网络出版时间:2022-11-28 19:43 网络出版地址:https://kns.cnki.net/kcms/detail/34.1065.R.20221128.1718.025.html

基于 cfDNA 甲基化和机器学习的男性胃癌早筛预测模型的建立

季 杰[1,2,3],齐 健[1,2,3],洪 波[1,3],王姝洁[1,3],孙瑞芳[4],曹雪玲[4],孙晓君[1,3],聂金福[1,3]

摘要 **目的** 利用新型的细胞游离 DNA(cfDNA)甲基化检测技术,建立中国男性人群胃癌患者 cfDNA 甲基化模型。**方法** 使用 cfDNA 甲基化免疫沉淀和高通量测序技术(cfMeDIP-seq)开展胃癌患者的全基因组甲基化的检测,利用生物信息学的方法定位胃组织来源的 cfDNA,提取区分胃癌患者的特异甲基化标签,通过随机森林算法建立诊断模型,开展胃癌早期筛查的临床验证研究。**结果** 基于胃癌样本和正常对照选取了前 63 个最为显著的差异甲基化区段,构建了 cfDNA 甲基化模型,并将此模型应用于胃癌早期筛查,灵敏度达到 85% 以上,特异性达到 95% 以上。验证集的灵敏度和特异性分别为 98.7% 和 99.0%,曲线下方面积大小(AUC)为 0.999。**结论** 该研究构建的 cfDNA 甲基化模型具有良好的胃癌预测性能。

关键词 胃癌;液体活检;cfDNA 甲基化;MeDIP-seq;机器学习

中图分类号 R 735.2

文献标志码 A **文章编号** 1000-1492(2022)12-1991-06

doi:10.19405/j.cnki.issn1000-1492.2022.12.024

胃癌在所有新发癌症病例中名列第 5,男性的发病率远高于女性[1]。通常胃癌确诊时已处于晚期[2],治疗与预后情况很不乐观,有效的早期筛查是提高治愈率和生存率的关键。液体活检相比于传统检测方法具有微创以及更易于检测出早期肿瘤等优点[3],但需要合适的肿瘤标志物。细胞游离 DNA(cell-free DNA,cfDNA)是在细胞凋亡、坏死等过程释放到血液中的核酸混合物,其组分中包含循环肿瘤 DNA(circulating tumor DNA,ctDNA),ctDNA 来自肿瘤 DNA 片段并具有甲基化等特征,水平与肿瘤大小之间具有相关性[4]。通过对血浆中的 cfDNA 进行分析可以获得肿瘤相关信息,用于癌症早期检测、预测治疗反应和预后等方面[5]。cfDNA 甲基化免疫沉淀和高通量测序技术(cell-free methylated DNA immunoprecipitation and high-throughput sequencing,cfMeDIP-seq)可以富集并分析血液中全基因组的 CpG 甲基化 cfDNA[6]。该研究计划利用 cfMeDIP-seq 技术建立一个具有高精度的胃癌早筛 cfDNA 甲基化模型。

2022-10-10 接收

基金项目:国家自然科学基金(编号:81872438);中国科学院合肥物质科学研究院院长基金青年“火花”项目(编号:YZJJ2022QN43)

作者单位:[1]中国科学院合肥物质科学研究院健康与医学技术研究所,医学物理与技术安徽省重点实验室,合肥 230031

[2]中国科学技术大学合肥物质科学研究院,合肥 230026

[3]中国科学院合肥肿瘤医院,合肥 230031

[4]中国医学科学院肿瘤医院山西医院肿瘤生物样本库,太原 030013

作者简介:季 杰,男,硕士研究生;

聂金福,男,研究员,硕士生导师,责任作者,E-mail:jeffnie@cmpt.ac.cn

1 材料与方法

1.1 病例资料 收集 2020 年 4 月—2021 年 12 月来自山西肿瘤医院的 76 例受试者临床资料(60 例胃癌患者,16 例健康者作为对照),胃癌患者的入组条件为已经病理诊断确诊为胃癌,包括病理学诊断、血常规、生化指标在内的所有临床检测的结果,都通过常规的临床检测来完成。在伦理委员会通过后开展临床试验(伦理委员会论证报告编号为 Y-2020-16),患者均签署了知情同意书。

