

第十章 表型组学

陈迪俊， 蒋杭进， 黄涛， 陈超

中南大学

- 第一节 表型组学概述
- 第二节 表型组的测量技术
- 第三节 表型组的数据分析
- 第四节 细胞与类器官表型组学
- 第五节 人类表型组
- 第六节 动植物表型组学
- 第七节 未来展望

重点内容

- 表型组学的定义和范围;
- 数据采集和处理;
- 数据分析方法;
- 基因-表型关联研究;
- 表型组学在人类与动植物的应用。

难点内容

- 复杂数据的处理与分析;
- 多学科知识的融合;
- 大数据管理与存储;
- 数据的标准化与规范化;
- 生物学意义的解释。



09 ▶

人类表型组微观与整体的复杂关联及其机制解密

破解宏观 - 微观表型之间的复杂关联与互作机制一直是生命科学前沿探索的方向。以破解宏观—微观表型之间的复杂关联与互作机制为目标，通过大规模表型全景精密测量与数据系统解析，贯穿和破解基因—表型—环境、宏观—微观表型之间以及宏观表型项目之间的关联与调控机制，有助于对生物系统的全面理解，最终帮助实现对疾病与健康的精准干预。

推荐单位：中国认知科学学会

第一节 表型组学概述

什么是**表型**？

表型包括生物体从分子到个体的多尺度空间特征，如形态、结构、功能、行为、生理、代谢等方面，以及从胚胎发育到衰老死亡的全生命周期时间特征，可以通过观察和测量获得。

什么是**表型组学**？

表型组学是一门研究生物体在不同环境条件下表现出的所有性状特征（即表型）的学科。

按研究对象分类

- 人类表型组学 (human phenomics)
- 植物表型组学 (plant phenomics)
- 动物表型组学 (animal phenomics)
- 生态学与环境表型组学 (ecological and environmental phenomics)

按数据类型分类

- 分子表型组学 (molecular phenomics)
- 细胞表型组学 (cellular phenomics)
- 图像表型组学 (image phenomics)
- 功能表型组学 (functional phenomics)

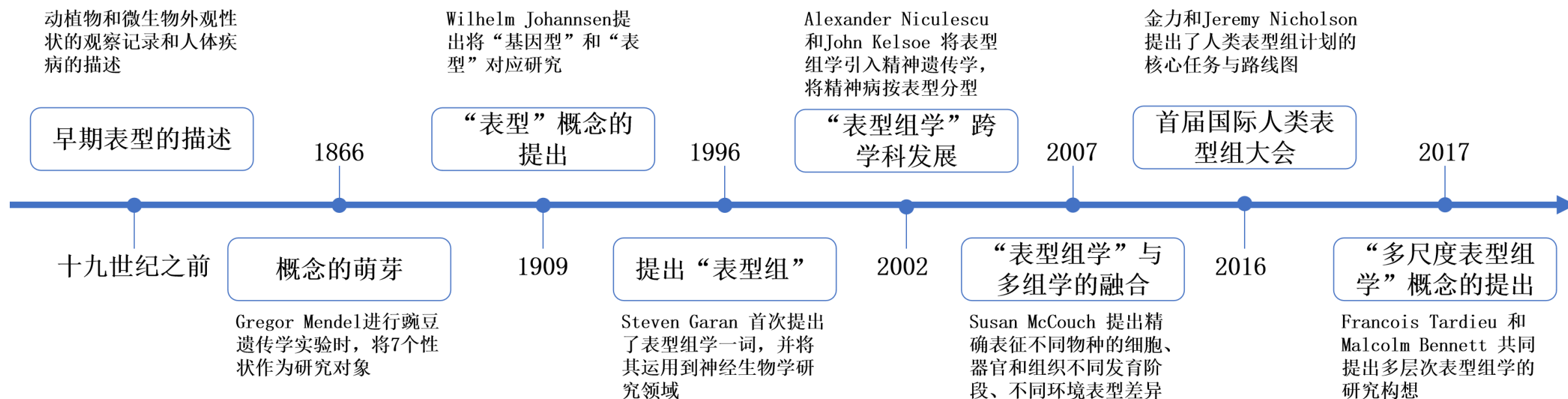


第一阶段：早期表型的描述（古代）

第二阶段：遗传学的兴起（19世纪末至20世纪初）

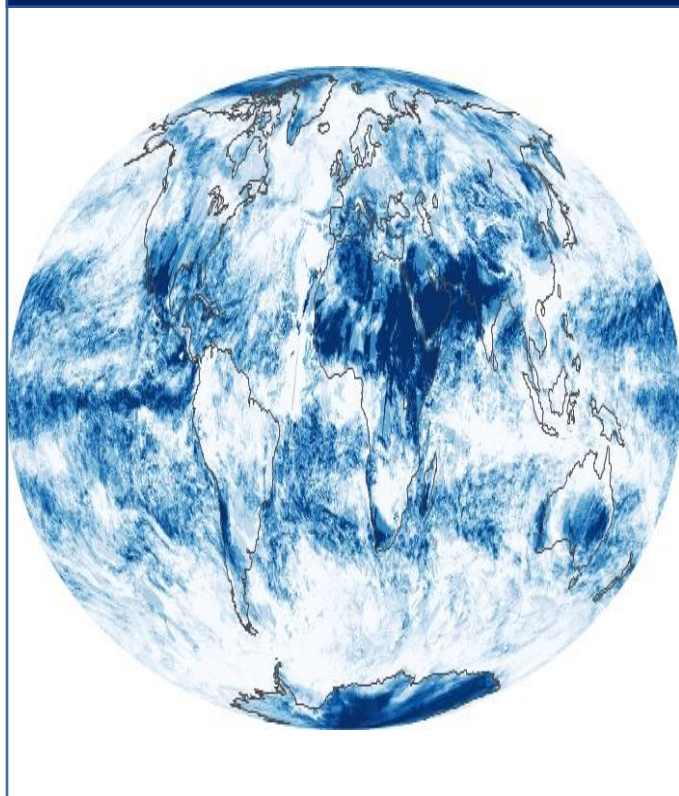
第三阶段：表型组学的崛起（20世纪中后叶）

第四阶段：表型组学的蓬勃发展（21世纪）

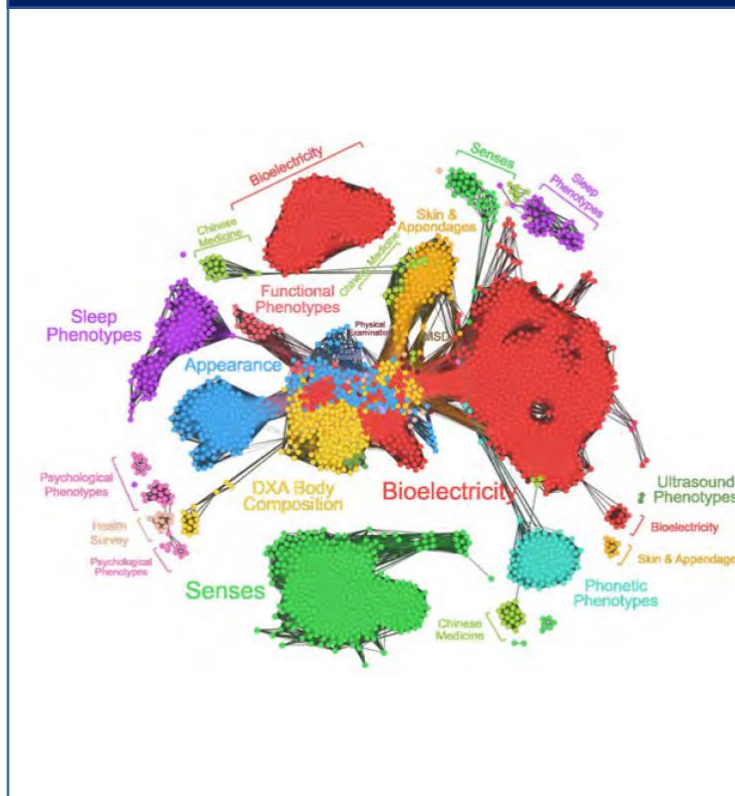


表型组学发展历程

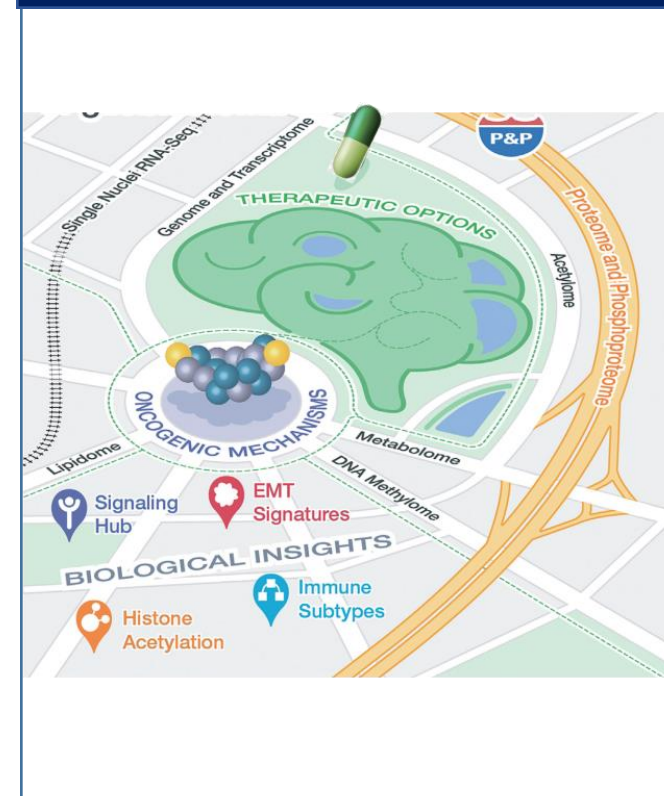
人类的复杂生命系统



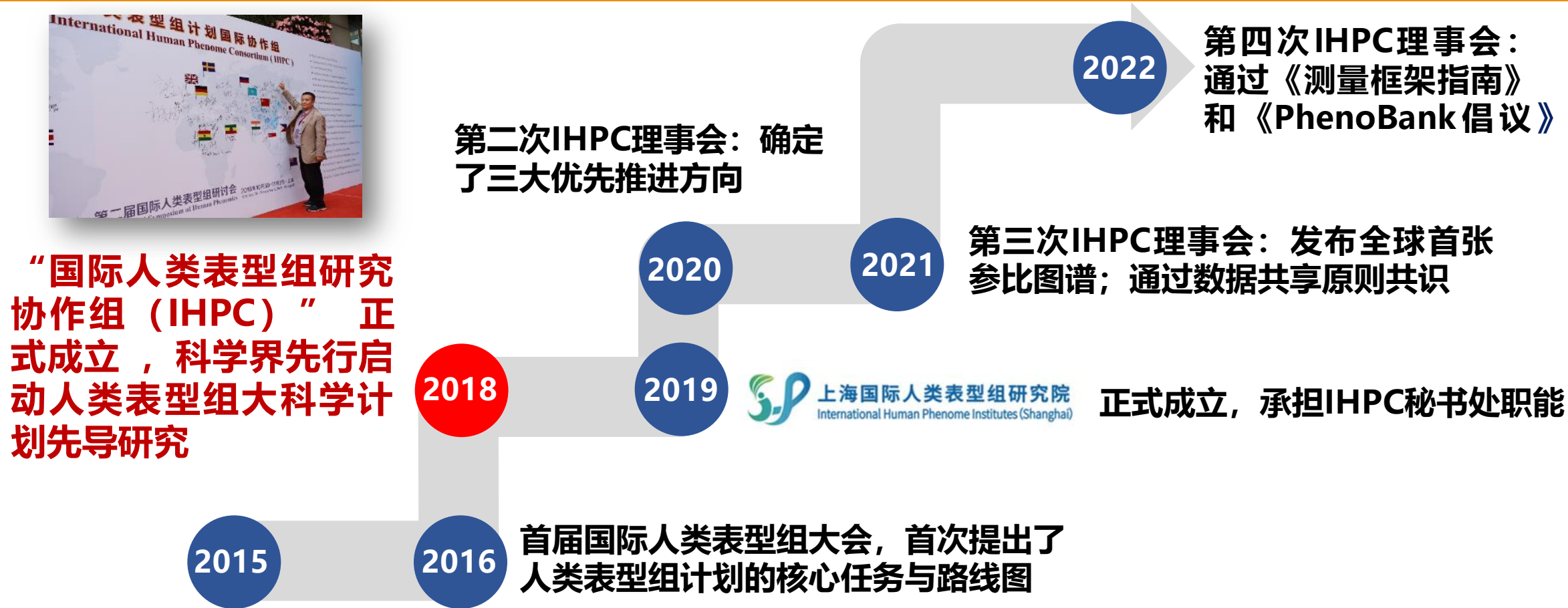
人类表型组网络



人类表型组导航图

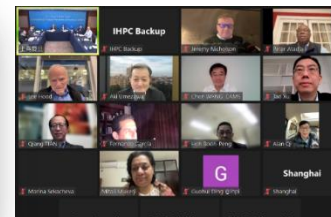


图片来源：金力报告



香山科学会议上最早正式倡议发起“国际人类表型组计划”

图片来源：金力报告



人类表型组国际大科学计划

分布型的国际大科学计划模式：统一技术标准，共享数据资源

美洲人群
1万人

欧洲人群
1万人

中国人群
1万人

非洲人群
1万人

东南亚人群
1万人

I 期：上海示范人群1千人全景测量、1万人应用示范测量

创新的
理论体系

领先的
核心技术

公认的
国际标准

卓越的
领军团队

共享的
数据平台

实现国际引领、推动范式变革、加速创新策源

第二节 表型组数据来源

相关技术/维度	研究方法
分子水平	转录组学：基因表达、可变剪接等高通量定量分析。 高级质谱：蛋白质的高通量定量分析。 蛋白质芯片技术：连续动态监测蛋白质的状态变化。 代谢组学：结合核磁共振和质谱技术，精确测定代谢物种类和浓度，探索代谢途径。 微生物组：全面描绘人体微生态系统成分。 表观组学：包括全基因组甲基化分析、染色质免疫沉淀以及单细胞表观组学等高分辨率下检测表观遗传学变化。
细胞/组织水平	流式细胞术：同时对数万个细胞的多个特征进行定量和定性评估，深入分析细胞表型和信号通路。 高通量微电极阵列：同时记录多细胞电信号，提供高时空分辨率数据。 单细胞测序：分析单细胞基因表达特征，揭示细胞间异质性。 空间转录组：结合空间信息，研究组织内基因表达的空间模式和细胞间相互作用。 高通量筛选：快速识别和分析大规模样本中的目标分子或细胞，应用于药物发现和基因分析。 O2k能量代谢分析：通过测量氧气消耗率实时评估细胞呼吸和能量产生，用于能量代谢研究。

相关技术/维度	研究方法
图像水平	高分辨率显微镜：如荧光显微镜、共聚焦显微镜，用于观察细胞和微观结构，深入了解细胞功能和疾病机制。 三维细胞成像：利用共聚焦和多光子显微镜获取三维高分辨率图像，揭示细胞结构和相互关系。 医学影像技术：如核磁共振成像、电子发射断层扫描等，提供深层次组织结构的清晰、精确视图，用于疾病早期诊断和病理研究。 可见光成像：用于监测动物的行为、生活习性及生命周期，或者捕捉植物外观和结构信息，观察植物生长发育和形态变化。 红外成像：利用红外光波段捕捉热辐射或反射信号，以研究植物的生理、健康和化学状态，或者监测动物体温分布、活动模式和生理适应性。 荧光成像：基于植物组织中荧光素分子的光能吸收和荧光发射，通过特定设备转化为图像，研究植物的生物过程。
功能水平	传感系统：包括可穿戴设备和非接触式设备，如智能手环、红外成像技术，实时监测人或者动物的生理、行为信号及环境互动。 心理测量工具：通过问卷、量表和任务测验等检测认知、人格、幸福感、心理应激，以及焦虑、抑郁状态等个体在心理特征和行为模式上的可观察表现。 中医生理特征：对体质、病症、脉象、舌象等方面进行系统的描述和分类。 脑电波检测：分析计算信号在不同频率上的功率分布，识别不同频段的脑电活动。 皮肤及附属物检测：测量皮肤基础生理指标，如温度、pH、油脂含量、色素沉着、肤质、表皮含水量、厚度、纹理、弹性，以及附属器指甲、毛发等的形态、功能等特征。以及对皮肤附属器表型测量，如脱发、毛发性状、粗细、密度、汗腺密度及功能、指甲形态、指甲厚度、指甲纹理及甲襞微循环等。

特点	描述
多样性	涵盖从人类到微生物的各种生物体的形态学、生理学等表型特征
高维度	包含多种类型的特征或变量（数值型、分类型、文本型等），导致数据结构高维复杂
大规模	涉及成千上万乃至数百万的样本，对计算能力和存储资源要求极高
时序性	包含随时间变化的数据，如生长曲线、代谢活性等，要求跟踪时序变化
异质性	数据来源多样和技术差异带来的质量、精度、可信度差异，需要仔细清洗和标准化
数据关联性	特征间存在密切关联，如基因型与表型、环境因素的相互作用
复杂性	表型特征受基因型、环境等多因素影响，分析时需考虑这些综合因素

编码系统	主要用途	主要实体种类	规模	使用许可
人类表型本体论 (human phenotype ontology, HPO)	标准化描述人类表型异常，支持临床诊断和生物医学研究	人类表型异常的描述，包括症状、体征、异常实验室结果等	约16,000个概念，13,000个表型条目	开源
国际疾病分类 (international classification of disease, ICD)	国际通用的疾病分类方法，用于流行病学、健康管理和临床工作	疾病、伤害、死因	约55,000个编码，分为22个章节	开源
医学系统术语标准化命名-临床术语 (systematized nomenclature of medicine - clinical terms, SNOMED CT)	提供临床术语标准，支持电子健康记录和临床决策	疾病、过程、观察结果等临床概念	约340,000个活跃概念，超过1,360,000个概念关系	需要许可
疾病本体论 (disease ontology, DO)	提供疾病分类框架，支持生物医学研究	人类疾病	约7,000种疾病概念	开源
Orphanet罕见病数据库	提供罕见病信息，支持诊断和研究	罕见病及相关基因	约7,000种罕见病，3,700种基因	部分开源

第三节 表型组的数据分析

数据预处理与质量控制

明确数据来源
识别数据类型
异常值检测与处理
评估数据缺失情况
数据填充
去除重复数据
数据标准化
混杂因素校正
数据降维
...



单模态表型组数据分析

图像表型
声音表型
菌群表型
心理表型
...

多模态表型组学数据分析

全基因组关联研究
精细定位
表达数量性状基因座
共定位分析
孟德尔随机化
全转录组关联研究
全表型组关联研究
复合表型分析
网络分析
多基因风险评分
...

常用分析方法

- (1) 明确数据来源
- (2) 识别数据类型
- (3) 异常值检测与处理
- (4) 评估数据缺失情况
- (5) 数据填充
- (6) 去除重复数据
- (7) 数据标准化
- (8) **混杂因素校正**

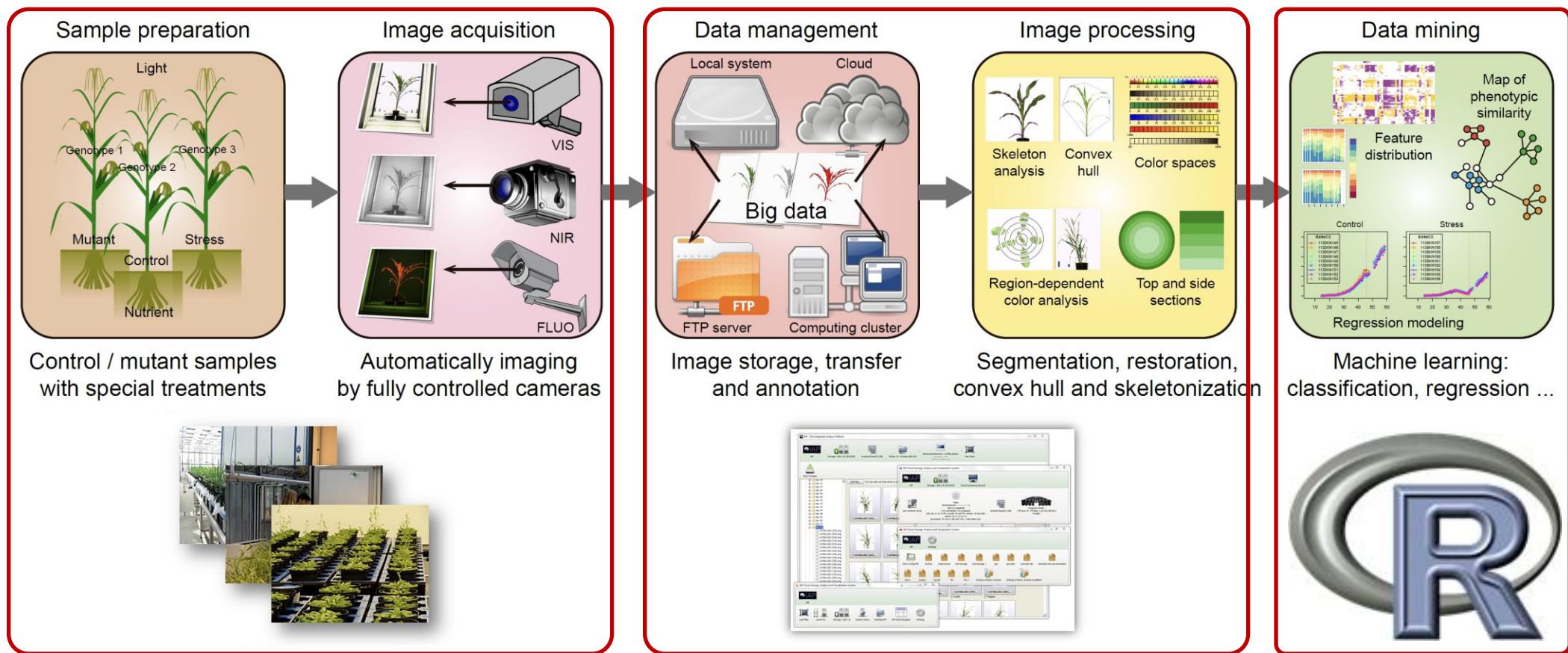
混杂因素校正

混杂变量分析 (surrogate variable analysis, SVA)

$$\begin{aligned}x_{ij} &= \mu_i + f_i(y_i) + \sum_{l=1}^L \gamma_{li} g_{lj} + e_{ij}^* \\ &= \mu_i + f_i(y_i) + \sum_{k=1}^K \lambda_{ki} h_{kj} + e_{ij}^*\end{aligned}$$

SVA通常应用于高维数据分析，比如微阵列或RNA-seq实验中的基因表达数据分析，目的是为了了解释除了研究中的生物条件以外的变异因素。

- (1) 宏观表型组数据分析 (图像、菌群、心理等表型)**
- (2) 微观表型组数据分析 (转录组、蛋白组、表观组等表型)**
- (3) 多模态数据的关联和整合性分析 (GWAS、QTL、PWAS、复合表型)**



表型平台 → 图像数据

图像分析 → 表型特征

特征分析 → 统计建模

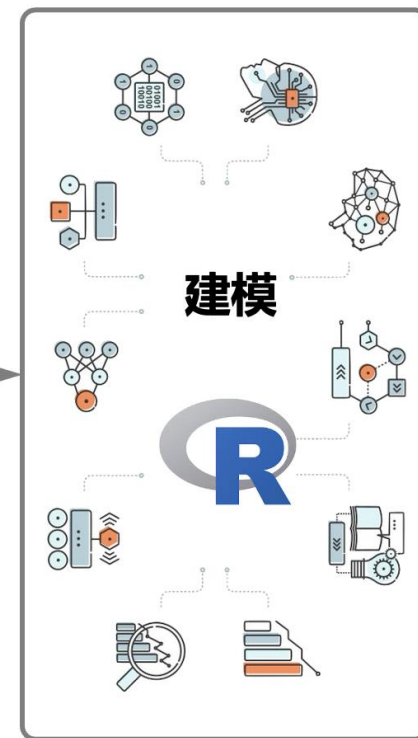
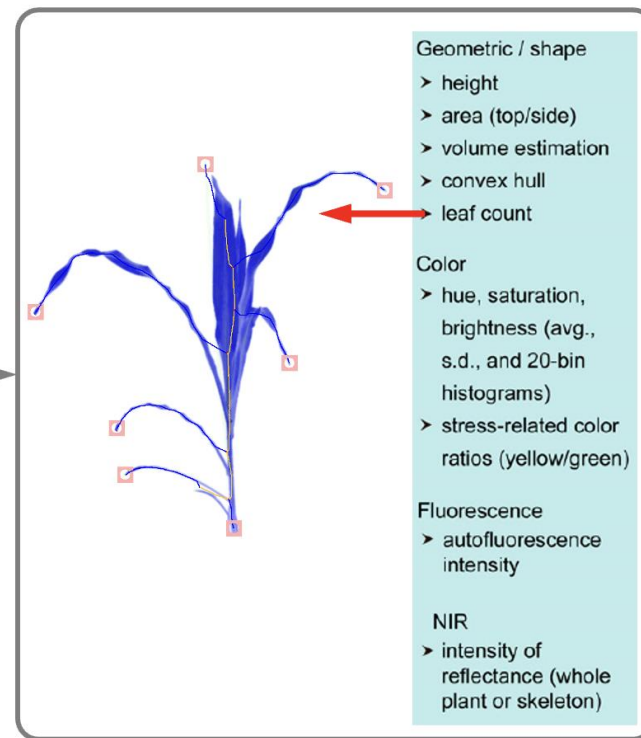
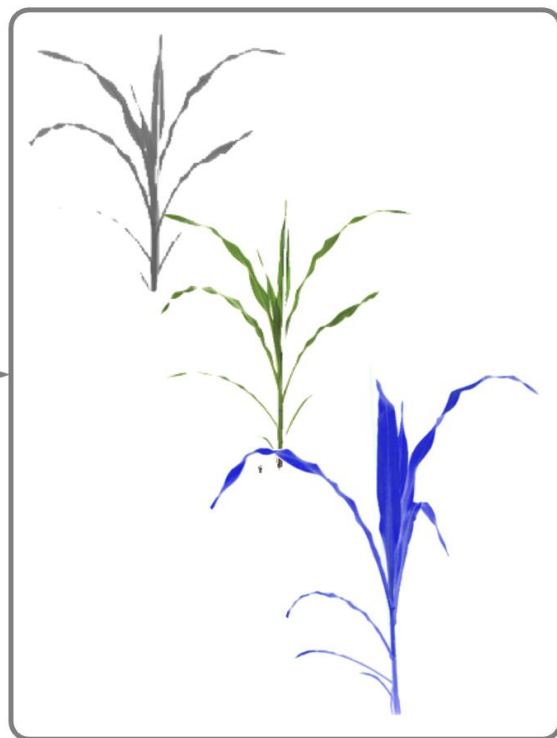
• 图像数据分析基本步骤

