**iMeta: 海南大学张家超组建立一种基于肠道微生物区系单核苷酸变异的基因标记物的炎症性肠病预测模型**

**一种新的基于肠道微生物区系单核苷酸变异的基因标记物的炎症性肠病预测模型的建立**

**Establishing a novel inflammatory bowel disease prediction model based on gene markers identified from single nucleotide variants of the intestinal microbiota**

DOI：<https://doi.org/10.1002/imt2.40>

发表时间：2022年7月24日

第一作者：Shuaiming Jiang (姜帅铭)；Denghui Chen (陈登辉)

通讯作者：Jiachao Zhang (张家超) zhjch321123@163.com；Shi Huang (黄适) shihuang@hku.hk

合作作者：Chenchen Ma (马臣臣)；Huanwei Liu (刘桓玮)

主要单位：

海南大学食品科学与工程学院

美国加州大学圣地亚哥分校精神病学系

香港大学牙科学院

**亮点**



* 本研究在炎症性肠病(IBD)患者的肠道微生物组中发现了基于单核苷酸变异的特异性基因特征
* 本研究建立了一种新的准确预测IBD的诊断方法
* 普氏栖粪杆菌 (*Faecalibacterium prausnitzii*) 在不同疾病中均丰度降低，但在基因标志物特征上呈现疾病特异性

**摘要**

肠道微生物区系是炎症性肠病(Inflammatory bowel disease, IBD)发生发展的重要环境因素。研究发现IBD患者体内普氏栖粪杆菌 (*Faecalibacterium prausnitzii*) 丰度显著降低，这种特征也被用来作为IBD诊断的生物标志物。然而，这种特征不只在IBD疾病中被发现，在大肠癌 (Colorectal cancer, CRC)中也出现了普氏栖粪杆菌丰度降低的情况，以此特征作为生物标志物可能会混淆结果。因此，我们基于单核苷酸变异(Single nucleotide variants, SNV)的基因特征建立了高精度预测克罗恩病(Crohn's disease, CD)的新模型。接下来，我们找到了五个富集在CD组的基因标志物用于建立CD预测模型，分别属于普氏栖粪杆菌和直肠真杆菌两个菌种。模型在区分克罗恩病和健康队列上具有较高准确度，其中在发现队列中AUC值为91.13%，在验证队列中也达到了79.55%的准确度。扩大健康队列后，模型仍保持较高的准确率(AUC=89.75%)。同时，模型也具有较好的疾病特异性，在区分CD和CRC人群队列中，AUC值达到了95.74%。本研究建立了一种预测IBD的新型诊断方法，也为其他非传染性疾病的早期无痛诊断提供了新的视角。

**引言**

炎症性肠病(IBD)是一种慢性的免疫介导性炎症性疾病，包括溃疡性结肠炎(UC)和克罗恩病(CD)，由肠道微生物区系和肠道免疫系统相互作用的改变引起。现已发现IBD患者体内普氏栖粪杆菌丰度显著降低，可作为IBD诊断的生物标志物。然而，普氏栖粪杆菌的减少不仅发生在IBD患者中，也发生在其他炎症性疾病的患者中，如结直肠癌、Ⅱ型糖尿病和牛皮癣等。此外，属于同一物种的菌株在遗传上可能存在5%到30%或更多的差异。因此，基于物种水平细菌丰度的IBD生物标志物诊断可能会导致结果的混淆。前期研究表明，建立基于单核苷酸变异(SNV)的预测模型来区分结直肠癌(CRC)和健康队列是可靠的。此外，为了减少肠道微生物SNV稀疏性的影响，我们对SNV进行注释后，将其进行归一化到基因水平上，以此寻找基于单核苷酸变异的基因标志物用于建模，我们认为基于基因水平的结果更符合适应性进化的生物学特征。

在此，我们基于对SNV和肠道微生物相应注释基因的分析，对IBD的发病机制提出了一个新的视角。首先，我们仔细挑选了测序深度大于10×的菌株，并从 NCBI数据库下载了标准菌株的完整基因组。接着，我们甄别发现队列中肠道微生物SNV，并计算每个菌株每个基因的上的SNV数量，其中包括CD(n=68)和对照组(n=34)。根据这些基因特征，我们建立了一个高精度的CD预测模型。模型的验证队列包括CD组(n=20)和对照组(n=22)。此外，使用UC组(n=76)和结直肠癌组(n=126)检测CD基因标志物的特异性。本研究建立了一种预测IBD的新型诊断方法，它可以更深入地评估肠道微生物区系在疾病过程中的作用，也为其他非传染性疾病的早期无痛诊断提供了新的视角。

**方法**

我们收集了IBD患者和健康人群粪便样本的鸟枪式宏基因组数据。原始数据下载自NCBI数据库，测序数据的详细信息可以在表S1中找到。根据PRJNA400072项目中的原始数据分类，将数据分为发现队列和验证队列。68例CD患者和34例健康对照作为发现队列进行建模，22例健康对照和20例CD患者用于验证。我们还收集了用于二次验证的公共IBD数据集。此外，76例溃疡性结肠炎(UC)和126例结直肠癌(CRC)患者粪便样本也被用于验证这些基因标记的特异性。本研究中使用的所有数据集都可以在表1中找到。

