

# 上海交通大学医学院



学者介绍  
Author introduction



陈豪燕 博士

研究员、硕士生导师

ORCID ID: 0000-0001-6722-8534

CHEN Hao-yan

M.D, Ph.D

Professor, Master's Supervisor

ORCID ID: 0000-0001-6722-8534

**陈豪燕** (1980—), 上海交通大学医学院附属仁济医院、上海市消化疾病研究所研究员。2009年毕业于上海交通大学医学院并获得博士学位。2009—2012年赴美国加利福尼亚大学旧金山分校从事医学遗传学博士后研究工作。现任美国癌症学会会员、北美人类遗传学会会员，担任*International Journal of Cancer*等数十本国际知名期刊审稿专家。

• 主要研究方向为消化道肿瘤基因组生物信息学分析。以第一作者或通信作者在*Cell*、*Cancer Discovery*、*Clinical Cancer Research*、*Oncogene*、*PLoS Genetics*等著名杂志发表高影响力论文数十篇，主持多项国家自然科学基金和上海市自然科学基金课题，并入选上海市教育委员会高峰高原学科建设计划和青年东方学者计划。

**CHEN Hao-yan** born in 1980, principal investigator of Renji Hospital, Shanghai Jiao Tong University School of Medicine and Shanghai Institute of Digestive Disease. He graduated from Shanghai Jiao Tong University School of Medicine in 2009 with a doctor degree, and then worked on medical genetics as a postdoctor at the University of California in San Francisco. He is a member of the American Association of Cancer Research and the American Society of Human Genetics. Meanwhile, he is appointed as a reviewer of several journals about cancer genomics such as *International Journal of Cancer*.

• His project mainly focuses on the bioinformatics analysis of gastrointestinal cancer genomics. He has published several high impact papers including *Cell*, *Cancer Discovery*, *Clinical Cancer Research*, *Oncogene*, *PLoS Genetics* etc. He has been in charge of the projects funded by National Natural Science Foundation of China, Shanghai Natural Science Foundation. He was also enrolled into “Shanghai Municipal Education Commission—Gaofeng Clinical Medicine Grant Support”, and “Youth Eastern Scholar” at Shanghai Institutions of Higher Learning.



论著·基础研究

## 基于宏基因组学分析构建诊断大肠癌的肠道菌群标签

张昕雨<sup>1\*</sup>, 张 璟<sup>2\*</sup>, 朱小强<sup>1</sup>, 曹颖颖<sup>1</sup>, 陈豪燕<sup>1</sup>

1. 上海交通大学医学院附属仁济医院消化科, 上海市消化疾病研究所, 上海 200001; 2. 上海交通大学医学院附属仁济医院病案统计中心, 上海 200001

**[摘要]** 目的 · 根据粪便样本宏基因组学数据建立肠道菌群标签, 探索用于筛查与诊断大肠癌的非侵入性方法。方法 · 共纳入 285 例样本, 根据随机森林分类算法筛选出与大肠癌发生密切相关的特征细菌; 利用 6 种机器学习分类模型建立大肠癌的诊断模型, 并进行内部和外部验证。结果 · 首先筛选出了 9 种与大肠癌发生密切相关的特征细菌, 利用这 9 种细菌建立了 6 种诊断模型。其中随机森林模型准确率最高 (达 0.847 7), 其在内部验证集和外部验证集中的准确率分别为 0.815 8 和 0.734 4, 在全集中受试者工作特征 (receiver operating characteristic, ROC) 曲线下面积 (area under curve, AUC) 为 0.894。结论 · 根据粪便样本的宏基因组学数据, 利用随机森林算法建立了由 9 种细菌组成的诊断大肠癌的菌群标签, 能够有效对健康者与大肠癌患者进行区分。

**[关键词]** 大肠癌; 诊断; 肠道菌群; 机器学习; 随机森林

**[DOI]** 10.3969/j.issn.1674-8115.2018.09.004 **[中图分类号]** R446.5; R735.3 **[文献标志码]** A

### Bacterial signatures for diagnosis of colorectal cancer by fecal metagenomics analysis

ZHANG Xin-yu<sup>1\*</sup>, ZHANG Jing<sup>2\*</sup>, ZHU Xiao-qiang<sup>1</sup>, CAO Ying-ying<sup>1</sup>, CHEN Hao-yan<sup>1</sup>

1. Department of Gastroenterology and Hepatology, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai Institute of Digestive Disease, Shanghai 200001, China; 2. Medical Record Statistics Center, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200001, China

**[Abstract]** Objective · To construct bacterial signatures by analyzing fecal metagenomics for the screening and diagnosis of colorectal cancer (CRC). Methods · A total of 285 samples were included in the study. Diagnostic models for CRC according to six different machine learning algorithms were developed using the featured bacteria selected by random forest algorithm, and validated in validation sets. Results · Nine bacteria that differentiated CRC and the control were identified, with which 6 models were established. The best model was random forest model, with an accuracy of 0.847 7 in the training set. Its accuracy in two test sets was 0.815 8 and 0.734 4, respectively. The area under curve (AUC) of receiver operating characteristic of the random forest model in the set including all samples was 0.894. Conclusion · Bacterial signatures based on random forest algorithm for the diagnosis of CRC can differentiate patients with CRC and the control effectively, which suggests the potential clinical value of the bacterial signatures.