1.2 主要试剂与仪器 Qubit dsDNA 高灵敏度检测试剂盒(美国 Thermo Fisher Scientific 公司);Qiagen 循环核酸试剂盒、QiagenMinElute PCR 纯化试剂盒(德国 Qiagen 公司);KAPA HyperPrep 试剂盒、KAPA HiFi 高保真热启动 DNA 聚合酶预混液(美国 KAPA Biosystem 公司);NEBNext Illumina 二代测序接头、USER 酶(美国 New England BioLabs 公司);MagMeDIP 试剂盒(美国 Diagenode 公司);AMPure XP beads 磁珠(美国 Beckman Coulter 公司);Illumina Nova 6000 测序仪(美国 Illumina 公司)

1.3 样本处理 使用 EDTA 真空试管(BD)收集受试者 8 ml 外周血,13 000 r/min 离心 10 min 后取上层血清,转移到微量离心管中再次在室温下 13 000 r/min 离心 10 min,将上清液分为 1~2 ml 的若干份并于 -80 ℃保存直到 DNA 提取。DNA 提取时,取 3~4 ml 血浆,通过 Qiagen 循环核酸试剂盒提取血液 cfDNA,并用 Qubit dsDNA 高灵敏度检测试剂盒

检测 cfDNA 浓度。

1.4 MeDIP-seq 流程 首先使用 KAPA HyperPrep 试剂盒对 cfDNA 样本进行末端修复和 a－尾化，然后对 cfDNA 样本添加 NEBNext 接头。然后用 USER 酶消化之后，再用 Qiagen MinElute PCR 纯化试剂盒进行纯化。

随后进行 MeDIP 过程，这里使用 MagMeDIP 试剂盒，处理后的样品分装到 2 个 0.2 ml PCR 管内：10% 的样品用于输入控制，其余 90% 用于免疫沉淀。然后使用试剂盒内的 5-mC 单克隆抗体 33D3 和磁珠进行免疫沉淀，清洗后样品用 AMPure XP 磁珠进行纯化。

文库扩增则使用 KAPA HiFi 高保真热启动 DNA 聚合酶预混液和 NEBNext Illumina 二代测序接头，这个过程通过 PCR 进行，之后再次使用 AMPure XP 磁珠对文库进行纯化。文库的片段大小分布通过核酸片段毛细管电泳分析仪进行分析并由 Qubit dsDNA 高灵敏度检测试剂盒进行定量。最后在 Illumina Nova 6000 平台进行测序，生成长度为 150 bp 的测序片段。

1.5 差异甲基化区域的检测 使用 FastQC version 0.11.7（http://www.bioinformatics.babraham.ac.uk/projects/fastqc/）和 MultiQC v1.9（https://multiqc.info/）来进行质控，去掉接头和不符合质量要求的序列，然后使用 bowtie2 v2.2.9 软件将序列与参考基因组（hg19）进行比对。通过 samtools v1.7 软件（http://samtools.sourceforge.net/）中的 markup 工具去除 PCR 重复序列，最后通过 Macs2 Version 2.2.7.1 软件（https://github.com/macs3-project/MACS）显示甲基化峰，以便进一步寻找差异甲基化区域（difeffential methylated genes，DMR）。最后对甲基化峰值进行处理并寻找 DMR，这个过程中还用到了 Bedtools Version v2.22.0（https://bedtools.readthedocs.io/en/latest/content/bedtools－suite.html）和 limma-voom（R package edgeR version 3.26.8 and limma version 3.6.0），最终通过 limma 软件检测 DMR。

DMR 的差异分析中，以 |logFC| >1，adj.p.value <0.001 为阈值筛选差异基因，每次选择前 100 个 DMR，该过程共重复 100 次，并从所有 10 000 个结果中筛选出重复次数超过 50 次的 DMR。

