GWAS分析

生信帮
云南农业大学
云南省生物大数据重点实验室
2023年8月27日

目录

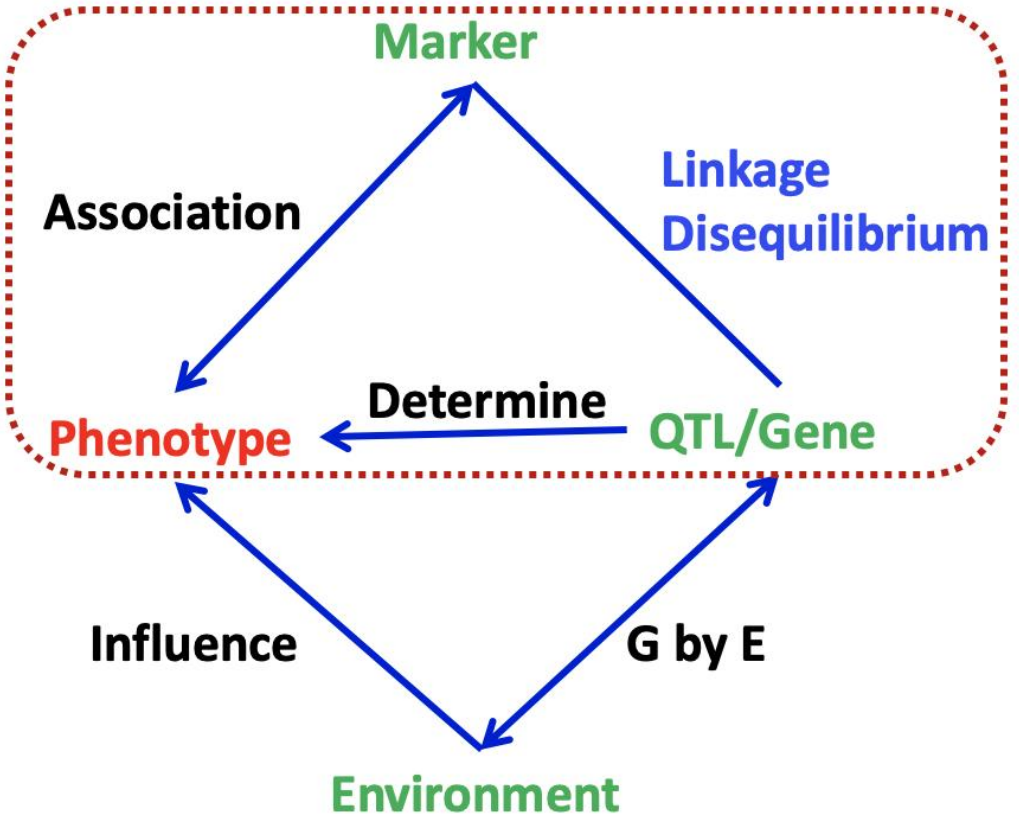
目录

GWAS基本原理

群体遗传学中，目前主要基于二三代测序技术进行功能基因定位与克隆，根据研究的群体类型，将研究方法分为两大类：

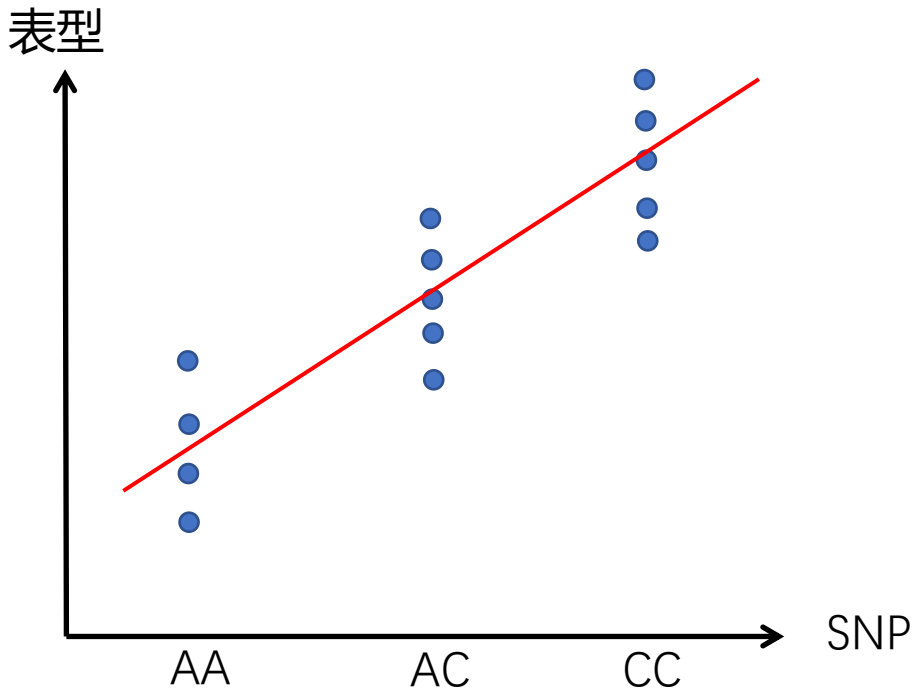
GWAS基本原理

全基因组关联分析（GWAS)是多个个体在全基因组范围内遗传标记（如SNP, InDel, SV)多态性进行检测（获得基因型），进而将基因型与可观测的性状（即表型）进行群体水平的统计学分析，根据统计量显著性p值筛选出最有可能影响表型的遗传标记，挖掘与性状变异相关的基因。

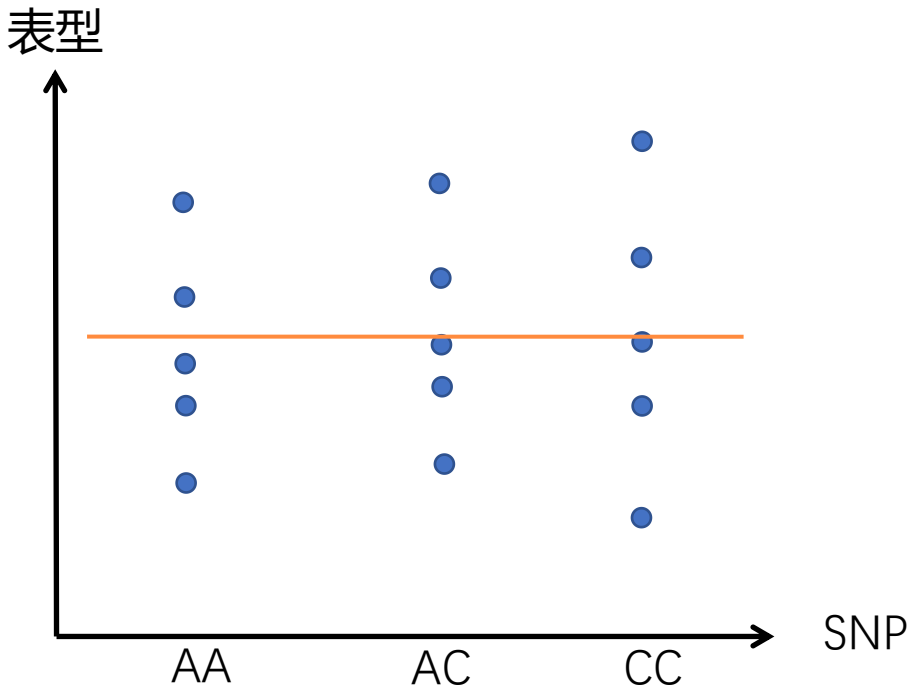


GWAS基本原理

什么是相关性?



