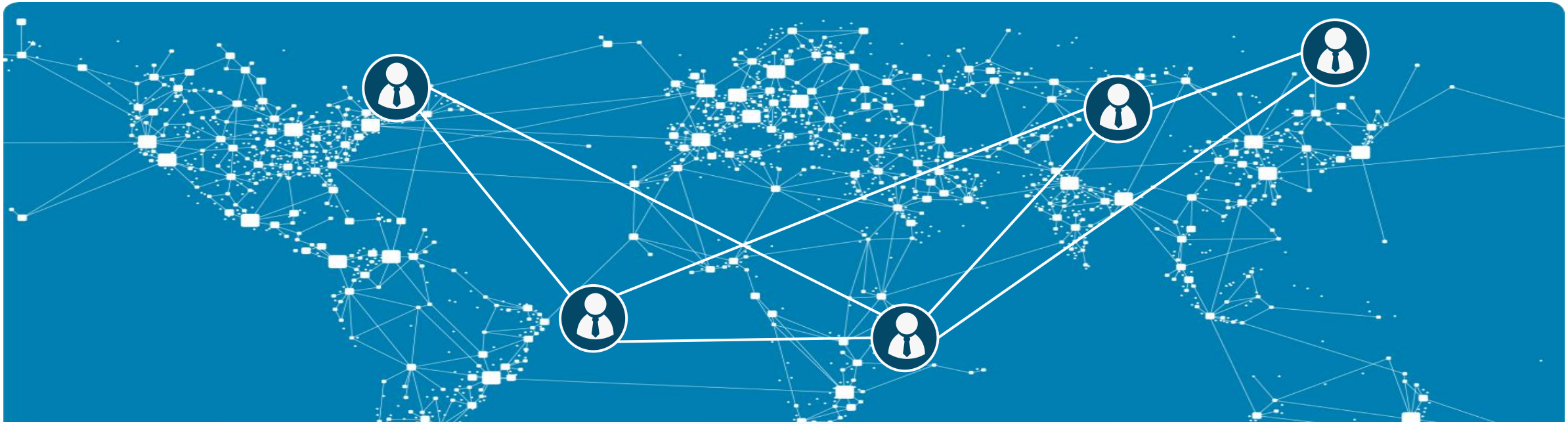
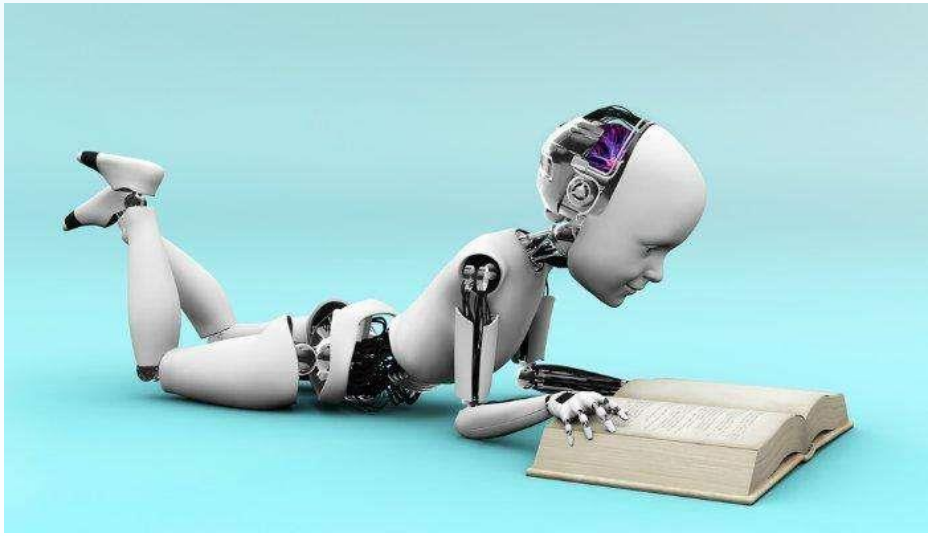
微生物组—扩增子16S分析研讨会第21期



32机器学习分类与回归

易生信

2024年4月14日



易生信，毕生缘；培训版权所有。

题 纲

- 什么是机器学习？
- 随机森林和Adaboost的基本思想
- 两篇Nature结果解读
- RandomForest、Adaboost分析实战

易生信 生信宝典 宏基因组

题 纲

- **什么是机器学习？**
- 随机森林和Adaboost的基本思想
- 两篇Nature结果解读
- RandomForest、Adaboost分析实战

易生信 生信宝典 宏基因组

机器学习

- 机器学习(Machine Learning，ML)是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样**模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。**

Nature近年来最高引的3篇文章全是深度学习

Deep learning.
Y LeCun, Y Bengio, G Hinton
Nature 521 (7553), 436-444
16750

Human-level control through deep reinforcement learning.
V Mnih, K Kavukcuoglu, D Silver, AA Rusu, J Veness, MG Bellemare, ...
Nature 518 (7540), 529-533
6101

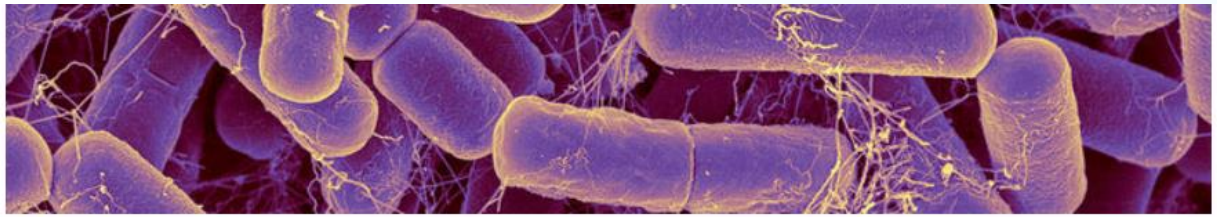
Mastering the game of Go with deep neural networks and tree search.
D Silver, A Huang, CJ Maddison, A Guez, L Sifre, J Schrittwieser, ...
Nature 529 (7587), 484-489
5212

Mastering the game of go without human knowledge
D Silver, J Schrittwieser, K Simonyan, I Antonoglou… - nature, 2017 - nature.com
A long-standing goal of artificial intelligence is an algorithm that learns, tabula rasa, superhuman proficiency in challenging domains. Recently, AlphaGo became the first program to defeat a world champion in the game of Go. The tree search in AlphaGo evaluated positions and selected moves using deep neural networks. These neural networks were trained by supervised learning from human expert moves, and by reinforcement learning from self-play. Here we introduce an algorithm based solely on
☆ Cited by 5077 Related articles All 37 versions

谷歌学术指数发布——看看综合、生物、生信、微生物领域高引文章和杂志

未来最热门的方法——深度学习

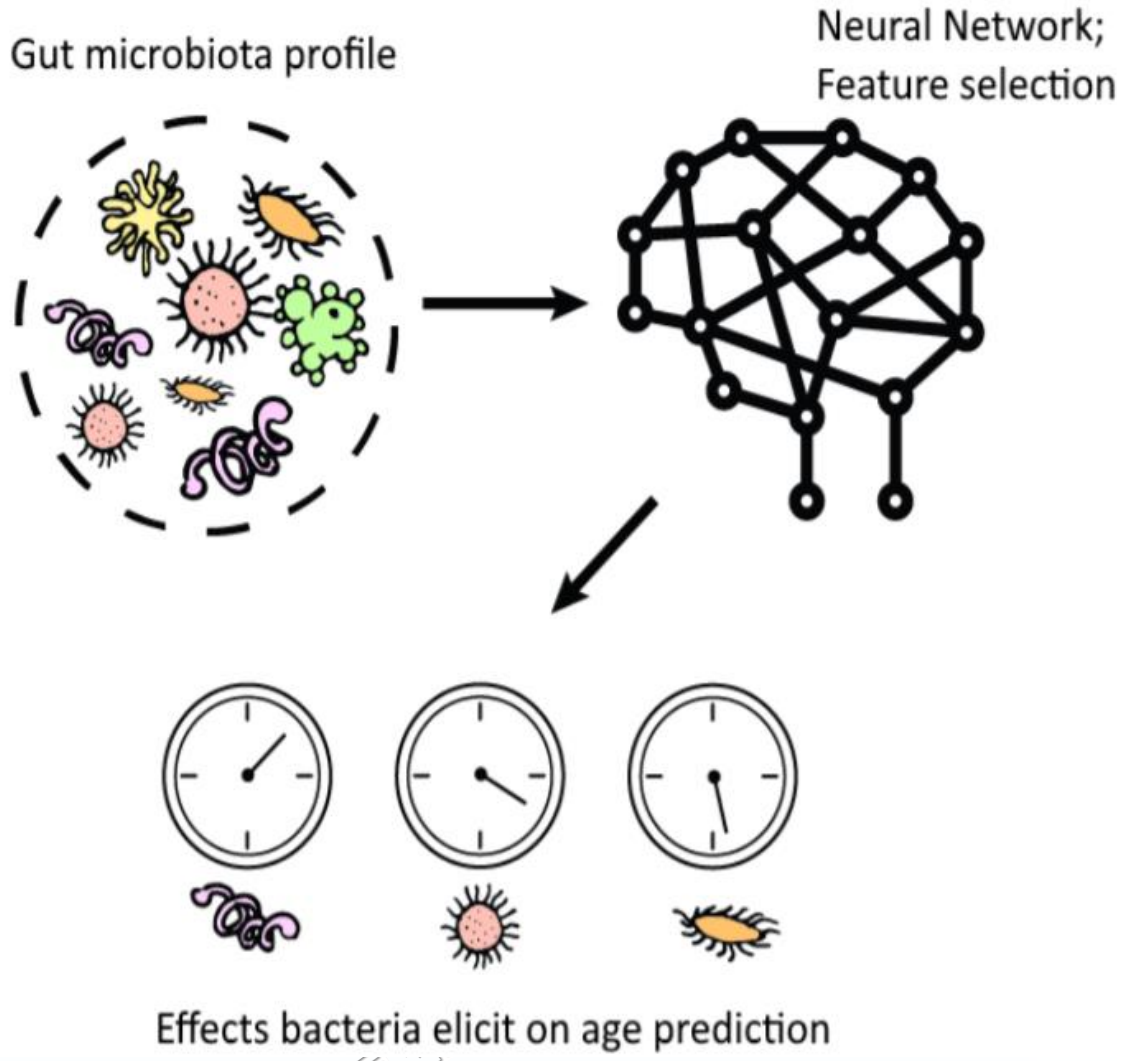
众所周知，目前最火的研究领域，莫过于**肠道菌群和深度学习**，而此项研究的作者，完美的将这两项研究结合，为我们带来一项非常有趣的研究。该研究通过**深度学习的算法分析肠道菌群组成来预测人类年龄，误差在4岁以内，说明菌群可作为生物标志物用于衰老相关的研究。**



Bacteroides are the most common bacteria species found in the human intestinal tract. DENNIS KUNKEL MICROSCOPY/SCIENCE SOURCE