**[Key words]** colorectal cancer; diagnosis; intestinal bacteria; machine learning; random forest

目前, 在世界范围内, 大肠癌是第三常见的肿瘤, 其致死率居恶性肿瘤前 5 位<sup>[1]</sup>。很多大肠癌患者在疾病进展至中晚期才得到确诊, 这在很大程度上导致了大肠癌患者的高死亡率。近年来, 随着一些筛查方法的应用, 大肠癌患者的死亡率有所下降<sup>[2]</sup>。目前, 肠镜和免疫化学法便隐血检测 (fecal immunochemical test, FIT) 是筛查和诊断大肠癌比较常用的方法<sup>[3-4]</sup>。但是, 由于经济原因以及肠镜检查侵入性的检查过程, 很多人无法接受, 人们在接受肠镜检查前通常会感觉到恐惧、尴尬、不适<sup>[5]</sup>。因此, 大部分人都没有接受过结肠镜检查。此外, 便隐血筛查大肠癌

也存在局限性: 其敏感性不稳定, 变化范围较大<sup>[6]</sup>, 而且难以诊断没有出血的病灶<sup>[7]</sup>, 对一些有痔疮等疾病的患者可能导致假阳性的结果<sup>[8]</sup>。因此, 需要探索新的非侵入性检查方法来筛查大肠癌。

近年来肠道微生物在大肠癌发生和发展中的作用引起了广泛关注。许多研究发现, 大肠癌患者与健康者相比, 肠道菌群存在差异<sup>[9-11]</sup>。也有许多研究发现, 肠道菌群在大肠癌的发生、发展中具有一定作用<sup>[12-14]</sup>。本课题组研究发现, 具核梭杆菌 (*Fusobacterium nucleatum*) 能够通过自噬途径促进大肠癌耐药的发生<sup>[15]</sup>。大肠癌患者和健康者肠道菌

**[基金项目]** 国家自然科学基金 (31371273); 上海市教育委员会高校“青年东方学者”(QD2015003); 上海市教育委员会高峰高原学科建设计划 (20161309) (National Natural Science Foundation of China, 31371273; “Youth Eastern Scholar” at Shanghai Institutions of Higher Learning, QD2015003; Shanghai Municipal Education Commission—Gaofeng Clinical Medicine Grant Support, 20161309)。

**[作者简介]** 张昕雨 (1994—), 女, 硕士生, 电子信箱: stella941@126.com。张 璟 (1972—), 女, 高级统计师, 学士, 电子信箱: 13611793563@126.com。  
\* 为共同第一作者。

**[通信作者]** 陈豪燕, 电子信箱: haoyanchen@shsmu.edu.cn。



群的差异及肠道菌群在大肠癌发生、发展过程中的作用提示, 或许检测粪便中的肠道菌群可以作为一种新的筛查大肠癌的非侵入性方法。已有研究检测了某些肠道细菌诊断大肠癌的能力<sup>[16-17]</sup>, 但是单一细菌的诊断能力有限。也有研究利用多种细菌建立模型进行诊断<sup>[18-20]</sup>, 但用于构建模型的细菌较多, 难以应用于临床。宏基因组学 (metagenomics) 的概念由 Handelsman 等<sup>[21]</sup>首次提出, 宏基因组指环境中所有微生物的基因组的总体。二代和三代高通量测序技术的发展极大推动了宏基因组学的研究。宏基因组学为充分认识和利用肠道菌群, 进一步研究人肠道菌群与健康的关系提供了平台。本研究通过生物信息学分析, 利用 2 组独立的粪便样本的宏基因组学数据训练模型并进行验证, 建立了用于非侵入性筛查与诊断大肠癌的肠道菌群标签。

## 1 材料与方法

### 1.1 实验材料

下载来自 NCBI SRA 数据库的 2 个较大样本的大肠癌粪便宏基因组测序数据集, 分别为 ZellarG 数据集 (PRJEB6070) 和 YuJ 数据集 (PRJEB10878)。ZellarG 数据集包括健康者 66 名、大肠癌患者 91 名、结直肠腺瘤患者 42 名; YuJ 数据集包括健康者 53 名、大肠癌患者 75 名。