由于宏基因组测序的深度和覆盖率对SNV注释的限制，我们使用MetaPhlan2对微生物物种进行注释，并选择平均相对丰度大于0.5%的菌株进行SNV注释，以确保能够注释到。表2列出了筛选得到的17个参考基因组，以及NCBI数据库的代表菌株和GenBank登录号的信息。随后，使用SamTools (v1.11) 和BCFTools (v1.8) 将宏基因组测序的reads与参考菌株基因组对比甄别SNV，并计算属于每个菌株中每个基因的SNV总数。由于菌株的丰度直接影响被注释的SNV的数量，我们还通过将基因中的SNV数量除以菌株的相对丰度和测序深度的乘积来标准化基因中的SNV。有关代码的更多详细信息，请访问GitHub (<https://github.com/jsming1996/IBD_project>)。

我们把基因中注释SNV的数量作为特征使用R中的“随机森林”程序包筛选潜在生物标记物。用Wilcoxon秩和检验计算不同组间的富集基因 (p<0.05)。使用R中的“proc”程序包进行 ROC分析以评估基因生物标志物的准确性。用“Circlize”和“GRID”程序包绘制用于注释基因位置的基因组圈图。用KAKS\_Calculator 2.0计算非同义突变与同义突变（Ka/Ks）的比率。

**表1 粪便宏基因组测序样本收集**



**结果**

**SNV甄别与IBD模型中基因标记物的构建和验证**

我们使用17个菌株的全基因组构建基因组文库，并将每个样本的宏基因组数据与基因组文库比较来甄别SNV。属于PRJNA400072项目的样本被分为发现队列 (n=102) 和验证队列 (n=42)。发现队列包括68名CD患者和34名健康对照，总共有558,738个非冗余SNV被注释到28,816个基因上。SNV在基因中的位置和数量可能直接影响基因的功能，进而影响微生物的进化。我们以28,816个基因上的SNV数量为特征，进行随机森林建模，得到不同基因特征的重要度分数并进行排序。在CD组和对照组样本比较过程中，根据发现队列的随机森林结果选择了重要性分数大于0.01的5个基因 (见表S3)。我们也选择不同数量的基因标志物进行建模并进行了十倍交叉验证，当使用5个选择的基因时，错误率低于0.175 (见图1C)。最后，基于所选择的5个基因标志物和随机森林算法建立了CD预测模型。最终模型中每个基因的相应重要性分数如图1D所示。所选择的5个基因标志物：基因C4Q21\_RS08950 (FP\_RS08950)、基因C4Q21\_RS08935 (FP\_RS08935)、基因C4Q21\_RS06070 (FP\_RS06070)、基因C4Q21\_RS10895 (FP\_RS10895)和基因EUBREC\_RS1558 (Er\_RS15585) 属于两个种，即普氏栖粪杆菌和直肠真杆菌 (见表S4)。表S5显示了这五个基因中每个基因中的特定SNV突变信息。在发现队列中，模型区分CD组和对照组的准确率为91.13% (见图1A)，经Wilcoxon秩和检验比较，这 5个基因标志物均富集于CD组 (图1B)。

在验证队列中，AUC值也达到了79.55% (见图1E)。再次扩展的验证的公开IBD数据集也支持基因标志物的准确性，CD和对照样本之间的AUC值达到71.41% (见图1F)。而根据5种基因标志物建立的预测模型区分UC (n=76)和CD (n=88)组的准确率仅为58.46% (见图1H)，提示IBD患者UC和CD均为发生在肠道的慢性炎症反应，两种疾病个体肠道微生物之间存在重叠或相似的突变位点。发现队列中的CD样本 (n=68) 和扩展健康人群队列中的CD样本 (n=112) 之间的模型准确率高达89.75%，如图1G所示。此外，由于IBD患者发生大肠癌 (CRC)的风险增加，我们还使用了CRC队列 (n=126)来验证这些基因标记物在不同肠道疾病中的特异性，CD基因标志物可用于区分CD (n=88) 和CRC (n=126) 组，准确率高达95.74%，如图1H所示。



**图 1. 基于SNVs的基因标志物模型预测与验证。**

(A) 在发现队列中，CD和对照样本之间的预测准确率为91.13%。(B) 所选5个基因标志物的重要性分数和平均丰度。(C) 所选五个基因的10倍交叉验证误差图。(D) 所选5个基因的随机森林重要性得分。(E) 验证队列中，模型在区分 CD和对照样本的预测结果。(F) 在扩展的验证队列中，CD与健康对照样本的ROC曲线。(G) 模型在区分CD与其他健康样本的预测结果。(H) 模型在区分CD与其他疾病 (UC和CRC) 队列的预测结果。

