机器学习及疾病分型

Machine Learning and its Application to Disease Classification

www.genechem.com.cn

郝慧渊

2022-5-11

上海吉凯基因医学科技股份有限公司
SHANGHAI GENECHEM CO.,LTD.

GENE 吉凯基因

蛋白基因组与机器学习助力癌症研究迈入精准医疗时代

郝慧渊 瓦赫宁根大学生物信息学硕士

吉凯基因生信售前顾问

硕士学习期间深入掌握生物统计学、生物信息学和机器学习方法，致力于基于in silico方法的微生物特征和功能基因的发现。曾参与基于云计算的病原体感染诊断全自动化流程的研发。

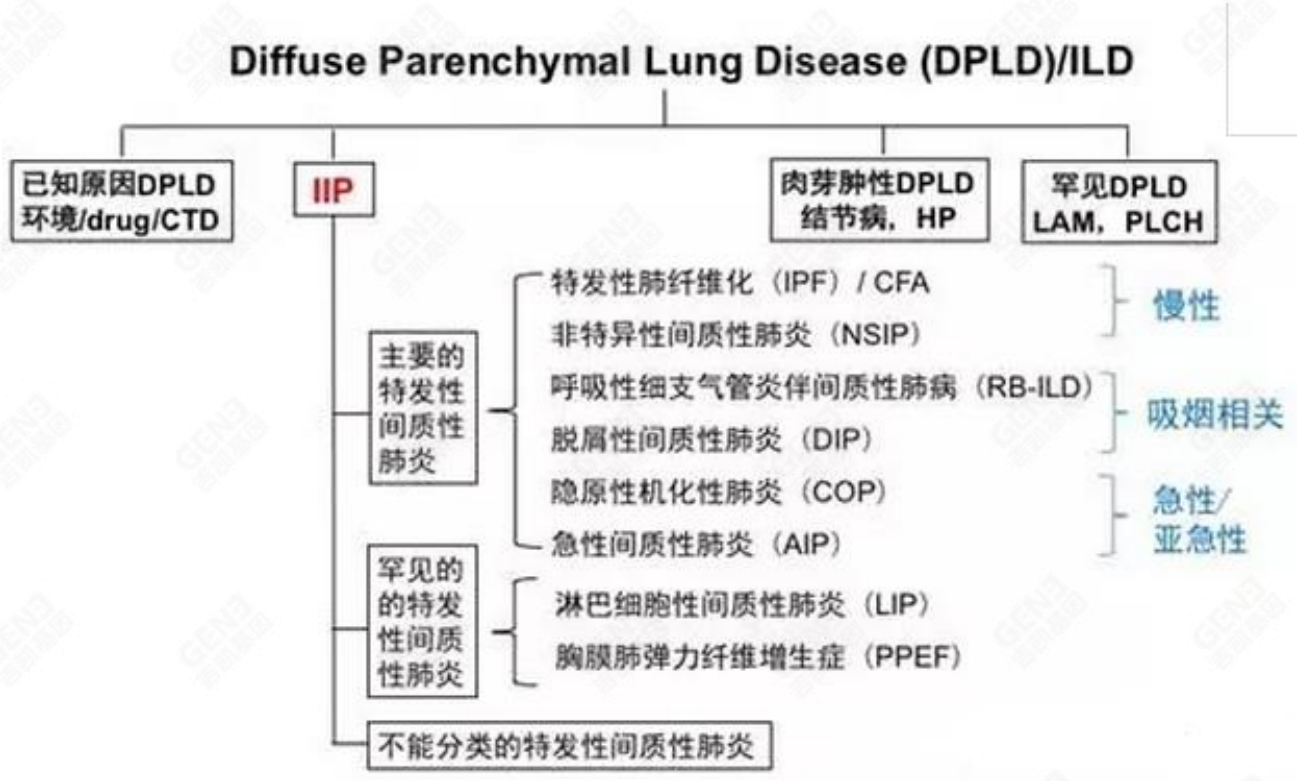
目前负责组学和单细胞领域前沿技术追踪和市场调研，以及生物信息学产品设计和方案设计。

目录 CONTENTS

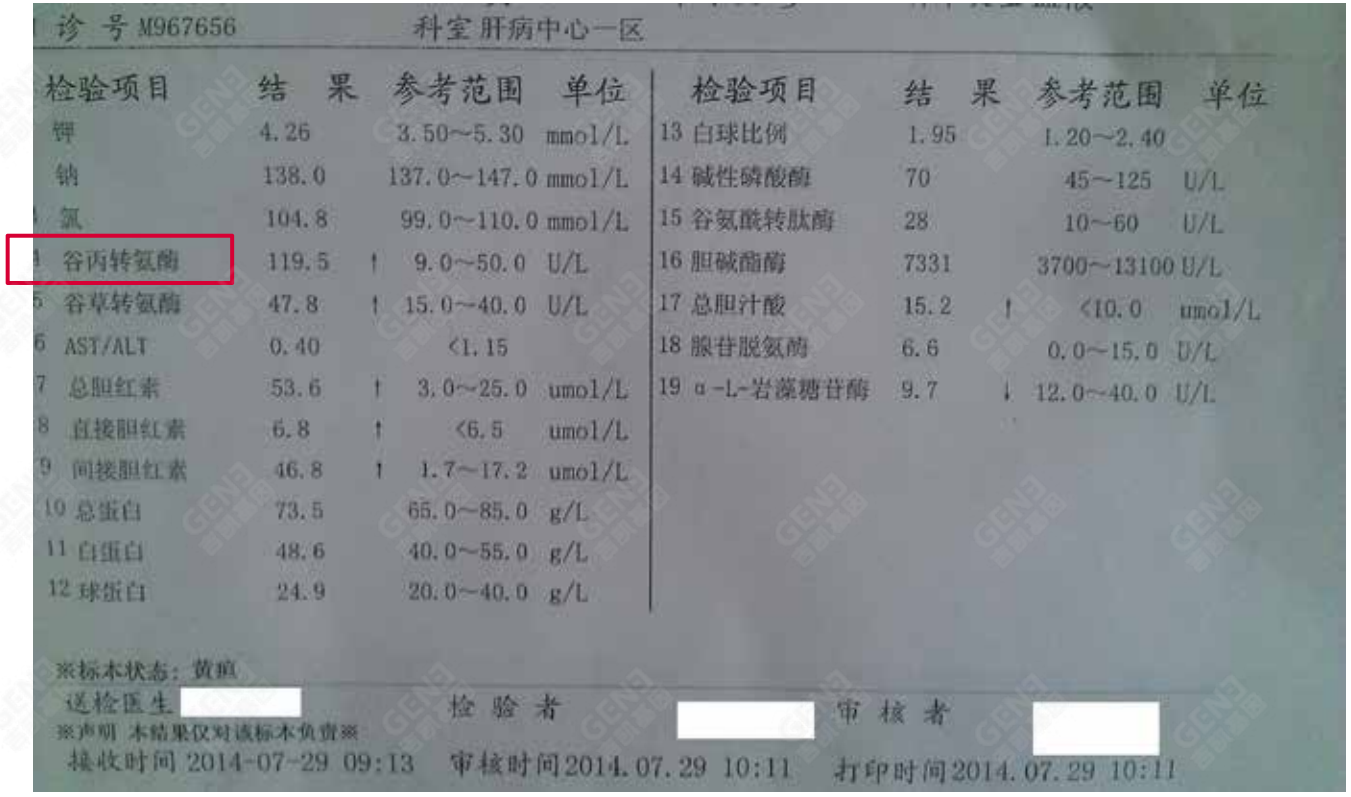
疾病分型

疾病分型：将疾病进行统一分类

间质性肺疾病的分类



什么是生物标志物



诊 号 M967656　　科室 肝病中心一区

检验项目	结 果		参考范围	单位
钾	4.26		3.50~5.30	mmol/L
钠	138.0		137.0~147.0	mmol/L
氯	104.8		99.0~110.0	mmol/L
4 谷丙转氨酶	119.5	↑	9.0~50.0	U/L
5 谷草转氨酶	47.8	↑	15.0~40.0	U/L
6 AST/ALT	0.40		<1.15	
7 总胆红素	53.6	↑	3.0~25.0	umol/L
8 直接胆红素	6.8	↑	<6.5	umol/L
9 间接胆红素	46.8	↑	1.7~17.2	umol/L
10 总蛋白	73.5		65.0~85.0	g/L
11 白蛋白	48.6		40.0~55.0	g/L
12 球蛋白	24.9		20.0~40.0	g/L

检验项目	结 果		参考范围	单位
13 白球比例	1.95		1.20~2.40	
14 碱性磷酸酶	70		45~125	U/L
15 谷氨酰转肽酶	28		10~60	U/L
16 胆碱酯酶	7331		3700~13100	U/L
17 总胆汁酸	15.2	↑	<10.0	umol/L
18 腺苷脱氨酶	6.6		0.0~15.0	U/L
19 α-L-岩藻糖苷酶	9.7	↓	12.0~40.0	U/L

※标本状态：黄疸
送检医生：　　检 验 者：　　审 核 者：
※声明 本结果仅对该标本负责※
接收时间 2014-07-29 09:13　审核时间2014.07.29 10:11　打印时间2014.07.29 10:11

生物标志物

定义

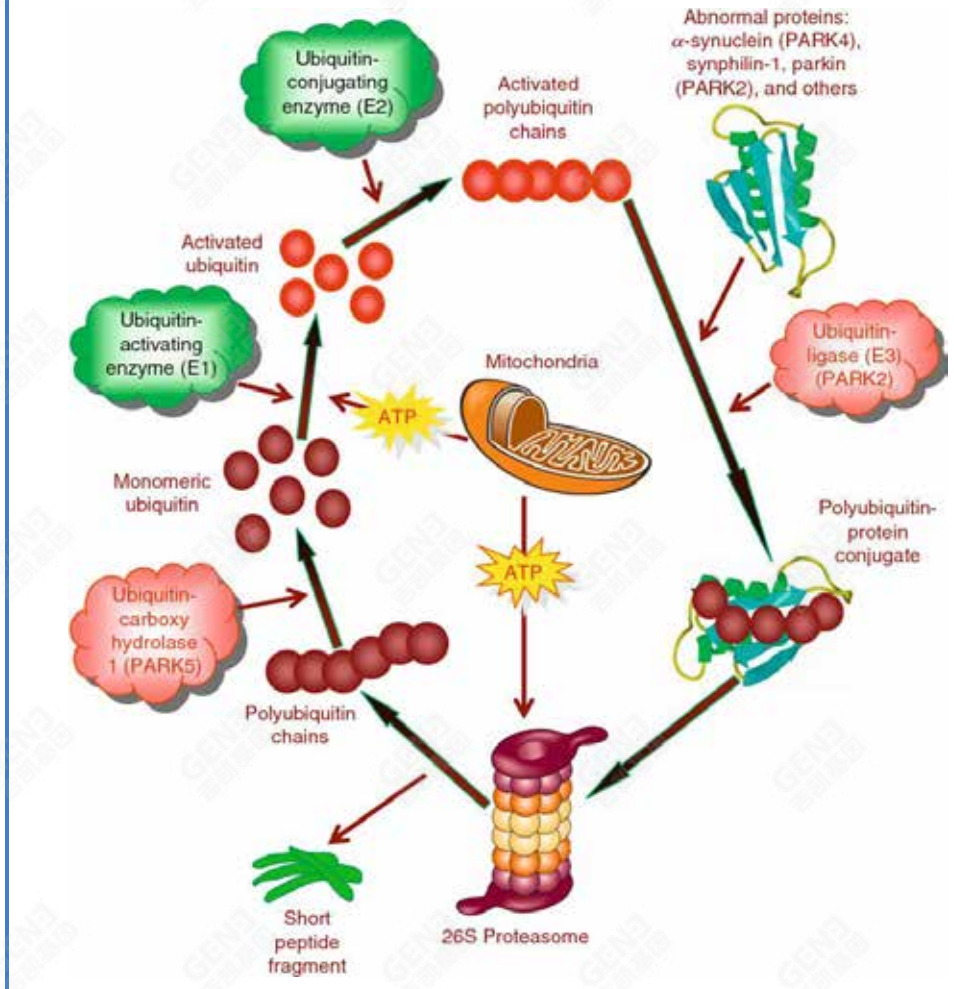
标记系统、器官、组织、细胞及亚细胞结构或功能的改变或可能发生的改变的生化指标

作用

疾病诊断；判断疾病分期；评价新药或新疗法在目标人群中的安全性及有效性。

传统生物标志物的发现要求分子机制图

参与帕金森病发病机制的基因



传统生物标志物的发现

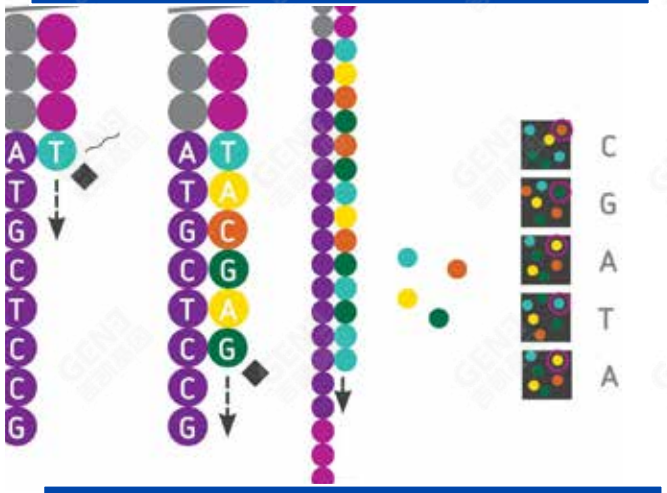
- 绘制疾病的分子机制图
- 推断某个环节存在异常，导致生化反应无法正常进行或受到抑制