预处理

分割

特征提取

特征分析



菌群多样性和丰度包括Alpha多样性和Beta多样性。

Alpha多样性是衡量一个样本内部的物种多样性，常用指标包括香农指数 (Shannon Index) 和辛普森指数 (Simpson Index) 。

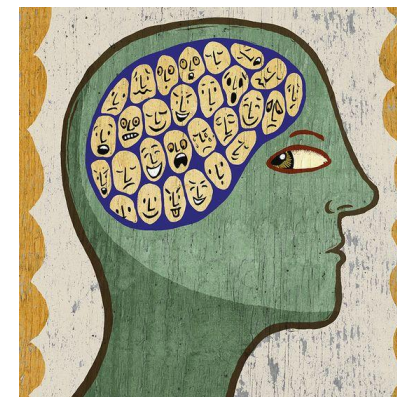
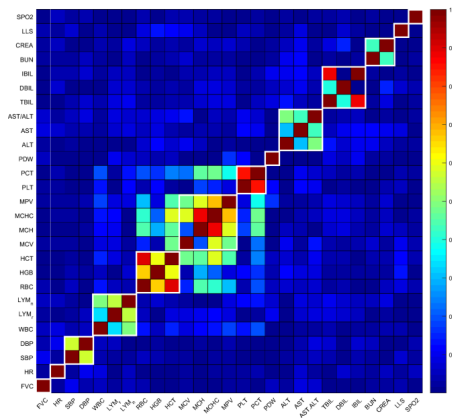
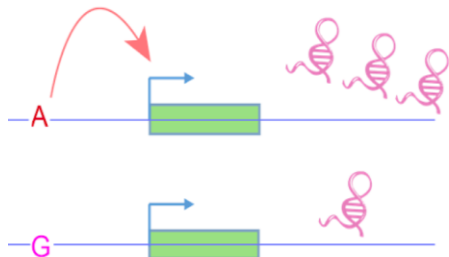
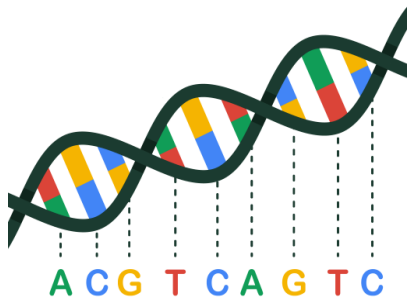
Shannon Index

$$H' = - \sum_{i=1}^S p_i \ln(p_i) \quad H' = - \sum_{i=1}^S p_i \ln(p_i)$$

Simpson Index

$$D = 1 - \sum_{i=1}^S p_i^2 \quad D = 1 - \sum_{i=1}^S p_i^2 \quad L$$

Beta多样性是比较不同样本之间的物种多样性，常用方法包括主坐标分析 (Principal Coordinates Analysis, PCoA) 和非度量多维尺度分析 (NMDS)



基因组

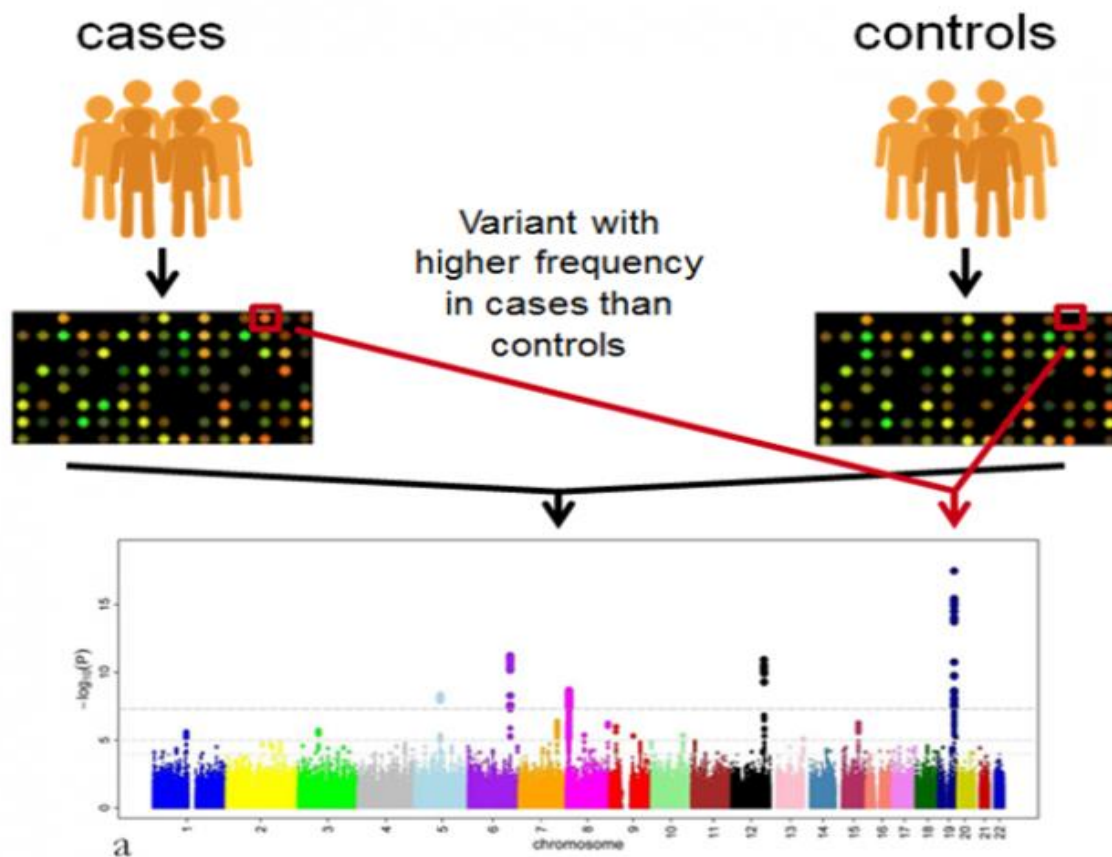
微观表型

复合/中间表型

疾病/性状

GWAS

识别特定遗传变异与表型特征的关联



拓展知识点:

- (1) 复杂性状的多基因特点
- (2) 变异类型
- (3) 连锁不平衡开展关联研究
- (4) 常见疾病-常见变异假说
- (5) 缺失的遗传度

大多数人类复杂疾病/特征是多基因的

融合遗传 Blending inheritance



离散遗传单位 Discrete inherited units



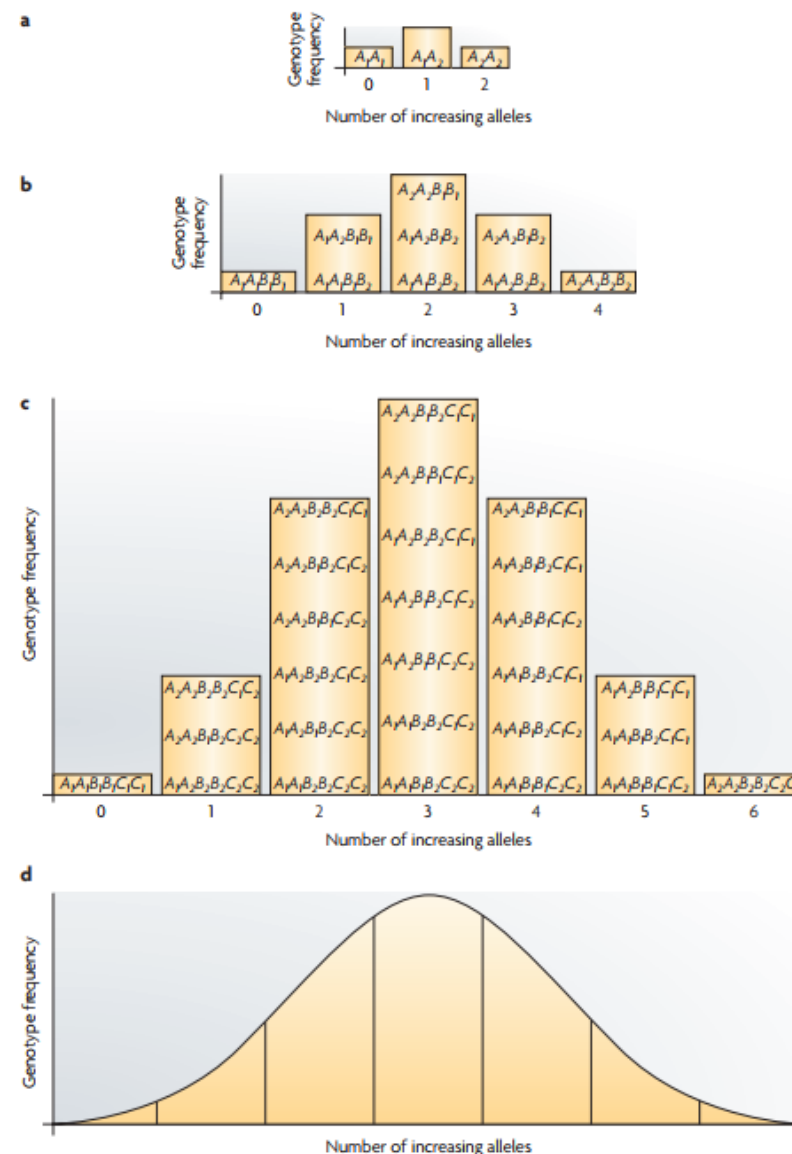
多基因性状 Polygenic trait

Opinion | Published: 27 October 2009

Common disorders are quantitative traits

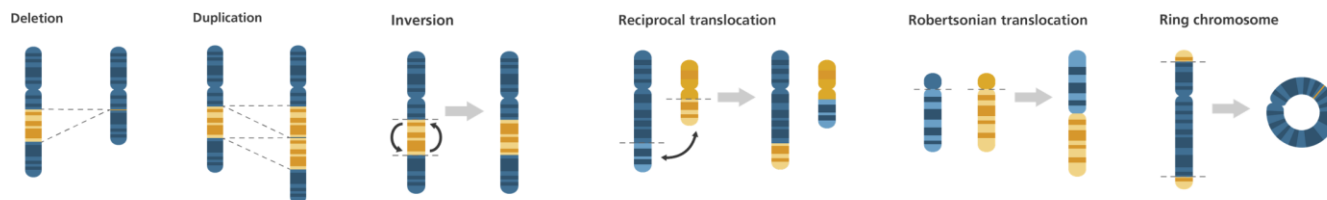
Robert Plomin , Claire M. A. Haworth & Oliver S. P. Davis

Nature Reviews Genetics **10**, 872–878 (2009) | [Cite this article](#)

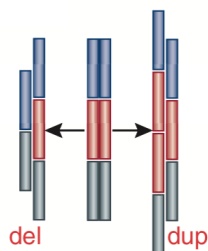


人类基因组变异

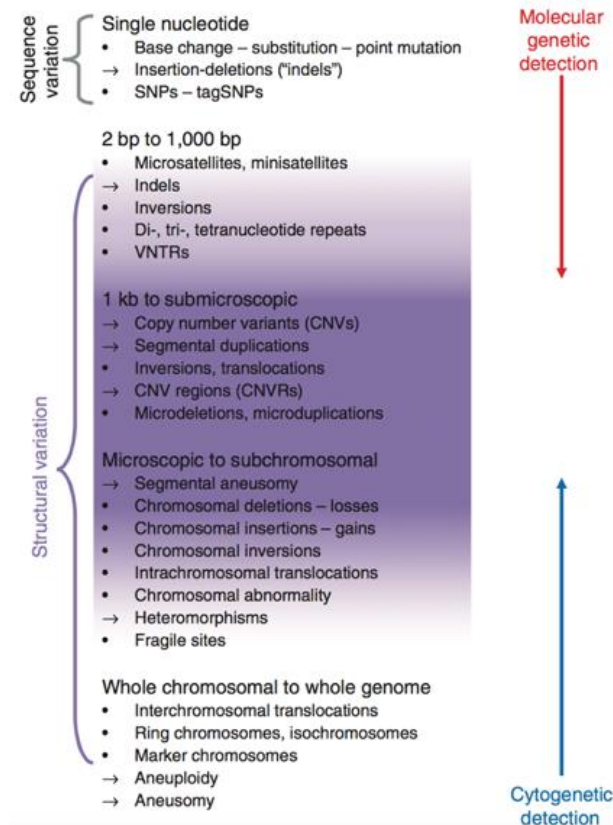
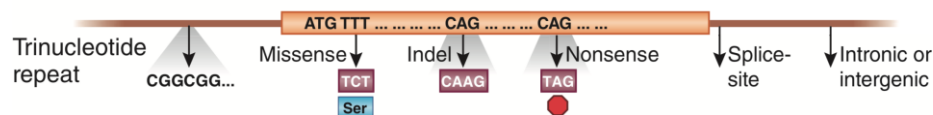
染色体异常 chromosome abnormalities



拷贝数变异 CNVs

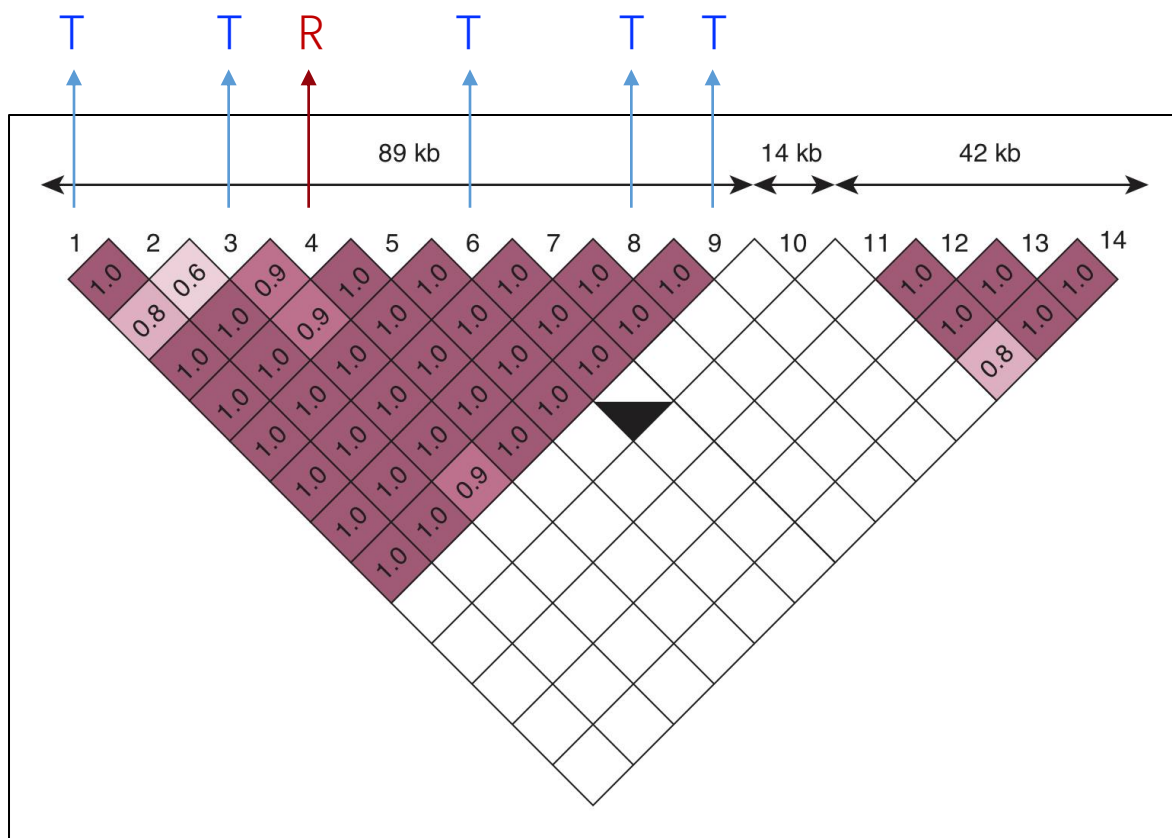


单核苷酸变异/小的插入缺失 SNV/indel



关联研究的遗传学原理 – 连锁不平衡

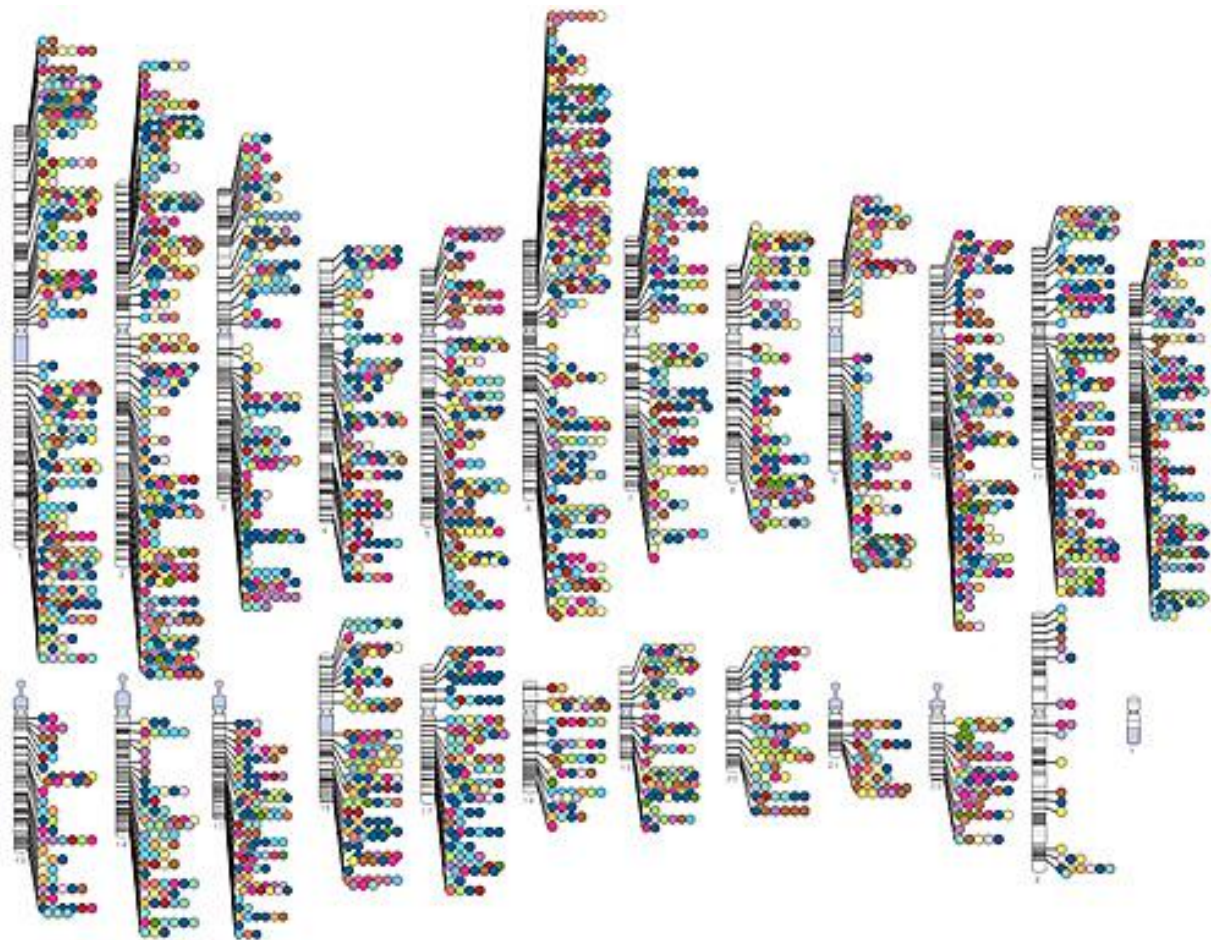
The principle of association study – Linkage Disequilibrium



T: Typed variant/marker

R: Risk variant

GWAS catalog



根据 2024 年 6 月 GWAS Catalog 的数据，共有 6,907 项研究被发表，揭示了 622,784 个遗传变异与表型特征之间的联系。

关联分析：线性回归模型

$$Y = Wa + X_s\beta_s + g + e$$

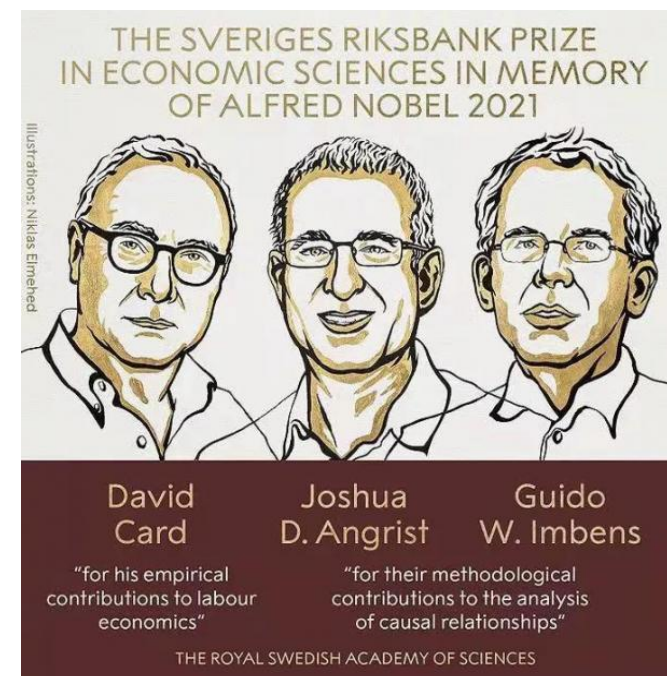
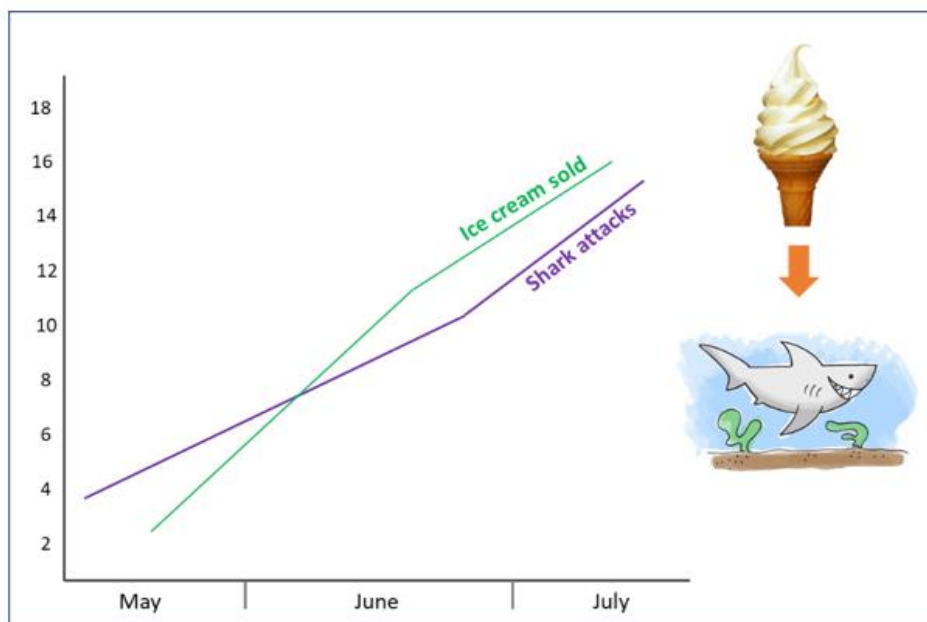
$$g \sim N(0, \sigma_A^2 \psi)$$

$$e \sim N(0, \sigma_e^2 I)$$

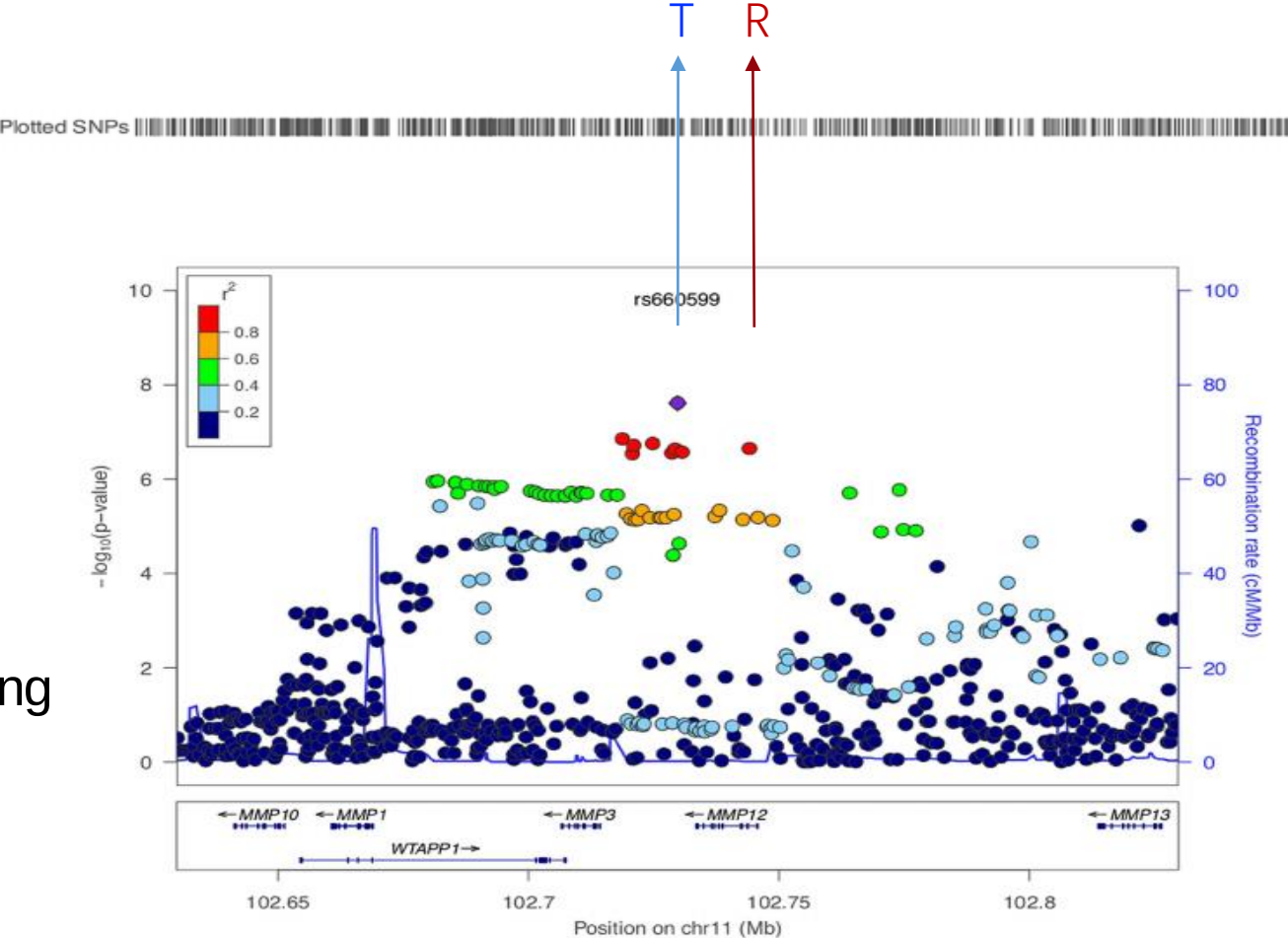
- Y 是表型值的向量，比如疾病状态或者生理特征。
- W 是协变量矩阵，如年龄、性别等。 α 是对应的效应量，表示协变量对表型值的影响。
- X_s 是SNPs下个体基因型值的向量。 β_s 是遗传变异 s 对应的固定效应量。
- g 是除了 X_s 外的其他SNP对表型加性影响， σ_A^2 表示加性遗传变异， ψ 是标准遗传关系矩阵。
- e 是剩余误差的随机效应， σ_e^2 表示剩余方差， I 是一个单位矩阵，用于加入剩余误差效应。

结果指标: β ; P-value

We want causality, not associations

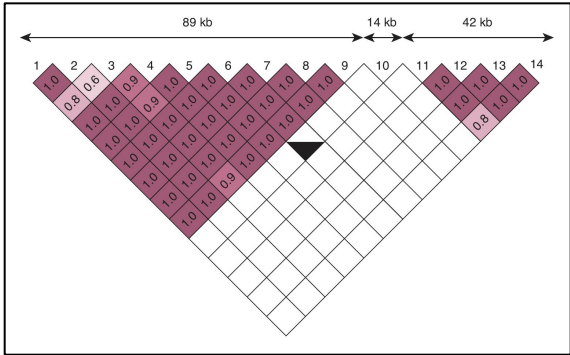


Fine mapping



T: Typed/Tag variant
R: Risk variant

Linkage Disequilibrium



连锁不平衡模块 Linkage Disequilibrium Block

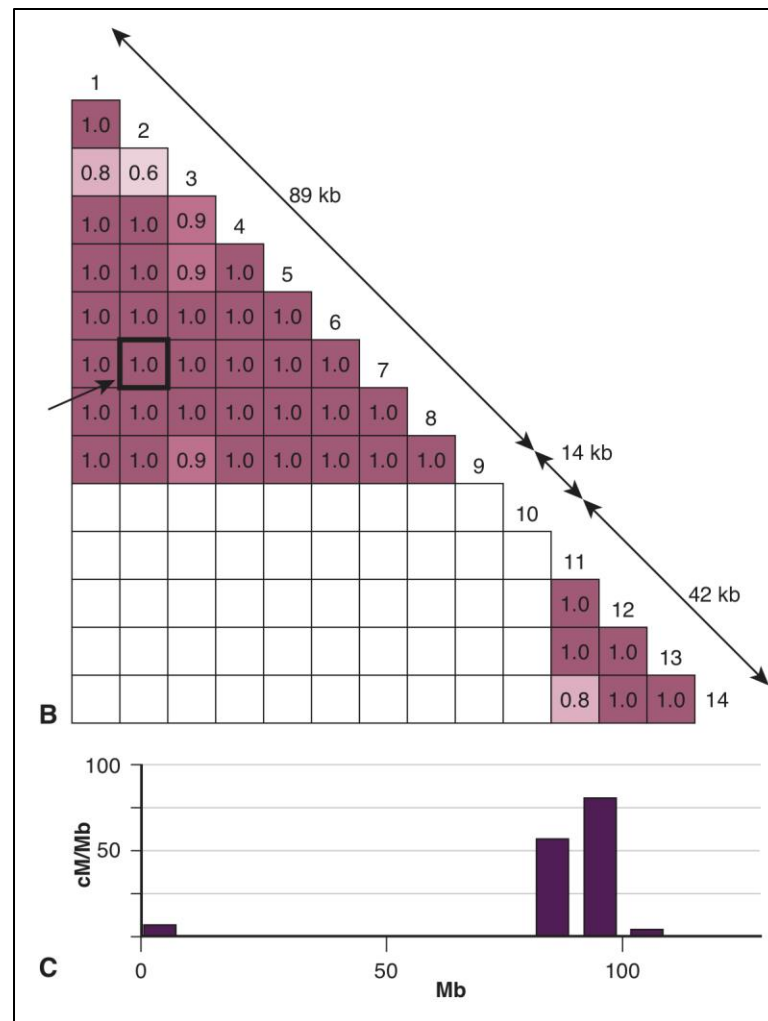
基因组中单体型分布呈模块结构，同一单体型模块中的SNP之间连锁不平衡

连锁不平衡: 朋友 or 敌人?