1.6 cfDNA 甲基化预测模型的构建 通过机器学习在 60 例胃癌患者样本和 16 例对照样本的甲基化测序数据中提取胃癌特异性甲基化指纹，并建立基于 cfDNA 甲基化的胃癌早筛模型。首先从总样本按照 80% 和 20% 的比例进行随机分组，这 80% 的样本再次随机按照 4∶1 的比例分为训练集（n = 49）和测试集（n = 12），最初分组中的另外 20% 的样本则作为验证集（n = 15）。该研究用随机森林算法来建立模型，ntree 设置为 500，mtry 设置为 10，这个过程重复进行使用随机选择的训练测试集进行 100 次，最后通过受试者工作特征曲线（ROC）来评估构建模型的效果。

1.7 统计学处理 数据处理基于 R project version 4.1.2 软件进行，连续变量差异通过 Wilcoxon 秩和检验进行分析，以 P < 0.05 为差异有统计学意义。绘制 ROC 曲线并通过 pROC 包计算曲线下面积（area under curve，AUC），通过该 AUC 值量化反映模型的预测性能。此外，DMR 富集分析基于基因本体论（gene ontology，GO）和京都基因与基因组百科全书（kyoto encyclopedia of genes and genomes，KEGG）进行。

2 结果

2.1 肿瘤组与对照组的临床信息和 cfDNA 浓度的比较 收集了共 76 例受试者（其中 60 例肿瘤患者，16 例为正常对照）的血浆 cfDNA 样本，所有受试者的年龄与 cfDNA 浓度见表 1。对照组样本与肿瘤组样本的年龄差异无统计学意义（P >0.05），而对照组样本的 cfDNA 平均浓度为 0.69 ng/ml，低于肿瘤组样本的 1.12 ng/ml。结合图 1 也可以看出，两组的年龄差异并不显著，且肿瘤组的 cfDNA 水平明显高于对照组。

表 1 所有肿瘤组和对照组的年龄和 cfDNA 浓度（$\bar{x} \pm s$）

项目	全体样本	对照组	肿瘤组	P 值
年龄（岁）	59 ± 8	56 ± 9	60 ± 8	0.077
DNA 浓度（ng/ml）	1.03 ± 0.71	0.69 ± 0.33	1.12 ± 0.75	0.023

2.2 DMR 的筛选与分布情况 通过 MeDIP-seq 技术对 cfDNA 样品进行检测，去除重复序列后在正常样本与肿瘤样本之间进行差异甲基化区域（DMR）的比较，使用 limma-trend 检验确定了共计 63 个出现超过 50 次且具有显著性差异的 DMR，并绘制成热图（图 2），纵坐标代表不同的 DMR，横坐标为样本，颜色越深说明高/低甲基化的程度越明显，由 60 例肿瘤组样本和 16 例对照组样本的 cfDNA 中筛选出差异最为显著的共 63 个 DMR。这些 DMR 在基因组中的定位情况是：44% 位于远端基因间区、40%