### 1.2 实验方法

#### 1.2.1 工作流程 工作流程示意图见图 1。

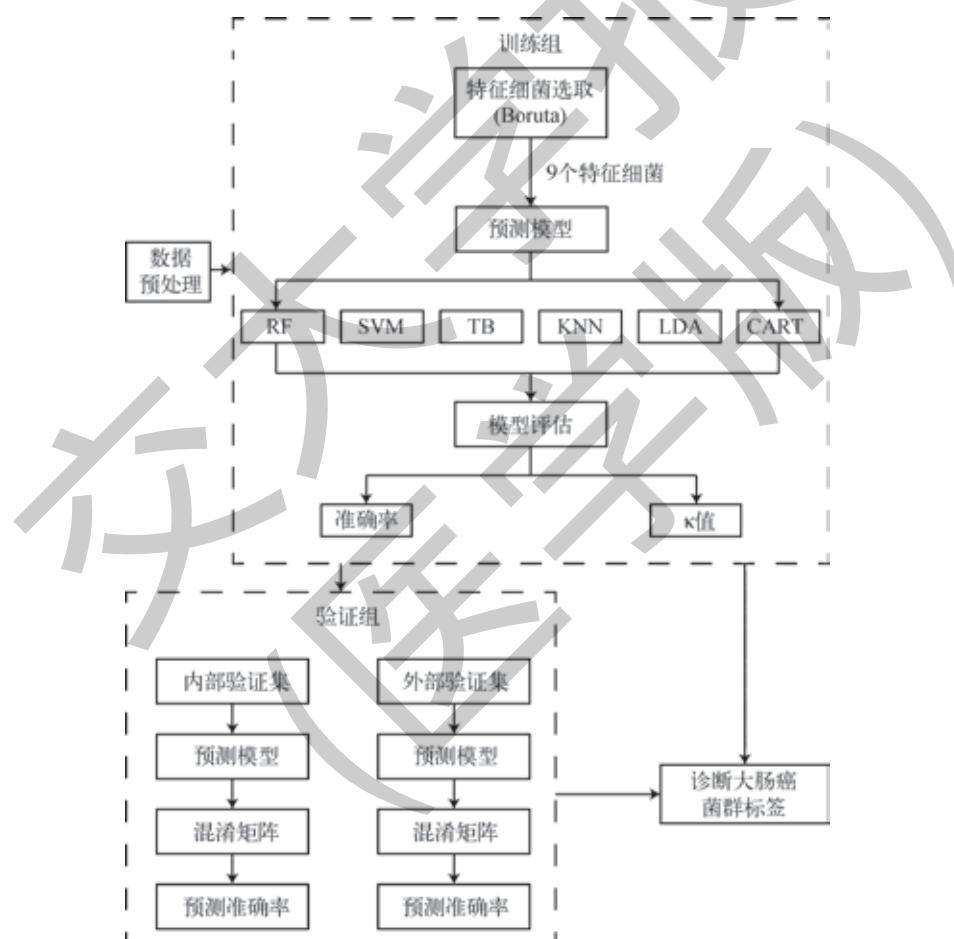


图 1 大肠癌菌群标签构建工作流程

Fig 1 Flow chart of construction of bacterial signatures for diagnosis of colorectal cancer

**1.2.2 数据预处理** 所有样本均按照人类微生物计划 SOP 所述进行标准预处理<sup>[22]</sup>。在预处理的数据上以默认参数运行 MetaPhlAn2 (v2.0) 产生微生物群落谱 (从界水平到种水平)<sup>[22]</sup>, 建立数据集。

由于本研究主要为构建诊断大肠癌的细菌标签, 故

删除 ZellarG 数据集中 42 名结直肠腺瘤患者的数据; 删除 ZellarG 数据集与 YuJ 数据集中病毒 (viruses)、真核生物 (eukaryota) 和古细菌 (archaea) 的数据, 仅保留细菌 (bacteria) 的数据; 再删除界 (kindom)、门 (phylum)、纲 (class)、目 (order)、科 (family) 及属 (genus) 水



平的数据，仅保留种（species）水平的数据。取筛选后 ZellarG 数据集与 YuJ 数据集中共有的细菌作为研究变量，分别对 2 个数据集重新规范化数据，使每个样本所有细菌总和为 100，再进行 z-scores 标准化。利用 caret 包中 createDataPartition 函数将 ZellarG 数据集删除结直肠癌患者后的 157 个样本按 75% 与 25% 的比例分为训练集（N=119）和内部验证集（N=38），将 YuJ 数据集作为外部验证集（N=128）。

**1.2.3 筛选用于构建诊断模型的特征细菌** 利用 Boruta R 包通过 Boruta 算法对 1.2.2 中筛选出的特征变量进行进一步筛选，选取特征细菌。Boruta 算法是一种基于随机森林（random forest, RF）分类算法的特征变量选择算法，这种算法迭代删除统计检验证明比随机成分与分类变量相关性更低的特征变量，最终获得与分类变量密切相关的特征变量，运算速度较快，不需要调整参数，能定量估计特征变量的重要性<sup>[23]</sup>。

