时需要将可能由遗传漂变所产生的位点筛去。因此我们还需要通过 QQ 图来完成阈值界定。QQ 图是一种根据分位数对两概率分布所作的图，该图来比较两分布差别的方法。[39] GWAS 中的 QQ 图横轴为遗传漂变所产生的分布，纵轴为实际分布，当图像偏离对角线时认为这类数据并非随机漂变。因此，根据 QQ 图??本文确定 p 阈值为 10^{-5} 。

等位基因频率也是进行基因型数据质控时常见的评判指标。在 GWAS 研究中往往会出现一些出现频率很低的突变。受限于 GWAS 原理，大多数研究不能很好计算低等位频率位点与性状的相关性。因此在实际研究时，还需要通过等位基因频率对所得位点进行进一步质控。根据文献我们将该阈值设定为 0.05[40]。

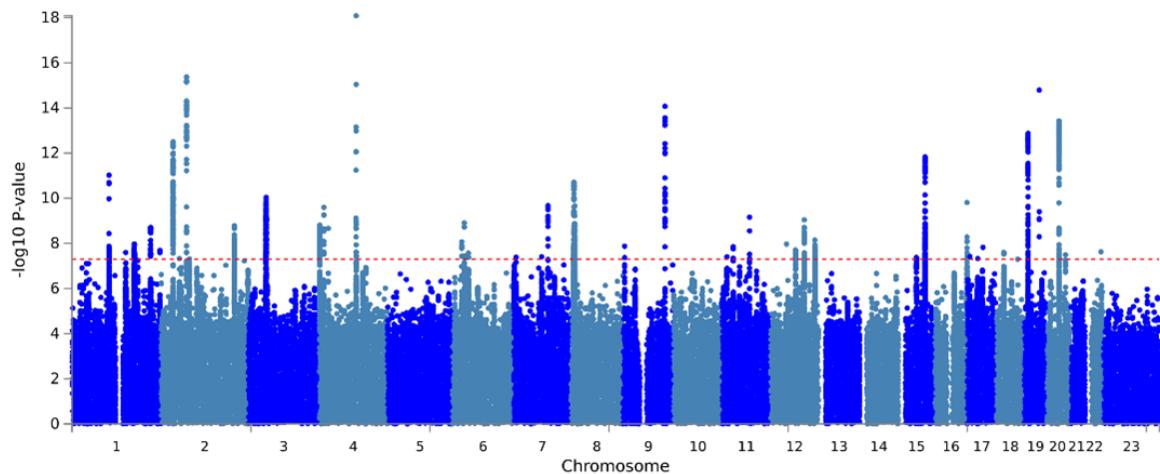


图 2-1 GWAS 研究的 Manhattan 图

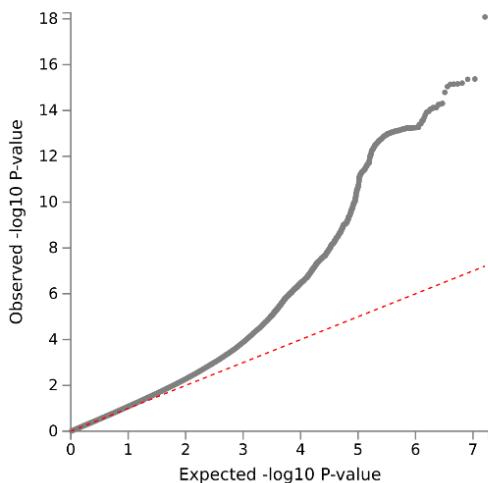


图 2-2 GWAS 研究的 Q-Q 图

综上，本文根据关联分析显著值与等位基因频率对所得 SNP 位点进行筛选，最终每个个体选取选取 8687 个 SNP 位点作为待分类基因型位点。同时，我们根据位点目标突变的出现频率对数据进行编码，将基因型信息转化为算法可直接读取的数值信息。

2.4 特征筛选

目前我们已经从 UKB 数据库中选取了 6706 位骨关节炎患者作为阳性样本，每个患者包括 8687 个基因型特征。考虑到特征数大于样本数，可能对预测模型的准确率产生不利影响。^[41] 本文还使用卡方法与支持向量机法对现有数据进行了特征筛选

2.4.1 基于卡方的特征筛选

卡方法使用统计学方法对输入特征与样本标注的关联进行计算，并根据计算结果筛选与样本标注相关程度较高的特征。其具体过程如下：

1. 确定零假设与备择假设：对于特征与样本标注，定义零假设 H_0 为该特征与样本标注无关；定义备择假设 H_1 为该特征与样本特征相关
2. 计算特征卡方值：根据公式计算特征与样本标注的卡方值

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (2-1)$$

其中：

- c 为该卡方分布的自由度
- O 为观测值
- E 为期望值

3. 查卡方表，拒绝或接受假设：对于计算所得卡方值，查卡方表。对照计算所得卡方值与一定自由度与置信度下的标准卡方值。当计算卡方值大于标准卡方值时接受零假设，认为该特征与样本标注无关；当计算卡方值小于标准卡方值时拒绝零假设，认为该特征与样本标注相关。基于此选择该特征

2.4.2 基于支持向量机的特征筛选

支持向量机（Support Vector Machine, SVM）是一种著名的机器学习算法。其通过构建数据空间内的分类平面来实现数据分类算法。其算法核心为以下优化过程

$$\min_{W,b} [C \sum_i \max(0, 1 - y_i(X_i^T W + b)) + l(W)] \quad (2-2)$$

其中：

- X_i 为样本数据
- W, b 为超平面参数
- y_i 为样本标注
- C 为惩罚系数，即错误分类情况下对优化函数的惩罚
- $l(W)$ 为超平面拟合程度的评价函数

从上式可以看出，该优化目标主要分为两部分：一部分衡量超平面对数据点的分类能力；另一部分衡量超平面对数据的拟合能力。目前常用的 SVM 算法中

$$l(W) = \frac{1}{2} w^T w \quad (2-3)$$

此时该拟合函数评价分离两类数据点超平面在超空间上的距离。距离越大，证明该超平面对数据的拟合能力越好。但是，当该函数变为

$$l(W) = e^T |w| \quad (2-4)$$

此时该拟合函数又称 w 的 l_1 范数，又称 Lasso 惩罚 [42]，该范数可以用来评价特征在整体分类效果中的贡献 [43]。因此本文可通过该范数对特征进行筛选。

2.5 数据集描述

经过以上步骤，我们制得包括 13706 位个体、每个体选取 200 个基因型位点的基因型数据集。数据集大小 2.3GB。其中患病正样本 6706 例，正常负样本 7000 例。每个基因型位点我们通过 8 个字段描述，字段及其含义见表2-3。

表 2-3 基因型位点字段信息

字段	含义
RSID	SNP ID
Chromosome	SNP 所在染色体
BP	以 GRCh37 为参考的碱基对位置
Effect Allele	与表型相关的效应等位位点
Non-Effect Allele	与表型无关的等位位点
Beta	效应值
P	统计显著性
MAF	等位基因频率

同时为了计算与处理方便，我们对患者基因型位点 SNP 数据根据效应等位位点 E 与非效应等位位点 N 进行编码，编码方式见表2-4

表 2-4 基因型位点编码方式

SNP 信息	编码
EE	(1,1)
EN	(1,0)
NN	(0,0)

2.6 本章小结

本章介绍了本文所使用数据集获取与预处理过程。首先，根据 ICD 编码我们从 UKB 数据库中选取了 6706 名患病个体（阳性集）与 7000 名非骨关节炎个体（阴性集）作为研究对象。其次，根据关联分析显著性及基因频率从已发表的骨关节炎 GWAS 中筛选出感兴趣的位点并获取研究样本相应的基因型数据，之后再根据位点信息对其进行

编码。最后，为了进一步减少无关的样本特征，加快计算速度，本文通过卡方法与支持向量机法对数据进行特征筛选。最终完成了由每个样本包含 200 个基因型信息位点的 13706 样本所构成的骨关节炎风险预测模型基因型数据集的构建。

3 骨关节炎风险预测模型构建

第二章我们获取了骨关节炎患者的数据并对其进行了预处理与特征筛选。本章则主要对骨关节炎风险预测模型做以介绍。该模型首先根据基于变分期望最大化算法的图估计器构建了输入基因型数据的关联并通过图来表示该关联关系。之后构建了图神经网络对建立的图进行学习与处理并最终完成图的分类。该模型还通过融合患者表型进一步增强模型性能。最后本模型还构建了基于预测值与输入数据的模型解释器。

3.1 图估计器

3.1.1 图神经网络中无结构数据的处理方法

图神经网络只适用于图类型的数据，对常见无结构数据（例如文本，图像）的处理较为困难。本文目前获得的基因型数据也属于无结构数据，原理上不适用于图神经网络。但是，考虑到图神经网络优秀的性能与可解释性，我们需要基于该无结构基因型数据构建图，继而输入图神经网络。目前根据无结构数据建图的方法主要有两种方法：一种是通过分析数据构建静态图，神经网络训练过程中不对该图结构加以更新 [44, 45]，例如在处理无结构的图像数据时，Monti[45] 提出了一种通过超像素将图像数据转为图的方法，使得图神经网络能够处理图像数据；第二种是通过学习方法构建动态图结构，并在网络训练过程中不断更新图结构 [46]。这类通过学习方法获取图结构特征的过程也被称为图结构学习（Graph Structure Learning, GSL）。目前的图结构学习过程主要分为三个步骤：首先通过已知数据通过 k 近邻法 [47] 或者阈值法 [48] 构建出一个初始图结构或观测图结构；然后使用诸如随机块模型（Stochastic Block Model, SBM）等方法 [49] 对该观测图结构进行建模；最后通过算法估计出数据可能具有的图结构。在图估计的过程中，基于统计学的算法受到了广泛关注：例如 Zhang[50] 等人提出了一种基于贝叶斯思想的图估计器。其认为图结构根据依赖一定参数的分布所产生，如果通过蒙特卡洛方法对该参数进行估计，就可因此得到该分布的具体信息，继而对图结构进行估计。Elinas[51] 同样基于贝叶斯思想提出了一种使用变分推断法的图估计方法。但是，以上研究仅适用于图神经网络中的节点分类问题，同本研究所属图分类问题不符，不能直接应用于本研究中。但考虑到这类贝叶斯方法思路简洁，原理清晰。因此，本文也将基于贝叶斯思想构建适用于本文图分类问题的图结构估计器。

3.1.2 问题描述

贝叶斯思想认为：事件的观测值并不能反映事件的真实特性。贝叶斯方法因此主要解决由事件观测值向事件真实值的推断过程。但是在讨论推断过程前，我们首先需要对观测值与真实值加以定义。

3.1.2.1 观测值构建

对于本文研究对象基因型数据而言，未经图神经网络处理的原始数据所含信息量较少，较难依据初始数据构建观测值。但是有研究显示，经过谱图神经网络处理之后，具有较强关联的节点具有相似的值。^[52]我们因此认为在图神经网络的输出值中，对任一节点，如果有一其他节点与其有着相似的值，则两节点之间可能存在关联。这种关联可以通过 k 近邻算法^[47]计算从而得到 k 近邻网络，而该 k 近邻网络可以很好地描述节点的局部特征。在 Wang^[53]等人的研究中，也有通过根据图神经网络输出值构建 k 近邻网络的描述。我们因此根据每个样本经图神经网络处理后的输出值建立 k 近邻网络，再将所有网络相加求平均以获得单个观测值。同时为了防止信息丢失，我们将首次经图神经网络处理时输入的图初始化为全连接图，即认为每个节点都同其他节点存在关联。

3.1.2.2 真实值构建

本研究中所感兴趣的真实值为实际的未知图结构。但是对图结构的直接计算过于复杂，我们因此需要将图结构参数化，即使用少数几个参数来描述图结构。目前在图论领域常用的图参数化模型为随机块模型^[49]，该模型认为图中的节点属于某几个簇，节点之间是否相连仅与两节点所在的簇相关且服从某一参数化分布。因此我们只需要得出支配该分布的参数就能对图结构加以复现，继而用少数参数描述庞大的网络。因此本文将该类参数作为推断的目标，而由其产生的图结构作为模型输出值。

3.1.2.3 推断过程

界定了观测值与真实值之后，我们便可对图估计器所解决的问题加以严谨描述。

图估计器首先根据式3-1有图神经网络的输出构建观测值 O 。

$$O^{ij} = \frac{1}{N} \sum_N o_n^{ij} \quad (3-1)$$

其中 O^{ij} 描述了节点 i, j 之间关联的观测值， N 为样本数， o_n^{ij} 为由第 n 个样本神经网络输出值构建的 k 近邻网络中节点 i, j 之间的关联。通过该观测值我们构建图 $G = (V, O)$ ，其中 $V = \{v_1, \dots, v_n\}$ 描述图中节点，即样本 SNP 位点。

为了通过 SBM 模型参数化图结构，我们假定图中共含有 C 个簇，每个节点 i 从属于簇 m 的概率由式3-2来描述，且 z_i 相互独立并满足由式3-3所描述的多项分布。

$$z_{im} \in \mathbf{Z}^{N \times C}, i \in \{1, \dots, n\}, m \in \{1, \dots, C\} \quad (3-2)$$

$$z_i \stackrel{iid}{\sim} M(1, \alpha) \quad (3-3)$$

基于此，我们使用 z_i 来描述节点之间关联。我们认为任意两节点之间的关联相互独立并满足如式3-4所描述二项分布。

$$o_{ij} | z_{im}, z_{jn} = 1 \sim B(1, \pi_{z_i z_j}) \quad (3-4)$$

可以看出，只要解出参数 π 与 α ，就可以此构建出如图3-1所示的 SBM 模型，继而对图结构加以推断；因此本文给出以下目标：

Representation of Stochastic Block Model

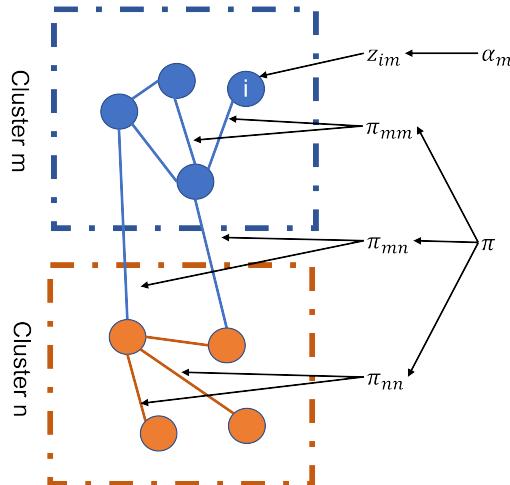


图 3-1 随机块模型及其参数化示意图

基于观察到的节点关联信息 O 以及隐变量 Z ，通过式3-5估计参数 $\theta = \{\alpha, \pi\}$ ，并根据该参数给出推断图结构

$$\hat{\theta} = \arg \max l_\theta(O) \quad (3-5)$$

3.1.3 基于变分期望最大化的图估计

3.1.3.1 推导与整理

在节点关联与节点簇信息均已知的情况下，我们首先由式3-6计算似然函数 l_θ ，根据边缘概率公式，有

$$l_\theta(O) = \sum_Z \log p_\theta(O, Z) = \sum_Z (\log p_\pi(O|Z) + \log p_\alpha(Z)) \quad (3-6)$$

其中

$$\sum_Z (\log p_\pi(O|Z) + \log p_\alpha(Z)) = \sum_Z \left(\sum_{i,j}^n \sum_{m,n}^C \log p_\pi(o_{ij}|z_{im}z_{jn}) + \sum_i^N \sum_m^C \log p_\alpha(z_{im}) \right) \quad (3-7)$$

给定参数时，有

$$p_\pi(o_{ij}|z_{im}z_{jn}; \pi) = \binom{1}{O_{ij}} [\pi_{mn}(1 - \pi_{mn})^{1-O_{ij}}]^{z_{im}z_{jn}} \quad (3-8)$$

$$p_\alpha(z_{im}) = \alpha_m^{z_{im}} \quad (3-9)$$

将式3-8与3-9代入式3-7可得式3-10。

$$\begin{aligned} l_\theta(O; \theta) &= \sum_Z \left(\sum_{i,j}^n \sum_{m,n}^C \log p_\pi(o_{ij}|z_{im}z_{jn}) + \sum_i^N \sum_m^C \log p_\alpha(z_{im}) \right) \\ &= \sum_Z \left(\sum_{i,j}^n \sum_{m,n}^C \log \binom{1}{O_{ij}} [\pi_{mn}(1 - \pi_{mn})^{1-O_{ij}}]^{z_{im}z_{jn}} + \sum_i^N \sum_m^C \log \alpha_m^{z_{im}} \right) \end{aligned} \quad (3-10)$$

3.1.3.2 期望最大化算法

考虑到似然函数中含隐变量 Z ，通常我们采用期望最大化算法 [54] 计算使得似然函数取极大的参数。我们对似然函数式3-6做如下处理

$$\begin{aligned} l(O; \theta) &= \log p(O; \theta) \\ &= \log \int_Z p(O, Z; \theta) dz \\ &= \log \int_Z q(Z) \frac{p(O, Z; \theta)}{q(Z)} dz \\ &= \log \mathbb{E}\left[\frac{p(O, Z; \theta)}{q(Z)}\right] \end{aligned} \quad (3-11)$$

其中 $q(Z)$ 为一辅助函数且满足 $\int q(Z) = 1$ 。根据 Jensen 不等式 [55]，对任意一凸函数，有

$$\mathbb{E}[f(x)] \leq f(\mathbb{E}[x]) \quad (3-12)$$

当且仅当 $x = \mathbb{E}[x]$ 时等号成立。又因为 $\log x$ 为一凸函数，因此将式3-12代入式3-11，有

$$\begin{aligned}
 l(O; \theta) &= \log p(O; \theta) \\
 &= \log \mathbb{E}\left[\frac{p(O, Z; \theta)}{q(Z)}\right] \\
 &\geq \mathbb{E}\left[\log \frac{p(O, Z; \theta)}{q(Z)}\right]
 \end{aligned} \tag{3-13}$$

当且仅当式3-13满足式3-14所示条件时等号成立。

$$\frac{p(O, Z; \theta)}{q(Z)} = c \tag{3-14}$$

对式3-14进行变化，有：

$$q(Z) = c \cdot p(O, Z; \theta) \tag{3-15}$$

又因为 $\int q(Z) dz = 1$ ，有：

$$\begin{aligned}
 \int_Z q(Z) dz &= \int_Z c \cdot p(O, Z; \theta) dz = 1 \\
 c &= \frac{1}{\int_Z p(O, Z; \theta) dz}
 \end{aligned} \tag{3-16}$$

因此：

$$q(Z) = \frac{p(O, Z; \theta)}{\int_Z p(O, Z; \theta) dz} = \frac{p(O, Z; \theta)}{p(O; \theta)} = p(Z|O; \theta) \tag{3-17}$$

将式3-17代回原不等式3-13，我们由式3-18构造辅助函数 Q 。

$$l(O; \theta) \geq \mathbb{E}\left[\log \frac{p(O, Z; \theta)}{q(Z)}\right] = \int_Z p(Z|O; \theta) \log \frac{p(O, Z; \theta)}{p(Z|O; \theta)} dz = Q(\theta^{t+1}; \theta) \tag{3-18}$$

显然，在取等条件成立时， Q 为似然函数 $l(O; \theta)$ 的下界。期望最大化算法因此首先计算给定预期参数 θ 时下界 Q 的值，再通过优化参数 θ 提高下界 Q ，通过不断迭代计算使得似然函数极大的参数 $\hat{\theta}$ 。具体步骤描述如下：

- E 步骤：计算 $Q(\theta^{t+1}; \theta)$
- M 步骤：优化参数 θ

但在下界 Q 的计算过程中我们由式3-18设 $q(Z) = p(Z|O; \theta)$ ，但是在本问题中， $p(Z|O; \theta)$ 过于复杂，无法计算，因此也无法完成后续步骤。我们需要采取其他方法来估计参数 θ 。

3.1.3.3 变分期望最大化

3.1.3.3.1 变分推理 变分推理是贝叶斯统计领域遇到无法计算的分布时常见的处理方法 [56]。它通过构建辅助函数来模拟无法计算的分布，并通过优化过程使得辅助函数逐渐逼近真实分布。[57] 该方法与本文所处理问题较为契合，我们因此结合变分推理来处理期望最大化过程中无法计算的分布 $p(Z|O; \theta)$ 。

3.1.3.3.2 证据下限与 KL 散度 在进一步解释之前我们需要先对可能使用到的概念加以说明。在上一节中我们已经证明，根据 Jensen 不等式，含隐变量的似然函数具有一下界。我们将该下界定义为似然函数的证据下界 (Evidence Lower Bound, ELBO)[58]。其具有如式3-19所示形式。

$$\begin{aligned} l(O; \theta) &= \log \int_Z q(Z) \frac{p(O, Z; \theta)}{q(Z)} dz \\ &\geq \int_Z q(Z) \log \frac{p(O, Z; \theta)}{q(Z)} dz \\ &= J(q(Z); \theta) \end{aligned} \tag{3-19}$$

我们同时根据式3-20定义似然函数与证据下界之间的差为分布 $q(Z)$ 与分布 $p(Z|O)$ 之间的 KL 散度 (Kullback-Leibler divergence)。[59]

$$\begin{aligned} l(O; \theta) - \int_Z q(Z) \log \frac{p(O, Z; \theta)}{q(Z)} dz &= \int_Z q(Z) \log p(O; \theta) dz - \int_Z q(Z) \log \frac{p(O, Z; \theta)}{q(Z)} dz \\ &= \int_Z q(Z; \theta) \log \frac{p(O; \theta)q(Z; \theta)}{p(O, Z; \theta)} dz \\ &= - \int_Z q(Z; \theta) \log \frac{p(Z|O; \theta)}{q(Z; \theta)} dz \\ &= KL(q(Z)||p(Z|O)) \end{aligned} \tag{3-20}$$

可以看出，该散度衡量了两分布间的差异，可以用于后续辅助函数的构造。

3.1.3.3.3 辅助函数构建 基于变分推理思想，我们构建辅助分布 $q_\psi(Z)$ 来模拟无法计算的 $p(Z|O)$ 。基于期望最大化思想，我们构建如式3-21所示似然函数。

$$l(O; \theta) = J(q_\psi(Z); \theta) + KL(q_\psi(Z)||p(Z|O)) \tag{3-21}$$

式3-21中首项描述了在使用辅助分布替换 $p(Z|O)$ 时原始似然函数的下界，第二

项描述了辅助分布与原始分布之间的差值。值得注意的是，KL 散度具有非负性，即 $KL(\cdot||\cdot) \geq 0$ 。我们因此给出如式3-22所示不等式。

$$l(O; \theta) \geq J(q_\psi(Z); \theta) \quad (3-22)$$

借助期望最大化思想，我们如果能够通过迭代计算辅助函数与参数使得式3-22中下界 $J(q_\psi(Z); \theta)$ 不断升高，就能使得该似然函数最终取最大。展开该下界，得式3-23。

$$\begin{aligned} J(q_\psi(Z); \theta) &= \int_Z q(Z) \log \frac{p(O, Z; \theta)}{q(Z)} dz \\ &= \int_Z q(Z) \log p(O, Z; \theta) dz - \int_Z q(Z) \log q(Z) dz \\ &= \int_Z q(Z) \left(\sum_{i,j}^n \sum_{m,n}^C \log \binom{1}{O_{ij}} [\pi_{mn}(1 - \pi_{mn})^{1-O_{ij}}]^{z_{im}z_{jn}} + \sum_i^N \sum_m^C \log \alpha_m^{z_{im}} \right) dz \\ &\quad - \int_Z q(Z) \log q(Z) dz \\ &= \mathbb{E}_{q_\psi(Z)} \left[\sum_i^N \sum_m^C z_{im} \log \alpha_m \right] - \sum_i^N \sum_m^C \mathbb{E}_{q_\psi(Z)} [z_{im}] \log \mathbb{E}_{q_\psi(Z)} [z_{im}] \\ &\quad + \mathbb{E}_{q_\psi(Z)} \left[\sum_{i < j}^N \sum_{m,n}^C z_{im} z_{jn} \log \left(\binom{1}{O_{ij}} \pi_{mn}(1 - \pi_{mn})^{1-O_{ij}} \right) \right] \end{aligned} \quad (3-23)$$

此时，我们构造基于服从参数 τ 的多项分布的辅助函数，使得该辅助函数满足式3-24。

$$q_\psi(Z) = \prod_i^N q_\psi(Z_i) = \prod_i^N M(Z_i, \tau) \quad (3-24)$$

对于该变分参数 τ ，我们依据多项分布的性质给出如下限制条件 $\tau \in [0, 1]^C, \sum_m \tau_{im} = 1$ 。可以看出，该参数同原始隐变量 Z 的意义相同，都描述了节点属于簇的概率。因此我们给出如式3-25与3-26所示关系。

$$\tau_{im} = p(q_\psi(z_{im} = 1)) = \mathbb{E}_{q_\psi(Z)} [z_{im}] \quad (3-25)$$

$$\tau_{im}\tau_{jn} = p(q_\psi(z_{im} = 1, z_{jn} = 1)) = \mathbb{E}_{q_\psi(Z)} [z_{im}z_{jn}] \quad (3-26)$$

将式3-25与3-26代入式3-23，有式3-27所示形式。

$$\begin{aligned} J(O; \theta) = & - \sum_i \sum_m \tau_{im} \log \tau_{im} + \sum_i \sum_m \tau_{im} \log \alpha_m \\ & + \sum_{i < j} \sum_{m,n}^C \tau_{im} \tau_{jn} \log \left(\binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) \end{aligned} \quad (3-27)$$

此时我们可以发现，虽然该下界的形式较为复杂，但是在给定辅助函数与参数的情况下依然能够给出结果，我们采用变分法的目的也就达到了。因此我们设计如下变分期望最大化算法。

- VE 步骤：固定参数 θ ，根据 J 更新变分参数 τ ，并计算更新参数后下界 J 的值
- M 步骤：固定变分参数 τ ，根据 J 更新参数 θ ，直至收敛。

3.1.3.3.4 参数求解

VE 步 此步骤中主要基于式3-28根据式3-33求解变分参数 τ 。

$$\hat{\tau} = \arg \max_{\tau} J(O, \theta; \tau) \quad (3-28)$$

$$\begin{aligned} J(O, \theta; \tau) = & - \sum_i \sum_m \tau_{im} \log \tau_{im} + \sum_i \sum_m \tau_{im} \log \alpha_m \\ & + \sum_{i < j} \sum_{m,n}^C \tau_{im} \tau_{jn} \log \left(\binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) \end{aligned} \quad (3-29)$$

考虑到式3-33中的参数存在限制条件 $\sum_m \tau_{im} = 1$ ，我们采用拉格朗日乘数法 [60] 对该式加以计算得3-30。

$$\begin{aligned} L(J, \lambda) = & J(O, \theta; \tau) + \lambda_i (\sum_m \tau_{im} - 1) \\ = & - \sum_i \sum_m \tau_{im} \log \tau_{im} + \sum_i \sum_m \tau_{im} \log \alpha_m + \lambda_i (\sum_m \tau_{im} - 1) \\ & + \sum_{i < j} \sum_{m,n}^C \tau_{im} \tau_{jn} \log \left(\binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) \end{aligned} \quad (3-30)$$

将式3-30分别对 τ_{im} 与 λ 求偏导，得式3-31与3-41。

$$\frac{\partial L}{\partial \tau_{im}} = \sum_j^N \sum_n^C \tau_{jn} \log \left(\binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) + \log \alpha_m - \log \tau_{im} + \lambda_i - 1 = 0 \quad (3-31)$$

$$\frac{\partial L}{\partial \lambda_i} = \tau_{im} - 1 = 0 \quad (3-32)$$

对式3-31加以变换，我们得到式3-33。

$$\log \tau_{im} = \sum_j^N \sum_n^C \tau_{jn} \log \left(\binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) + \log \alpha_m - \lambda_i + 1 \quad (3-33)$$

我们因此基于式3-33给出 τ_{im} 的解析解

$$\begin{aligned} \tau_{im} &= e^{-\lambda_i+1} \alpha_m \left[\prod_j^N \prod_n^C \left[\binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right]^{\tau_{jn}} \right] \\ &\quad \forall i \in \{1 \dots N\}; \forall m \in \{1 \dots C\} \end{aligned} \quad (3-34)$$

可以看出，每个变分参数的计算都依赖于其他变分参数，虽然可以通过迭代至收敛的方法计算，但这也对参数初始值的选取提出了要求，具体细节将于后文讨论。

M步 此步骤中主要基于式3-35求解目标参数 θ 。

$$\hat{\theta} = \arg \max_{\theta} J(O, \tau; \theta) \quad (3-35)$$

首先我们根据式3-36计算 θ 中的 π 。由于 π 在式中相对较为独立，我们直接对其求偏导得式3-37。

$$\begin{aligned} J(O, \tau, \alpha; \pi) &= - \sum_i \sum_m \tau_{im} \log \tau_{im} + \sum_i \sum_m \tau_{im} \log \alpha_m \\ &\quad + \sum_{i < j} \sum_{m, n} \tau_{im} \tau_{jn} \log \left(\binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) \end{aligned} \quad (3-36)$$

$$\frac{\partial J(O, \tau, \alpha; \pi)}{\partial \pi_{mn}} = \sum_{i < j} \tau_{im} \tau_{jn} \left(\frac{O_{ij}}{\pi_{mn}} - \frac{1 - O_{ij}}{1 - \pi_{mn}} \right) \quad (3-37)$$

求解以上方程，得如式3-38所示的 π 的解析解。

$$\hat{\pi}_{mn} = \frac{\sum_{i < j} \tau_{im} \tau_{jn} O_{ij}}{\sum_{i < j} \tau_{im} \tau_{jn}} \quad (3-38)$$

对于 θ 中的 α ，由于其也存在限制条件 $\sum_m \alpha_m = 1$ ，我们依旧采用拉格朗日乘数法计算得式3-39。

$$\begin{aligned}
 L(J, \lambda) &= J(O, \pi, \tau; \alpha) + \lambda_i (\sum_m \alpha_m - 1) \\
 &= - \sum_i \sum_m \tau_{im} \log \tau_{im} + \sum_i \sum_m \tau_{im} \log \alpha_m + \lambda_i (\sum_m \alpha_m - 1) \\
 &\quad + \sum_{i < j}^N \sum_{m,n}^C \tau_{im} \tau_{jn} \log \left(\binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right)
 \end{aligned} \tag{3-39}$$