传统方式存在的问题

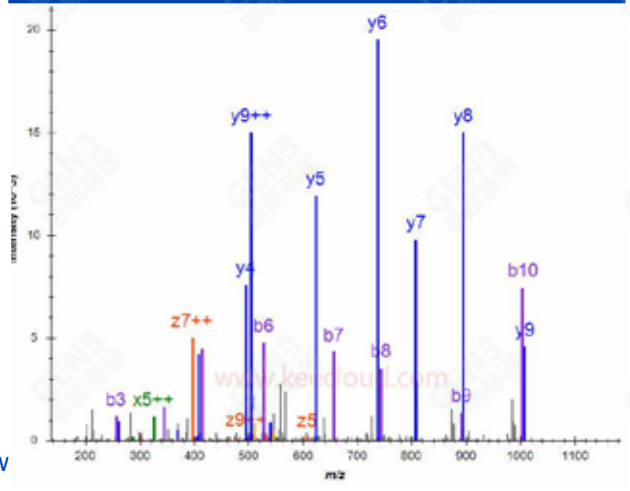
- 疾病的分子机制尚不明确
- 标志物不一定直接参与疾病的分子过程

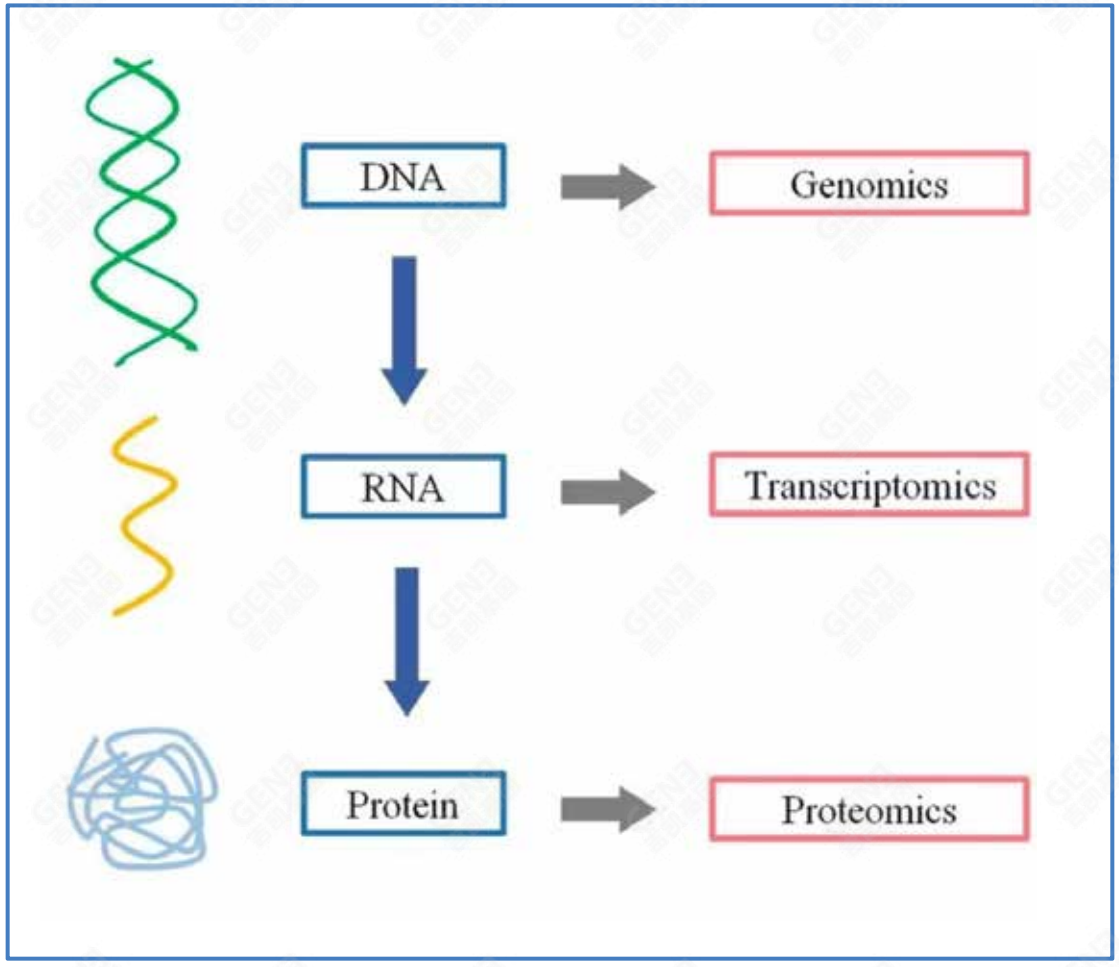
蛋白质组更适合疾病分型和标志物发现

二代测序



质谱





目录 CONTENTS

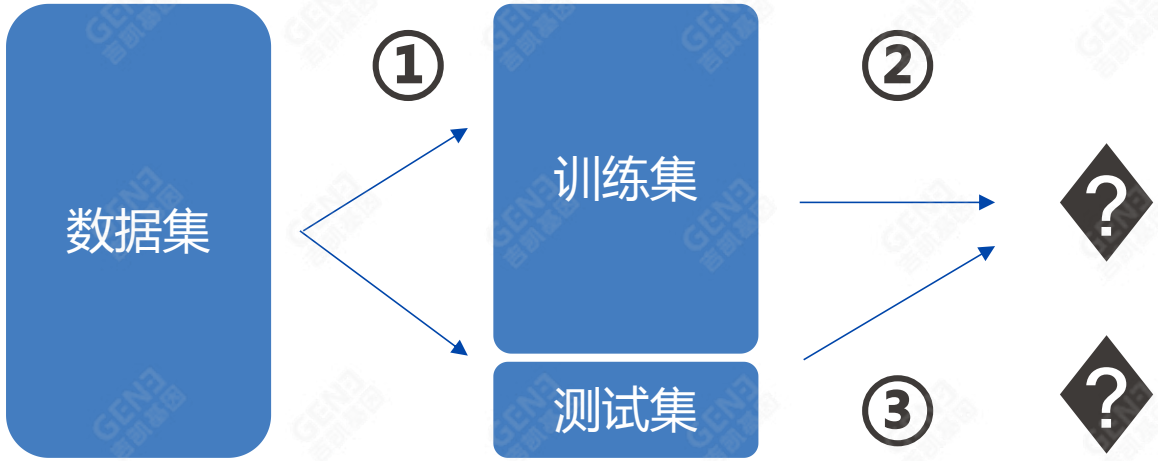
机器学习模型的作用是预测

模型

5

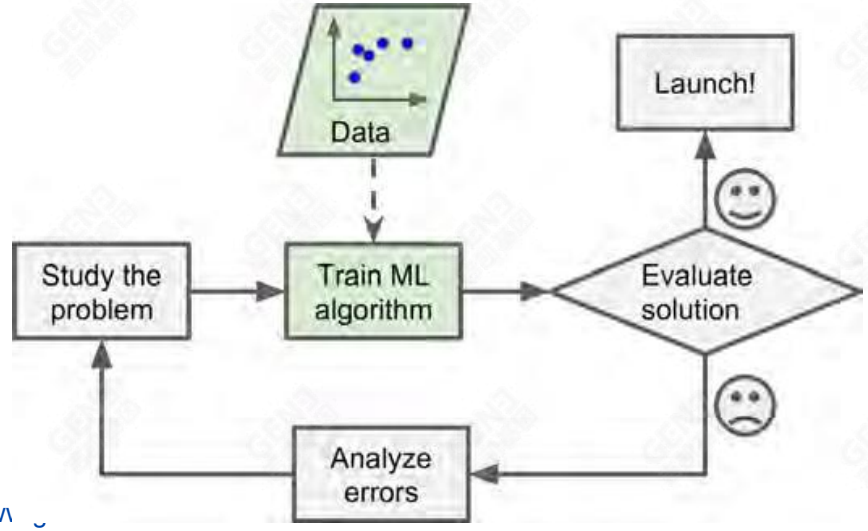
- 一个经典的机器学习任务是识别手写体数字（MNIST），及根据手写数字图像预测其表示的数字
- 预测模型分为两类，预测分类结果的分类模型和预测数值结果的回归模型
- 预测0~9的数字，获益与否都属于分类结果。上海的房价和患者预期生存期属于数值结果
- 机器学习和疾病分型

如何构建一个好的模型



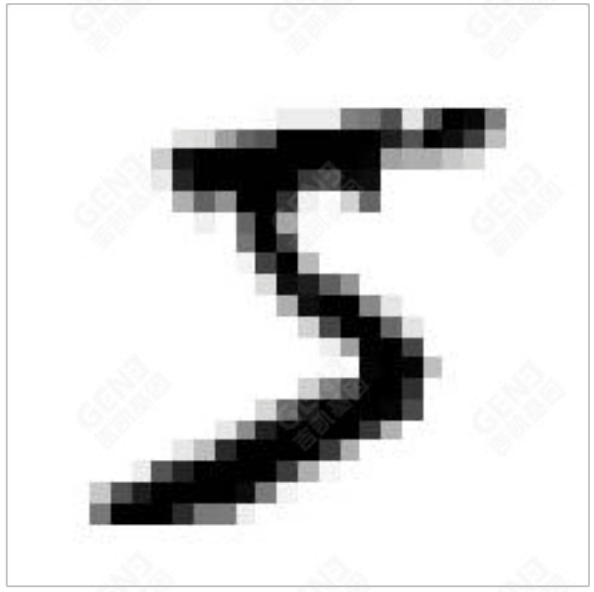
- 机器学习分为**三个步骤**：
  1. 将数据集拆分为训练集（training set）和测试集（test set）
  2. 使用训练集训练模型
  3. 使用测试集评价模型

- 模型训练是一个trial and error的过程



- 需要注意区分的是，训练和评价模型使用的数据都属于发现队列（discovery cohort）。验证队列（validation cohort）是在此以外重新收集的真实世界的数据
- 一个类比，发现队列的作用好比模拟，可以反复查看答案，以此提升能力（模型性能）。验证队列的作用好比高考，答案只用于评分（性能评价），不能用于提升模型。

数据由特征和标签组成



特征

	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	pixel9	pixel10	...	pixel775	pixel776	pixel7
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
69995	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
69996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
69997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
69998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
69999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	