优点:

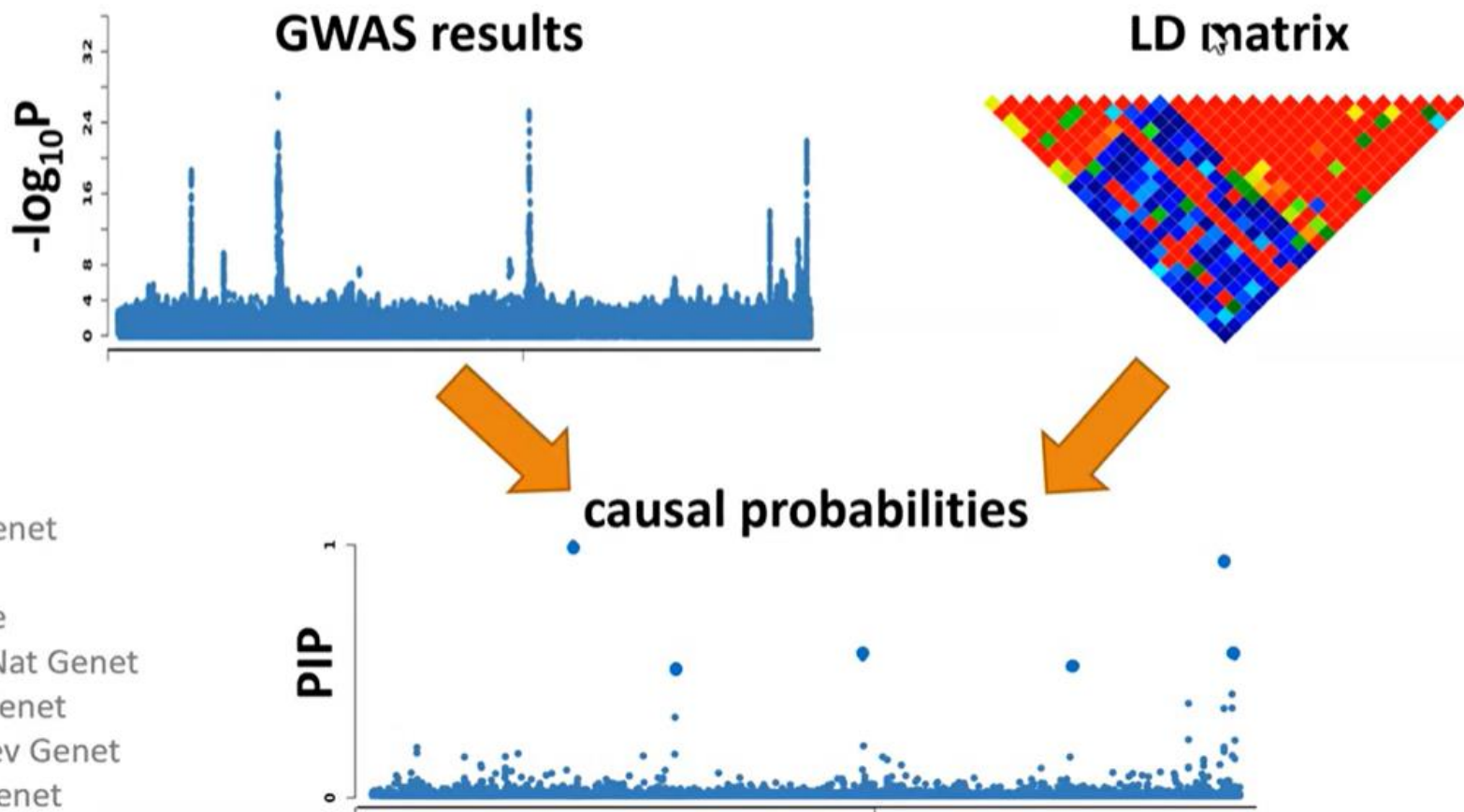
- ① 同block选择一个tag SNP进行检测，降低GWAS检测成本，提高效率。
- ② LD模块的存在提高了GWAS发现因果SNP的可能性。

缺点: 和致病位点关联的位点较多且关联信号强，难以分辨哪个是真正的致病变异。



如何通过Fine-mapping寻找致病位点?

样本量；基因组的全面覆盖



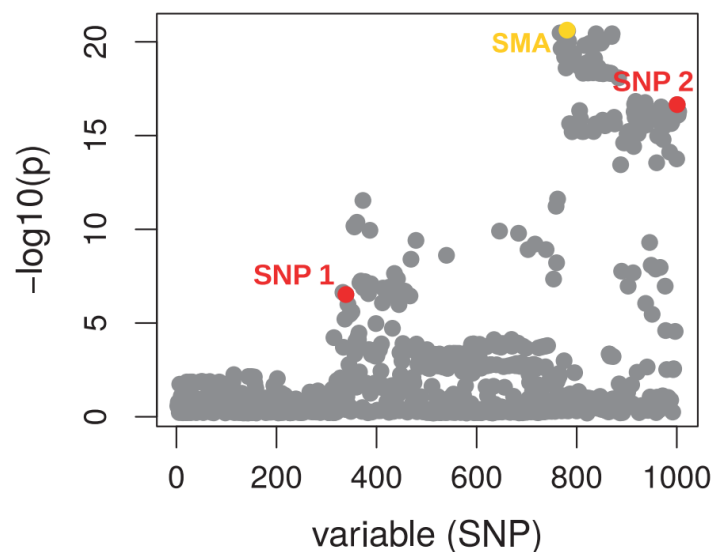
Maller *et al.* 2012 Nat Genet
Farh *et al.* 2015 Nature
Huang *et al.* 2017 Nature
Mahajan *et al.* 2018a,b Nat Genet
Westra *et al.* 2018 Nat Genet
Schaid *et al.* 2018 Nat Rev Genet
Ulirsch *et al.* 2019 Nat Genet

[nature](#) > [nature reviews genetics](#) > [review articles](#) > [article](#)

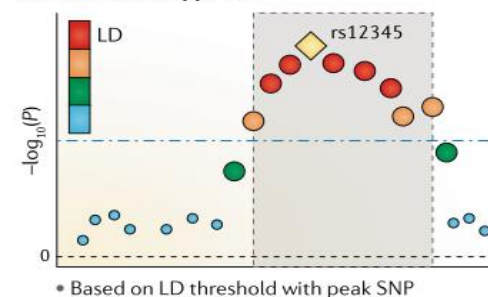
Review Article | Published: 29 May 2018

From genome-wide associations to candidate causal variants by statistical fine-mapping

[Daniel J. Schaid](#) , [Wenhan Chen](#) & [Nicholas B. Larson](#)

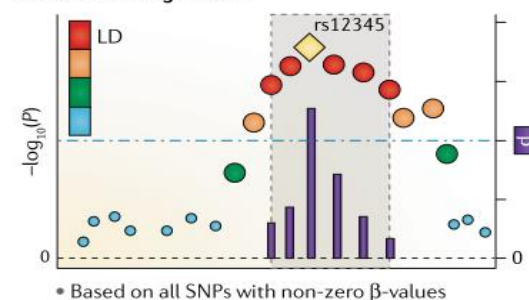


A Heuristic LD approach



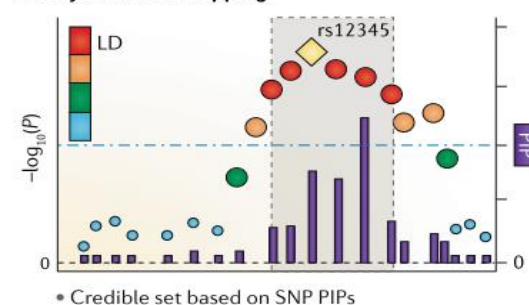
启发式LD算法

B Penalized regression



惩罚回归

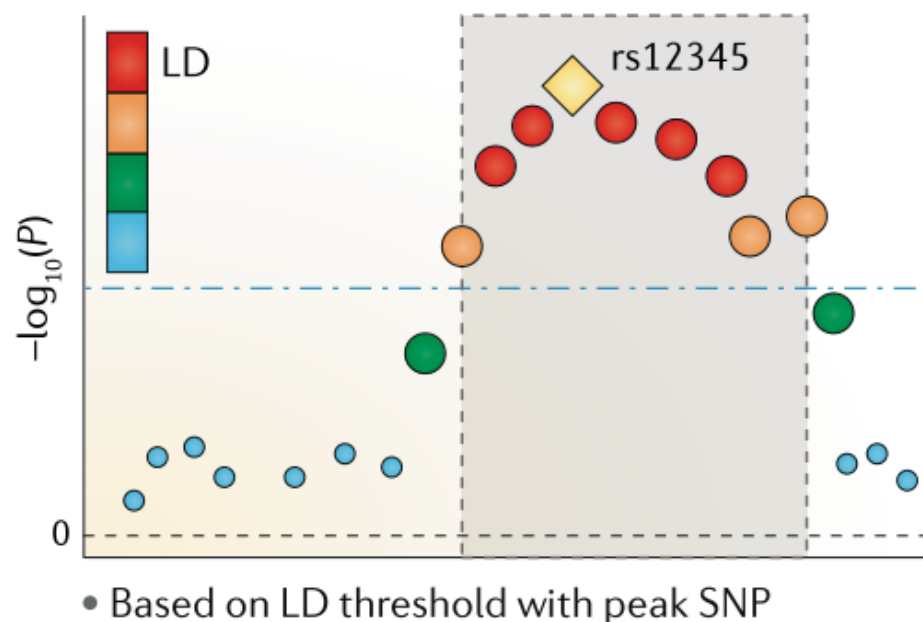
C Bayesian fine-mapping



贝叶斯精细定位

启发式LD算法 Heuristic methods

A Heuristic LD approach

**原理：**

来自 GWAS 的最显著 SNP 周围，具有强 LD 关系的 SNPs 可能具有重要作用。

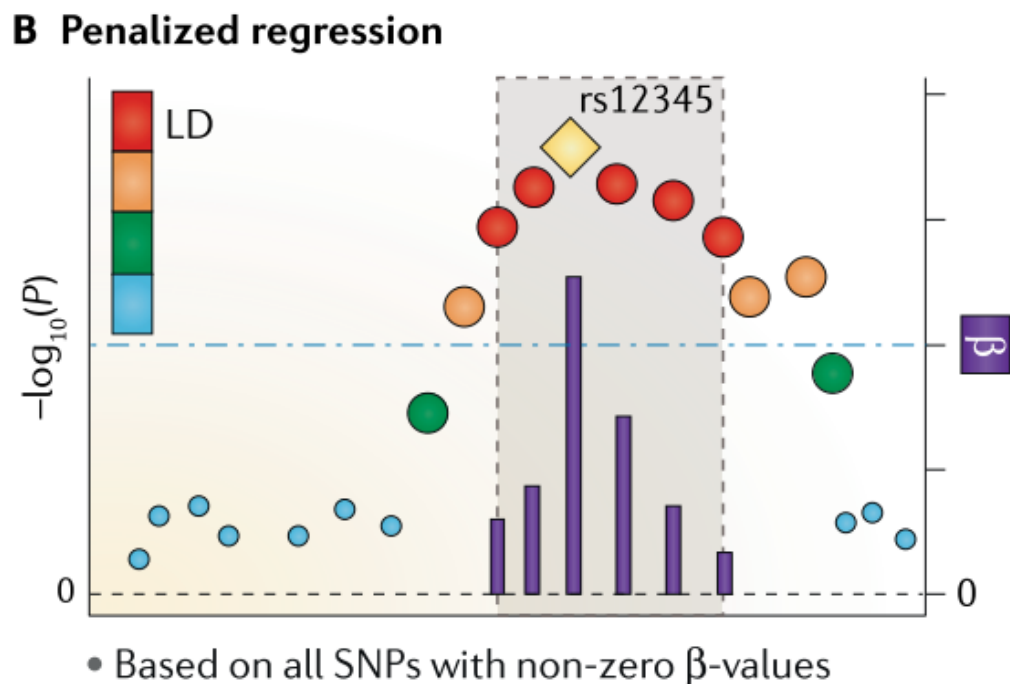
方法：

1. 根据与 tag SNP LD关系 (r^2) 过滤。
2. 与同一单倍型块中的 SNP选择。

局限性：

没有考虑 SNP 对性状的联合影响，没有客观衡量 SNP 是因果关系的置信度，单纯依赖于某种程度的和 SNP 之间相关性的阈值。

惩罚回归 Penalized regression



惩罚回归方法选择所有效应 (β , 右轴) 未被缩减为零的SNP。

原理:

使用回归模型来联合分析一个区域内的所有SNP，但大量SNP和SNP之间的高度相关性会使传统的回归模型变得不稳定。惩罚回归模型（例如LASSO、弹性网络）提供了更稳健的方法。

方法:

通过对所有变量系数进行回归惩罚，使得相对不重要的独立变量系数变为0，从而排除在建模之外。

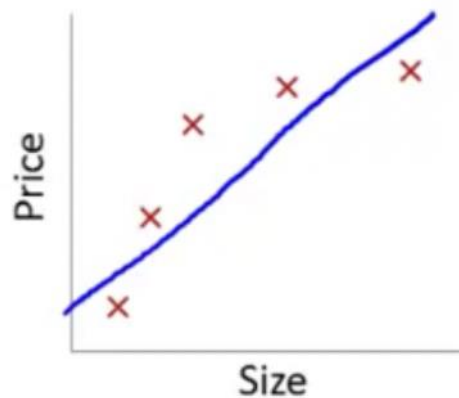
局限性:

无法排除很强LD相关的非因果SNP。

惩罚回归 Penalized regression

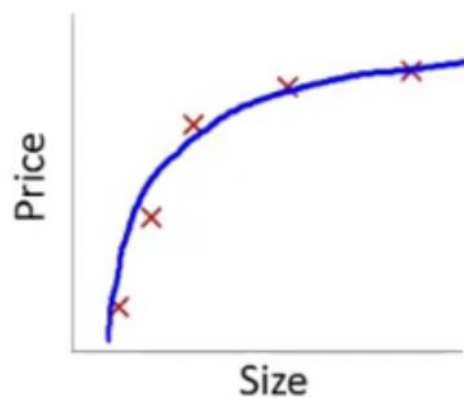
① 拟合曲线

欠拟合 underfitting



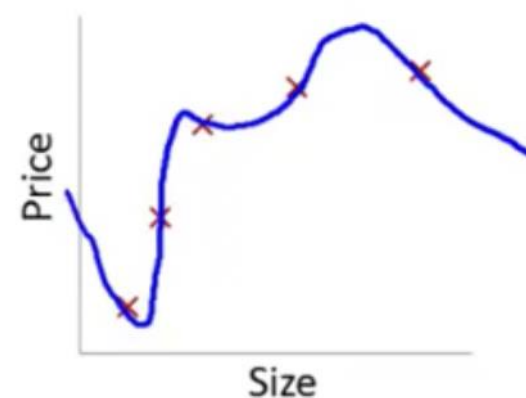
$$\theta_0 + \theta_1 x$$

合适拟合 fitting



$$\theta_0 + \theta_1 x + \theta_2 x^2$$

过拟合 overfitting



$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

② 正则化

(带入 λ , 计算拟合参数 θ)

$$J(\theta) = \frac{1}{2m} \cdot \left[\sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

$$J(\theta) = \text{MSE}(y, \hat{y}; \theta) + r\alpha \sum_{i=1}^n |\theta_i| + \frac{1-r}{2} \alpha \sum_{i=1}^n \theta_i^2$$

惩罚回归 Penalized regression

③ Lasso回归、岭回归、弹性网络回归

Lasso回归：线性回归 + L1正则化

$$\min_{\omega} \sum_{i=1}^m (y_i - \omega^T x_i)^2 + \lambda \|\omega\|_1$$

多数系数为零的稀疏解；内置特征选择

岭回归：线性回归 + L2正则化

$$\min_{\omega} \sum_{i=1}^m (y_i - \omega^T x_i)^2 + \lambda \|\omega\|_2^2$$

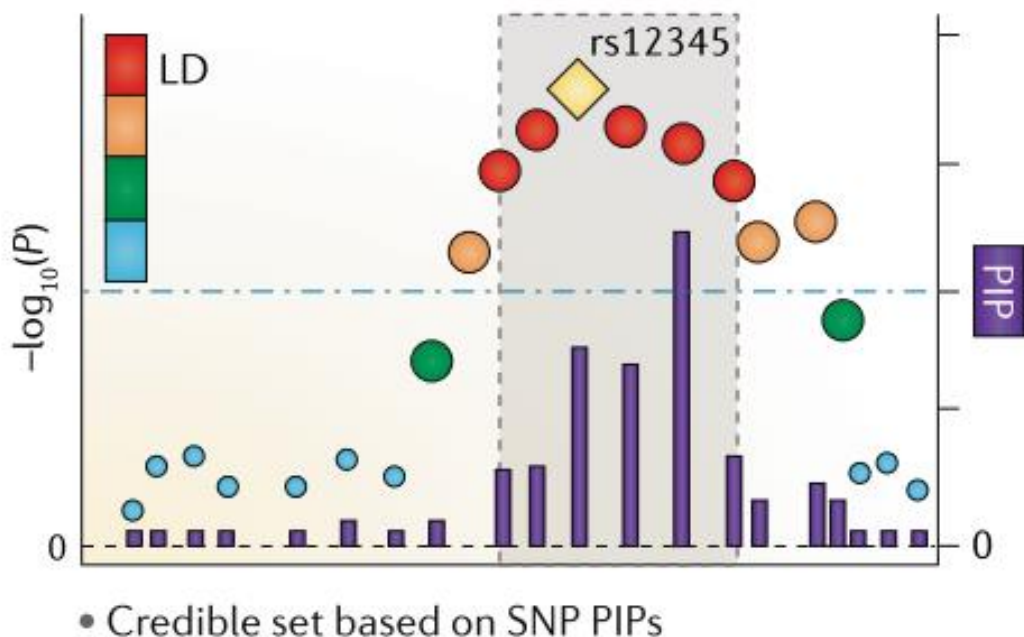
系数均匀压缩；有助于处理多重共线性

弹性网络回归：线性回归 + L1正则化 + L2正则化

$$\min_{\omega} \sum_{i=1}^m \left(y_i - \sum_{j=1}^d x_{ij} \omega_j \right)^2 + \lambda \sum_{j=1}^d |\omega_j| + \lambda \sum_{j=1}^d \omega_j^2$$

贝叶斯方法 Bayesian methods

C Bayesian fine-mapping



方法:

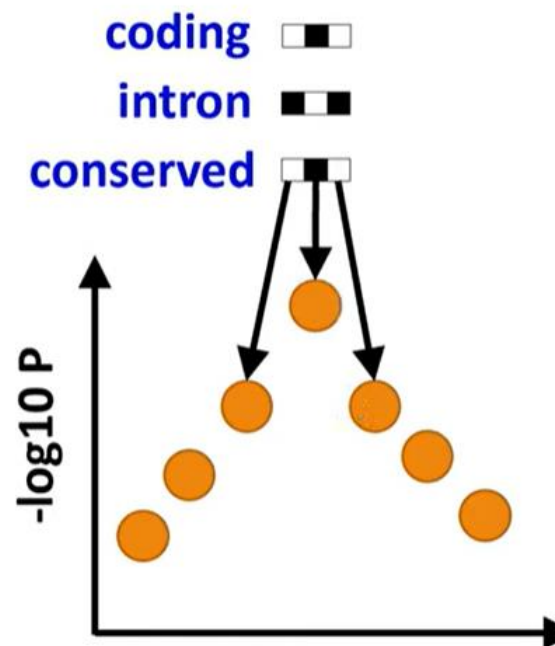
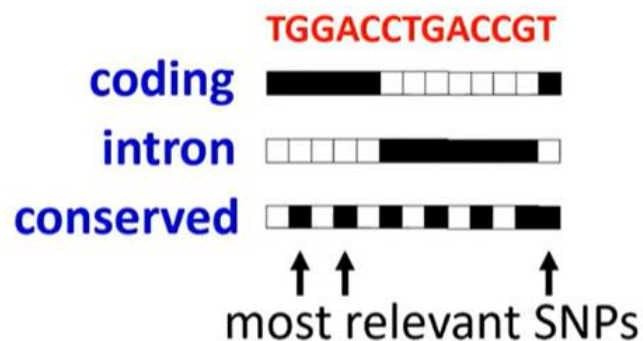
借助于特定假设或特定模型的先验概率，为感兴趣模型的概率解释。需要额外的独立的信息，如调控元件，eQTL等。

输出:

每个 SNP 的后验概率 (posterior inclusion probability, PIP)，以及确定捕获可能的因果 SNP 所需的最小 SNP 集 (Credible Set)。

这个例子表明，峰值SNP并不对应于具有最高PIP的SNP。

贝叶斯方法 Bayesian methods

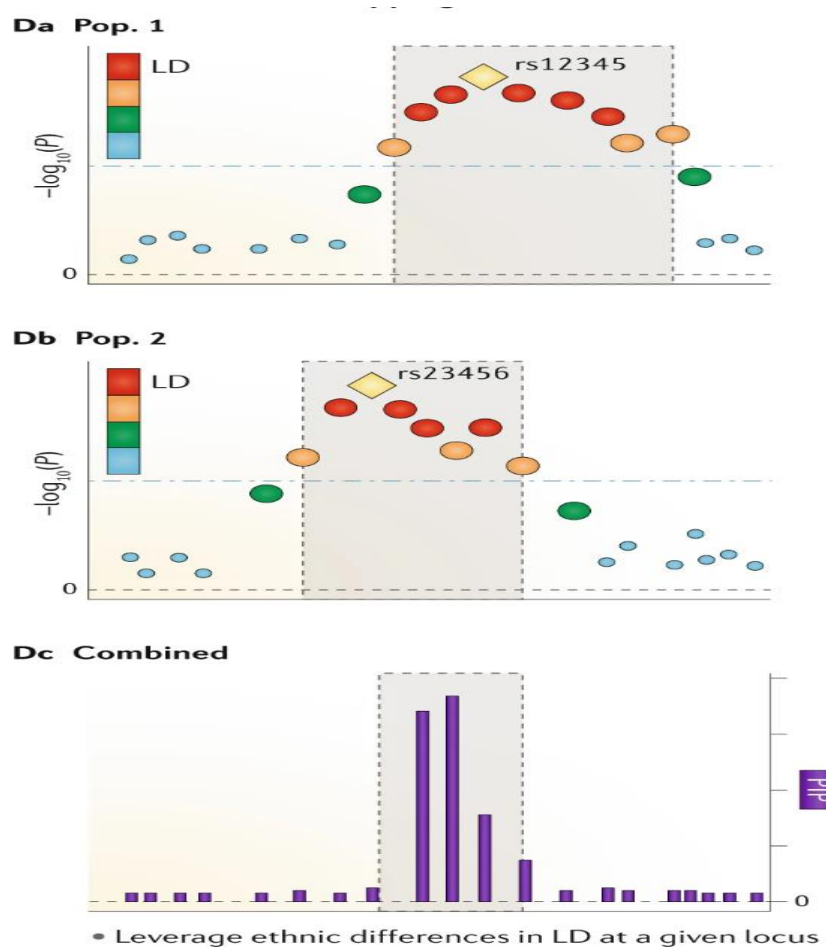


Kichaev *et al.* 2014 PLOS Genet
Chen *et al.* 2016 Genetics
Wen *et al.* 2016 AJHG
Mahajan *et al.* 2018b Nat Genet

贝叶斯方法 Bayesian methods

- 与 P 值不同，可以**直接比较** SNP 的后验概率。
- 与根据同主要 SNP 相关性选择 SNP 相比，该方法**倾向于选择较少**的 SNP 作为潜在致病因素。
- 模拟研究表明，贝叶斯方法的**性能优于条件逐步回归和惩罚回归模型**。
- 贝叶斯模型**基于 SNP 的联合效应**，它们控制具有较大效应的 SNP，提高检测具有较小效应的 SNP 的能力。

跨人群精细定位 Trans-ethnic fine-mapping



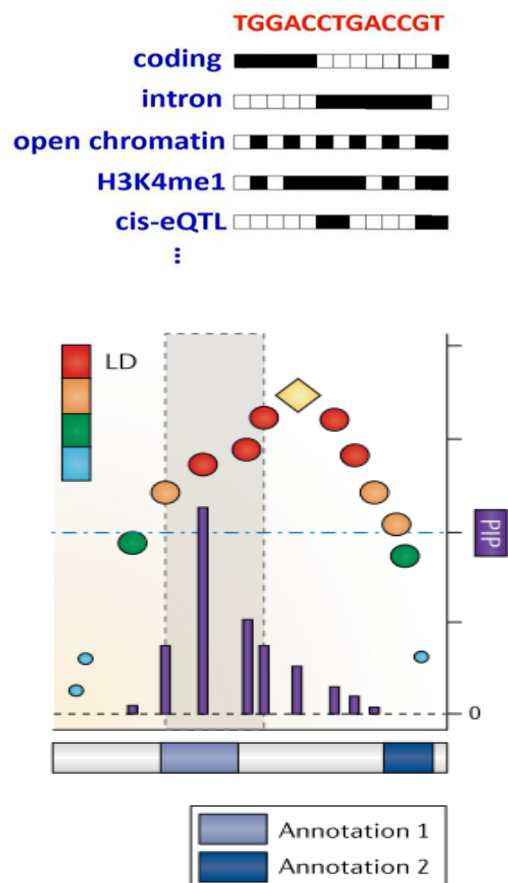
原理:

跨种族荟萃分析结合了遗传多样性人群中相同特征的 GWAS 结果，可以通过利用 LD 模式中的种族差异来帮助精细定位。

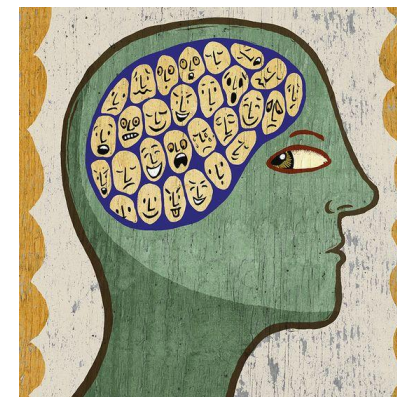
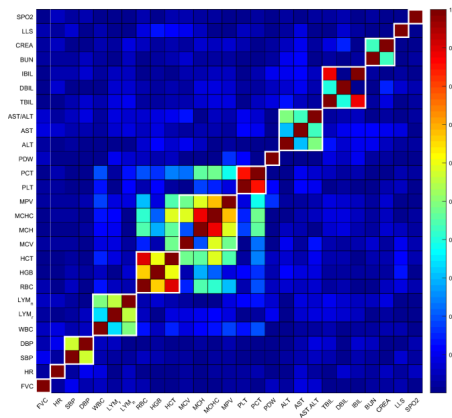
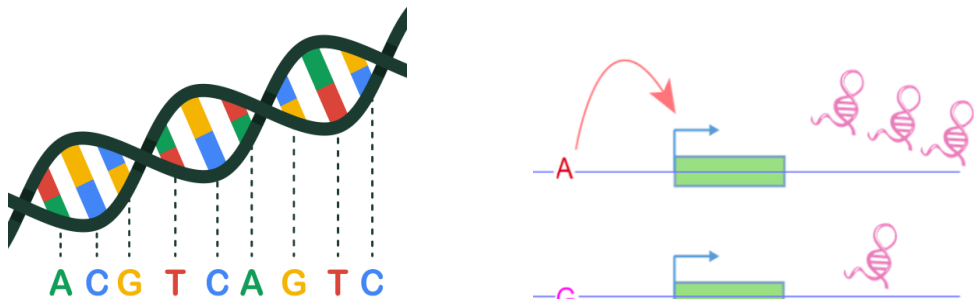
软件:

METASOFT, PAINTOR, CAVIAR, SuSiEx等。

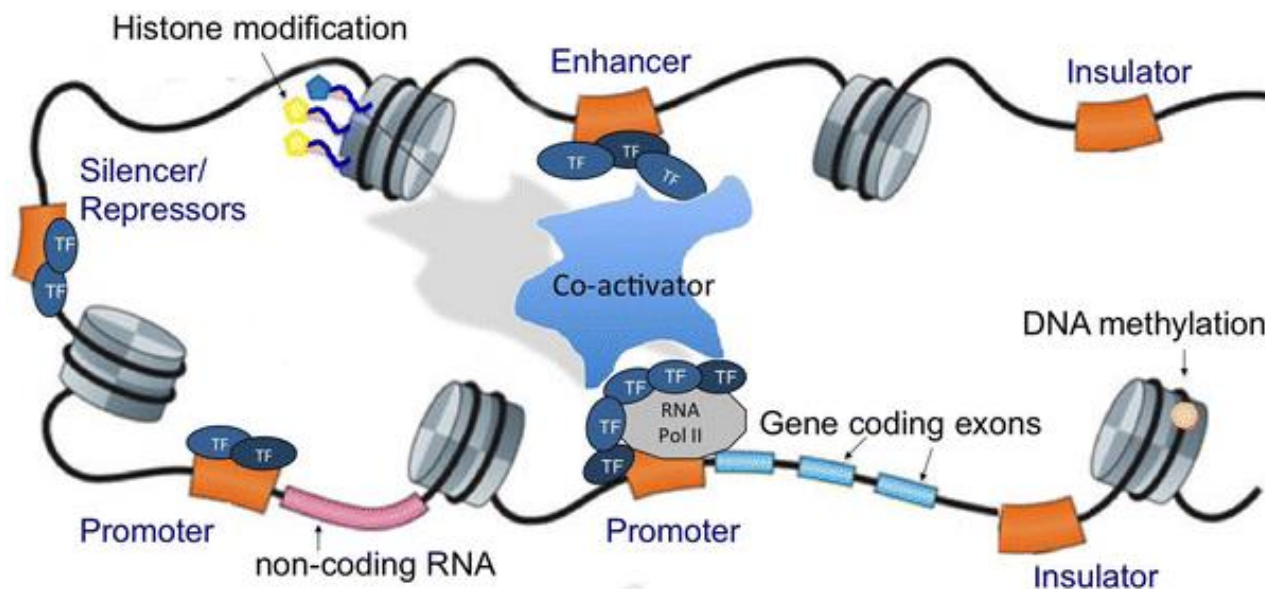
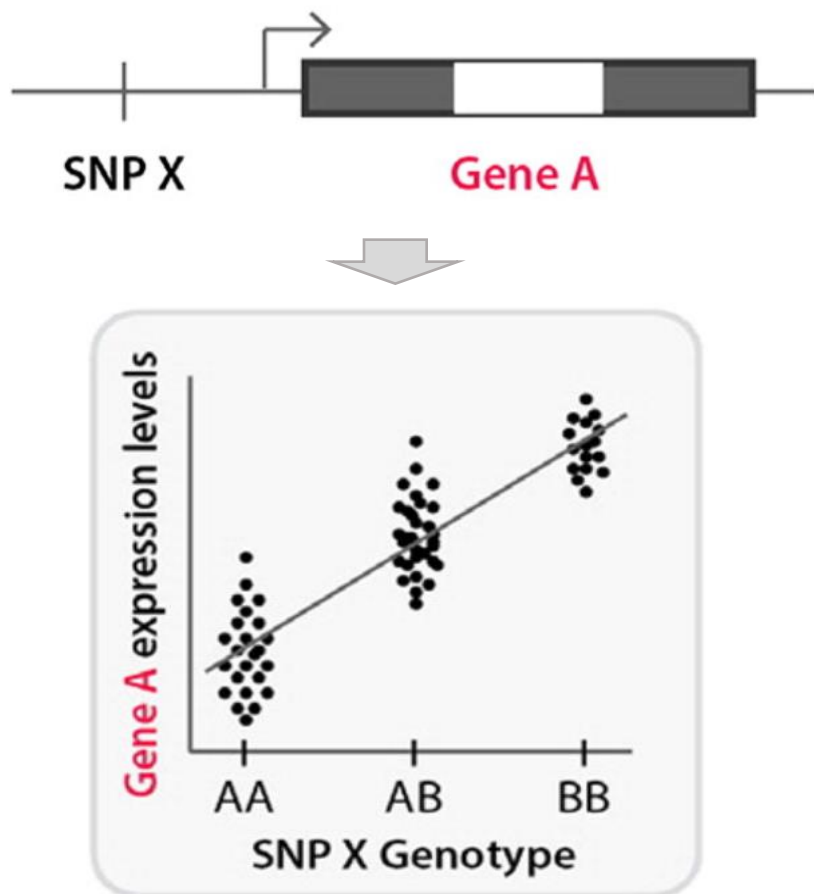
通过借助其他调控信息，增强精细定位结果的可靠性



- Genomic region (coding / intron / promoter...)
- Chromatin state (DHS / predictions)
- Histone marks (methylation / acetylation...)
- Conserved regions
- Enhancers / super-enhancers
- Bins of minor allele frequency
- LD-annotations (Levels of LD, recombination rate...)
- Cis-eQTL SNPs



eQTL分析：研究个体的基因型与基因表达水平之间的关联，鉴定与基因表达水平显著相关的遗传变异

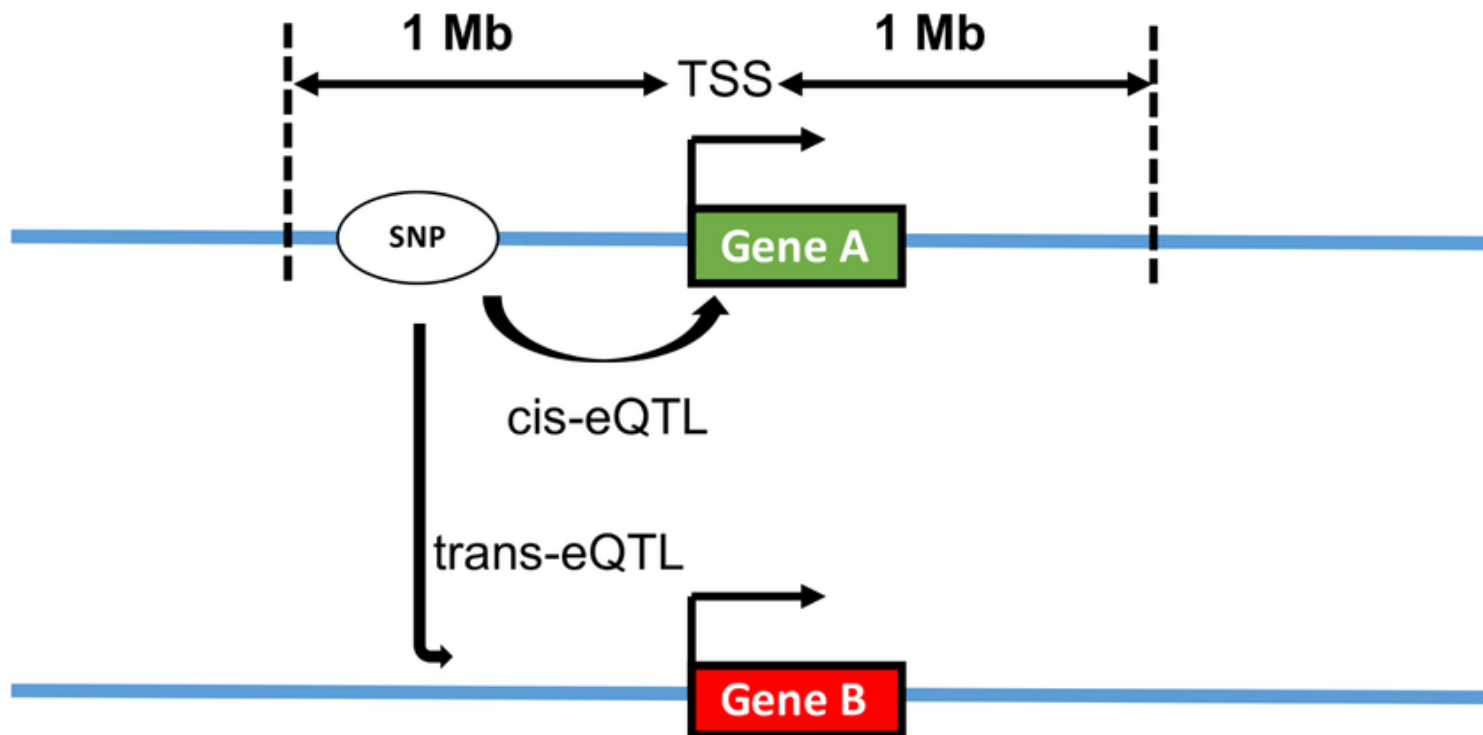


主要的调控元件：（1）启动子，（2）增强子，（3）组蛋白修饰和DNA甲基化，（4）沉默子、抑制子和绝缘子，以及（5）非编码RNA。

— 植物QTL分析

顺式-eQTLs (cis-eQTLs)

反式-eQTLs (trans-eQTLs)



- 顺式-eQTLs（通常在1Mb以内）通常比反式-eQTLs具有更大的效应。
- 大多数研究都集中在顺式-eQTLs上。

关联分析：线性回归模型

$$Y = \beta X + C_a + \epsilon$$

Y 代表基因表达量,

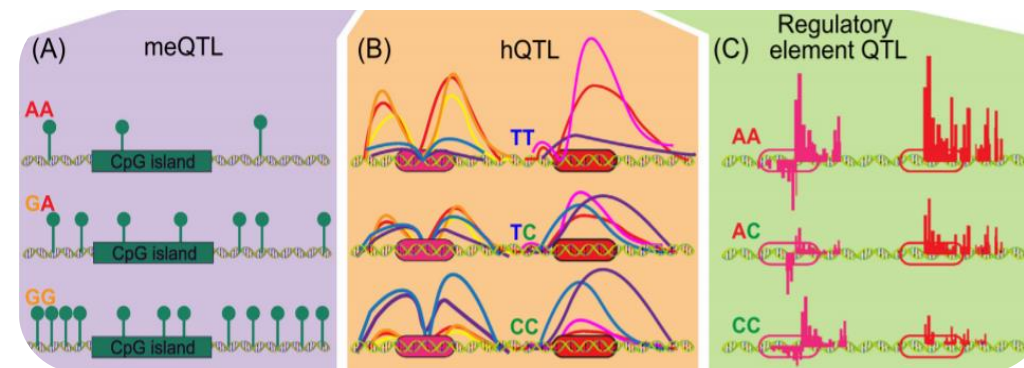
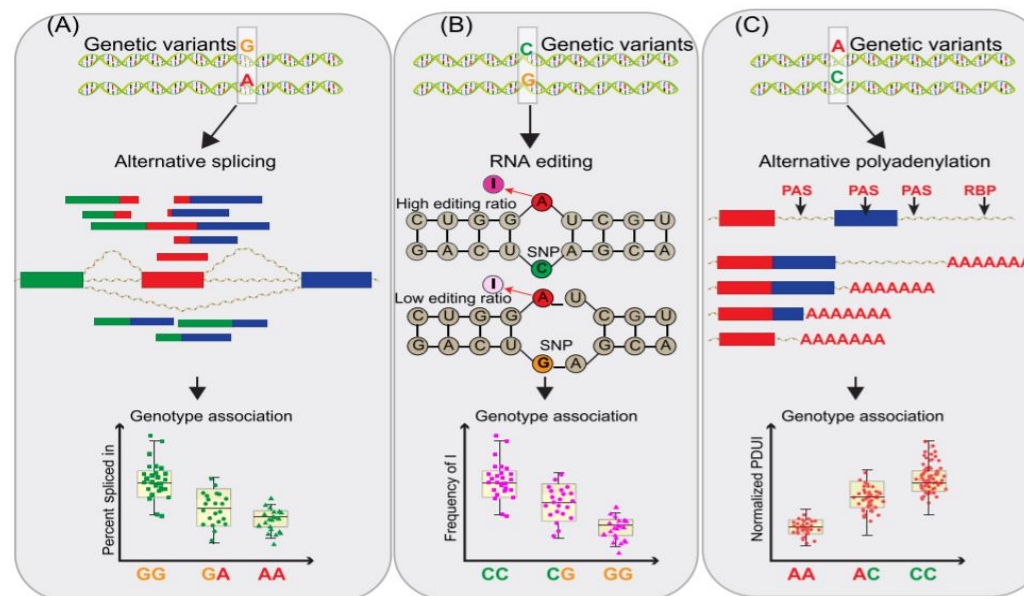
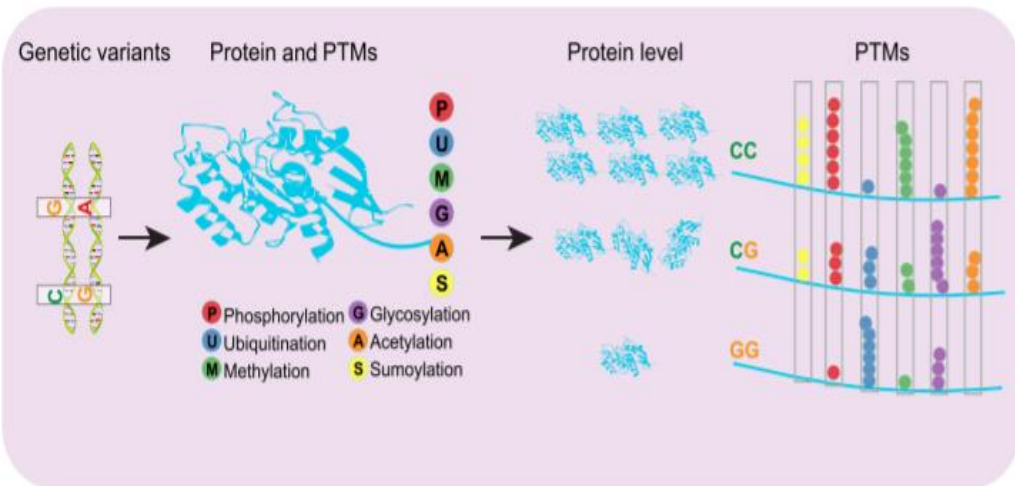
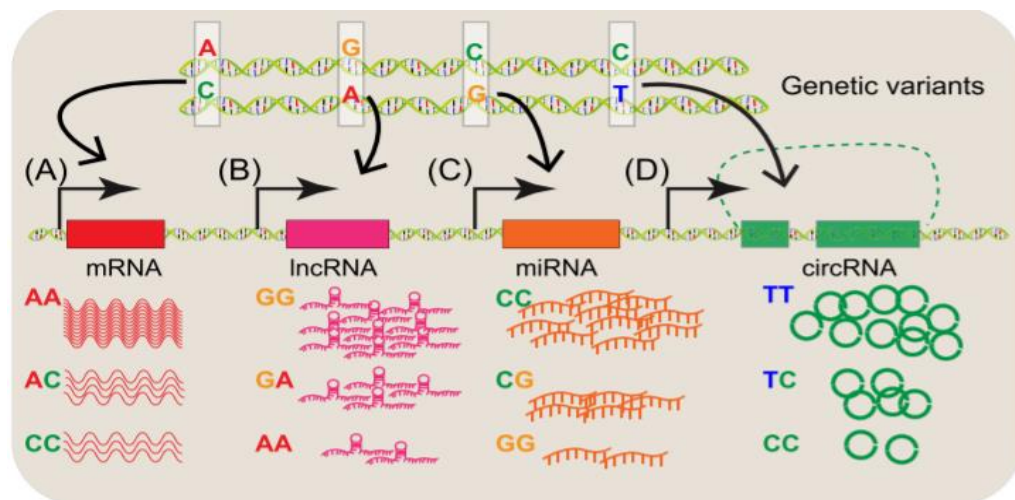
X 代表0, 1 和 2 编码的 SNP,

β 代表线性回归的斜率, 代表了SNP对基因表达的调控效应,

C 为协变量因素, 包括群体的主成分, 年龄, 性别及等选择的其他潜在协变量。

ϵ 代表残差项。

- 软件: Matrix EQTL, FastQTL, QTLtools ...
- 数据库: GTEx, Braineac, Blood eQTL Browsers ...



- Expression QTL (eQTL)
- Methylation QTL (meQTL)
- Splicing QTL (sQTL)
- Chromatin accessibility QTL (caQTL)
- Protein QTL (pQTL)
- ...

连接基因型到表型

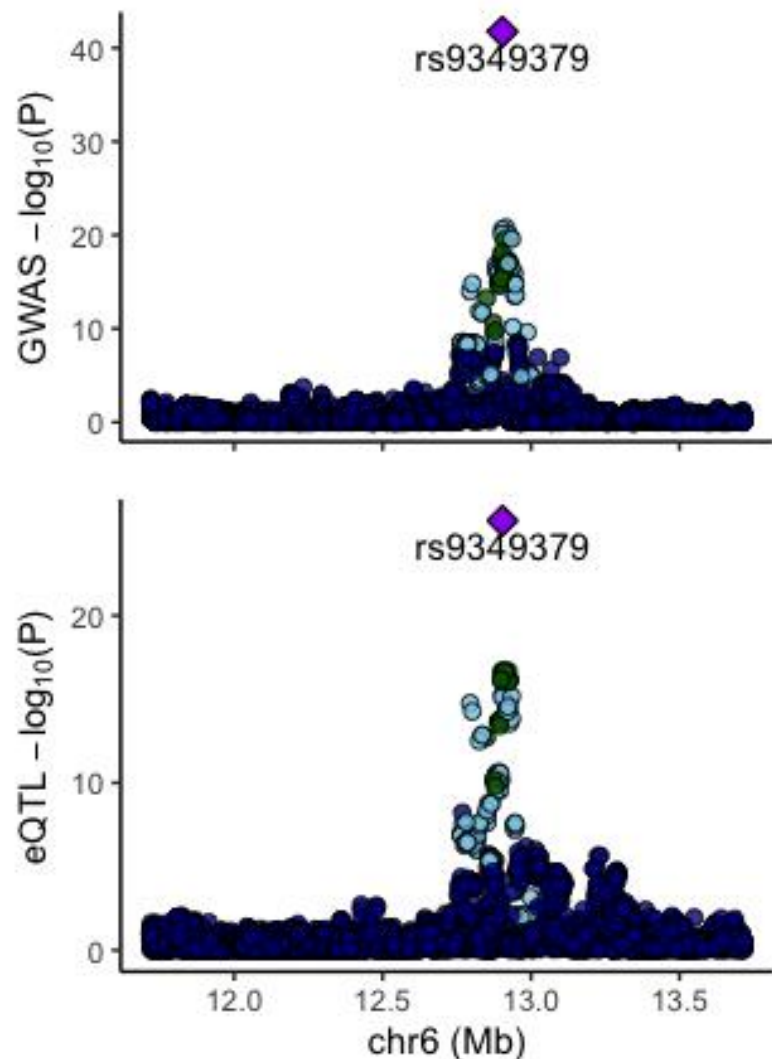
Connect Genotypes to Phenotypes

Methods

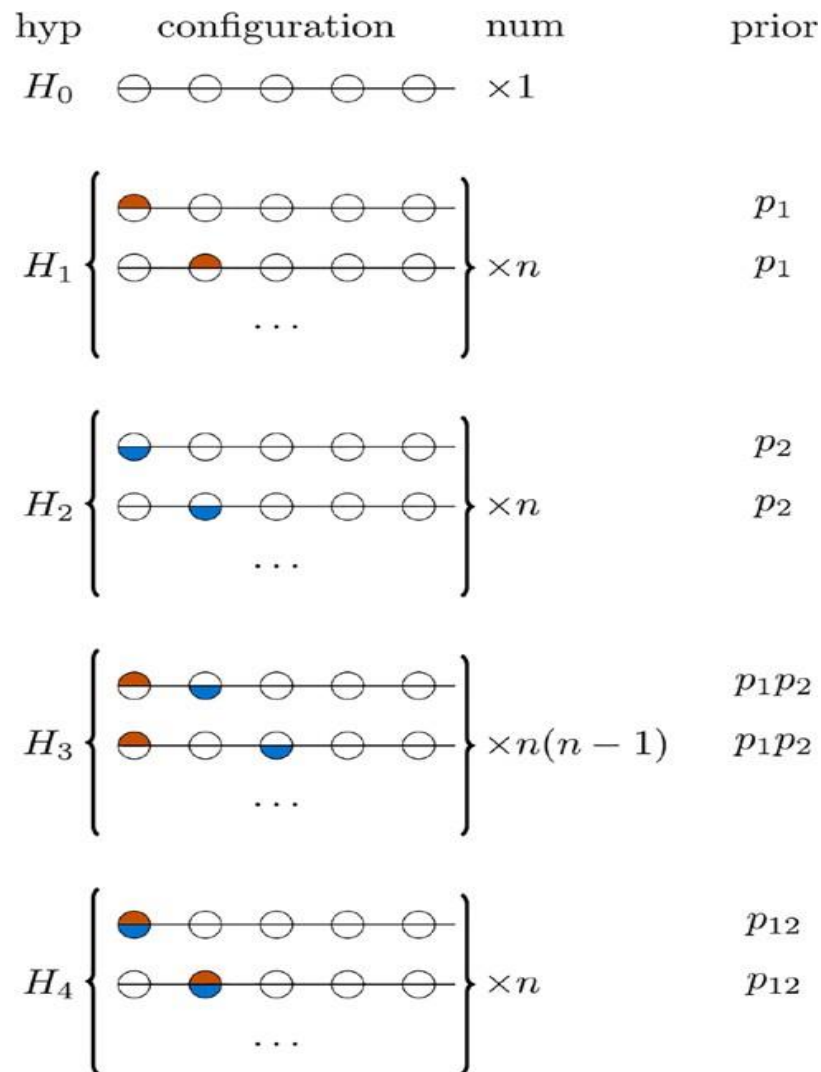
- 共定位方法
- Co-localization
- 孟德尔随机化
- Mendelian randomization
- 基因表达填充
- Gene-expression imputation

Name	Description	Account for LD	Programming Language	Operating system	Link
MetaXcan	Gene expression Y imputation	Y	Python	Unix/Linux	https://github.com/hakyimlab/MetaXcan
PrediXcan	Gene expression N imputation	N	Python	Unix/Linux	https://github.com/hakyimlab/PrediXcan
TWAS	Gene expression Y imputation	Y	R	Unix/Linux, Mac OS, Windows	http://bogdan.bioinformatics.ucla.edu/software/twas/
COLOC	Co-localization	Y	R	Unix/Linux, Mac OS, Windows	https://github.com/chriswallace/coloc
MOLOC	Co-localization	Y	R	Unix/Linux, Mac OS, Windows	https://github.com/clagiamba/moloc
Sherlock	Co-localization	Y	-	Web interface	http://sherlock.ucsf.edu/
JEPEG	Joint analysis eQTL	N	C	Unix/Linux	https://dleeelab.github.io/jepeg/
CAVIAR	Co-localization	Y	C	Unix/Linux, Mac OS, Windows	http://genetics.cs.ucla.edu/caviar/
eCAVIAR	Co-localization	Y	C	Unix/Linux, Mac OS, Windows	http://genetics.cs.ucla.edu/caviar/
GMAC	Mediation analysis	N	R	Unix/Linux, Mac OS, Windows	https://cran.r-project.org/web/packages/GMAC
FINEMAP	Co-localization	N	C	Unix/Linux, Mac OS	http://www.christianbenner.com/
NEO	Structural Equation Model	N	R	Unix/Linux, Mac OS, Windows	https://labs.genetics.ucla.edu/horvath/htdocs/aten/NEO/
SMR	Mendelian randomization	Y	C++	Unix/Linux, Mac OS, Windows	http://cnsgenomics.com/software/smr
GSMR	Mendelian randomization	Y	R	Unix/Linux, Mac OS, Windows	http://cnsgenomics.com/software/gsmr/

Note: TWAS, transcriptome-wide association study; JEPEG, joint effect on phenotype of eQTLs/functional SNPs associated with ; CAVIAR, causal variants identification in associated regions; eCAVIAR, eQTL and GWAS causal variants identification in associated r GMAC, Genomic Mediation Analysis with Adaptive Confounding Adjustment; NEO, Network Edge Orienting; SMR, summary-dat