将该式分别对 α_m 与 λ 求偏导，得式3-40与式3-41。

$$\frac{\partial L}{\partial \alpha_m} = \frac{\sum_i \tau_{im}}{\alpha_m} + \lambda_i = 0 \tag{3-40}$$

$$\frac{\partial L}{\partial \lambda_i} = \alpha_m - 1 = 0 \tag{3-41}$$

对式3-40加以处理，得：

$$\begin{aligned}
 \frac{\sum_i \tau_{im}}{\alpha_m} + \lambda_i &= 0 \\
 \sum_i \tau_{im} + \alpha_m \lambda_i &= 0 \\
 \sum_m \sum_i \tau_{im} + \sum_m \alpha_m \lambda_i &= 0 \\
 \lambda_i &= - \sum_m \sum_i \tau_{im}
 \end{aligned} \tag{3-42}$$

代入式3-40，得如式3-43所描述的 α 的解析解。

$$\begin{aligned}
 \hat{\alpha}_m &= \frac{\sum_i \tau_{im}}{\sum_m \sum_i \tau_{im}} \\
 &= \frac{\sum_i \tau_{im}}{N}
 \end{aligned} \tag{3-43}$$

至此，我们已经完成了所有参数的估计。确定参数后只需将其代回原始模型即可得出估计的节点之间关联，继而完成图结构的构建。

3.1.3.4 起始与终止条件

之前在对变分参数 τ 的计算中我们发现，任一参数的计算都依赖于其他参数。但是如果以零为初始值，参数在计算时只会考虑项 $e^{-\lambda_i+1} \alpha_m$ ，无法对参数 π 形成有效估计。因此我们需要通过一定方法给予该变分参数初值，并通过迭代的方法计算 τ 直至收敛。目前的研究中常使用 k-Means 法 [61] 先对节点进行聚类，根据聚类结果对变分参数进行赋值。但是 k-Means 法需要预先确定图中节点簇数 C ，且对于一般问题而言

该簇数未知，因此需要一种能够评估所选簇数对最终似然函数影响的方法。目前研究中，一种叫做贝叶斯信息指标（Bayesian Information Criterion, BIC）[62] 的方法可以用来完成类似工作，其定义为式3-44。

$$BIC(C) = \log P(O; \theta) - \frac{V_C}{2} \log N \quad (3-44)$$

其中 $\log P(O; \theta)$ 描述对应模型下的似然函数， V_C 描述了选择对应簇数 C 时模型参数数量，但是该方法涉及到无法求解似然函数的计算，在本研究中无法实现。但是，基于类似的思想，Daudin 等人提出另一种评价指标：整合分类似然（Integrated Classification Likelihood, ICL）[63]，该指标并不直接计算原始似然函数，而是通过计算变分时使用的变分似然函数给出评判依据。该指标定义如式3-45。

$$ICL(C) = \sum_i \sum_m \tau_{im} \log \tau_{im} + \sum_{i < j}^N \sum_{m,n}^C \tau_{im} \tau_{jn} \log \left(\binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) - \frac{1}{2} \left(\frac{C(C+1)}{2} \log \frac{N(N-1)}{2} - (C-1) \log N \right) \quad (3-45)$$

基于以上信息，我们给出图3-2描述的基于变分期望最大化的图估计算法全流程。

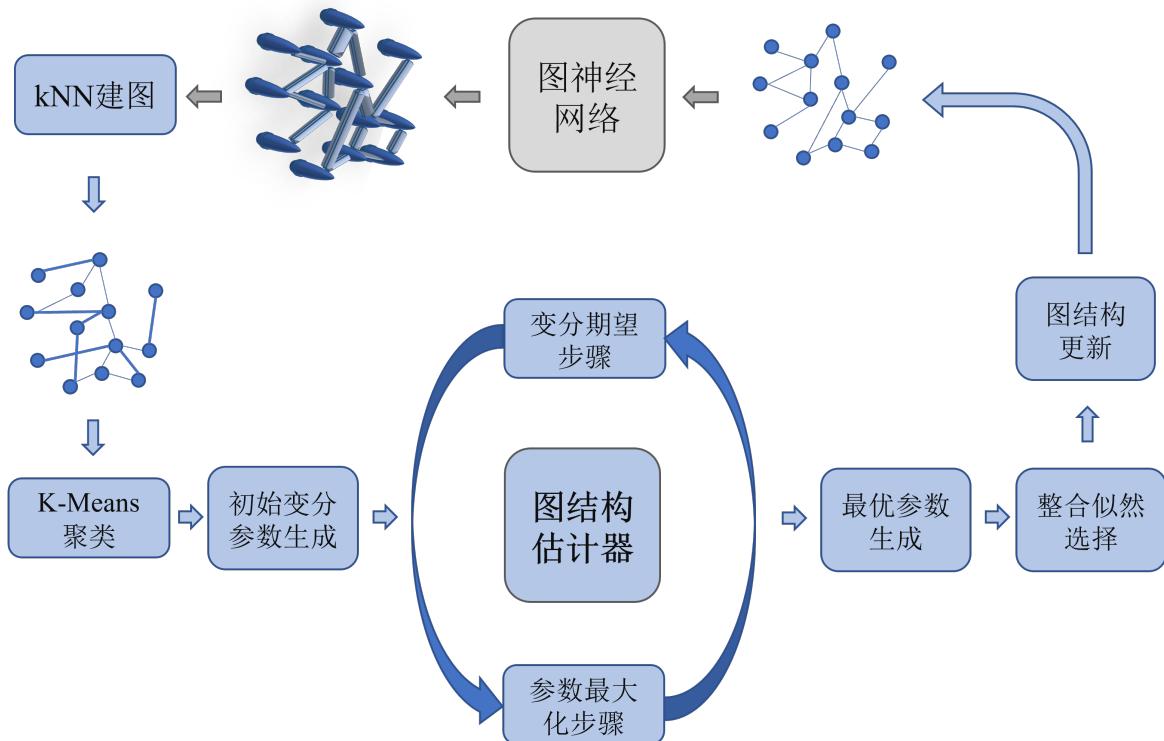


图 3-2 基于变分期望最大化的图估计算法流程

Data: 观测值 O 与聚类数目 c

Result: 模型参数 θ

Input: 观测值, 确定簇数范围

1 使用 k-Means 法对输入值进行聚类, 给出初始变分参数 τ_0

2 **while** θ 不收敛 **do**

3 VE 步

4 **while** 变分参数 τ 不收敛 **do**

5 | 循环计算 $\hat{\tau} = \arg \max_{\tau} J(O, \theta; \tau)$

6 **end**

7 根据该变分参数计算 J

8 M 步

9 根据 $\hat{\theta} = \arg \max_{\theta} J(O, \tau; \theta)$ 计算参数 θ

10 计算最优似然下界 \hat{J} 与该簇数下的整合分类似然 $ICL(C)$

11 **end**

Output: 选择使得整个分类似然最小的簇数 C , 选择该簇数计算出的参数 $\hat{\theta}$

算法 3-1 基于变分期望最大化的图估计算法

3.2 图神经网络

在完成了输入图结构的估计后, 我们便可开始设计模型中图处理的核心——图神经网络。基于绪论中对图神经网络的讨论与综述, 本文决定使用谱图神经网络中的切比雪夫层来处理生成的图数据。

3.2.1 基本定义

为了高效处理图数据, 我们将图估计器所生成的节点关系通过矩阵表示, 这类描述图节点之间关联的矩阵被称为邻接矩阵 (Adjacency Matrix), 用 A 来表示

$$A_{ij} = \{0, 1\}^{N \times N} \quad (3-46)$$

$A_{ij} = 1$ 意味着节点 i, j 之间存在相关, 反之 $A_{ij} = 0$ 意味着两节点之间无相关关系。我们同时定义无向图中节点的度 (Degree) 为与该节点相关节点的数量并定义图的度矩阵 D 为

$$D_{ii} = \sum_{i \neq j} A_{ij} \quad (3-47)$$

可以看出, 该度矩阵为一对角矩阵。同时为了后续矩阵的特征分解操作, 我们定义图的拉普拉斯矩阵 L 为度矩阵与邻接矩阵之差, 即

$$L = D - A \quad (3-48)$$

3.2.2 切比雪夫层

切比雪夫网络是一种谱图神经网络，其通过卷积方式来从数据中学习复杂信息 [64]。但是同常见的卷积神经网络不同，谱图神经网络通过傅里叶变换的方式来实现图这种非欧数据上的卷积。根据卷积定理 [65]，两个函数的卷积为他们傅里叶变换后的结果的点积的傅里叶逆变换，即

$$x * y = F^{-1}\{F\{x\} \cdot F\{y\}\} \quad (3-49)$$

因此如果我们能够在图上定义一傅里叶变换，我们就能以此定义图数据上的卷积。通常来说，函数的傅里叶变换是函数在拉普拉斯算子特征函数这一组标准正交基上的投影。类似地，我们根据图的拉普拉斯矩阵定义该变换。

$$Lu = \lambda u \quad (3-50)$$

其中 L 为图的拉普拉斯量， u 为该正交基， λ 为特征值。我们同时定义矩阵 $U = [u_1, \dots, u_N]$ ，使得

$$L = U^T \Lambda U \quad (3-51)$$

因此我们定义图 ϕ 上的傅里叶变换为

$$\begin{aligned} \hat{\phi} &= U^T \phi \\ &= \text{diag}(\hat{\phi}(\lambda_l)) \\ &= \text{diag}\left(\sum_i^N \phi_i u_{li}\right) \end{aligned} \quad (3-52)$$

因此根据卷积定理，对于滤波器 g 与图信号 x 而言，在图 G 上定义的卷积为

$$\begin{aligned} (g * x)_G &= U(U^T g \cdot U^T x) \\ &= U g_\theta(\Lambda) U^T x \\ &= g_\theta(L)x \end{aligned} \quad (3-53)$$

其中 g_θ 为卷积核在图中的傅里叶变换， θ 为该卷积核的参数。因此我们可以定义谱图神经网络中的一层为

$$y = \sigma(g_\theta(L)x) \quad (3-54)$$

但是，此时 g_θ 较难确定。Deffend[65] 等人依靠切比雪夫多项式对 g_θ 如式3-55进行了近似。

$$g_\theta(\Lambda) = \sum_{k=0}^K \theta_k T_k(\tilde{\Lambda}), \quad \tilde{\Lambda} = 2\Lambda_n/\lambda_{max} - I_n \quad (3-55)$$

其中 θ 代表需要在训练中学习的参数， T_k 代表 k 阶切比雪夫行列式， $\tilde{\Lambda}$ 代表特征矩阵。其中切比雪夫行列式计算方式如式3-56。

$$\begin{aligned}
 T^{(0)} &= X \\
 T^{(1)} &= \tilde{L}X \\
 T^{(k \geq 2)} &= 2 \cdot \tilde{L}T^{(k-1)} - T^{(k-2)},
 \end{aligned} \tag{3-56}$$

因此切比雪夫网络中的一层表示为式3-57所示形式。

$$Y = \sigma \left(\sum_{k=0}^{K-1} \theta_k T_k(\tilde{L})x \right) \tag{3-57}$$

3.2.3 结构设计

基于切比雪夫层的特点，我们构建如表3-1的图神经网络。

表 3-1 图神经网络结构

层	功能	超参数	输出形式
切比雪夫层	图卷积	$k=1, \text{Channel}=64, \text{activation}=\text{'ReLU'}$	
切比雪夫层	图卷积	$k=1, \text{Channel}=64, \text{activation}=\text{'ReLU'}$	
读出层	将图卷积结果转为矩阵	-	(3200)
全连接层	学习图数据	$N=512, \text{activation}=\text{'ReLU'}$	(256)
Drop-Out 层	抑制网络过拟合	Dropout rate = 0.5	
全连接层	学习图数据	$N=256, \text{activation}=\text{'ReLU'}$	(64)
全连接层	学习图数据	$N=64, \text{activation}=\text{'Sigmoid'}$	(1)

对于切比雪夫层，我们首先选择参数 $k = 1$ ，这意味着该层只取图切比雪夫矩阵的前两项。由式3-56与式3-57我们可知，在该参数条件下，每个节点输出值至于该节点及其相邻节点相关。这与传统卷积神经网络中选取的 (3×3) 卷积核功能类似。同时为了充分学习图中潜藏的信息，我们设定该层的通道数为 16，这意味着同时有 16 个卷积核在一图中学习，增强了模型对潜藏信息的挖掘能力。同时根据文献报道，我们选择如式3-58所示 *ReLU* 函数作为该层的激活函数。[29]

$$\text{ReLU}(x) = \max\{0, x\} \tag{3-58}$$

在对图信息充分学习与处理后，我们将图数据再次通过读出层转化为数据矩阵，准备后续通过全连接层对疾病风险进行预测。考虑到读出层共产生 3200 个特征，我们选择 $512 - 256 - 64$ 的全连接层组合以求最大限度提高模型性能。同时为了防止该规模下模型对数据的过拟合，我们加入 Dropout 层以在训练中随机屏蔽某些神经元。输出层的最后，本模型通过如式3-59所示 *Sigmoid* 激活函数来给出根据患者基因型信息的骨关节炎患病风险。

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (3-59)$$

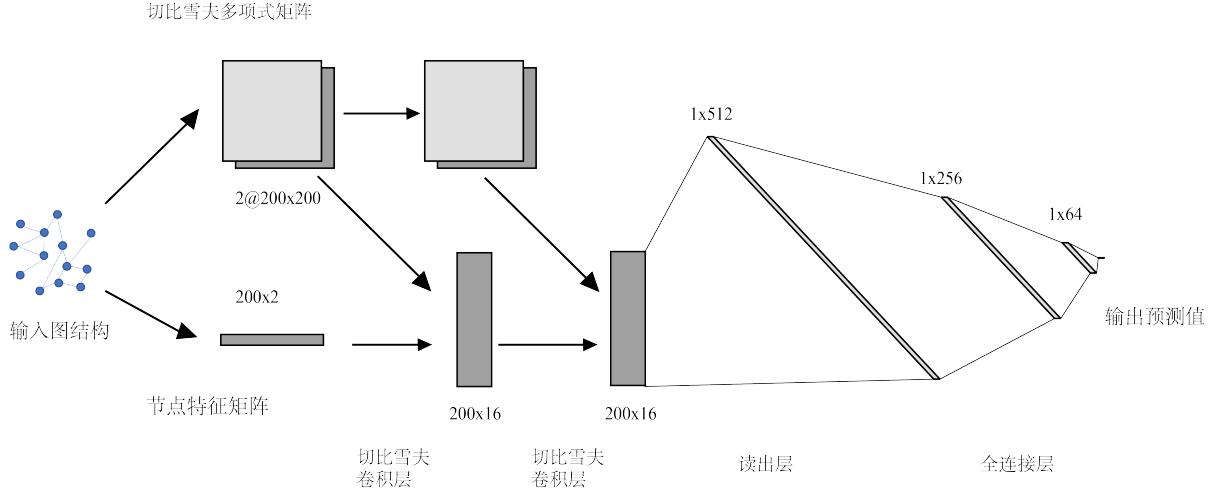


图 3-3 图神经网络架构

在完成了工作过程如图3-3的图神经网络构建后，我们将其通过算法2与已经搭建好的图估计器相融合。使其具备根据输入基因型信息给出风险预测的能力。

Data: 基因型数据 X

Result: 患病风险预测值 y , 最优图结构 G , 神经网络参数 Θ

Input: 基因型数据 X

- 1 依据全连接法构建初始图结构 G
- 2 **while** 未到达设定迭代数 **do**
- 3 神经网络训练
- 4 $\Theta = \arg \min_{\Theta} l(y, G)$
- 5 图估计
- 6 根据图神经网络输出估计可能图结构并更新 G
- 7 **end**

Output: 输出最优图结构及图神经网络参数, 根据该参数给出风险预测

算法 3-2 图估计切比雪夫图神经网络

3.3 表型融合

考虑到骨关节炎作为一种复杂多基因疾病，单纯以基因型为输入可能导致风险预测效果不佳。因此有必要同时融合表型数据。根据 Boer[66] 等人的报道，我们选择同骨关节炎密切相关的表型做为协变量并从 UKB 数据库中获取相关表型数据。具体表型如表3-2。同时我们选择多模态数据融合中的中间融合法 [67] 将基因型数据与表型数据

表 3-2 选取协变量表型

UKBID	表型
32883-31	性别
32883-48	髌围
32883-49	腕围
32883-2463	是否发生过骨折
32883-21001	BMI
32883-21002	体重
32883-21003	年龄

相融合。我们将已构建的图神经网络的中间输出层与所获取的表型数据相融合，并输入新的分类器进行训练，最终给出融合表型后的预测结果。

3.4 图解释器

图是一种包含了节点信息与节点之间关系的数据结构，但是，本文目前为止建立的模型只能直接给出对患病风险的预测，与传统机器学习方法无异。如果能够充分利用图中的关联信息并对产生该预测结果的原因加以解释，不仅能够增强预测结果可信性，还能根据其解释内容发掘数据中潜藏的其他信息。目前对于图神经网络解释器的研究已有颇多进展，本文基于其中备受关注的 GNNExplainer [35] 来构建模型解释器。

GNNExplainer 通过输入的已训练图神经网络模型及其预测结果从输入数据中选择对预测结果影响最大的子图。其数学描述如下

$$\max_{G_s} MI(Y, G_s) = H(Y) - H(Y|G = G_s) \quad (3-60)$$

其中

- MI 表示两变量之间的互信息（Mutual Information），其代表两变量之间相互依赖的程度。
- Y 表示模型给出的预测结果
- G_s 表示同预测结果相关的子图
- H 表示变量的信息熵

该文献描述：对于给定的 GNN 模型， $H(Y)$ 恒定，此时只需要最小化 $H(Y|G = G_s)$ 就可达到解释目的。该文献同时描述了一种变分近似方法来实现该优化过程，最终生成了与预测结果相关的最小子图。

3.5 本章小结

本章描述了骨关节炎风险预测模型的构建过程与基本模块。本文首先针对现存基因型数据结构缺失的问题基于变分期望最大化算法构建了一个图结构估计器。该估计器能根据神经网络的输出动态更新与预测图结构，同时能够对图中的节点聚类分析。其次，本文根据谱图神经网络中的切比雪夫层构建了包裹卷积层、读出层、全连接层在内的图神经网络用于图数据的处理以及患病风险的预测。考虑到单纯通过基因型数据预测患病风险结果可能较差，本文还基于文献描述的同骨关节炎密切相关的表型构建了表型融合风险预测模型。最后，为了尽最大可能挖掘出图数据中的潜藏信息，本文还根据已有的图神经网络解释器构建了预测结果解释器，在给出风险预测结果的同时获取同预测结果相关的子图。以上模块共同构成了本文预期的骨关节炎风险预测模型。

4 结果讨论与分析

本章将对搭建的骨关节炎风险预测模型的图估计效果、患病风险预测能力进行分析与评估。同时将结合具体案例分析模型解释器所产生的解释结果。

4.1 评估指标

为了对模型的风险预测能力进行科学的评估，本文依照真实样本标签与预测结果构建如图4-1混淆矩阵（Confusion Matrix）并根据该矩阵计算如表4-1所示常用评价指标[67]。

		True Class	
		Positive	Negative
Predicted Class	Positive	TP True Positive	FP False Positive
	Negative	FN False Negative	TN True Negative

图 4-1 混淆矩阵构成

表 4-1 混淆矩阵相关指标

指标	计算公式	含义
准确率 Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	模型整体判断准确程度
召回率 Recall	$\frac{TP}{TP+FN}$	正确预测阳性样本占总阳性样本的比例
特异度 Specificity	$\frac{TN}{TN+FP}$	正确预测阴性样本占总阴性样本的比例
精确度 Precision	$\frac{TP}{TP+FP}$	正确预测阳性样本占预测结果为真的样本比例
F1 分数	$\frac{2*Precision*Recall}{Precision+recall}$	综合召回率与精确度评估模型预测准确程度

此外，在评估模型预测效果时还有另一种不依赖混淆矩阵的指标，本文使用了其中

的两种——接收器操作特性曲线下面积 (Area Under Receiver Operating Curve, AUROC) 以及柯尔莫哥洛夫-斯米尔诺夫指标 (Kolmogorov-Smirnov Statistic, K-S Statistic)。接收器操作特性曲线 (ROC) 是用来评价二元分类器分类能力的一种曲线，其以真阳性率为纵轴假阳性率为横轴绘制模型不同阈值下真假阳性率的坐标。其曲线与横轴所围成面积被称为曲线下面积 (AUC)，对于一类模型而言，AUC 越高表示其做出正确判断的能力越强。^[68] 并且同准确率相比，AUC 评价时不受训练集正负样本比例所影响。^[69] 因此本文使用该指标衡量模型预测能力。具有相似功能的指标还有 K-S^[70]，它是一种用来检验两经验分布之间是否相同的统计量，K-S 量越大表明模型的区分能力越强。

4.2 模型预测效果评价

4.2.1 基准计算

在对本模型效果进行分析前，我们首先需要计算使用传统 PRS 方法时骨关节炎的风险预测效果。PRS 方法，又称多基因风险评分 (Polygenic Risk Score) 方法^[71]，是一种评估 SNP 位点对某一特定性状的累计影响的量化指标。它由已有 GWAS 研究结果中 SNP 对表型影响的效应值以及统计显著性构建。其计算公式为如式4-1

$$S = \sum_i^N X_i \beta_i \quad (4-1)$$

其中 S 为个体的风险评分， N 为总潜在 SNP 的数量， X_i 为第 i 个 SNP 中潜在效应等位基因的数量， β_i 为第 i 个 SNP 对形状的效应值。我们根据该公式对现有基因型数据进行计算得到个体评分，再通过逻辑回归法构建基于 PRS 的风险预测模型并得到如图4-2-a 所示 ROC 曲线。可以发现，传统 PRS 模型的风险预测效果较差，AUC 仅为 0.51，与随机预测相当。

同时我们也使用了包括决策树算法在内的若干传统机器学习模型对基因型数据进行处理与预测，所得结果如图4-2-b 与基于 PRS 的模型无显著差异。我们还通过混淆矩阵相关指标对两种模型进行评估，得到表4-2。从表中我们可以看出两模型预测性能均较差，这意味着使用 PRS 模型与传统机器学习模型对骨关节炎的预测几乎没有任何实际意义。而这也与 Boer^[5] 研究中构建的 PRS 模型结果相符。因此本文以该 PRS 模型为基础讨论本文提出模型对预测效果的改善。

4.2.2 特征筛选方法对比

之前的讨论中我们提到，特征筛选对于特征数与样本数之比较高的数据集有着很好的提高模型性能的效果。本文也因此提到了两种特征筛选方法——基于卡方的样本筛选与基于支持向量机的样本筛选。对于使用同样超参数的决策树模型，我们输入使

表 4-2 混淆矩阵相关指标

指标	基于 PRS 的模型	决策树模型
auc	0.51	0.50
f1	0.45	0.29
accuracy	0.52	0.51
precision	0.50	0.50
recall	0.41	0.21

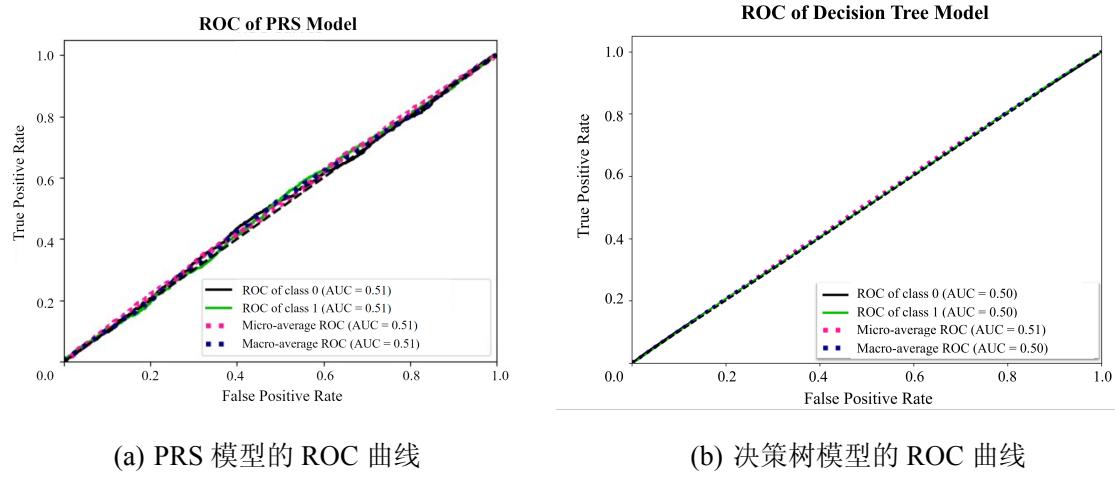


图 4-2 基准模型 ROC 曲线

用不同特征筛选方式处理后的数据集并比较两种特征筛选方法对模型效果的影响。我们首先通过混淆矩阵相关指标对该结果进行评估，结果如表4-3。

表 4-3 特征筛选方法对比

指标	经卡方法筛选特征	经支持向量机法筛选特征
f1	0.38	0.43
accuracy	0.52	0.54
precision	0.51	0.55
recall	0.29	0.35

可以看出，使用两者特征筛选方法所得结果虽然均同 PRS 法没有明显提高，但是相比较而言支持向量机法筛选特征在决策树模型中的效果更好。我们又对两种方法所得数据在决策树模型中的性能通过非混淆矩阵指标进行评估，得到图4-3。图中我们可以看到，支持向量机法筛选所得数据训练出的模型在 AUC、KS 指标方面均优于通过卡方法筛选所得数据，本文因此使用支持向量机法筛选特征。

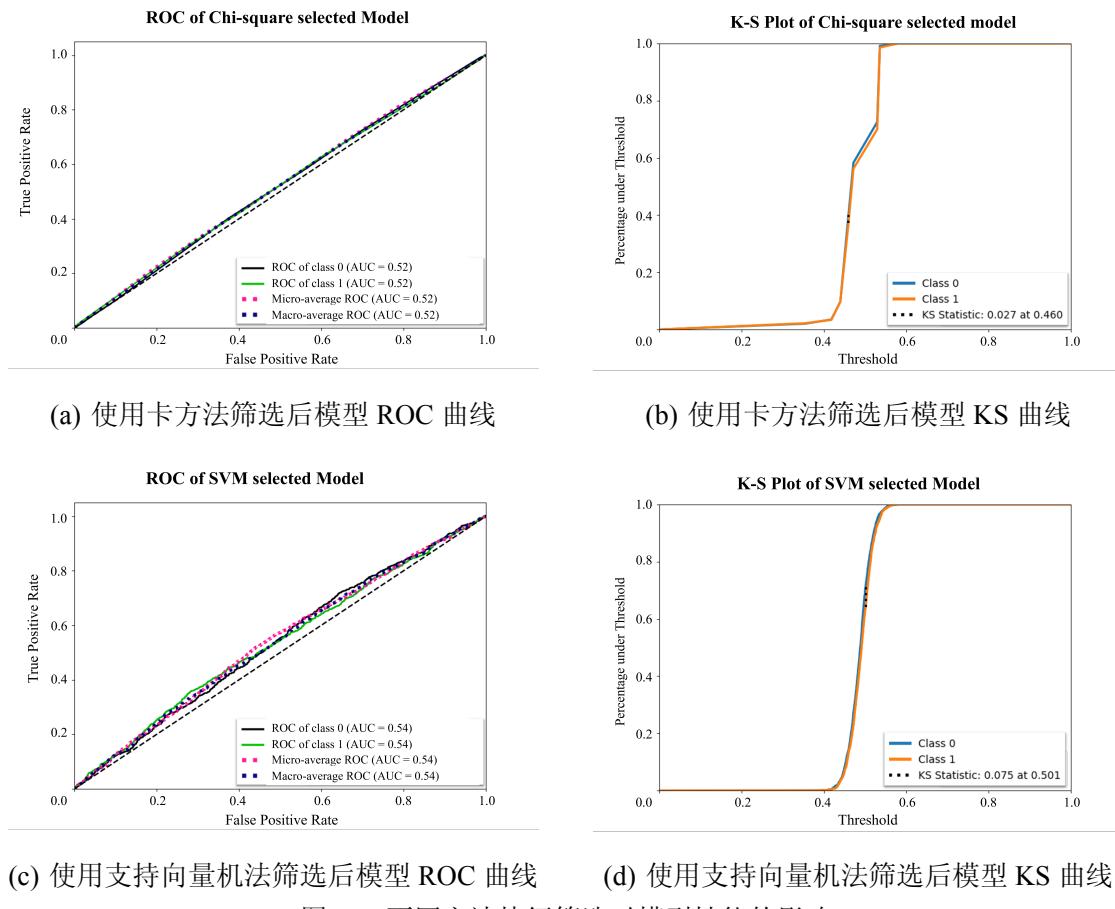


图 4-3 不同方法特征筛选对模型性能的影响

4.2.3 图神经网络处理

将经过支持向量机法筛选所得特征输入本文提出的风险预测模型中进行图估计与图神经网络处理，选择最优的图估计结构，并完成参数调优后，本文训练出根据个体基因型数据完成骨关节炎风险预测的图神经网络模型。我们首先通过混淆矩阵相关指标对该结果进行评估，结果如表4-4。

表 4-4 图神经网络处理

指标	图神经网络预测结果	PRS 模型
f1	0.56	0.45
accuracy	0.60	0.52
precision	0.60	0.50
recall	0.53	0.41

根据该结果我们发现，训练所得图神经网络在预测准确性、f1 分数、精准度、灵敏度上均优于传统 PRS 模型。我们还通过非混淆矩阵指标对图神经网络性能加以评估，

得到图4-4。

可以看出，本文建立的图神经网络较 PRS 模型在 AUC 方面有了明显改善。同时模型在模型 Precision-Recall 曲线中我们也发现随着召回率的提升，模型准确率下降速度较慢这也预示着模型分类性能较好。在模型 Lift 曲线中，随着深度增加，模型 lift 值在深度较大时下降较快，意味着模型良好的分类性能。综上，本文使用的单纯依靠基因型数据的图神经网络模型相较于传统模型而言有了明显的性能进步，然而，该模型绝对性能依然不如人意。