标签

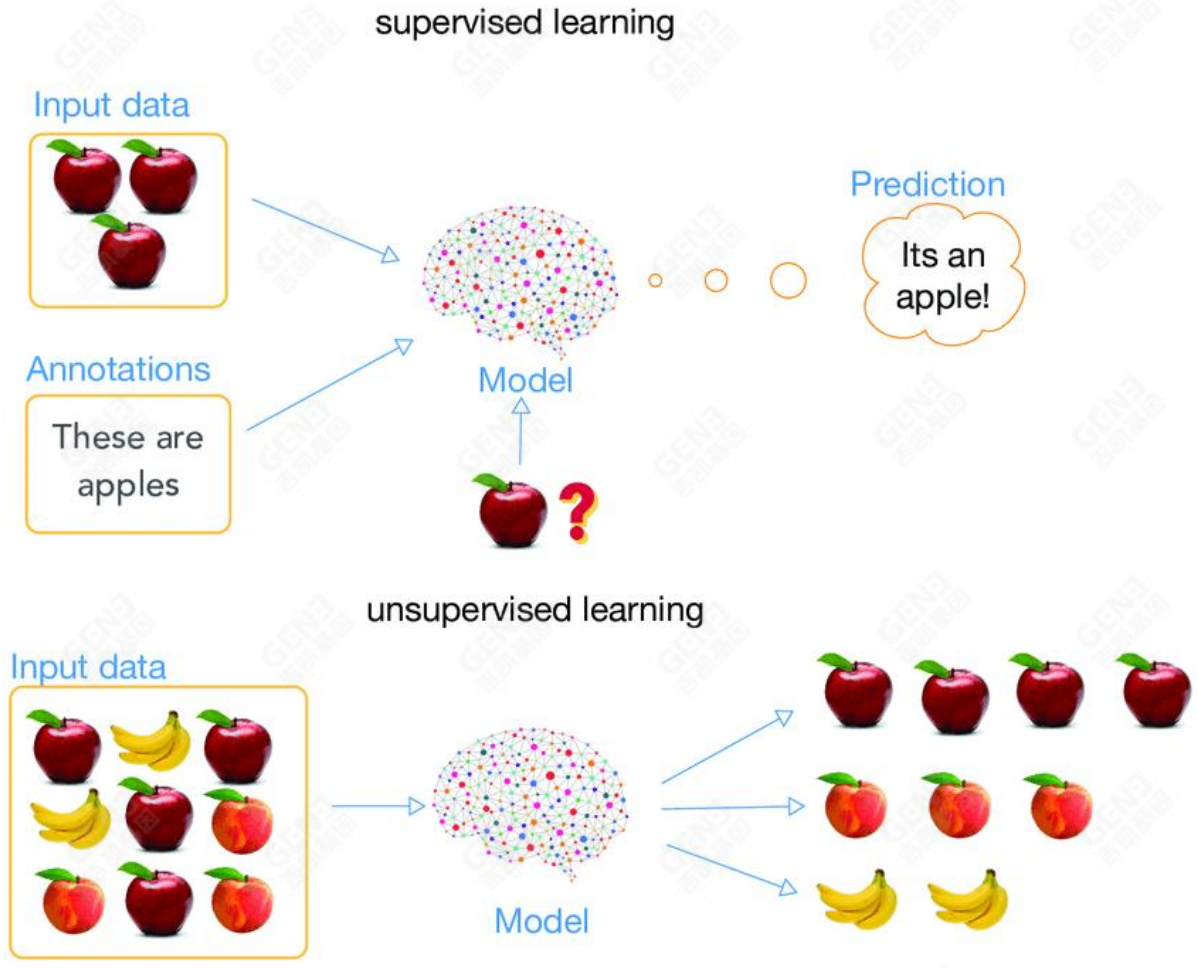
```
0        5
1        0
2        4
3        1
4        9
        ..
69995    2
69996    3
69997    4
69998    5
69999    6
Name: class, Length: 70
Categories (10, object)
```

1. 数据集中的每个样本，或者称为观测对象，包含特征和标签：
   特征是可以量化的观测对象的属性。例如MNIST中，每个数字图像可以分割成784（28×28）个像素点作为特征，而每个像素点的色彩强度则作为特征的值。
   标签是人为每个观测对象标记好的，需要模型预测的属性，例如数字5

2. 因为标签的存在，分类和回归模型也被称为有监督（supervised）的机器学习模型。与之相对的是无监督（supervised）的机器学习模型，如聚类和降维。

有监督学习vs无监督学习

- 有监督的机器学习中需要标记的数据
- 两者相比，有监督的机器学习通常用于对数据进行分类或预测，而无监督的机器学习是在原始和无标签的训练数据上训练模型。它经常被用来识别原始数据集的模式和趋势，或将类似的数据聚类到特定数量的组中。它也经常是在早期探索阶段使用的一种方法，以更好地理解数据集。
- 由于需要标记的数据，监督式机器学习的资源密集度更高
- 在无监督的机器学习中，由于少了人类的监督，可能更难达到足够的可解释性水平



常见有监督和无监督的机器学习模型

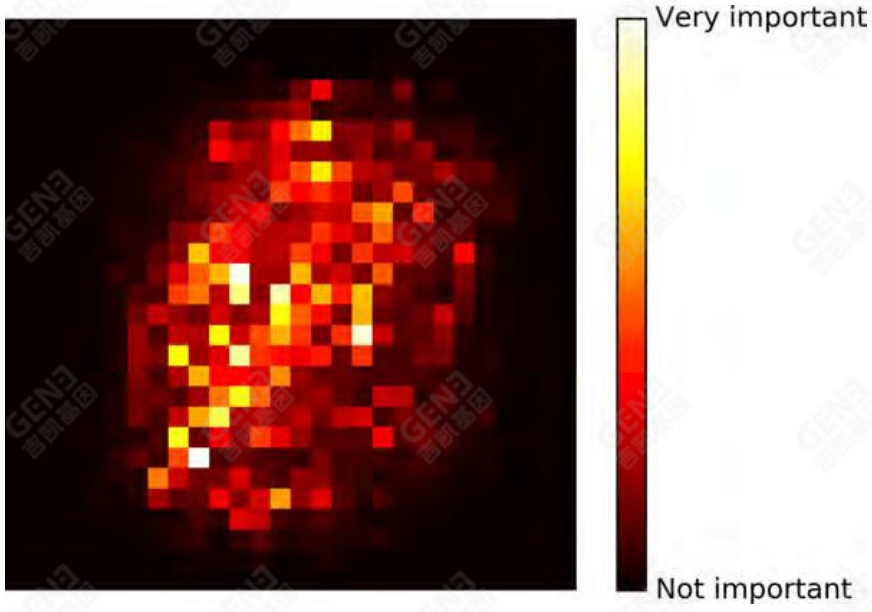
有监督机器学习模型

- 回归，岭（Ridge）回归和Lasso回归可防止过拟合
- Logistic regression
- LDA (Linear Discriminant Analysis), QDA (Quadratic Discriminant Analysis)
- bootstrap
- 决策树（Decision Tree），
- 集成学习 (随机森林, Random Forest; boosting, XGBoost)
- 支持向量机 (Support Vector Machine, SVM)

无监督机器学习模型

- 降维、聚类

特征重要性和生物标志物



在模型训练过程中，可以获得一个“副产物”，就是每个特征的重要性评分。图片显示的是MNIST数据中，每个像素点的重要性。可以发现外围的像素点的重要性几乎为0，因此可以再下一轮的模型优化环节中排除，以加快训练速度。有时这种手段还可以减少噪音。

在医学研究中，重要性得分高的特征往往具有作为诊断或筛查生物标志物的潜力。但需要注意的是：

1. 这样的重要性可能与技术方法相关，例如一个通过质谱蛋白组筛选出来的重要特征，在酶联免疫方法中的表现却不好
2. 这样的标志物通常不能够解释疾病的发生机制，因为事物的发生存在因果关系，一个标志物发生显著改变往往是果而不是因。预测疾病发生的驱动基因需要用到其他机器学习算法，如因果推断模型。

混淆矩阵（Confusion Matrix）评价模型好坏

混淆矩阵		预测值	
		正类	负类
实际值	正类	真正类（TP）	假负类（FN）
	负类	假正类（FP）	真负类（TN）

$$Accuracy = \frac{4096 + 53057}{4096 + 1522 + 1352 + 53057} = 95.2\%$$

$$Precision = \frac{4096\ (TP)}{4096\ (TP) + 1522\ (FP)} = 72.9\%$$

$$Recall = \frac{4096}{4096 + 1325} = 75.6\%$$

对SDG分类器的精度和召回率进行计算，性能远不如准确性那么光鲜亮眼了。

精度（precision）：用于评价正类预测的准确性

$$precision = \frac{TP}{TP + FP}$$

召回率（recall）也称灵敏度，或真正类率，灵敏度高，不会错放一个患病的人（FN），漏诊率低。

$$recall = \frac{TP}{TP + FN}$$

特异度（specificity） 特异度高，不会冤枉一个没病的人（FP），误诊率低。

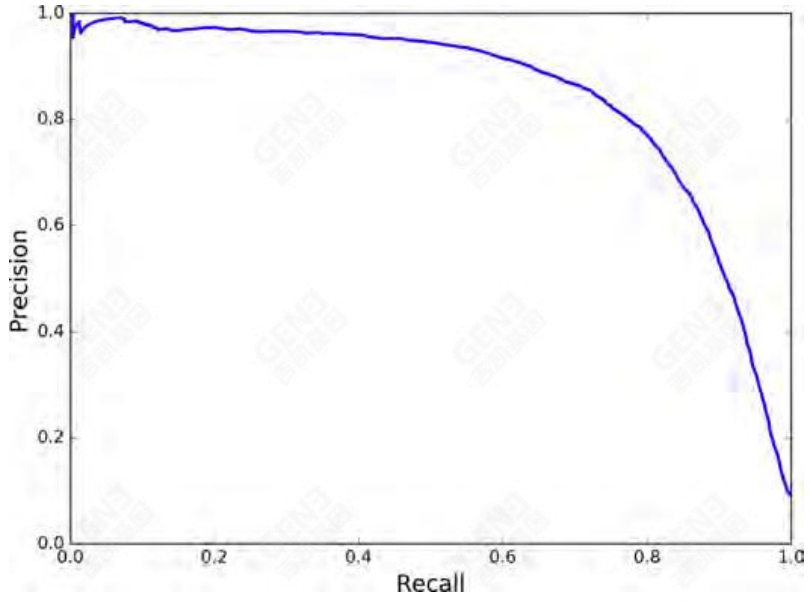
$$specificity = \frac{TN}{TN + FP}$$

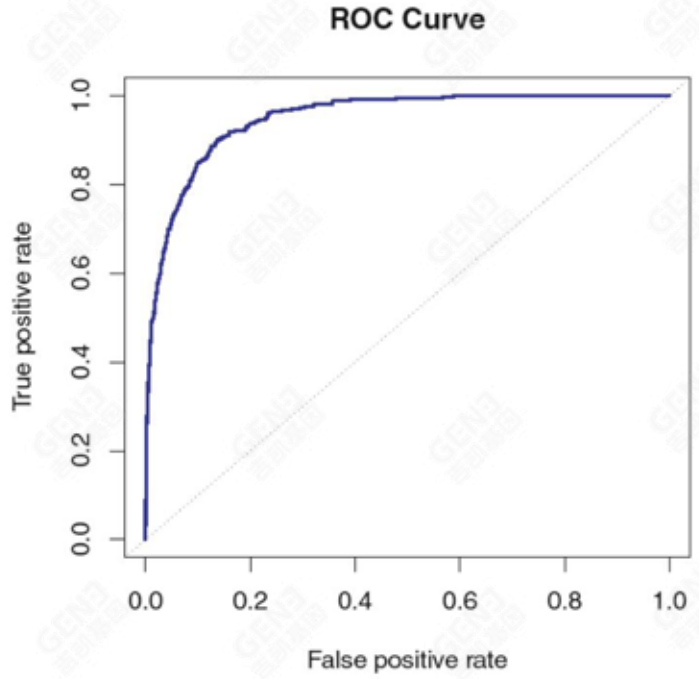
F_1分数：将精度和召回率组合在一起

$$\mathrm{F}_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

精度/召回率曲线（PR）与受试者操作曲线（ROC）

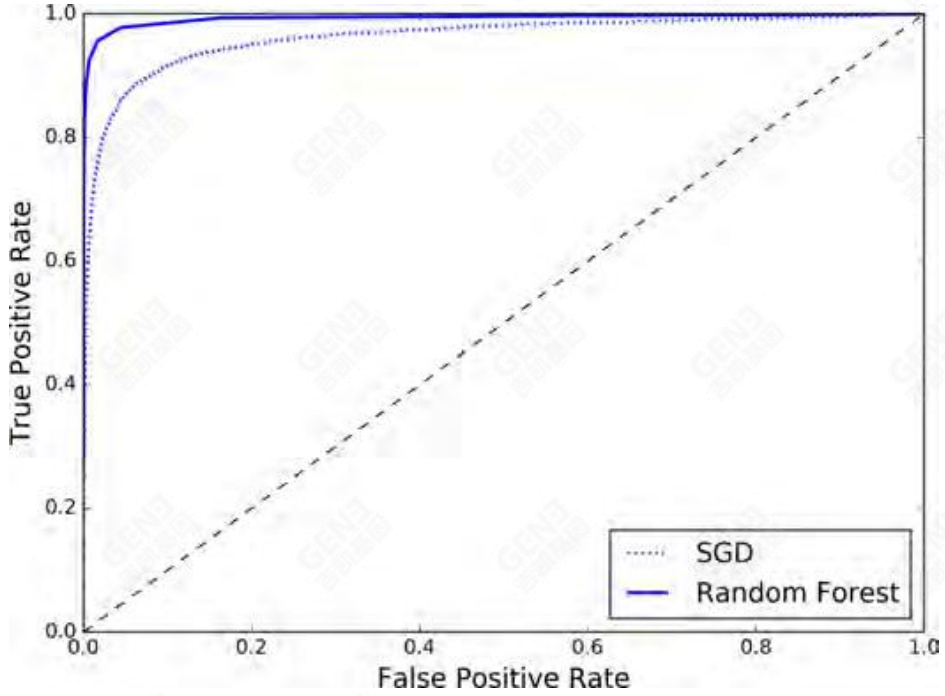
精度/召回率曲线（PR）





- 权衡精度和召回率可以采用精度/召回率曲线（PR）。它以召回率为横坐标，精度为纵坐标绘制而成。
- 于此类似的是受试者操作曲线（ROC）。其横坐标表示假正率，纵坐标表示真正率。虚线表示纯随机的分类器曲线。