The bacteria in your gut may reveal your true age

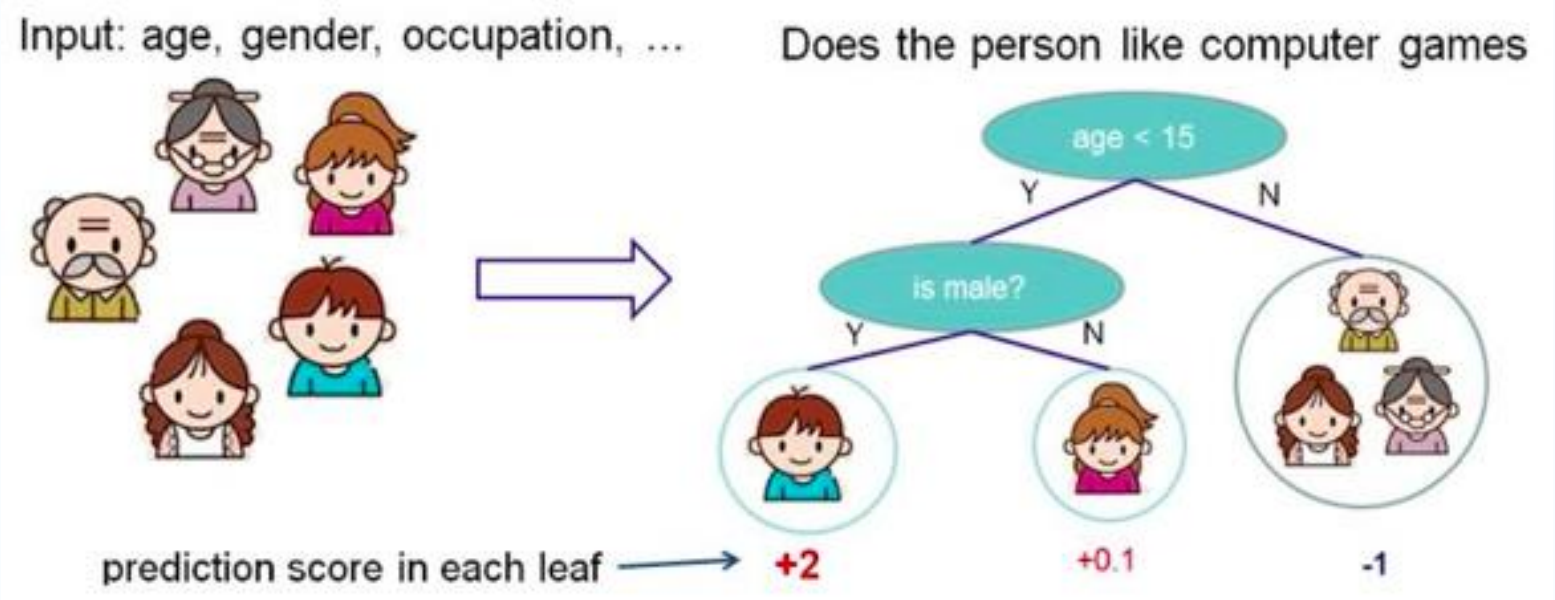
By Emily Mullin | Jan. 11, 2019 , 11:50 AM



Science评论：肠道菌群又添一大功能，揭示你的真实年龄 Fedor Galkin, Polina Mamoshina, Alex Aliper, Evgeny Putin, Vladimir Moskalev, Vadim N. Gladyshev, Alex Zhavoronkov. 2020. Human Gut Microbiome Aging Clock Based on Taxonomic Profiling and Deep Learning. ***iScience*** 23: 101199. https://doi.org/10.1016/j.isci.2020.101199 Nature Medicine同期8篇论文，聚焦人工智能在医学领域的应用

机器学习：十大常用算法

- **决策树、随机森林、逻辑回归、SVM、朴素贝叶斯、K最近邻、K均值、Adaboost 、神经网络、马尔可夫**



决策树：根据一些特征(feature)进行分类，每个节点提一个问题，通过判断，将数据分为两类，再继续提问。这些问题是根据已有数据学习出来的，再投入新数据的时候，就可以根据这棵树上的问题，将数据划分到合适的叶子上。

宏基因组

题 纲

- 什么是机器学习？
- **随机森林和Adaboost的基本思想**
- 两篇Nature结果解读
- RandomForest、Adaboost分析实战

易生信 生信宝典 宏基因组

随机森林

- **随机森林** 是一种**基于决策树的高效的机器学习算法**，可以用于**对样本进行分类或回归分析**。它属于非线性分类器，因此可以**挖掘变量之间复杂的非线性的相互依赖关系**。通过随机森林分析，可以找出能够区分两组样本间差异的关键成分(ASV/OTU/属/科等物种分类单元)。
- **优点：准确率极好**；高效地**分析大数据**集上；**处理高维特征**的输入样本；**缺省值结果很好**；**无需对它进行交叉验证**或者用一个独立的测试集来获得误差的一个无偏估计(它可以在内部进行评估，也就是说在生成的过程中就可以对误差建立一个无偏估计)。
- 注：高水平文章为什么还要二次、三次验证呢？

易生信

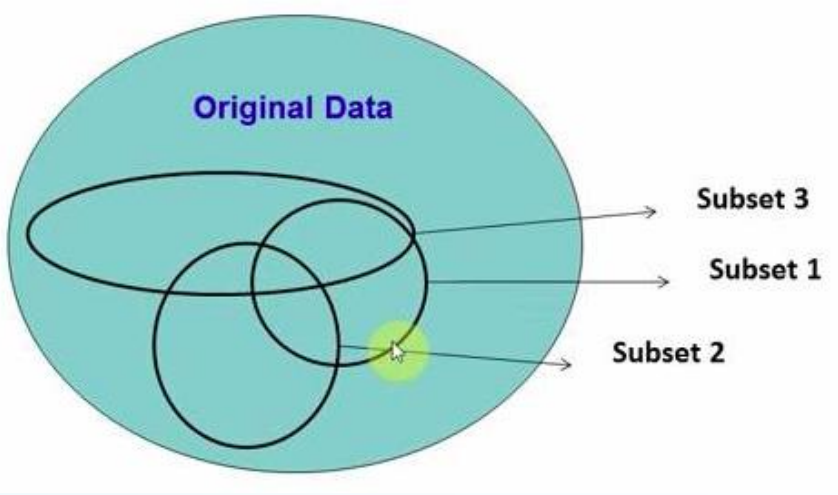
随机森林的原理

S矩阵是特征表，有1-N个样品，A B C是特征，最后一列C是分组信息

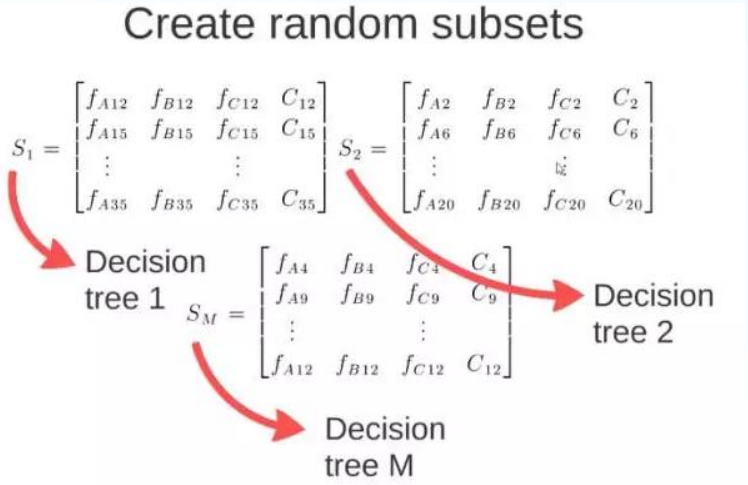
feature A of the 1st sample

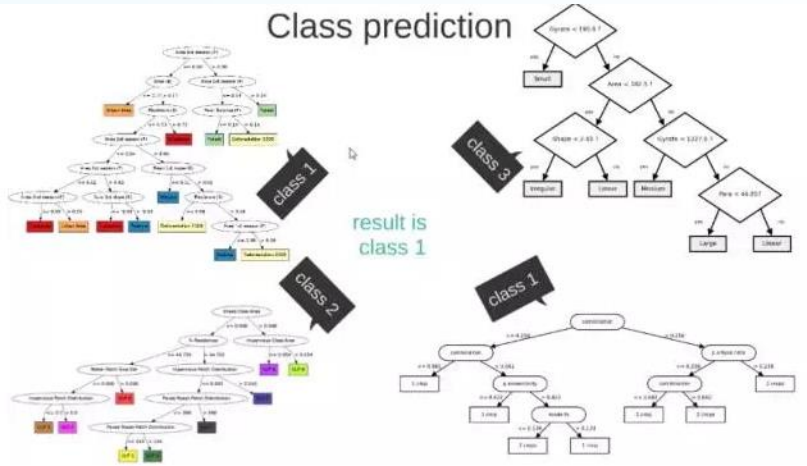
$$S = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & C_1 \\ \vdots & & \vdots & \\ f_{AN} & f_{BN} & f_{CN} & C_N \end{bmatrix}$$