**1.2.4 建立预测模型** 在训练集中利用 caret R 包根据 1.2.3 中筛选出的特征细菌进行模型训练。用 10 折交叉验证重复抽样拟合模型，将训练集分为 10 组，依次以其中 9 组作训练，另外 1 组作测试；如此重复 10 次，将 10 次交叉验证的结果取平均值评价分类模型准确率。

#### 1.2.5 选取 6 种分类模型建立预测模型并比较

(1) RF 在机器学习中，RF 分类器是最常用的算法之一。这种算法通过建立多个决策树最后整合成 RF，以各决策树判别结果的众数作为最后判别的结果，抗噪能力较强，分类的准确率相对较高，且很少出现过度拟合。设置 RF 中树数为 500（ntree=500），每棵树每个节点预选 5 个特征变量（mtry=5），通过进一步的计算确定一个最具有分类能力的变量，最终建立 RF 模型。

(2) 支持向量机（support vector machines, SVM） SVM 是一种有指导的学习模式，通过将原来有限维的数据映射到高维或者无限维的空间中，选择以最大间隔把样本分开的超平面。选择径向核函数（svmRadial）， $\Sigma$  值取 0.55，惩罚参数 C 取 0.50。

(3) 装袋决策树（treebag, TB） 装袋算法通过有放回的抽样训练多棵决策树，最终根据各决策树预测结果的众数确定预测结果。

(4) k 近邻算法（k nearest neighbours, KNN） KNN 是一种非参数分析方法，对某一样本的分类由其周围最相邻的 k 个邻居“投票”决定，是机器学习中较为简单的算法之一。k 值取 5。

(5) 线性判别分析（linear discriminant analysis, LDA） LDA 根据 Fisher 准则，尝试找出一个两类样本的特征的线性组

合以区分样本。

(6) 分类回归树（classification and regression tree, CART） CART 是一种非参数分析方法，是一种二叉决策树，通过剪枝避免过度拟合。惩罚参数 cp 值取 0。

**1.2.6 评估预测准确率** 在训练集中，用交叉验证准确率的平均数表示预测的准确率。 $\kappa$  值用于一致性检验；通常情况下， $\kappa$  值在 0.00 ~ 0.20 表示一致性极低，在 0.21 ~ 0.40 表示一致性一般，在 0.41 ~ 0.60 表示一致性中等，在 0.61 ~ 0.80 表示一致性较好，在 0.81 ~ 1.00 表示几乎完全一致。用内部验证集和外部验证集验证预测模型的准确程度，通过混淆矩阵可直接比较真实值与预测值，利用准确率和  $\kappa$  值比较各模型的性能。绘制受试者工作特征（receiver operating characteristic, ROC）曲线。

### 1.3 统计学分析

通过 R 3.5.0 软件 (<https://cran.r-project.org/>) 进行数据分析及作图。划分训练集与验证集以及训练、验证预测模型主要通过 caret 包 (<https://cran.r-project.org/web/packages/caret/>) 实现。筛选特征细菌主要通过 Boruta 包 (<https://cran.r-project.org/web/packages/Boruta/>) 实现。caret 包包括一组简化创建预测模型处理过程的功能，其中 createDataPartition 函数基于结果创建数据的平衡分割，利用其建立训练集和内部验证集；train 函数采用重采样方法评估模型的调整参数对模型性能的影响，从而选出最佳模型并估计模型性能，利用其选择参数并训练模型；predict 函数可在验证集中建立混淆矩阵，评估模型。pROC 包 (<https://cran.r-project.org/web/packages/pROC/>) 采用非参数分层或非分层采样方法进行自举操作，并建立 ROC 曲线、计算 ROC 曲线下面积（area under curve, AUC）。

通过准确率、 $\kappa$  值及混淆矩阵评估模型性能。训练集中的准确率为 10 次交叉验证所得准确率的平均值，验证集中的准确率为真阳性率与真阴性率之和，准确率置信区间取 95%； $\kappa$  值代表根据模型分类与完全随机分类相比减少的分类错误的比例，具体评价标准如 1.2.5 所述；混淆矩阵将模型的预测值与真实值进行比较，能够直接观察各模型的效果。

## 2 结果

### 2.1 数据预处理结果

筛选后 ZellarG 数据集与 YuJ 数据集均剩余 436 个共同特征变量，ZellarG 数据集有 157 例样本，YuJ 数据集有 128 例样本。训练集 119 例样本（健康者 50 例，大肠癌患



者 69 例), 内部验证集 38 例样本 (健康者 16 例, 大肠癌患者 22 例), 外部验证集 128 例样本 (健康者 53 例, 大肠癌患者 75 例)。