使用曲面下面积（AUC）比较分类器



- 测量曲面下面积（AUC）是常用的比较分类器的方法，完美的分类器的ROC AUC等于1，而纯随机的ROC AUC等于0.5。
- 思考纯随机的ROC AUC为什么不是0？
- 如图所示，使用随机森林构建的分类器在性能上要优于SGD。

目录 CONTENTS

supervised - 尿液蛋白组用于家族性帕金森病患者的分层分析

Urinary proteome profiling for stratifying patients with familial Parkinson's disease

Sebastian Virreira Winter [1], Ozge Karayel [1], Maximilian T Strauss [1], Shalini Padmanabhan [2], Matthew Surface [3], Kalpana Merchant [4], Roy N Alcalay [3], Matthias Mann [1 5]

EP Q1 IF: 12.137 Cited by: 13 Sci-Hub Link PDF(Full Text) Citation Collect

背景：帕金森病 (PD) 是第二常见的神经退行性疾病，早期检测PD的方法缺失
样本： 235 个尿液样本，来自两个队列 (Columbia, LCC), 分为四类，

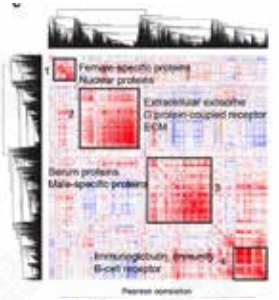
- 健康对照 (*LRRK2* $^-$ /PD$^-$, HC)
- *携带LRRK2* G2019S 突变 (*LRRK2* $^+$ / PD$^-$, NMC) 的非显性携带者;
- 特发性 PD 患者 (*LRRK2* $^-$ /PD $^+$, iPD) ;
- *LRRK2* G2019S显性的 PD 患者 (*LRRK2* $^+$ /PD$^+$, LLRK2 PD)

supervised - 尿液蛋白组用于家族性帕金森病患者的分层分析

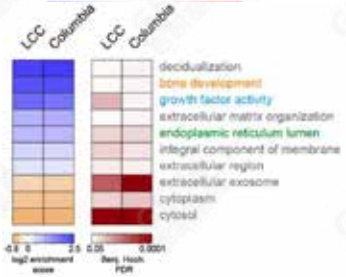
不一致的样品处理、样本采集

排除11个样本

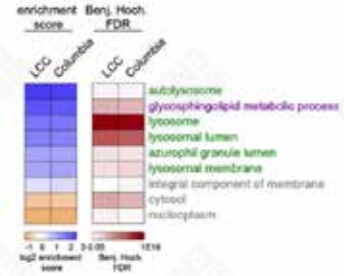
1. 几个共调节蛋白簇，富集到细胞外泌体，免疫球蛋白，B 细胞受体



2. PD vs 非PD：361个差异蛋白质，与蛋白折叠，典型核糖核酸酶相关，bone development相关



3. *LRRK2* G2019S 突变vs 非突变：237 种差异蛋白质，与溶酶体、及鞘脂代谢过程相关的term显著富集

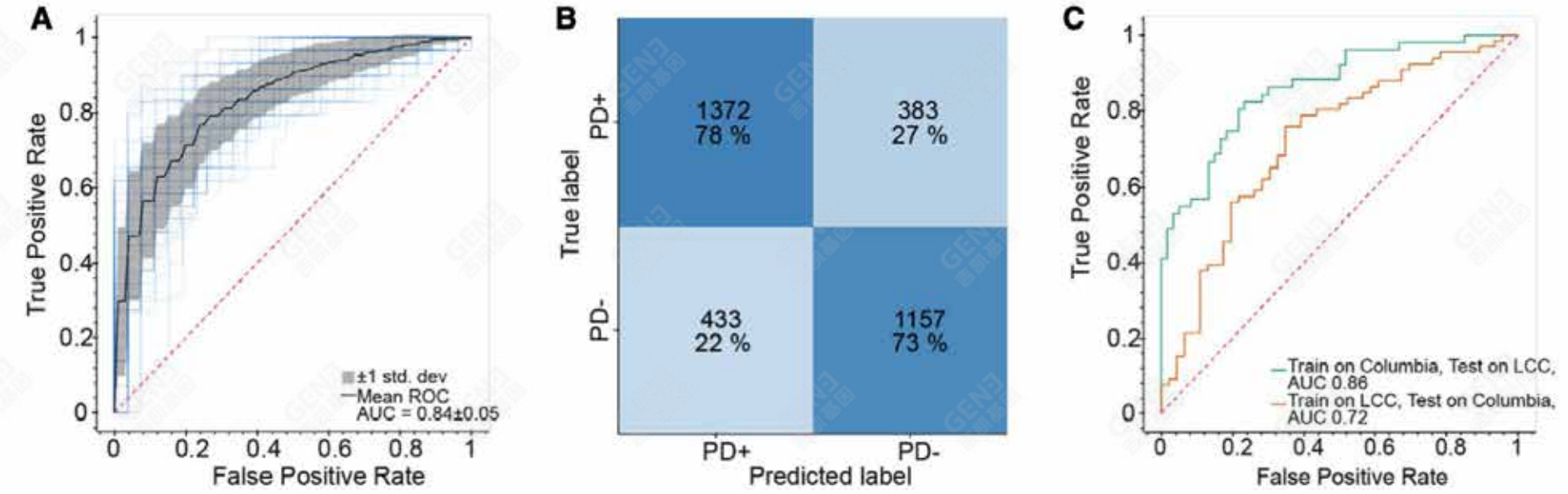


4. 蒙特利尔认知评估（MoCA）测试评估的参与者的认知能力，TNR和FURIN，与PD患者的MoCA评分显示出强烈的负相关

supervised - 尿液蛋白组用于家族性帕金森病患者的分层分析

supervised - 尿液蛋白组用于家族性帕金森病患者的分层分析

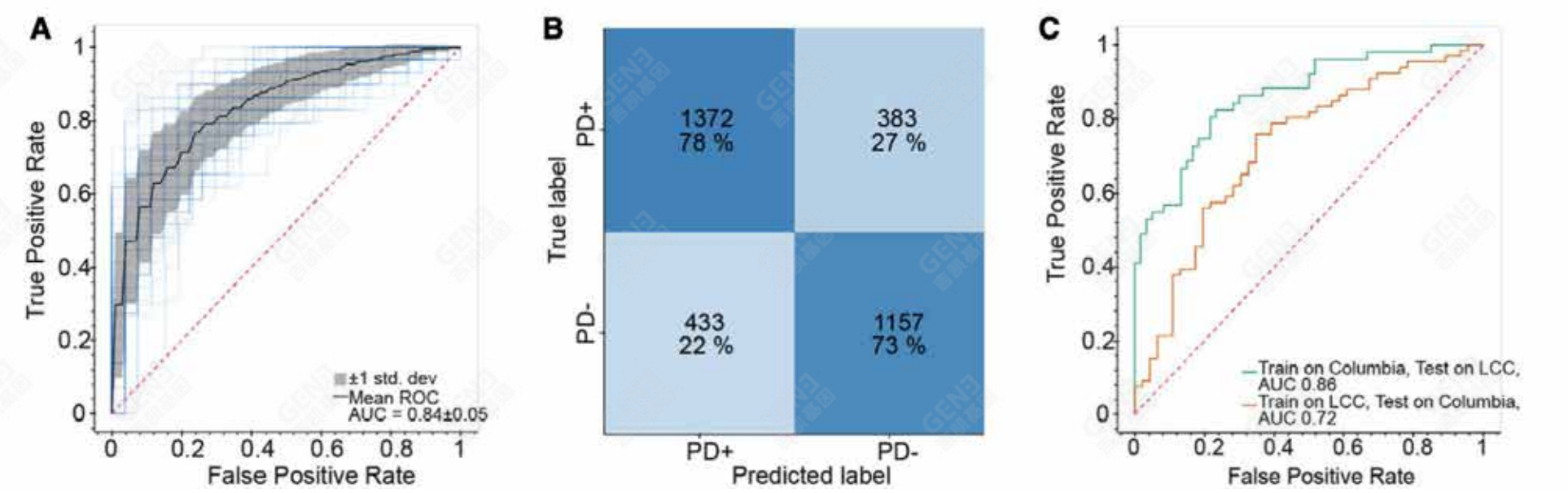
1. PD+ vs. PD- 2. LRRK2 G2019S vs. LRRK2 WT carriers 3. PD+ vs. PD- in LRRK2 G2019S carriers



- 决策树选择了15个最重要蛋白，作为特征，其中PPIB的强度位居榜首，这是PD样本与对照组相比丰度差异最大的蛋白质之一
- AUC = 0.84±0.05
- 灵敏度 = 78%，特异性 = 73%
- 在一个队列中训练模型并在另一个队列中测试时，我们获得了0.86或0.72的AUC

supervised - 尿液蛋白组用于家族性帕金森病患者的分层分析

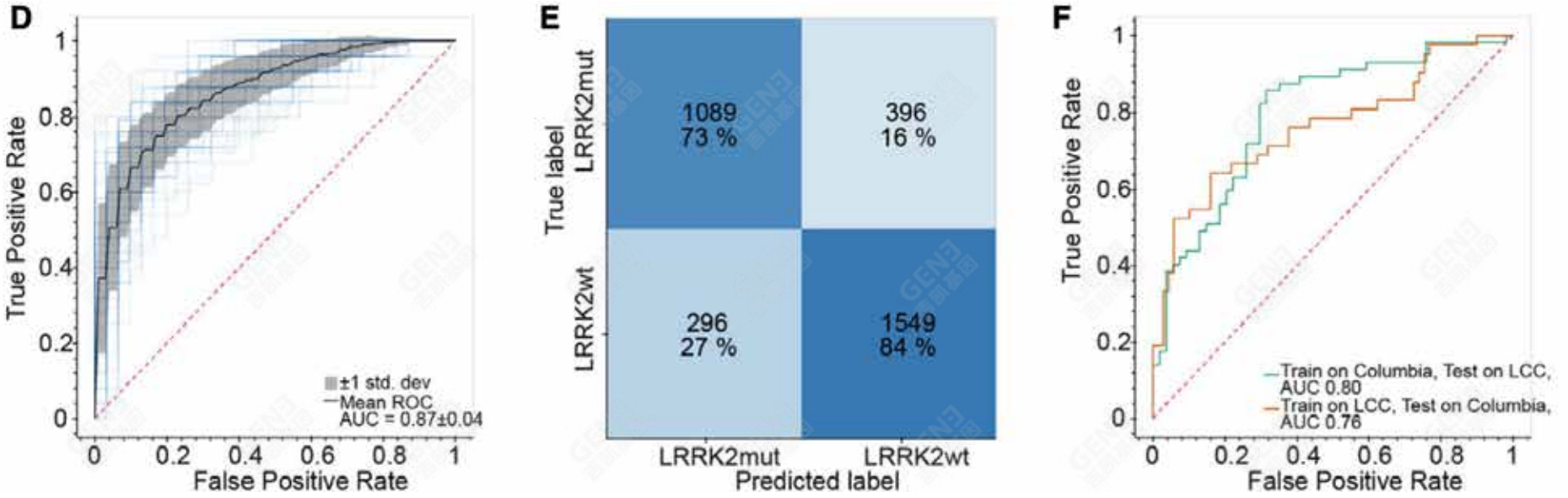
1. PD+ vs. PD



- 决策树选择了15个最重要蛋白，作为特征，其中PPIB的强度位居榜首，这是PD样本与对照组相比丰度差异最大的蛋白质之一
- AUC = 0.84±0.05
- 灵敏度 = 78%，特异性 = 73%
- 在一个队列中训练模型并在另一个队列中测试时，我们获得了0.86或0.72的AUC