SNP与表型显著相关



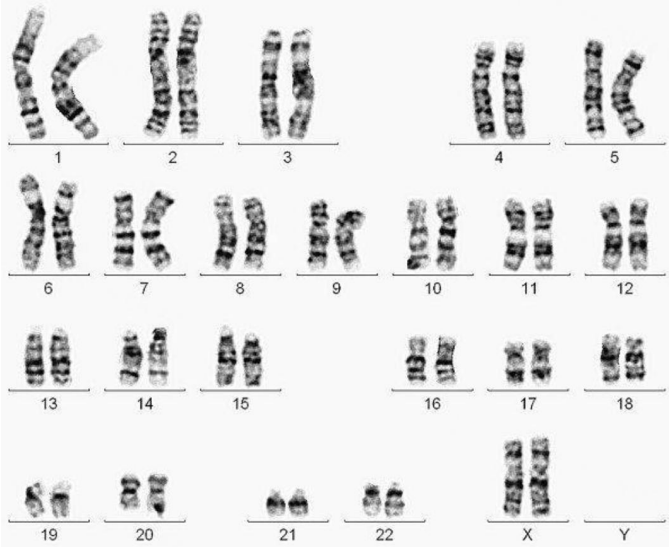
SNP与表型无关

GWAS基本原理

云南省生物大数据重点实验室

生信帮

DNA是GWAS分析的遗传物质基础



核型
Karyotype

C	T	A	G	T	A
C	T	T	G	T	A
G	T	A	C	T	A
C	T	A	G	T	A
G	T	A	G	T	A
C	T	A	G	T	A
C	T	A	C	T	A
C	T	T	G	T	A

单倍型
Haplotypes

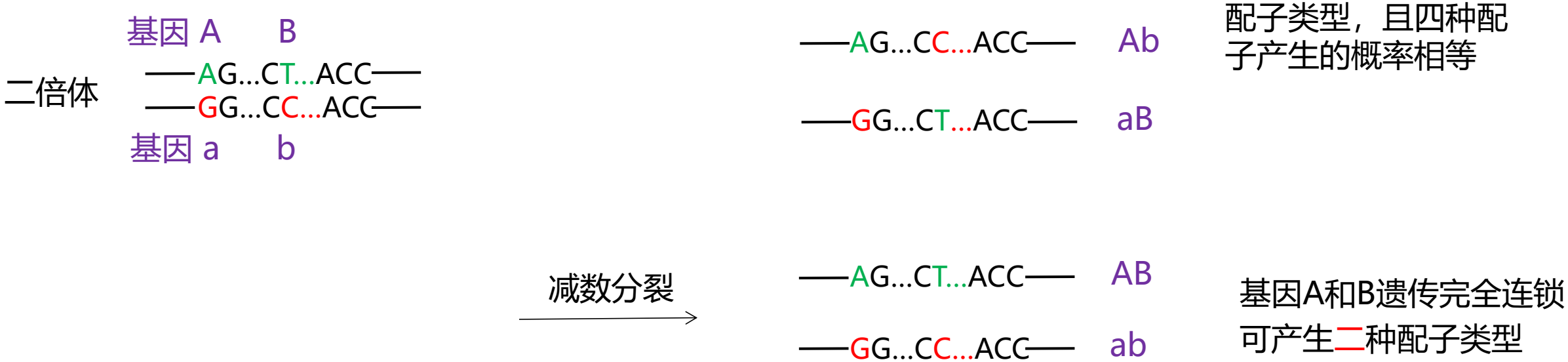
C/C	T/T	A/T	G/G	T/T	A/A
G/C	T/T	A/A	C/G	T/T	A/A
G/C	T/T	A/A	G/G	T/T	A/A
C/C	T/T	A/T	C/G	T/T	A/A

基因型
Geneotypes

表型
Phenotypes

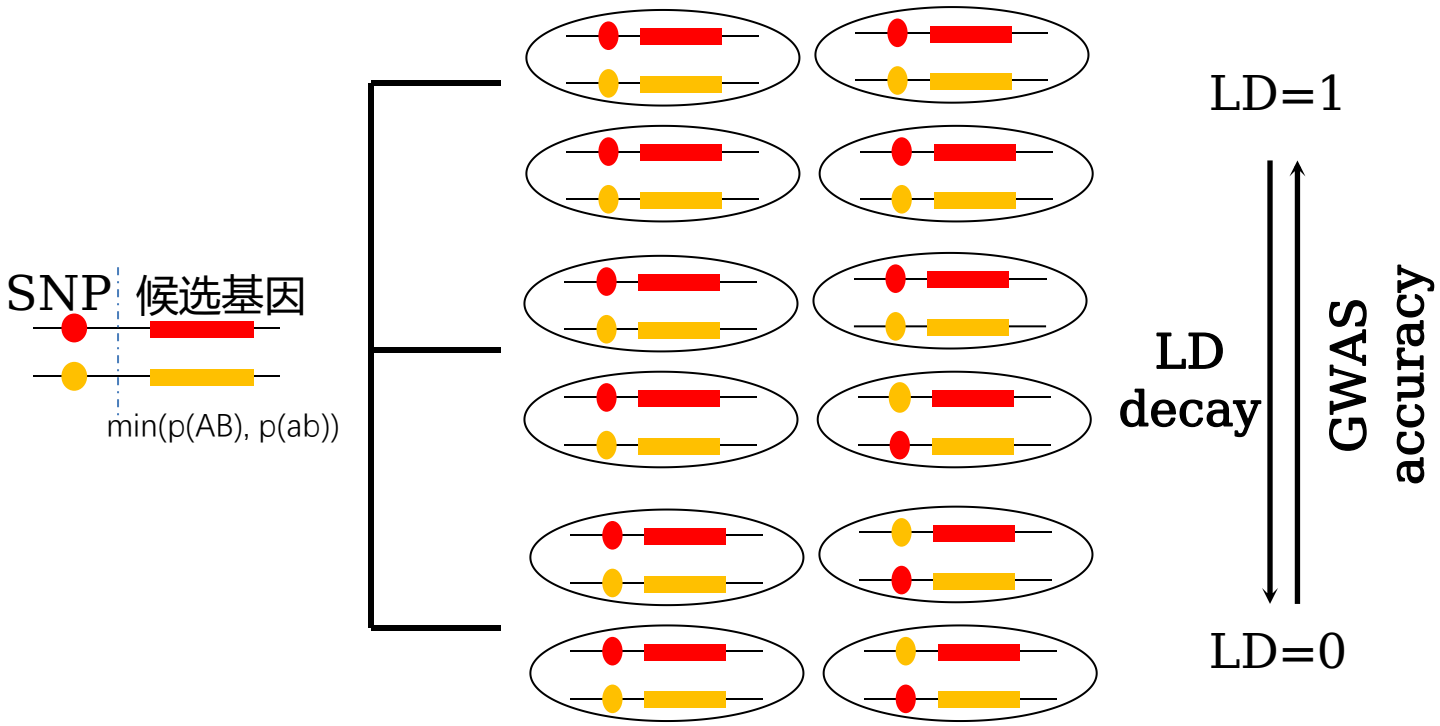
GWAS基本原理

- **遗传连锁：** 个体的不同基因在同一条染色体上的物理位置相近，倾向于一同遗传



GWAS基本原理

连锁不平衡 (linkage disequilibrium, LD)：是指同一条染色体上，等位基因与SNP位点之间的非随机相关。即当位于同一条染色体的等位基因与邻近的SNP位点同时存在的概率大于群体中因随机分布同时出现的概率时，就称这该等位基因与SNP位点处于LD状态。



LD

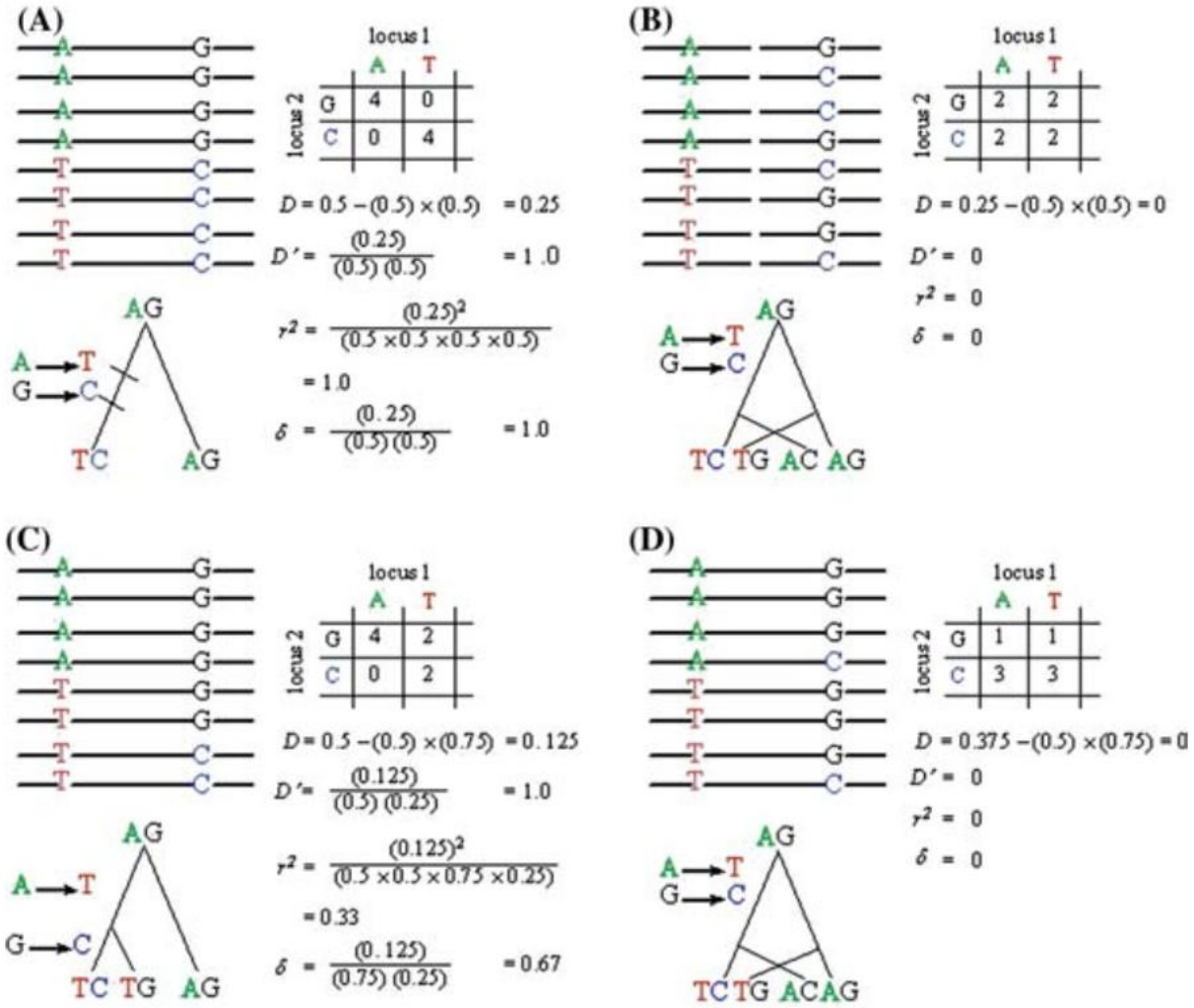
LD（连锁不平衡） 的基本单位D, 是度量观察到的单倍型频率与平衡状态下期望频率的偏差(D=P(AB)-P(A)*P(B))。虽然D能够很好的表达LD的基本含义，但是由于其严格依赖于等位基因频率（allele frequency），故不适合应用于表述实际的LD强度，尤其是进行不同研究的LD值的相互比较。几个常用于度量LD的符号中，最重要的是D'和r^2，两者都是基于D，各有各的特点及用途。