在源数据中随机选取数据，组成M个子集



由 S 随机生成 M 个子矩阵



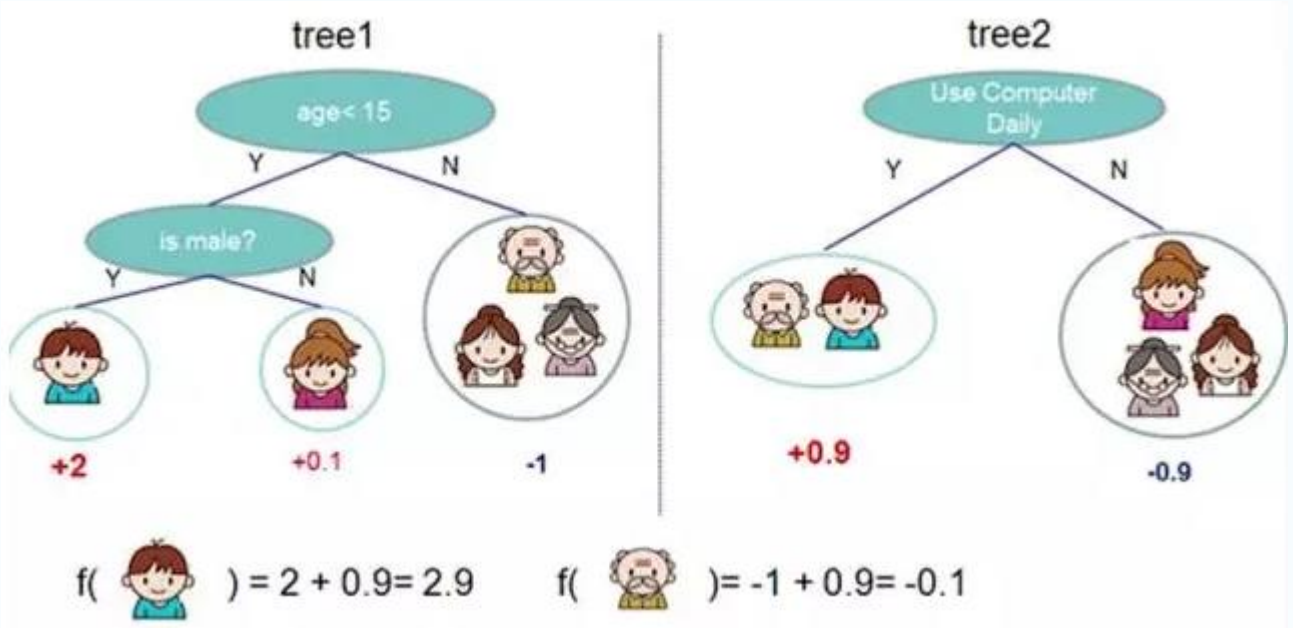


M 个子集得到 M 个决策树，将新数据投入得到 M 个分类结果

寻找哪些特征决策与已知分组一致率最高，即分组贡献重要特征

Adaboost ——迭代算法的一种

- 核心思想是针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器（强分类器）。



左右两个决策树，单个看是效果不怎么好的，但是把同样的数据投入进去，
把两个结果加起来考虑，就会增加可信度

易生信宝典 宏基因组

题 纲

- 什么是机器学习？
- 随机森林和Adaboost的基本思想
- **两篇Nature结果解读**
- RandomForest、Adaboost分析实战

易生信 生信宝典 宏基因组

doi:10.1038/nature24628

Salt-responsive gut commensal modulates T_H17 axis and disease

Nicola Wilck[1,2,3,4,5], Mariana G. Matus[6,7], Sean M. Kearney[6], Scott W. Olesen[6], Kristoffer Forslund[8], Hendrik Bartolomaeus[1,2,3,4], Stefanie Haase[9], Anja Mähler[1,5], András Balogh[1,2,3,4,5], Lajos Markó[1,2,3,4,5], Olga Vvedenskaya[3,10,11], Friedrich H. Kleiner[1], Dmitry Tsvetkov[1,2], Lars Klug[1,5], Paul I. Costea[8], Shinichi Sunagawa[8,12], Lisa Maier[13], Natalia Rakova[1,9], Valentin Schatz[14], Patrick Neubert[14], Christian Frätzer[15], Alexander Krannich[5], Maik Gollasch[1,2,3], Diana A. Grohme[16], Beatriz F. Côrte-Real[17], Roman G. Gerlach[18], Marijana Basic[19], Athanasios Typas[13], Chuan Wu[20], Jens M. Titze[21], Jonathan Jantsch[14], Michael Boschmann[1,5], Ralf Dechend[1,2,5], Markus Kleinewietfeld[16,17,22], Stefan Kempa[3,5,10], Peer Bork[3,8,23,24], Ralf A. Linker[9]§, Eric J. Alm[6]§ & Dominik N. Müller[1,2,3,4,5]§

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

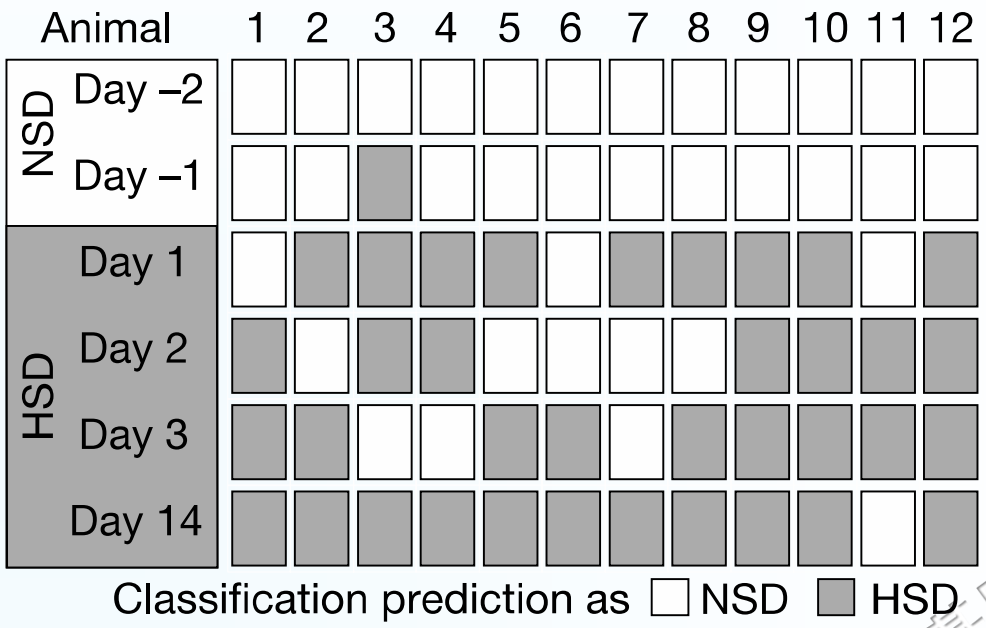
Nature：如何做一篇肠道菌群免疫的顶级文章

预测人高盐/低盐组生物标记(biomarker)+动物验证

a

OTU	Classifier importance (%)	Max. relative abundance (%)
Lactobacillus	25.0	1.24
Prevotellaceae	20.1	0.04
Pseudoflavonifractor	12.5	0.85
Clostridia	12.5	0.58
Parasuterella	9.7	7.33
Akkermansia	9.3	4.22
Bacteroidetes	6.0	0.37
Alistipes	4.9	19.5

b



- 随机森林和Adabooster分类，Adabooster结果重要性importance值更大，且对动物验证数据结果准确性也很高
- 软件：使用Python3的slime2计算决定分组特征的贡献率/重要性

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

Slime2主页：https://github.com/swo/slime2
原文分析实例：https://github.com/almlab/analysis-salt-responsive/

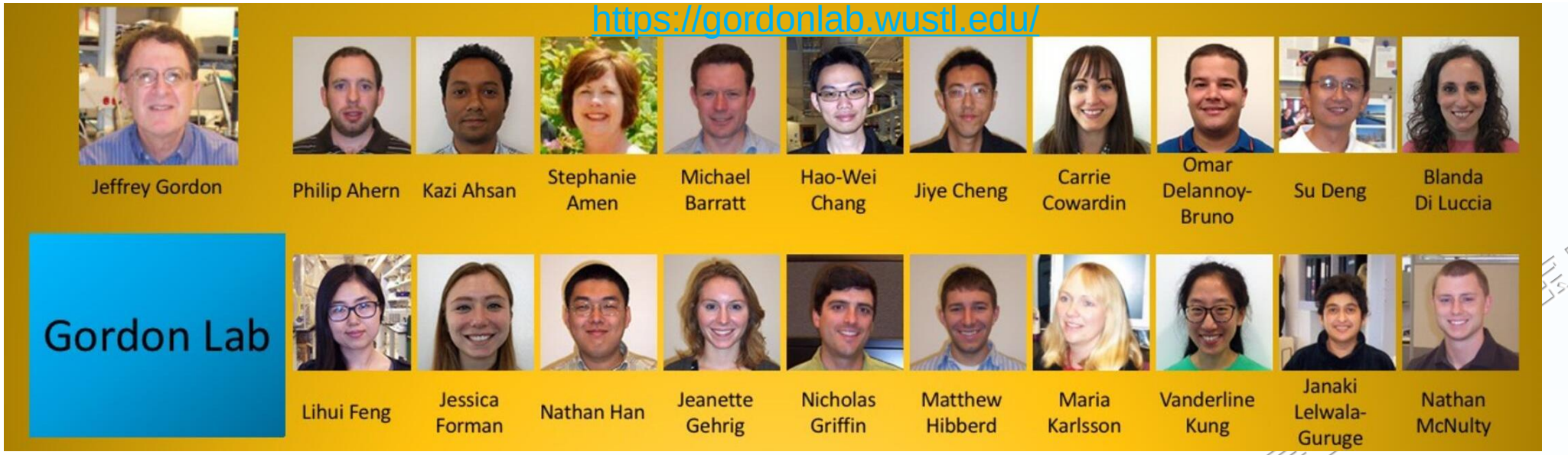
LETTER

doi:10.1038/nature13421

Persistent gut microbiota immaturity in malnourished Bangladeshi children

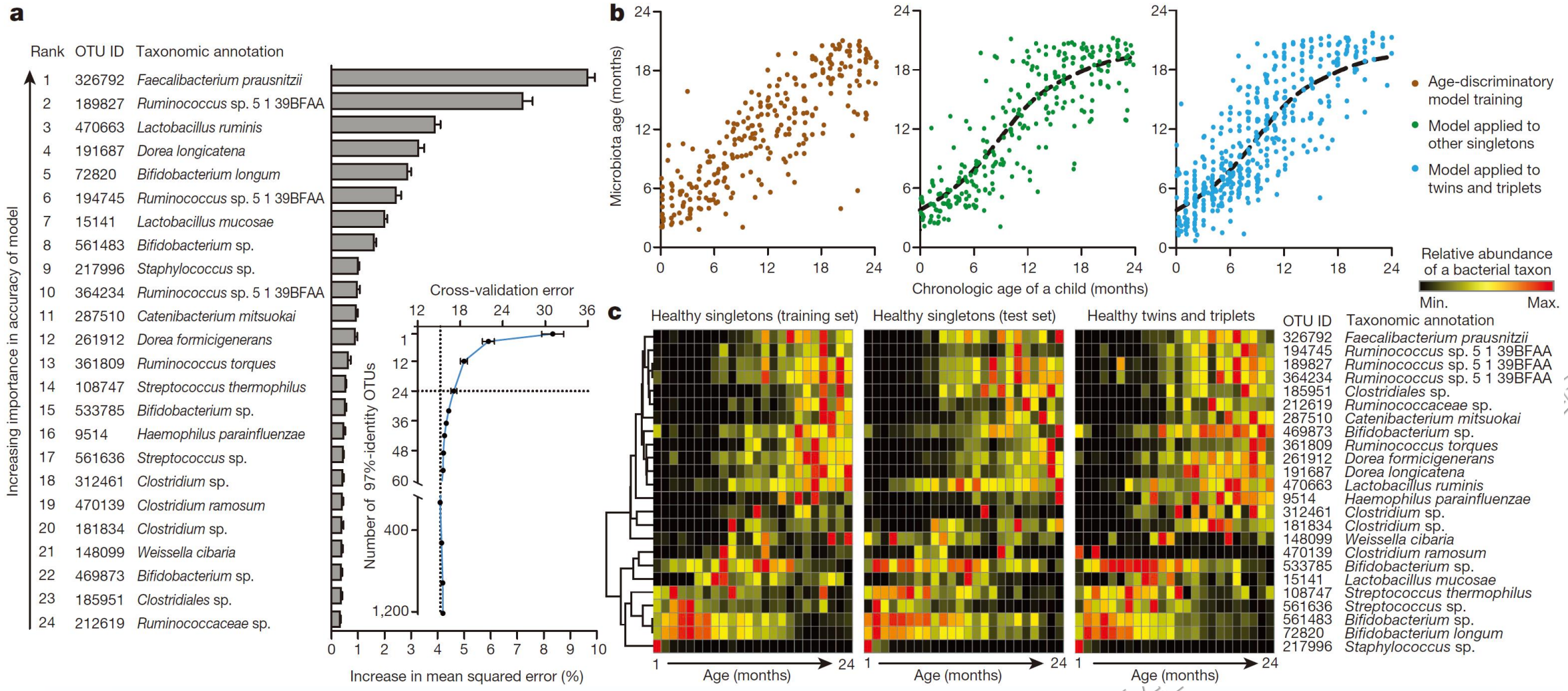
Sathish Subramanian[1], Sayeeda Huq[2], Tanya Yatsunenko[1], Rashidul Haque[2], Mustafa Mahfuz[2], Mohammed A. Alam[2], Amber Benezra[1,3], Joseph DeStefano[1], Martin F. Meier[1], Brian D. Muegge[1], Michael J. Barratt[1], Laura G. VanArendonk[1], Qunyuan Zhang[4], Michael A. Province[4], William A. Petri Jr[5], Tahmeed Ahmed[2] & Jeffrey I. Gordon[1]

https://gordonlab.wustl.edu/



下次诺贝尔奖会是他吗？肠道微生物组领域开创者Jeffrey Gordon
Nature: 甘露糖苷选择性抑制致病性大肠杆菌
科普纪录片 ： 肠道菌群在人体中的作用