supervised - 尿液蛋白组用于家族性帕金森病患者的分层分析

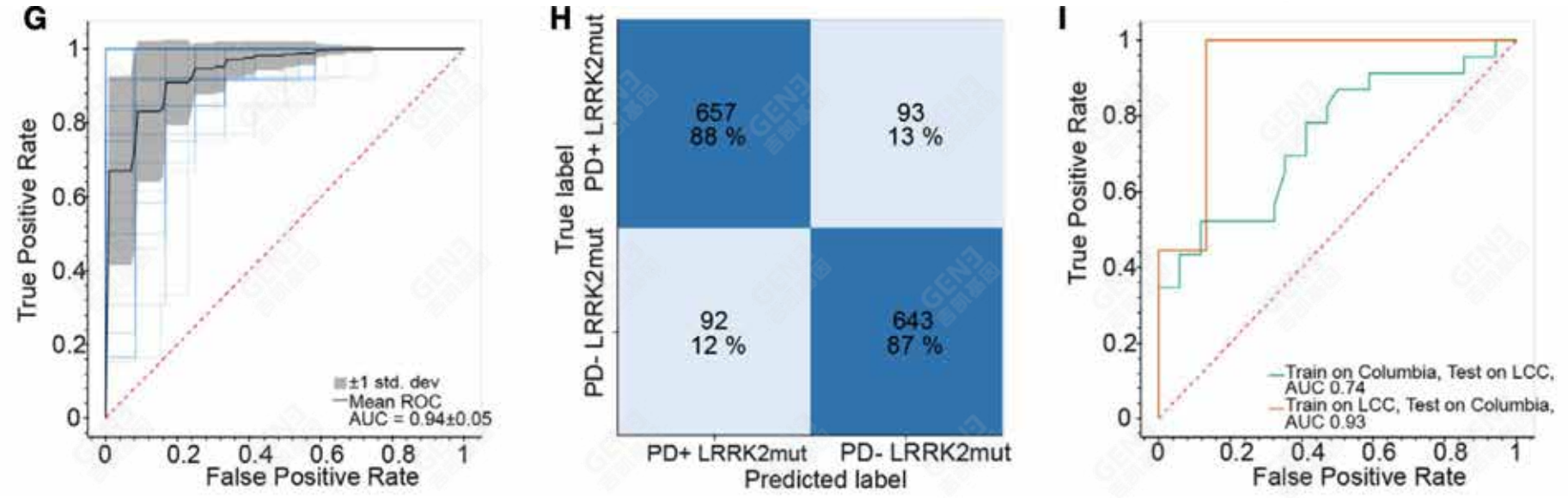
2. LRRK2 G2019S vs. LRRK2 WT carriers



- 决策树选择了15个最重要蛋白，作为特征，其中ENPEP是最重要的一个
- AUC = 0.87 ± 0.04
- 灵敏度 = 74%，特异性 = 84%
- 在一个队列中训练模型并在另一个队列中测试时，我们获得了0.76或0.80的AUC

supervised - 尿液蛋白组用于家族性帕金森病患者的分层分析

3. PD+ vs. PD- in LRRK2 G2019S carriers



- 决策树选择了7个最重要蛋白，作为特征，其中VGF，一种神经营养因子，是最重要的一个
- AUC = 0.94±0.05
- 灵敏度 = 88%，特异性 = 88%
- 在一个队列中训练模型并在另一个队列中测试时，我们获得了0.93和0.74的AUC

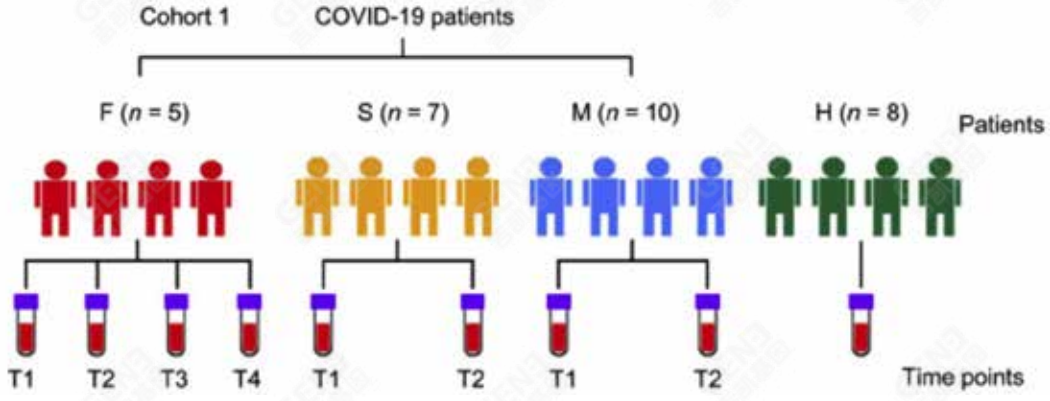
supervised -血浆蛋白质组学确定COVID-19的生物标志物和致病机制

Plasma Proteomics Identify Biomarkers and Pathogenesis of COVID-19

Ting Shu [1], Wanshan Ning [2], Di Wu [3], Jiqian Xu [4], Qiangqiang Han [5], Muhan Huang [3], Xiaojing Zou [4], Qingyu Yang [1], Yang Yuan [6], Yuanyuan Bie [7], Shangwen Pan [4], Jingfang Mu [3], Yang Han [1], Xiaobo Yang [4], Hong Zhou [6], Ruiting Li [4], Yujie Ren [3], Xi Chen [5], Shanglong Yao [8], Yang Qiu [9], Ding-Yu Zhang [10], Yu Xue [11], You Shang [12], Xi Zhou [13]

EP Q1 IF: 31.745 Cited by: 69 Sci-Hub Link PDF(Full Text) Citation Collect

背景：COVID-19
样本类型：血液样本
样本来源：5 fatal (F), 7 severe (S), 10 mild (M), 8 health (H)



supervised -血浆蛋白质组学确定COVID-19的生物标志物和致病机制

1

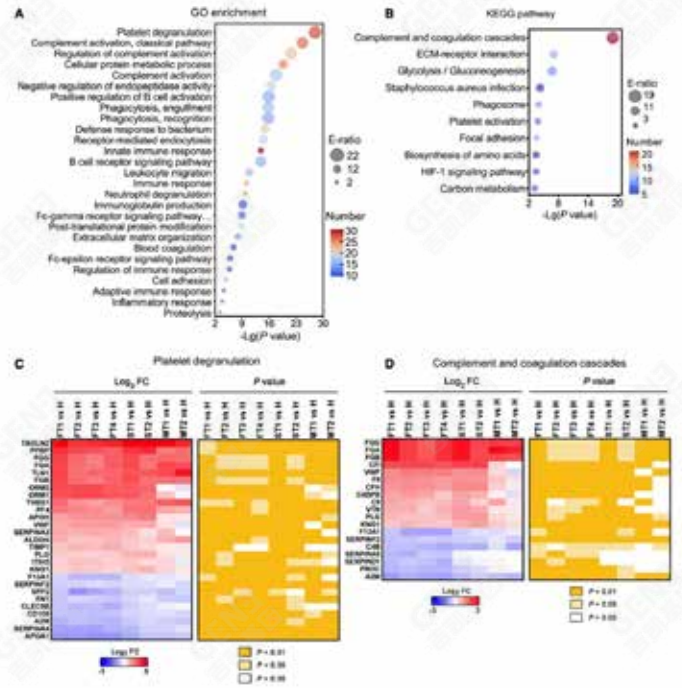
F与H组之间存在195个差异蛋白，与S,F组相比，M组的差异表达程度降低。

这些基因富集在炎症、免疫细胞迁移和脱颗粒、补体系统、凝血级联和能量代谢，血小板脱颗粒以及补体和凝血级联等。

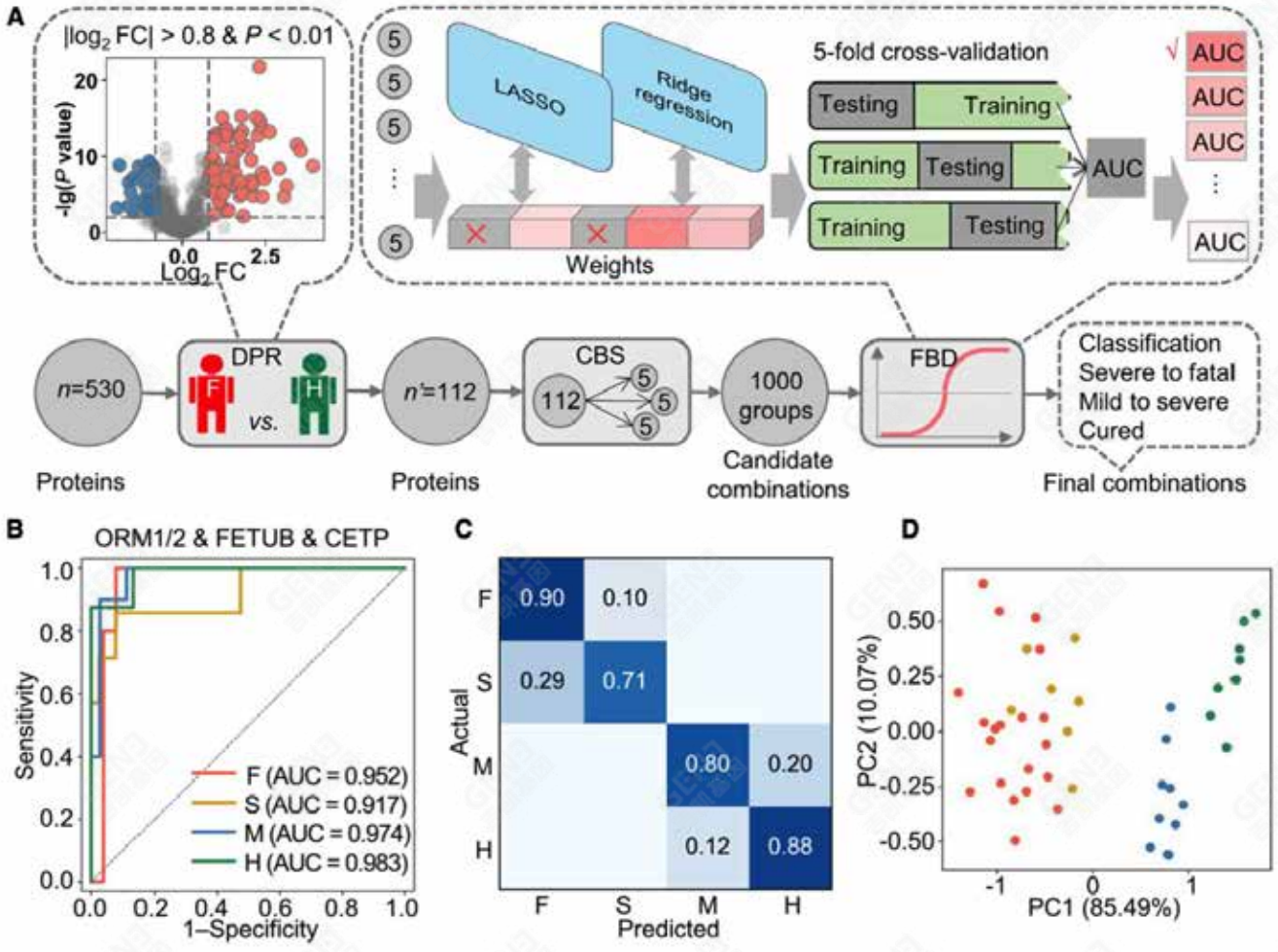
S组（5/7）和F组（2/5）分别发现了两种SARS-CoV-2编码的蛋白nsP2和nsP7，但在M组和H组的样本中均未发现

2

机器学习



supervised -血浆蛋白质组学确定COVID-19的生物标志物和致病机制



建立模型：

1. 差异蛋白保留（DPR）112个DEPs
2. 候选生物标志物选择（CBS）以生成1,000组初始生物标志物组合
3. 最终生物标志物确定（FBD）以得到蛋白质组合
4. 预测：惩罚逻辑回归 (PLR)

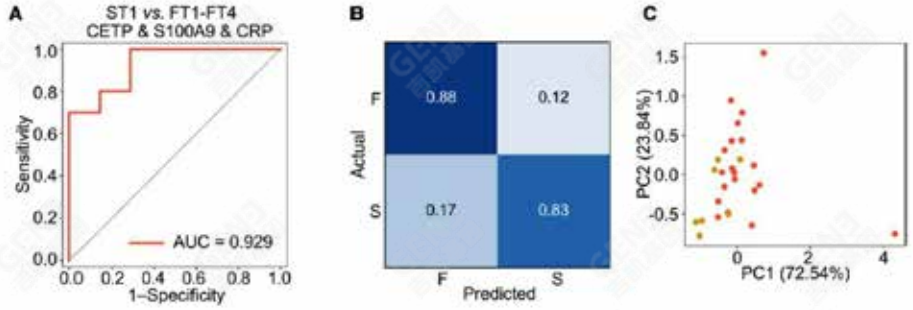
Biomarker包括： ORM1/AGP1, ORM2, FETUB, CETP