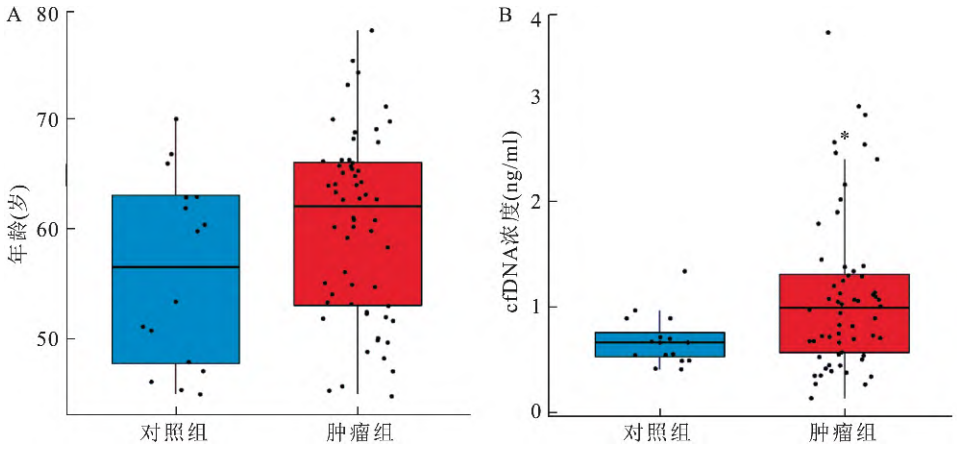

图1 对照组与肿瘤组年龄及血浆 cfDNA 浓度箱线图

A: 对照组与肿瘤组年龄分布; B: 对照组与肿瘤组 cfDNA 浓度分布; 与对照组比较: * $P<0.05$

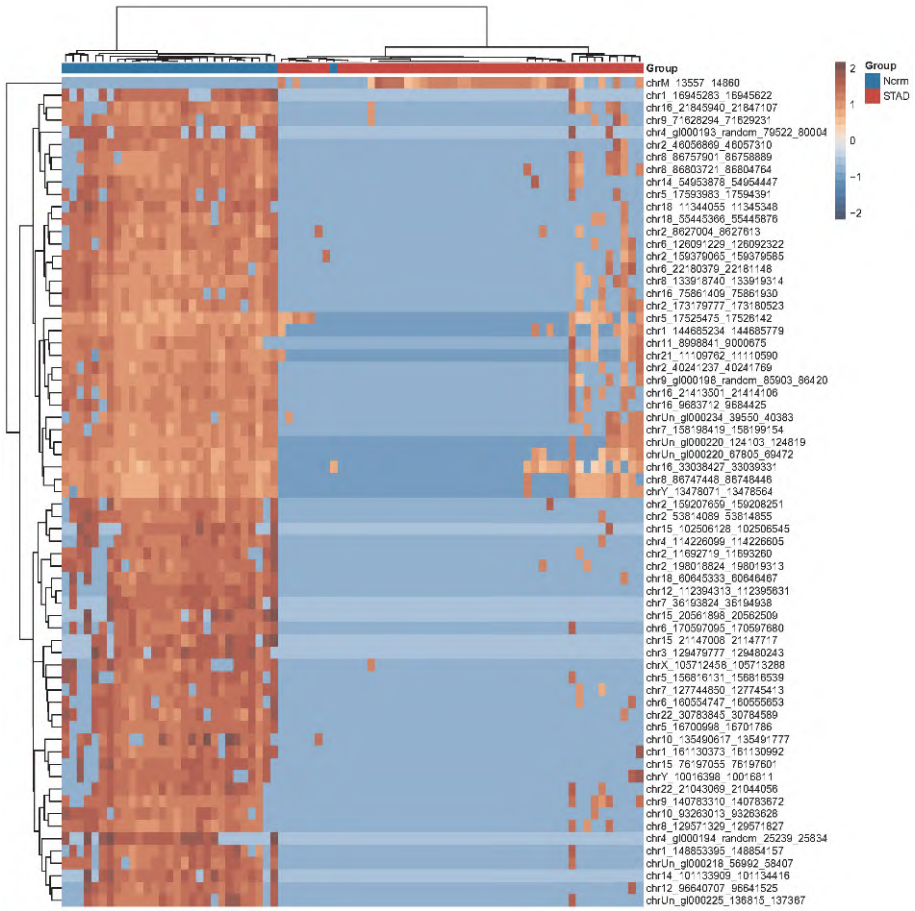

图2 筛选出的 DMR 热图

红色: 高甲基化; 蓝色: 低甲基化

位于内含子区、9% 位于启动子区,其余分布在外显子区、3′ UTR 区、下游区和 5′ UTR 区(图 3A)。绝大多数 DMR 在肿瘤组中都表现为低甲基化,仅在远端间区和内含子区存在少量的 DMR 表现为高甲基化(图 3B)。