定义健康人1-24月菌群生物标记(Bacterial taxonomic biomarkers)



Subramanian et al. 2014. Persistent gut microbiota immaturity in malnourished Bangladeshi children. ***Nature*** 510: 417. https://doi.org/10.1038/nature13421

题 纲

- 什么是机器学习？
- 随机森林和Adaboost的基本思想
- 两篇Nature结果解读
- **RandomForest、Adaboost分析实战**

易生信 生信宝典 宏基因组

扩展学习：Slime2鉴定分组贡献特征

易生信

使用adabooster计算10000次(16.7s)，推荐千万次

slime2.py otutab.txt design.txt --normalize --tag ab_e4 ab -n 10000

```
Root;Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus      0.2331
Root;Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae                0.1661
Root;Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Pseudoflavonifractor       0.1154
Root;Bacteria;Bacteroidetes          0.1071
Root;Bacteria;Firmicutes;Clostridia0.0859
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Verrucomicrobiaceae;Akkermansia 0.0701
```

使用RandomForest计算10000次(14.5s)，推荐百万次，支持多线程

slime2.py otutab.txt design.txt --normalize --tag rf_e4 rf -n 10000

```
Root;Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus      0.08616193880686196
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Sutterellaceae;Parasutterella              0.07444763171040213
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Verrucomicrobiaceae;Akkermansia 0.06552857763500938
Root;Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Pseudoflavonifractor       0.04065535850958773
Root;Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Oscillibacter  0.031399910016908436
Root;Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae       0.02733110442598093
```

分析需要OTU表和实验设计(分组信息)两个文件(演示运行)

https://github.com/swo/slime2

•Nature：如何做一篇肠道菌群免疫的顶级文章

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

宏基因组 易生信 生信宝典

R语言中的RandomForest包

[PDF] Classification and regression by randomForest

A Liaw, M Wiener - R news, 2002 - journal.r-project.org

Recently there has been a lot of interest in “ensemble learning”—methods that generate many classifiers and aggregate their results. Two well-known methods are boosting (see, eg, Shapire et al., 1998) and bagging Breiman (1996) of classification trees. In boosting, successive trees give extra weight to points incorrectly predicted by earlier predictors. In the end, a weighted vote is taken for prediction. In bagging, successive trees do not depend on earlier trees—each is independently constructed using a bootstrap sample of the data set. In …

☆ Save 99 Cite Cited by 24069 Related articles All 11 versions »

- 可以实现基于随机森林的分类(Classification)和回归(Regression)
- # 安装

  install.packages("randomForest")
- 查看帮助

  ?randomForest

易生信 生信宝典 宏基因组

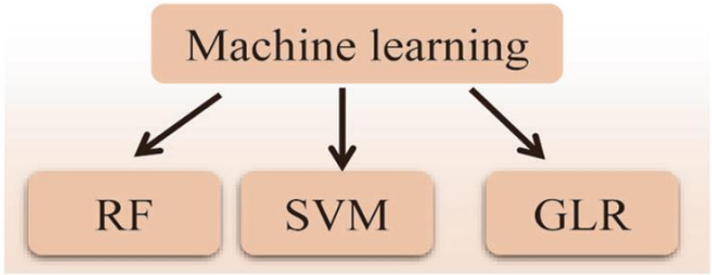
随机森林分类预测枯萎病的发生

A

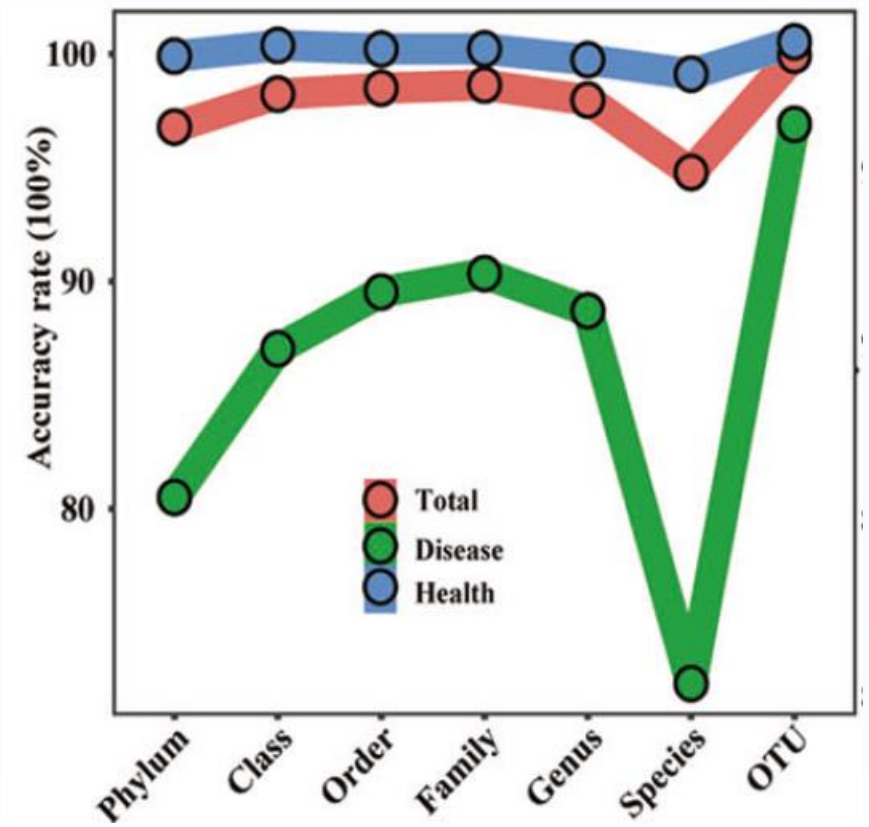
Bacterial	
37 studies	
Disease	Health
148	957