迁移、突变、选择、有限的群体大小以及其他引起等位基因频率改变的因素都会引起LD的改变。

LD常见分析包括LD衰减距离分析和LD block分析。

GWAS基本原理

LD计算



D=P(AB)-P(A)*P(B);

P(AB)表示实际观察到的AB频率，P(A)*P(B)表示AB频率的期望值。

D'=D/Dmax

当D<0, Dmax=min{P(A)P(B),P(a)P(b)};

当D>0, Dmax=min{P(A)P(b),P(a)P(B)};

r2=D*D/(P(A)P(a)P(B)P(b))

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

LD衰减距离分析(LD decay)

- LD衰减是指位点间由连锁不平衡到连锁平衡的演变过程，一般以LD下降到最高值的一半时经过的距离作为LD的衰减距离（也可以用LD下降到特定值，如0.2或者0.1，经过的距离作为衰减距离）。

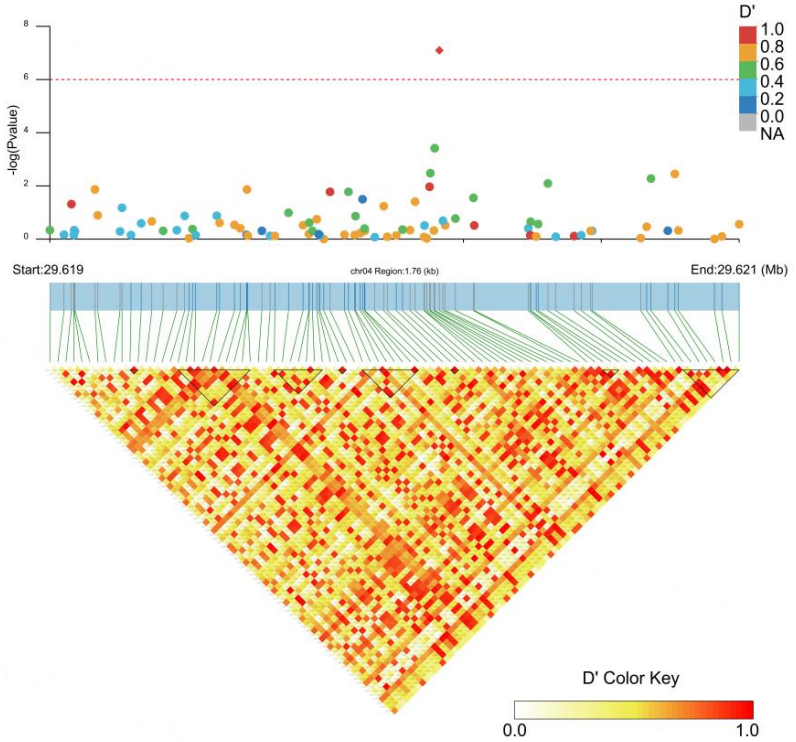
- LD的衰减距离决定关联分析所需标记密度，也在一定程度上决定关联分析的精度。理论上，GWAS最低饱和标记密度=基因组大小/LD衰减距离，衰减距离越短，所需要的标记量就越大，关联的精度越高。

- 较大的样本量，均匀的标记分布是获得平滑和真实LD衰减的前提，重测序数据进行LD分析时效果一般好于简化测序和芯片结果，因为重测序数据获得的标记在基因组上的分布较为均匀。

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

LD block分析

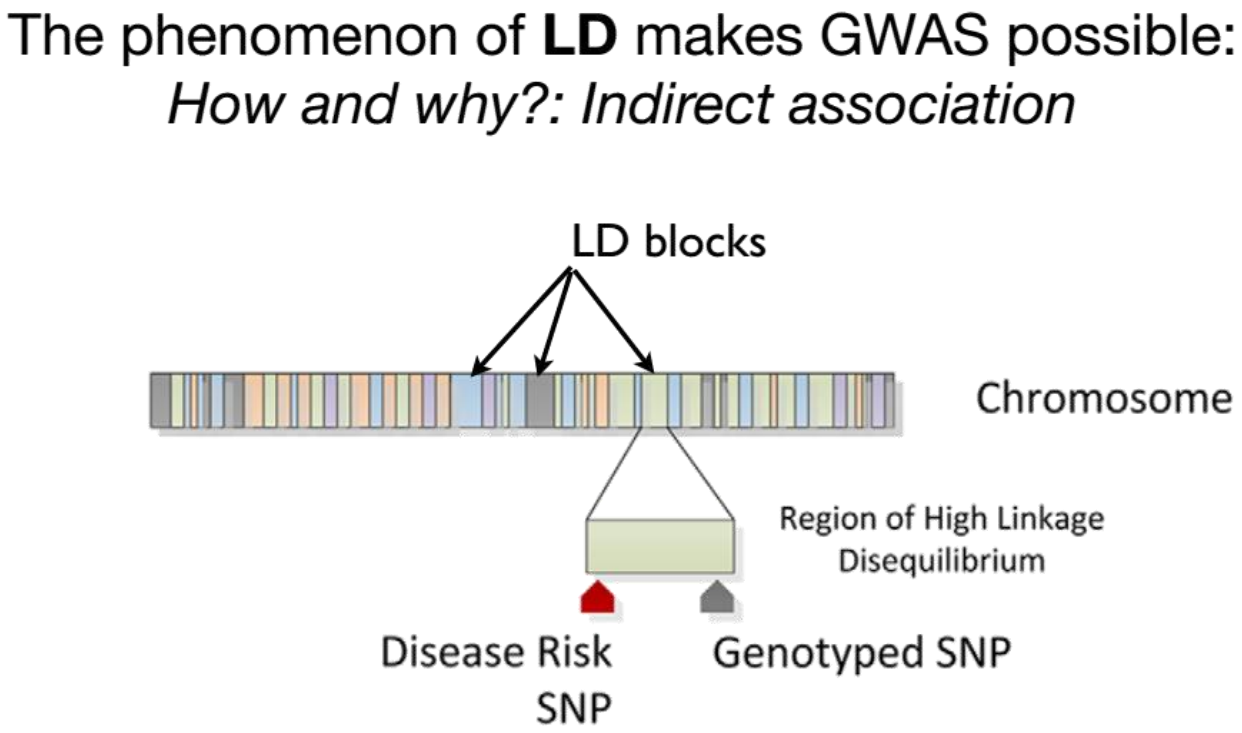
1. LD block也称单倍型块（Haplotype block)，是指同一条染色体上处于连锁不平衡状态的一段连锁区域。
2. 单体型块分析可以用于筛选tag SNP、确定候选基因的范围等。



GWAS基本原理

LD与GWAS

从理论上说，关联分析的分辨率能达到单碱基水平，但很多的实际项目中，很难直接关联到功能SNP，这时就只能以显著的SNP为线索，在其附近找到对性状起决定作用的功能变异，由于LD的存在，使得这种做法成为可能