2.3 甲基化模型预测效果评估 接着通过交叉验证来评估基于 MeDIP 谱预测肿瘤情况的能力,将受试者随机分为80% 的训练集和20% 的测试集。利

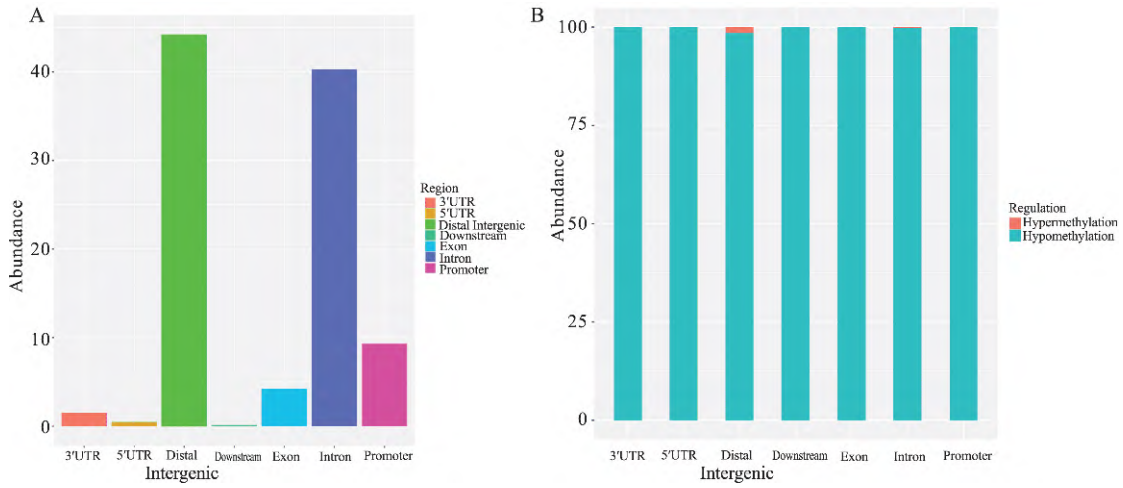

图3 DMR 在基因组中的分布情况

A: DMR 在基因组中的位置及比例; B: 各位置 DMR 的甲基化修饰情况

用训练集样本，通过筛选出的 63 个差异最显著的 DMR 对样本进行分类，并基于随机森林算法用这些 DMR 构建了一个可以用来评估受试者胃癌风险的模型，对测试样本进行风险评分，该过程重复 100 次，绘制风险分数箱线图（图 4），肿瘤组与对照组之间存在明显的差异。通过 ROC 曲线图来对模型的预测效果进行评估（图 5），利用该预测模型得出的对胃部肿瘤情况作预测的测试集敏感性为 89.0%，特异性为 98.5%，AUC 为 0.980；验证集的敏感性为 98.7%，特异性为 99.0%，AUC 为 0.999。