Fungal	
26 studies	
Disease	Health
143	301

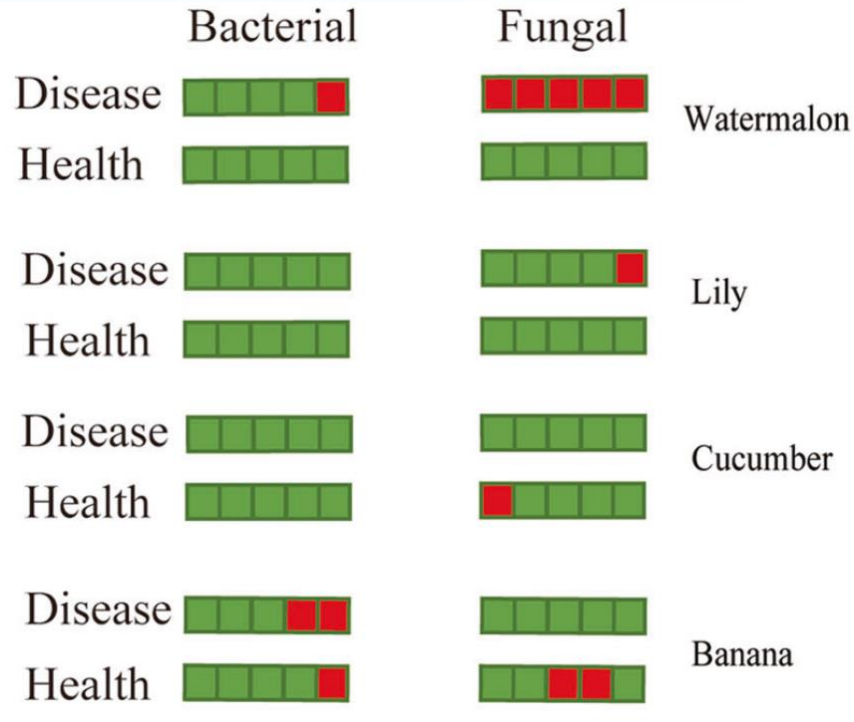
B



随机森林（RF）、支持向量机（SVM）和逻辑回归（LR）
数据和方法选择



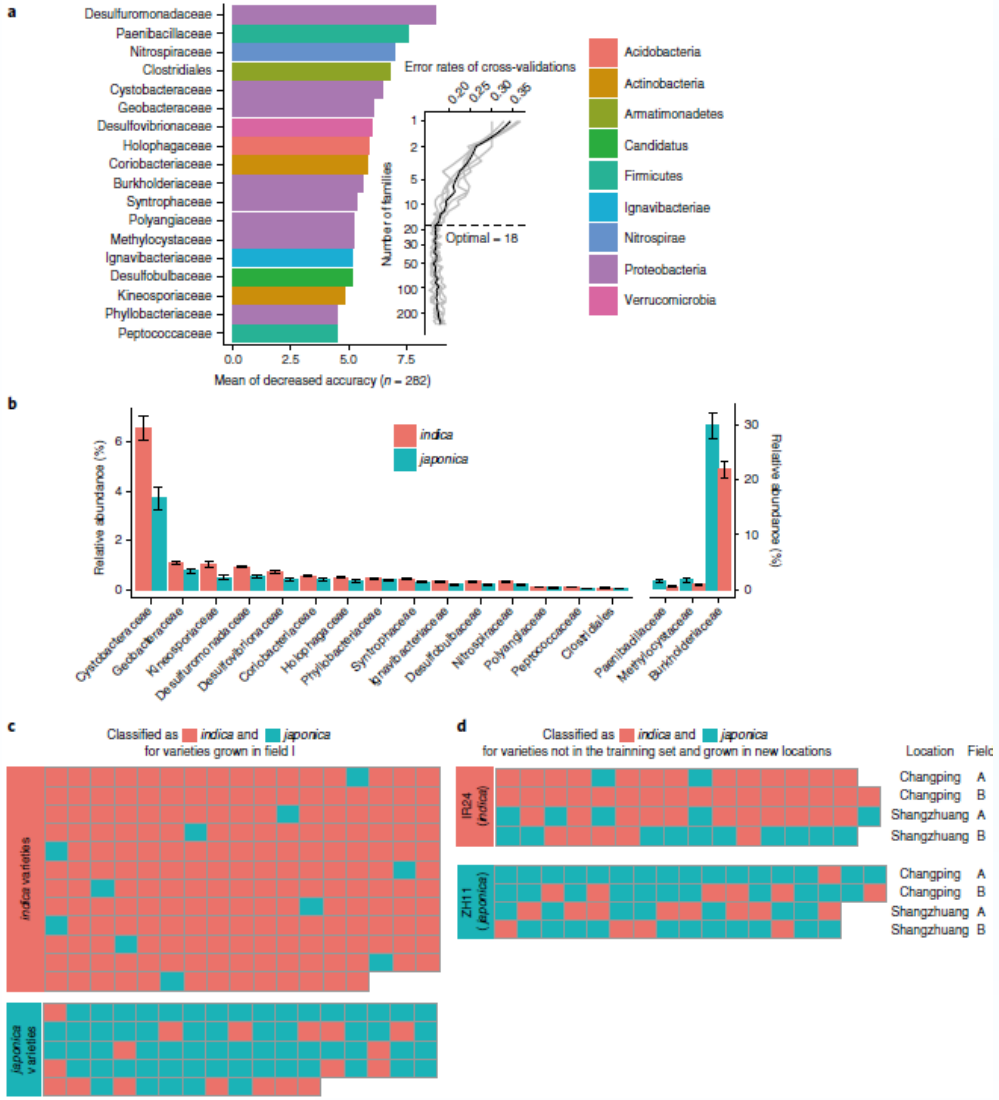
层级选择——分类预测评估准确度

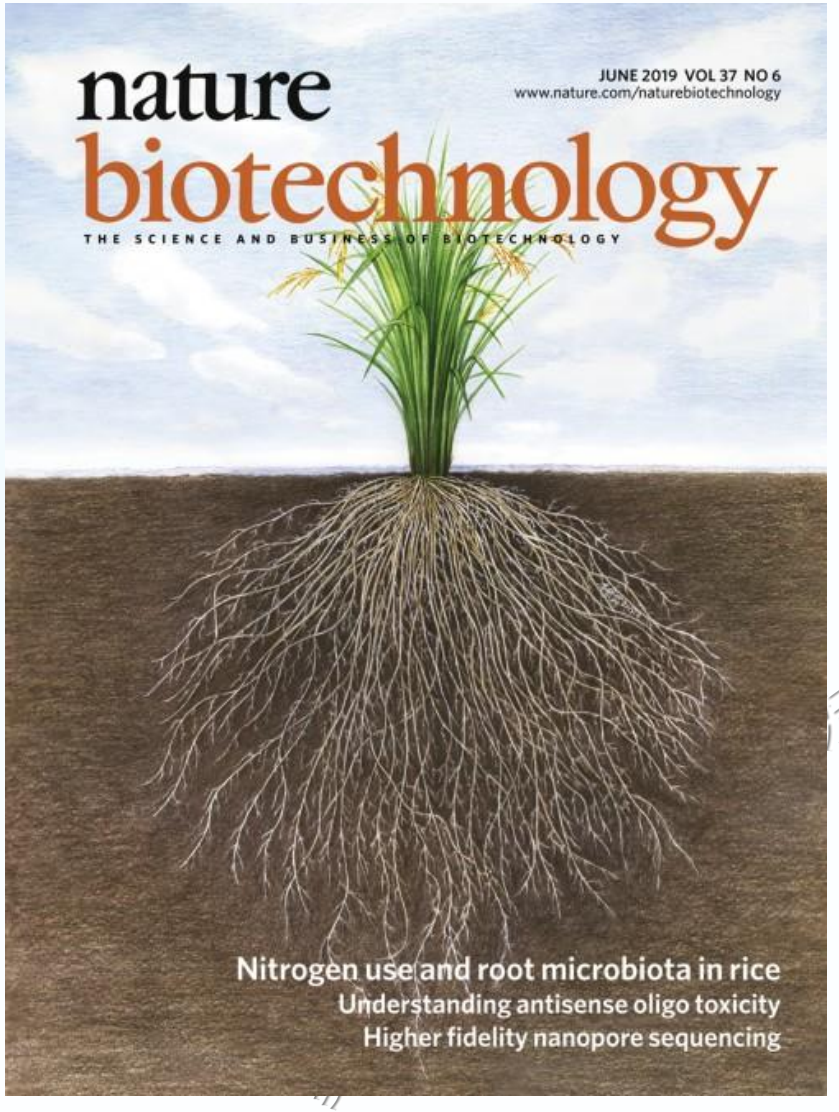


验证模型的有效性

ISME：南农沈其荣团队基于大数据准确预测土壤的枯萎病发生 纯生信发ISME
https://github.com/taowenmicro/Wen-etal-200214-paper-code
南京袁军/文涛ISME：基于大数据整合准确预测土壤的枯萎病发生(视频)

随机森林分类实战：NBT基于微生物预测水稻亚种



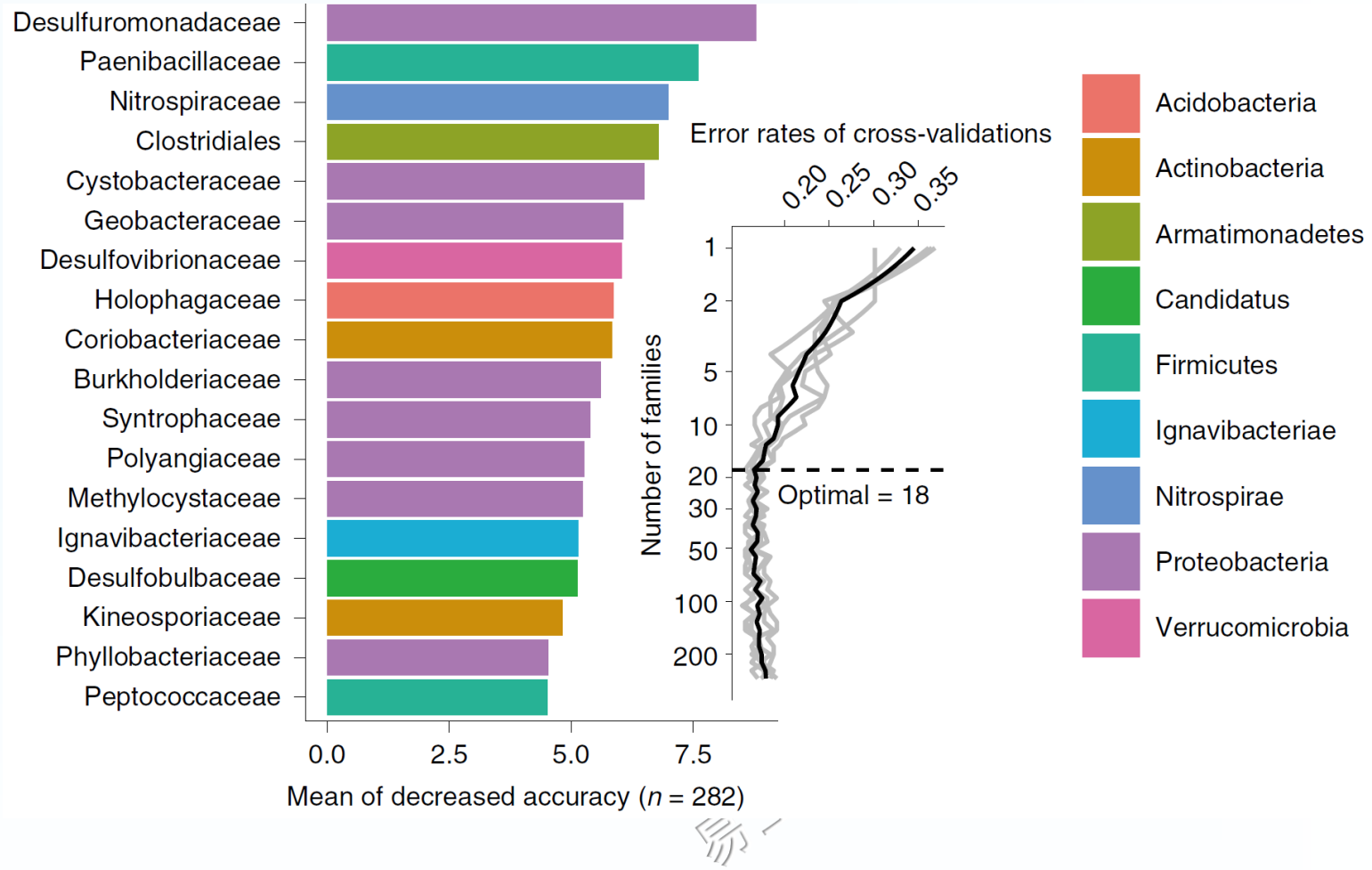


- NBT封面：水稻NRT1.1B基因调控根系微生物组参与氮利用
- R语言大会：宏基因组数据分析和可视化套路总结(视频)

挖掘重要分类特征

randomForest包对样本离散型进行分类，连续型进行回归

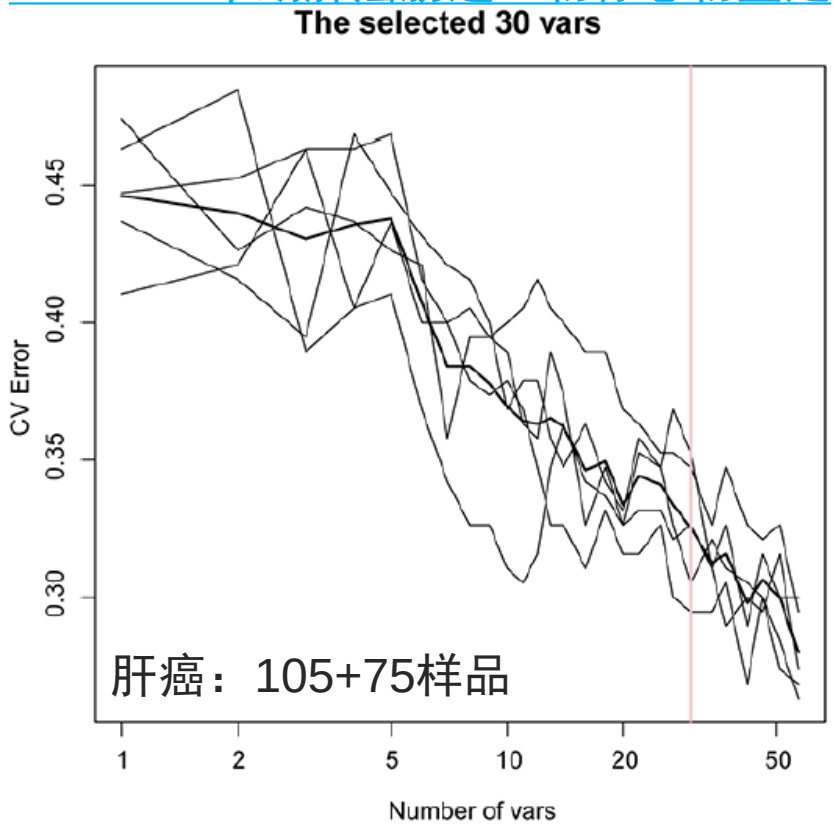
1. randomForest建模和导出特征贡献度importance，ggplot2柱状图可视化
2. 多次交叉验证，人为选择合适的特征组合，ggplot2折线图可视化



Jingying Zhang, **Yong-Xin Liu**, et. al. ***Nature Biotechnology***. 2019. Fig 2a