**表 1 9 种特征细菌**  
Tab 1 Names of nine featured bacteria

序号	界	门	纲	目	科	属	种
1	细菌	厚壁菌	梭菌	梭菌	真杆菌	真杆菌	霍氏真杆菌
2	细菌	厚壁菌	梭菌	梭菌	位置未定的梭菌目 11	微单胞菌	未分类微单胞菌
3	细菌	厚壁菌	梭菌	梭菌	消化链球菌	消化链球菌	口炎消化链球菌
4	细菌	厚壁菌	梭菌	梭菌	真杆菌	真杆菌	挑剔真杆菌
5	细菌	拟杆菌	拟杆菌	拟杆菌	紫单胞菌	副杆菌	粪拟杆菌
6	细菌	厚壁菌	梭菌	梭菌	颤螺旋菌	颤螺旋菌	未分类颤螺旋菌
7	细菌	厚壁菌	杆菌	乳杆菌	链球菌	链球菌	唾液链球菌
8	细菌	厚壁菌	梭菌	梭菌	梭菌	梭菌	共生梭菌
9	细菌	厚壁菌	梭菌	梭菌	梭菌	梭菌	哈氏梭菌

### 2.3 建立预测模型

根据选出的 9 种特征细菌, 在训练集中训练模型并进行 10 折交叉验证计算不同模型的预测准确率和  $\kappa$  值。6 种模型的准确率和  $\kappa$  值见表 2。

**表 2 6 种模型的准确率与  $\kappa$  值**  
Tab 2 Accuracy and  $\kappa$  value of 6 different models

模型	准确率	$\kappa$ 值
RF	0.847 7	0.684 1
SVM	0.769 7	0.516 3
TB	0.781 8	0.566 8
KNN	0.824 2	0.631 5
LDA	0.816 6	0.615 3
CART	0.790 9	0.571 5

### 2.4 在验证集中验证各模型的分类效果

在内部验证集和外部验证集中对各模型进一步进行验证, 通过混淆矩阵和准确率判断各模型的识别能力。在 ZellarG 数据集与 YuJ 数据集组成的合集中计算出全集的 AUC 值。在验证集中, 各模型准确率及混淆矩阵如下。

**2.4.1 RF** 内部验证集准确率 (95% CI) 为 0.815 8 (0.656 7, 0.922 6); 外部验证集准确率 (95% CI) 为 0.734 4 (0.649 1, 0.808 5)。混淆矩阵见表 3。

**2.4.2 SVM** 内部验证集准确率 (95% CI) 为 0.657 9 (0.486 5, 0.803 7); 外部验证集准确率 (95% CI) 为 0.671 9 (0.583 3, 0.752 2)。混淆矩阵见表 4。

### 2.2 特征细菌选取

在训练集中通过 Boruta 算法选取了 9 种特征细菌用于建立模型 (表 1)。

**表 3 验证集中 RF 模型分类结果 (n)**  
Tab 3 Discrimination results of RF in test sets (n)

预测分类	实际分类	
	对照	CRC
<b>内部验证集</b>		
对照	11	2
CRC	5	20
<b>外部验证集</b>		
对照	31	12
CRC	22	63

**表 4 验证集中 SVM 模型分类结果 (n)**  
Tab 4 Discrimination results of SVM in test sets (n)

预测分类	实际分类	
	对照	CRC
<b>内部验证集</b>		
对照	6	3
CRC	10	19
<b>外部验证集</b>		
对照	17	6
CRC	36	69

**2.4.3 TB** 内部验证集准确率 (95% CI) 为 0.736 8 (0.569 0, 0.866 0); 外部验证集准确率 (95% CI) 为 0.734 4 (0.649 1, 0.808 5)。混淆矩阵见表 5。

**表 5 验证集中 TB 模型分类结果 (n)**  
Tab 5 Discrimination results of TB in test sets (n)

预测分类	实际分类	
	对照	CRC
<b>内部验证集</b>		
对照	9	3
CRC	7	19
<b>外部验证集</b>		
对照	28	11
CRC	25	64



**2.4.4 KNN** 内部验证集准确率( $95\% CI$ )为 $0.6316$ ( $0.4599, 0.7819$ )；外部验证集准确率( $95\% CI$ )为 $0.5469$ ( $0.4565, 0.6350$ )。混淆矩阵见表6。

**表6 验证集中KNN模型分类结果( $n$ )**  
Tab 6 Discrimination results of KNN in test sets( $n$ )

预测分类	实际分类	
	对照	CRC
<b>内部验证集</b>		
对照	9	7
CRC	7	15
<b>外部验证集</b>		
对照	28	33
CRC	25	42

**2.4.5 LDA** 内部验证集准确率( $95\% CI$ )为 $0.7105$ ( $0.5410, 0.8458$ )；外部验证集准确率( $95\% CI$ )为 $0.6797$ ( $0.5915, 0.7594$ )。混淆矩阵见表7。