测试：

1. 5-fold cross-validation: AUC (0.952, 0.917, 0.974, 0.983); 混淆矩阵; PCA
2. 验证队列：26个血浆样本 (9F, 6S 和 6M, 5H), AUC (0.941, 0.825, 0.842, 1.000)

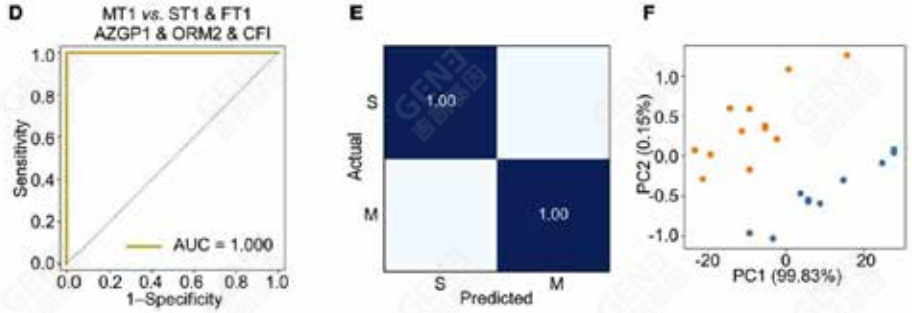
supervised -血浆蛋白质组学确定COVID-19的生物标志物和致病机制

F vs S



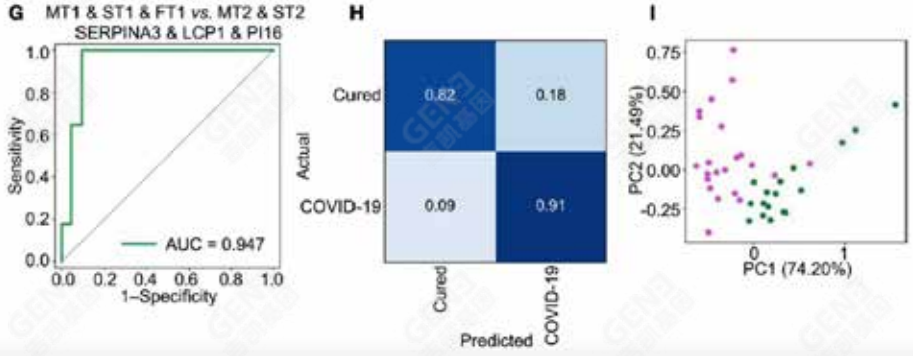
- biomarker: CETP, S100A9, CRP
- AUC: 0.929

S vs M



- biomarker: AZGP1, ORM2, CFI
- AUC: 1，但M 组的患者比 S 或 F 组的患者年轻，以年龄 marker，AUC=0.792

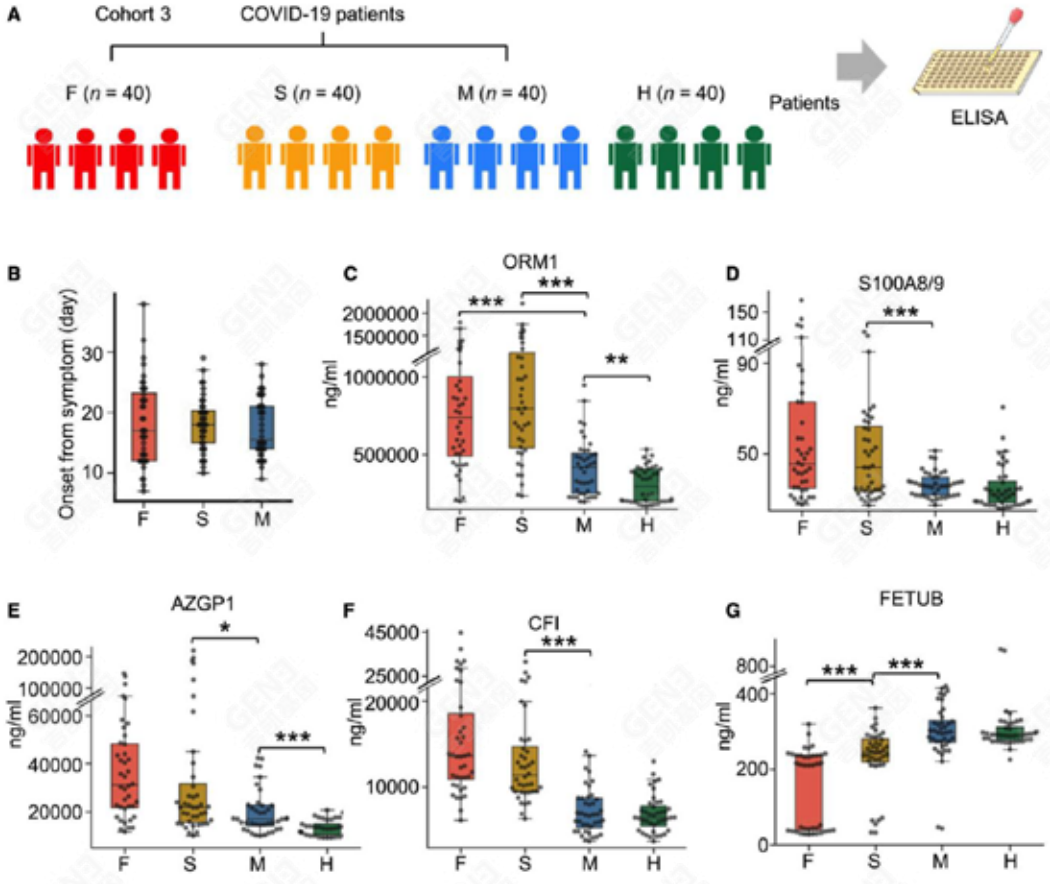
Cured vs COVID-19



- biomarker: SERPINA3/ACT, LCP1/LPL, PI16)
- AUC: 0.947，单个蛋白质作为 biomarker：AUC：0.832 至 0.941

supervised -血浆蛋白质组学确定COVID-19的生物标志物和致病机制

验证biomarker



- 方法：ELISA: 酶联免疫吸附测定
- 结果：ORM1、AZGP1、CFI、FETUB 和 S100A8/S100A9 的血浆水平在不同患者中有显著差异

supervised - 鉴定胰腺导管腺癌的循环蛋白生物标志物

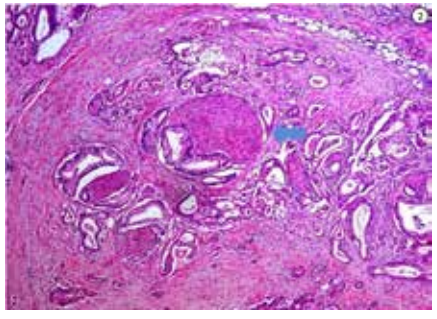
Multicenter Study > Clin Cancer Res. 2021 May 1;27(9):2592-2603.
doi: 10.1158/1078-0432.CCR-20-4215. Epub 2021 Mar 18.

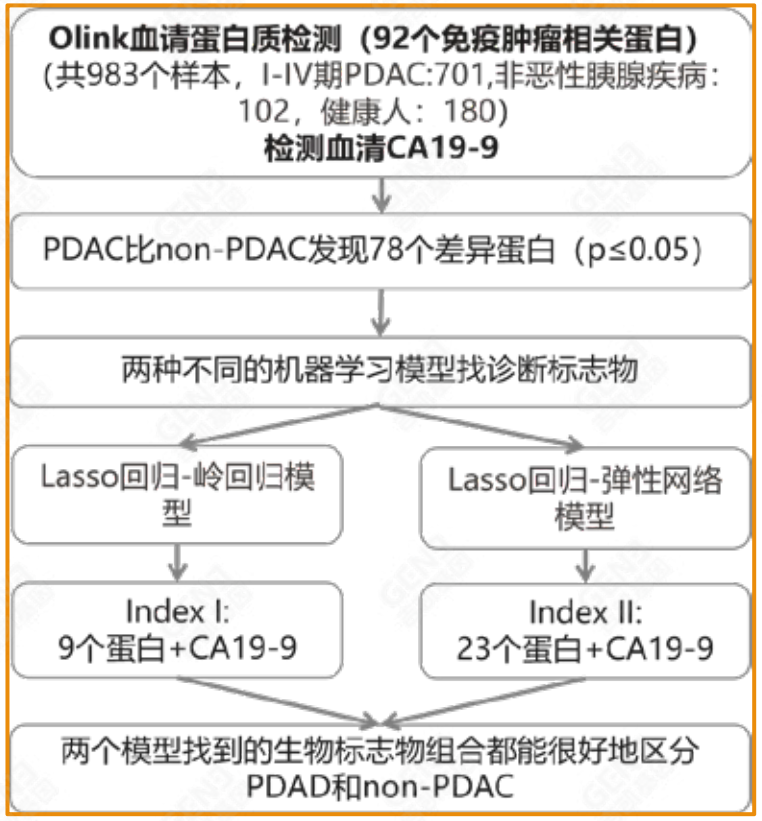
Circulating Protein Biomarkers for Use in Pancreatic Ductal Adenocarcinoma Identification

Sidsel C Lindgaard [1], Zsófia Sztupinszki [# 2], Emil Maag [# 3], Inna M Chen [4], Astrid Z Johansen [4], Benny V Jensen [4], Stig E Bojesen [5 6], Dorte L Nielsen [4 6], Carsten P Hansen [7], Jane P Hasselby [8], Kaspar R Nielsen [9], Zoltan Szallasi [2 10], Julia S Johansen [4 6 11]

Affiliations + expand

PMID: 33737308 DOI: 10.1158/1078-0432.CCR-20-4215





unsupervised -利用机器学习发现肺动脉高压的独特免疫表型

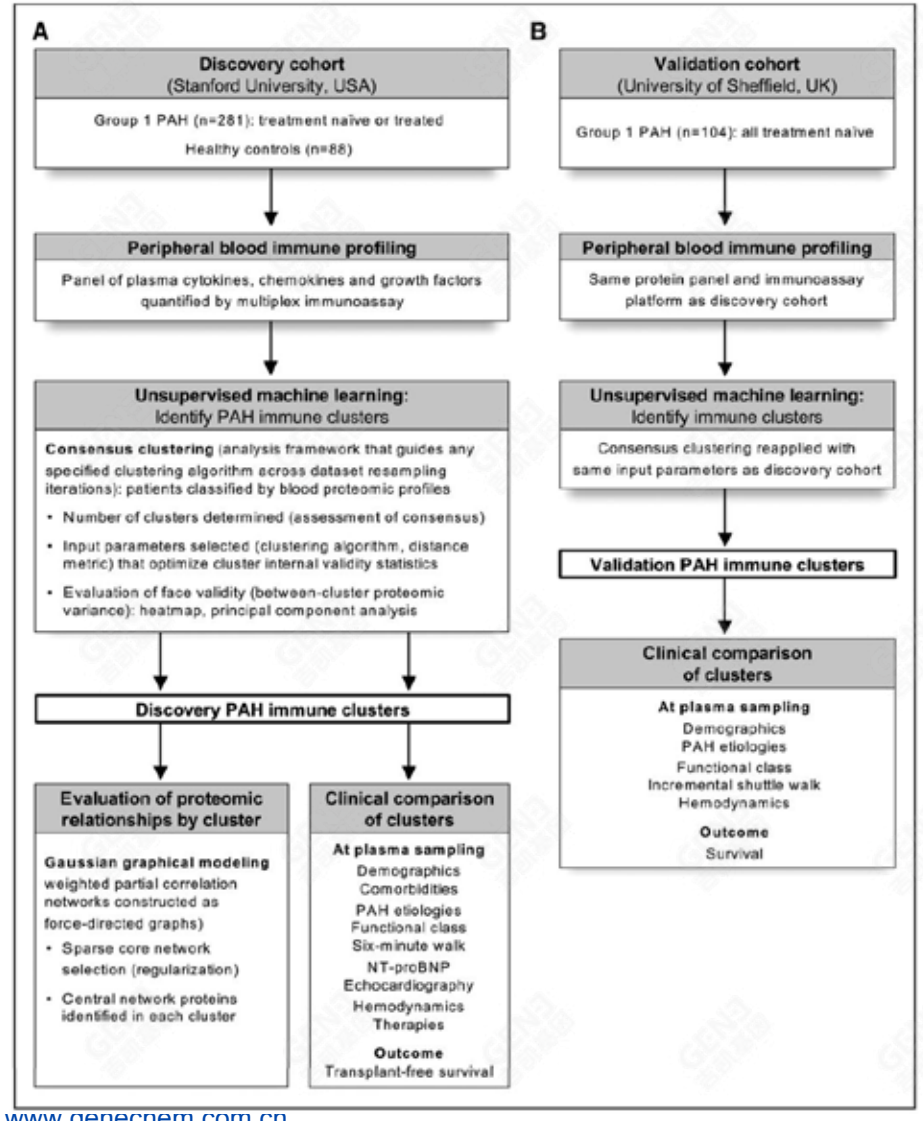
Discovery of Distinct Immune Phenotypes Using Machine Learning in Pulmonary Arterial Hypertension