共定位分析通过计算 **GWAS (表型1)** 显著位点和 **QTL (表型2)** 显著位点共同定位在同一基因组区域的后验概率，来估计GWAS信号上的位点通过改变基因表达等特征的生物学过程从而影响疾病的可能性。



$$P(H_h|D) \propto \sum_{S \in S_h} P(D|S)P(S) \quad \text{表型1-GWAS; 表型2-QTL}$$

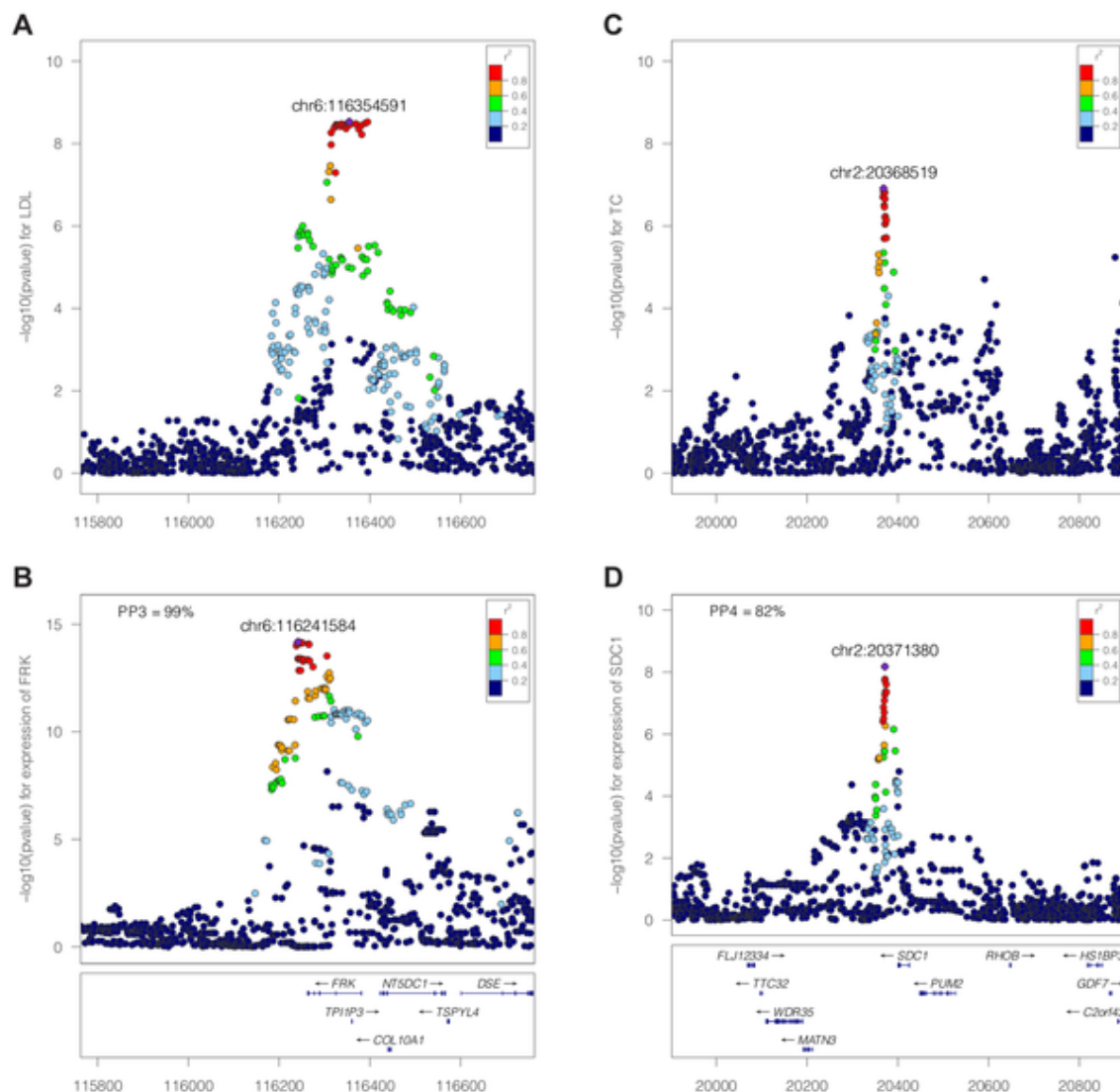
H_0 : 表型1 和表型2 与某个基因组区域的所有SNP位点无显著相关。

H_1 : 表型1 与某个基因组区域的SNP位点显著相关。

H_2 : 表型2 与某个基因组区域的SNP位点显著相关。

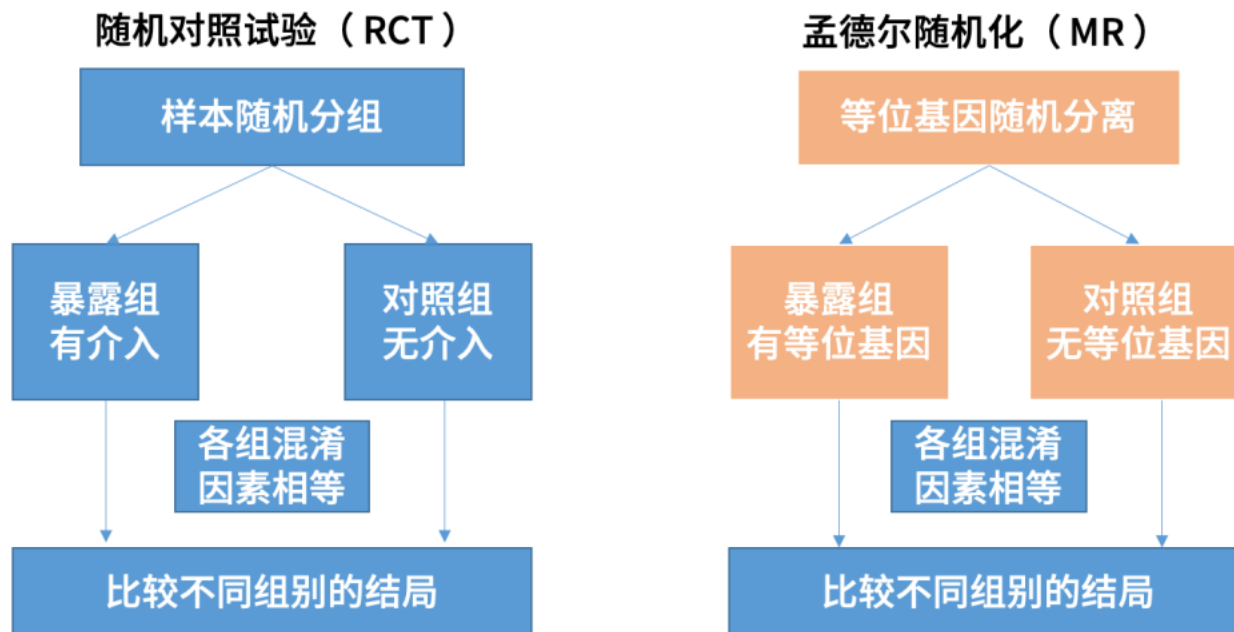
H_3 : 表型1 和 表型2 与某个基因组区域的SNP位点显著相关，但由不同的因果变异位点驱动。

H_4 : 表型1 和 表型2 与某个基因组区域的SNP位点显著相关，且由同一个因果变异位点驱动。

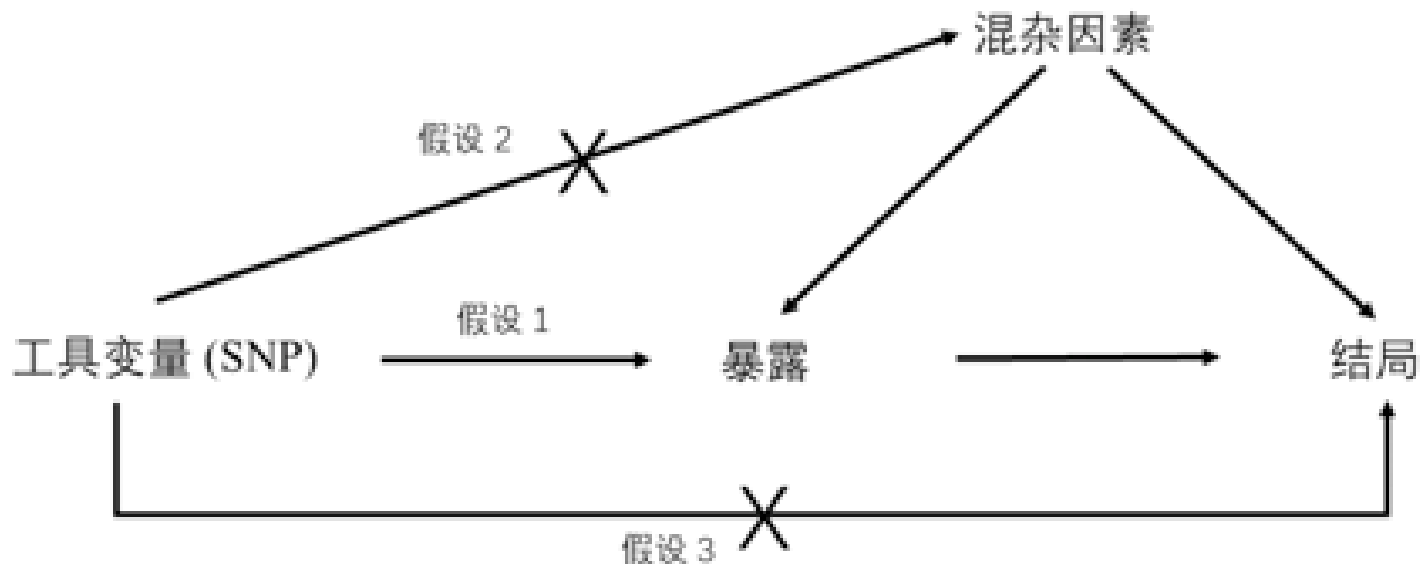


- PPH4 值的范围在0 - 1之间。
- 当 PPH4 大于 **0.75** 时，提示了 GWAS 和 eQTL 在同一基因组区域的共定位显著遗传关联信号。

- 工具变量法在生物信息学中的重要应用。SNP位点符合因果时序关系，不会受到后天环境等混杂因素的干扰，选择特定的SNP位点作为工具变量，可以对观察性研究中的因果关系进行推断。
- 随机对照试验 (Randomized Controlled Test, RCT)
 - 多中心，大样本，双盲
- 天然造成不同的“组” 随机对照试验 (RCT) 孟德尔随机化 (MR)



前提/假说

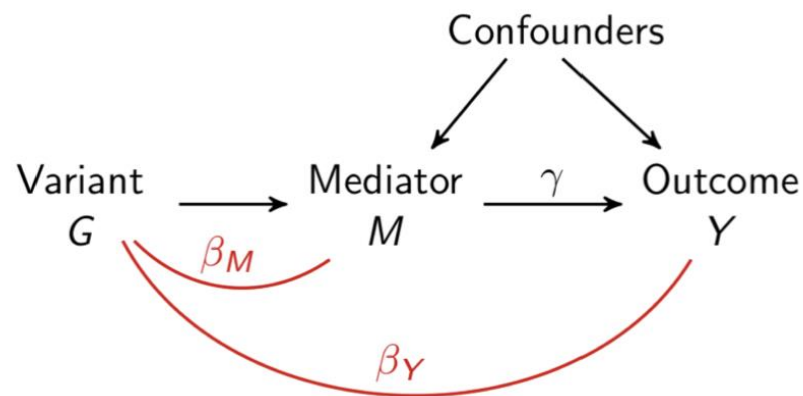


假说1： 关联性假设，即工具变量与暴露因素之间是强相关的

假说2： 独立性假设，即工具变量与混杂因素之间是独立的

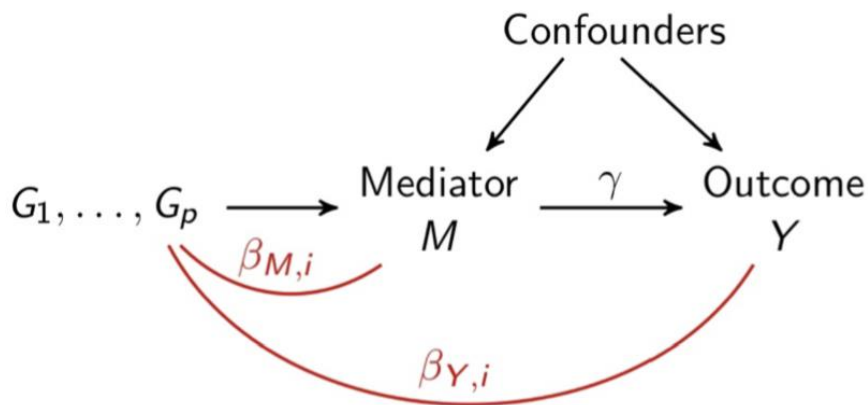
假说3： 排他性假设，即工具变量只能通过暴露因素对结局产生作用

单个工具变量



$$\frac{\Delta Y}{\Delta G} = \frac{\Delta Y}{\Delta M} \times \frac{\Delta M}{\Delta G} \Rightarrow \beta_Y = \gamma \cdot \beta_M$$

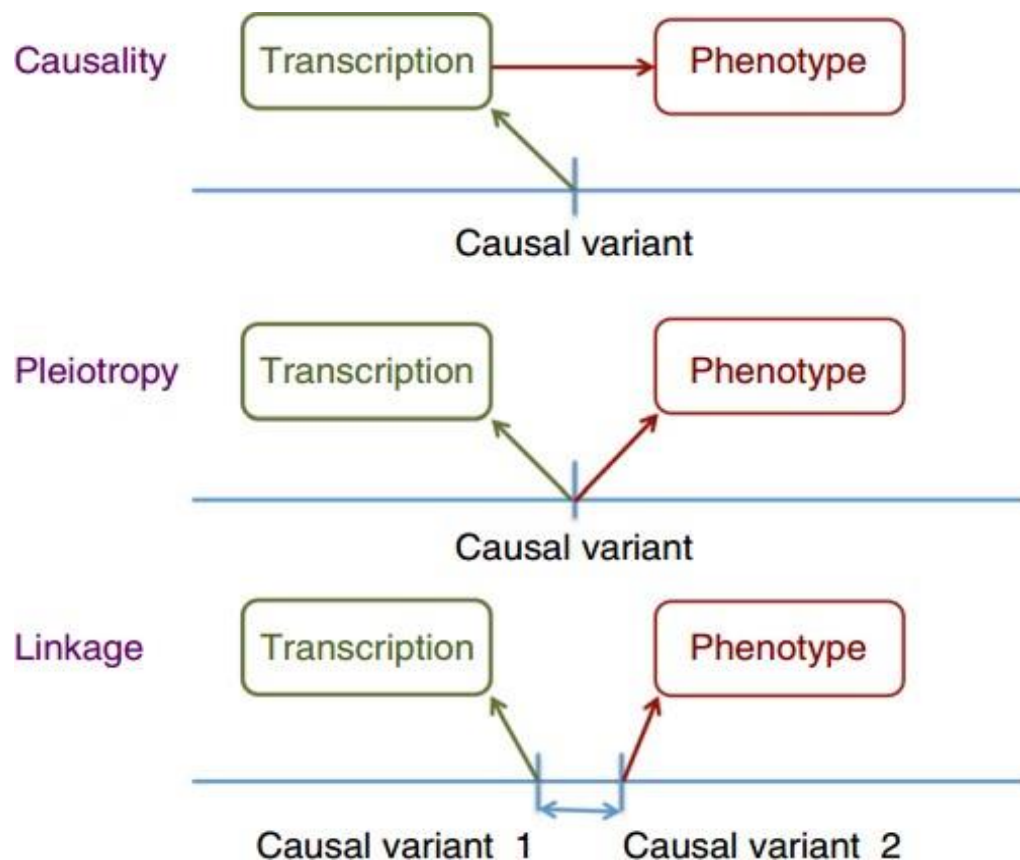
多个工具变量



$$\beta_{Y,i} = \gamma \cdot \beta_{M,i}, 1 \leq i \leq p$$

- **逆方差加权 (inverse variance weighted, IVW):**
汇总 p 个遗传工具变量 (instrumental variables, IVs) 的效应, 提供暴露变量对结果变量的总体因果效应估计

基因型作为工具变量，观察到的性状与基因表达之间的关联的可能解释



- MR-Egger
- HEIDI

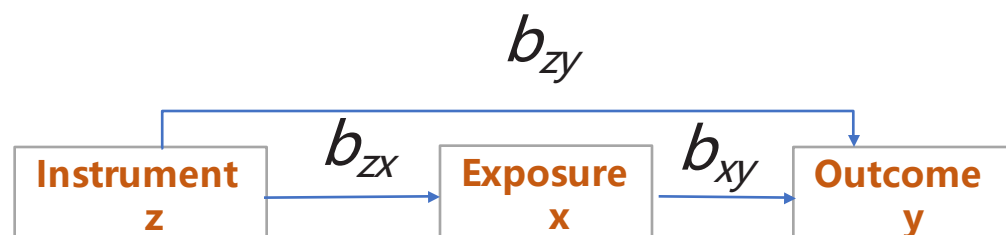
- Harmonize test
- MR-LDP

违背工具变量核心假设的因素

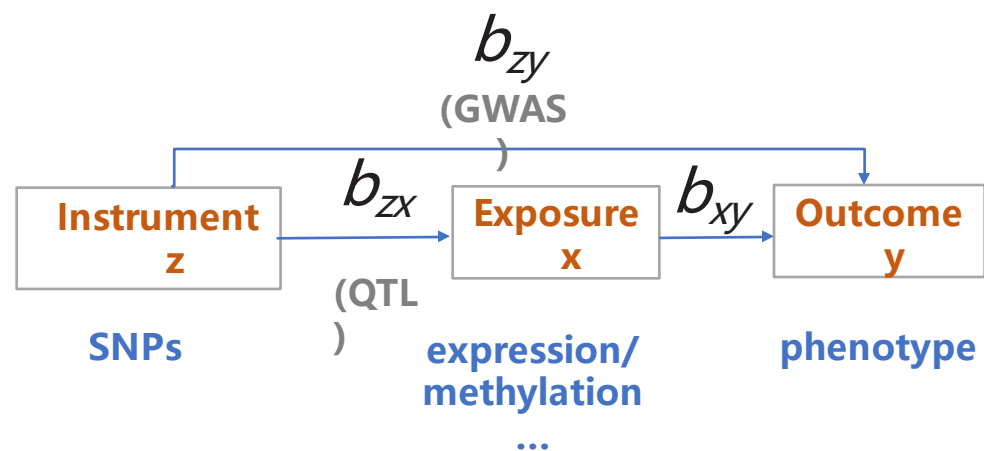
- **弱工具变量 (weak instrument)**
遗传变异与暴露因素不是强相关关系；遗传变异仅解释小部分的表型变异
一是增加研究样本量，工具变量F值 > 10 ；二是增加表型解释度，综合多个遗传变异效应
- **遗传多效性/异质性 (horizontal pleiotropy/heterogeneity)**
某些遗传变异能够影响暴露因素的多个方面，作为工具变量不影响暴露因素整体与结局之间的因果推断，但却难以推断暴露因素的某个方面与结局之间的因果关系
- **连锁不平衡 (linkage disequilibrium)**
- **人群分层 (population stratification)**
- **Beavis效应 (Beavis effect) : Winner's curse** 高估工具变量效应
- **Collider偏倚 (Collider bias)** : 暴露和结果通过一个共同变量中间变量，导致虚假关联

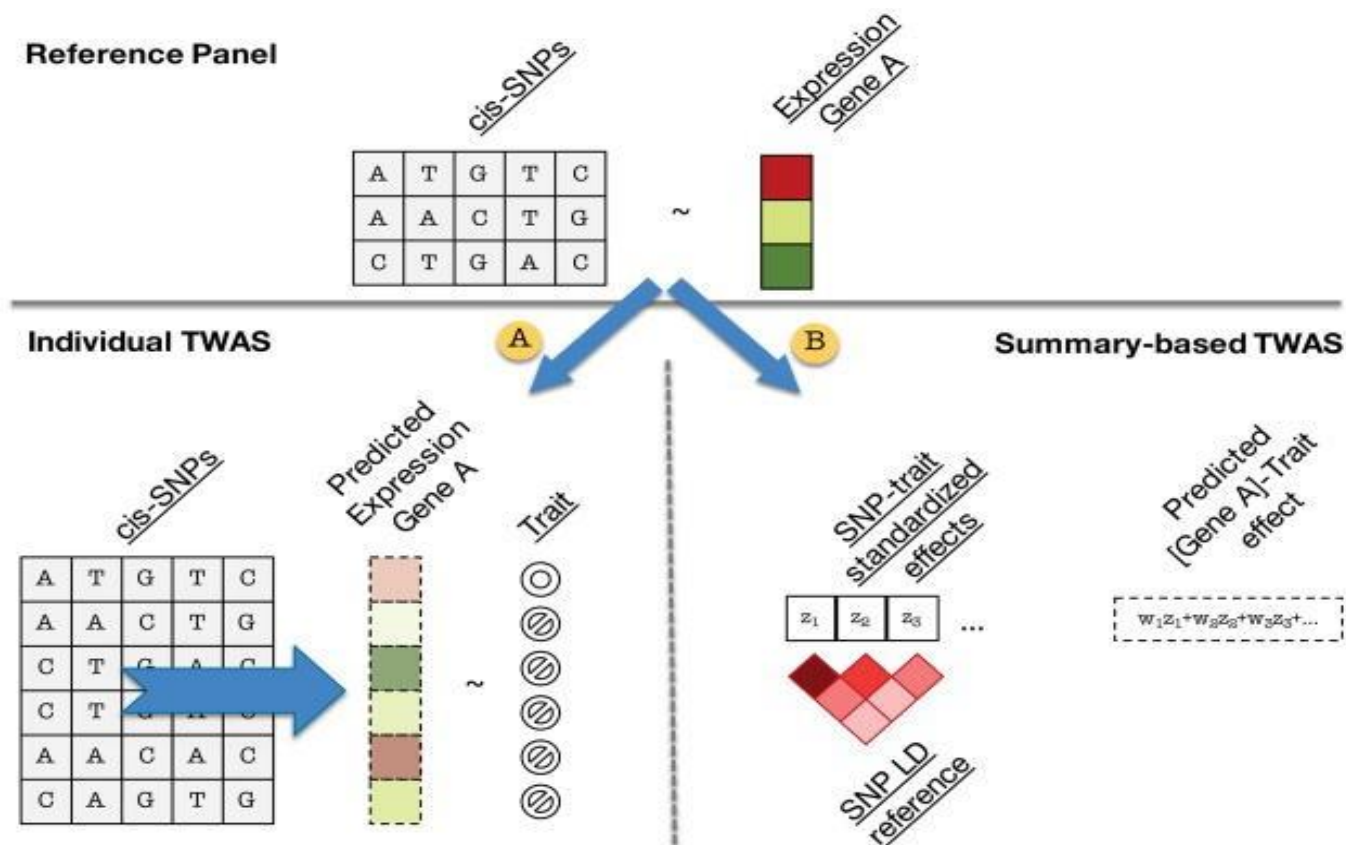
汇总孟德尔随机化 Summary-data-based Mendelian randomization (SMR)

一种特定形式的MR分析，主要利用公开的GWAS汇总数据



In an MR analysis, b_{xy} (defined as $b_{xy} = b_{zy}/b_{zx}$)





- A: 使用参考面板中的效应大小直接预测基因测序样本的表达，并测量预测表达与性状之间的关联。
- B: 利用 SNP 性状标准化效应大小的加权线性组合，同时考虑 SNP 之间的 LD，间接估计预测表达和性状之间的关联。

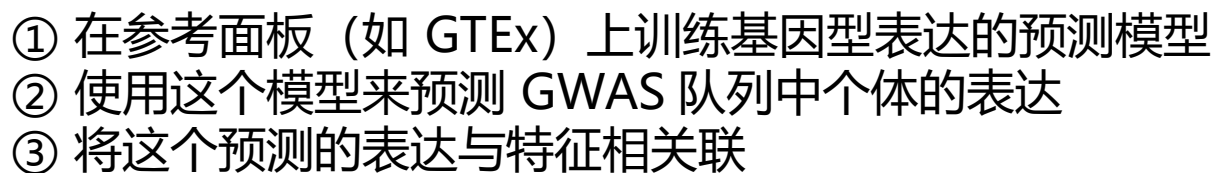


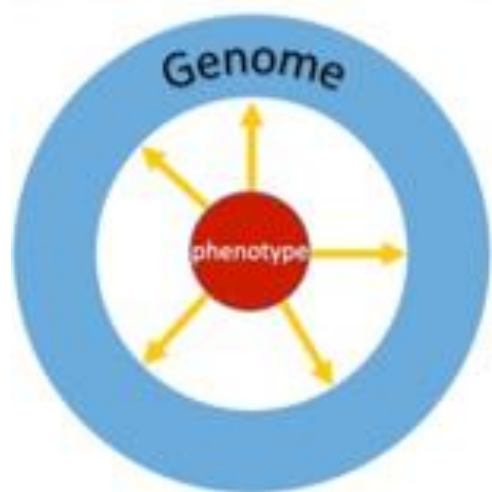
Table 2 A summary of thirteen TWAS approaches examined in the present review

Methods	Design	Tissue	Data type	Pleiotropy	Model assumptions	URLs
PrediXcan	Two-stage	Single	Individual	No	Elastic net	https://github.com/hakyimlab/PrediXcan
S-PrediXcan	Two-stage	Single	Summary	No	Elastic net	https://github.com/hakyimlab/MetaXcan
TWAS	Two-stage	Single	Individual /Summary	No	BSLMM	https://bogdan.dgsom.ucla.edu/pages/twas/
DPR	Two-stage	Single	Individual	No	DPR	http://www.xzlab.org/software.html
TIGAR	Two-stage	Single	Individual /Summary	No	DPR	https://github.com/yanlab-emory/TIGAR
CoMM	Likelihood-based	Single	Individual	No	LMM	https://github.com/gordonliu810822/CoMM
CoMM-S ²	Likelihood-based	Single	Summary	No	LMM	https://github.com/gordonliu810822/CoMM
PMR	Likelihood-based	Single	Individual /Summary	Yes	LMM	https://github.com/yuanzhongshang/PMR
UTMOST	Two-stage	Multiple	Summary	No	LASSO & Ridge	https://github.com/Joker-Jerome/UTMOST
MultiXcan	Two-stage	Multiple	Individual /Summary	No	Elastic net	https://github.com/hakyimlab/MetaXcan
TisCoMM	Likelihood-based	Multiple	Individual /Summary	No	LMM	https://github.com/XingjieShi/TisCoMM
fQTL	Two-stage	Multiple	Summary	No	BVSR	https://github.com/ypark/fqtl
FOCUS	Gene-mapping	Multiple genes	Summary	No		https://github.com/bogdanlab/focus

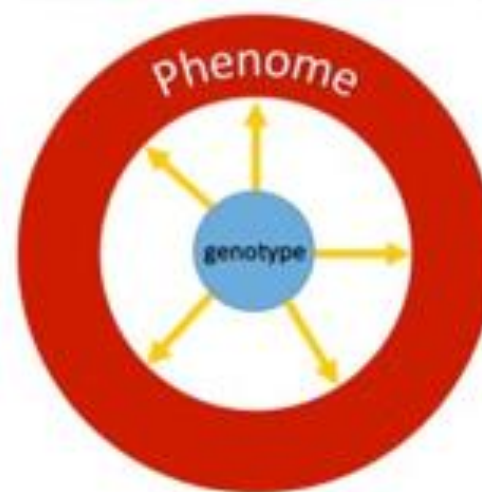
建模方法包括

弹性网络 (PrediXcan) ,
贝叶斯稀疏线性混合模型
(BSLMM) ,
狄利克雷过程回归 (DPR,
TIGAR) ,
线性混合回归 (CoMM,
CoMM-S2,PMR) ,
贝叶斯变量选择回归
(fQTL)

GWAS与PheWAS的区别:

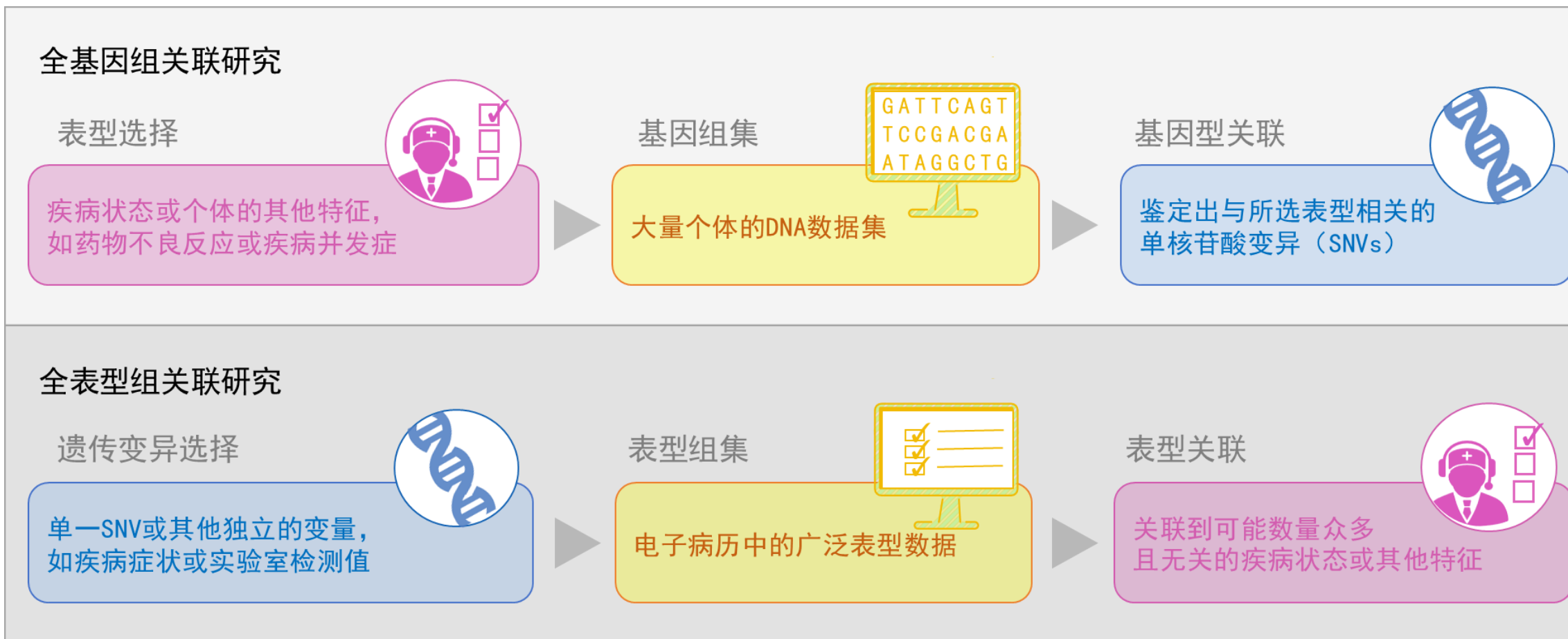


GWAS: 寻找与特定表型相关的未知遗传变异。

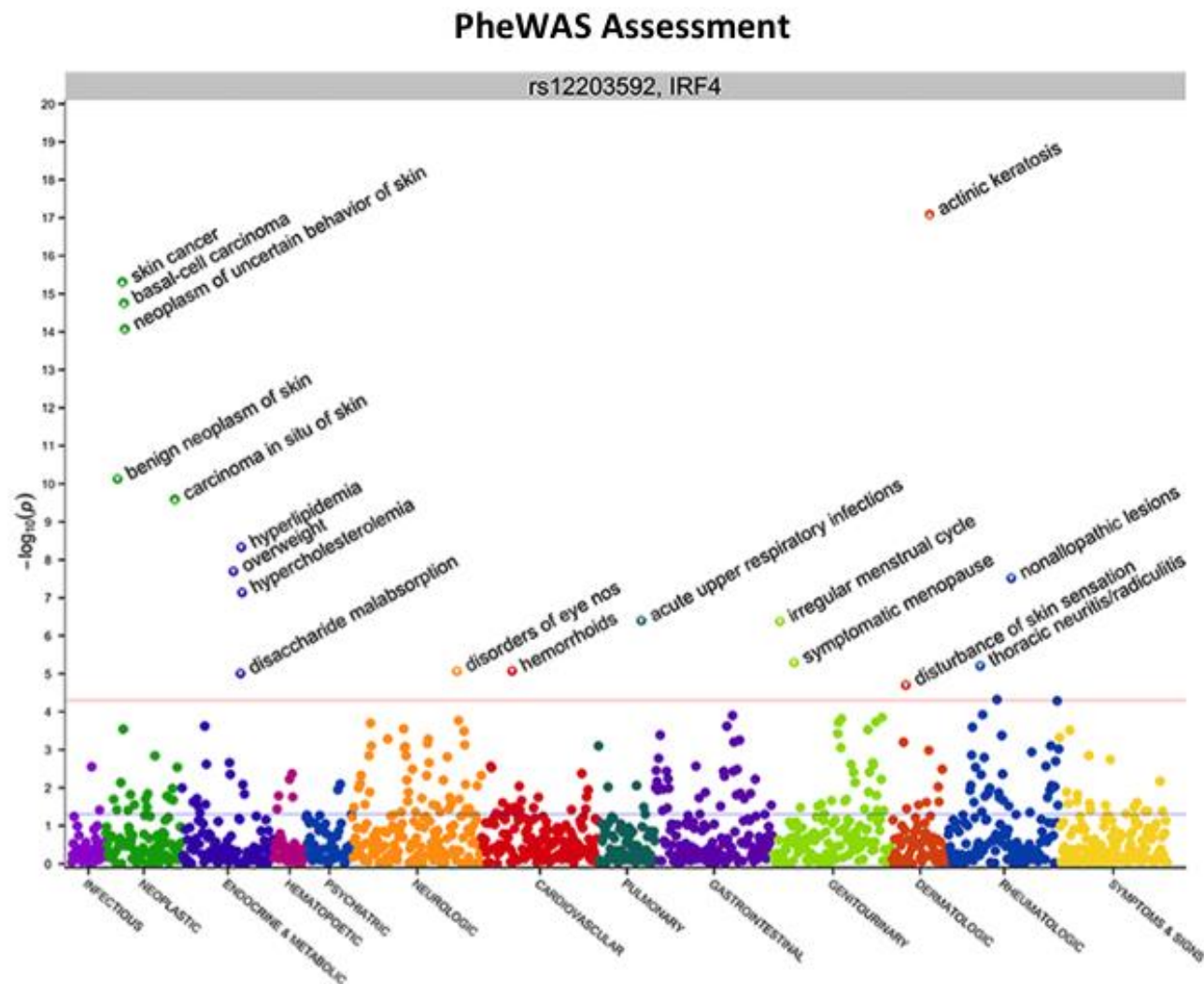


Phenome-wide association study (PheWAS): 从已知遗传变异出发, 探索其与**多种表型**特征的复杂联系。

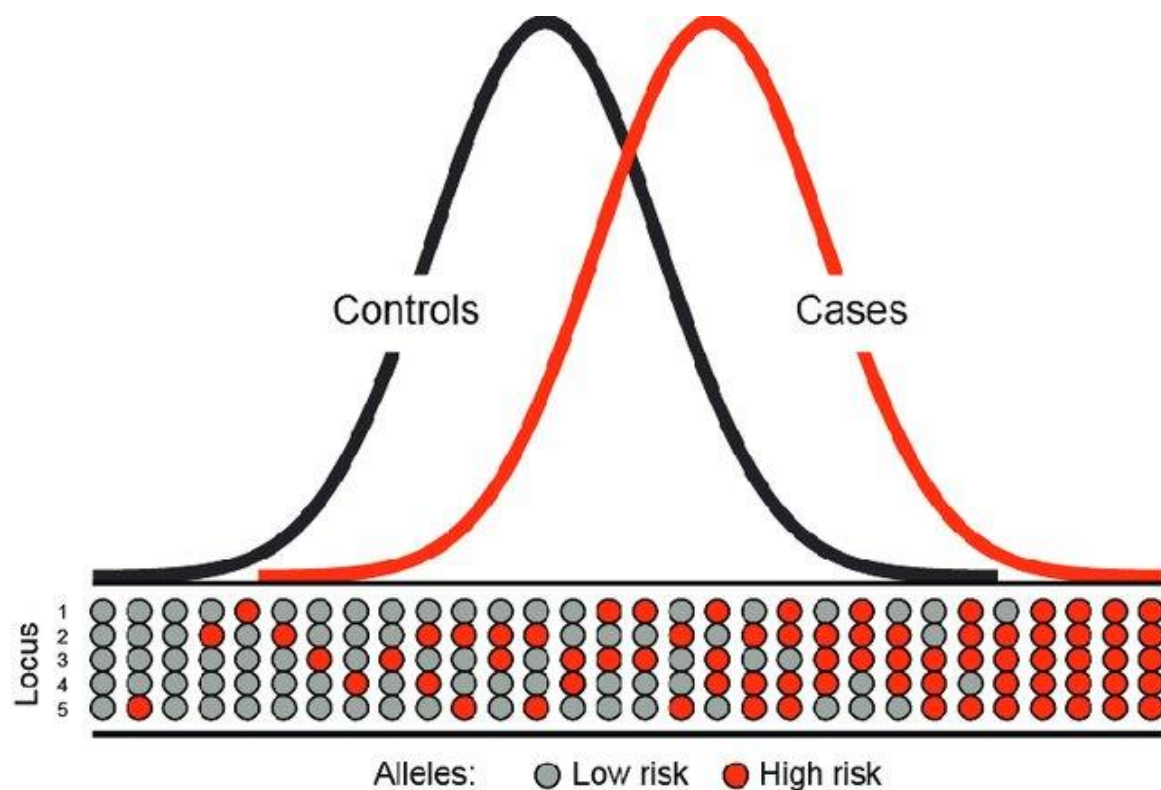
- 分类表型: 卡方检验、比值比
- 连续表型: 方差分析、线性回归



全基因组关联研究与全表型组关联研究的对比



揭示遗传位点的多效应 (pleiotropy)



- 许多遗传变异与疾病风险相关，效应各不相同且较小。
- 一个人的遗传风险是其所有风险等位基因效应的汇总。
- 通常情况下，病例个体所携带的风险等位基因数量较多。

多基因疾病易感性模型

PRS的计算

$$S = \sum_{k=1}^m x_k \beta_k$$

S 表示个体的多态性风险评分, m 是 SNP 的数量, x_k 代表第 k 个 SNP 的风险等位基因的个数, 通常取值为0、1或2, 分别表示该SNP上有0个、1个或2个风险等位基因。 β_k 是该SNP的效应权重 (通常来自GWAS的效应大小)

两个关键问题:

- 权重选择;
- 变异位点的纳入标准;

PRS的计算

传统方法：利用GWAS中所有的位点及对应的效应量进行PRS计算

- 未考虑不同疾病的遗传结构（问题1）
- 未考虑LD结构（问题2）

改进方法一：聚集+阈值法（clumping + thresholding）

- 阈值：控制位点和性状之间的相关性；不同阈值都有一个对应的位点集合；在独立数据集上验证，选择最佳预测阈值集合；弥补了问题1。
- 聚集：选择LD模块最显著的SNP，删除周围有强LD关系的SNP（可尝试0.1, 0.3, 0.5等不同阈值）；选择第二显著的SNP，删除周围有强LD关系的SNP；重复此步骤，避免重复计算；弥补了问题2。
- **潜在问题：**选择的最显著SNP可能不是真正影响最大的；同一区域只选一个信号，可能忽略多个真实信号。

改进方法二：贝叶斯多基因预测 (Bayesian Polygenic Prediction)

- 不进行阈值筛选，构建大型回归模型。
- 将单一正态分布扩展为多个正态分布，更灵活地描述性状的遗传结构。

假设：

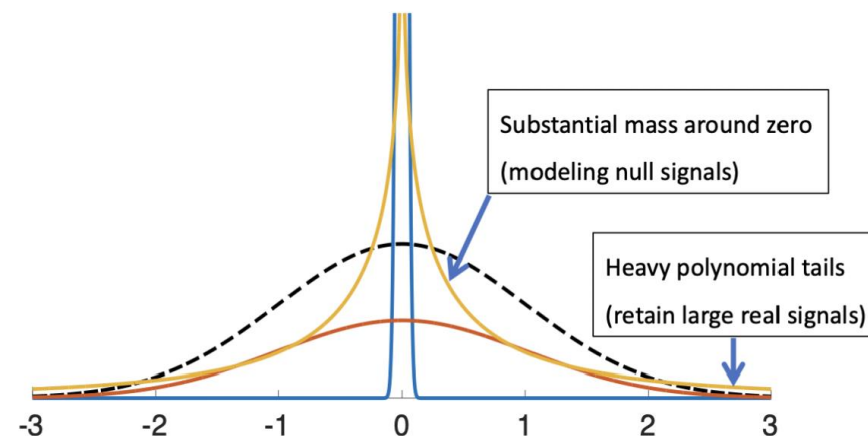
- SNPs是独立的
- 所有SNPs都是因果的
- SNP效应大小呈正态分布
- 罕见SNPs倾向于有更大的等位基因效应大小

Article | [Open access](#) | Published: 16 April 2019

Polygenic prediction via Bayesian regression and continuous shrinkage priors

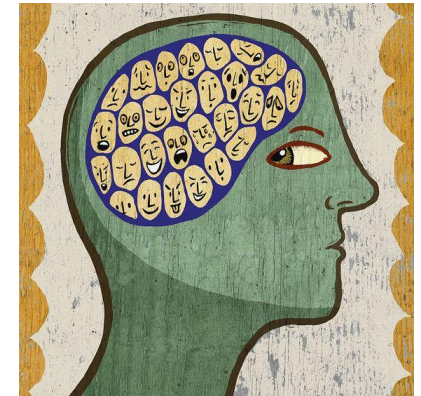
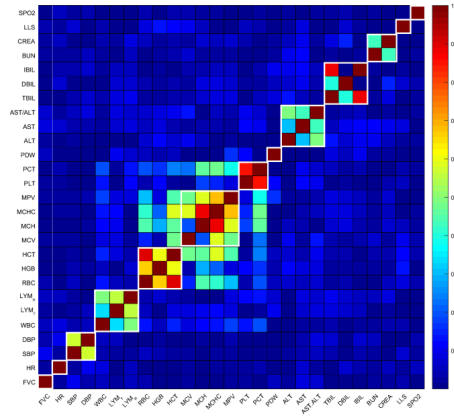
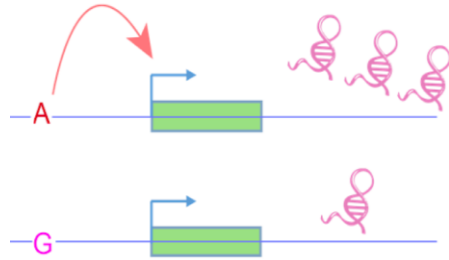
Tian Ge , Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng & Jordan W. Smoller

Nature Communications 10, Article number: 1776 (2019) | [Cite this article](#)

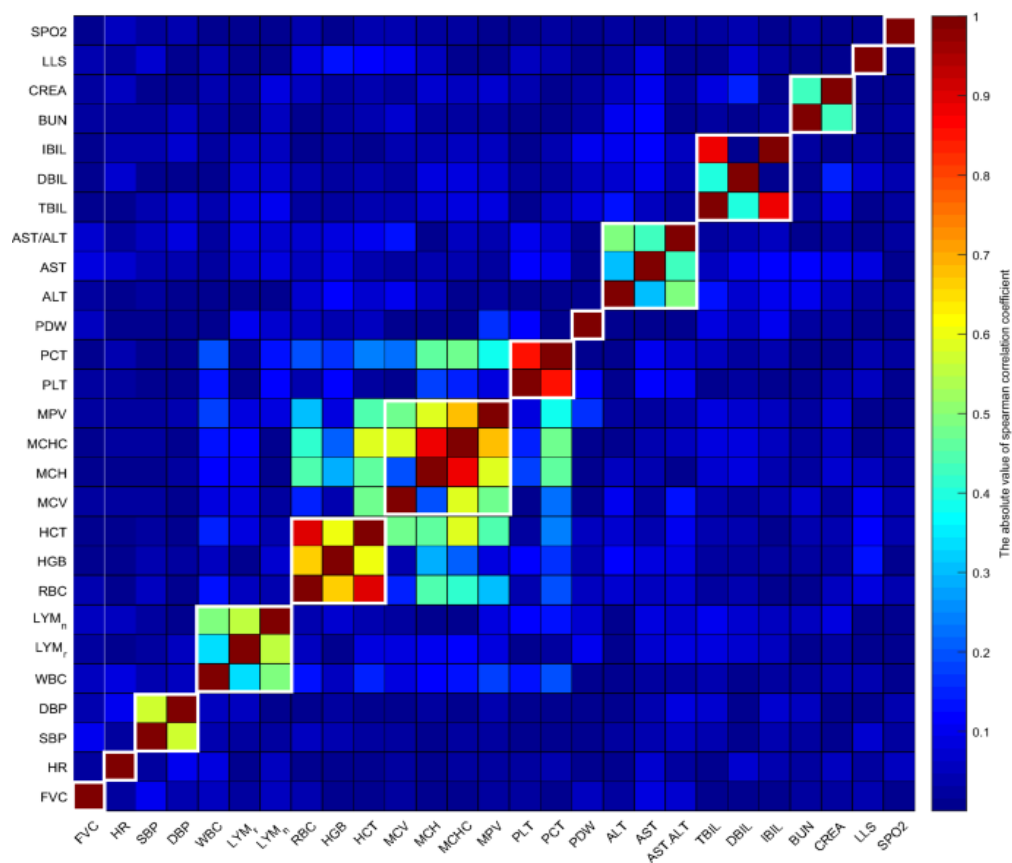


影响PRS统计功效和预测精度的几个关键因素

- 受到表型遗传力的影响
- GWAS的样本量应尽可能大，以提升SNP效应大小估计的准确性
- 目标数据集的样本量需要足够大，以确保性能指标的精确性和稳定性
- 包含的SNP数量（依方法而定）
- 训练样本与目标样本之间的相似性



复合表型 (Composite Phenotype) : 由多个单独的表型特征或指标组合而成的综合特征



光谱聚类 (Spectral Clustering)

不同表型之间的相似度矩阵

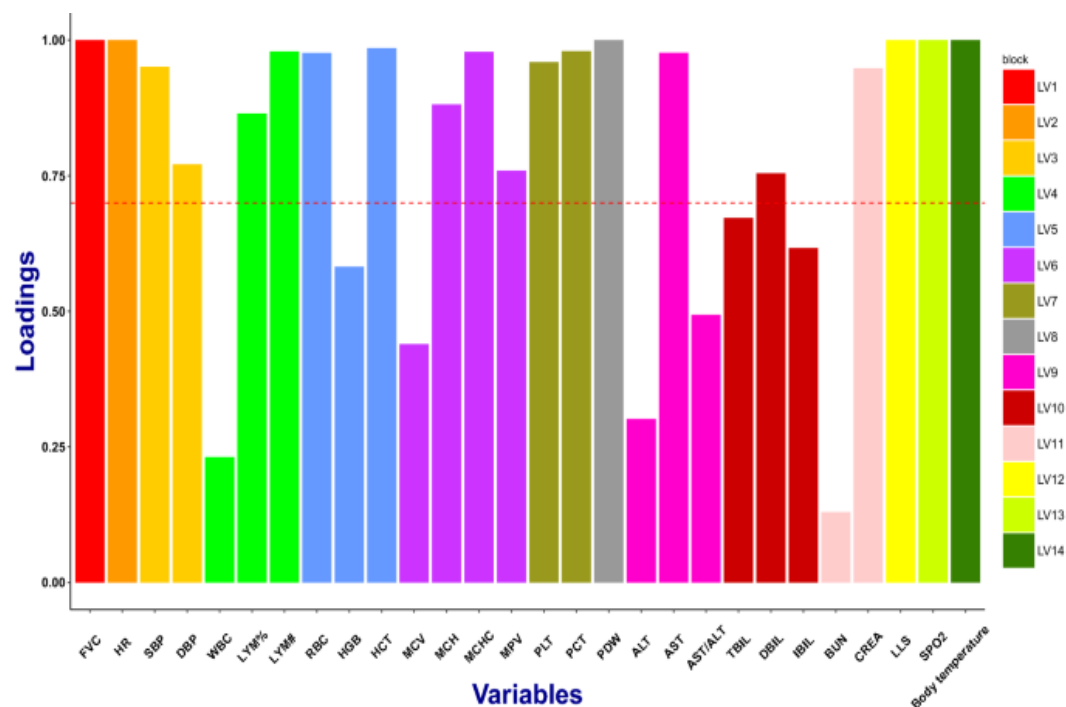
$$S_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

构建拉普拉斯矩阵 L

$$L = D - S$$

特征值分解和聚类, 构成复合表型的初步结构

复合表型 (Composite Phenotype) : 由多个单独的表型特征或指标组合而成的综合特征



偏最小二乘路径建模

(Partial Least Squares Path Modeling, PLSPM)

外部模型描述观测变量与潜变量之间的关系

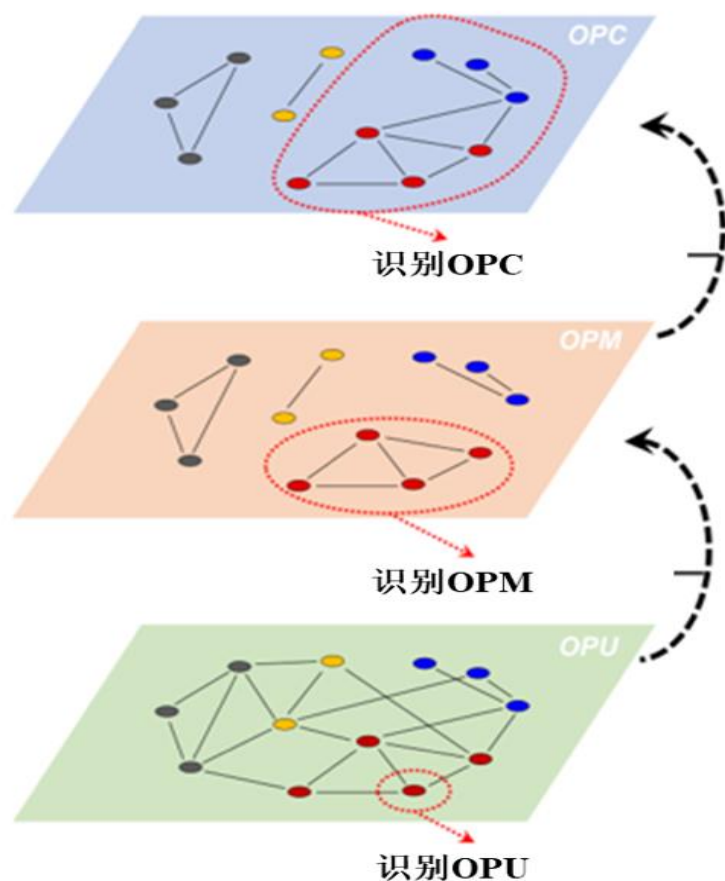
$$x_{ij} = \lambda_{ij}\varepsilon_j + \epsilon_{ij}$$

内部模型描述潜变量之间的关系

$$\xi_j = \sum_{k \neq j} \beta_{jk} \xi_k + \zeta_j$$

载荷矩阵和潜在变量得分来解释模型。

载荷矩阵显示了每个指标变量对每个潜在变量的贡献程度, 潜在变量得分被用于定义复合表型, 代表了不同生理方面

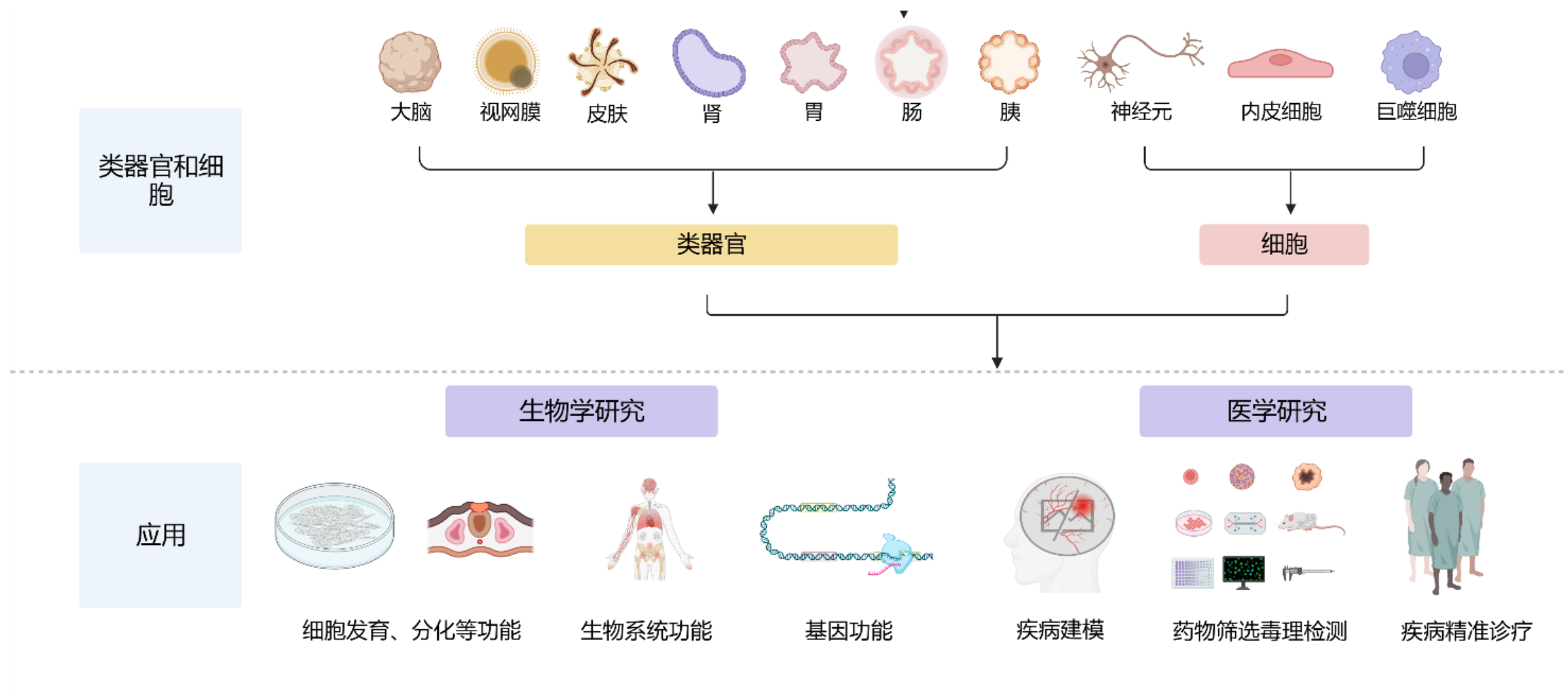


表型-表型交互作用网络构建框架

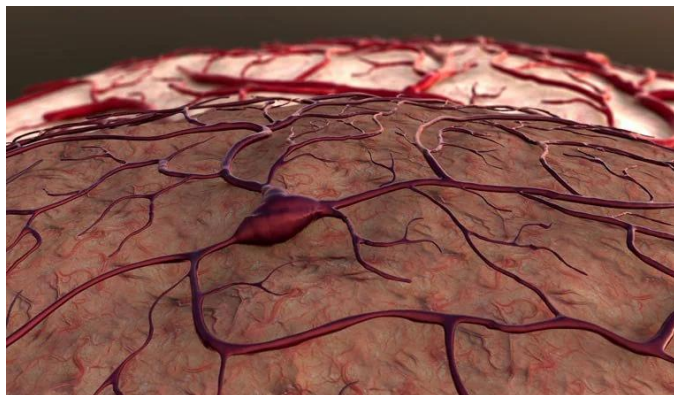
三个核心构成要素：

- **可操作表型单位 (operational phenotypic unit, OPU)**
- 独立的表型特征，如遗传特征或临床指标。可以表现为不同的方向性和强度，各表型特征之间的复杂相互作用
- **可操作表型模块 (operational phenotypic module, OPM)**
- 相互作用的OPU集合构成，具有相似表型模式的个体群体，体现了不同生物学过程的协同或相互抑制
- **可操作表型群落 (operational phenotypic community, OPC)**
- OPC由相互作用的OPM集合组成，反映不同层次的表型数据之间的关联性和动态相互作用