**表7 验证集中LDA模型分类结果( $n$ )**  
Tab 7 Discrimination results of LDA in test sets( $n$ )

预测分类	实际分类	
	对照	CRC
<b>内部验证集</b>		
对照	8	3
CRC	8	19
<b>外部验证集</b>		
对照	26	14
CRC	27	61

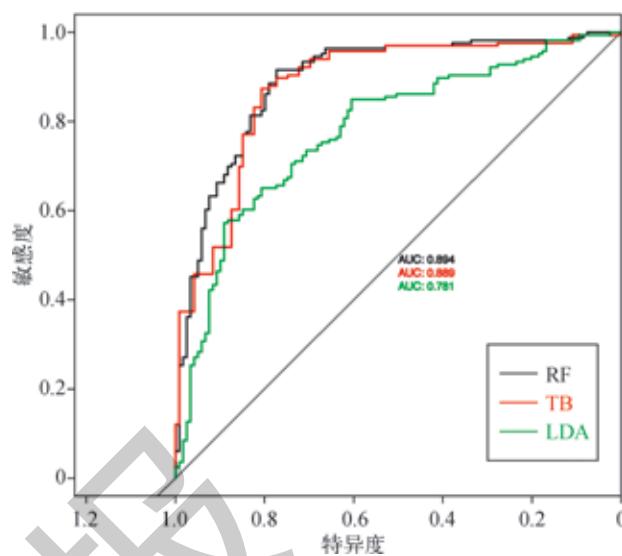
**2.4.6 CART** 内部验证集准确率( $95\% CI$ )为 $0.7895$ ( $0.6268, 0.9045$ )；外部验证集准确率( $95\% CI$ )为 $0.6172$ ( $0.5272, 0.7017$ )。混淆矩阵见表8。

**表8 验证集中CART模型分类结果( $n$ )**  
Tab 8 Discrimination results of CART in test sets( $n$ )

预测分类	实际分类	
	对照	CRC
<b>内部验证集</b>		
对照	10	2
CRC	6	20
<b>外部验证集</b>		
对照	24	20
CRC	29	55

## 2.5 在全集中的ROC曲线及AUC值

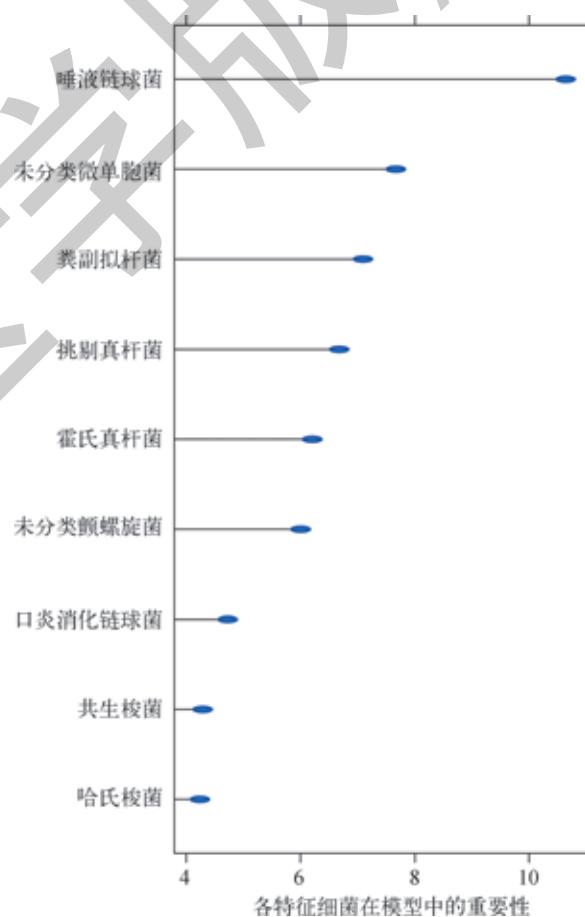
RF、TB及LDA这3种模型准确率相对较高且相对稳定，图2示RF、TB及LDA3种模型在全集中的ROC曲线及AUC值。



**图2 ROC曲线比较RF、TB及LDA3种模型在全集中的判别效能**  
Fig 2 ROC analysis of sensitivity and specificity of diagnosing CRC by RF, TB and LDA models