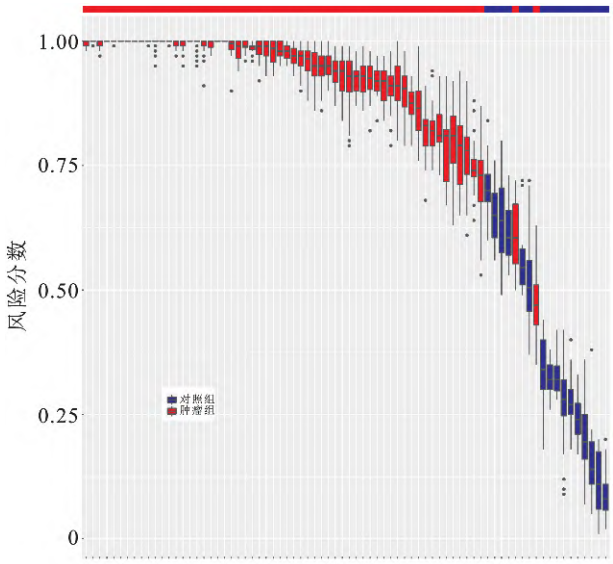

图4 正常对照和胃癌患者的单个血浆样本的预测风险分数箱线图

2.4 DMR 基因的富集分析 利用 GO 和 KEGG 对选择的 DMR 基因进行富集分析，根据 GO 的结果

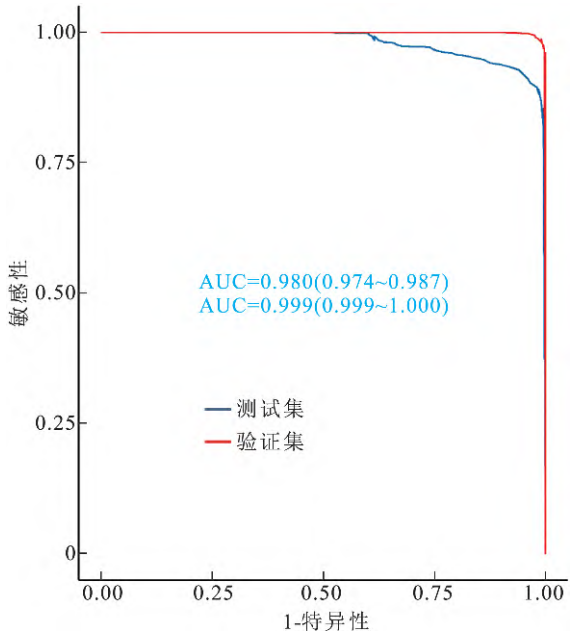

图5 测试集和验证集的 ROC 曲线

（图 6A），这些基因主要与神经和信号转导过程相关，在生物过程（biological process，BP）方面，参与了轴突形成、蛋白质去磷酸化、膜电位调节等过程，细胞组分（cellular component，CC）方面主要为细胞间连接、细胞基膜以及神经突触的组分，在分子功能（molecular function，MF）方面主要调节了肌动蛋白的结合、β-连环蛋白结合以及磷酸酯酶的活性。基于 KEGG 的分析结果显示（图 6B），差异 DMR 基因主要参与 Rap1 信号通路，并且与甲状旁腺激素的合成分泌以及吗啡成瘾性有关，差异均有统计学意