GWAS基本原理

GWAS材料群体类型

➢ 自然群体材料

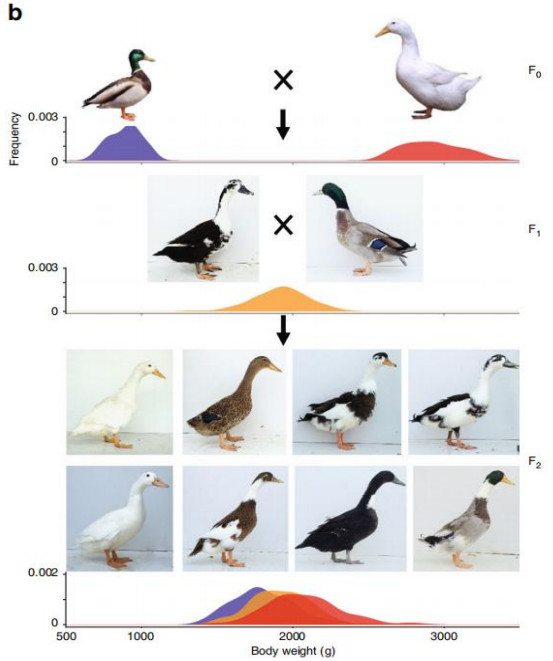
特点：

1. 遗传变异丰富，可以同时对多个性状进行分析；
2. 群体结构复杂，稀有变异多，遗传信息丢失明显。

➢ 遗传群体材料

特点：

1. 背景单纯，检测功效高；
2. 可以放大稀有变异；
3. 遗传变异不够丰富，重组事件有限，定位精度较低。

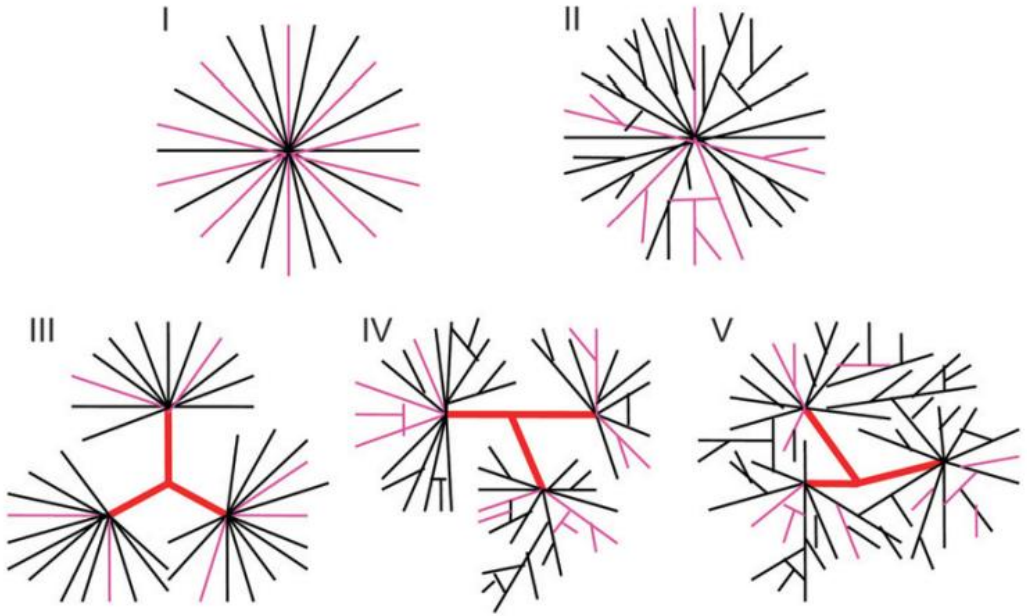


GWAS基本原理

GWAS材料选择的基本原则

1. 遗传变异和表型变异丰富，材料代表性强；

2. 群体结构分化，亲缘关系不能过于明显；



(I) 具有微弱人群结构和家族关系的理想样本；
(II) 具有家族关系的样本；
(III) 具有人群结构的样本；
(IV) 既具有人群结构又具有家族关系的样本；
(V) 具有分化、混合和家族关系的样本；

GWAS基本原理

样本群体足够大

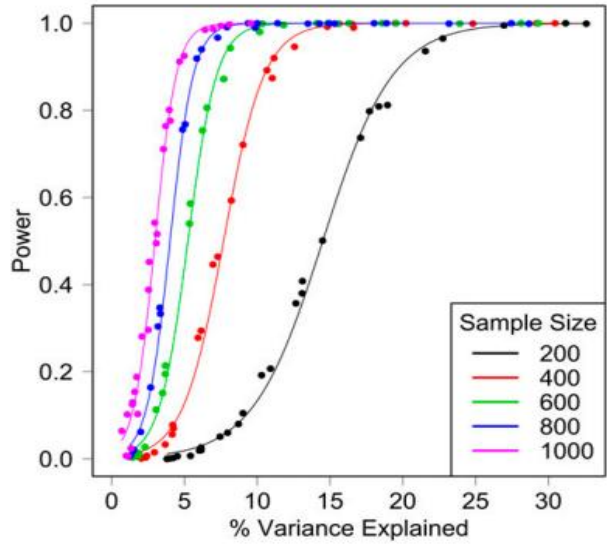


Figure 4 Power simulations demonstrate the relationship between power, sample size, and percentage variance explained. The percentage variance explained by the simulated QTL is plotted *vs.* the power to detect the simulated QTL for five different sample sizes. Points are the mean value from 1000 simulations and curves are logistic regression models fit to the data.

物种	年份	杂志	测序方法	群体数量	研究内容
水稻	2018	*Nature*	重测序	3010	产量和抗性
棉花	2018	*Nature Genetics*	重测序	243	多个农艺性状
棉花	2018	*Nature Genetics*	重测序	419	纤维品质与产量
棉花	2017	*Nature Genetics*	重测序	318	纤维与产量性状
棉花	2017	*Nature Genetics*	重测序	352	纤维品质
水稻	2016	*Nature Genetics*	重测序	176	抽穗期，株高，产量等性状
水稻	2016	*Nature Genetics*	重测序	342	籽粒
水稻	2015	*Nature*	重测序	10072	杂种优势
芝麻	2015	*Nature Communication*	重测序	705	油份相关性状
疟原虫	2015	*Nature Genetics*	重测序	1612	抗药相关
大豆	2014	*Nature Biotechnology*	重测序	302	产油和形态相关
水稻	2014	*Nature Genetics*	重测序	520	代谢相关
黄瓜	2014	*Science*	重测序	115	黄瓜苦味
牛	2014	*Nature Genetics*	重测序	234	经济性状
香菇	2014	*Nature Genetics*	重测序	360	果实颜色

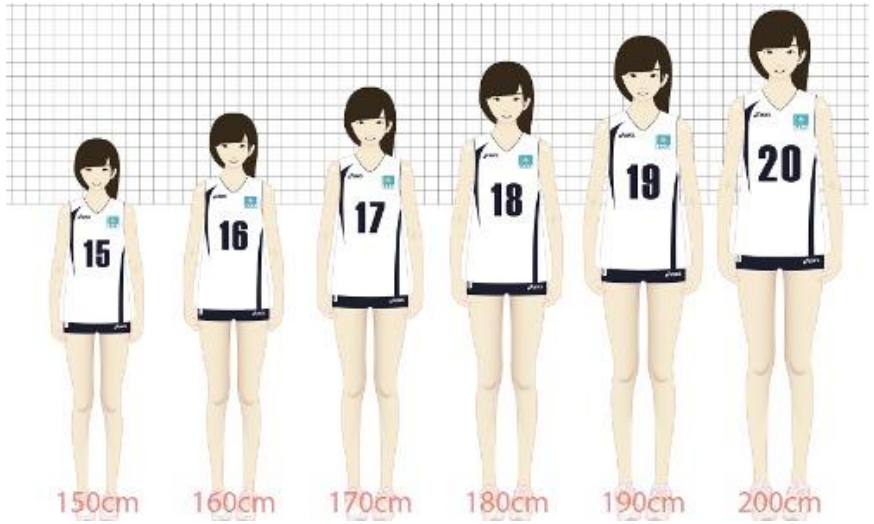
1.位点的效应越低，需要的样本量越大；

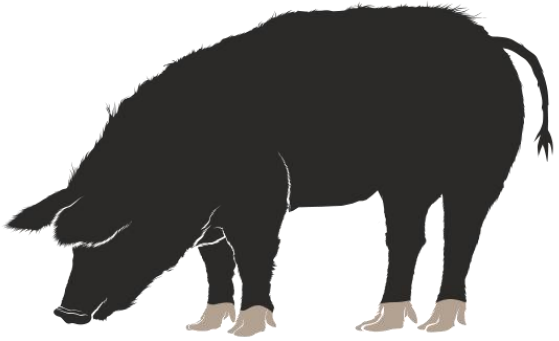
2.非稀有变异中，对中等变异解释率（10%左右）的位点的检测功效要达到80%以上时，需要的样本量在400左右