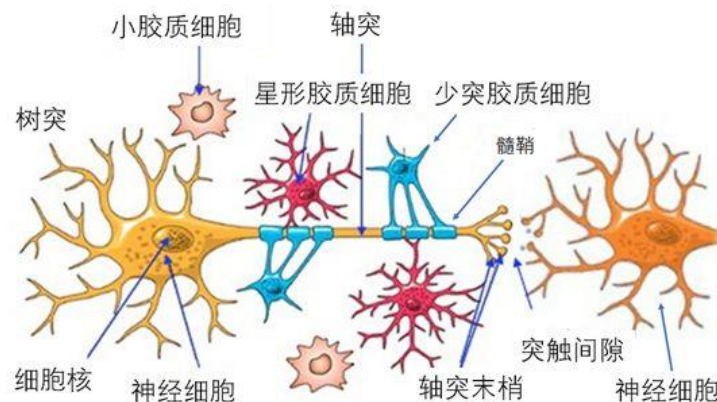
第四节 细胞与类器官表型组学



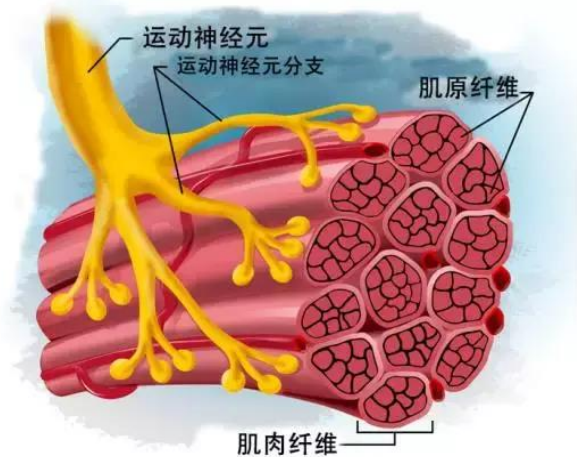
数据库名称	资源类型	网址
OrganoidDB	提供类器官转录组测序和单细胞测序数据等内容。	http://www.inbirg.com/organoid_db/
UCSC Cell Browser	提供类器官单细胞测序数据可视化分析。	https://cells.ucsc.edu/?ds=organoidreportcard+organoids10X&meta=Sample
Gene Expression in Cortical Organoids	提供脑皮层类器官不同培养时期基因表达数据。	http://solo.bmap.ucla.edu/shiny/GECO/
Organoid Cell Atlas	建立类器官的综合单细胞图谱。	https://hca-organoid.eu/
OrganoidBase	患者来源的肿瘤衍生类器官及其所有相关的组织病理学、IC50、基因组、转录组和单细胞分析数据。	https://www.crownbio.cn/databases/organoidbase/
Huborganoids	提供类器官基因组测序、表达谱分析和对已知药物和实验药物的敏感性等信息的数据库。	https://www.huborganoids.nl/
OrganoidID	提供自动识别和分析类器官形态的工具。	https://github.com/jono-m/OrganoidID



缺乏功能性的血管系统



缺乏完备的免疫系统



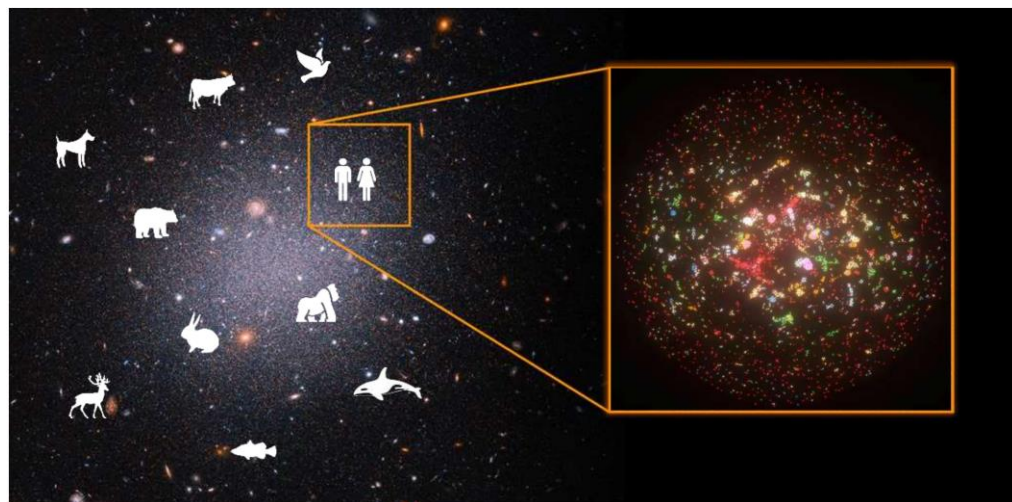
缺乏功能性的神经网络

- 类器官培养流程缺乏标准化，研究结果可靠性和可重复性存在挑战
- 检测表型的时空变化，不同时间点和空间尺度的异质性
- 低通量和缺乏样本资源——类器官芯片和类器官微流控培养技术
- 类器官的使用仍然存在伦理争议

第五节 人类表型组学

人类表型组 (human phenome) 是描述和分类人类个体在形态、生理、行为和疾病等方面的可观察特征的集合。它涵盖了从分子水平到整个个体的各个方面，旨在为理解人类的生理学、行为学、遗传学等提供一个全面的框架。

“人类表型组计划” (human phenome project, HPP) 提出要“测一切之可测”，进行全尺度、全周期精密系统的表型测量，从而解析基因、表型、环境之间的关系，以及宏观和微观表型之间的关联，进一步破解各种表型与人类健康和疾病的关系。



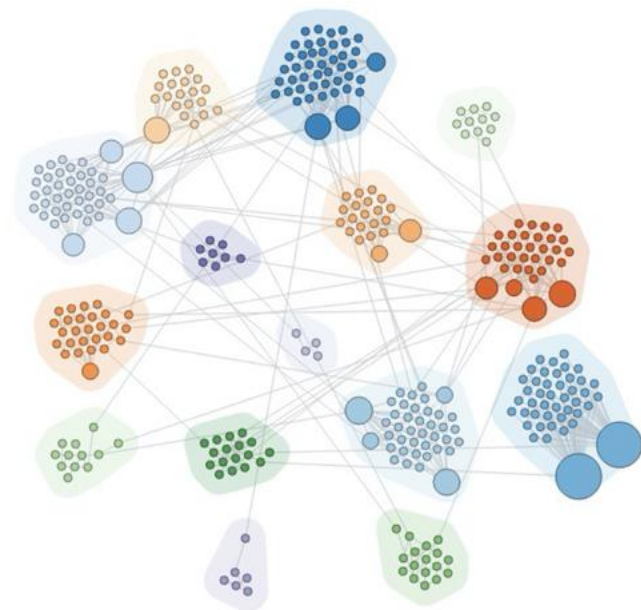
导航图：
生命星空中人类表型组关键网络
图片来源：金力报告

□ 表型测量的标准化

□ 明确表型特征三个分布

- 个体的多次跟踪测量 (n of 1)
- 个体间的差异分布
- 群体间的差异分布

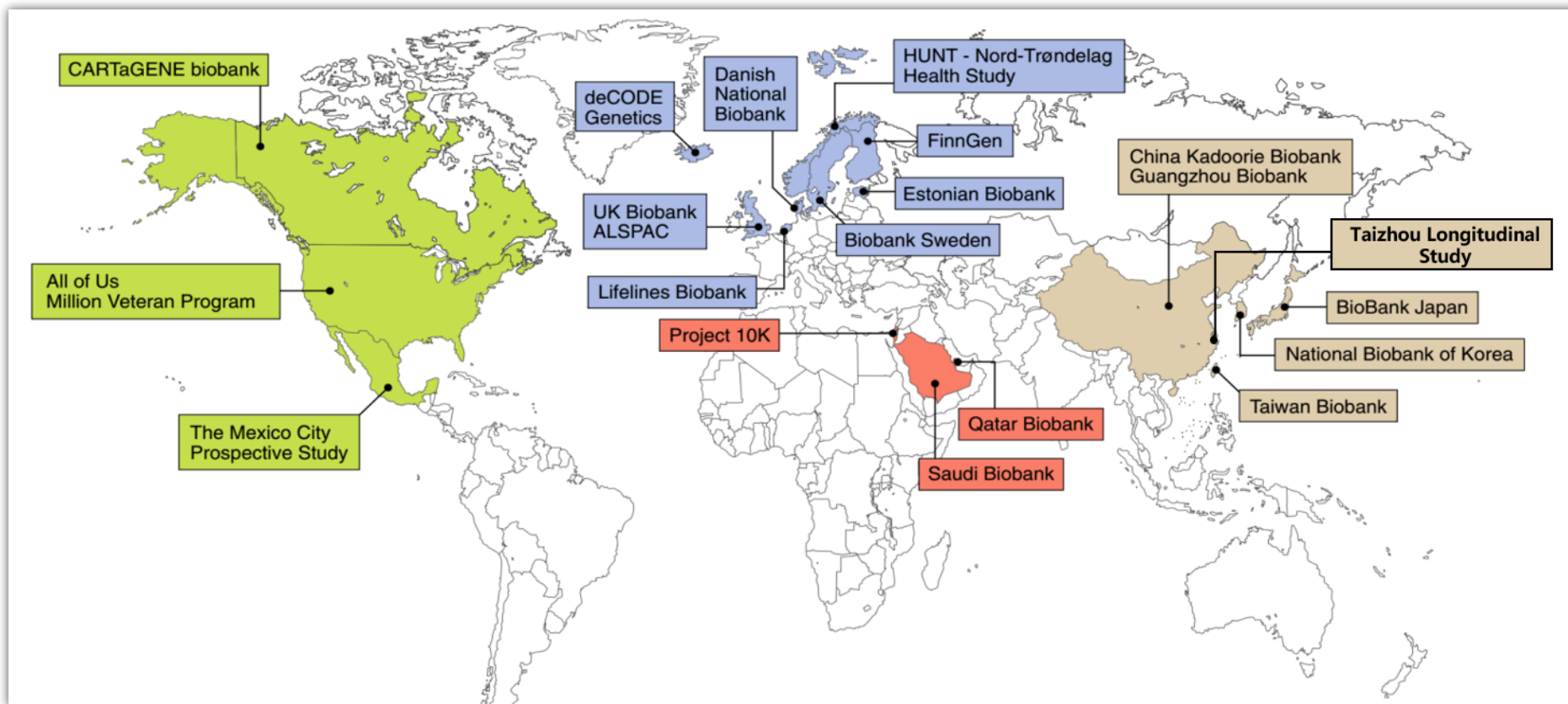
□ 构建表型组参比关系图谱 (导航图)



图片来源：金力报告

数据名称	样本量	人群	地区	年代	描述
环境暴露与疾病表型数据					
英国生物样本库 (UK biobank, UKB)	50万人	50岁以上成人	英国	2006年起	提供人口统计学信息、身体测量数据和生物样本，旨在分析影响健康和疾病的因素。
美国“全民健康”百万人群队列 (all of us, AoU)	40万人	不同种族、年龄、性别、社会背景和健康状况	美国	2018年起	包括人口统计学信息、身体测量数据和生物样本。旨在探索影响健康和疾病的机制，预防、诊断和治疗提供个性化方案。
美国合作性围产期项目 (collaborative perinatal project, CPP)	5.8万	孕妇及新生儿	美国	1959年至1974年	包括产科、儿科、病理、血清学、社会和家庭、历史、心理、语言、语言和听力等多方面信息。
社区动脉粥样硬化风险队列 (the atherosclerosis risk in communities study, ARIC)	1.6万人	45-64岁成人	美国	1987年起	包括医学、社会和行为学信息，用于研究动脉粥样硬化的病因学和自然史，临床动脉粥样硬化疾病的病因学
欧洲癌症与营养前瞻性队列 (European prospective investigation into cancer and nutrition, EPIC)	52.1万人	10个国家	欧洲	1992年起	包括饮食、身体测量数据、生物标志物、医学历史和生物样本等信息，旨在研究饮食、营养状况、生活方式和环境因素与癌症和其他慢性疾病发病率之间的关系。

世界主要国家均已布局大型自然人群队列建设，并持续投入



图片来源：
陈兴栋报告

- ◆ 体量多在10万人以上
- ◆ 长期随访
- ◆ 表型数据丰富

大型自然人群队列

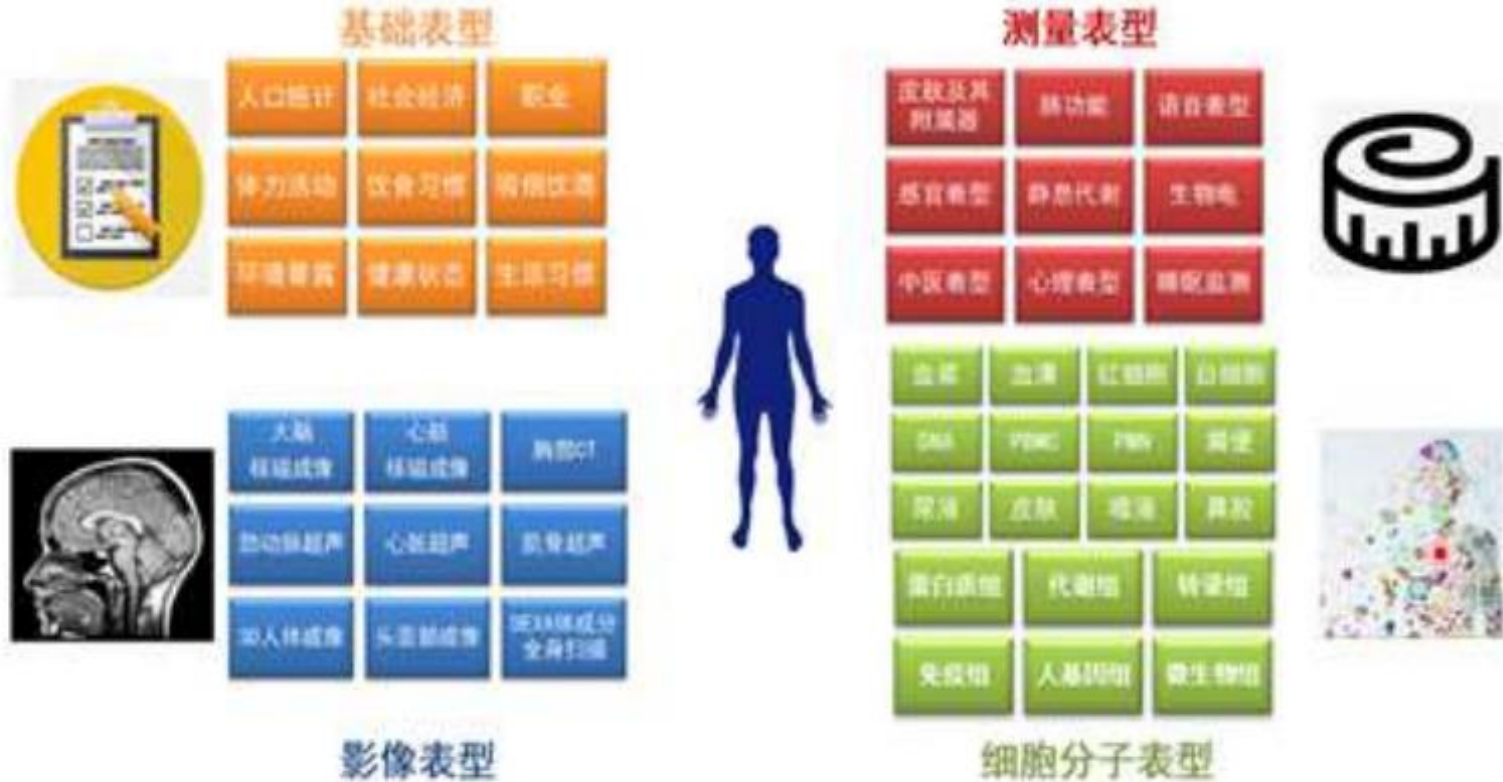


大型生物银行

多维度健康表型网络



完成1005人检测，每人采集约2万个表型指标



人类表型组精密测量平台

分子
表型
测量
子平台

一站式蛋白质组测量平台



亚洲最大代谢组质谱平台



孙春兰副总理、李强书记视察

细胞
表型
测量
子平台

细胞成分与分析平台



单细胞分析平台



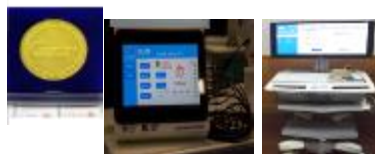
时任上海市市长应勇视察

影像
表型
测量
子平台

三维成像与测量



骨超声诊疗仪与影像



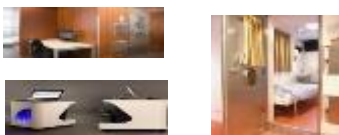
CT



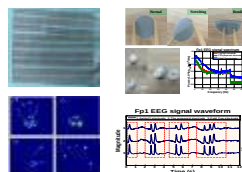
MRI

功能
表型
测量
子平台

人体能量代谢平台



睡眠与生物电



皮肤及其附属器



领域名称	表型词条数
分子蛋白质组	10230
分子代谢组	2153
核酸组	3854
宏基因组	990
细胞	1469
睡眠	294
人体能量代谢	4
生物电	302
皮肤及其附属器	60
感官	1423
语音	181
心理认知	126
生化检测	18
中医	120
影像	942
超声	228
人体外观测量	324
DXA	545
功能	17
体检	99
健康问卷	665
合 计	24044

图表来源：金力报告

第六节 动植物表型组学

动物表型组学 (animal phenomics) 是研究动物表型 (即生物学上可观察的特征和性状) 的综合学科, 它结合了生物学、遗传学、基因组学、统计学和计算科学等领域的知识和技术。

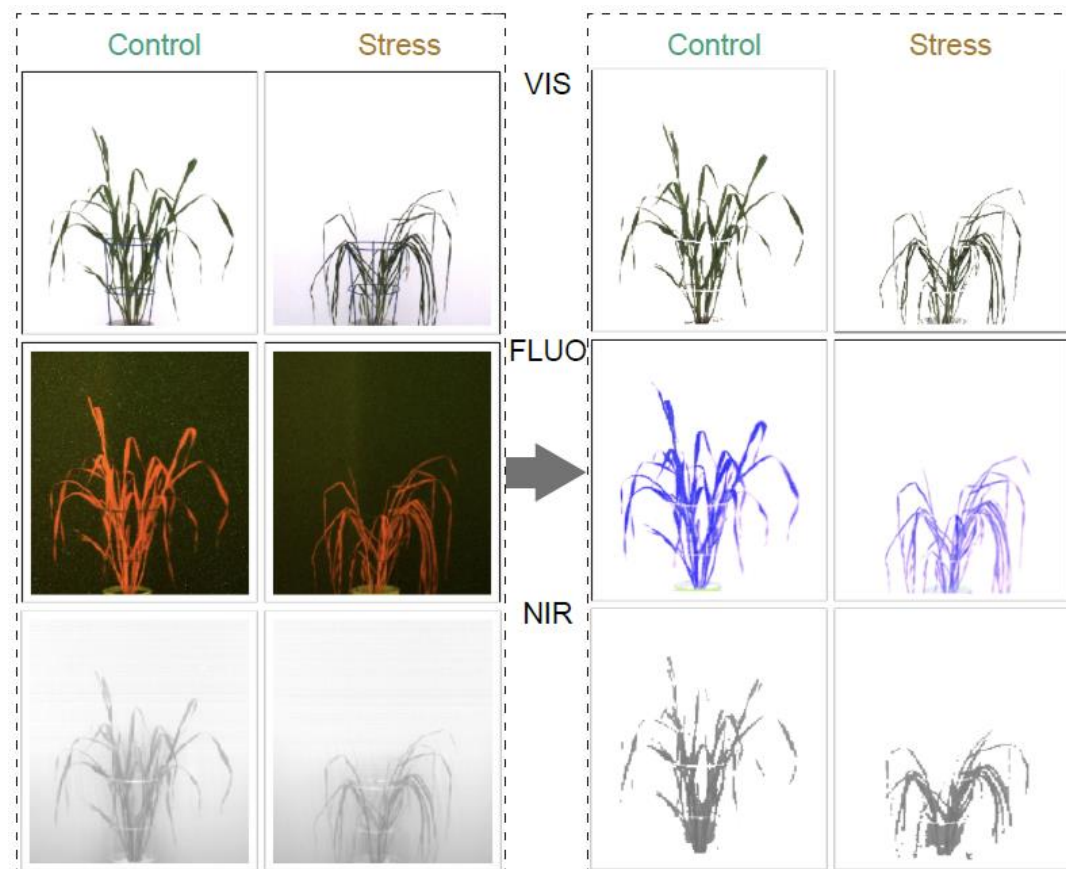
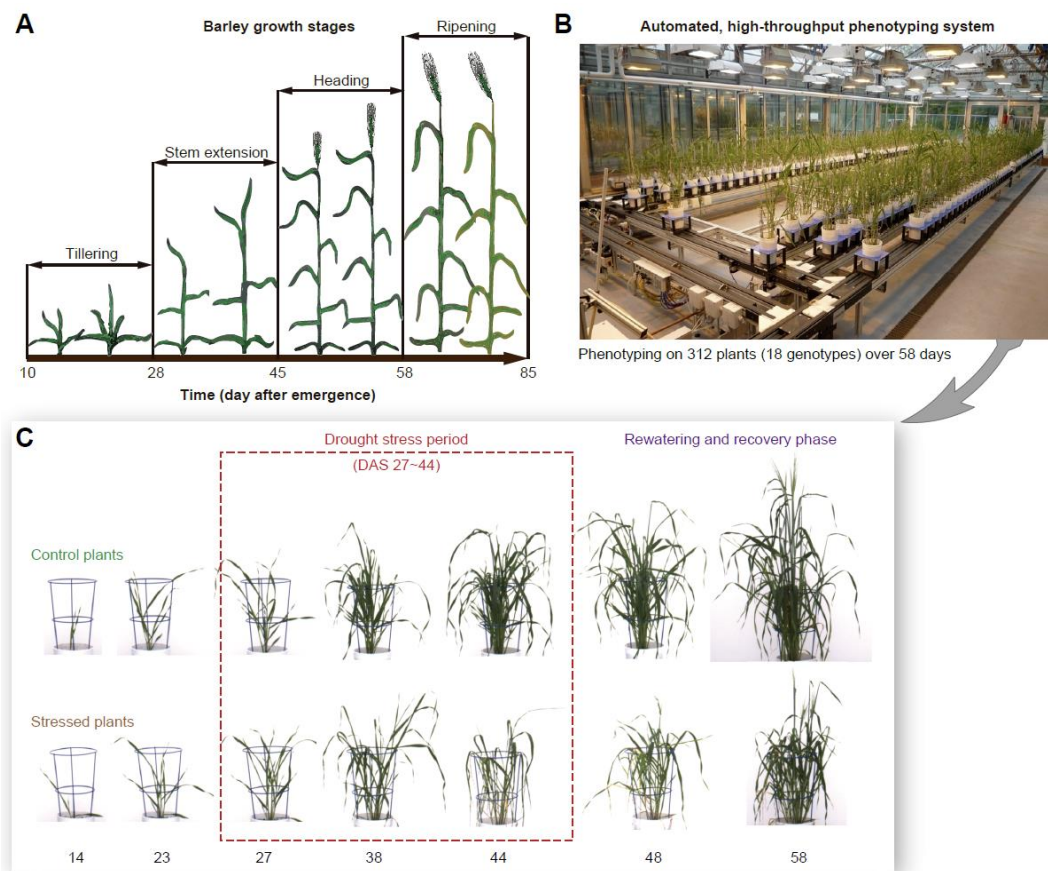
植物表型组学 (plant phenomics) 是一门研究植物的可观测特征和性状的学科, 它利用先进的高通量技术和数据分析方法, 系统性地描述、量化和分析植物的生物学特性, 包括形态、结构、生长、发育、生理、代谢、抗性等各方面的性状。

• 植物表型组研究中心和学术组织

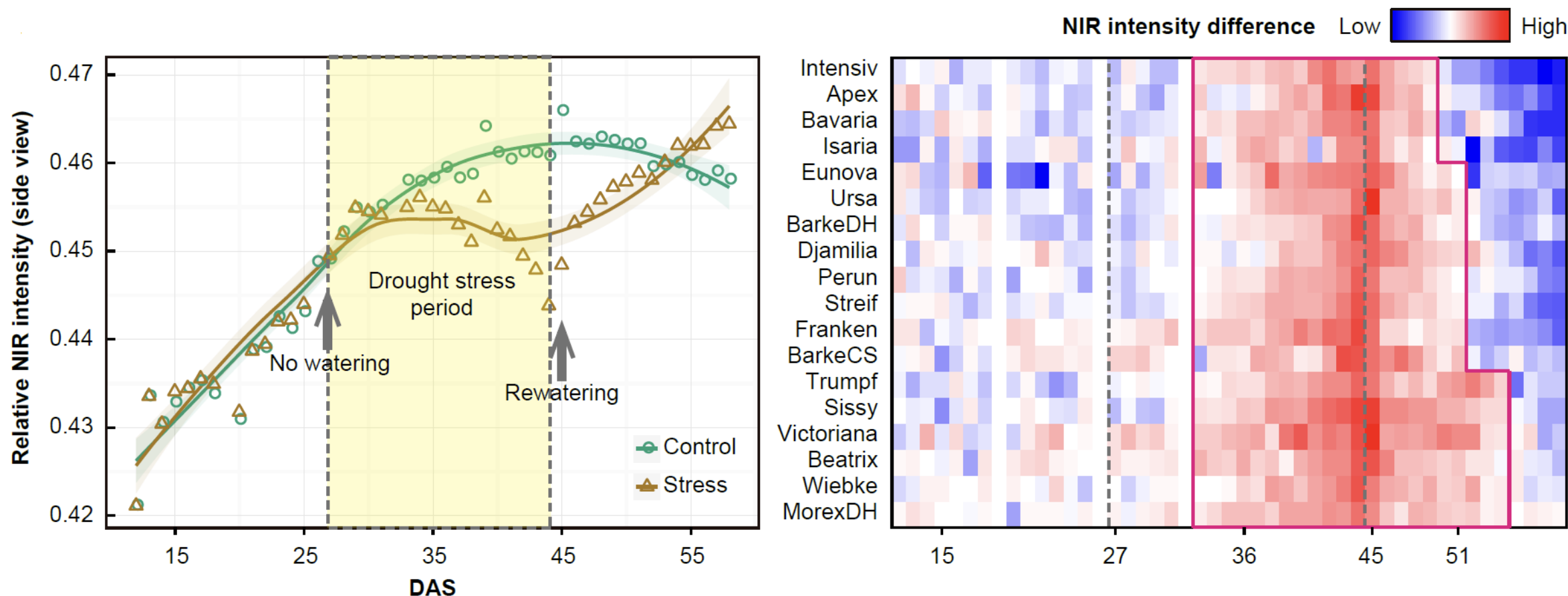
- 德国植物遗传和作物研究所 (IPK)
- 澳大利亚植物表型研究中心 (APPF)
- 德国Jülich植物表型研究中心 (JPPC)
- 荷兰植物生态表型研究中心 (NPEC)
- 英国国家植物表型研究中心 (NPPC)
- 加拿大植物表型与影像研究中心 (P2IRC)
- 国际植物表型协会 (IPPN)
- 欧盟植物表型协会 (EPPN)
- 德国植物表型协会 (DPPN)
- 华中农业大学作物表型中心
- 南京农业大学植物表型学研究中心
- . . .