**五个富集于CD的基因功能注释**

在5个基因标记中，Fp\_RS08950、Fp\_RS08935、Fp\_RS06070和Fp\_RS10895属于普氏栖粪杆菌，1个基因Er\_RS15585属于直肠真杆菌。图2的A-D显示了普氏栖粪杆菌和直肠真杆菌中SNV的突变类型和位置，其中不同颜色的点代表不同的突变类型，而点的位置指示了基因组中突变位点位置，外圈代表CD组，内圈代表对照组。Er\_RS15585基因功能注释为含有螺旋转角的结构域蛋白 (WP\_003505382.1)，而FP\_RS08950基因表达为假想蛋白(AXB29042.1) 。FP\_RS08935基因可表达ZF-HC2结构域蛋白(AXB29039.1)，而FP\_RS06070基因可产生多肽酶S51 (AXB28508.1)，FP\_RS10895基因功能为复制蛋白 (AXB29383.1)。这5个基因的Ka/Ks比率如图2E所示。Fp\_RS08950、Fp\_RS08935和Fp\_RS06070的Ka/Ks值在0.4~0.6之间，而Fp\_RS10895和Er\_RS15585的Ka/Ks值远小于0.1，基因受纯化选择。这可能是因为一般情况下非同义替换常常会导致有害的特征，只在少数情况下会导致进化优势。这些基因的突变正在进行纯化选择意味着表明这些突变正在被消除。

**图 2. 五个突变基因标志物的功能注释。**

(A，B) 普氏栖粪杆菌SNV的突变类型和位置；(C，D) 直肠真杆菌SNV的突变类型和位置；(E) Ka/Ks是非同义替换率 (Ka) 与同义替换率 (Ks) 的比率。Ka/Ks比值为1表明所研究的基因是中性进化；小于1表明基因受纯化选择；而比值大于1则被认为基因受正向选择。SNV，单核苷酸位点变异。

**讨论**

普氏栖粪杆菌 (*F. prausnitzii*) 和直肠真杆菌 (*E. rectale*) 均属于厚壁菌门，它们很容易发酵可溶性膳食纤维在肠道中产生短链脂肪酸，从而增加宿主抗炎细胞因子的水平并预防炎症性疾病。在本研究中，选取的基因标志物中有4个基因属于普氏栖粪杆菌。其中，Fp\_RS06070基因功能为肽酶S51，被认为具有营养吸收功能。在IBD患者肠道中普氏栖粪杆菌的丰度会下降，我们推测其在有限的肠道生态位中发生更多与营养摄取相关的突变，使其得以生存。此外，普氏栖粪杆菌还可以产生丁酸盐来维持肠道平衡，增强肠道免疫力，进而影响IBD的病程。此外，Fp\_RS08935基因还能产生含有zf-HC2结构域的蛋白。有趣的是，在大肠癌队列中，SNV标记也在Fp\_RS08935基因中丰富，这可能是由于CD和CRC患者的肠道菌群相似，普氏栖粪杆菌均呈下降趋势。但SNV突变位置和数量不同，这也证明了基于SNV在不同疾病队列中建立预测模型是可行的，特别是当菌株具有相似的丰度趋势变化时。

此外，突变多富集在CD组中，这表明在疾病组生态位受限的情况下，更多的突变更有可能发生。这5个基因标志物可以通过基因扩增检测其突变，实现对CD的无痛快速预测。本研究建立了一种预测IBD的新型诊断方法，可以更深入地解释肠道微生物区系在疾病过程中的作用，也同时为其他疾病的早期无痛诊断提供了一种新的视角。

**作者简介**



姜帅铭，海南大学食品科学与工程专业在读博士，目前研究方向为益生菌肠道菌群互作效应及定殖机制研究，相关学术成果已发表于Microbiome、Food & Function、Frontiers in Microbiology等期刊。



陈登辉，加州大学圣地亚哥分校生物信息学与系统生物学在读博士。目前研究方向为遗传学、微生物组大数据分析。

**通讯作者简介**



黄适，香港大学教授，博士生导师。研究方向为微生物预警人体慢病和衰老的机制及应用，已在Nature Methods、Microbiome、Genome Biol、Cell Host Microbe、ISMEJ、mBio等杂志发表论文共30余篇。先后主持国家自然科学基金项目1项，获中国博士后科学基金-中国科学院优秀博士后联合资助，中国博士后国际交流派出计划支持等。



张家超，海南大学教授，博士生导师。主要研究方向为肠道微生物与人类健康和基于肠道微生物组模型的食品营养及功能评价，擅长多组学及生物信息学技术。已在Microbiome、ISME J、Mol Nutr Food Res、Food Res Int、mSystems、npj Biofilms等高水平国际期刊发表SCI论文共52篇。主持及参与国家级课题5项和省重点研发计划2项。

**引文**

Jiang, Shuaiming, Denghui Chen, Chenchen Ma, Huanwei Liu, Shi Huang, and Jiachao Zhang. 2022. “Establishing a Novel Inflammatory Bowel Disease Prediction Model Based on Gene Markers Identified from Single Nucleotide Variants of the Intestinal Microbiota.” iMeta e40. <https://doi.org/10.1002/imt2.40>