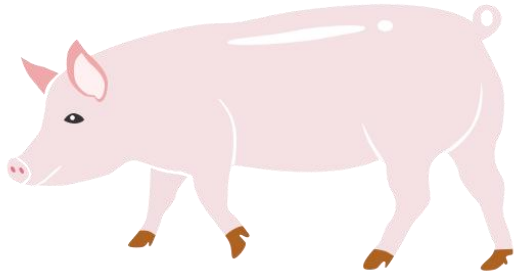
表型考察及质控

性状的定义　　任何可以表征或者量化的表型特征

Seed		Flower	Pod		Stem	
Form	Cotyledons	Color	Form	Color	Place	Size
Round	Yellow	White	Full	Yellow	Axial pods	Tall
Wrinkled	Green	Violet	Constricted	Green	Terminal pods	Short
1	2	3	4	5	6	7







表型考察及质控

质量性状

特点：

- ✓ 一般受单基因或者少数几对基因控制；
- ✓ 性状相对稳定，不易受到环境影响；
- ✓ 其变异在群体内的分布是间断的，即使出现有不完全显性杂合体的中间类型也可以区别归类；
- ✓ 遗传关系较简单，一般服从三大遗传定律（分离定律，自由组合定律和分离组合定律）；
- ✓ 每类性状在群体中的样本量尽量接近，比例不要过于悬殊；
- ✓ 性状一般可以用文字描述，无法用具体数值衡量，可转换成0、1等表示；

例如大豆种皮颜色；大豆圆皱；角的有无；毛色；血型；花药、籽粒或者颖壳等器官的颜色；芒的有无；绒毛的有无等。

数量性状

特点：

- ✓ 一般是受多基因控制的复杂性状；
- ✓ 表型指标从一个极端到另一个极端的变异，差异呈连续状态，中间无明显界限或者中断，界限不清楚，不易分类；
- ✓ 性状相对不稳定，容易受到环境影响；
- ✓ 性状一般可以直接用测量的数值表示；
- ✓ 性状一般符合正态分布；

例如身高，体重和动植物的许多重要经济性状都是数量性状（作物的产量、成熟期，奶牛的泌乳量，棉花的纤维长度、细度）等等。

其它性状

- ✓ 计数性状（Count Traits）：这些性状表示某个事件或特征的计数，表型为离散型，例如某种疾病的发病次数、某个特定行为的频率等。在GWAS中，研究人员可以使用泊松回归模型或负二项回归模型等方法，来分析基因变异与计数性状之间的关联；

- ✓ 顺序性状（Ordinal Traits）：这些性状具有有序分类的取值，有两种或者少数几种表型类别，但是遗传上由多基因控制，如植物抗病能力、教育水平（初中、高中、大学等）、疼痛程度（轻度、中度、重度等）。在GWAS中，研究人员可以使用有序回归模型或序数logistic回归模型等方法，来研究基因与顺序性状之间的关联；

- ✓ 生存性状（Survival Traits）：这些性状涉及到时间的测量，例如生存时间、事件发生的时间等。在GWAS中，研究人员可以使用生存分析方法（如Cox比例风险模型），来评估基因变异与生存性状之间的关联。

表型考察及质控

高通量表型分析技术和平台

1、作物地上部分表型自动化测量技术

2、作物根系表型自动化测量

3、种子表型自动化测量

综合自动化控制、光学成像、图像分析及计算机技术等多种现代科学技术，表型组学研究可追踪分析基因型、环境与表型的关系。基于图像技术，可在稳定一致的环境下，定量、准确、无损地获取植株表型，除了可对产量、抗性等复杂性状进行动态测量外，还可获得人工无法测量的新性状，如植株密度、谷粒投影面积等。

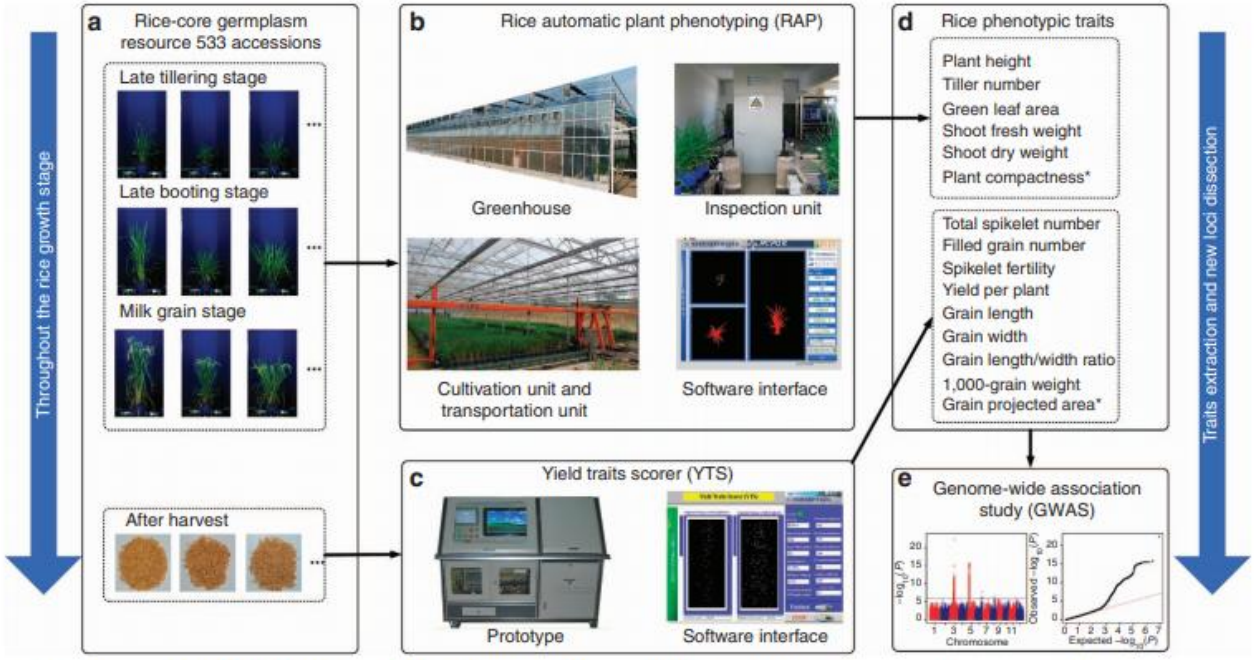


Figure 1 | Combination of the HRPF (RAP and YTS) and genome-wide association study (GWAS). To automatically screen the rice-core germplasm resource throughout the growth period (**a**), the entire HRPF was designed with two main elements: a rice automatic plant phenotyping device (RAP, **b**) and a YTS (**c**). These novel phenotyping tools were able to extract not only the traditional agronomic traits but also several novel phenotypic traits (such as plant compactness and grain-projected area). After the rice phenotypic traits (**d**) were extracted with the RAP and YTS, new loci were dissected using GWAS (**e**). *New traits are those that cannot be defined and extracted using traditional measurement techniques.