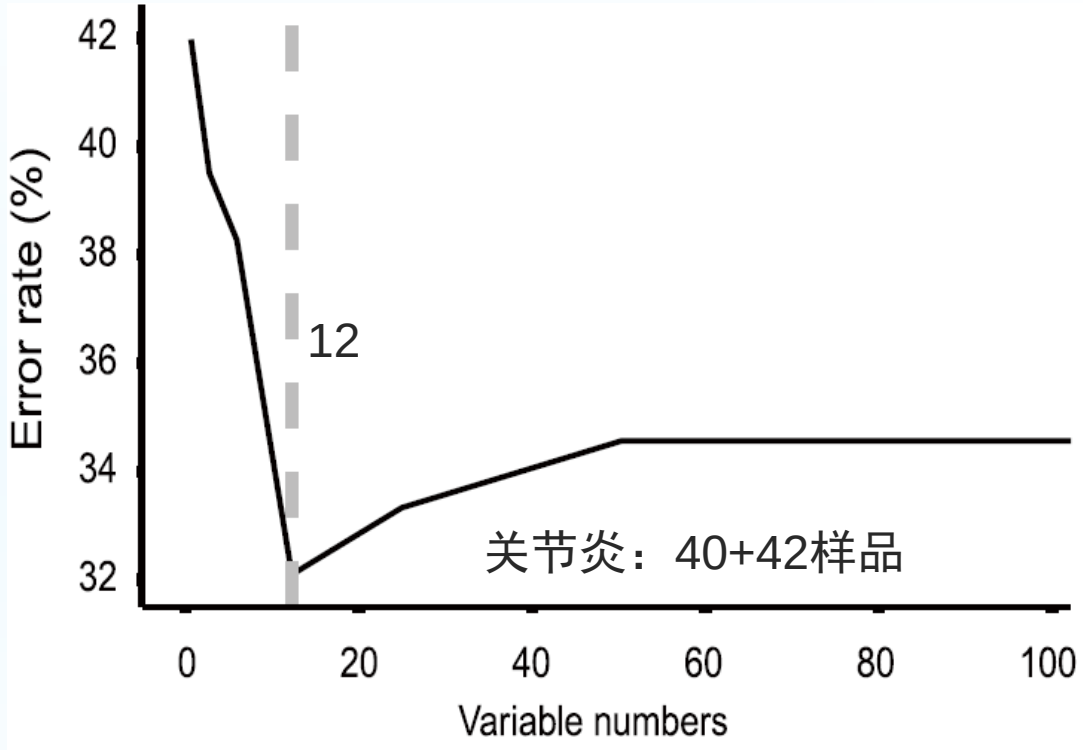
寻找生物标志物：交叉验证确定较优数量

Gut-2018-早期肝癌肠道生物标志物鉴定



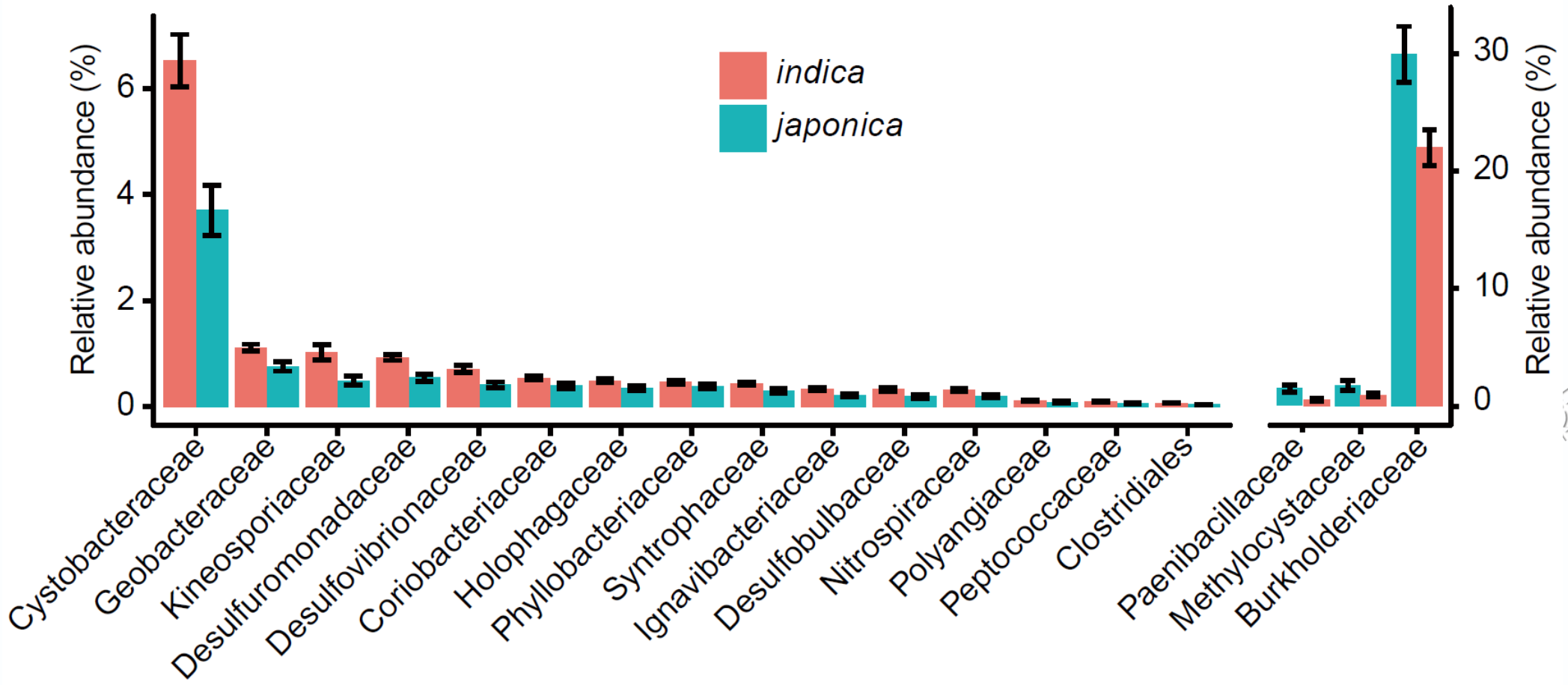
Zhigang Ren, Ang Li, Jianwen Jiang, Lin Zhou, Zujiang Yu, Haifeng Lu, Haiyang Xie, Xiaolong Chen, Li Shao, Ruiqing Zhang, Shaoyan Xu, Hua Zhang, Guangying Cui, Xinhua Chen, Ranran Sun, Hao Wen, Jan P. Lerut, Quancheng Kan, Lanjuan Li & Shusen Zheng. (**2019**). Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early hepatocellular carcinoma. ***Gut*** 68, 1014-1023, doi: https://doi.org/10.1136/gutjnl-2017-315084

BMC：幼儿关节炎患儿肠道菌群的特征



Xubo Qian†, **Yong-Xin Liu†**, Xiaohong Ye, Wenjie Zheng, Shaoxia Lv, Miaojun Mo, Jinjing Lin, Wenqin Wang, Weihan Wang, Xianning Zhang & Meiping Lu. (**2020**). Gut microbiota in children with juvenile idiopathic arthritis: characteristics, biomarker identification, and usefulness in clinical prediction. ***BMC Genomics*** 21, 286, doi: https://doi.org/10.1186/s12864-020-6703-0

重要分类特征在组间差异

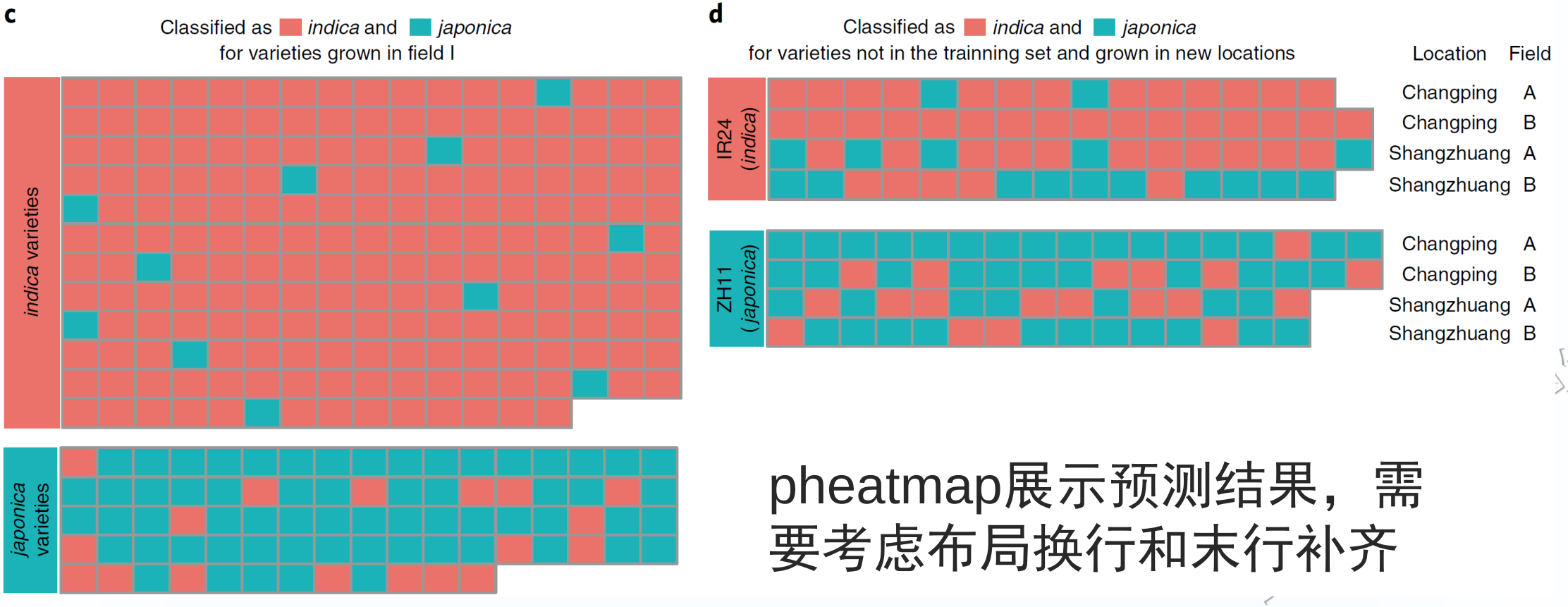


柱状图+标准误展示特征丰度的组间差异，分类再排序

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

Jingying Zhang, **Yong-Xin Liu**, et. al. ***Nature Biotechnology***. 2019. Fig 2b