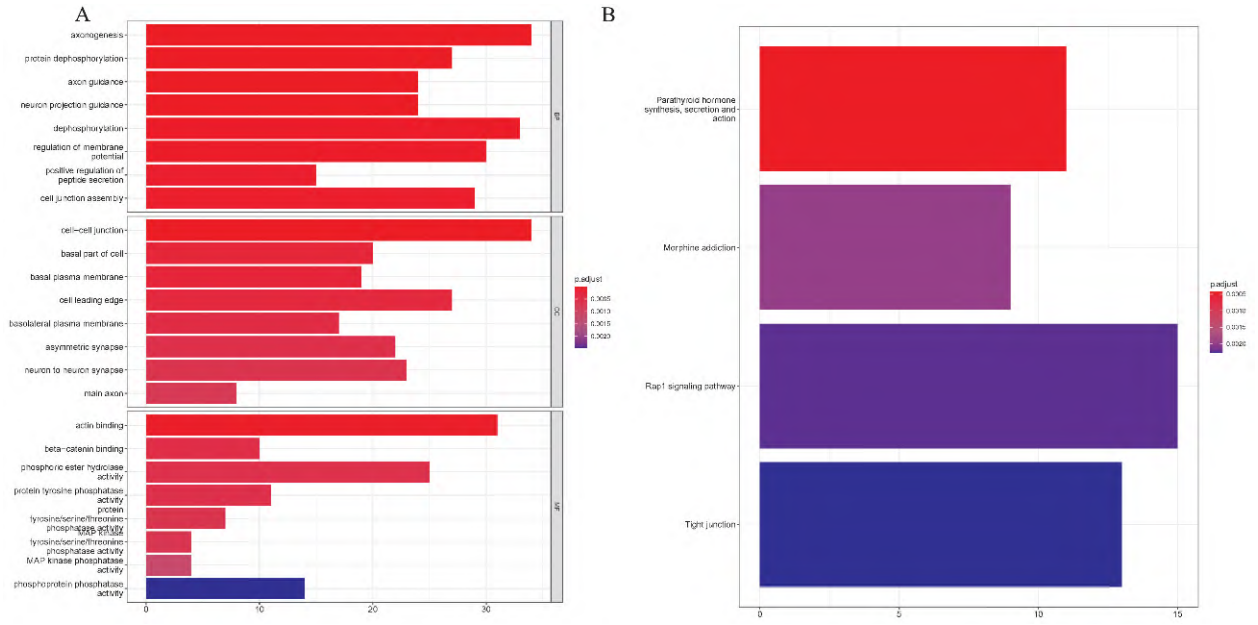

图6 胃癌相关 DMR 的富集分析

A: DMR 基因的 GO 分析结果，由上至下为 BP、CC、MF; B: DMR 基因的 KEGG 分析结果

义($P<0.05$)。

3 讨论

尽早发现并治疗处于早期胃癌的患者是提高胃癌治愈率和患者生存水平的关键。在该项研究中，首先比较了对照组和肿瘤组样本的 cfDNA 浓度差异情况，表明肿瘤患者的 cfDNA 水平要高于健康人群。有研究[7]表明，这可能与肿瘤组织和微环境的相互作用以及肿瘤代谢特性有关。但 cfDNA 的浓度不足以作为评估肿瘤风险的独立因素，因此，还需要寻找其他特征，比如甲基化。该研究对 cfDNA 基因组的甲基化情况作了分析，结果显示肿瘤组中大部分 DMR 都表现为低甲基化，有研究[8]表明，GpG 岛中的总体甲基化程度和去甲基化程度在不同的肿瘤间有很大差异，一种癌症表现出甲基化或是去甲基化取决于这些 DMR 基因参与的生物学过程，例如 BMP 信号通路和 LPA 信号通路。在特定基因中的去甲基化可以激活原癌基因并促进肿瘤发生[9]，在另一些基因中则是由甲基化来扮演这一角色。

有基于 cfDNA 等新型肿瘤标志物而进行的癌症早筛或预后的研究，如 Liggett et al[10]通过 Meth-Det56 甲基化技术对胰腺癌和慢性胰腺炎患者的 cfDNA 进行分析并确认了 17 个具有差异甲基化的基因启动子，并可以用于区分胰腺癌和慢性胰腺炎。Phallen et al[11]开发了一种被称为靶向纠错测序(TEC-seq)的方法，可以通过对 58 个癌症相关基因的检测来预测结直肠癌、乳腺癌、肺癌和卵巢癌这 4 种常见癌症。该研究选择 cfDNA 的甲基化作为预测胃癌的标志物也是建立在前期研究[6]的基础之上。在 cfDNA 的分析方法中该研究采用了 MeDIP-seq 法，由于血浆中 cfDNA 的丰度并不高，MeDIP-seq 法相比于亚硫酸氢盐修饰法受到更少的限制。该研究最终建立了一个基于 DMR 的胃癌早筛预测模型，其 AUC 达到 0.999 且敏感性与特异性均较高，结果显示，该研究建立的预测模型可以根据风险分数来将肿瘤样本与正常样本区分开来，具有良好的预测性能，有运用于辅助胃癌诊断或开发胃癌早筛试剂盒的潜力，对胃癌的早期筛查以及临床资源管理有重要的意义。

该研究具有以下几个优点：① 提供了一个仅基于血液 cfDNA 的胃癌早期筛查方案，材料易于从血液中获得，这意味着该方案相比于胃镜等常规检测手段更为便捷且适用于大规模筛查。尽管目前体液检测尚无法代替常规检测手段，但具有高准确度的肿瘤标志物在辅助诊断和潜在人群筛查中也能发挥不可忽视的作用；② 该研究是从中国招募受试者，建立的模型能更具针对性地应用于中国的胃癌防治。

综上所述，该研究表明基于 cfDNA 甲基化差异区域构建的胃癌早筛预测模型可以帮助识别和预测

患胃癌风险的患者，并且可以基于患者的风险评分进行分层管理，有助于降低胃癌的病死率和改善临床管理策略。但该研究仍然有一些局限性，例如所有样本均来自中国且目前样本总数较少。此外，由于该研究面向中国男性人群构建预测模型，不适用于女性风险人群，但仍具有一定参考价值。

参考文献

[1] Sung H, Ferlay J, Siegel R L, et al. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries [J]. CA Cancer J Clin, 2021, 71(3): 209 - 49.