4.2.4 表型融合

为进一步增强模型性能，本文引入同骨关节炎密切相关的表型参与风险预测，该模型其给出如表4-5混淆矩阵相关指标结果。

表 4-5 融合表型信息对模型性能的影响

指标	融合表型	单纯图神经网络	PRS 模型
f1	0.66	0.57	0.45
accuracy	0.67	0.58	0.52
precision	0.66	0.57	0.50
recall	0.69	0.58	0.41

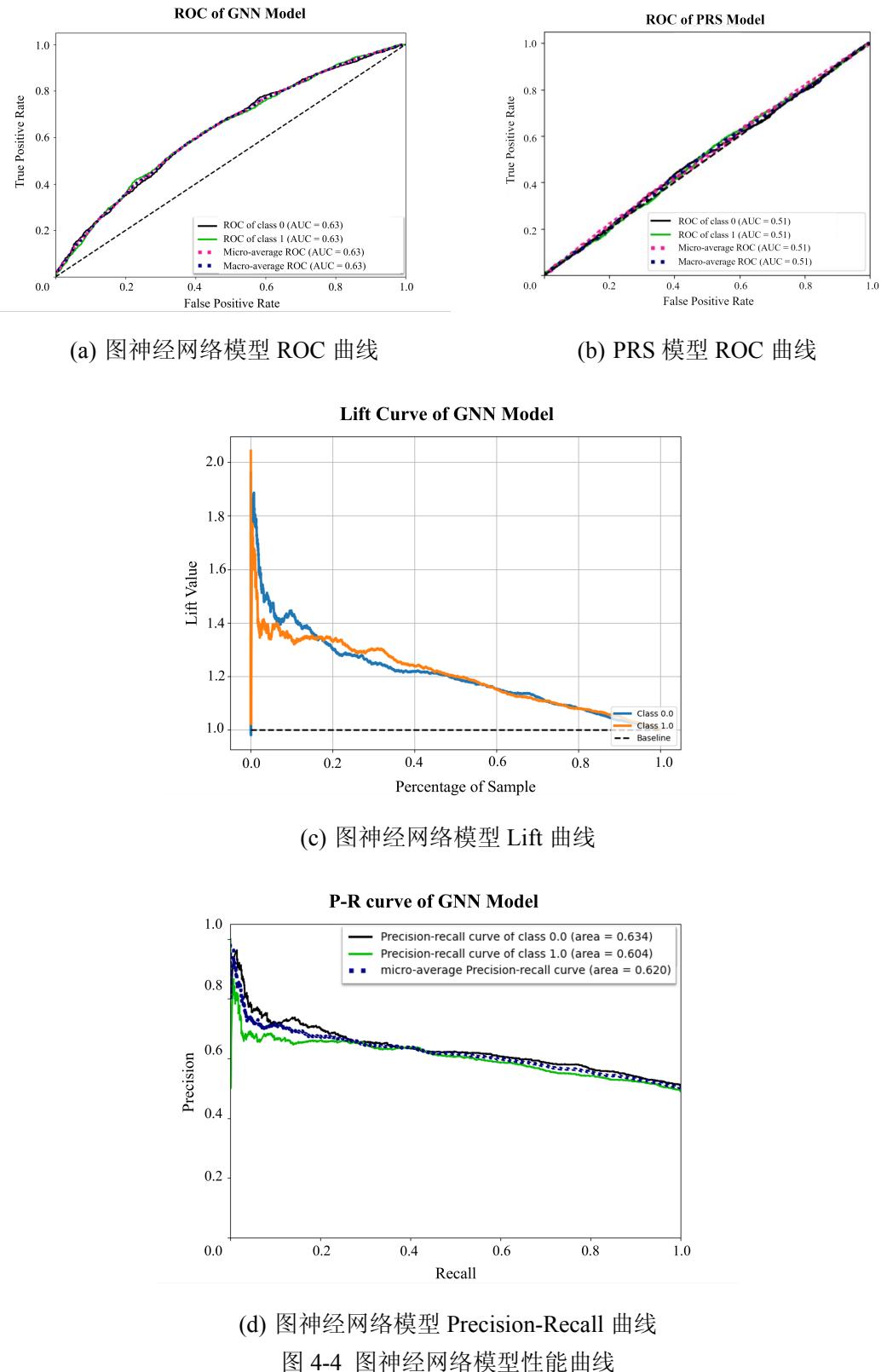
可以看出融合表型之后模型性能相较于单纯使用基因型数据的图神经网络模型有了进一步提升。我们还对该模型性能的非混淆矩阵指标进行了计算并同图神经网络模型加以比较。在融合表型模型中，预测 AUC 已经达到 0.74，相较于单纯利用基因型数据的图神经网络预测准确能力显著增强。同时模型的 KS 值（0.356）相较于单纯利用基因型数据的图神经网络 KS 值（0.202）也有明显提高。Precision-Recall 曲线方面，融合表型模型的 PR 曲线下降较单纯使用基因型的图神经网络模型更慢，意味着模型良好的分类性能。而在 Lift 曲线中，融合表型模型 Lift 曲线较图神经网络模型曲线更为陡峭，也提醒模型预测性能的增强。综上，通过表型信息的融合，本文提出的骨关节炎风险预测模型较传统模型相比取得了可观的性能提升，已经具有较好的风险预测能力，可以应用于实际的风险预测之中。

4.3 图估计效果

4.3.1 估计器在训练过程中的效果

4.3.1.1 可行性验证

为了验证本文设计图估计器对无结构数据的处理能力，在使用该估计器估计本研究所使用骨关节炎患者基因型数据结构之前，本文先在 MNIST 数据集上进行测试。MNIST



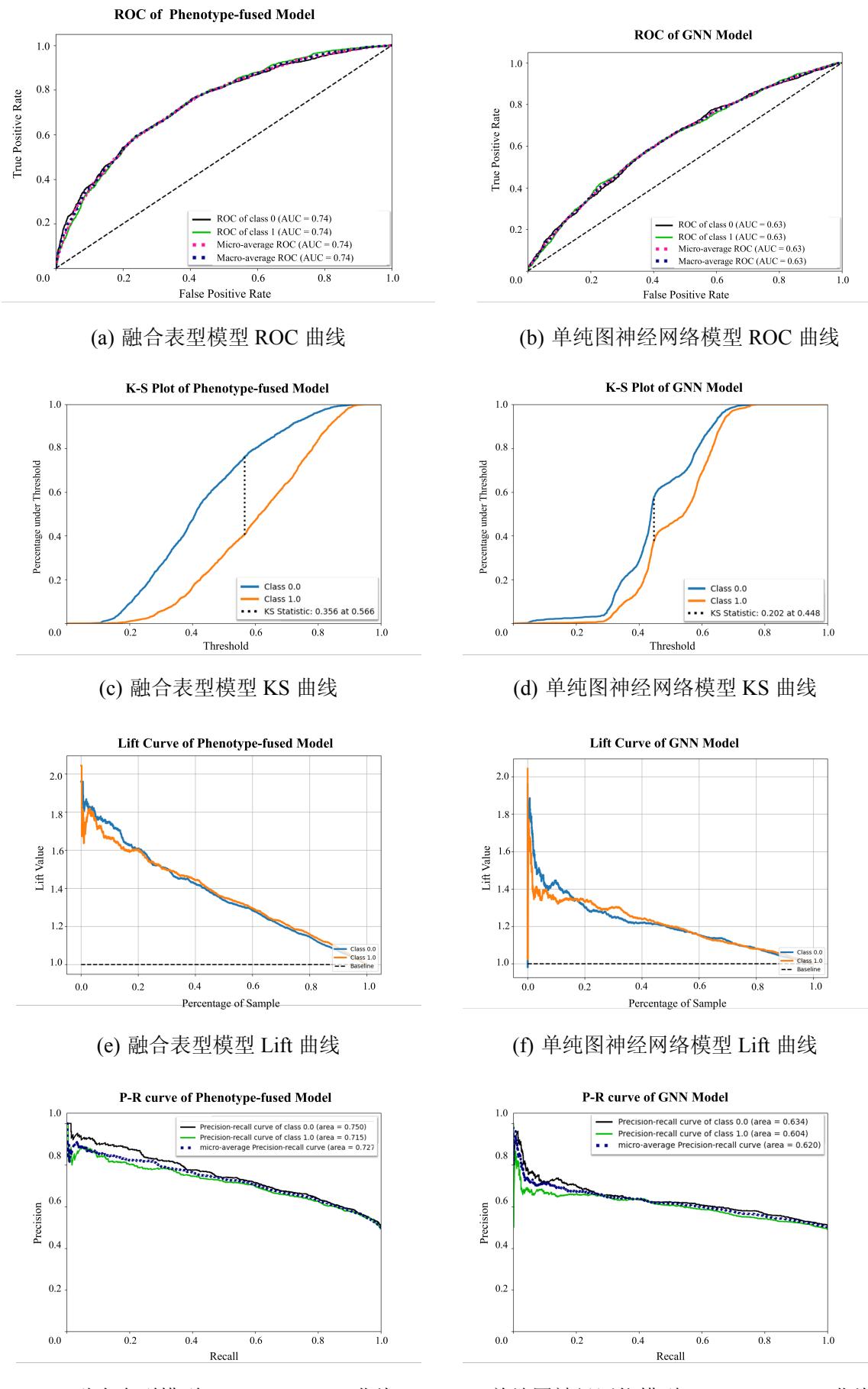


图 4-5 融合表型信息对模型性能的影响

数据集 [72] 由若干已标注手写数字图片组成，也是一种无结构数据。我们因此在该数据集上应用我们的模型并将结果加以记录如图4-6。可以看到，在每个迭代的图估计器工作之后，模型损失有着明显下降，同时分类准确率也有显著提升。这证明了本文提出图估计器对图神经网络在无结构数据上的性能有着明显改善，为后续对骨关节炎患者基因型数据的处理工作提供可行性基础。

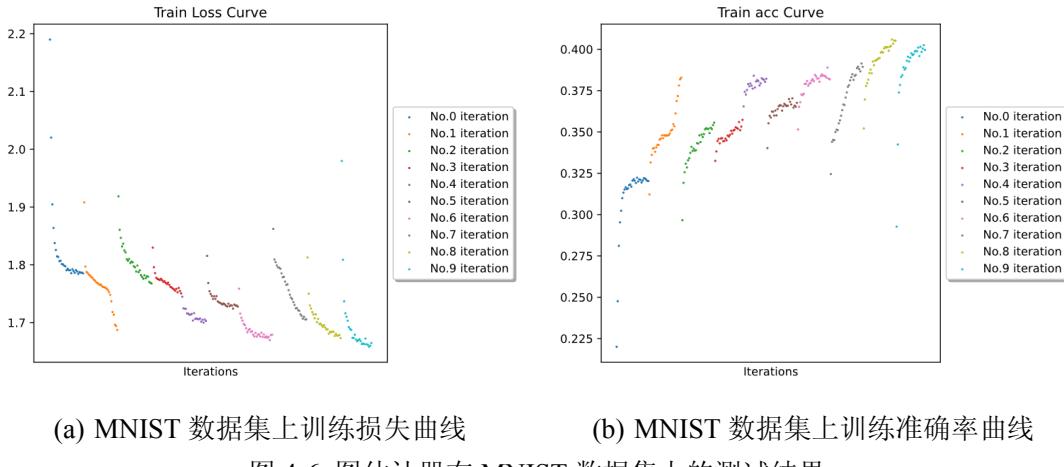


图 4-6 图估计器在 MNIST 数据集上的测试结果

4.3.1.2 对照组

图估计器的工作过程涉及到图神经网络参数与图结构的循环更新。为了排除多次重启训练对模型性能的潜在影响，我们设计对照组，即使用全连接法构建图矩阵并且每次图神经网络训练之后不更新图结构。并对对照组的模型训练情况做以记录如图4-7。可以看到，在不更新图结构时，每次迭代中模型损失与准确率的变化趋势基本一致，同时不同迭代末期模型 AUC 变化不大。因此我们认为模型效果与无图结构更新过程的迭代次数无关。

4.3.1.3 实际数据

该组试验中我们正式将图估计器与图神经网络运用于骨关节炎患者基因型数据的处理过程中。我们首先使用全连接法构建初始图，在每个迭代中我们训练神经网络并给出数据在该网络下的输出，再将输出作为图估计器的观察值并基于此更新图结构，再将该图结构作为数据的结构重新输入图神经网络中训练。如此迭代若干次，记录模型性能如图4-8。

可以看出，首次迭代中由于使用全连接网络，模型性能变化同对照组中模型相仿。但是在首次迭代结束生成新图结构后并以此进行第二次迭代的图神经网络训练时我们可以发现，训练末期模型损失明显下降，模型准确率明显上升。模型预测能力 AUC 也有明显提高。最终我们选择预测能力最好的迭代作为最优模型，参与到骨关节炎风险预测过程之中。

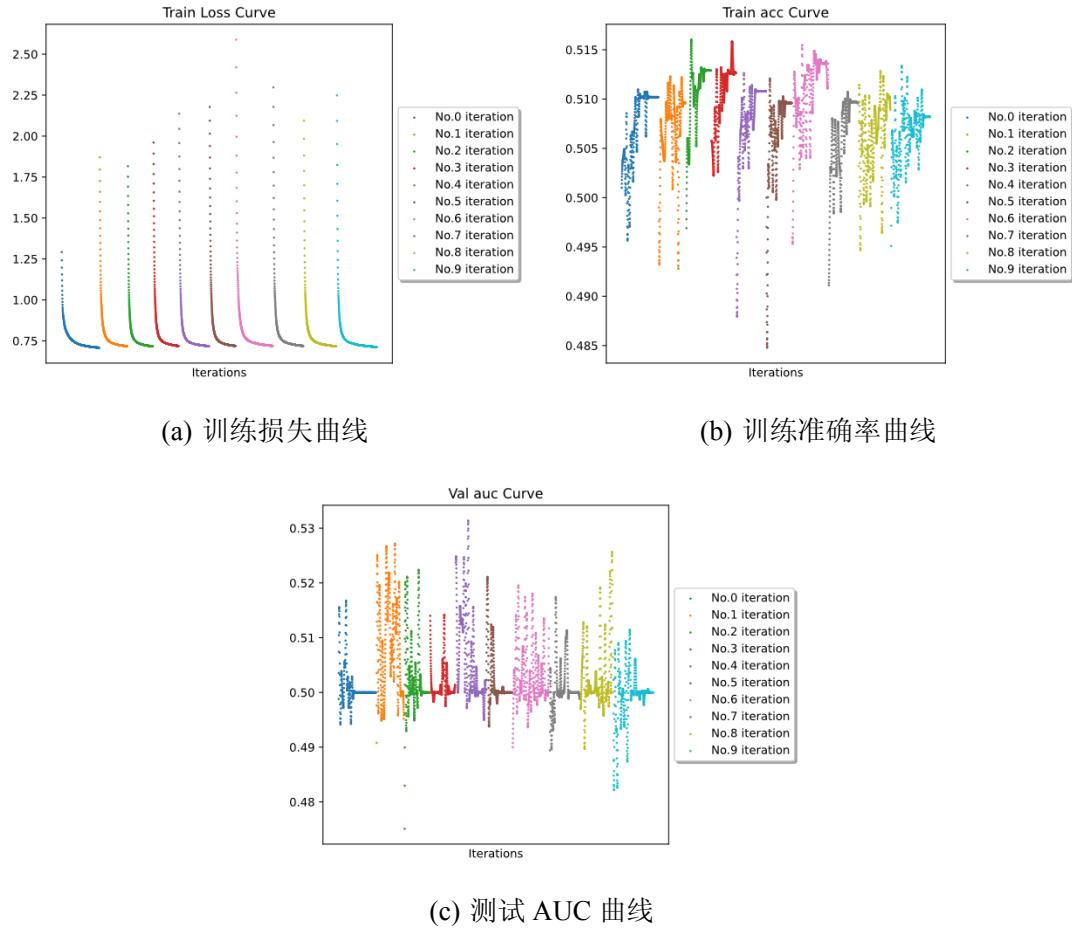


图 4-7 对照组训练结果

4.3.2 图估计器工作过程分析

4.3.2.1 聚类参数选择

在讨论图估计器时我们提到，图估计器在输入观察值的同时还需要人工选择聚类参数 k 以初始化变分参数，且该聚类参数会对估计结果产生直接影响，该影响可通过整合分类似然评估。为了选择最优聚类参数，我们计算了使用一定范围内的聚类参数时的模型整合分类似然，结果如图4-9。

可以发现， k 取 6 时模型的整合分类似然最低，性能最好。同时我们考虑了在使用该范围内参数时变分期望最大化算法所给出的证据下界 ELBO，训练中 ELBO 的变化如图4-10。我们发现虽然 k 取 6 的组证据下界并非最高，但是同其他组相比也处于高位。因此我们将 k 定为 6 并估计该参数下可能的图结构。

4.3.2.2 生成图结构分析

根据前文对变分期望最大化算法的讨论我们可以发现， k 取 6 意味着最终生成的图结构可以分为 6 簇。为了探究该网络结构以及其潜藏的信息，我们将估计图结构中的

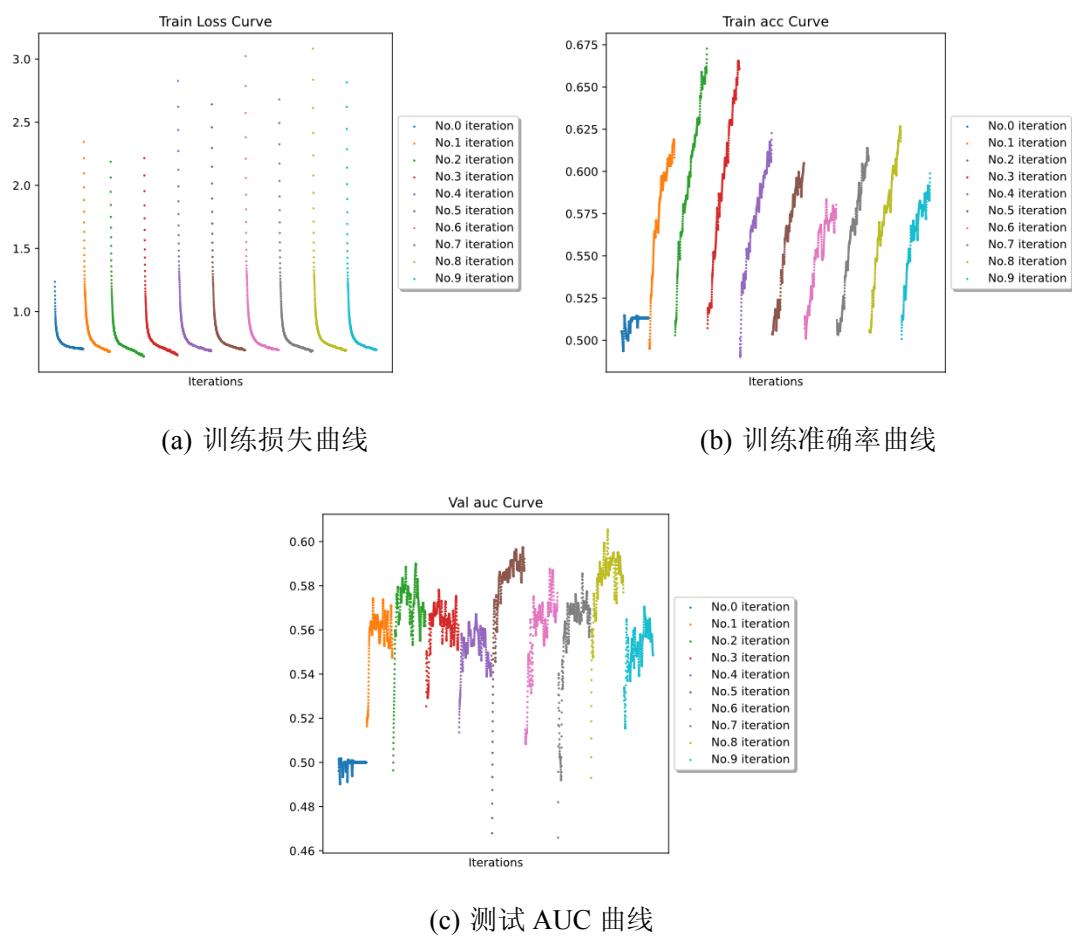


图 4-8 实际数据训练结果

ICL curve on different initial clusters

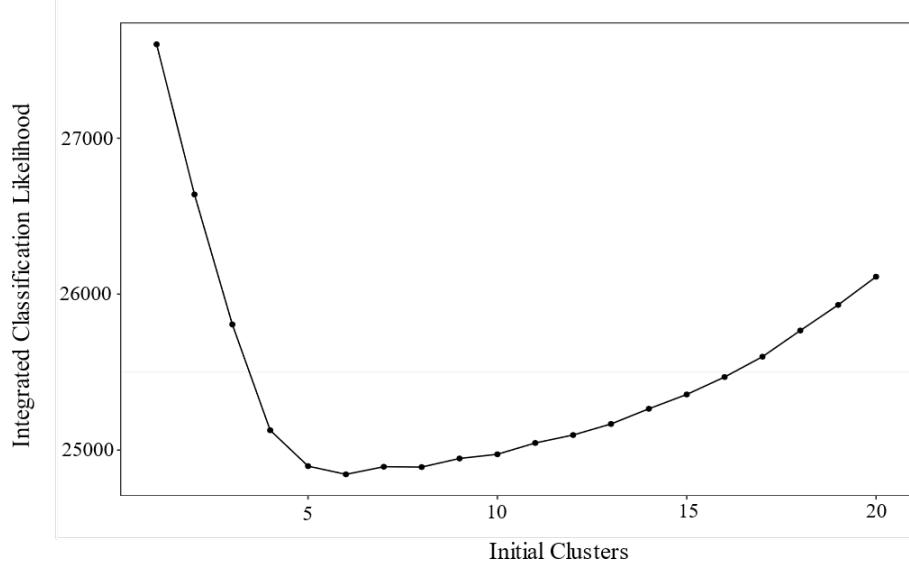


图 4-9 整合分类似然变化曲线

节点依照其所属的簇加以归类并表示如图4-11。

我们同时将计算所得的簇与簇之间的关系表示如图4-12。可以发现，簇3与簇4、

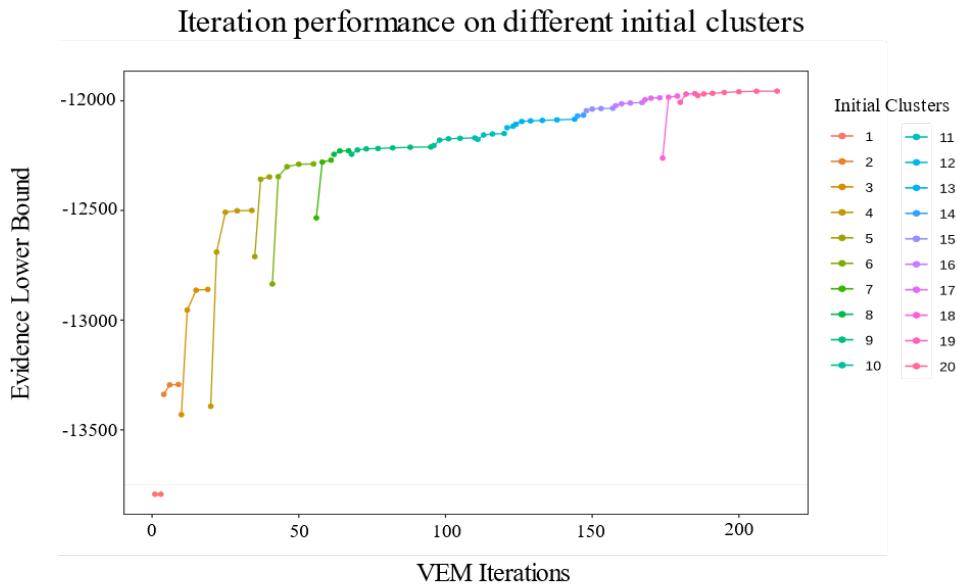


图 4-10 证据下界曲线

Estimated Graph Structure

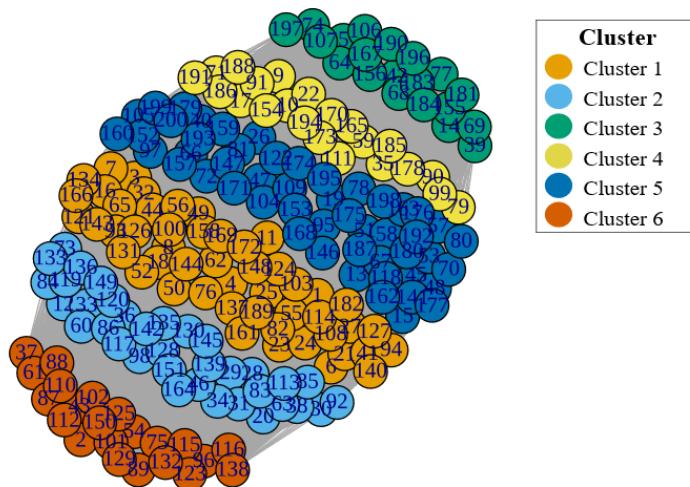


图 4-11 节点簇信息

簇 2 与簇 6 之间存在着很强的相关信号。且簇 6 与除簇 1, 2 之外的其它簇关联都不强。我们同时发现簇 3、5 之间也存在一相关信号。除此以外数学分析无法从该图中获取更多信息，我们因此需要结合图中节点的生物学意义对该图结构与簇关系加以分析。

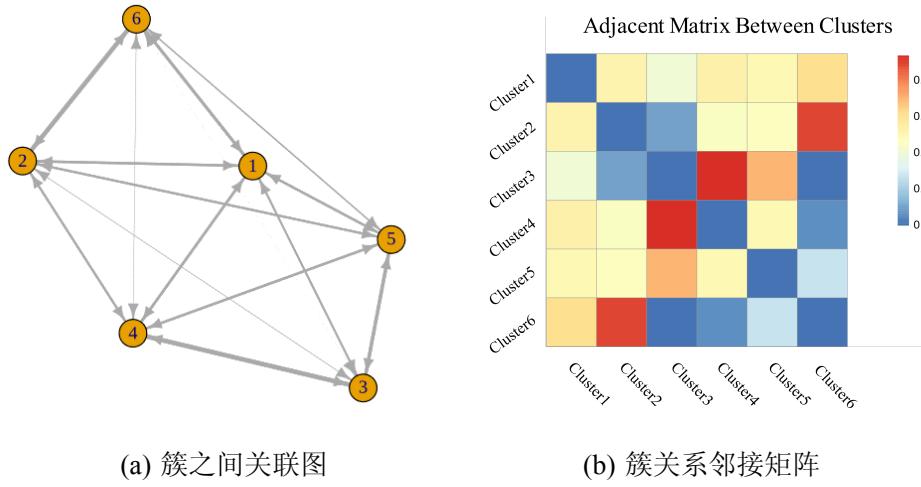


图 4-12 簇之间关系

4.3.2.3 结合表型分析

为了进一步探究生成图结构的生物学意义，我们需要了解其中节点，也就是 SNP 位点对人的不同表型的贡献。同 GWAS 研究类似的，全表型关联研究（Phenome-wide Association Study, PheWAS）主要研究某一 SNP 位点与已知性状的关联 [73]，因此我们对图中每个节点进行 PheWAS 研究。我们对节点的研究主要在 GWASATLAS 数据库内进行 [74]，该数据库包含了在 UKB 数据上进行的对 3302 个表型的约 4756 个 GWAS 研究，可以较为全面地展现 SNP 位点对表型的影响。对于每个 SNP，我们编写爬虫由数据库获取与其相关的表型以及该相关的统计学显著性，经过统计学显著性筛选 ($p < 10^{-5}$) 后将与每个 SNP 相关的表型按照领域分为 15 类。对于每个簇，我们统计簇内 SNP 位点对每类性状的关联并依照式4-2进行簇内标准化，得出如图4-13所示热图。

$$x = \frac{x - \mu}{\sigma / \sqrt{n}} \quad (4-2)$$

可以看出，每个簇内与簇内 SNP 相关联程度最高的表型均集中在代谢与免疫领域，这也与骨关节炎本身炎症性质相关。为了能进一步分析簇间差异，我们还依照式4-2对每个领域内关联进行簇间标准化，得到如图4-13所示热图。

相较于簇内标准化，我们可以从此图中发现簇间的明显差异，尤其是簇一、簇五在环境、内分泌、免疫等领域表型关联同其他簇存在明显差异。我们对六簇在这三个领域中所观察到的关联事件绘制提琴图，得到图4-15。提琴图中的簇间差别也印证了在热图中观察到的簇一、簇五的明显差别，因此本文对该两簇进行深入研究。

我们首先对簇一内布关联性状分布、数量、统计显著性进行分析并得出图4-16。通过该组图可以看到同其他簇相同，与簇一内 SNP 位点相关表型主要集中在代谢与免疫领域。但是我们从热图中发现簇一有着较强的环境相关信号。我们对该环境相关信号加以统计，结果如表4-6所示。出乎我们意料的是，与簇一内 SNP 相关的环境相关表型