模型普适性评估——预测样本分类



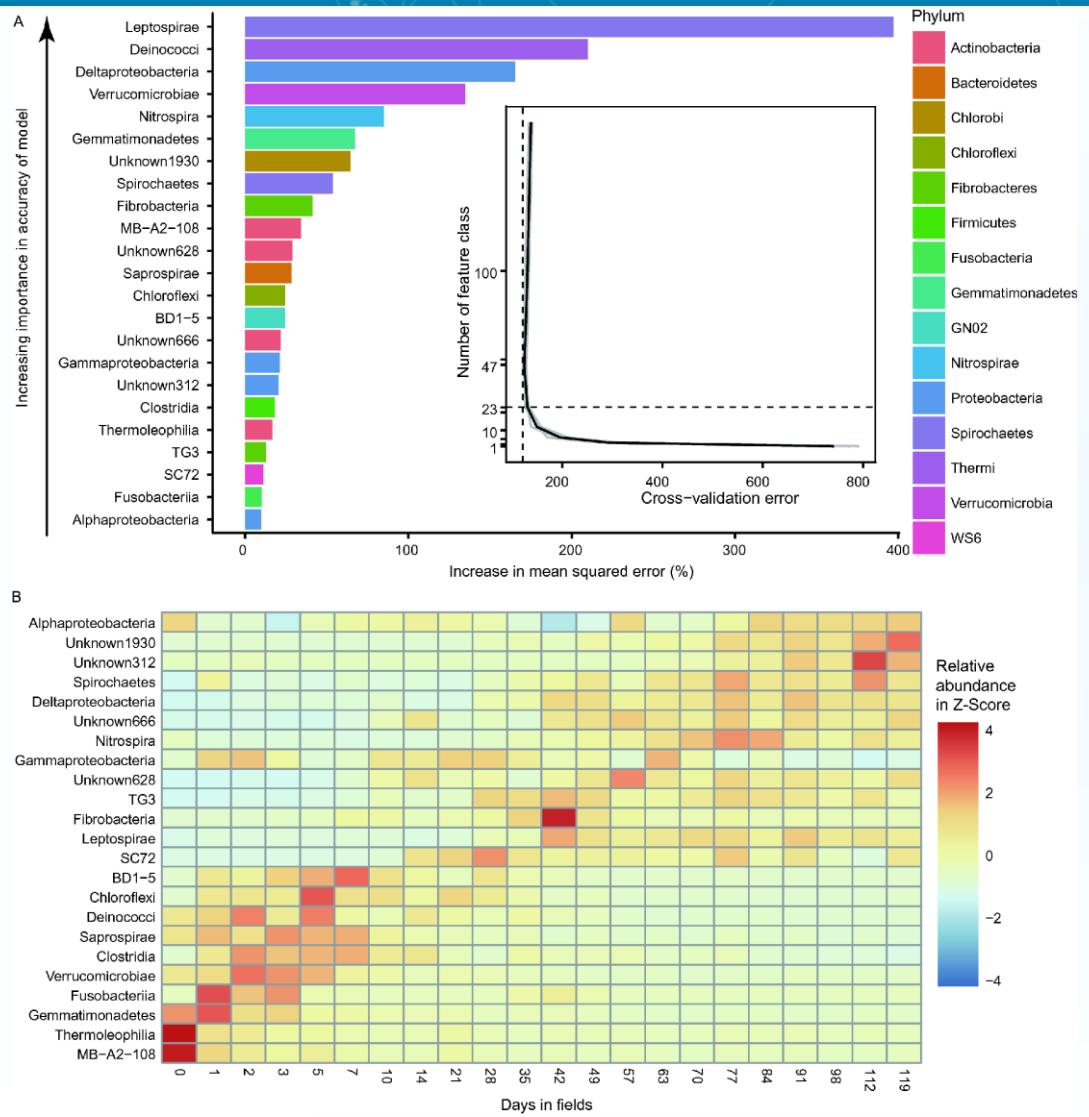
pheatmap展示预测结果，需要考虑布局换行和末行补齐

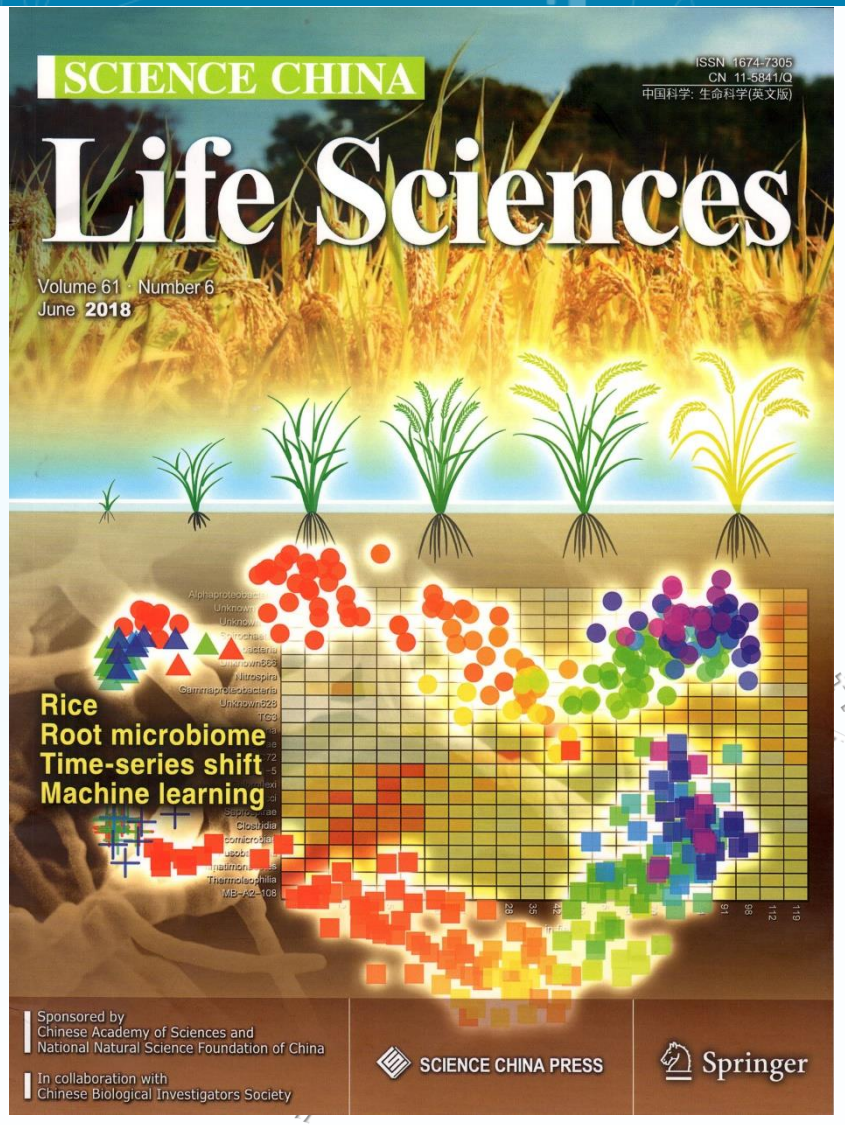
模型对另一块地相同品种(c)和不同地点不同品种(d)预测

操作详见：RF_classification目录

Jingying Zhang, **Yong-Xin Liu**, et. al. ***Nature Biotechnology***. 2019. Fig 2c/d

R包RandomForest包分析回归实战





易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

•手把手带你重现菌群封面文章图表

开始前准备

- randomForest包
- 实验设计：design.txt

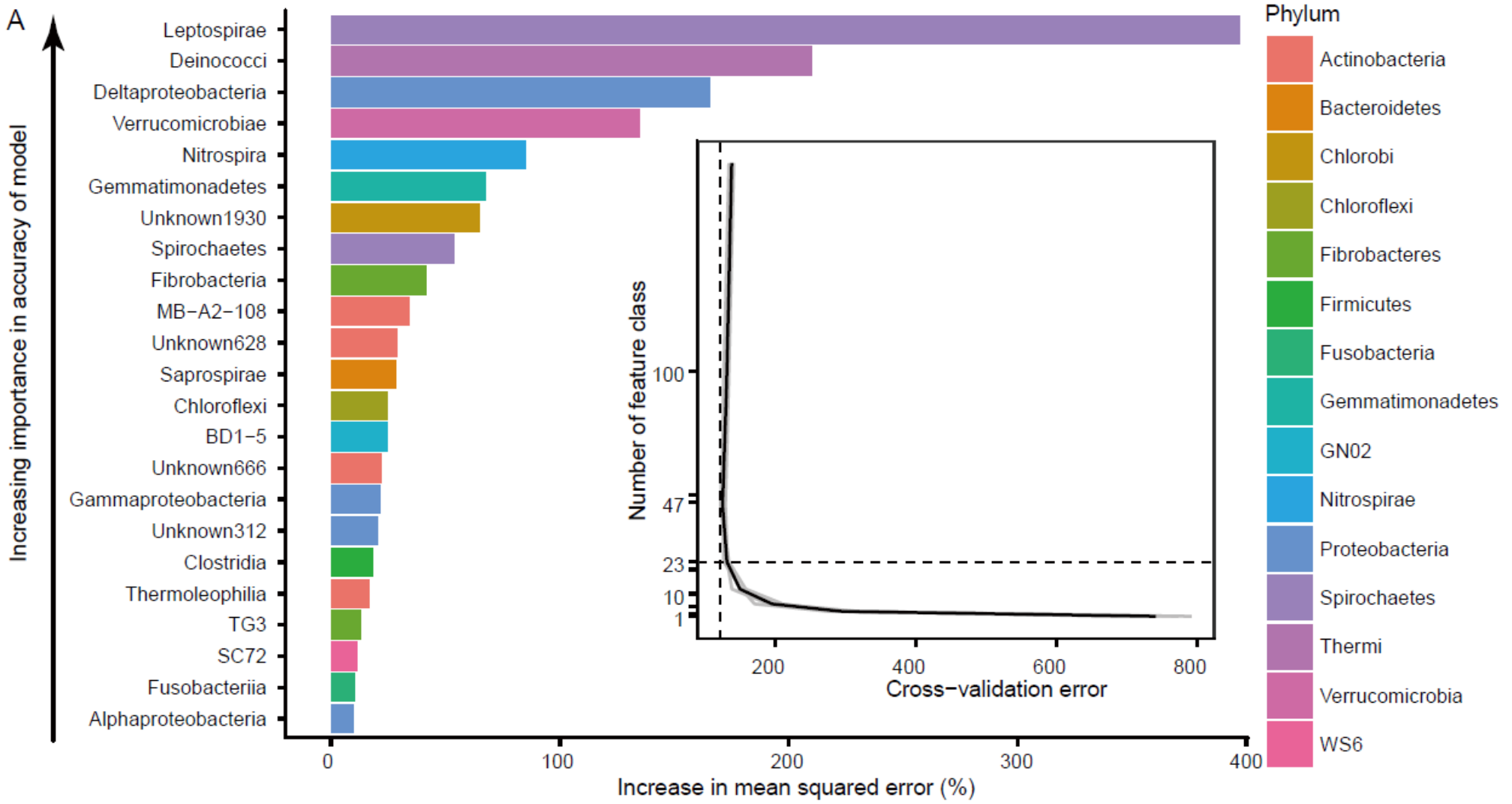
SampleID	groupID	genotype	site	day	replicate	compartment
A50Cp0r1	A50Cp0	A50	Cp	0	1	root
A50Cp0r2	A50Cp0	A50	Cp	0	2	root
A50Cp0r3	A50Cp0	A50	Cp	0	3	root
IR24Cp0r1	IR24Cp0	IR24	Cp	0	1	root

- OTU表/分类级：otu_table_tax_L3.txt

ID	A50Cp10r2	A50Cp3r3	IR24Cp2r1
k__Bacteria;p__Proteobacteria;c__Betaproteobacteria	0.29122	0.396278	0.316639
k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria	0.204275	0.162716	0.158932
k__Bacteria;p__Actinobacteria;c__Actinobacteria	0.218686	0.04667	0.12337
k__Bacteria;p__Proteobacteria;c__Deltaproteobacteria	0.035686	0.04625	0.046986
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria	0.078237	0.159094	0.211824

Jingying, Zhang(#); Na, Zhang(#); **Yong-Xin, Liu**(#); Xiaoning, Zhang; Bin, Hu; Yuan, Qin; Haoran, Xu; Hui, Wang; Xiaoxuan, Guo; Jingmei, Qian; Wei, Wang; Pengfan, Zhang; Tao, Jin; Chengcai, Chu(*); Yang, Bai(*), Root microbiota shift in rice correlates with resident time in the field and developmental stage[J]. ***Science China Life Sciences***, 2018: 1-9.