[2] Van Cutsem E, Sagaert X, Topal B, et al. Gastric cancer [J]. Lancet, 2016, 388(10060): 2654 - 64.

[3] Chen M, Zhao H. Next-generation sequencing in liquid biopsy: cancer screening and early detection [J]. Hum Genomics, 2019, 13(1): 34.

[4] Newman A M, Bratman S V, To J, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage [J]. Nat Med, 2014, 20(5): 548 - 54.

[5] Luo H, Wei W, Ye Z, et al. Liquid biopsy of methylation biomarkers in cell-free DNA [J]. Trends Mol Med, 2021, 27(5): 482 - 500.

[6] Qi J, Hong B, Tao R, et al. Prediction model for malignant pulmonary nodules based on cfMeDIP-seq and machine learning [J]. Cancer Sci, 2021, 112(9): 3918 - 23.

[7] Kustanovich A, Schwartz R, Peretz T, et al. Life and death of circulating cell-free DNA [J]. Cancer Biol Ther, 2019, 20(8): 1057 - 67.

[8] Lee S T, Wiemels J L. Genome-wide CpG island methylation and intergenic demethylation propensities vary among different tumor sites [J]. Nucleic Acids Res, 2016, 44(3): 1105 - 17.

[9] Wilson A S, Power B E, Molloy P L. DNA hypomethylation and human diseases [J]. Biochim Biophys Acta, 2007, 1775(1): 138 - 62.

[10] Liggett T, Melnikov A, Yi Q L, et al. Differential methylation of cell-free circulating DNA among patients with pancreatic cancer versus chronic pancreatitis [J]. Cancer, 2010, 116(7): 1674 - 80.

[11] Phallen J, Sausen M, Adleff V, et al. Direct detection of early-stage cancers using circulating tumor DNA [J]. Sci Transl Med, 2017, 9(403): eaan2415.

Construction of prediction model for early screening in male patients with gastric cancer based on cell - free DNA methylation and machine learning

Ji Jie[1,2,3], Qi Jian[1,2,3], Hong Bo[1,3], Wang Shujie[1,3], Sun Ruifang[4], Cao Xueling[4], Sun Xiaojun[1,3], Nie Jinfu[1,3]

([1]*Anhui Province Key Laboratory of Medical Physics and Technology, Center of Medical Physics and Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei* 230031; [2]*University of Science and Technology of China, Hefei Institutes of Physical Science, Hefei* 230026; [3]*Hefei Cancer Hospital, Chinese Academy of Sciences, Hefei* 230031; [4]*Shanxi Provincial Cancer Hospital, Biobank of Tumor, Taiyuan* 030013)

Abstract ***Objective*** To construct a cell-free DNA(cfDNA) methylation model for early screening in male patients with gastric cancer by using novel cfDNA methylation detection technology. ***Methods*** Methylation information of the whole genome of gastric cancer patients were detected by cell-free methylated DNA immunoprecipitation and highthroughput sequencing (cfMeDIP-seq) technology and locate gastrogenic cfDNA. Then bioinformation methods were used to extract specific methylation labels which could distinguish GC patients and establish diagnosis model by random forest algorithm. Related validation clinical researches were also conducted. ***Results*** 63 most significant DMR were selected to construct the cfDNA methylation model based on GC samples and normal control samples, the goal sensitivity was above 85 percent while the goal specificity was above 95%. The sensitivity and specificity of the validation set were 98.7% and 99.0% while the area under curve(AUC) was 0.999. ***Conclusion*** The cfDNA methylation model constructed in this study has good performance in predicting GC.

Key words gastric cancer; liquid biopsy; cfDNA methylation; MeDIP-seq; machine learning