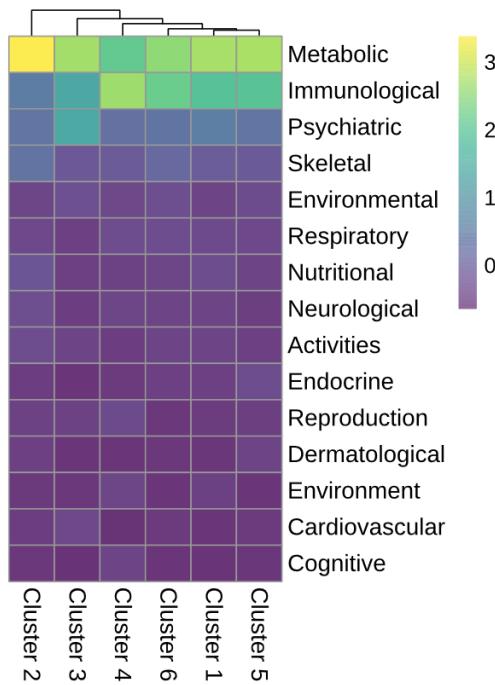


图 4-13 簇与性状关联簇内标准化热图

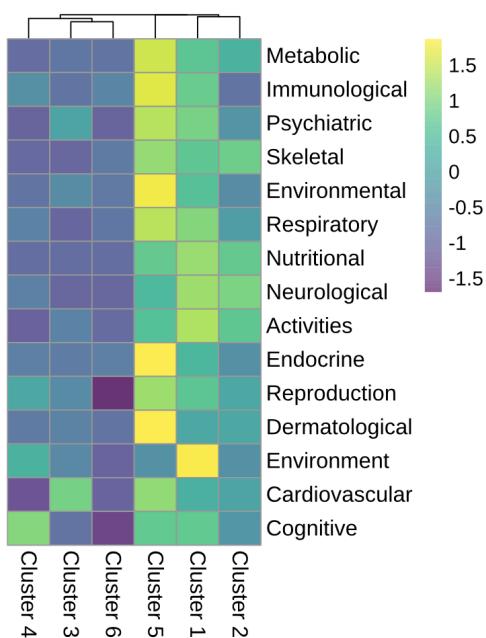
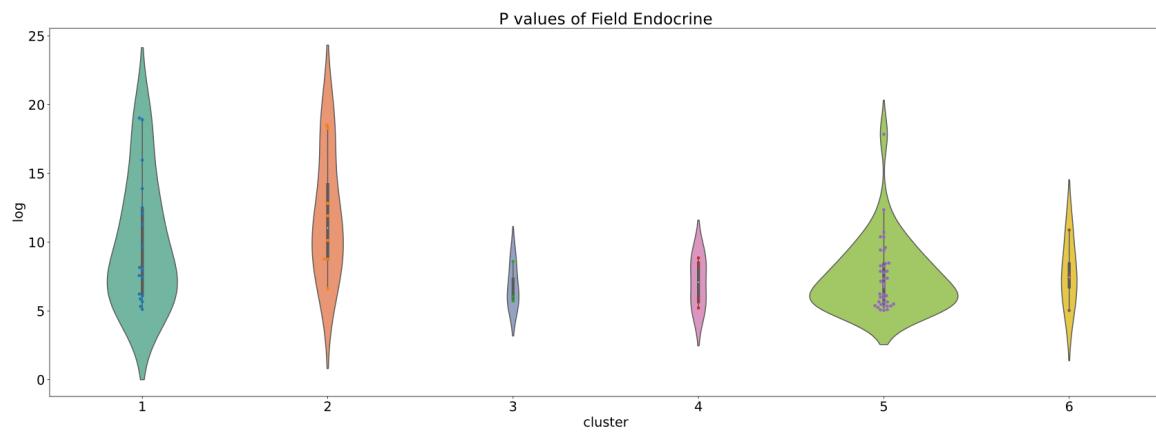
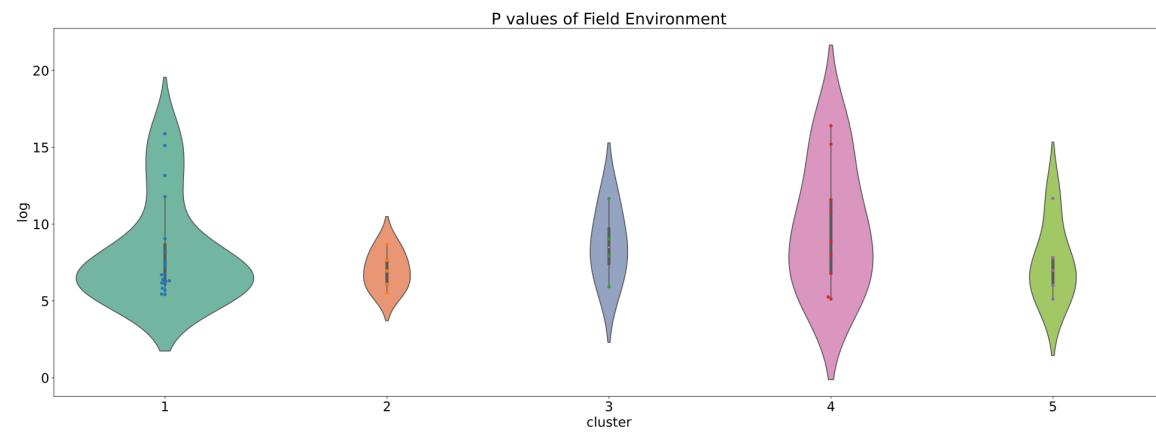


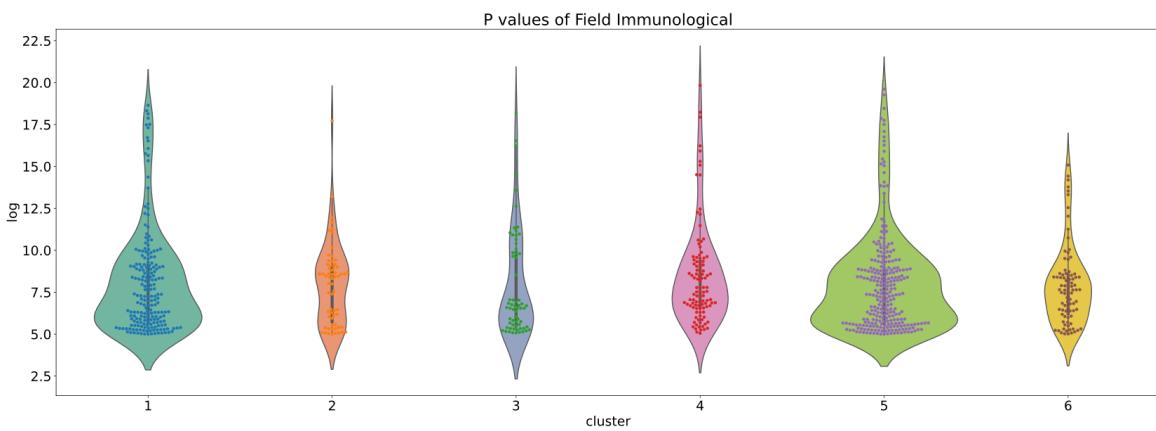
图 4-14 簇与性状关联簇间标准化热图



(a) 内分泌领域表型关联分布提琴图

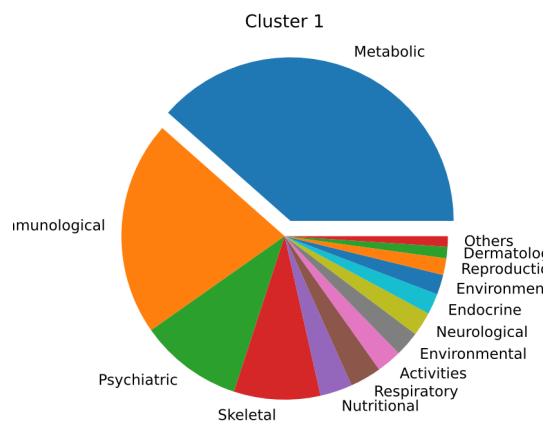


(b) 环境领域表型关联分布提琴图

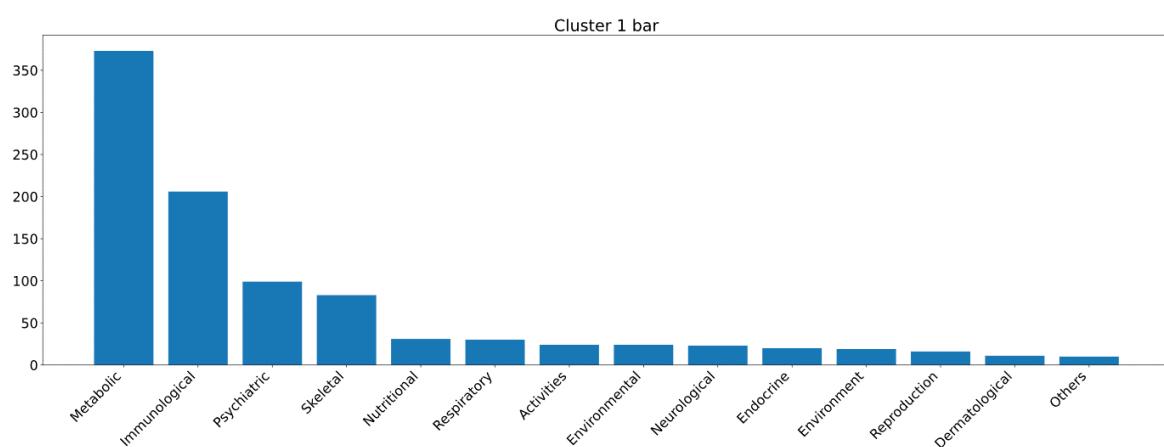


(c) 免疫领域表型关联分布提琴图

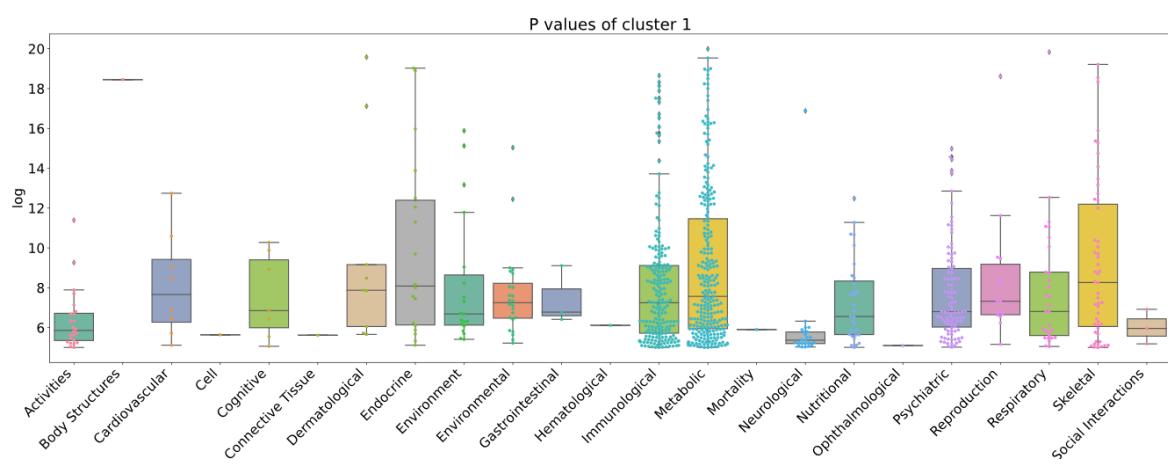
图 4-15 不同领域性状关联分布提琴图



(a) 簇相关性状分布饼图



(b) 簇相关性状分布柱图



(c) 簇相关性状 p 值图

图 4-16 簇一相关性状分领域图

出现了受教育程度、工作是否涉及重体力劳动等看起来与骨关节炎毫不相干的性状。但是，在查阅文献 [75] 后我们了解到，骨关节炎作为一种退化性疾病有可能与关节过度磨损相关。因此我们做出猜测，受教育程度较低的个体趋向于从事体力型劳动，导致关节加速磨损，最终导致骨关节炎的发生。而该簇内的 SNP 可能在受到这种环境影响时更易导致骨关节炎的发生。我们猜测该簇内 SNP 可能与环境因素导致关节磨损继而导致的骨关节炎相关。

表 4-6 簇一相关表型

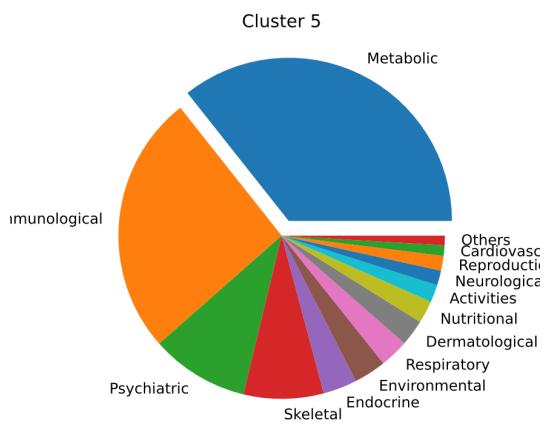
领域	表型	计数
Environment	Educational attainment	10
Environment	Education - Qualifications	3
Environment	Attendance/disability/mobility allowance: Blue badge	1
Environment	Job involves heavy manual or physical work	1
Environment	Maternal smoking around birth	1

我们也对簇五进行了类似分析，得到图4-17. 相较于其他簇，簇五在免疫、内分泌、皮肤病领域有着很强相关。我们对这些领域内的典型性状加以统计得到表4-7。可以看出，簇五内的 SNP 与二型糖尿病、巨噬细胞、白细胞与粒细胞计数相关表型都存在着相关。二型糖尿病是一种由于机体胰岛素抗性而生成的糖尿病，其主要由超重乃至肥胖而导致。[76] 而肥胖导致的高血脂又会诱发体内免疫系统包括巨噬细胞与粒细胞的激活。[77] 这也与簇内的观察一致，因此我们猜测簇五内的 SNP 位点主要与个体自身肥胖导致的糖尿病与超重继而诱发的自身型骨关节炎相关。

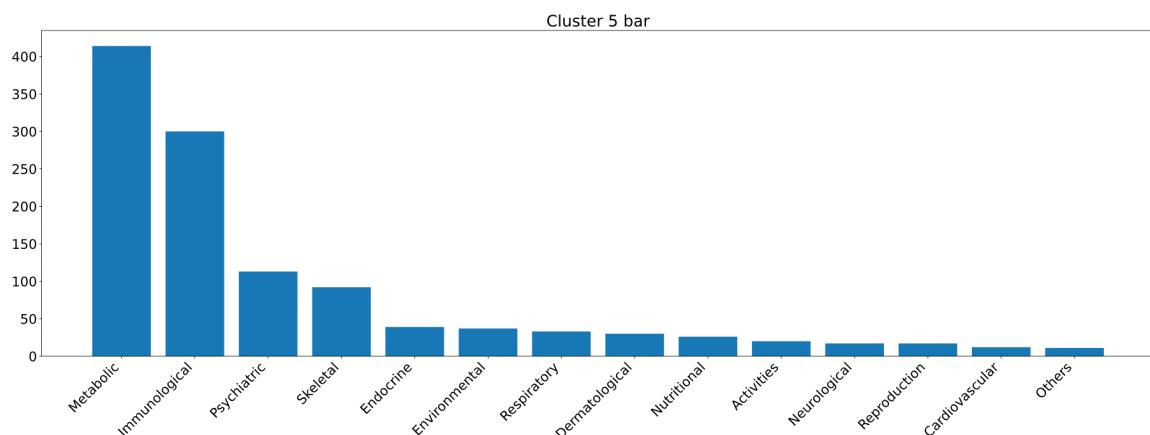
表 4-7 簇五相关表型

领域	表型	计数
Endocrine	Type 2 Diabetes	25
Immunological	Myeloid white cell count (three-way meta)	14
Immunological	White blood cell count (three-way meta)	14
Immunological	Granulocyte count (three-way meta)	13
Dermatological	Male pattern baldness	10

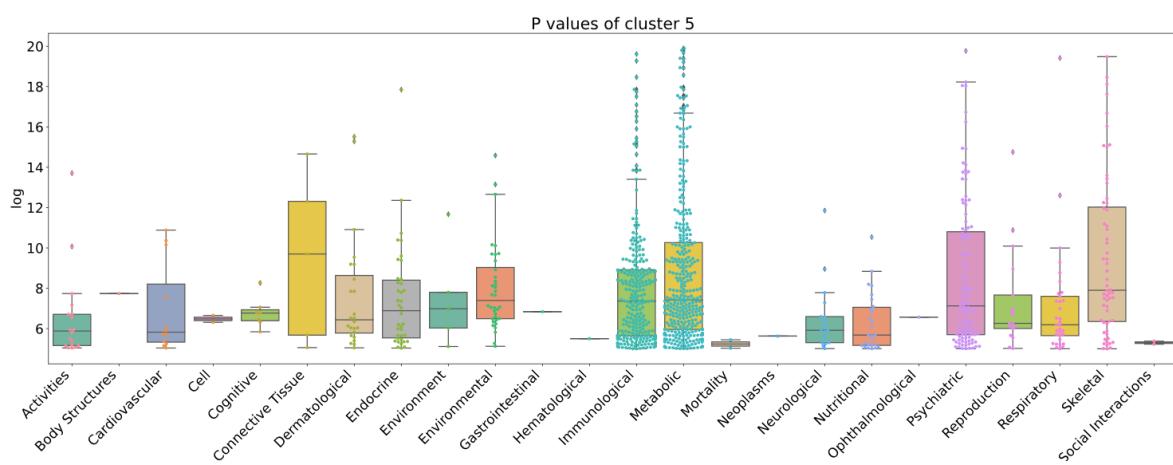
以上我们通过对图估计生成图结构的分析将输入的致病 SNP 分为六簇并对其中两个典型簇进行具体生物学意义分析。同时根据两簇内 SNP 的相关性状提出了骨关节炎



(a) 簇相关性状分布饼图



(b) 簇相关性状分布柱图



(c) 簇相关性状 p 值图

图 4-17 簇五相关性状分领域图

的两种诱因，即由环境因素导致关节过度磨损诱发的骨关节炎以及由于超重诱发的自身型骨关节炎。而这类分析是传统机器学习模型所无法实现的，这也进一步印证了图神经网络在学习数据深度信息时的出众能力。

4.4 图解释器结果案例分析

我们之前提到本风险预测模型中的图解释器能对产生预测结果的原因，即输入基因型图中同预测结果相关的最大子图，加以分析与计算。本节通过模型对 UKB 数据库中一个案的分析结果展示该能力。本文从 UKB 数据库中选取样本 (UID=5125713)，并将其基因型数据根据图估计器给出的最佳图结构转为图数据。利用训练好的图神经网络模型对该图进行处理，预测结果为该样本有很高概率患骨关节炎。我们将该预测结果与模型输入图解释器中，最终图解释器给出如图4-18相关子图。

Subgraph from GNN Explainer

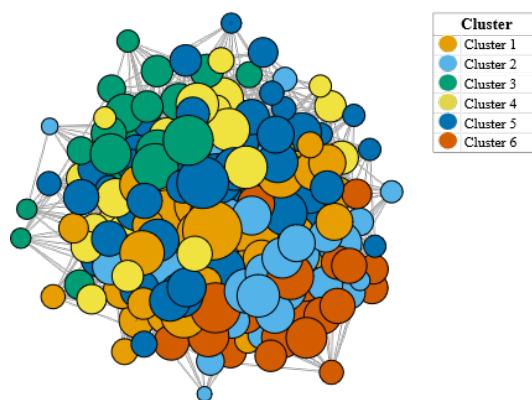


图 4-18 图解释器生成与预测结果相关子图

在该子图中，边的权重为该关联对预测值的贡献，权重越高代表该边对预测值的生成越“重要”。我们将该图按照边的权重进行稀疏化，得到如图4-19系列子图。可以看到，随着权重阈值的不断提升，子图的骨干逐渐显现，而该主干则主要由簇一、五中的节点构成。上一节的分析中我们提到簇五中所含位点可能与二型糖尿病相关，我们因此推测该个体可能由肥胖造成的关节过度磨损诱发骨关节炎，同时该患者可能合并二型糖尿病。我们从 UKB 数据库中提取患者其他表型得到表4-8。

患者实际 BMI 为 35.2，体脂率为 46%，已为肥胖体型，并且合并确诊糖尿病。可以看出，患者实际情况同我们根据图结构以及图解释器做出的解释一致。体现了本文提出模型除高效预测骨关节炎风险之外对疾病分型以辅助治疗的能力。

4.5 本章小结

本章对本文搭建的骨关节炎风险预测模型的预测准确性通过若干指标加以评估并同现有的常见疾病风险预测模型算法进行比较，发现本文搭建模型较传统 PRS 方法与机器学习算法而言在准确性上有着显著提高，已可作为临床辅助工具参与骨关节炎的

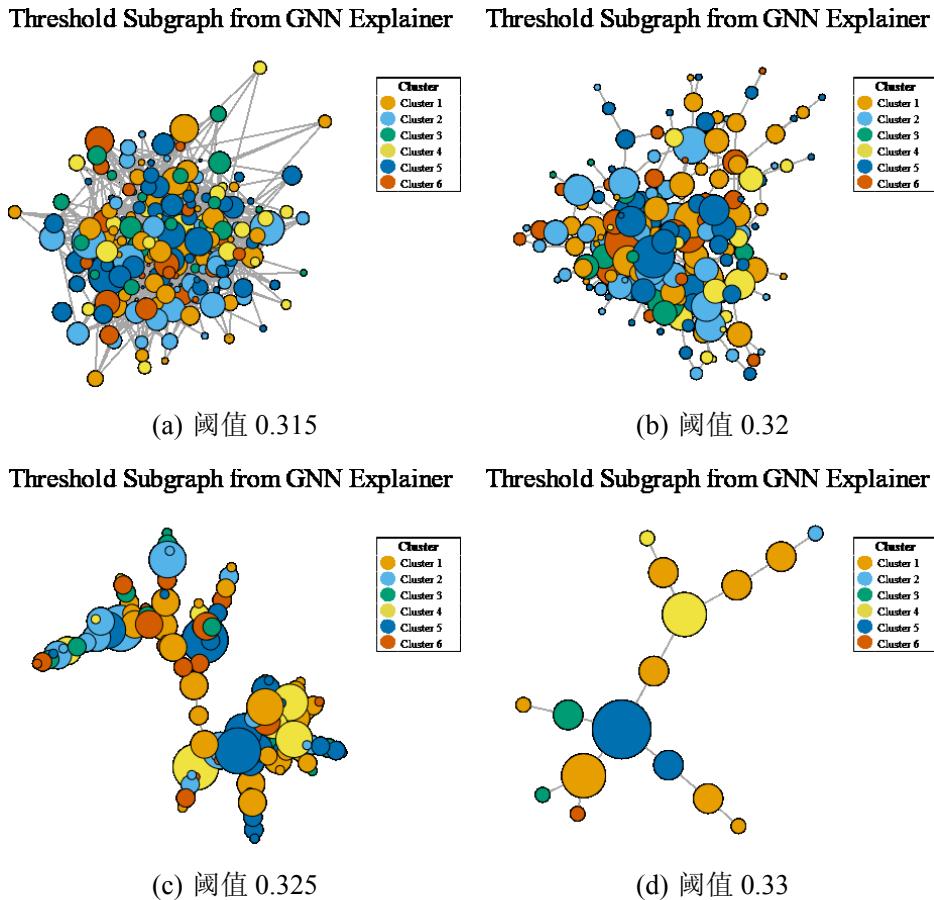


图 4-19 子图阈值化结果

表 4-8 该个体其他表型

UKB-ID	表型	值
31	性别	男
48	腰围	113
49	胯围	121
2443	确诊糖尿病	是
21001	BMI	35.2
23098	体重	105.5
23099	体脂率	46.1

早期诊断。本章同时对图估计器的工作过程加以记录与展示，一方面验证了算法的可行性；另一方面也证实了图估计器可以显著改善无结构数据在图神经网络中的表现。为了进一步挖掘潜藏在估计器所得图结构中的信息，本章还结合图估计器输出数学模型与图中节点的 PhEWAS 研究结果对图中的簇聚节点之间关系加以分析并且识别出了两个典型簇，并且根据典型簇内节点信息提出了关节炎的两种诱因。最后本章利用模型解释器进行案例分析，在给出预测结果的同时还利用解释器的结果与识别出的典型簇

预测了疾病的分型。以上内容完整展现了本模型作为高效准确可解释骨关节炎风险预测模型的功能与性能。

5 总结与展望

5.1 工作总结

构建高效可解释骨关节炎对骨关节炎的早期诊断与防治有着重要的意义。本文首先根据目前已发表的 GWAS 研究及公共数据库 UK BioBank 获取患者表型与基因型数据并进行数据预处理；之后构建了包括图结构估计器、图卷积神经网络、表型信息融合与图解释器四个模块在内的骨关节炎风险预测模型；最后本文对该模型的性能以及预测结果进行了进一步的分析和解读，继而对模型的预测准确性，可解释性等指标进行评估。本文对主要完成的工作总结如下。

1. 构建了适用于图分类问题的图结构估计器：目前提出的用来处理无结构数据的图结构估计器多适用于图节点分类问题，对适用于图分类问题的图结构估计器研究较少。本文提出了一种基于变分期望最大化算法的图结构估计器，该估计器通过对图神经网络输出的处理通过参数估计的方法预测图节点之间关联可能性与图整体结构，继而在图神经网络训练中动态更新图结构。经测试，该估计器能明显改善图分类问题中无结构数据在图神经网络上的表现。同时，参数估计方法使得估计器生成的图结构具有明显的局部结构，能够用于图深层信息的挖掘。该估计器弥补了图分类问题图估计器的空白，对无结构数据的图神经网络处理具有积极意义。

2. 构建了具有实用价值的骨关节炎风险预测模型：现有的骨关节炎风险预测模型效果差，不能满足骨关节炎患者早期诊断与筛查的需求。本文所构建的风险预测模型通过结合表型信息与基因型信息输入图神经网络进行处理，最终实现了较好的预测准确率，具备实用风险预测价值。同时本文提出模型通过结合图解释器对风险预测结果结合图神经网络给出预测值解释，使得模型在输出预测患病风险的同时能够对潜在的病因进行预测，对疾病分型与精准治疗具有积极意义。

3. 通过估计器所得图结构界定出骨关节炎两可能诱因：通过对图估计器所得图结构并结合 PheWAS 表型信息研究，我们发现了图结构中的两典型簇。其中一簇主要与受教育程度、是否从事中体力劳动等环境因素相关，本文推测该簇内 SNP 可能与环境因素导致关节磨损继而导致的骨关节炎相关；另一簇主要与免疫、内分泌等自身因素相关，本文推测该簇主要与个体自身肥胖导致的糖尿病与超重继而诱发的自身型骨关节炎相关。本文因此认为骨关节炎存在包括重体力劳作在内的环境诱因与包括肥胖、二型糖尿病在内的自身诱因，与目前对骨关节炎研究的观点一致。以上分析均是在单纯依靠图结构的基础上进行的，可以推广到其他疾病风险预测模型的研究之中，对复杂多基因疾病的风险预测与分型诊断有着积极意义。

5.2 展望

虽然本文所提出风险预测模型在给出可信预测结果的同时兼具可解释性，但是本文中还存在若干问题需要进一步研究

1. 模型超参数与结构还存在调整空间：本模型中四个模块都有着数量较多的决定模型性能的超参数，由于时间限制本文未能对这些超参数进行细致调整，可能对最终模型的性能产生不利影响。后续研究中仍要对其中一些参数加以分析调整以求继续提高模型性能。同时随着图神经网络研究的快速发展，本研究进行之时不断有效果更好的图卷积方式出现，后续研究中还需基于这些研究对图卷积层进行优化。

2. 图估计器仍需继续研究：为了简化问题，本模型图估计器所生成的图结构中各边的权重一致。但是现实生活中的图数据中各边权重通常不一致，并且有着具体意义，忽略该权重意味着潜藏信息的损失。如何通过参数描述各边的权重并通过算法估计该参数仍需后续研究。

3. 图估计产生图结构仍需进一步分析：本研究中图估计器生成的图结构共分为六簇，但是由于 SNP 位点生物学意义的复杂性，本文只对其中两簇进行了浅显的分析，并且没有对簇内节点之间与簇与簇之间的关联加以深入解释。后续研究中仍要对产生该结构的具体原因与意义加以具体分析

参考文献

- [1] Martel-Pelletier J, Barr A J, Cicuttini F M, et al. Osteoarthritis[J/OL]. Nature Reviews Disease Primers, 2016, 2(1):16072[2022-06-08]. <http://www.nature.com/articles/nrdp201672>. DOI: 10.1038/nrdp.2016.72.
- [2] James S L, Abate D, Abate K H, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017[J/OL]. The Lancet, 2018, 392 (10159):1789-1858[2022-06-08]. <https://linkinghub.elsevier.com/retrieve/pii/S0140673618322797>. DOI: 10.1016/S0140-6736(18)32279-7.
- [3] Vos T, Flaxman A D, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010[J/OL]. The Lancet, 2012, 380(9859):2163-2196[2022-06-08]. <https://linkinghub.elsevier.com/retrieve/pii/S0140673612617292>. DOI: 10.1016/S0140-6736(12)61729-2.
- [4] Hiligsmann M, Cooper C, Arden N, et al. Health economics in the field of osteoarthritis: An Expert's consensus paper from the European Society for Clinical and Economic Aspects of Osteoporosis and Osteoarthritis (ESCEO)[J/OL]. Seminars in Arthritis and Rheumatism, 2013, 43 (3):303-313[2022-06-08]. <https://linkinghub.elsevier.com/retrieve/pii/S0049017213001510>. DOI: 10.1016/j.semarthrit.2013.07.003.
- [5] Cooper C, Snow S, McAlindon T E, et al. Risk factors for the incidence and progression of radiographic knee osteoarthritis[J/OL]. Arthritis & Rheumatism, 2000, 43(5):995[2022-06-08]. [https://onlinelibrary.wiley.com/doi/10.1002/1529-0131\(200005\)43:5<995::AID-ANR6>3.0.CO;2-1](https://onlinelibrary.wiley.com/doi/10.1002/1529-0131(200005)43:5<995::AID-ANR6>3.0.CO;2-1).
- [6] Zhang Y, Niu J, Felson D T, et al. Methodologic challenges in studying risk factors for progression of knee osteoarthritis[J/OL]. Arthritis Care & Research, 2010, 62(11):1527-1532[2022-06-08]. <https://onlinelibrary.wiley.com/doi/10.1002/acr.20287>.
- [7] Veronese N, Cereda E, Maggi S, et al. Osteoarthritis and mortality: A prospective cohort study and systematic review with meta-analysis[J/OL]. Seminars in Arthritis and Rheumatism, 2016, 46 (2):160-167[2022-06-08]. <https://linkinghub.elsevier.com/retrieve/pii/S0049017216300087>. DOI: 10.1016/j.semarthrit.2016.04.002.
- [8] Styrkarsdottir U, Lund S H, Thorleifsson G, et al. Meta-analysis of Icelandic and UK data sets identifies missense variants in SMO, IL11, COL11A1 and 13 more new loci associated with osteoarthritis[J/OL]. Nature Genetics, 2018, 50(12):1681-1687[2022-06-08]. <http://www.nature.com/articles/s41588-018-0247-0>. DOI: 10.1038/s41588-018-0247-0.
- [9] arcOGEN Consortium, Tachmazidou I, Hatzikotoulas K, et al. Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data[J/OL]. Nature Genetics, 2019, 51(2):230-236[2022-06-08]. <http://www.nature.com/articles/s41588-018-0327-1>. DOI: 10.1038/s41588-018-0327-1.
- [10] Zengini E, Hatzikotoulas K, Tachmazidou I, et al. Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis[J/OL]. Nature Genetics, 2018, 50(4): 549-558[2022-06-08]. <http://www.nature.com/articles/s41588-018-0079-y>. DOI: 10.1038/s41588-018-0079-y.