精确的表型检测是关联分析的核心

表型考察及质控

不同性状的处理与考察

质量性状

- 给予每类性状以相对数量的方法；如小麦颜色的红与白，可令白色数量值为0，红色数量值为1。
- 每类性状在群体中样本尽量接近，比例不能过于悬殊；

数量性状

- 尽量符合正态分布；
- 表型数据离散时建议进行数据转换；
- 删除异常表型值样本；
- 多年多点重复观测，单独分析或者取BLUE值进行一次分析；

表型考察及质控

表型的基本处理

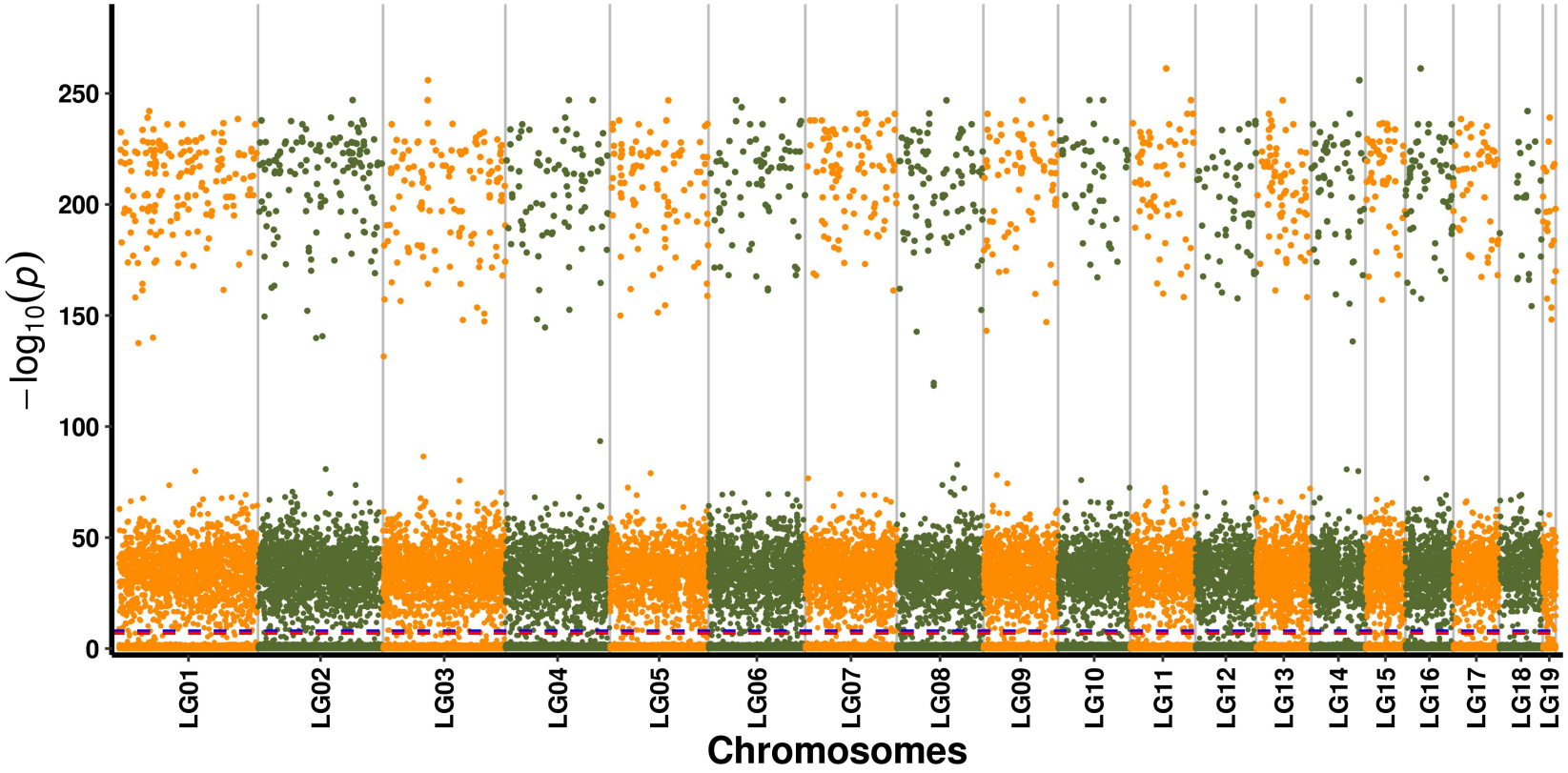
多年多点表型处理

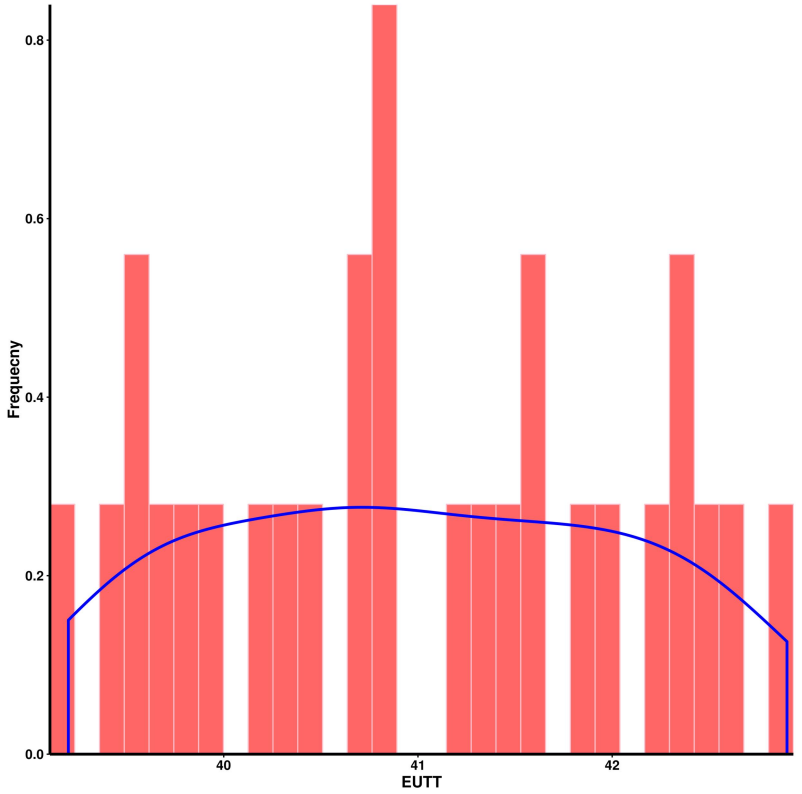
- 为获得可靠的关联结果，我们通常会对同一个性状观测多次，多次观测可能是相同年份不同地点，或者不同年份相同或者不同地点；
- 对于此类数据，我们可以根据性状的遗传机制的不同选择不同的处理方式；
- 如果性状的遗传力高，受环境影响不大，我们可以根据多年多点的结果取均值或者BLUE/BLUP值作为该性状的代表值进行分析；
- 如果性状遗传力低，受环境影响大，我们可以每年每点单独分析后综合评判结果，在获得定位结果的同时可以进行GXE分析。

表型考察及质控

雲南農業大學
Yunnan Agricultural University
云南省生物大数据重点实验室

异常表型





表型质控

（1）正态性检验

关联分析的模型属于线性模型，线性模型要求数据符合正态分布，所以从理论上说，只有我们的表型数据是正态分布才可以进行关联分析。

正态性检验简单直观的方法为绘制频率分布图，观测数据分布情况，数据大致为钟形分布且没有极端异常值存在时，即可认为大致符合正态分布。如果需要判断数据是否符合正态分布，可以用Shapiro-Wilk方法进行检验，该方法可以通过R函数shapiro.test()实现