## 2.6 最佳模型中各特征细菌的重要性

在RF模型中检测9种特征细菌对预测分类的重要性，见图3。



**图3 RF模型中特征细菌的重要性**  
Fig 3 Needle plot of variable importance values in RF model



### 3 讨论

近年来,越来越多的研究证明肠道菌群参与了大肠癌的发生和发展<sup>[12-14]</sup>。以往已有研究利用肠道菌群作为生物标志物建立诊断大肠癌的菌群标签<sup>[16-20]</sup>;但是,单一细菌诊断能力不理想,或者用于构建模型的菌群太多,临床应用较为困难。本研究通过对2组肠道菌群宏基因组测序数据进行数据挖掘,筛选出9种与发生大肠癌密切相关的特征细菌;并基于这9种特征细菌,根据不同的机器学习的模型,建立了诊断大肠癌的菌群标签,同时比较了各模型的判别效果。结果显示,RF模型建立的菌群标签的预测准确率最高,且其在内部验证集和外部验证集中均能保持相对较好的预测准确率。通过这样的模型,能够进一步联系肠道菌群与临床表型,通过获取更多的信息完善肠道菌群标签,能够促进菌群标签在临床的应用。

筛选出的9种细菌中,相较于大肠癌患者,真杆菌属在健康者的肠道中更为常见<sup>[11,18]</sup>,而消化链球菌属、微单胞菌属、梭菌属在大肠癌患者肠道中更为常见<sup>[16,19,24-26]</sup>。这9种特征细菌参与大肠癌发生和发展的机制已有报道。真杆菌与链球菌属均与肠道微环境有关。真杆菌通过利用乳酸和乙酸生成丁酸从而减少乳酸的堆积,维持肠道的微生态稳态<sup>[27]</sup>,因而在健康者肠道中更为常见;而大肠癌早期发生的Wnt通路的激活及APC基因的突变引起肠道特异性改变,形成有利于链球菌属的解没食子酸链球菌的微环境,从而促进其在肠道的定植<sup>[28]</sup>,因而在大肠癌患者中更为常

见。梭菌、消化链球菌、颤螺旋菌可能通过不同的机制促进大肠癌的发生和发展。梭菌属中的酪酸梭菌可以通过调节miR-200c介导促炎因子肿瘤坏死因子α和白介素12的产生,从而促进炎症相关的大肠癌的发生和发展<sup>[29]</sup>;消化链球菌通过诱导细胞内胆固醇合成,促进结肠细胞异型增生<sup>[25]</sup>;颤螺旋菌能够直接调节维护肠道屏障完整性的组分从而影响肠道的渗透性<sup>[30]</sup>,也可能与大肠癌的发生和发展有关。副杆菌属在大肠癌中的作用尚存在争议。有研究显示副杆菌属与肿瘤负荷呈正相关<sup>[26]</sup>,也有报道称其与肿瘤负荷呈负相关<sup>[31]</sup>,其具体作用有待进一步探索。这些研究结果均进一步证明了本研究中建立的菌群标签的合理性。

最终选择的随机森林模型是一种基于决策树由多个树组成的分类器算法<sup>[32]</sup>,结合了Breimans<sup>[33]</sup>的自举汇聚法和Ho<sup>[32]</sup>的随机子空间法,可以处理大量输入变量,学习速度快,抗噪能力强。Tsuji等<sup>[34]</sup>利用RF筛选出对FOLFOX疗法有反应的大肠癌患者的相关基因,为大肠癌的个性化治疗奠定了基础。

本研究存在一定的不足。由于所获取的临床信息的限制,没有进一步分析大肠癌患者的分期、饮食、血脂等因素对肠道菌群的影响。此外还需扩大样本量,对菌群标签进行进一步的验证。在此基础上,可进行临床试验,以期将该肠道菌群标签应用于临床。