Andrew J Sweatt [1 2], Haley K Hedlin [3], Vidhya Balasubramanian [3], Andrew Hsi [2], Lisa K Blum [4], William H Robinson [4], Francois Haddad [5 6], Peter M Hickey [7], Robin Condliffe [8], Allan Lawrie [7], Mark R Nicolls [1 9 2], Marlene Rabinovitch [2 10], Purvesh Khatri [9 11], Roham T Zamanian [1 2]

EP Q1 IF: 17.367 Cited by: 57 Sci-Hub Link PDF(Full Text) Citation Collect

背景：越来越多的证据表明，炎症与肺动脉高压（PAH）有关，针对免疫的治疗方法也在调查之中，但仍不知道是否存在不同的免疫表型。
样本类型：外周静脉血，蛋白质组
样本来源： PAH患者（发现队列：n=281；验证队列：n=104）；健康对照组（n=88）
测序：Bio-plex多重免疫测定，测量了48种细胞因子、趋化因子和生长因子

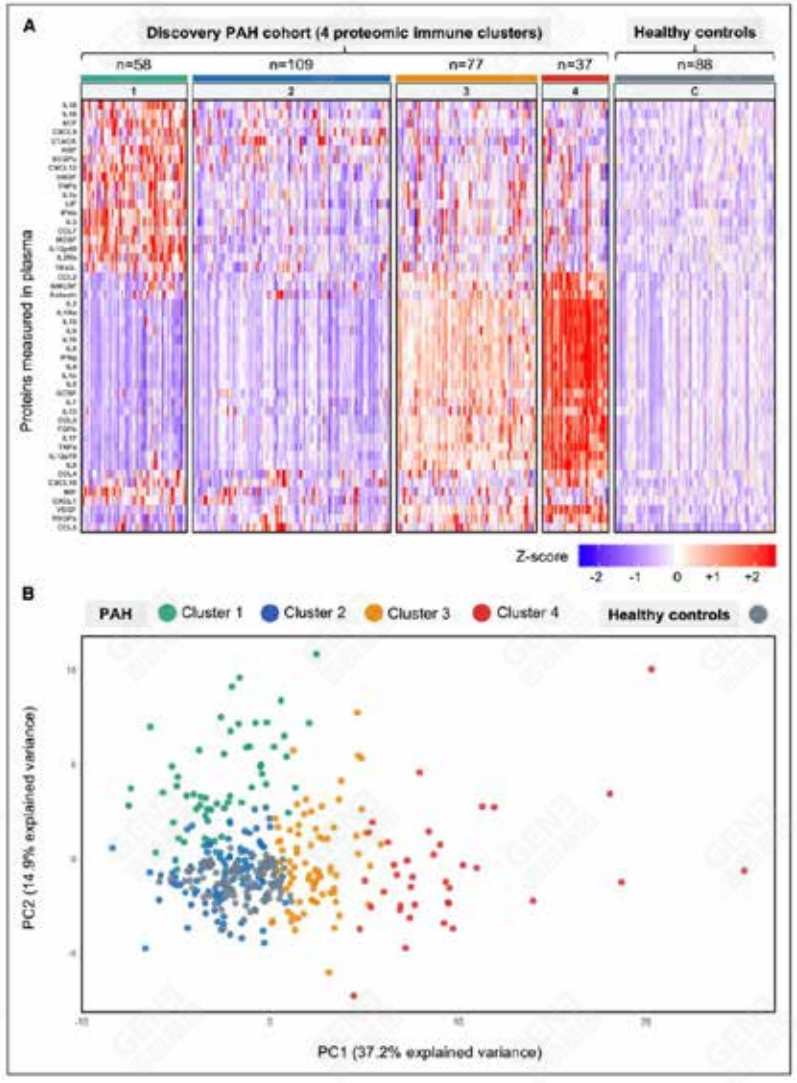
unsupervised -利用机器学习发现肺动脉高压的独特免疫表型



发现队列：
（1）无监督的机器学习（共识聚类），在没有临床数据指导的情况下，根据蛋白质组免疫图谱确定PAH集群，
（2）蛋白质组网络分析，确定所发现集群中的中心蛋白
（3）比较各集群的临床特征。

验证队列：
重新应用无监督共识聚类法，确定该方法是否产生了具有细胞因子谱和临床特征的免疫集群，与发现阶段确定的集群相似。

unsupervised -利用机器学习发现肺动脉高压的独特免疫表型



发现队列：

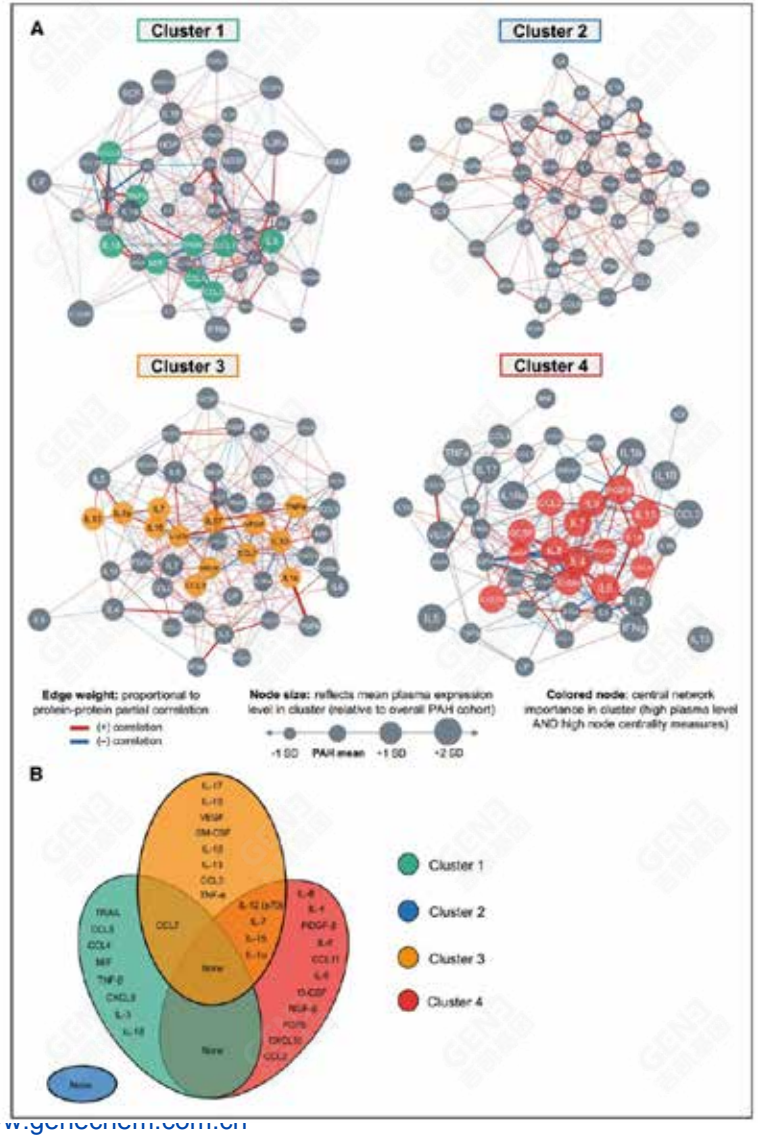
（1）共识聚类

确定了4个病人集群，每个集群都表达了独特的血液蛋白质组免疫特征（图A）

- 集群1、3和4: 与循环炎症相关， 172人（61.2%）
- 集群2: 低细胞因子，与健康对照组相似， 109名（38.8%）
- 集群4: 几种免疫介质的水平最高，类似的蛋白质在第3组中被较小幅度上调
- 集群1: 一组完全不同的细胞因子和因素被上调

主成分分析证实了集群之间的蛋白质组差异（图B）

unsupervised -利用机器学习发现肺动脉高压的独特免疫表型

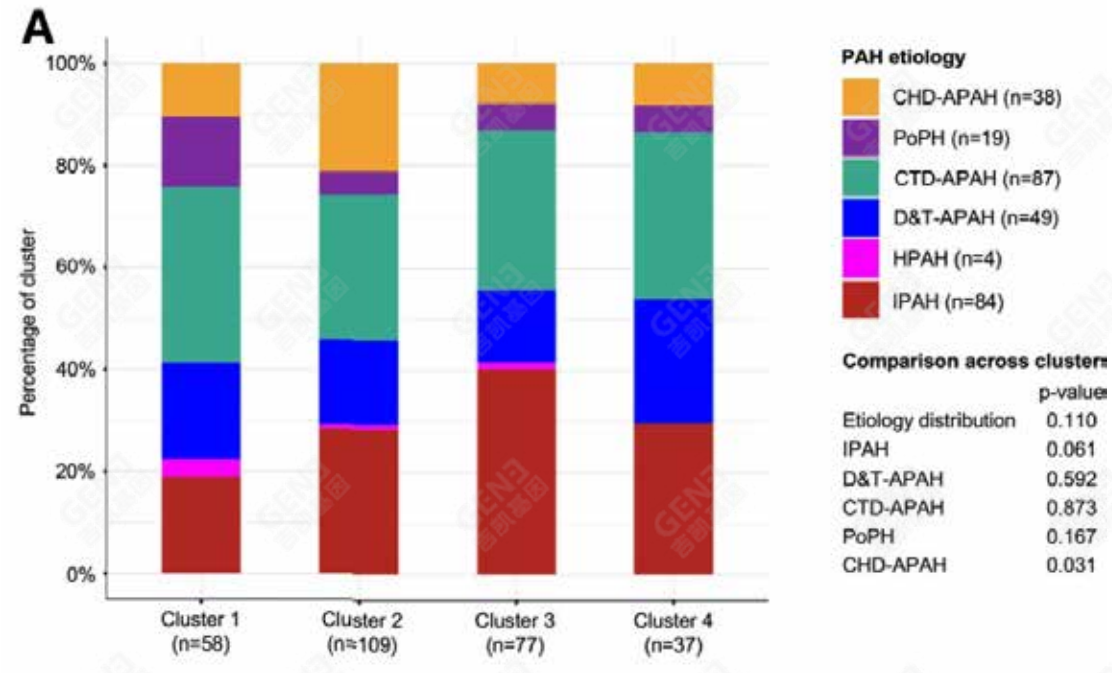


（2）确定集群的中心蛋白

部分相关网络表明，每个聚类都显示了一个特征性的蛋白质组网络结构（图A）,聚类也被独特的中心网络特征所区分（图B）- 由具有高网络中心度的上调蛋白所定义

- 集群1: 上调9个蛋白: TRAIL, CCL5, CCL7, CCL4, MIF, TNF-β
- 群集2: 没有确定中心网络特征
- 集群3: 上调13个蛋白: IL-12, IL-17, IL-10, IL-7, VEGF, IL-15
- 集群4: 上调15个蛋白: IL-8, IL-4, PDGF-β, IL-6, CCL11, and IL-9