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室

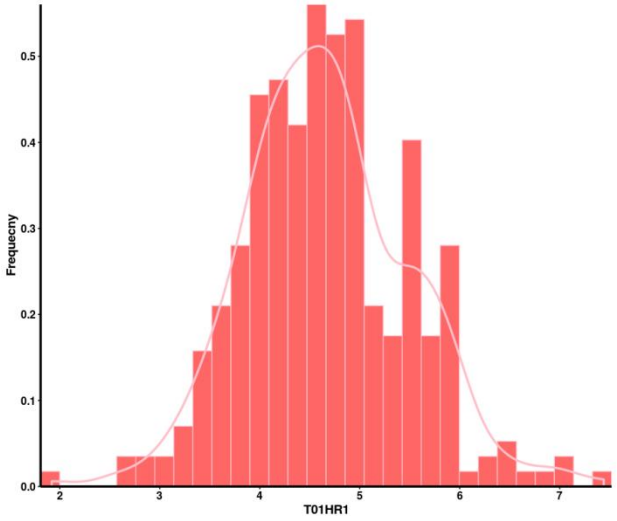
表型质控

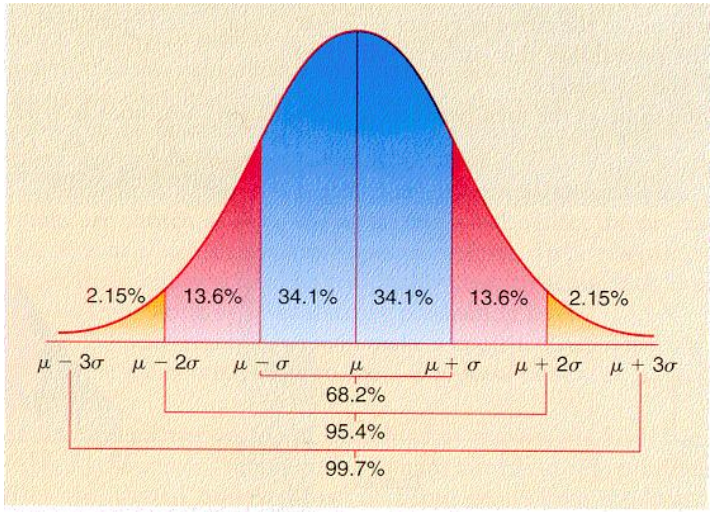
（2）极端异常值的去除

极端异常值（极大或者极小）的存在可能引起关联结果的异常，因此在分析关联分析前应该将其去除。

异常值的检测方法：

1. 排序观察法（人工法）；
2. 绘制表型数据的频率分布图；
3. 正态分布的3σ准则，即在均值加减3倍标准差范围内的值为正常值，其他值为异常值；

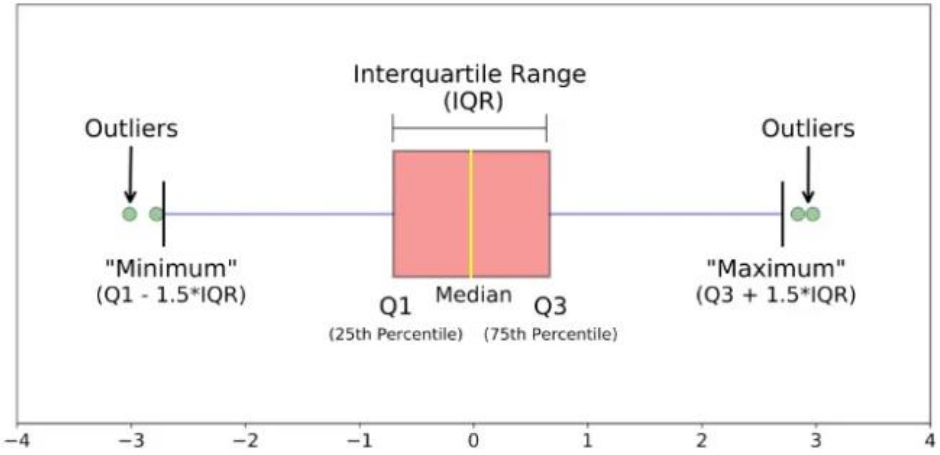




表型质控

异常值的检测方法：

4. 箱线图方法，绘制箱线图；



最小数：Q1 -1.5 * IQR

第一四分位数（Q1 / 25百分位数）：最小数（不是“最小值”）和数据集的中位数之间的中间数；

第二四分位数：中位数（Q2 / 50th百分位数）：数据集的中间值；

第三四分位数（Q3 / 75th Percentile）：数据集的中位x数和最大值之间的中间值（不是“最大值”）；

最大数：Q3 + 1.5 * IQR

四分位间距（IQR）：第25至第75个百分点的距离；

晶须（蓝色显示）

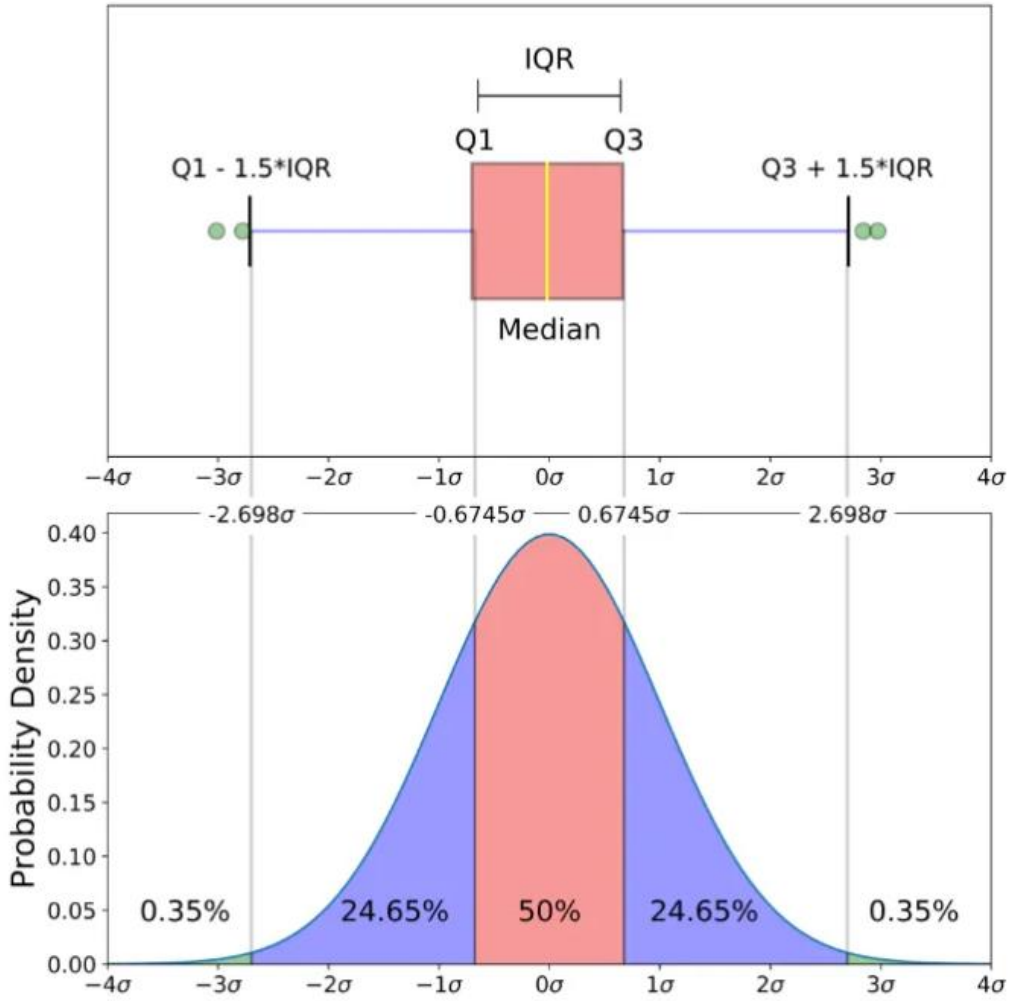
表型考察及质控

表型质控

（2）极端异常值的去除

离群值（显示为绿色圆圈）

将> Q3 + 1.5 * IQR 或者< Q1 -1.5 * IQR的离群赋值为NA



表型质控

(3) 表型数据标准化

数据标准化一般针对绝对值较大，且有明显梯度间隔的表型，绝对值较小的比较连续的表型可以不标准化，直接用于关联分析。标准化目的是为了降低由于量刚不同、自身变异或者表型值相差较大所引起的误差，不会改变原有数据本身的大小趋势，也不会改变关联结果。

常用的标准化方法有两种：z-score标准化和min-max标准化

- **z-score标准化**：以标准差为单位的中心化，原始值x高于均值时，z为正值，反之则为负值

公式：$z=(x-\mu)/\sigma$

x为原始值，μ为平均值，σ为标准差

- **min-max标准化**：离差标准化，归一化，绝对值较大且具有明显梯度的数据常采用该方法进行标准化；

公式：$y=(x-\min(x))/(\max(x)-\min(x))$

y为标准化后的值，x为原始值，标准化所有值均在0-1之间

基因型分型及填补

基因分型

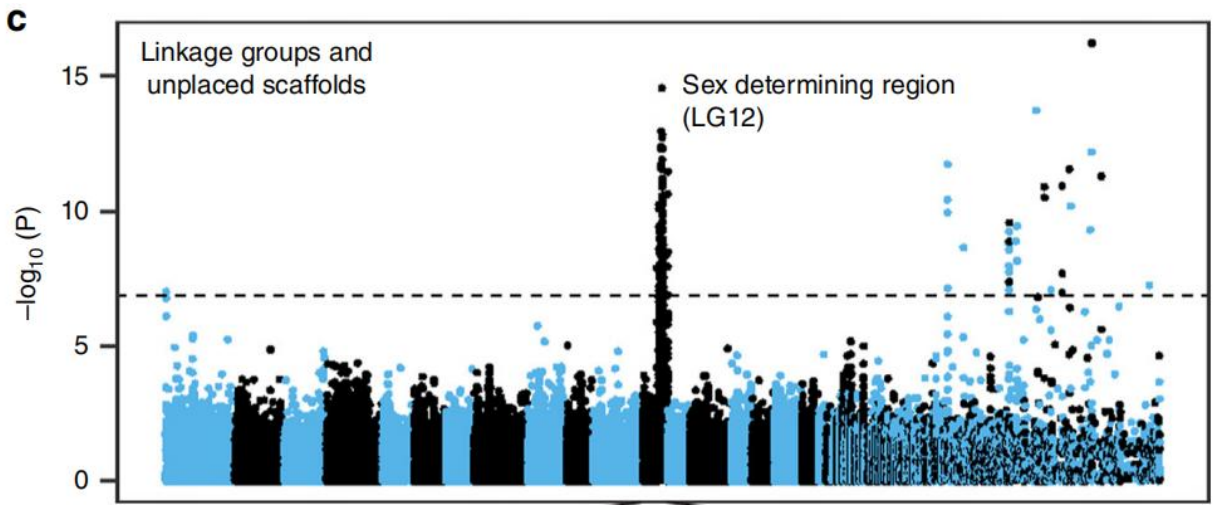
```
##fileformat=VCFv4.2
##fileDate=20171105
##source=PLINKv1.90
##contig=<ID=1,length=2>
##INFO=<ID=PR,Number=0,Type=Flag,Description="Provisional reference allele, may not be based on real reference genom
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  -9_CZTB0004     -9_CZTB0006     -9_CZTB0008
1       1       10000235        A       C       .       .       PR      GT      0/1     0/0     0/0     0/0     0/0
1       1       10000345        A       G       .       .       PR      GT      0/0     0/0     0/0     0/0     1/1
1       1       10004575        G       .       .       .       PR      GT      0/0     0/0     0/0     0/0     0/0
1       1       10006974        C       T       .       .       PR      GT      0/0     0/0     0/1     1/1     0/1
1       1       10006986        A       G       .       .       PR      GT      0/0     0/0     0/1     ./.     1/1
```