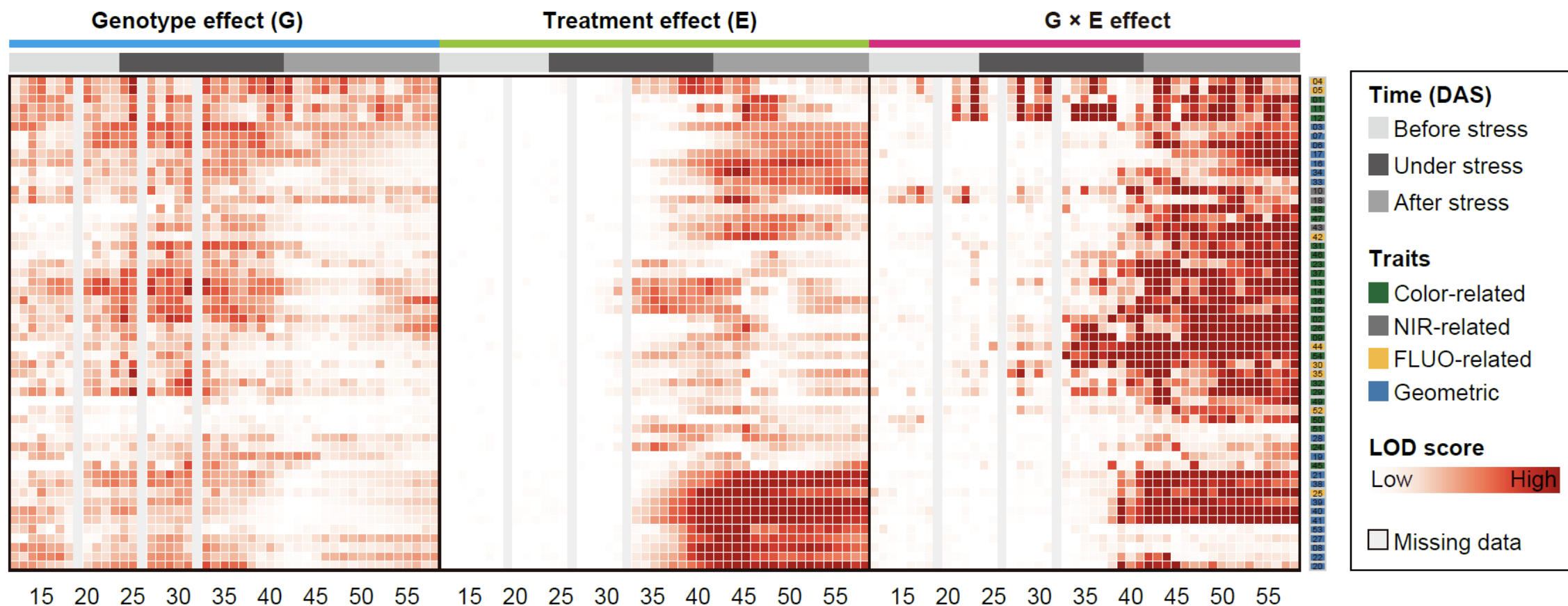
• 植物生长及抗逆反应 (大麦)



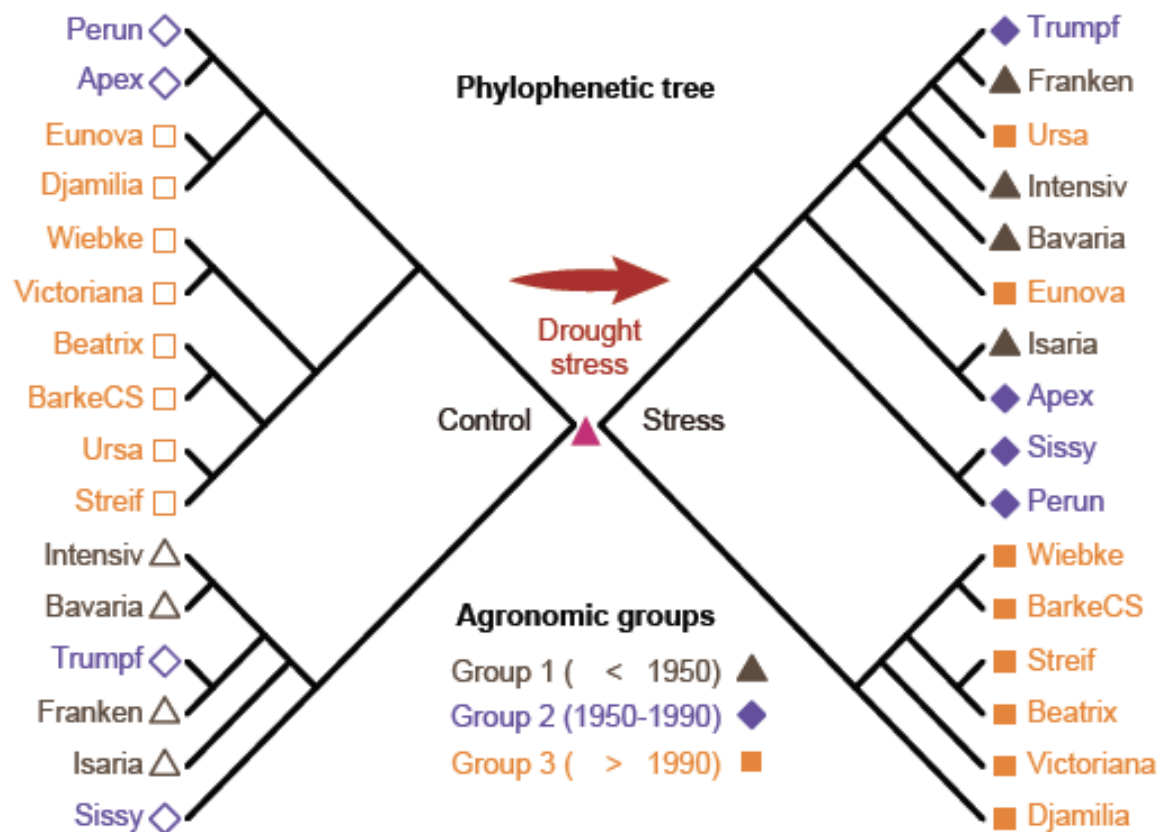
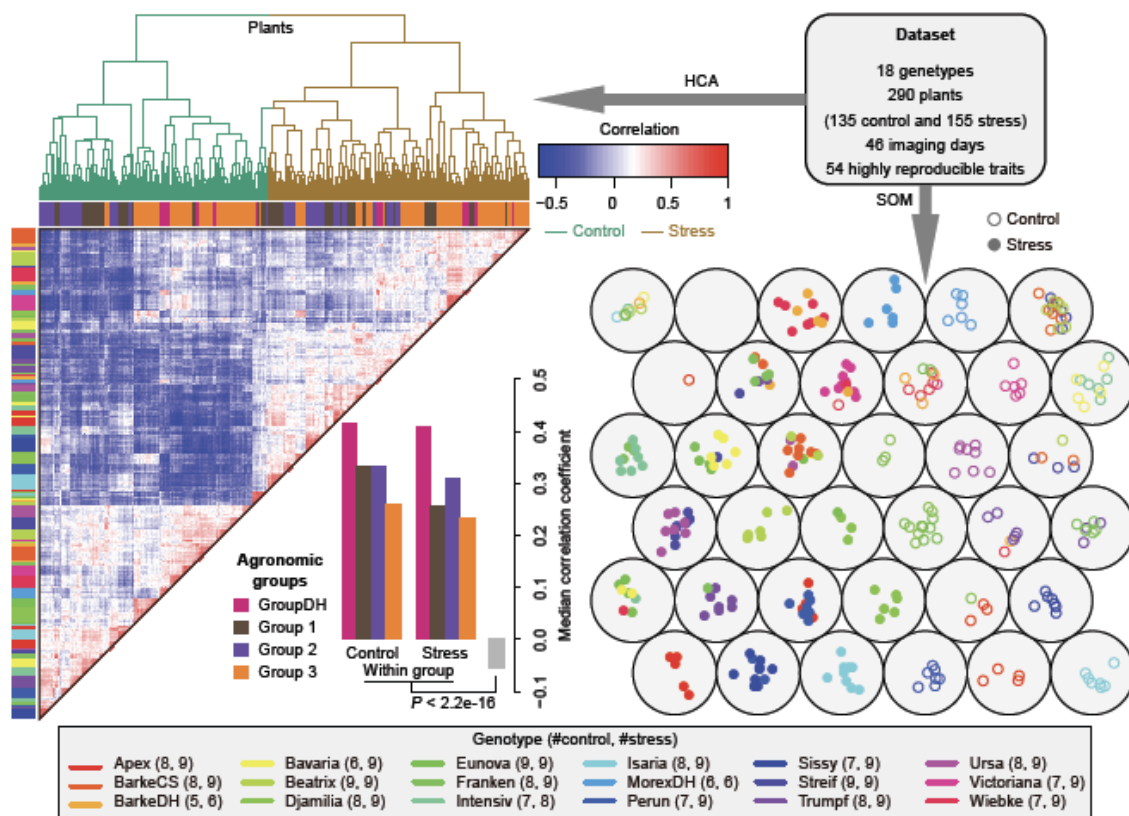
• 图像特征可以模拟植物抗旱反应（从近红外光谱图像数据推断）



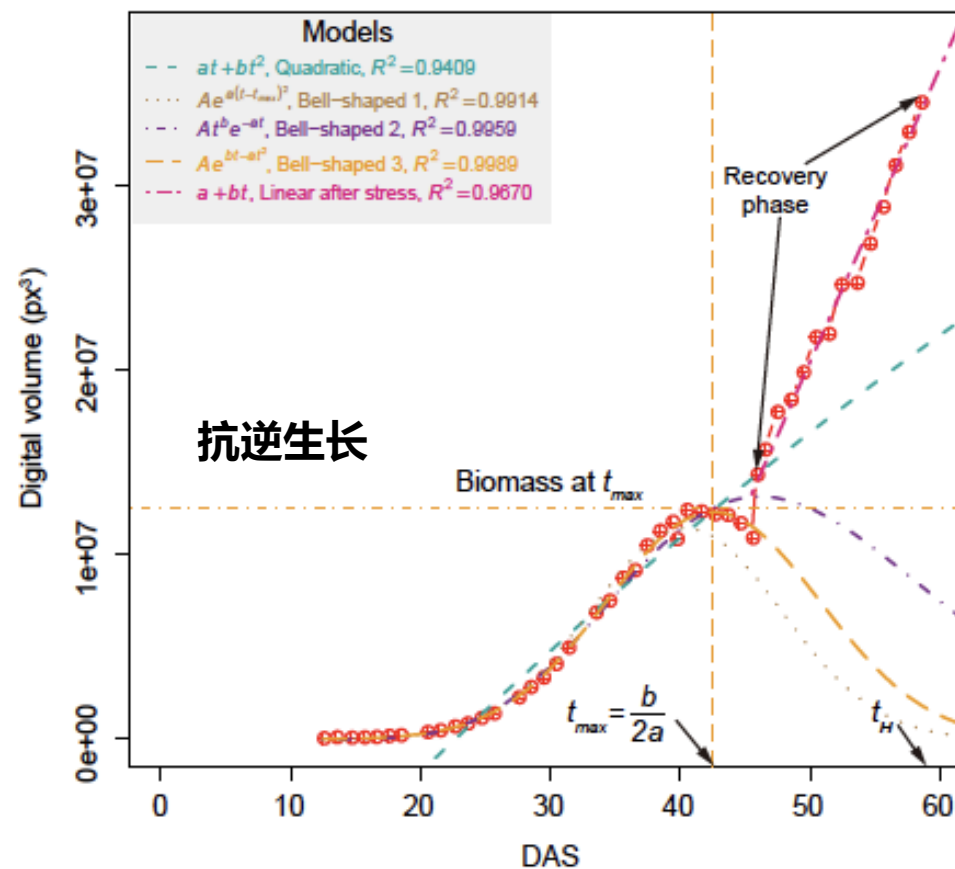
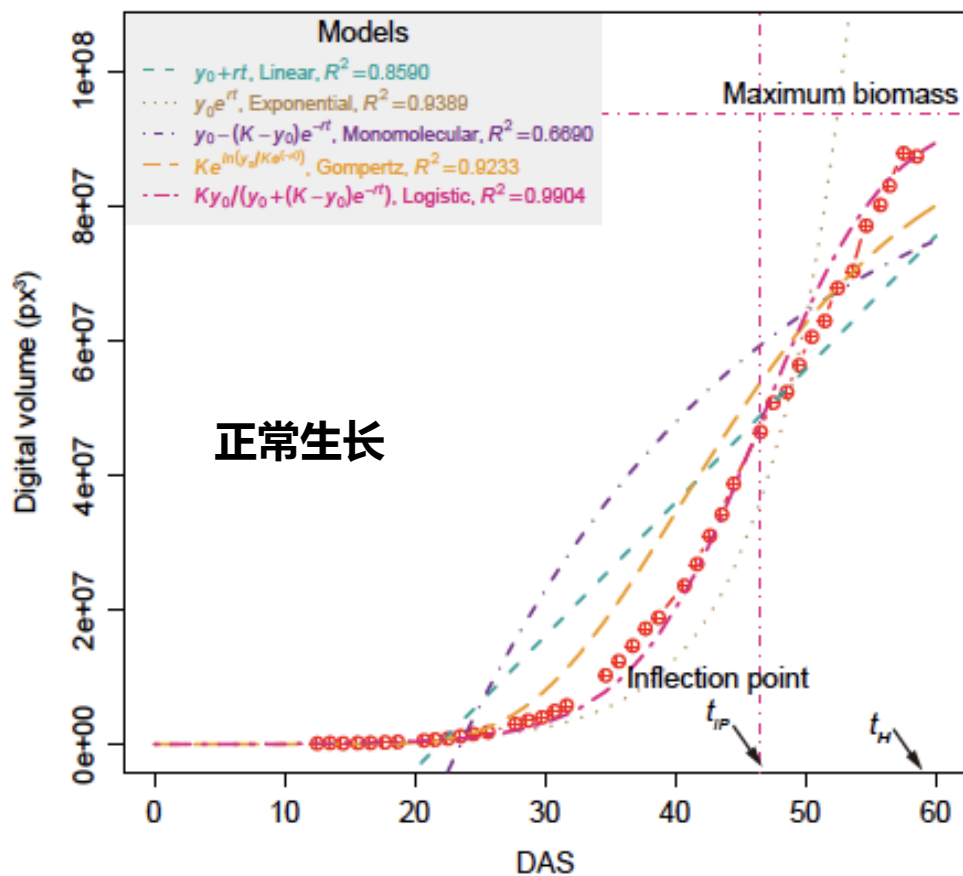
• 图像特征可以模拟植物抗旱反应 (线性混合模型LMM)



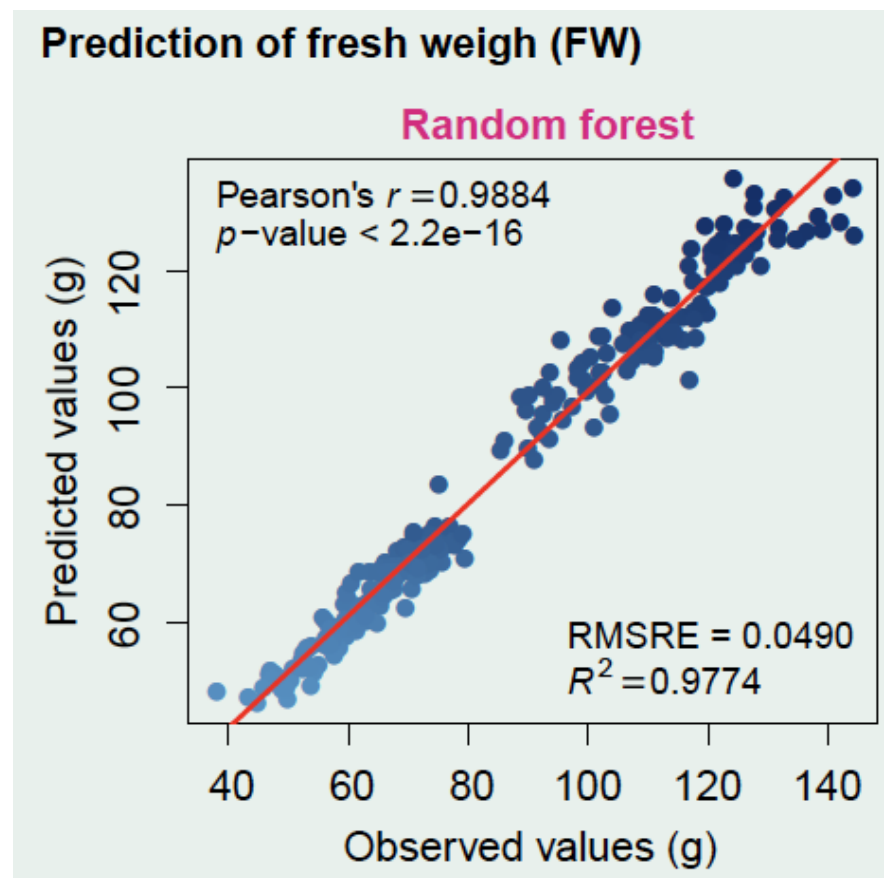
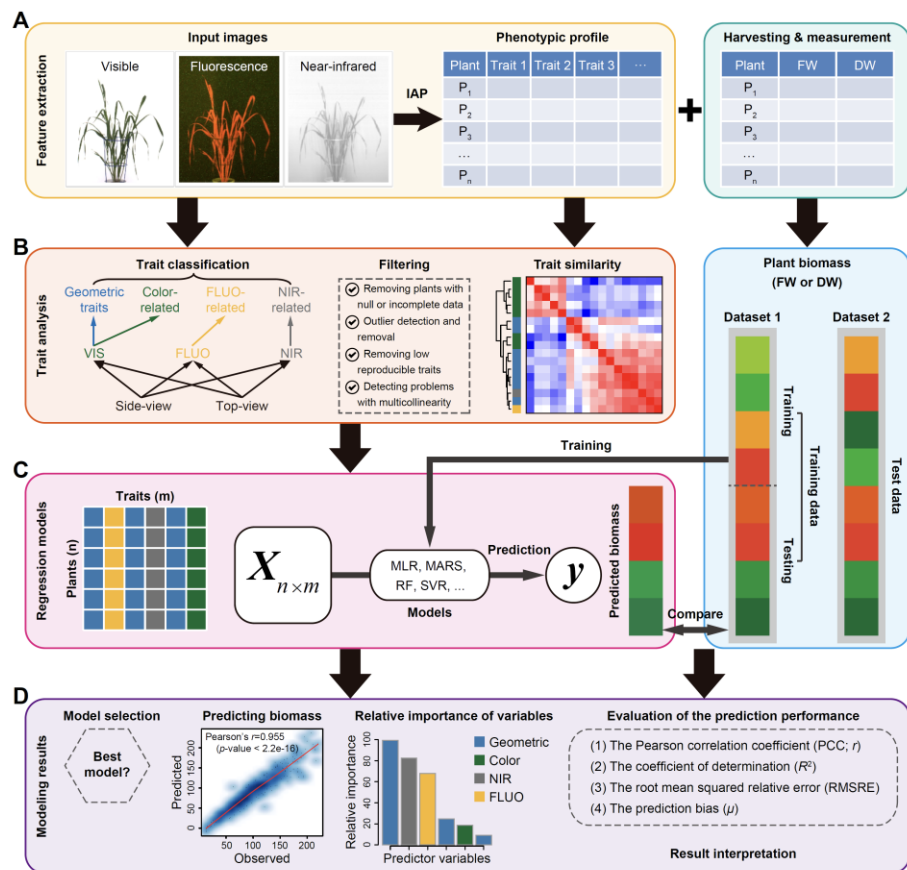
• 图像特征可反映不同的品种**对抗旱的反应差异** (无监督学习)

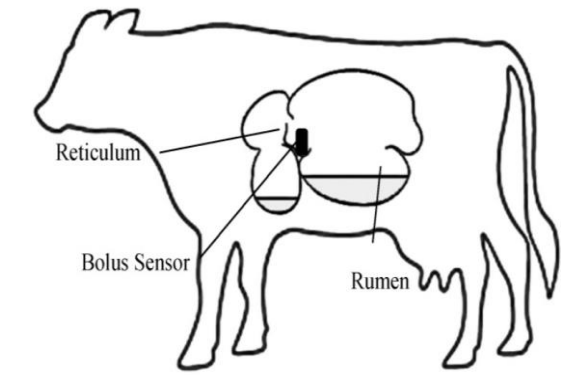


• 图像特征可用于**模拟植物生长** (数学建模: 生长模型)

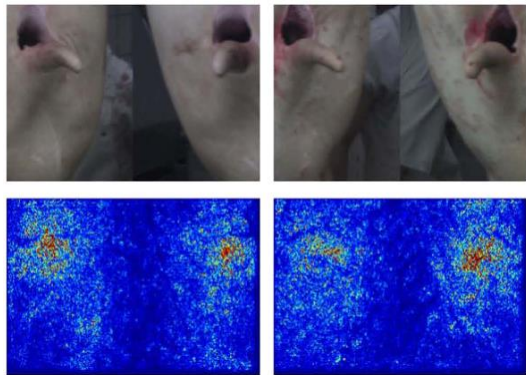


• 图像特征可用于预测植物生物量 (有监督学习)





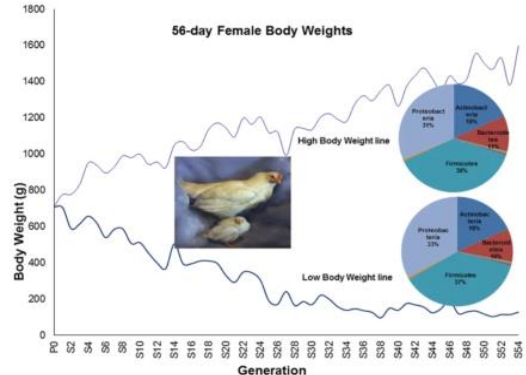
生物标志物检测



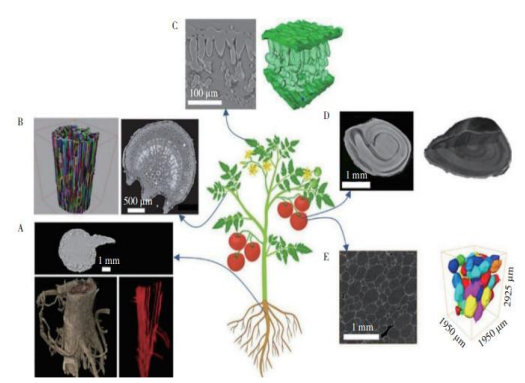
牲畜行为监测



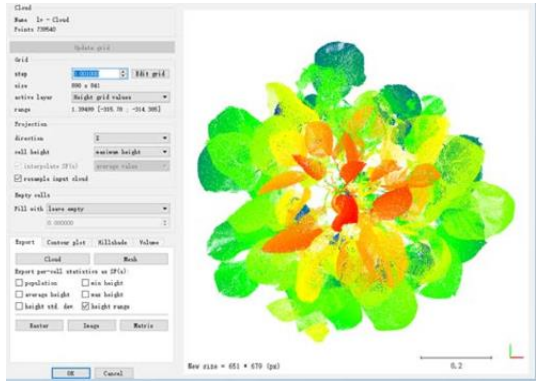
牲畜饲喂系统设计



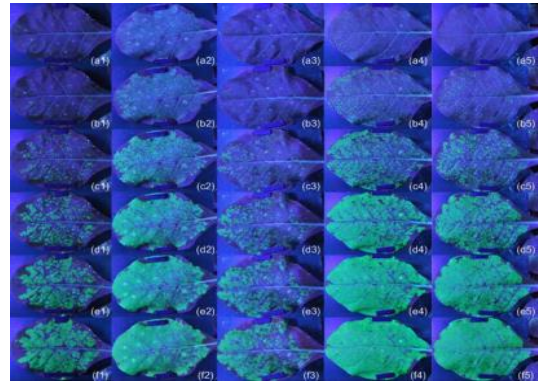
牲畜肠道微生物组分析



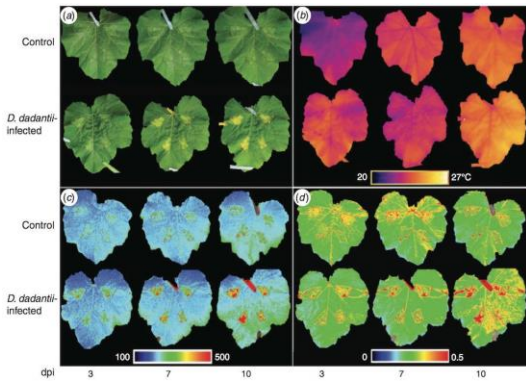
作物根系研究



作物冠层生物量估测



光合作用效率和叶绿素含量



受生物或非生物胁迫研究

第七节 未来展望

未来趋势：

- 多功能、高通量的表型鉴定平台，快速、精准和全面采集数据。
- 需要不同学科领域的知识融合
- 大数据技术和人工智能挖掘表型数据中的深层次信息，发现新的生物学规律和关联。
- 为个性化医学提供重要支持，制定更加精准的健康干预策略。

面临挑战：

- 需要高度重视隐私保护和伦理原则
- 科研监管方面的规范和指导不足，对研究进行监督和评估
- 数据整合、标准化和共享方面面临技术和政策上的挑战
- 将人类表型组学的研究成果应用于医疗实践需要克服技术和方法上的难题

表型组学将在技术、合作、数据应用等方面取得显著的进展，为生物学、医学和环境科学等领域带来新的突破和发展机遇。

请各位专家提出宝贵意见！

谢 谢 ！

1. **版权声明：**本PPT及其所有内容（以下简称“本PPT”）仅用于教育和教学用途，版权归属于本PPT作者。
2. **使用要求：**任何使用本PPT的行为均须遵守以下条件：
 - 1) **致谢和标注：**若部分或全部使用本PPT的内容，请在使用内容的适当位置标注出处，并致谢本PPT作者。
 - 2) **修改和再分发：**未经作者书面许可，不得对本PPT进行修改或再分发。
3. **禁止商业化使用：**严禁将本PPT用于任何形式的商业化用途，包括但不限于：
 - 1) 通过网络或其他途径进行付费使用或分发；
 - 2) 在商业培训、广告或其他商业活动中使用本PPT的内容。
4. **法律责任：**任何违反上述条款的行为，作者保留追究法律责任的权利，包括但不限于：
 - 1) 要求停止侵权行为；
 - 2) 追究侵权使用者的经济赔偿责任。
5. **其他规定：**
 - 1) 本使用条款的解释权归本PPT作者所有。
 - 2) 作者保留随时更新本使用条款的权利，更新后的条款将即时生效。