- [11] Choi S W, Mak T S H, O' Reilly P F. Tutorial: a guide to performing polygenic risk score analyses[J/OL]. *Nature Protocols*, 2020, 15(9):2759-2772[2022-06-08]. <http://www.nature.com/articles/s41596-020-0353-1>. DOI: 10.1038/s41596-020-0353-1.
- [12] Jostins L, Barrett J C. Genetic risk prediction in complex disease[J/OL]. *Human Molecular Genetics*, 2011, 20(R2):R182-R188[2022-06-08]. <https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddr378>.
- [13] Wray N R, Lee S H, Mehta D, et al. Research Review: Polygenic methods and their application to psychiatric traits[J/OL]. *Journal of Child Psychology and Psychiatry*, 2014, 55(10):1068-1087[2022-06-08]. <https://onlinelibrary.wiley.com/doi/10.1111/jcpp.12295>.
- [14] So H C, Sham P C. Exploring the predictive power of polygenic scores derived from genome-wide association studies: a study of 10 complex traits[J/OL]. *Bioinformatics*, 2017:btw745[2022-06-08]. <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw745>.
- [15] López B, Torrent-Fontbona F, Viñas R, et al. Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction[J/OL]. *Artificial Intelligence in Medicine*, 2018, 85:43-49[2022-06-08]. <https://linkinghub.elsevier.com/retrieve/pii/S0933365717300684>. DOI: 10.1016/j.artmed.2017.09.005.
- [16] Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information[J/OL]. *Bioinformatics*, 2006, 22(22):2729-2734[2022-06-08]. <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl423>.
- [17] Cruz J A, Wishart D S. Applications of Machine Learning in Cancer Prediction and Prognosis [J/OL]. *Cancer Informatics*, 2006, 2:117693510600200[2022-06-08]. <http://journals.sagepub.com/doi/10.1177/117693510600200030>.
- [18] Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques [C/OL]//2008 IEEE/ACS International Conference on Computer Systems and Applications. 2008: 108-115. DOI: 10.1109/AICCSA.2008.4493524.
- [19] Yu W, Liu T, Valdez R, et al. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes[J/OL]. *BMC Medical Informatics and Decision Making*, 2010, 10(1):16[2022-06-08]. <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-10-16>.
- [20] Zhang D, Shen D. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease[J/OL]. *NeuroImage*, 2012, 59(2): 895-907[2022-06-08]. <https://linkinghub.elsevier.com/retrieve/pii/S105381191101144X>. DOI: 10.1016/j.neuroimage.2011.09.069.
- [21] Montañez C A C, Fergus P, Montañez A C, et al. Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs[J/OL]. arXiv:1804.03198 [cs, q-bio], 2018[2022-06-08]. <http://arxiv.org/abs/1804.03198>.
- [22] Bronstein M M, Bruna J, LeCun Y, et al. Geometric deep learning: going beyond Euclidean data[J/OL]. *IEEE Signal Processing Magazine*, 2017, 34(4):18-42[2022-06-08]. <http://arxiv.org/abs/1611.08097>. DOI: 10.1109/MSP.2017.2693418.
- [23] Sperduti A, Starita A. Supervised neural networks for the classification of structures[J/OL]. *IEEE Transactions on Neural Networks*, 1997, 8(3):714-735[2022-06-08]. <https://ieeexplore.ieee.org/document/572108/>. DOI: 10.1109/72.572108.

- [24] Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains[C/OL]//Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.: volume 2. 2005: 729-734 vol. 2. DOI: 10.1109/IJCNN.2005.1555942.
- [25] Xu K, Hu W, Leskovec J, et al. How Powerful are Graph Neural Networks?[J/OL]. arXiv:1810.00826 [cs, stat], 2019[2022-06-08]. <http://arxiv.org/abs/1810.00826>.
- [26] Veličković P, Cucurull G, Casanova A, et al. Graph Attention Networks[J/OL]. arXiv:1710.10903 [cs, stat], 2018[2022-06-08]. <http://arxiv.org/abs/1710.10903>.
- [27] Atwood J, Towsley D. Diffusion-Convolutional Neural Networks[J/OL]. arXiv:1511.02136 [cs], 2016 [2022-06-08]. <http://arxiv.org/abs/1511.02136>.
- [28] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks[J/OL]. arXiv:1609.02907 [cs, stat], 2017[2022-06-08]. <http://arxiv.org/abs/1609.02907>.
- [29] Defferrard M, Bresson X, Vandergheynst P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering[J/OL]. arXiv:1606.09375 [cs, stat], 2017[2022-06-08]. <http://arxiv.org/abs/1606.09375>.
- [30] Li R, Wang S, Zhu F, et al. Adaptive Graph Convolutional Neural Networks[J/OL]. arXiv:1801.03226 [cs, stat], 2018[2022-06-08]. <http://arxiv.org/abs/1801.03226>.
- [31] Veselkov K, Gonzalez G, Aljifri S, et al. HyperFoods: Machine intelligent mapping of cancer-beating molecules in foods[J/OL]. Scientific Reports, 2019, 9(1):9237[2022-06-08]. <http://www.nature.com/articles/s41598-019-45349-y>. DOI: 10.1038/s41598-019-45349-y.
- [32] Knyazev B, Lin X, Amer M R, et al. Spectral Multigraph Networks for Discovering and Fusing Relationships in Molecules[J/OL]. arXiv:1811.09595 [cs, stat], 2018[2022-06-08]. <http://arxiv.org/abs/1811.09595>.
- [33] Yan S, Xiong Y, Lin D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition[J/OL]. arXiv:1801.07455 [cs], 2018[2022-06-08]. <http://arxiv.org/abs/1801.07455>.
- [34] Ghosal S, Chen Q, Pergola G, et al. A Biologically Interpretable Graph Convolutional Network to Link Genetic Risk Pathways and Neuroimaging Markers of Disease[R/OL]. Bioinformatics, 2021 [2022-06-08]. <http://biorxiv.org/lookup/doi/10.1101/2021.05.28.446066>.
- [35] Ying R, Bourgeois D, You J, et al. GNNExplainer: Generating Explanations for Graph Neural Networks [J/OL]. arXiv:1903.03894 [cs, stat], 2019[2022-06-08]. <http://arxiv.org/abs/1903.03894>.
- [36] Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age[J/OL]. PLOS Medicine, 2015, 12(3):e1001779[2022-06-08]. <https://dx.plos.org/10.1371/journal.pmed.1001779>.
- [37] [J/OL]. World Health Organization. <https://icd.who.int/en>.
- [38] Chen Z, Boehnke M, Wen X, et al. Revisiting the genome-wide significance threshold for common variant GWAS[J/OL]. G3 Genes|Genomes|Genetics, 2021, 11(2):jkaa056[2022-06-08]. <https://academic.oup.com/g3journal/article/doi/10.1093/g3journal/jkaa056/6080665>.
- [39] Wilk M B, Gnanadesikan R. Probability plotting methods for the analysis for the analysis of data [J/OL]. Biometrika, 1968, 55(1):1-17[2022-06-08]. <https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/55.1.1>.
- [40] Marees A T, de Kluiver H, Stringer S, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis[J/OL]. International Journal of Methods in Psychiatric Research, 2018, 27(2):e1608[2022-06-08]. <https://onlinelibrary.wiley.com/doi/10.1002/mpr.1608>.

- [41] Janes H. The optimal ratio of cases to controls for estimating the classification accuracy of a biomarker [J/OL]. Biostatistics, 2005, 7(3):456-468[2022-06-08]. <https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxj018>.
- [42] Tibshirani R. Regression Shrinkage and Selection Via the Lasso[J/OL]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1):267-288[2022-06-08]. <https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1996.tb02080.x>.
- [43] Fung G M, Mangasarian O. A Feature Selection Newton Method for Support Vector Machine Classification[J/OL]. Computational Optimization and Applications, 2004, 28(2):185-202[2022-06-08]. <https://doi.org/10.1023/B:COAP.0000026884.66338.df>.
- [44] Hu J, Cheng R, Huang Z, et al. On Embedding Uncertain Graphs[C/OL]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore Singapore: ACM, 2017: 157-166[2022-06-08]. <https://dl.acm.org/doi/10.1145/3132847.3132885>.
- [45] Monti F, Boscaini D, Masci J, et al. Geometric deep learning on graphs and manifolds using mixture model CNNs[J/OL]. arXiv:1611.08402 [cs], 2016[2022-06-08]. <http://arxiv.org/abs/1611.08402>.
- [46] Zhu Y, Xu W, Zhang J, et al. A Survey on Graph Structure Learning: Progress and Opportunities [J/OL]. arXiv:2103.03036 [cs], 2022[2022-06-08]. <http://arxiv.org/abs/2103.03036>.
- [47] Preparata F P, Shamos M I. Computational Geometry[M/OL]. New York, NY: Springer New York, 1985[2022-06-08]. <http://link.springer.com/10.1007/978-1-4612-1098-6>.
- [48] Bentley J L, Stanat D F, Williams E. The complexity of finding fixed-radius near neighbors[J/OL]. Information Processing Letters, 1977, 6(6):209-212[2022-06-08]. <https://linkinghub.elsevier.com/retrieve/pii/0020019077900709>. DOI: 10.1016/0020-0190(77)90070-9.
- [49] Holland P W, Laskey K B, Leinhardt S. Stochastic blockmodels: First steps[J/OL]. Social Networks, 1983, 5(2):109-137[2022-06-08]. <https://linkinghub.elsevier.com/retrieve/pii/0378873383900217>. DOI: 10.1016/0378-8733(83)90021-7.
- [50] Zhang Y, Pal S, Coates M, et al. Bayesian Graph Convolutional Neural Networks for Semi-Supervised Classification[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33:5829-5836[2022-06-08]. <https://aimagazine.org/ojs/index.php/AAAI/article/view/4531>. DOI: 10.1609/aaai.v33i01.33015829.
- [51] Elinas P, Bonilla E V, Tiao L. Variational Inference for Graph Convolutional Networks in the Absence of Graph Data and Adversarial Settings[J/OL]. arXiv:1906.01852 [cs, stat], 2020[2022-06-08]. <http://arxiv.org/abs/1906.01852>.
- [52] Hamilton W L, Ying R, Leskovec J. Inductive Representation Learning on Large Graphs[J/OL]. arXiv:1706.02216 [cs, stat], 2018[2022-06-08]. <http://arxiv.org/abs/1706.02216>.
- [53] Wang R, Mou S, Wang X, et al. Graph Structure Estimation Neural Networks[C/OL]//Proceedings of the Web Conference 2021. Ljubljana Slovenia: ACM, 2021: 342-353[2022-06-08]. <https://dl.acm.org/doi/10.1145/3442381.3449952>.
- [54] Dempster A P, Laird N M, Rubin D B. Maximum Likelihood from Incomplete Data Via the EM Algorithm[J/OL]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1): 1-22[2022-06-08]. <https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1977.tb01600.x>.
- [55] Jensen J L W V. Sur les fonctions convexes et les inégalités entre les valeurs moyennes[J/OL]. Acta Mathematica, 1906, 30(0):175-193[2022-06-08]. <http://projecteuclid.org/euclid.acta/1485887155>. DOI: 10.1007/BF02418571.

- [56] Jordan M I, Ghahramani Z, Jaakkola T S, et al. An Introduction to Variational Methods for Graphical Models[J/OL]. *Machine Learning*, 1999, 37(2):183-233[2022-06-08]. <https://doi.org/10.1023/A:1007665907178>.
- [57] Jaakkola T S, Jordan M I. Bayesian parameter estimation via variational methods[J/OL]. *Statistics and Computing*, 2000, 10(1):25-37[2022-06-08]. <https://doi.org/10.1023/A:1008932416310>.
- [58] Tzikas D G, Likas A C, Galatsanos N P. The variational approximation for Bayesian inference[J]. *IEEE Signal Processing Magazine*, 2008, 25(6):131.
- [59] Neal R M, Hinton G E. A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants[M/OL]//Jordan M I. *Learning in Graphical Models*. Dordrecht: Springer Netherlands, 1998: 355-368[2022-06-08]. https://doi.org/10.1007/978-94-011-5014-9_12.
- [60] Haj A E. Stochastics blockmodels, classifications and applications[D/OL]. Université de Poitiers ; Université Libanaise, 2019[2022-06-08]. <https://tel.archives-ouvertes.fr/tel-02926379>.
- [61] Haj A E, Slaoui Y, Louis P Y, et al. Estimation in a binomial stochastic blockmodel for a weighted graph by a variational expectation maximization algorithm[J/OL]. *Communications in Statistics - Simulation and Computation*, 2020:1-20[2022-06-08]. <https://www.tandfonline.com/doi/full/10.1080/03610918.2020.1743858>.
- [62] Schwarz G. Estimating the Dimension of a Model[J/OL]. *The Annals of Statistics*, 1978, 6(2)[2022-06-08]. <https://projecteuclid.org/journals/annals-of-statistics/volume-6/issue-2/Estimating-the-Dimension-of-a-Model/10.1214/aos/1176344136.full>.
- [63] Daudin J J, Picard F, Robin S. A mixture model for random graphs[J/OL]. *Statistics and Computing*, 2008, 18(2):173-183[2022-06-08]. <https://doi.org/10.1007/s11222-007-9046-7>.
- [64] Proakis J G, Manolakis D G. Prentice Hall international editions: Digital signal processing: principles, algorithms, and applications[M]. 3. ed ed. London: Prentice-Hall International (UK), 1996.
- [65] Stahlschmidt S R, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review[J/OL]. *Briefings in Bioinformatics*, 2022, 23(2):bbab569[2022-06-08]. <https://academic.oup.com/bib/article/doi/10.1093/bib/bbab569/6516346>.
- [66] Boer C G, Hatzikotoulas K, Southam L, et al. Deciphering osteoarthritis genetics across 826,690 individuals from 9 populations[J/OL]. *Cell*, 2021, 184(18):4784-4818.e17[2022-06-08]. <https://linkinghub.elsevier.com/retrieve/pii/S0092867421009417>. DOI: 10.1016/j.cell.2021.07.038.
- [67] Gaudillo J, Rodriguez J J R, Nazareno A, et al. Machine learning approach to single nucleotide polymorphism-based asthma prediction[J/OL]. *PLOS ONE*, 2019, 14(12):e0225574[2022-06-08]. <https://dx.plos.org/10.1371/journal.pone.0225574>.
- [68] Fawcett T. An introduction to ROC analysis[J/OL]. *Pattern Recognition Letters*, 2006, 27(8): 861-874[2022-06-08]. <https://www.sciencedirect.com/science/article/pii/S016786550500303X>. DOI: 10.1016/j.patrec.2005.10.010.
- [69] Zou K H, O' Malley A J, Mauri L. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models[J/OL]. *Circulation*, 2007, 115(5):654-657[2022-06-08]. <https://www.ahajournals.org/doi/10.1161/circulationaha.105.594929>. DOI: 10.1161/CIRCULATIONAHA.105.594929.
- [70] Naaman M. On the tight constant in the multivariate Dvoretzky - Kiefer - Wolfowitz inequality [J/OL]. *Statistics & Probability Letters*, 2021, 173:109088[2022-06-08]. <https://linkinghub.elsevier.com/retrieve/pii/S016771522100050X>. DOI: 10.1016/j.spl.2021.109088.
- [71] Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores[J/OL]. *PLoS Genetics*, 2013, 9(3):e1003348[2022-06-08]. <https://dx.plos.org/10.1371/journal.pgen.1003348>.

- [72] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J/OL]. Proceedings of the IEEE, 1998, 86(11):2278-2324. DOI: 10.1109/5.726791.
- [73] Pendergrass S, Brown-Gentry K, Dudek S, et al. The use of genome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery[J/OL]. Genetic Epidemiology, 2011, 35(5):410-422[2022-06-08]. <https://onlinelibrary.wiley.com/doi/10.1002/gepi.20589>.
- [74] Watanabe K, Stringer S, Frei O, et al. A global overview of pleiotropy and genetic architecture in complex traits[J/OL]. Nature Genetics, 2019, 51(9):1339-1348. DOI: 10.1038/s41588-019-0481-0.
- [75] Jensen L K. Hip osteoarthritis: influence of work with heavy lifting, climbing stairs or ladders, or combining kneeling/squatting with heavy lifting[J/OL]. Occupational and Environmental Medicine, 2008, 65(1):6-19[2022-06-08]. <https://oem.bmjjournals.org/lookup/doi/10.1136/oem.2006.032409>.
- [76] Maruthur N M, Tseng E, Hutfless S, et al. Diabetes Medications as Monotherapy or Metformin-Based Combination Therapy for Type 2 Diabetes: A Systematic Review and Meta-analysis[J/OL]. Annals of Internal Medicine, 2016, 164(11):740[2022-06-08]. <http://annals.org/article.aspx?doi=10.7326/M15-2650>.
- [77] Barrett T J, Murphy A J, Goldberg I J, et al. Diabetes-mediated myelopoiesis and the relationship to cardiovascular risk: Diabetes myelopoiesis and cardiovascular risk[J/OL]. Annals of the New York Academy of Sciences, 2017, 1402(1):31-42[2022-06-08]. <https://onlinelibrary.wiley.com/doi/10.1111/nyas.13462>.

附录 A 外文文献原文



Peter Kraft,³² Steven A. Lietman,³³ Dino Samartzis,^{30,34} P. Eline Slagboom,⁵ Kari Stefansson,^{3,14} Unnur Thorsteinsdottir,^{3,14} Jonathan H. Tobias,^{9,7} André G. Uitterlinden,¹ Bendik Winsvold,^{10,35,36} John-Anker Zwart,^{10,35} George Davey Smith,^{7,37} Pak Chung Sham,³⁸ Gudmar Thorleifsson,³ Tom R. Gaunt,⁷ Andrew P. Morris,³⁹ Ana M. Valdes,⁴⁰ Aspasia Tsezou,⁴¹ Kathryn S.E. Cheah,⁴² Shiro Ikegawa,²⁴ Kristian Hveem,^{10,43} Tõnu Esko,⁹ J. Mark Wilkinson,⁴⁴ Ingrid Meulenbelt,⁵ Ming Ta Michael Lee,^{4,45} Joyce B.J. van Meurs,¹ Unnur Styrkársdóttir,³ and Eleftheria Zeggini^{2,46,48,*}

¹⁷Research and Communication Unit for Musculoskeletal Health (FORMI), Department of Research, Innovation and Education, Division of Clinical Neuroscience, Oslo University Hospital, 0424 Oslo, Norway

¹⁸Department of Medicine, Landspítali The National University Hospital of Iceland, 108 Reykjavík, Iceland

¹⁹Departments of Rheumatology and Clinical Epidemiology, Leiden University Medical Center, 9600, 23OORC Leiden, the Netherlands

²⁰Department of Twin Research and Genetic Epidemiology, Kings College London, London SE1 7EH, UK

²¹Department of Orthopaedics, Leiden University Medical Center, 9600, 23OORC Leiden, the Netherlands

²²Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

²³Daffodil Centre, The University of Sydney, a joint venture with Cancer Council NSW, Sydney, NSW 1340, Australia

²⁴Laboratory for Bone and Joint Diseases, RIKEN Center for Integrative Medical Sciences, Tokyo 108-8639, Japan

²⁵Department of Orthopedic Surgery, Shimane University, Shimane 693-8501, Japan

²⁶Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, 7491 Trondheim, Norway

²⁷BioCore-Bioinformatics Core Facility, Norwegian University of Science and Technology, 7491 Trondheim, Norway

²⁸Clinic of Laboratory Medicine, St. Olavs Hospital, Trondheim University Hospital, 7030 Trondheim, Norway

²⁹²¹ Department of Orthopaedics, National and Kapodistrian University of Athens, Medical School, Nea Ionia General Hospital Konstantopoulio, 14233 Athens, Greece

³⁰Department of Orthopaedics and Traumatology, The University of Hong Kong, Pokfulam, Hong Kong, China

³¹Department of Medicine, Brigham and Women's Hospital, 181 Longwood Ave, Boston, MA 02115, USA

³²Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA

³³Musculoskeletal Institute, Geisinger Health System, Danville, PA 17822, USA

³⁴Department of Orthopaedic Surgery, Rush University Medical Center, Chicago, IL 60612, USA

³⁵Department of Research, Innovation and Education, Division of Clinical Neuroscience, Oslo University Hospital and University of Oslo, 0450 Oslo, Norway

³⁶Department of Neurology, Oslo University Hospital, 0424 Oslo, Norway

³⁷Population Health Sciences, Bristol Medical School, University of Bristol, Bristol BS8 2BN, UK

³⁸Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong, China

³⁹Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, University of Manchester, Manchester M13 9LJ, UK

⁴⁰Faculty of Medicine and Health Sciences, School of Medicine, University of Nottingham, Nottingham, Nottinghamshire NG5 1PB, UK

⁴¹Laboratory of Cytogenetics and Molecular Genetics, Faculty of Medicine, University of Thessaly, Larissa 411 10, Greece

⁴²School of Biomedical Sciences, The University of Hong Kong, Pokfulam, Hong Kong, China

⁴³HUNT Research Center, Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, 7600 Levanger, Norway

⁴⁴Department of Oncology and Metabolism and Healthy Lifespan Institute, University of Sheffield, Sheffield S10 2RX, UK

⁴⁵Institute of Biomedical Sciences, Academia Sinica, 115 Taipei, Taiwan

⁴⁶TUM School of Medicine, Technical University of Munich and Klinikum Rechts der Isar, 81675 Munich, Germany

⁴⁷These authors contributed equally

⁴⁸Lead contact

*Correspondence: eleftheria.zeggini@helmholtz-muenchen.de
<https://doi.org/10.1016/j.cell.2021.07.038>

Osteoarthritis is a complex degenerative disease of the whole joint, characterized by cartilage degeneration, subchondral bone thickening, osteophyte formation, synovial inflammation, and structural alterations of the joint capsule, ligaments, and associated muscles (Hunter and Bierma-Zeinstra, 2019). Recently, advances were made in elucidating the genetic background of osteoarthritis, using genome-wide association studies (GWAS) (Styrkarsdóttir et al., 2018; Tachmazidou et al., 2019; Tachmazidou et al., 2017; Zengini et al., 2018), with 96 statistically independent risk variants reported to date. These variants only explain a small proportion of the phenotypic variance (Tachmazidou et al., 2019) and are mainly associated with osteoarthritis affecting the knee and hip joints.

Osteoarthritis can affect every synovial joint and an increase in body mass index (BMI) is associated with risk of disease (Geu-

sens and van den Berg, 2016). A better understanding of the genetic differences between weight bearing (knee, hip, and spine) and non-weight bearing joints (hand, finger, and thumb) is needed to help disentangle the metabolic and biomechanical effects contributing to disease development. Here, we conducted a GWAS meta-analysis across knee, hip, finger, thumb, and spine osteoarthritis phenotypes in 826,690 individuals of European and East Asian descent. We integrated functional genomics analyses from disease-relevant tissue, including gene expression, protein abundance and genome-wide methylation, mouse knockout model and monogenic human disease phenotyping data, and complementary computational fine-mapping, colocalization, and causal inference approaches to identify likely effector genes and facilitate much-needed translation into therapies by enhancing our understanding of disease etiopathology.

RESULTS

Genetic architecture

Identification of osteoarthritis risk variants

We performed GWAS meta-analyses for osteoarthritis across 13 international cohorts stemming from 9 populations (Table S1), in up to 826,690 individuals (177,517 osteoarthritis patients). This is a substantial (2.3-fold) increase of osteoarthritis patient numbers compared to the largest osteoarthritis GWAS to date. Two of the cohorts are of East Asian and 11 of the cohorts are of European descent. We defined 11 phenotypes encompassing all major sites for osteoarthritis (Figure 1; Table S1; STAR Methods). We found 11,897 genome-wide significantly associated single nucleotide variants (SNVs) using a threshold of $p < 1.3 \times 10^{-8}$, to account for the effective number of independent tests. We applied conditional analyses within phenotype and identified 223 independent associations, some of which overlap across phenotypes (Figure 1; Table 1). Eighty-four variants have not been associated with osteoarthritis before. We investigated the previously reported osteoarthritis-loci and found that 87 out of 96 replicated in the same direction at nominal significance (Table S2).

We used conditional analyses to identify associations that do not overlap across disease phenotype definitions. We identified 100 unique and independent variant associations, 60 of which were associated with more than one osteoarthritis phenotype. Fifty-two variants have not been associated with any osteoarthritis phenotype before (Tables 2 and S3). For each of the 100 association signals, we defined the lead SNV as the risk variant with the strongest statistical evidence for association. Six lead SNVs are coding (all missense), 59 reside within a gene transcript, and 35 are intergenic.

Here, we report signals for spine ($n = 1$) and thumb ($n = 2$) osteoarthritis and increase the number of risk SNVs for hand (5 new, 3 previously reported) and finger (3 new, 2 previously reported) osteoarthritis, phenotypes that had not been studied at scale before (Tables 1, 2, and S3). Of the 100 SNVs, 90 are common (minor allele frequency [MAF] $\geq 5\%$) and 4 are low-frequency variants (MAF $< 5\%$ and $\geq 0.5\%$). We detected 6 rare variant associations (MAF 0.03%–0.11%) with large effect sizes (odds ratio [OR] range = 3.03–9.52) (Table 2); 1 variant association was previously reported and 5 variant associations are new findings. All of the new rare variant associations are primarily driven by a large extended family in Iceland.

Signals from 4 osteoarthritis phenotypes (spine, knee, knee and/or hip, and osteoarthritis at any site) included individuals of non-European ancestry (between 0.9%–2.8% of cases were of East Asian descent). Even though sample sizes in the East Asian cohorts are small, we observed that 62% of the signals have supportive evidence in East Asian ancestry-only analysis, with the same direction of effect, and 20% of these signals are also nominally significant (binomial test $p = 2.27 \times 10^{-5}$, 95% confidence interval [CI] = 7%–100%) (STAR Methods).

We investigated the predictive power of polygenic risk scores (PRS) and found significantly higher odds of developing disease in individuals at the higher decile of the PRS distribution for several osteoarthritis phenotypes (Table S4; STAR Methods).

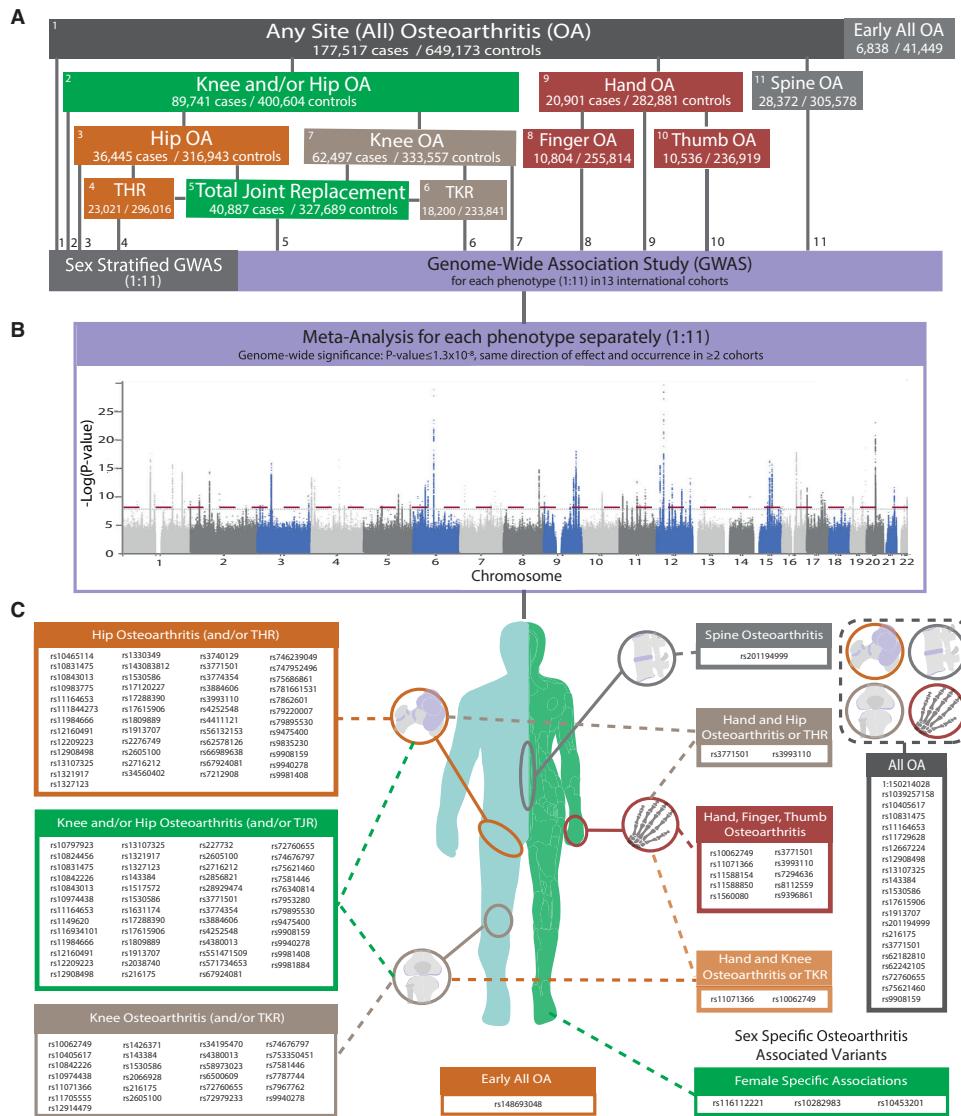
Female-specific osteoarthritis risk variants

To investigate the presence of osteoarthritis signals specific to males only, females only, or with effects of opposite direction in men and women, we performed a sex-differentiated test of association and a test of heterogeneity in allelic effects (Mägi et al., 2010; Mägi and Morris, 2010). We identified 3 new female-specific independent SNVs, two of which showed significant (Phet-diff < 0.016) differences in effect size between sexes (Tables 2 and S5). rs116112221 (Psex-diff = 3.20×10^{-9} , Phet-diff = 4.09×10^{-4} ; female OR = 1.95, 95% CI = 1.58–2.41, P-female = 4.61×10^{-10} ; male OR = 1.06, 95% CI = 0.82–1.38, P-male = 0.64) is significant in the female-only total hip replacement phenotype and is located in a region containing long intergenic non coding RNAs with the closest protein coding gene being FANCL. FANCL mutations are potentially causative for premature ovary insufficiency in humans (Yang et al., 2020), a condition that leads to early menopause, which has been suggested to be linked to increased prevalence of osteoarthritis, although definitive evidence for this hypothesis is still lacking (Jung et al., 2018; Srikanth et al., 2005). Preclinical and clinical studies indicate that selective estrogen receptor modulators (SERMs) treatment has consistently positive effects on osteoarthritis, especially for postmenopausal patients with early-stage or osteoporotic osteoarthritis (Xiao et al., 2016).