本研究根据肠道菌群的宏基因组测序数据建立了诊断大肠癌的肠道菌群标签,为大肠癌的非侵入性筛查诊断新策略的发展提供了思路。

### 参·考·文·献

- [1] Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012[J]. Int J Cancer, 2015, 136(5): E359-E386.
- [2] Edwards BK, Ward E, Kohler BA, et al. Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates[J]. Cancer, 2010, 116(3): 544-573.
- [3] Rex DK, Johnson DA, Anderson JC, et al. American college of gastroenterology guidelines for colorectal cancer screening 2008[J]. Am J Gastroenterol, 2009, 104(10): 739.
- [4] Sung JJY, Ng SC, Chan FKL, et al. An updated Asia Pacific Consensus Recommendations on colorectal cancer screening[J]. Gut, 2015, 64(1): 121-132.
- [5] Jones RM, Devers KJ, Kuzel AJ, et al. Patient-reported barriers to colorectal cancer screening: a mixed-methods analysis[J]. Am J Prev Med, 2010, 38(5): 508-516.
- [6] Lee JK, Liles EG, Bent S, et al. Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis[J]. Ann Intern Med, 2014, 160(3): 171.
- [7] Eklöf V, Löfgren-Burström A, Zingmark C, et al. Cancer-associated fecal microbial markers in colorectal cancer detection[J]. Int J Cancer, 2017, 141(12): 2528-2536.
- [8] van Turenhout ST, Oort FA, Terhaar sive Droste JS, et al. Hemorrhoids detected at colonoscopy: an infrequent cause of false-positive fecal immunochemical test results[J]. Gastrointest Endosc, 2012, 76(1): 136-143.
- [9] Feng Q, Liang S, Jia H, et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence[J]. Nat Commun, 2015, 6: 6528.
- [10] Kostic AD, Gevers D, Pedamallu CS, et al. Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma[J]. Genome Res, 2012, 22(2): 292-298.
- [11] Wang T, Cai G, Qiu Y, et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers[J]. ISME J, 2012, 6(2): 320-329.
- [12] Arthur JC, Perez-Chanona E, Mühlbauer M, et al. Intestinal inflammation targets cancer-inducing activity of the microbiota[J]. Science, 2012, 338(6103): 120-123.
- [13] Kostic AD, Chun E, Robertson L, et al. *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment[J]. Cell Host Microbe, 2013, 14(2): 207-215.
- [14] Zackular JP, Baxter NT, Iverson KD, et al. The gut microbiome modulates colon tumorigenesis[J]. MBio, 2013, 4(6): e00692-13.
- [15] Yu T, Guo F, Yu Y, et al. *Fusobacterium nucleatum* promotes chemoresistance to colorectal cancer by modulating autophagy[J]. Cell, 2017, 170(3): 548-563.
- [16] Xie YH, Gao QY, Cai GX, et al. Fecal *Clostridium symbiosum* for noninvasive detection of early and advanced colorectal cancer: test and validation studies[J]. EBioMedicine, 2017, 25: 32-40.
- [17] Wong SH, Kwong TNY, Chow TC, et al. Quantitation of faecal *Fusobacterium* improves faecal immunochemical test in detecting advanced colorectal neoplasia[J]. Gut, 2017, 66(8): 1441-1448.
- [18] Zeller G, Tap J, Voigt AY, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer[J]. Mol Syst Biol, 2014, 10(11): 766.



- [19] Yu J, Feng Q, Wong SH, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer[J]. Gut, 2017, 66(1): 70-78.
- [20] Baxter NT, Ruffin MT, Rogers MAM, et al. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions[J]. Genome Med, 2016, 8(1): 37.
- [21] Handelsman J, Rondon MR, Brady SF, et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products[J]. Chem Biol, 1998, 5(10): R245-R249.
- [22] Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling[J]. Nat Methods, 2015, 12(10): 902.
- [23] Kursa MB, Rudnicki WR. Feature selection with the Boruta package[J]. J Stat Softw, 2010, 36(11): 1-13.
- [24] Drewes JL, White JR, Dejea CM, et al. High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia[J]. NPJ Biofilms Microbiomes, 2017, 3: 34.
- [25] Tsoi H, Chu ESH, Zhang X, et al. *Peptostreptococcus anaerobius* induces intracellular cholesterol biosynthesis in colon cells to induce proliferation and causes dysplasia in mice[J]. Gastroenterology, 2017, 152(6): 1419-1433.
- [26] Baxter NT, Zackular JP, Chen GY, et al. Structure of the gut microbiome following colonization with human feces determines colonic tumor burden[J]. Microbiome, 2014, 2: 20.
- [27] Louis P, Flint HJ. Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine[J]. FEMS Microbiol Lett, 2009, 294(1): 1-8.
- [28] Aymeric L, Donnadien F, Mulet C, et al. Colorectal cancer specific conditions promote *Streptococcus gallolyticus* gut colonization[J]. Proc Natl Acad Sci U S A, 2018, 115(2): E283-E291.
- [29] Xiao Y, Dai X, Li K, et al. Clostridium butyricum partially regulates the development of colitis-associated cancer through miR-200c[J]. Cell Mol Biol, 2017, 63(4): 59-66.
- [30] Lam YY, Ha CWY, Campbell CR, et al. Increased gut permeability and microbiota change associate with mesenteric fat inflammation and metabolic dysfunction in diet-induced obese mice[J]. PLoS One, 2012, 7(3): e34233.
- [31] Pfalzer AC, Nesbeth PDC, Parnell LD, et al. Diet- and genetically-induced obesity differentially affect the fecal microbiome and metabolome in Apc(1638N) mice[J]. PLoS One, 2015, 10(8): e0135758.
- [32] Ho TK. Random decision forests[M]// Kavarnaugh M, Storms P. Proceedings of the third international conference on document analysis and recognition. Los Alamitos: IEEE Computer Society Press, 1995: 278-282.
- [33] Breiman L. Random Forests[J]. Mach Learn, 2001, 45(1): 5-32.
- [34] Tsuji S, Midorikawa Y, Takahashi T, et al. Potential responders to FOLFOX therapy for colorectal cancer by Random Forests analysis[J]. Br J Cancer, 2012, 106(1): 126-132.

[ 收稿日期 ] 2018-05-31

[ 本文编辑 ] 吴 洋