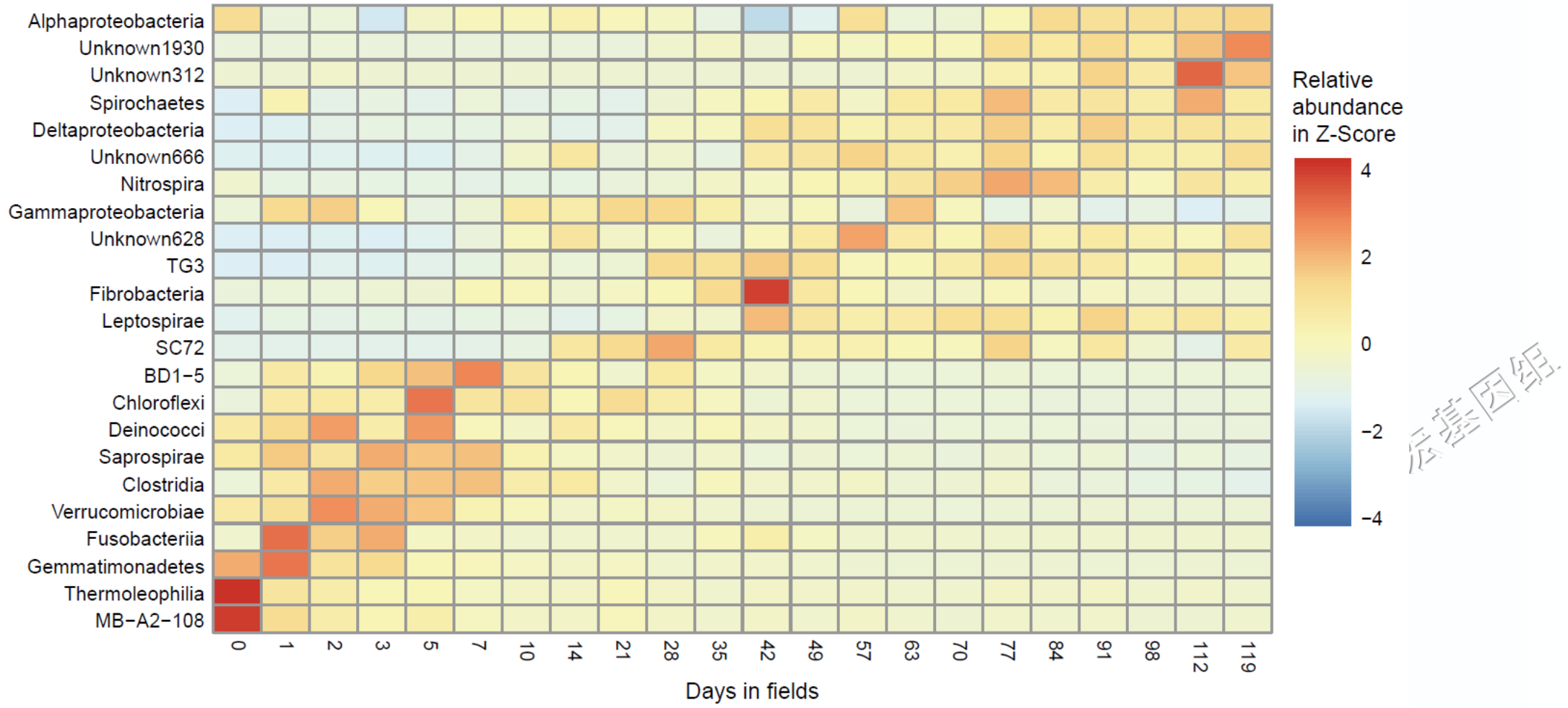
水稻年龄(时间序列)相关Biomarkers挖掘



宏基因组

Jingying, Zhang(#); Na, Zhang(#); **Yong-Xin, Liu**(#); Xiaoning, Zhang; Bin, Hu; Yuan, Qin; Haoran, Xu; Hui, Wang; Xiaoxuan, Guo; Jingmei, Qian; Wei, Wang; Pengfan, Zhang; Tao, Jin; Chengcai, Chu(*); Yang, Bai(*), Root microbiota shift in rice correlates with resident time in the field and developmental stage[J]. ***Science China Life Sciences***, 2018: 1-9.

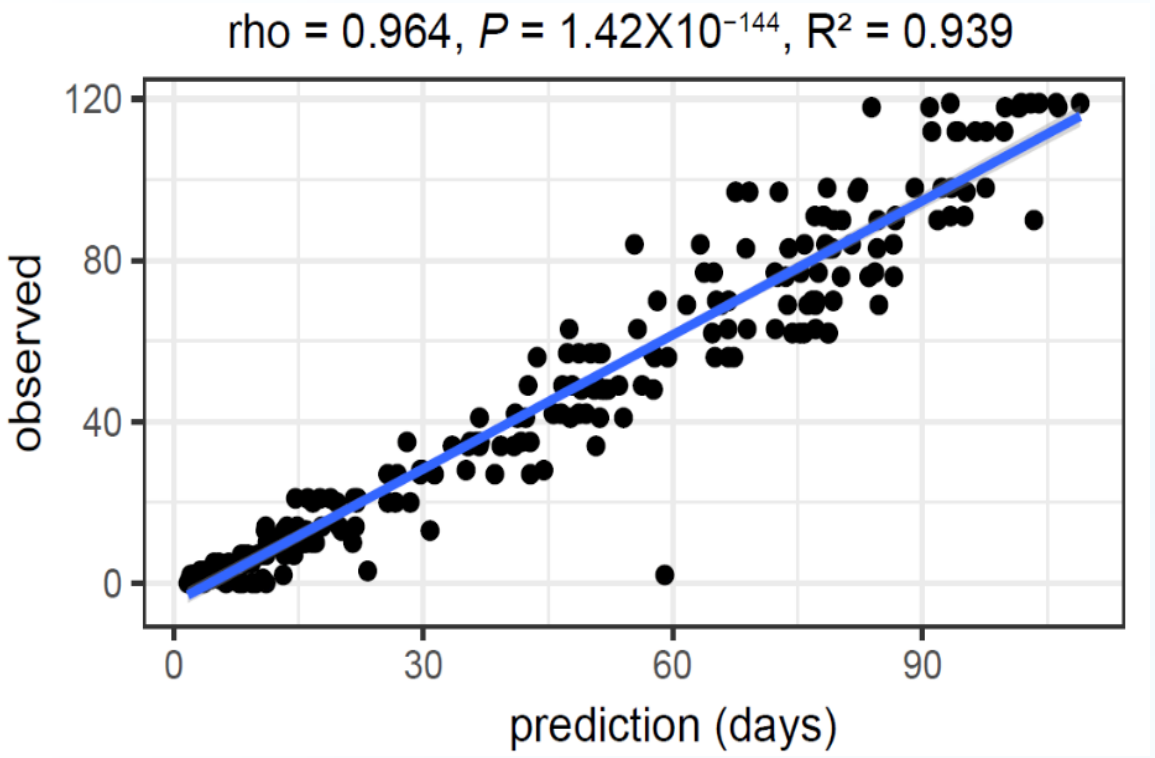
生物标志物在时间梯度上分布特征



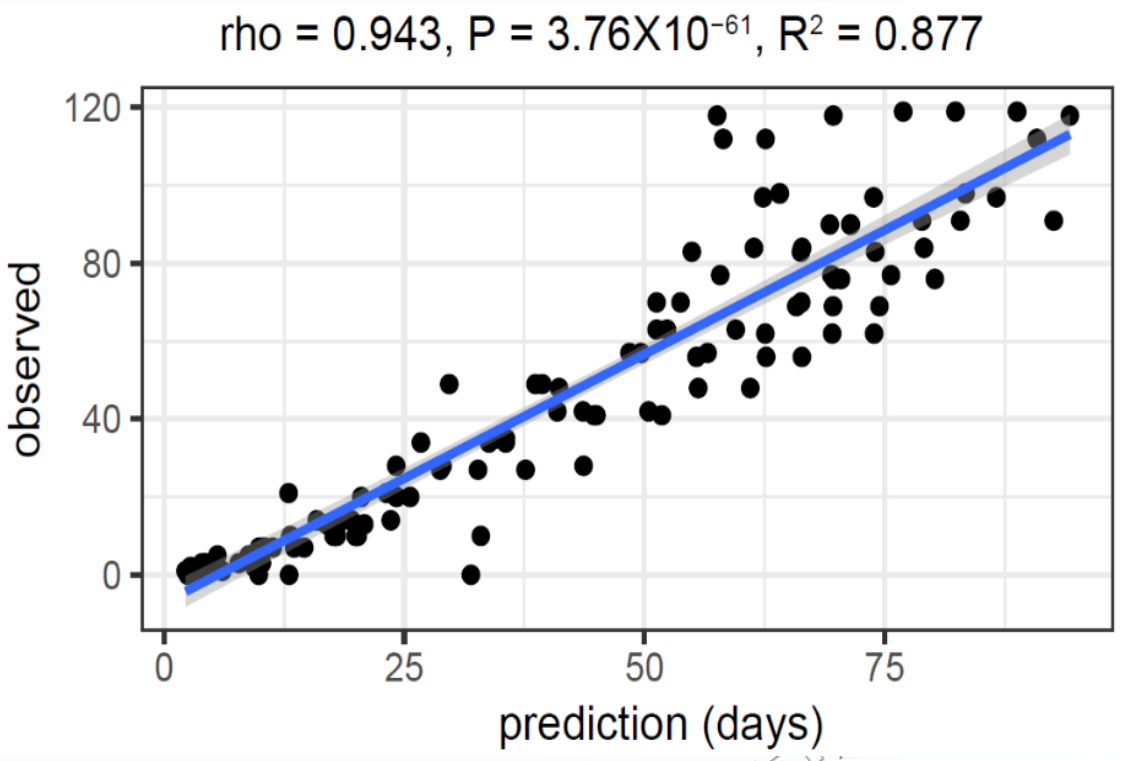
易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

Liu et al. 2023. EasyAmplicon: An easy-to-use, open-source, reproducible, and community-based pipeline for amplicon data analysis in microbiome research. *iMeta* 2: e83. https://doi.org/10.1002/imt2.83

基于细菌标志物模型预测水稻生育时期



训练集模型预测相关性



测试集模型预测相关性

Liu et al. 2023. EasyAmplicon: An easy-to-use, open-source, reproducible, and community-based pipeline for amplicon data analysis in microbiome research. *iMeta* 2: e83. https://doi.org/10.1002/imt2.83

易生信，毕生缘；培训版权所有。

总结

- 机器学习领域是当前计算方法中热门的方法，其中深度学习将成本未来的研究和应用热点；
- 本领域目前已经开始应用深度学习，但使用最广泛的仍然是随机森林，效果不佳时可尝试Adaboost；
- Slime2可以实现随机森林和Adaboost分析，可在Linux下运行；
- R语言的RandomForest包引用超万次，可以Win/Mac/Linux跨平台分析，可实现随机森林分类与回归，同时结果R语言的统计和绘图完成全套的分析、统计与可视化——推荐。
- QIIME 2：15样品分类和回归q2-sample-classifier

Liu et al. 2023. EasyAmplicon: An easy-to-use, open-source, reproducible, and community-based pipeline for amplicon data analysis in microbiome research. *iMeta* 2: e83. https://doi.org/10.1002/imt2.83

扫码关注生信宝典，学习更多生信知识

扫码关注宏基因组，获取专业学习资料

易生信，没有难学的生信知识