We further identified a signal associated with total hip replacement with opposite direction of effects between men and women, rs10282983 (Psex-diff = 4.93×10^{-16} , Phet-diff = 7.66×10^{-14} ; female OR = 1.15, 95% CI = 1.11–1.19, P-female = 2.21×10^{-14} ; male OR = 0.92, 95% CI = 0.88–0.96, P-male = 5.16×10^{-4}). rs10282983 resides in an intron of C8orf34, which has been associated with waist-to-hip ratio (Kichaev et al., 2019; Pulit et al., 2019) and heel bone mineral density (Kichaev et al., 2019), both risk factors for osteoarthritis (Hardcastle et al., 2015; Lohmander et al., 2009). rs10453201 is significantly associated with female osteoarthritis at any site (Psex-diff = 5.67×10^{-9} , Phet-diff = 0.049; female OR = 1.05, 95% CI = 1.03–1.06, P-female = 1.05×10^{-8} ; male OR = 1.02, 95% CI = 1.003–1.04, P-male = 0.02) and is located 5' of UBAP2, which has been associated with Parkinson's disease (Nalls et al., 2019), type 2 diabetes (Xue et al., 2018), BMI (Kichaev et al., 2019), and heel bone mineral density (Morris et al., 2019) in humans.

Early-onset osteoarthritis

Genome-wide meta-analysis identified a new risk variant for early osteoarthritis with large effect size and low allele frequency (rs148693048; effect allele frequency = 0.12%, $p = 3.37 \times 10^{-8}$, OR = 6.26, 95% CI = 3.26–12.00) (Tables 2 and S3). The variant is nominally significantly associated in all contributing studies and with the same direction of effect. rs148693048 has not been associated with osteoarthritis before. Two protein-coding genes in the vicinity show significantly different expression levels in intact compared to degraded cartilage (NEFM and DOCK5). NEFM (neurofilament medium) is relevant to the elongation of neuronal structures (Pezzini et al., 2017), and the expressed protein is commonly used as a biomarker of neuronal damage (Khalil et al., 2018). The guanine nucleotide exchange activity of DOCK5

**Figure 1. Genetic architecture**

Graphical summary of the Genetics of Osteoarthritis Consortium workflow and results.

(A) Overview of the 11 defined osteoarthritis phenotypes, sex specific analysis, their relationship with each other and their sample sizes (cases/controls). TKR, total knee replacement; THR, total hip replacement.

(B) Merged Manhattan-plot of all individual meta-analysis results of all 11 examined osteoarthritis phenotypes. The dashed line represents the genome-wide significance threshold $p = 1.3 \times 10^{-8}$.

(C) Graphical overview of all lead genome-wide significant independent osteoarthritis associated single nucleotide variants (SNVs) and the osteoarthritis phenotypes with which they are associated.

See also [Table S1](#).

Table 1. Summary results for all genome-wide significant osteoarthritis associated SNVs

Genome-wide association study	Cases/controls	Signals ^b	New signals ^b	Known signals ^b
All osteoarthritis ^a	177,517/649,173	21	8	13
Knee and/or hip osteoarthritis	89,741/400,604	31	9	22
Hip osteoarthritis	36,445/316,943	45	17	28
Knee osteoarthritis	62,497/333,557	24	11	13
Total hip replacement	23,021/296,016	38	12	26
Total knee replacement	18,200/233,841	10	4	6
Total joint replacement	40,887/327,689	37	12	25
Hand osteoarthritis	20,901/282,881	7	5	2
Finger osteoarthritis	10,804/255,814	5	3	2
Thumb osteoarthritis	10,536/236,919	4	2	2
Spine osteoarthritis	28,372/305,578	1	1	0
Total	223	84	139	
Total independent signals across phenotypes ^c	100	52	48	

Sex-specific analysis

Female total hip replacement	11,089/67,516	2	2	0
Female all osteoarthritis	90,838/192,697	1	1	0

Early-onset osteoarthritis analysis

Early all osteoarthritis	6,838/41,449	1	1	0
--------------------------	--------------	---	---	---

Signals reported here are genome-wide significant ($p < 1.3 \times 10^{-8}$) with the exception of the early-onset analysis ($p < 5 \times 10^{-8}$).

^aCases are any-site osteoarthritis: hip, knee, hand, finger, thumb, and spine.

^bSignals numbers are based per defined osteoarthritis phenotype, new known are based on previously reported osteoarthritis loci.

^cIndependence calculated within and across osteoarthritis phenotypes, the lead SNP is assigned to the most significant phenotype (Table S3).

(dedicator of cytokinesis 5) has been identified as a regulator of osteoclast function, playing an essential role in bone resorption (Vives et al., 2011). Pharmacological inhibition of its activity prevents osteolysis, while preserving bone formation in both humans and mice (Mounier et al., 2020). Intronic variation in DOCK5 also shows association ($p < 5.0 \times 10^{-8}$) with other bone phenotypes, such as heel bone mineral density (Kim, 2018) and adolescent idiopathic scoliosis (Liu et al., 2018).

Cross-phenotype analysis

Similarities and differences of signals across phenotypes

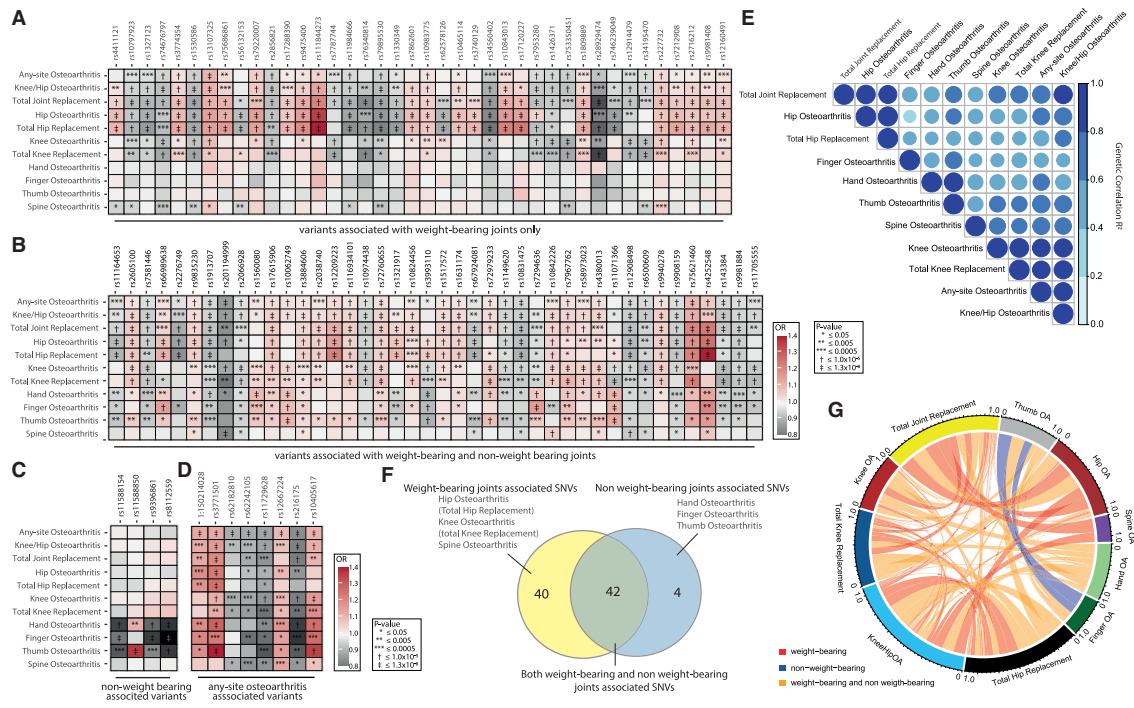
We observed that some variants demonstrate a joint-specific effect. We found that the majority of SNVs (60 out of the 100) were genome-wide significantly associated with more than one osteoarthritis phenotype (Figure 2). Forty of the identified SNVs show genome-wide significant associations with weight bearing joints only and 4 SNVs show genome-wide significant associations with non-weight bearing joints only (Figure 2; Table S3). We have over 80% power to detect all 4 non-weight bearing specific variants in the weight bearing joint analyses (at genome-wide significance). Further, we have over 80% power to detect 22 of the 40 weight bearing joint-specific effects in non-weight bearing joint analyses (hand osteoarthritis). Although several core pathways are known to underpin osteoarthritis pathology, regardless of joint site affected, no common genetic osteoarthritis SNVs have been found previously, with the exception of the *GDF5* locus (Reynard and Loughlin, 2013; Sandell, 2012). Here, we have identified 42 SNVs with strong association across both weight bearing and non-weight bearing joints. Several of these SNVs, rs3771501 (*TGFA*), rs3993110 (*TEAD1/DKK3*), rs72979233 (*CHRD1*), and rs7967762 (*PFKM/WNT10B*) (Figures 2B and 2D), are associated with multiple osteoarthritis joint sites. These signals likely represent a common underlying mechanism in osteoarthritis pathology. They have been shown to play a role in the transforming growth factor β (TGF- β)/bone morphogenic protein (BMP), Wnt/ β -catenin signaling pathways, the functional interaction of which has been implicated in the pathogenesis of osteoarthritis (Wu et al., 2012). These signaling pathways could be prime candidates for drug development.

Additional insights may also be gleaned from the comparison of association signals across osteoarthritis phenotypes. Most of the SNVs associated with knee, hip, and knee and/or hip osteoarthritis have a larger effect size on the respective joint replacement-defined phenotypes, all of which are notably of smaller sample size. This could be driven by homogeneity of phenotype definition (Manchia et al., 2013) (Table S1) or can represent a biological and functional relevance, indicating that these loci might play more important roles in receiving a joint replacement (i.e., pain and inflammation) than in osteoarthritis pathology itself. For example, rs76340814 (*PTCH1*) and rs28929474 (missense variant in *SERPINA1*) have stronger associations and larger effect sizes with total hip replacement (THR), total knee replacement (TKR), and total joint replacement (TJR), than with hip or knee osteoarthritis (Figure 2A). Indeed, *PTCH1* is thought to function in neurogenic and brain development (Mansilla et al., 2006; Ribeiro et al., 2010), and *SERPINA1* is thought to function in inflammation. Studies in rat osteoarthritis models have shown that early treatment with alpha-1-antiproteinase, encoded by *serpina1*, blocked the proteolytic activity of neutrophil elastase and caused lasting improvements in joint inflammation, pain, and saphenous nerve damage (Muley et al., 2017).

Genetic links between phenotypes

We found osteoarthritis subtypes to share substantial genetic components, albeit with a wide range (Figure 2E; Table S6).

We investigated if osteoarthritis genetic components are shared with other traits and found significant correlation with anthropometric traits (BMI, obesity, weight, and fat mass), type

**Figure 2. Similarities and differences of signals across phenotypes**

Correlation and overlap between osteoarthritis genetics

(A–D) Heatmap plots of osteoarthritis associated single nucleotide variants (SNVs). Effect sizes (OR, odds ratio) and p values are displayed for each lead SNP for each osteoarthritis phenotype GWAS results. OR are plotted as color, and p values are represented as symbols in the box. (A) Weight bearing joints only (hip, knee, and spine). (B) Both weight and non-weight bearing joints (hip, knee, spine, hand, finger, and thumb). (C) Non-weight bearing joints (hand, finger, and thumb). (D) Any-site osteoarthritis SNVs.

(E) Heatmap plot of the genetic correlation (R^2) between the examined osteoarthritis phenotypes.

(F) Venn diagram depicting the number and overlap of SNVs associated with weight bearing and non-weight bearing joints.

(G) Circos plot depicting the overlap in osteoarthritis associations of the 100 lead variants.

See also Table S6.

2 diabetes, education, depressive symptoms, smoking behavior, bone mineral density, reproductive phenotypes and intelligence as previously reported (Tachmazidou et al., 2019; Zengini et al., 2018), and several pain phenotypes (Table S6).

Pain is the most disabling symptom experienced by osteoarthritis patients and is one of the main reasons to proceed to physician consultation and total joint replacement (Schaible, 2018). The etiology of pain in osteoarthritis is multifactorial including significant soft tissue inflammation, the sensitization of pain pathways involving the joint nociceptors, the nociceptive processing in the CNS, and neuropathic pain components in osteoarthritis models (Dimitroulas et al., 2014; Fu et al., 2018; Hsia et al., 2018; Kidd, 2012). Although a main symptom, no genetic determinants of osteoarthritis pain have been discovered before. We found high correlation between osteoarthritis and sciatica, fibromyalgia, headaches, and other back pain phenotypes, where the highest correlation is with spine osteoarthritis (genetic correlation [rg] = 0.61, 0.87, 0.39, and 0.79, respectively). SOX5, one of the new signals, has been previously reported to be upregulated in human

osteoarthritis cartilage (Liu et al., 2020) and has been associated with back pain and with lumbar intervertebral disc degeneration (Suri et al., 2018). These findings are supported by animal model data, in which inactivation of SOX5 leads to defects in skeletogenesis such as in cartilage development, the notochord, and intervertebral discs in mice (Smits and Lefebvre, 2003; Smits et al., 2001). We also observed strong correlation between osteoarthritis and pain phenotypes in the LD-Hub database (all derived from the UK Biobank dataset), in particular between spine osteoarthritis and dorsalgia (rg = 0.87), leg pain on walking (rg = 0.82), knee pain (rg = 0.63), hip pain (rg = 0.76), back pain (rg = 0.75), and neck/shoulder pain (rg = 0.67) (Table S6). Thus, our data suggest that a proportion of the identified signals are also associated with osteoarthritis pain.

Effector genes and biological pathways

Identification of putative causal variants

We employed complementary computational approaches (STAR Methods) to fine-map the GWAS signals to a small set

of likely causal variants, identify relevant tissues based on signal enrichment (Figure S1), and provide mechanistic insights based on expression quantitative trait locus (eQTL) colocalization and causal inference analysis (Table S7). Twelve signals were fine-mapped to variant sets contained entirely within the transcript of a single gene with >95% posterior probability (PP), although we note that this does not provide conclusive evidence for the effector gene. Of note, *ALDH1A2*, which fine-maps to 6 intronic variants with 99% PP, is currently the target of approved drugs in use for other indications, providing a potential opportunity for drug repositioning (Sumita et al., 2017) (Table S8).

For 6 SNVs (3 new and 3 known), a single variant could be postulated as causal with >95% PP (Table S8).

Amassing evidence to identify effector genes

We assessed if any of the genes residing within 1 Mb of the osteoarthritis-associated lead variants showed differential gene expression and protein abundance in primary osteoarthritis-affected tissue in chondrocytes extracted from osteoarthritis patients undergoing joint replacement surgery. Similarly, we compared gene expression of subchondral bone tissue underneath the intact and degraded cartilage tissue (Tables S9 and S10). By combining results from the complementary functional genomics and computational approaches (outlined above), we identified 637 genes with at least one line of evidence pointing to a putative effector gene (Table S10). For these 637 genes, we combined supportive information from the fine-mapping, eQTL colocalization analyses, animal model data, human musculoskeletal and neuronal phenotype data, functional genomics, and causal inference analysis and identified 77 genes that have at least 3 different lines of evidence in support of their role as an effector gene (Tables 3 and S10). Of these 77 genes, 4 are supported by missense lead variants (rs2276749 in *VGLL4*, rs3740129 in *CHST3*, rs143083812 in *SMO*, and rs4252548 in *IL11*). Forty eight provide strong additional evidence for the likely effector gene at previously reported osteoarthritis-associated SNVs (Table 3) and 30 reside in newly associated signals.

CHST3, *SMAD3*, and *GDF5* accrued the highest levels of confidence, each with 6 different lines of evidence in support of their involvement in osteoarthritis. *CHST3* (carbohydrate sulfotransferase 3) represents a newly identified signal and encodes chondroitin sulfate, the major proteoglycan present in cartilage. Mutations in *CHST3* have been previously associated with short stature, congenital joint dislocations, clubfoot, Larsen syndrome, and elbow joint dysplasia (Superti-Furga and Unger, 1993; Unger et al., 2010). *CHST3* has also been shown to be associated with lumbar disc degeneration (Song et al., 2013).

To glean further insight into the biological role of the high-confidence effector genes in disease processes, we integrated additional information based on the analysis of endophenotypes more closely related to the underlying biology, monogenic and rare human disease data, genome-wide analyses, and additional functional genomics data (Tables S11 and S12; STAR Methods). By synthesizing all lines of evidence, we found that the assignment of several of the 77 high-confidence effector genes into likely mechanisms through which they exert their effect traverses multiple biological processes (Figure 3A). Here, we primarily focus on the newly associated genes that are re-

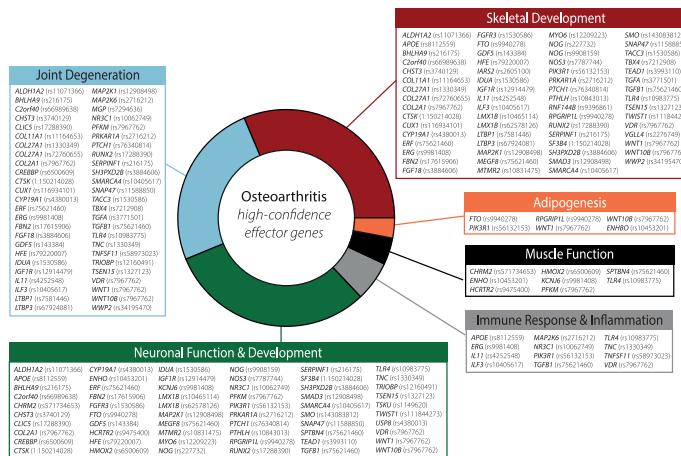
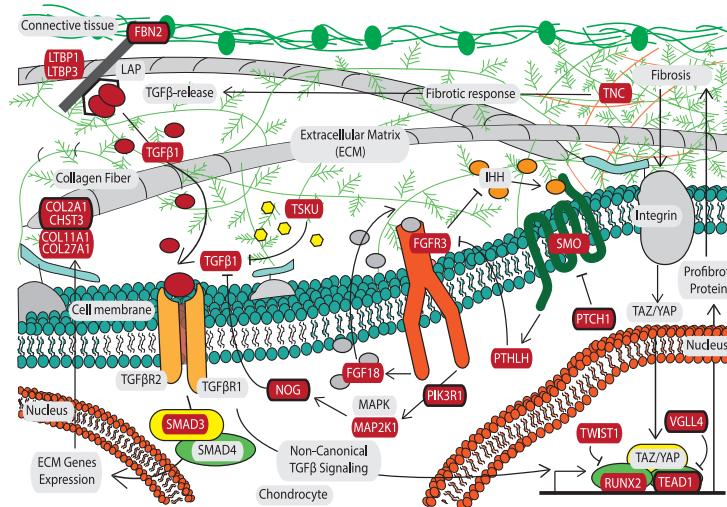
ported in this work. These represent high-value candidates for further mechanistic and clinical investigation.

The majority of high-confidence effector genes are associated with skeletal development (63 in total, 21 genes associated with newly reported signals) and joint degradation (50 in total, 18 genes associated with newly reported signals; 13 genes in common between the skeletal development and joint degradation categories) (Figure 3A). Three effector genes arising from new genetic signals encode structural proteins: *CHST3*, *COL2A1*, and *FBN2*. Collagen type II alpha 1 chain (*COL2A1*) codes for an essential structural component of cartilage and is important for joint formation and bone growth (Figure 3B). A wide spectrum of diseases is associated with *COL2A1*, including cartilage and bone abnormalities, such as spondyloepimetaphyseal dysplasia, Kniest dysplasia, and early onset osteoarthritis (Kuivaniemi et al., 1991; Löppönen et al., 2004; Wilkin et al., 1999; Xiong et al., 2018). Fibrillin 2 (*FBN2*) encodes a glycoprotein that forms microfibrils in the extracellular matrix and has a major role during early morphogenesis. Fibrillins potently regulate pathways of the immune response, inflammation, and tissue homeostasis (Zeyer and Reinhardt, 2015), are important in bone remodeling, and regulate local availability of BMP and TGF- β (Nistala et al., 2010) (Figure 3B). Mutations in *FBN2* cause contractual arachnodactyly (Putnam et al., 1995).

Several genes are connected with signaling pathways. Vestigial like family member 4 (*VGLL4*) functions via interacting with TEA domain (TEAD) transcription factors (Jiao et al., 2017; Lin et al., 2016). Notably, we identified another new THR and hand osteoarthritis-associated signal located in such a transcription factor, the *TEAD1* gene, indicating a common molecular pathway underlying both signals (Figure 3B). *TEAD1* functions in the Hippo signaling pathway and is transcriptionally regulated by the YAP1 and TAZ protooncogene proteins, which are involved in mechanosensing and mechanotransduction (Dupont et al., 2011; Low et al., 2014). Mechanoadaptation of articular cartilage is an important factor in osteoarthritis (Vincent and Wann, 2019; Zhao et al., 2020). Downregulation of *VGLL4* is linked to the upregulation of Wnt/ β -catenin pathway target genes (Jiao et al., 2017).

Wnt family member 1 (*WNT1*) and wnt family member 10B (*WNT10B*) are involved in the Wnt signaling pathway, which has an established role in osteoarthritis pathogenesis (Zhou et al., 2017). Mutations in *WNT10B* have been linked to limb defects and dental abnormalities (Kantaputra et al., 2018; Ullah et al., 2018; Yu et al., 2016), and mutations in *WNT1* are associated with osteogenesis imperfecta (Fahiminiya et al., 2013). Insulin-like growth factor 1 receptor (*IGF1R*) has tyrosine kinase activity, mediates the action of insulin-like growth factor, and regulates cartilage mineralization (Heilig et al., 2016).

Nitric oxide synthase 3 (*NOS3*) encodes the vascular endothelial isoform of nitric oxide synthase (eNOS). *NOS3* is associated with sporadic limb defects in mice (Gregg et al., 1998) and has been implicated in bone remodeling in rats (Hukkanen et al., 1999). LIM homeobox transcription factor 1 beta (*LMX1B*) is a transcription factor. Mutations in *LMX1B* cause a rare autosomal dominant disorder characterized by dystrophic nails, hypoplastic or absent patellae, and dysplasia of the elbows and iliac horn (Marini et al., 2010).

A**B**

Patched 1 (*PTCH1*) codes for a receptor for Hh ligands and regulates the activity of smoothened, frizzled class receptor (SMO, another effector gene associated with a known lead SNV). When bound, *PTCH1* relinquishes its inhibitory effect on SMO and activates the Hh signaling cascade, which plays an important role in controlling the proliferation of chondrocytes and also in stimulating osteogenesis during endochondral bone formation and longitudinal growth (Alman, 2015).

Several further newly identified high-confidence effector genes have a neuronal connection (Figure 3A). Augurin, the protein encoded by *C2orf40* (also called *ECRG4*), is involved in CNS development in animal models (Gonzalez et al., 2011) and shows association with neuropathologic features of Alzheimer's disease and related dementias in humans (Beecham et al., 2014). SNVs in the vicinity of *TSEN15* have been robustly associated with anthropometric traits that have epidemiological links to

Figure 3. High-confidence osteoarthritis effector genes

(A) Overview of the 77 high-confidence osteoarthritis effector genes and their broad biological classifications, as depicted in Tables 3 and S12. The lead SNV for each is given in brackets.

(B) Schematic representation of a chondrocyte and its extracellular matrix, highlighting exemplary osteoarthritis-implicated biological pathways (TGF- β signaling, FGFR3 signaling, and part of the fibrosis pathway) and the high-confidence effector genes (in red boxes), both established and newly identified (in red boxes with a black outline) that have been found to play a role.

osteoarthritis, such as height (Gudbjartsson et al., 2008), body fat distribution (Rask-Andersen et al., 2019), and waist circumference adjusted for BMI (Hübel et al., 2019). *CUX1* is a transcription factor involved in brain neuronal differentiation and synaptogenesis (Cubelos et al., 2010). *Cux1* expression was observed at chondrogenic interzones during limb development, suggesting also a regulatory role in joint formation (Lizarraga et al., 2002).

The TRIO and f-actin binding protein (*TRIOBP*) gene encodes multiple protein isoforms via 2 promoters (Park et al., 2018). *TRIOBP-1* is ubiquitously expressed and interacts with TRIO and f-actin binding protein that together play crucial roles in neuronal morphogenesis (Woo et al., 2019) and controlling actin cytoskeleton organization, cell motility, and cell growth (Zaharia et al., 2020).

Myotubularin related protein 2 (*MTMR2*) has an important role in membrane targeting, vesicular trafficking, and regulation of signal transduction pathways. Mutations in *MTMR2* cause Charcot-Marie-Tooth disease type 4B, which features a generalized loss of large myelinated nerve fibers and focally folded myelin sheaths giving rise to inadequate nerve signaling to muscles, resulting in muscle weakness and atrophy (Volpatti et al., 2019). The ubiquitously expressed protein encoded by CREB-binding protein (*CREBBP*) plays a critical role during development in particular with brain size regulation, correct neural cell differentiation, and neural precursor cell migration, as demonstrated in mouse models (Schoof et al., 2019).

Cholinergic receptor muscarinic 2 (*CHRM2*) is involved in the mediation of cellular responses. Analysis of rat tissues revealed expression in whole brain (Peralta et al., 1987) and in human neuroblastoma cells (Zhou et al., 2001). Variation in *CHRM2* predisposes to various neuropsychiatric diseases (Cannon et al., 2011; Rajji et al., 2012), and Alzheimer's disease (Mash et al., 1985). The protein encoded by synapsosome associated protein 47

(SNAP47) is a soluble N-ethylmaleimide-sensitive fusion protein attachment protein receptor (SNARE) protein involved in trafficking and membrane fusion. SNARE-mediated fusion is an essential mechanism that drives the synaptic transmission, neuron development, and growth. SNAP47 plays a role in exocytic mode and neuronal morphogenesis (Holt et al., 2006; Urbina et al., 2021).

Several of the effector genes have an immune or inflammatory role. For example, the protein encoded by toll like receptor 4 (*TLR4*) plays a fundamental role in pathogen recognition and activation of the innate immune response (Tatematsu et al., 2016). *TLR4* is also activated by host-derived molecules generated by damaged tissues related to different musculoskeletal pathologies (Abdollahi-Roodsaz et al., 2007; Goldring and Goldring, 2007). This, along with gene expression in chondrocytes (Wang et al., 2011), osteoblasts (Kikuchi et al., 2001), and synoviocytes (Midwood et al., 2009), has linked *TLR4* to diseases like rheumatoid arthritis (Abdollahi-Roodsaz et al., 2007), osteoarthritis (Gómez et al., 2015), and osteoporosis (Vijayan et al., 2014), where modulation or inhibition of *TLR4* has been suggested as a treatment. Activation of T cells can lead to osteoclastogenesis and bone resorption by influencing the expression of tumor necrosis factor ligand superfamily member 11 (*TNFSF11*) (Kong et al., 1999). *TNFSF11* encodes receptor activator of nuclear factor kappa-β ligand (also known as RANKL), a cytokine that has been linked to inflammatory bone remodeling in rheumatoid arthritis, with increased *TNFSF11* levels associated with worsening arthritis severity (Papadaki et al., 2019; Remuzgo-Martínez et al., 2016) and a well-established role in osteoclastogenesis (Kohli and Kohli, 2011).

Nuclear receptor subfamily 3 group C member 1 (*NR3C1*) encodes the glucocorticoid receptor (GR) which circulates in the cytoplasm and is involved in the inflammatory response (Escoter-Torres et al., 2019). In osteoarthritis, endogenous glucocorticoid signaling in osteoblasts and chondrocytes is detrimental (Macfarlane et al., 2020).

Phosphofructokinase (*PFKM*) has a role in muscle function. It encodes a muscle isozyme that catalyzes the phosphorylation of fructose-6-phosphate during glycolysis. Mutations in this gene result in Tarui's disease (glycogen storage disease type 7) that is an autosomal recessive metabolic disorder characterized clinically by exercise intolerance, muscle cramping, exertional myopathy, and compensated hemolysis (Raben and Sherman, 1995).

Drug target identification

We examined the druggability status of all 637 genes with at least one piece of supporting evidence from fine-mapping and functional analyses (Table S10; STAR Methods). Of these 637 genes, 205 were present in the druggable genome database (Finan et al., 2017), showing a 1.46-fold enrichment of genes with supporting evidence in the database (binomial test $p = 2.21 \times 10^{-9}$) (STAR Methods). From these osteoarthritis druggable target genes, 71 genes reside in tier 1, which incorporates the targets of approved (licensed) drugs and drugs in clinical development (Table S10; STAR Methods). Of the 77 genes with three different lines of evidence supporting causality, 20 are tier 1 candidates (18 of these are present in DrugBank) (Table 4; STAR Methods),

of which 7 correspond to new genetic signals discovered in this study (*CHST3*, *VDR*, *TNFSF11*, *IGF1R*, *NR3C1*, *CHRM2*, and *NOS3*).