unsupervised -利用机器学习发现肺动脉高压的独特免疫表型

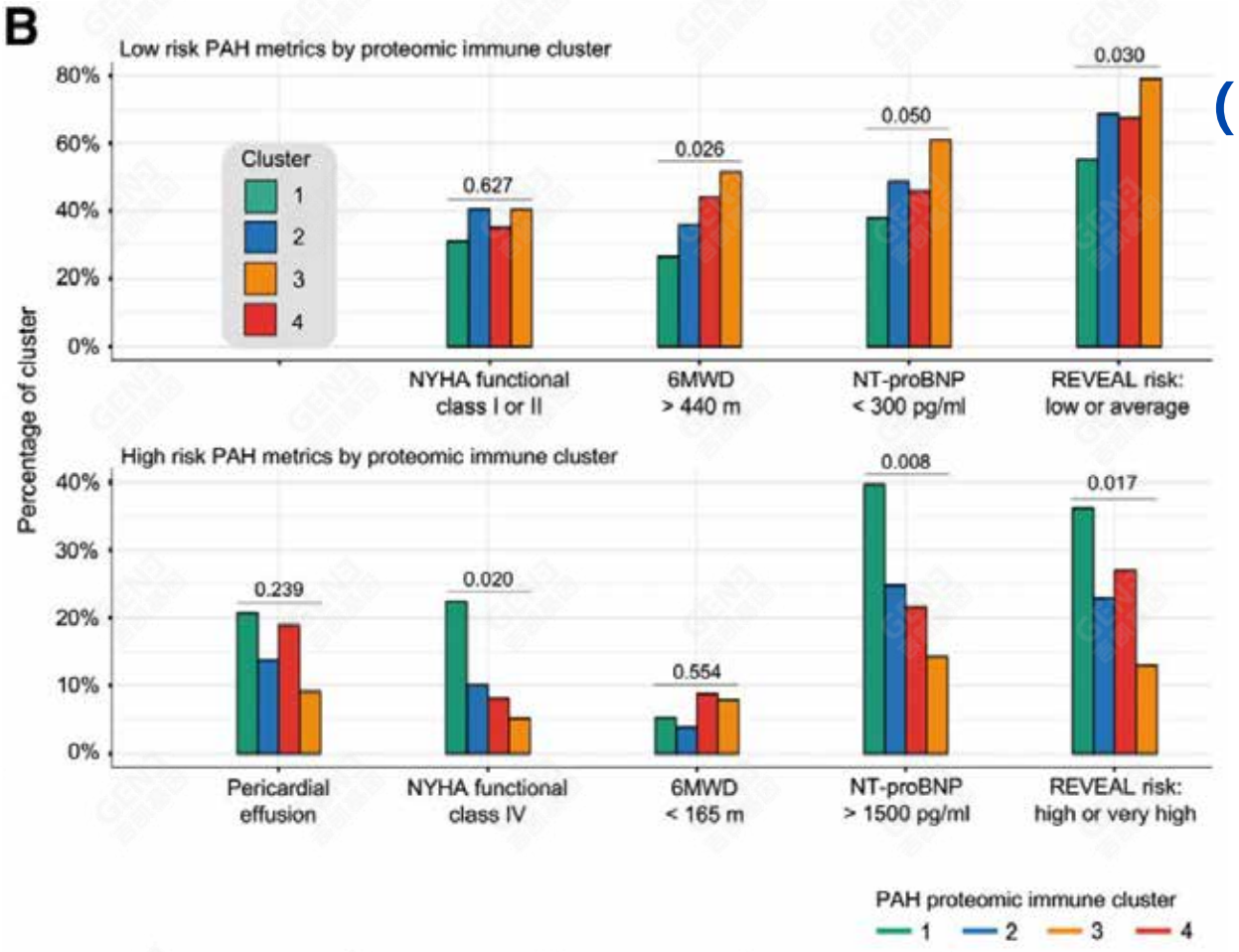


（3）比较各集群的临床特征

免疫集群与PAH亚型和病程无关

PAH临床亚型的总体分布在各群中相似（图A）。尽管特发性PAH在集群3中比例较高，先天性心脏病相关的PAH在第2组中更常见，但差异幅度相对较小。

unsupervised -利用机器学习发现肺动脉高压的独特免疫表型

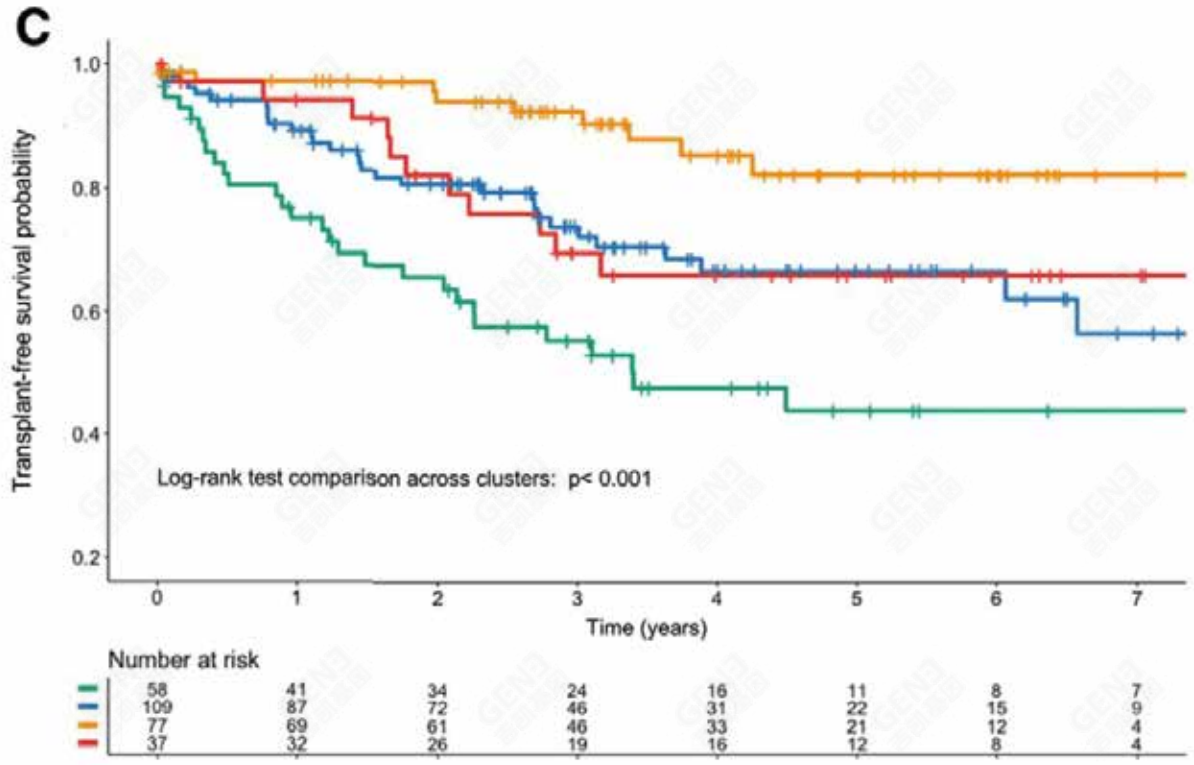


（3）比较各集群的临床特征

免疫集群具有不同的临床风险特征

对于PAH临床风险的多个既定代用指标，群组1是最高风险组，群组3是最低风险组，群组2和4是中等风险组

unsupervised -利用机器学习发现肺动脉高压的独特免疫表型



（3）比较各集群的临床特征

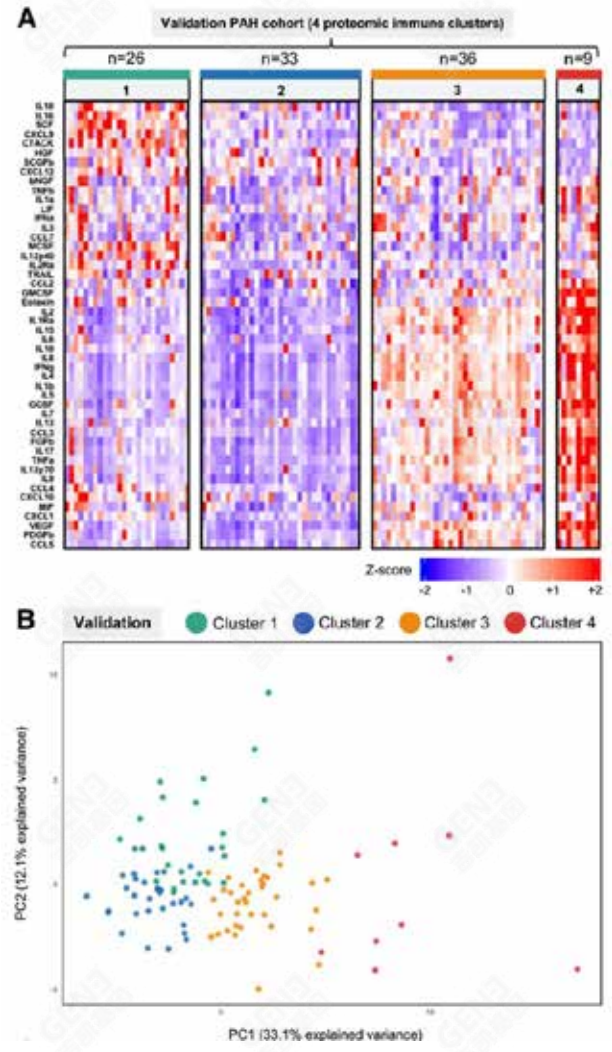
不同免疫组群的长期结果不同

从血浆采样开始，发现队列：

患者被随访的时间中位数为3.0年，
62名受试者死亡，
17名受试者接受了移植手术

5年后，Kaplan-Meier估计的无移植生存率在集群3中最高，集群1最低，集群2和集群4居中。

unsupervised -利用机器学习发现肺动脉高压的独特免疫表型



在验证队列中重新应用无监督共识聚类法

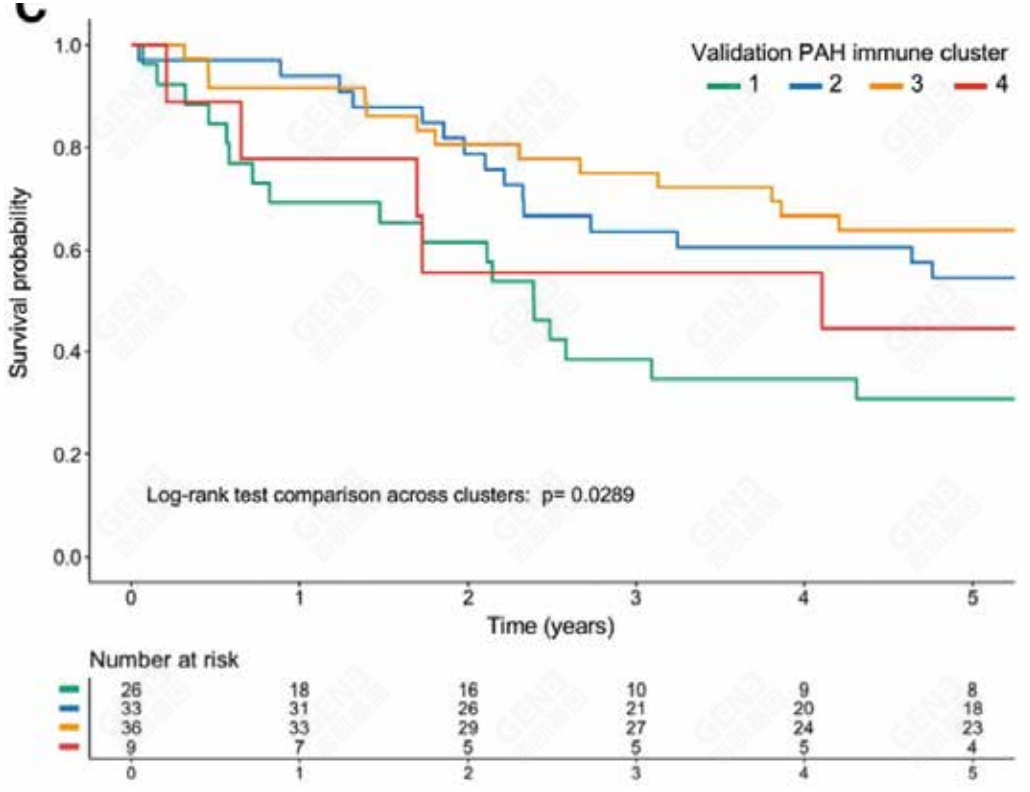
蛋白质组免疫集群的外部验证

无监督机器学习在验证队列中也发现了4个蛋白质组免疫集群，分子谱与之前相似（图A）。

集群2：具有低水平的细胞因子，
集群1、 3、 4：具有类似蛋白质的上调，这些蛋白质是发现队列中相应组别的特征。

验证集群的分离方式与发现队列中观察到的相似（图B），

unsupervised -利用机器学习发现肺动脉高压的独特免疫表型



在验证队列中重新应用无监督共识聚类法

验证群的生存差异与在发现群中观察到的相同，Kaplan-Meier估计的5年生存率在集群3中最好，集群1最差，集群2和集群4居中。

玩转转录组
——转录蛋白多组学专题讨论会

5/19

14：00-14：45
转录组测序---你想要的转录组测序问题都在这里了
童丹丹　博士　NGS线科研顾问

14：45-15：30
转录组+蛋白质组整合分析：如何助力更深机制探索，让科研不再难？
夏红蕾　质谱线售前科研顾问

扫码报名
讲座形式：微信群直播

THANK YOU !!

扫描二维码，获取更多资料

2022-5-11

上海吉凯基因医学科技股份有限公司
SHANGHAI GENECHEM CO.,LTD.