Within tier 1, ten candidates have previously been studied in clinical trials of efficacy or in cohort studies of osteoarthritis (six arising from new signals: *PPARD*, *NR3C1*, *VDR*, *MAPK14*, *IGF1R*, and *CHST3*). The *PPARD* antagonist sulindac has marketing authorization as a non-steroidal anti-inflammatory drug (NSAID) in osteoarthritis for its prostaglandin synthase activity. The *SLC1A1* agonist and neuropathic pain inhibitor pregabalin is commonly prescribed in osteoarthritis. Pregabalin has some supportive clinical trial data for its co-prescription with the NSAID meloxicam in the short-term treatment of pain in knee osteoarthritis (Ohtori et al., 2013). *NR3C1* encodes the glucocorticoid receptor, the activation of which has broad anti-inflammatory and immunomodulatory actions with marketing authorization for several agonist molecules. One of these, prednisolone, has long been used as a disease modifying agent in inflammatory arthritis and in the recent Heart Outcomes Prevention Evaluation (HOPE) study was found to be effective in reducing pain and synovitis in hand osteoarthritis (Kroon et al., 2019). Cathepsin K (encoded by the *CTSK* gene) is an enzyme that plays a critical role in collagen degradation within osteoclasts, and MIV-711 is a selective cathepsin K inhibitor that has recently been shown in a phase 2 clinical trial to be effective in reducing structural damage in patients with knee osteoarthritis (Conaghan et al., 2020). *VDR* encodes the vitamin D receptor, the activation of which is a major regulator of calcium metabolism. The results of clinical trials of vitamin D supplementation on symptoms and structural damage in knee osteoarthritis have been mixed (Arden et al., 2016; Jin et al., 2016; McAlindon et al., 2013; Sanghi et al., 2013; Zheng et al., 2017) but may suggest a small benefit in patients with vitamin D deficiency. *EGLN2* encodes Egl nine homolog 2, a prolyl hydroxylase that mediates hydroxylation of proline and thus contributes to collagen and proteoglycan synthesis. Supplementation of its agonist, ascorbic acid (vitamin C), has been associated with joint health in observational cohorts, although with mixed effects (Joseph et al., 2020; McAlindon et al., 1996; Peregoy and Wilder, 2011). Deficiency of the *HCAR2* agonist niacin (vitamin B3) was associated with knee osteoarthritis progression in the Japanese ROAD cohort (Muraki et al., 2014). The *MAPK14* antagonist PH-797804 has been studied in a phase 2 clinical trial to examine the pain relief of PH-797804 alone or with naproxen in subjects with osteoarthritis of the knee (NCT01102660), although we are not aware of any trial results reporting in PubMed or on ClinicalTrials.gov. Finally, the carbohydrate sulfotransferase 3 agonist thalidomide has been shown to attenuate early osteoarthritis development in a mouse medial meniscus destabilization model through a mechanism involving the downregulation of vascular endothelial growth factor (VEGF) expression (Seegmiller et al., 2019).

All of the 45 further tier 1 druggable targets have market authorization or are in clinical development for other indications (Table 4). Ten of these are high-confidence effector genes and 16 arise from new genetic signals. The functional and epidemiological evidence of their roles in clinical osteoarthritis presented here provides support for early repurposing investigation. One antibody small molecule, fostamatinib, appears multiple times

Gene	Encoded protein	Uniprot ID	Drug name	Drugbank ID	Molecule type	Development phase	Molecular mechanism of action	Mechanism of action	Current clinical indication(s)
VDR ^{b,c}	vitamin D receptor	P11473	Calcitriol ^a	DB00136	small molecule	approved	agonist	active metabolite of vitamin D	vitamin D deficiency, chronic kidney disease, hyperparathyroidism (secondary), investigation in osteoarthritis
<i>Genes are identified according to the Ensembl GeneName for the gene. Both agonists and antagonists of the target protein are shown. DrugBank information on the tier 1 likely effector genes.</i>									

^aIndicates that multiple drugs with similar mechanisms of action are identified for a given target. Here, an example drug from the class is shown to represent an identified mechanism of action on the target-encoded protein.

^bDenotes associated with newly reported signal.

^cDenotes effector genes with at least 3 lines of evidence.

in Table 4 as a tyrosine kinase inhibitor that targets AAK1, EPHA5, GAK, GSK3A, MAP2K1, MAP2K6, PAK1, and PRKCD, and has marketing approval as a biologic disease-modifying anti-rheumatic drug (DMARD). The JAK2 antibody baricitinib and the TYK2 antibody tofacitinib are both marketed as biologic DMARDs, and the MAP2K1 antibody binimetinib is currently in phase 3 clinical trials as a biologic DMARD. Each of these drugs therefore present a clinical opportunity and putative mechanism for repurposing studies in osteoarthritis. Of the remaining tier 1 and tier 2 druggable targets, the potentially actionable molecules are at an earlier stage of development and present a more distant repurposing opportunity.

DISCUSSION

Our findings have generated further knowledge on the differences between weight bearing and non-weight bearing joints and point to mechanisms that are common to disease development at any joint, and joint-type-specific. Indeed, bone and cartilage development pathways were enriched in signals traversing weight bearing and non-weight bearing joints, identifying joint development as a common mechanism for any form of osteoarthritis (Table S13).

We have been able to establish molecular links between the disease and its main symptom, pain. We demonstrate genetic correlation between osteoarthritis and pain-related phenotypes and identify signal enrichment in neurological pathways (Table S13). Furthermore, several of the high-confidence effector genes have a role in neuropathology. The majority of osteoarthritis cases in this study were defined as total joint replacement and/or self-reported osteoarthritis, and both of these disease phenotypes are highly driven by pain. Identification of these genes can also have implications for further joint pain-related disorders, for which insights have been limited to date.

A large number of the high-confidence effector genes converge on the endochondral pathway, playing an essential role in homeostasis of the chondrocyte (Figure 3) and osteophytosis. Several of the identified genes are important in TGF- β signaling and function. The newly identified fibrillin 2 (*FBN2*) signal, together with *LTPB1* and *LTPB3*, regulate the availability of active *TGFB1*. *TGFB1* is the major form of *TGFB* in cartilage and can activate a cascade of downstream genes through SMAD3-signaling, including ECM-genes which have been identified in our current study, such as carbohydrate sulfotransferase 3 (*CHST3*) (Zhou et al., 2020).

Our data provide evidence for the FGF-signaling cascade (*FGFR3*, *FGF18*, and *PIK3R1*) being causally involved in osteoarthritis (Figure 3). *FGF18* is currently being tested in clinical trials for its effectiveness in osteoarthritis (Hochberg et al., 2019). The newly identified molecular player phosphoinositide-3-kinase regulatory subunit 1 (*PIK3R1*) encodes the p85a, p55a, and p50a regulatory subunits of class IA phosphatidylinositol 3 kinases (PI3Ks), which are known to play a key role in the metabolic actions of insulin and are required for adipogenesis (Kim et al., 2014; Thauvin-Robinet et al., 2013). Mutations in *PIK3R1* cause agammaglobulinemia 7 (Conley et al., 2012), immunodeficiency 36 (Deau et al., 2014; Lucas et al., 2014), and SHORT syndrome (Dymont et al., 2013), which is characterized by short

stature, hyperextensibility of joints, ocular depression, Rieger anomaly, and teething delay (Dyment et al., 2013; Thauvin-Robinet et al., 2013). The balance between chondrocyte proliferation, differentiation, and hypertrophic conversion is controlled by crosstalk between several signaling pathways, of which we find causal evidence here: PTHLH and IHH-signaling (*SMO* and *PTCH1*) antagonize signaling through FGFR3. In addition, we identify two independent genetic variants implicating noggin (*NOG*) as an osteoarthritis effector gene. Noggin binds to TGFB, BMPs, and GDF5 and thereby prevents binding to the cognate receptor. Mutations in *NOG* cause a whole range of bone and cartilage phenotype depending on the severity of the mutation (Lehmann et al., 2007).

Several of the putative osteoarthritis causal genes are involved in developmental pathways (Figure 3). Skeletal development can be linked to osteoarthritis in several ways. First, skeletal developmental genes are involved in joint (tissue) characteristics before onset of disease such as cartilage thickness (*TGFA*, *FGFR3*, *RUNX2*, and *PIK3R1*) (Castaño-Betancourt et al., 2016) or joint shape (resulting in different loading of the joint). Second, the skeletal development pathway could be involved in the reaction to damage in the joint. Depending on the specific genetic makeup, each individual reaction to a damaging trigger to the joint could be different, thereby determining the risk of developing osteoarthritis upon trauma or mechanical overload. Pathway analysis performed on the current study signals further corroborated this, because it revealed evidence for enrichment of pathways typically involved in reaction to damage.

Our data also suggest that subtle changes in pivotal osteochondrogenic pathways lead to an adverse response to joint damage and/or overload. This may catalyze a fibrotic response both in cartilage and in the synovium. We identified tenascin C (*TNC*) as one of the high-confidence effector genes (Fu et al., 2017; Imanaka-Yoshida et al., 2020) (Figure 3). *TNC* is a component of the extracellular matrix and is involved with organ fibrosis, inflammation, and cardiovascular disease (Golledge et al., 2011; Yasuda et al., 2018). The formation of fibrocartilage and fibrosis in the joint is a major contributor to the degenerative changes in osteoarthritis (Rim and Ju, 2020). Further, elevated levels of TGF- β signaling are associated with the pathological and fibrosis changes (van der Kraan, 2017). TGF- β is also a potent inducer of epithelial-mesenchymal transition (EMT) (Nieto et al., 2016; Stone et al., 2016). EMT, a process whereby fully differentiated epithelial cells undergo transition to a mesenchymal phenotype giving rise to fibroblasts, is a driver of early fibrosis, which is a typical response to injury or pathological changes and inflammation, all common endpoint outcomes in osteoarthritis. The severity of fibrosis contributes to the degree of degenerative changes that lead to pain in osteoarthritis. We have identified significant association of variants in many of the genes involved in the induction (e.g., EMT genes, *CUX1* and multiple molecular components of the TGF- β pathway) and progression of fibrosis (ECM genes e.g., *TNC*, TGF- β signaling *FBN2*, *LTBP1*, *LTBP3*, *TGFB1*, and *SMAD3*) (Figure 3B). These findings indicate that combined variation in the regulation of these genes may collectively contribute to the susceptibility and severity of degenerative changes in osteoarthritis.

Seventy-one of the implicated genes code for molecules that are the targets of approved (licensed) drugs and drugs in clinical development. Our findings substantially strengthen the evidence for these potential therapeutics, provide drug repositioning opportunities, and offer a solid basis on which to develop, or repurpose, such interventions for osteoarthritis.

Our work provides a robust springboard for follow-up functional and clinical studies. We have demonstrated clear differences between distinct osteoarthritis patient populations, for example based on disease severity, joint site affected, and sex. We enhance our understanding of the genetic etiology of disease, shed biological insights, and provide a stepping stone for translating genetic associations into osteoarthritis drug development, ultimately helping to catalyze an improvement in the lives of patients suffering from osteoarthritis.

Limitations of the study

Enhancing population diversity in genetic association studies is important for discovering risk variants, pinpointing likely causal alleles, improving risk prediction, and ensuring the transferability of findings across global populations. In this work, less than 3% of contributing subjects were of non-European ancestry. Going forward, the identification and inclusion of diverse populations in osteoarthritis genetic association studies is urgently needed.

Disentangling mechanisms that are active at the point of disease initiation versus those activated during the natural history of disease warrant animal model studies in which disease dynamics can be studied in depth. Indeed, investment in mechanistic studies of the newly identified high-value targets will be important next steps. Clinical trials of intervention will be needed to take our findings forward into mechanism and clinical outcome, therefore elucidating how to target the implicated genes and proteins, how downstream events will be affected, and, ultimately, how these interventions will affect disease outcome in the patient.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
 - Lead contact
 - Materials availability
 - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
 - Study cohorts
 - Informed consent and study approval
- METHOD DETAILS
 - Cohorts and phenotype definition
 - Annotation of protein coding variants
 - Mouse and human phenotypes
 - Additional phenotypes and endophenotypes
 - Cartilage-type specific effect
 - Effect on intervertebral disc degeneration
 - Monogenic and rare human diseases
- QUANTIFICATION AND STATISTICAL ANALYSIS

附录 B 外文文献译文

题目：破译来自 9 个群体的 826690 个个体中的骨关节炎遗传密码

作者：Cindy G. Boer et al.

摘要：骨关节炎影响着全世界 3 亿多人。在此，我们对 826,690 名个体（177,517 名骨关节炎患者）进行了全基因组关联研究荟萃分析，在 11 种骨关节炎表型中确定了 100 个独立相关的风险变异，其中 52 个变异先前未曾报道与骨关节炎相关。我们报告了拇指和脊柱骨关节炎的风险变异，并确定了承重和非承重关节之间的遗传效应差异。我们确定了性别特异和早期发病的骨关节炎风险位点。我们整合了来自患者原始组织（包括关节软骨、软骨下骨和骨软骨）的功能基因组学数据，并确定了高置信度的效应基因。我们提供了与疼痛（主要疾病症状）相关的表型的遗传相关性证据，并确定了与神经元过程相关的可能的致病基因。我们的结果提供了对疾病过程中的关键分子角色的诠释，并强调了有吸引力的药物靶点。我们的结果提供了对疾病过程中关键分子角色的见解，并突出了有吸引力的药物目标，以加快骨关节炎药物的转化。

绪论

骨关节炎影响全世界超过 3 亿人。在本研究中，我们对 826,690 名个体（177,517 名患有骨关节炎）进行了全基因组关联研究 Meta 分析，并确定了 11 种骨关节炎表型中的 100 个独立的相关风险变异，其中 52 个以前从未被发现与该疾病相关。我们报告了拇指和脊柱骨关节炎的风险变异，并确定了负重和非负重关节之间遗传效应的差异。我们同时确定了性别特异性和早发性骨关节炎风险位点。我们整合了来自患者原发组织的功能基因组学数据（包括关节软骨，软骨下骨和骨赘软骨）并鉴定高置信度的效应基因。我们提供了与疼痛相关表型（主要疾病症状）的遗传相关性证据，并确定了与神经营养过程相关的可能致病基因。我们的结果提供了对疾病过程中关键分子通路的分析，并强调了可能的药物靶点。

背景

骨关节炎是全球范围内导致残疾和疼痛的主要原因之一，有超过 3 亿人受到影响，但目前尚无治愈方法。对骨关节炎的治疗侧重于通过缓解疼痛和通过关节成形术来缓解症状。因此，我们迫切需要详细了解疾病的病因和新的药物靶点。

骨关节炎是关节的复杂退行性疾病，其特征是软骨退化、软骨下骨增厚、骨赘形成、滑膜炎症以及关节囊、韧带和相关肌肉的结构改变。近年来，全基因组关联分析 (GWAS) 在阐明骨关节炎的遗传背景方面取得了进展。迄今为止研究者们报告了 96 个统计上独立的风险变体。但这些变异只解释了一小部分表型变异且主要与影响膝关节和髋关节的骨关节炎有关。

骨关节炎会影响每个滑膜关节。研究发现体重指数 (BMI) 的增加与疾病风险相关。因此我们需要更好地了解负重关节（膝关节、髋关节和脊柱）和非负重关节（手、手指和拇指）之间的遗传差异，以帮助解开导致疾病发展的代谢和生物机能影响。在本文中，我们对 826,690 名欧洲和东亚血统的个体的膝关节、髋关节、手指、拇指和脊柱骨关节炎表型进行了 GWAS Meta 分析。我们整合了来自疾病相关组织的功能基因组学分析，包括基因表达、蛋白质丰度和全基因组甲基化、小鼠敲除模型和单基因人类疾病表型数据，以及互补的 fine-mapping、共定位和因果推理方法，以识别可能的效应基因，并通过增强我们对疾病病因学的理解促进急需的治疗转化。

结果

基因结构

骨关节炎 SNV 的鉴定 我们对来自 9 个人群的 13 个国际队列进行了骨关节炎的 GWAS 萍萃分析，涉及多达 826,690 人（177,517 名骨关节炎患者）。与迄今为止最大的骨关节炎 GWAS 相比，骨关节炎患者人数大幅增加（2.3 倍）。其中两个队列是东亚人，其中 11 个队列是欧洲血统。我们定义了 11 种表型，涵盖骨关节炎的所有主要部位（图 1；表 S1；STAR 方法）。我们使用 $p < 1.3 \times 10^{-8}$ 的阈值发现了 11,897 个全基因组显着相关的单核苷酸变异 (SNV)，以说明独立测试的有效数量。我们在表型中应用条件分析并确定了 223 个独立关联，其中一些在表型之间重叠。84 种变体以前与骨关节炎无关。我们调查了先前报道的骨关节炎位点，发现 96 个中的 87 个在同一方向上以名义显着性复制。

我们使用条件分析来确定在疾病表型定义中不重叠的关联。我们确定了 100 个独特且独立的变异关联，其中 60 个与不止一种骨关节炎表型相关。这些变异位点中 52 个位点在先前报道中未发现与骨关节炎的关联。我们定义首要 SNV 为具有最强关联统计证据的易感 SNV 位点。我们发现其中的六个首要 SNV 都在编码区（都是错义突变），59 个 SNV 位于基因转录本中，35 个突变发现于基因间。

本研究中，我们报告了脊柱 ($n = 1$) 和拇指 ($n = 2$) 骨关节炎的 SNV 位点，并增加了之前未被广泛研究就的诸如手（5 个新的，3 个先前报告的）和手指（3 个新的，2 个先前报告的）骨关节炎的风险 SNV 的数量。在 100 个 SNV 中，90 个变异基因频率较高 ($MAF \geq 5\%$)，4 个变异属于低频变异 ($5\% > MAF \geq 0.5\%$)。本研究还检测到 6 个效应量较大 ($OR 3.03–9.52$) 的罕见的变异关联 ($MAF 0.03\%–0.11\%$)。除去一个已被研究发现的变异之外，其他 5 个变异关联是由本研究发现。而这五个新变异主要来自冰岛的一个大家系之中。

我们还对非欧洲人种 0.9%–2.8% 的病例是东亚人种）的个体进行了关于 4 种骨关节炎表型（包括脊柱、膝关节、膝关节和/或髋关节，以及任何部位的骨关节炎）的分析。尽管来自东亚人种的样本量很小，但我们观察到 62% 的变异在东亚人群基因型信息的分析中也具有一定意义，并且这些信号中的 20% 也具有显著的统计学意义 ($p = 2.27 \times$

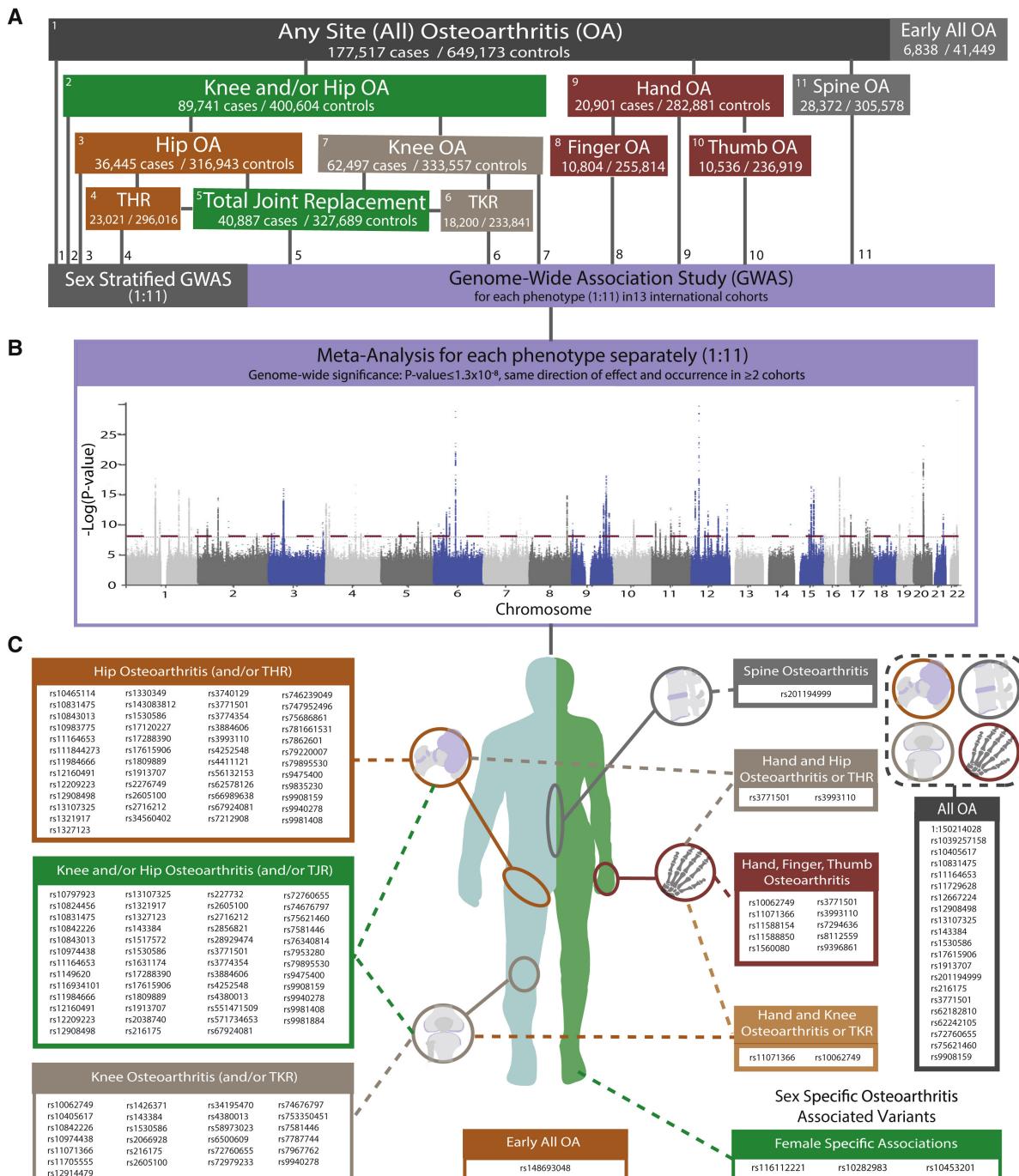


图 B-1 基因结构骨关节炎的基因结构图示 A. 11 种定义的骨关节炎表型的概述，特定性别分析，它们之间的关系以及它们的样本量（病例/对照）。TKR，全膝关节置换；THR，全髋关节置换。B. 所有 11 个被检查的骨关节炎表型的所有单独的元分析结果的合并曼哈顿图。虚线代表全基因组意义阈值 $P = 1.3 \times 10^{-8}$ 。C. 所有领先的全基因组意义的独立骨关节炎相关单核苷酸变异（SNVs）和与之相关的骨关节炎表型的图形概述。

10 -5)。

我们还测试了通过该变异信息构建的多基因风险评分 (PRS) 模型的预测能力，我们发现处于 PRS 分布的较高十分位数范围内的个体罹患骨关节炎的可能有显著升高。

女性特异性骨关节炎风险变异 为了研究仅针对男性、仅针对女性或在男性和女性中具有差异影响的骨关节炎相关变异，我们进行了性别相关关联测试以及等位基因效应的异质性测试。通过该测试我们发现了 3 个新的女性特异性 SNV，其中两个显示出性别间效应大小的显著差异 ($\text{Phet-diff} < 0.016$)。其中突变 rs116112221 在女性的全髋关节置换表型中具有显著性，并且位于包含长基因间非编码 RNA 的区域。同该区域最接近的蛋白质编码基因是 *FANCL*。*FANCL* 突变可能导致人类原发性卵巢功能不全，这种疾病同时会导致更年期提前，而虽然还没有有力研究论证，更年期的提前也被认为与骨关节炎患病率增加有关。临床研究表明，选择性雌激素受体调节剂 (SERMs) 治疗对该类型骨关节炎的预后有着积极作用，特别是对于绝经后早期或骨质疏松性骨关节炎患者。

我们进一步确定了与全髋关节置换相关的在男性和女性之间的影响具有明显差别的突变 rs10282983。rs10282983 位于 *C8orf34* 的内含子中，该基因被发现与腰臀比和足跟骨矿物质密度等性状有关，而两者都是可能导致骨关节炎的因素。另一突变 rs10453201 位于基因 *UBAP2* 的 5' 末端，该突变与任何部位的女性骨关节炎显著相关。而基因 *UBAP2* 与帕金森病、2 型糖尿病、BMI 和人体足跟骨矿物质密度都存在相关。

早发性骨关节炎 全基因组荟萃分析也确定了具有大效应量和低等位基因频率的早期骨关节炎的新风险变异 rs148693048。该变异在所有荟萃研究的来源中都被报道同骨关节炎形状存在一定的相关，但是之前并没有将其与骨关节炎相关联的研究。与退化的软骨相比，附近的两个蛋白质编码基因 (*NEFM* 和 *DOCK5*) 在完整的软骨中表现出显著不同的表达水平。*NEFM* (神经丝介质) 与神经元结构的延展有关，并且由其指导表达的蛋白质通常用作神经元损伤的生物标志物。*DOCK5* (胞质分裂作动蛋白 5) 的鸟嘌呤核苷酸交换活性已被确定为破骨细胞功能的调节剂，在骨吸收中发挥重要作用，对其活性的药理抑制可防止骨质溶解，同时保持人类和小鼠的骨形成速度。*DOCK5* 的其他内含子变异也显示出与其他骨表型的关联 ($p < 5.0 \times 10^{-8}$)，例如足跟骨矿物质密度和青少年特发性脊柱侧凸。

交叉表型分析

不同表型信号的异同 我们观察到一些变异表现出关节特异性效应。我们发现其中的 60 个 SNV 在全基因组范围内与一种以上的骨关节炎表型显著相关。其中 40 个 SNV 仅与负重关节骨关节炎存在着全基因组上的显著关联，4 个 SNV 仅与非负重关节骨关节炎存在显著关联。我们有超过 80% 的把握提取负重关节分析中的所有 4 种非负重特异性变异 (全基因组显著性)。此外，在非负重关节 (手骨关节炎) 分析中，我们有超过 80% 的把握来检测 40 个负重关节特异性效应中的 22 个。尽管已知有几种核心途径支持骨关节炎病理学，但无论受影响的关节部位如何，除了 *GDF5* 基因座外，之前没有发现常见的遗传性骨关节炎 SNV。而在本研究中我们已经确定了 42 个在承重和非承重关节中都有很强的关联性的 SNV。其中几个 SNV，包括 rs3771501 (*TGFA*)、rs3993110

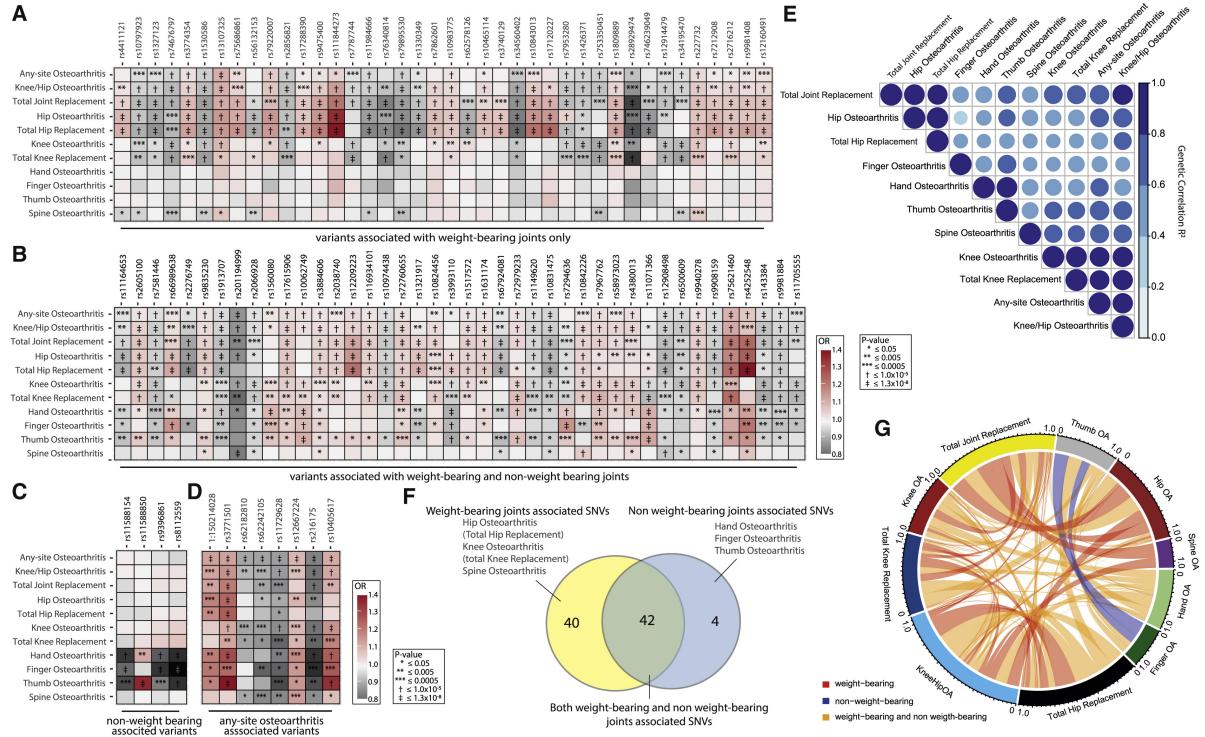


图 B-2 不同表型的信号的相似性和差异性骨关节炎遗传学之间的相关性和重叠性 A-D. 骨关节炎相关单核苷酸变异 SNV 的热图。每个骨关节炎表型 GWAS 结果的效果大小 (OR, 几率) 和 P 值都显示在每个主导 SNV 上。OR 以颜色表示, P 值以方框内的符号表示。A. 只有负重关节 (髋关节、膝关节和脊柱)。B. 承重和非承重关节 (髋关节、膝关节、脊柱、手、手指和拇指)。C. 非负重关节 (手、手指和拇指)。D. 任何部位的骨关节炎 SNVs。E. 受检的骨关节炎表型之间的遗传相关性 (R2) 的热图。F. 维恩图描述了与负重和非负重关节相关的 SNVs 的数量和重叠。(G) Circos 图描述了 100 个主导变异的骨关节炎关联的重叠情况。

(*TEADI/DKK3*)、rs72979233 (*CHRDL2*) 和 rs7967762 (*PFKM / WNT10B*) 与多个骨关节炎关节部位相关。这些变异可能指示了骨关节炎病理学中常见的潜在机制。它们已被证明在转化生长因子 β (TGF- β)/骨形态发生蛋白 (BMP)、Wnt/ β -catenin 信号通路中发挥作用，并且其功能相互作用可能与骨关节炎的发病机制有关。这些信号通路可能是药物开发的主要候选者。

从骨关节炎表型之间的关联信号的比较中也可以收集到更多的信息。大多数与膝关节、髋关节和膝关节和/或髋关节骨关节炎相关的 SNV 对各自的关节置换表型具有较大的影响，但是所有这些表型的样本量都较小。这可能是由表型定义的同质性驱动的，或者可以代表生物和功能的相关性，表明这些位点可能在接受关节置换 (即疼痛和炎症) 中发挥了比骨关节炎病理本身更重要的作用。例如，rs76340814 (*PTCH1*) 和 rs28929474 (*SERPINA1* 的错义变体) 与全髋关节置换 (THR)、全膝关节置换 (TKR) 和全关节置换 (TJR)，比与髋关节或膝关节骨性关节炎有更强的关联和更大的效应值。事实上，*PTCH1* 被认为在神经系统和大脑发育中发挥作用，而 *SERPINA1* 被认为在炎症中发挥作用。对大鼠骨关节炎模型的研究表明，用 *SERPINA1* 编码的 α -1-反蛋白酶进

行早期治疗，可以阻断中性粒细胞弹性蛋白酶的蛋白溶解活性，并使关节炎症、疼痛和隐性神经损伤得到持久改善。

表型之间的遗传联系 尽管骨关节炎的遗传特性范围较广，我们还是发现骨关节炎亚型共享大量遗传成分。

我们调查了骨关节炎的遗传成分是否与其他性状共享，发现与人体测量特征（BMI、肥胖、体重和脂肪量）、2型糖尿病、教育、抑郁症状、吸烟行为、骨矿物质密度、生殖表型和智力有明显的关系，如以前的报道，以及几种疼痛表型。

疼痛是骨关节炎患者经历的最多的致残症状，也是患者寻求医疗支持甚至全关节置换的主要原因之一。骨关节炎导致疼痛的病因是多因素的，包括明显的软组织炎症、涉及关节痛觉感受器的疼痛通路的敏感化、中枢神经系统的痛觉处理以及骨关节炎模型中的神经病理性疼痛成分。虽然疼痛是骨关节炎的主要症状，但之前的研究中没有发现骨关节炎与疼痛发生的遗传决定因素。本研究发现骨关节炎与坐骨神经痛、纤维肌痛、头痛和其他背痛表型之间有很高的相关性，其中与脊柱骨关节炎的相关性最高（遗传相关性 $[rg] = 0.61, 0.87, 0.39$ 和 0.79 ）。*SOX5* 是本文所新发现的疼痛信号之一，之前有报道称其在人类骨关节炎软骨中的表达被上调，并与背痛和腰椎间盘退化有关。这些发现得到了动物模型数据的支持，对小鼠的研究表明 *SOX5* 的失活导致小鼠包括软骨、脊索或椎间盘等骨骼发育的缺陷。我们还在 LD-Hub 数据库中观察到骨关节炎与背痛、走路时腿痛、膝痛、髋痛、背痛和颈/肩痛等一系列疼痛表型之间有很强的相关性（均来自 UKB）。因此，我们的数据表明，一部分已识别的骨关节炎相关信号也与骨关节炎疼痛有关。

效应基因和生物学途径

鉴定推定的因果变异 我们采用了互补计算方法，将 GWAS 信号细化到一小部分可能的因果突变上，根据信号的富集程度确定相关的组织，并根据表达定量性状位点（eQTL）的共定位和因果推断分析提供骨关节炎致病机理上的解释。12 个信号被精细地映射到完全包含在单个基因的转录本内的突变组，其后验概率大于 95%，但我们注意到这并没有为该基因为导致骨关节炎的效应基因这一推论提供结论性的证据。值得注意的是，目前是其他适应症使用的批准药物的靶点 *ALDH1A2* 以 99% 的后验概率精细映射到 6 个内含变异，这也为药物重利用提供了潜在的机会。对于其中 6 个 SNV（3 个新的和 3 个已知的），单个突变可以被假定为具有 >95% 后验概率的因果关系。

收集证据以识别效应基因 我们评估了在从接受关节置换手术的骨关节炎患者中提取的软骨细胞中，是否有位于骨关节炎相关变异 1Mb 距离内的基因在原发性骨关节炎影响的组织中表现出不同的基因表达和蛋白丰度。同样，我们比较了完整的和退化的软骨组织下的软骨下组织的基因表达。通过结合互补的功能基因组学和计算方法的结果，我们确定了 637 个至少有一条证据指向一个推定的效应基因的基于基因。对于这 637

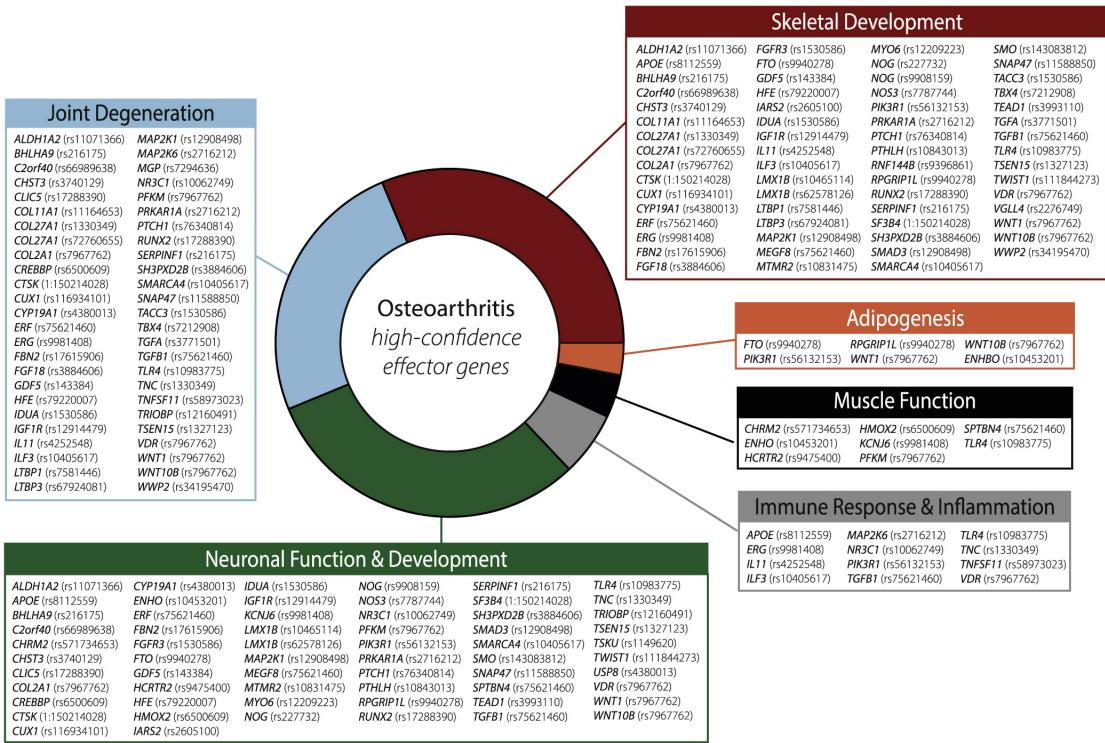
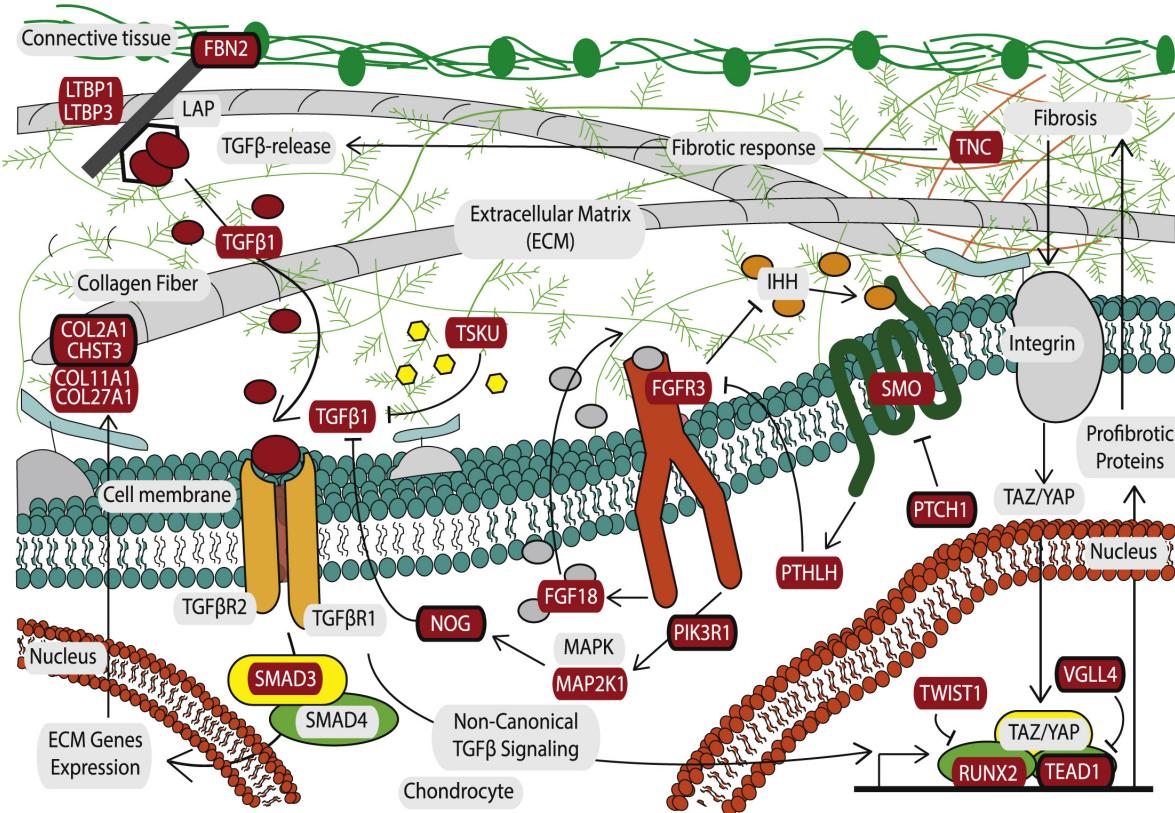
A**B**

图 B-3 高置信度骨关节炎效应基因 A.77 个高置信度骨关节炎效应基因及其广泛的生物分类概述，如表 3 和 S12 所描述。每个基因的主要 SNV 在括号中给出。B. 软骨细胞及其细胞外基质的示意图，突出了示范性的骨关节炎相关生物途径 (TGF- β 信号传导, FGFR3 信号传导，以及部分纤维化途径) 和高置信度的效应基因 (红框内)，包括已确定的和新确定的 (红框内有黑色轮廓)，已发现其发挥的作用。

个基因，我们结合了来自精细图谱、eQTL 共定位分析、动物模型数据、人类肌肉骨骼和神经元表型数据、功能基因组学和因果推理分析等支持性信息，确定了 77 个至少有 3 条不同证据支持其作为效应基因的基因。在这 77 个基因中，有 4 个是由错义线索变异支持的（*VGLL4* 中的 rs2276749, *CHST3* 中的 rs3740129, *SMO* 中的 rs143083812, 和 *ILII* 的 rs4252548）；48 个为以前报道的骨关节炎相关 SNVs 的可能效应基因提供了强有力的额外证据；30 个位于新的相关信号中。

在这些基因中，*CHST3*、*SMAD3* 和 *GDF5* 具有较高效应基因可信度，各有 6 条不同的证据支持它们同骨关节炎发展过程相关。*CHST3*（碳水化合物硫酸盐转移酶 3）是一个新发现的信号，它编码软骨中的主要蛋白多糖，即硫酸软骨素。*CHST3* 的突变在之前的研究中被发现与身材矮小、先天性关节脱位、足癣、Larsen 综合征和肘关节发育不良等性状有关。*CHST3* 同时也被证明与腰椎间盘退化有关。

为了进一步了解高置信度效应基因在疾病过程中的生物学作用，我们整合了基于内涵型分析的与基础生物学更密切相关单基因和罕见人类疾病数据、全表型分析和额外的功能基因组学数据等额外信息。通过综合所有的证据，我们发现 77 个高置信度效应基因中的几个基因被分配到穿越了多个生物学过程的可能机制中，并通过这些机制发挥其作用。我们主要关注本工作中新报道的相关基因。这些基因也是进一步机制探索和临床研究的高价值候选基因。

大多数高置信度的效应基因与骨骼发育（共 63 个，21 个与新报告的信号相关的基因）和关节退化（共 50 个，18 个与新报告的信号相关的基因；13 个基因在骨骼发育和关节退化类别中是共同的）有关。三个由新的遗传信号产生的效应基因，即 *CHST3*、*COL2A1* 和 *FBN2*，编码着结构蛋白。II 型胶原蛋白 $\alpha 1$ 链 (*COL2A1*) 编码的是软骨的基本结构成分，对关节的形成和骨骼的生长很重要。目前已发现许多疾病都同 *COL2A1* 有关，包括软骨异常和骨的异常，如脊柱骨骺线发育不良、Kniest 发育不良和早发性骨关节炎。纤维蛋白 2 (*FBN2*) 编码一种在细胞外基质中形成微纤维的糖蛋白，在个体的早期形态发生过程中具有重要作用。纤维蛋白能够有效地调节免疫反应、炎症和组织平衡，在骨重塑中很重要，并且可以调节 BMP 和 TGF- β 的局部可用性。*FBN2* 的突变会导致契约性蛛网膜病）。

同时有几个基因与信号传导途径有关。退化样家族成员 4 (*VGLL4*) 通过与 TEA 域 (TEAD) 转录因子相互作用而发挥作用。值得注意的是，我们发现另一个新的与关节置换和手部骨关节炎相关的信号位于这样的转录因子中，即 *TEADI* 基因，表明这两个信号背后有一个共同的分子途径。*TEADI* 在 Hippo 信号通路中发挥作用，并受参与机械感应和机械传导的 *YAP1* 和 *TAZ* 原基因蛋白的转录调控，而关节软骨的机械适应是骨关节炎的一个重要因素。同时有研究发现 *VGLL4* 的下调与 Wnt/ β -catenin 通路靶基因的上调有关。

Wnt 家族成员 1 (*WNT1*) 和 wnt 家族成员 10B (*WNT10B*) 参与 Wnt 信号通路，在骨关节炎发病机制中具有既定作用。*WNT10B* 的突变与肢体缺陷和牙齿异常有关，*WNT1* 的突变与成骨不全症有关。胰岛素样生长因子 1 受体 (IGF1R) 具有酪氨酸激酶活性，

介导胰岛素样生长因子的作用，并调节软骨矿化。

一氧化氮合酶 3 基因 (*NOS3*) 编码一氧化氮合酶 (eNOS) 的血管内皮异构体。*NOS3* 与小鼠的散发性肢体缺陷有关。LIM Homeobox 转录因子 1 β (*LMX1B*) 是一种转录因子。*LMX1B* 的突变导致一种罕见的常染色体显性疾病，其特征是指甲萎缩、髌骨发育不良或缺失，以及肘部和髂角发育不良。

Patched 1 (*PTCH1*) 编码 Hh 配体的受体，并调节平滑肌、毛细血管类受体 (*SMO*, 另一个与已知主导 SNV 有关的效应基因) 的活性。当结合时，*PTCH1* 放弃其对 *SMO* 的抑制作用，并激活 Hh 信号级联，这在控制软骨细胞的增殖中发挥了重要作用，也在软骨内层骨形成和纵向生长过程中刺激成骨作用。

还有几个新发现的高置信度效应基因与神经元有关。由 *C2orf40* (也叫 *ECRG4*) 编码的蛋白 Augurin 参与了动物模型的中枢神经系统发育，并与人类阿尔茨海默病和相关痴呆症的神经病理学特征有关。*TSEN15* 附近的 SNVs 与骨关节炎关联的人体测量特征有很强的相关性，如身高、体脂分布和根据 BMI 调整的腰围。*CUT* 样同源蛋白 1 (*CUX1*) 是参与大脑神经元分化和突触发生的转录因子。在肢体发育过程中被观察到的 *Cux1* 在软骨间的表达，也表明在关节形成中也有调节作用。

TRIO 和 f-actin 结合蛋白 (*TRIOBP*) 基因通过 2 个启动子编码多种蛋白异构体。*TRIOBP-1* 与 *TRIO* 和 f-actin 结合蛋白相互作用，共同在神经元形态发生中发挥关键作用并控制肌动蛋白细胞骨架组织、细胞运动和细胞生长。

肌管蛋白相关蛋白 2 (*MTMR2*) 在膜靶向、囊泡转运和信号转导途径的调节中具有重要作用。*MTMR2* 的突变导致 4B 型夏科-玛丽-托斯病，该病症状包含髓神经纤维普遍丧失，髓鞘局部折叠，神经对肌肉的神经信号传导不足，导致肌肉无力和萎缩。普遍表达的 CREB 结合蛋白 (*CREBBP*) 所编码的蛋白质在发育过程中起着关键作用，特别是与大脑大小调节、正确的神经细胞分化和神经前体细胞迁移有关，这些相关也在小鼠模型中得到了证明。

胆碱能受体毒蕈碱 2 (*CHRM2*) 参与了细胞反应的调节。对大鼠组织的分析亦发现其在全脑和人类神经母细胞瘤细胞中的表达。*CHRM2* 的变异易导致包括阿尔茨海默病在内的各种神经精神疾病。突触体相关蛋白 47 (*SNAP47*) 所编码的蛋白是一种可溶性 N-乙基马来酰亚胺敏感融合蛋白受体 (SNARE) 蛋白，参与囊泡转运和膜融合。SNARE 介导的融合是驱动突触传递、神经元发育和生长的一个重要机制。*SNAP47* 在物质出细胞模式和神经元形态发生中发挥作用。

有几个效应基因有免疫或炎症作用。例如，Toll 样受体 (*TLR4*) 编码的蛋白质在病原体识别和激活先天免疫反应中发挥着基本作用。*TLR4* 也可被与肌肉骨骼病症有关的受损组织产生的宿主分子激活。该基因与软骨细胞、成骨细胞和滑膜细胞的基因表达一起，将 *TLR4* 与类风湿性关节炎、骨关节炎和骨质疏松症联系在一起。而因此，*TLR4* 的调节或抑制已被建议作为这类疾病的一种治疗方法。T 细胞的激活可以通过影响肿瘤坏死因子配体超家族成员 11 (*TNFSF11*) 的表达而导致破骨细胞生成和骨吸收。*TNFSF11* 编码核因子卡帕-β 配体的受体激活剂 (也称为 RANKL)，这种细胞因子与类

风湿性关节炎的炎症性骨重塑有关, TNFSF11 水平的增加与关节炎严重程度的恶化有关, 并在破骨细胞生成中具有一定作用。核受体亚家族 3 组 C 成员 1 (*NR3C1*) 编码糖皮质激素受体 (GR), 在细胞质中循环, 参与炎症反应。在骨关节炎中, 成骨细胞和软骨细胞中的内源性糖皮质激素信号是有害的。

磷酸果糖激酶 (PFKM) 对肌肉功能有一定的作用。它编码一类肌肉同工酶, 在糖酵解过程中催化果糖-6-磷酸的磷酸化。该基因的突变导致 Tarui's 病 (糖原贮藏病 7 型), 这是一种常染色体隐性代谢疾病, 临幊上以运动不耐受、肌肉痉挛、劳累性肌病和代偿性溶血为特征。

药物靶点识别

我们检查了所有 637 个至少有一个来自精细映射和功能分析的支持证据的基因的成药性 (Druggability)。在这 637 个基因中, 有 205 个存在于可药性基因组数据库中, 致使数据库中具有支持证据的基因富集了 1.46 倍。从这些骨关节炎可药用的靶点基因中, 有 71 个基因位于第一类, 其中包括已批准 (许可) 的药物和临幊开发中的药物的靶点。在有三个不同证据支持因果关系的 77 个基因中, 有 20 个是一级候选基因 (其中 18 个存在于 DrugBank 中), 其中 7 个对应于本研究中发现的新基因信号 (*CHST3*、*VDR*、*TNFSF11*、*IGF1R*、*NR3C1*、*CHRM2* 和 *NOS3*)。

在第 1 类基因中, 有 10 种候选药物先前已在骨关节炎的临床疗效试验或队列研究中得到研究 (6 个作用于新信号: *PPARD*、*NR3C1*、*VDR*、*MAPK14*、*IGF1R* 和 *CHST3*)。*PPARD* 拮抗剂舒林酸因其前列腺素合成酶活性而被授权作为非甾体抗炎药 (NSAID) 用于骨关节炎。*SLC1A1* 激动剂和神经性疼痛抑制剂普瑞巴林是骨关节炎的常用药。普瑞巴林与非甾体抗炎药美洛昔康共同用于短期治疗膝关节骨性关节炎的疼痛, 并且有一些支持性的临床试验数据。*NR3C1* 编码糖皮质激素受体, 该受体的激活具有广泛的抗炎和免疫调节作用。类似的, 有几个激动剂分子获得了上市许可。其中, 泼尼松龙长期以来一直被用作炎症性关节炎的疾病调节剂, 在最近的心脏结果预防评估 (HOPE) 研究中, 发现它能有效减少手部骨关节炎的疼痛和滑膜炎。*Cathepsin K* (由 C TSK 基因编码) 是一种在破骨细胞内的胶原蛋白降解中起关键作用的酶, MIV-711 是一种选择性的 Cathepsin K 抑制剂, 最近在一項 2 期临床试验中被证明能有效减少膝关节骨性关节炎患者的结构损伤。*VDR* 编码维生素 D 受体, 它的激活是钙代谢的一个主要调节器。补充维生素 D 对膝关节骨性关节炎的症状和结构损伤的临床试验结果不一, 但可能对维生素 D 缺乏的患者稍有益处。*EGLN2* 编码一种脯氨酸羟化酶 EGL 九号同源物 2, 其主要介导脯氨酸的羟化, 从而促进胶原蛋白和蛋白多糖的合成。尽管效果不一, 补充其激动剂抗坏血酸 (维生素 C) 与观察人群中的关节健康有关。在日本 ROAD 人群中, *HCAR2* 激动剂烟酸 (维生素 B3) 的缺乏被发现与膝关节骨关节炎的进展有关。*MAPK14* 拮抗剂 PH-797804 已在一項 2 期临床试验中进行了研究以考察 PH-797804 单独或与萘普生一起对膝关节骨性关节炎受试者的疼痛缓解情况 (NCT01102660), 然而本文尚未在 PubMed 或 ClinicalTrials.gov 上发现任何试验结果报告。最后, 碳水化合

物碘化酶 3 激动剂沙利度胺已被证明可通过涉及血管内皮生长因子 (VEGF) 表达下调的机制，在小鼠内侧半月板失稳模型中减弱早期骨关节炎的发展。

另外 45 个一级可药用目标都有市场授权或正在进行其他适应症的临床开发。其中 10 个是高置信度的效应基因，16 个为本文发现的新的遗传信号。这里介绍的关于它们在临床骨关节炎中作用的功能和流行病学证据为早期再利用调查提供了支持。一种抗体小分子，福斯塔马提尼，作为一种酪氨酸激酶抑制剂多次出现在候选药物中，它针对 *AAK1*、*EPHA5*、*GAK*、*GSK3A*、*MAP2K1*、*MAP2K6*、*PAK1* 和 *PRKCD*，并作为一种生物性疾病修饰抗风湿药物 (DMARD) 获得了上市许可。*JAK2* 抗体 baricitinib 和 *TYK2* 抗体 tofacitinib 都作为生物 DMARDs 上市，而 *MAP2K1* 抗体 binimetinib 作为生物 DMARD 目前处于 3 期临床试验。因此，这些药物中的每一种都为骨关节炎的再利用研究提供了临床机会和假定的机制。在其余的一级和二级可药用目标中，潜在的可药用分子处于较早的开发阶段，提供了可期的再利用机会。

讨论

我们的研究结果产生了关于负重和非负重关节之间差异的进一步知识，并指出了任何关节的骨关节炎发展的共同机制，以及关节类型骨关节炎中的特异性。事实上，骨和软骨的发展途径在负重和非负重关节的信号中被富集，确定了关节生长是任何形式的骨关节炎的共同机制。

我们同时在该疾病和其主要症状—疼痛之间建立分子联系。我们证明了骨关节炎和疼痛相关表型之间的遗传相关性，并确定了神经系统通路的信号富集。此外，几个高置信度的效应基因在神经病理学中也有作用。本研究中的大多数骨关节炎病例被定义为全关节置换和/或自我报告的骨关节炎，而这两种疾病表型都是由疼痛高度驱动的。这些基因的鉴定也可以对进一步的关节疼痛相关疾病产生影响，迄今为止，对这些疾病的认识还很有限。

大量的高置信度效应基因汇聚在软骨细胞的平衡和骨质疏松中发挥着重要作用的软骨内径上。其中几个已被识别的基因在 TGF-β 信号传导和功能方面起到重要作用。新发现的纤维蛋白 2 (*FBN2*) 信号与 *LTBP1* 和 *LTBP3* 一起，调节活性 TGFB1 的可用性。TGFB1 是软骨中 TGFB 的主要形式，其可以通过 SMAD3 信号激活一连串的包括我们目前研究中已经识别的的 ECM-基因，如碳水化合物硫代转移酶 3 (*CHST3*) 在内的下游基因，。

我们的数据为 FGF 信号级联 (*FGFR3*、*FGF18* 和 *PIK3R1*) 参与骨关节炎的原因提供了证据。目前，*FGF18* 正在进行临床试验，以确定其对骨关节炎的有效性。新发现的分子选手磷脂酰肌醇-3-激酶调节亚单位 1 (*PIK3R1*) 编码 IA 类磷脂酰肌醇 3 激酶 (PI3Ks) 的 p85a、p55a 和 p50a 调节亚单位，已知其在胰岛素的代谢作用中起着关键作用，是脂肪生成的必要条件。*PIK3R1* 的突变导致 agmaglobulinemia 7、免疫缺陷 36 和 SHORT 综合征，其特点是身材矮小、关节过度伸展、眼球凹陷和出牙延迟。本文同时发现了软骨细胞的增殖、分化和肥大转化之间的平衡受几个信号通路的串扰控制因果

关系的证据。PTHLH 和 IHH 信号 (*SMO* 和 *PTCH1*) 通过 FGFR3 拮抗信号传导。此外，我们确定了两个独立的遗传变异，暗示 Noggin (*NOG*) 是一个骨关节炎效应基因。Noggin 与 TGFB、BMPs 和 GDF5 结合，从而阻止与同源受体的结合。*NOG* 的突变会导致一系列的骨和软骨表型，这取决于突变的严重程度。

一些假定的骨关节炎致病基因参与了发育途径。骨骼发育可以通过几种方式与骨关节炎联系起来。首先，骨骼发育基因同发病前的关节(组织)特征，如软骨厚度 (*TGFA*、*FGFR3*、*RUNX2* 和 *PIK3R1*) 或关节形状 (导致关节的不同负荷) 相关。其次，骨骼发育途径可能参与了对关节损伤的反应。根据具体的基因构成，每个人对关节的破坏性触发物的反应可能是不同的，从而决定了在创伤或机械过载时发生骨关节炎的风险。对目前的研究信号进行的路径分析进一步证实了这一点，因为它揭示了通常参与对损害反应的路径的富集证据。

我们的数据还表明，关键的骨软骨通路的细微变化导致了对关节损伤和/或超负荷的不良反应，这也可能催化了软骨和滑膜的纤维化反应。我们确定 tenascin C (*TNC*) 是高置信度的效应基因之一。*TNC* 是细胞外基质的一个组成部分，与器官纤维化、炎症和心血管疾病有关。关节中的纤维软骨和纤维化的形成是骨关节炎退行性变化的一个主要因素。此外，TGF-β 信号水平的升高与骨关节炎病理和纤维化变化有关。TGF-β 也是上皮-间质转化 (EMT) 的有效诱导剂。EMT 是一个完全分化的上皮细胞向间质表型过渡并产生成纤维细胞的过程，是早期纤维化的驱动因素，也是对损伤或病理变化和炎症的典型反应，而这些都是骨关节炎的常见症状。纤维化的严重程度有同导致骨关节炎疼痛的退行性改变的程度相关。我们已经确定了许多参与诱导 (如 EMT 基因、*CUXI* 和 TGF-β 途径的多种分子成分) 和纤维化进展 (ECM 基因如 *TNC*、TGF-β 信号传导的 *FBN2*、*LTPP1*、*LTPP3*、*TGFB1* 和 *SMAD3*) 的变异的显著联系。这些发现表明，这些基因调控的综合变异可能共同促成了骨关节炎退行性变化的易感性和严重性。

71 个相关基因编码的分子是已批准 (许可) 的药物和临床开发中的药物的目标。我们的发现大大加强了这些潜在治疗方法的证据，提供了药物重新定位的机会，并为开发或重新利用这些干预措施来治疗骨关节炎提供了坚实的基础。我们的工作为后续的功能和临床研究提供了一个强有力的跳板。我们已经证明了不同骨关节炎患者群体之间的明显差异，例如基于疾病严重程度、受影响的关节部位和性别的差别。我们加强了对疾病遗传病因果学的理解，揭示了其生物学的机制，并为将遗传关联转化为骨关节炎药物开发提供了基础，最终帮助催化改善骨关节炎患者的生活的药物出现。

研究的局限性

加强遗传关联研究中的人群多样性对于发现风险变异、确定可能的因果等位基因、改善风险预测以及确保研究结果在全球人群中的可转移性非常重要。在这项工作中，只有不到 3% 的受试者具有非欧洲血统。展望未来，在骨关节炎的遗传关联研究中，迫切需纳入不同人群的基因型信息。

分清疾病开始时的活跃机制和疾病自然史中的活跃机制亦需要进行动物模型研究，

以深入研究疾病的动态。事实上，对新发现的高价值目标的机制进行研究将是后续研究的重点。同时我们需要临床试验将我们的研究结果推向机制和临床结果，从而阐明针对相关的基因和蛋白质，下游性状将如何受到影响，以及最终这些干预措施将如何影响患者的疾病结果。

致 谢

致 谢