1. CHROM 染色体
2. POS 变异在参考基因组中的物理位置
3. ID - 变异的ID
4. REF 参考碱基，ATCGN
5. ALT 与参考序列比对发生突变的碱基
6. QUAL 表 示在该位点存在变异的可能性，值越高，可能性越大
7. FILTER 过滤结果中通过则该值为PASS
8. INFO 变异的详细信息，包括样本在这个位置的reads覆盖度等
9. 样本分型信息

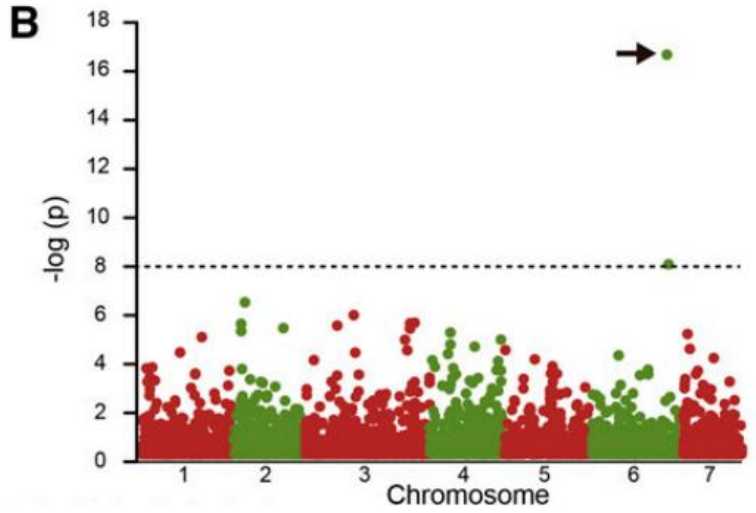
基因型分型及填补

变异类型

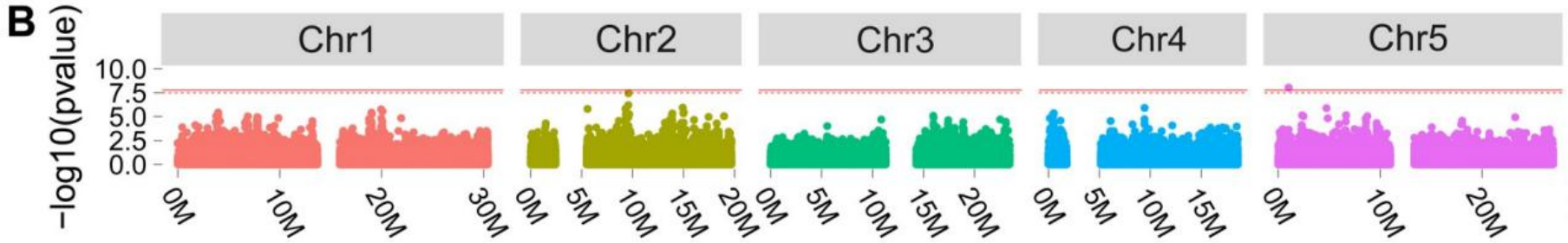
基于SNP（Hazzouri 2019）



基于SV（zhang 2015）



基于INDEL（song 2018）



基因型分型及填补

基因型填补

依据已分型位点的基因型数据对缺失的位点或者未分型位点进行基因型预测的方法

目的和意义：

1、针对分型缺失率高的数据进行填补，提高全基因组遗传标记的覆盖度，增加阳性关联位点的成功率；

2、应用于精细定位，填补已确认的关联位点附近的位点，以便评价相邻SNP位点关联证据，加快复杂疾病易感基因的定位；

3、降低测序成本，通过对低深度测序的群体进行基因型填充，大大提高数据的利用效率；

4、多数分析（部分软件）不允许存在缺失数据；

注：

（1）目前所有的填补方法对杂合位点分析能力较差，也就是说很难填补杂合位点，即使填补，也会存在填补错误的情况；

（2）当基因型密度很低时，才会考虑进行基因型填补，比如芯片数据或低深度测序数据。重测序数据一般无需进行基因型填充；

基因型分型及填补

测序深度与SNP的检出率

1、随着测序深度增加，纯合SNP数量反而略有降低。那是因为深度不足的情况下，被错误判断为纯合SNP的位点，被重新发现是杂合SNP位点；

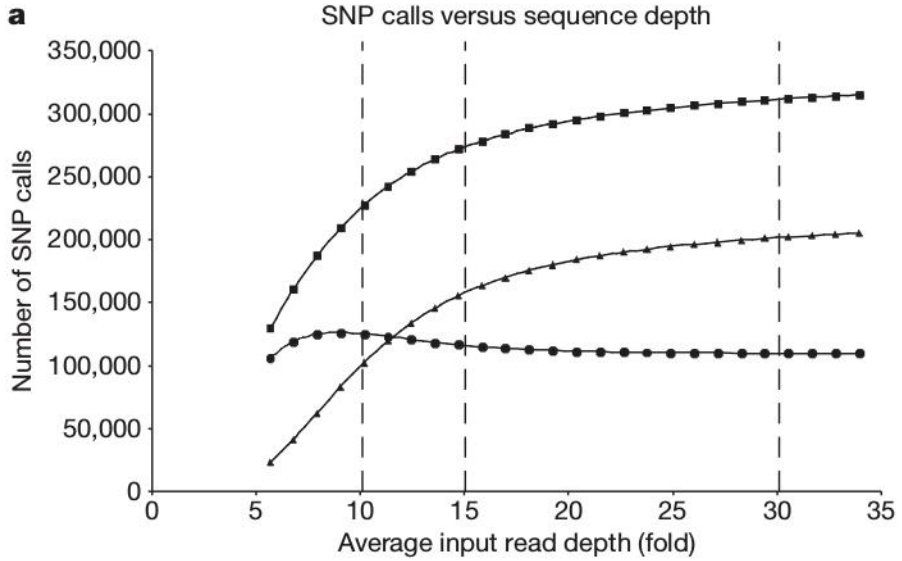
2、群体水平的研究（系统进化、选择压力分析等），10x测序深度基本可以；个体基因组重测序，需要30x的深度，才能更好的检测**SNV;**

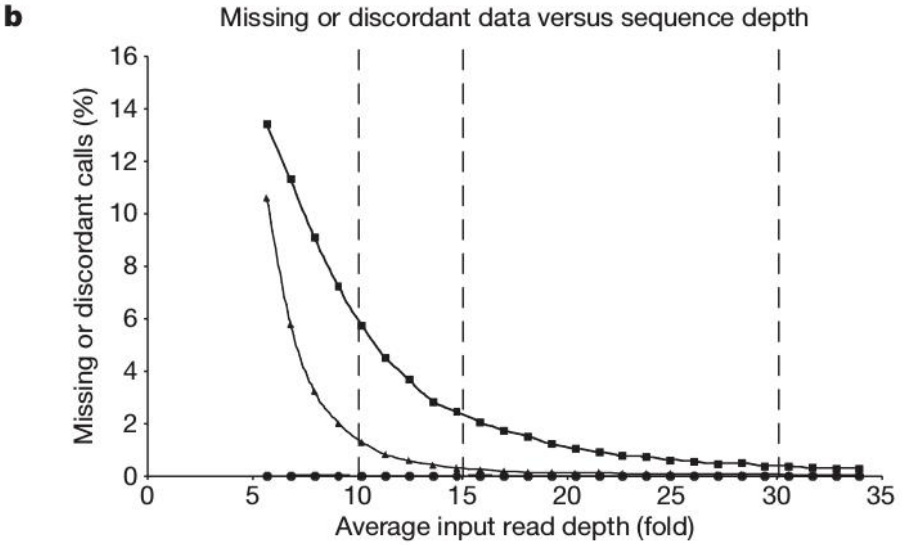
3、随着测序深度的提高，基因组未覆盖的区域/snp calling检出不一致性下降。

■ 所有SNP

▲ 杂合SNP

● 纯合SNP





雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

低深度填补 or 高深度测序

当然优先选择高深度测序

1）填补对杂合基因型的处理能力较弱；

2）依然有错误位点存在，假设正确率是95%，填补的数据点500万错误的数据点就25万，很可对关联结果产生影响

经费有限的情况下，大规模测序分析可以采用低深度重测序+基因型填补的策略

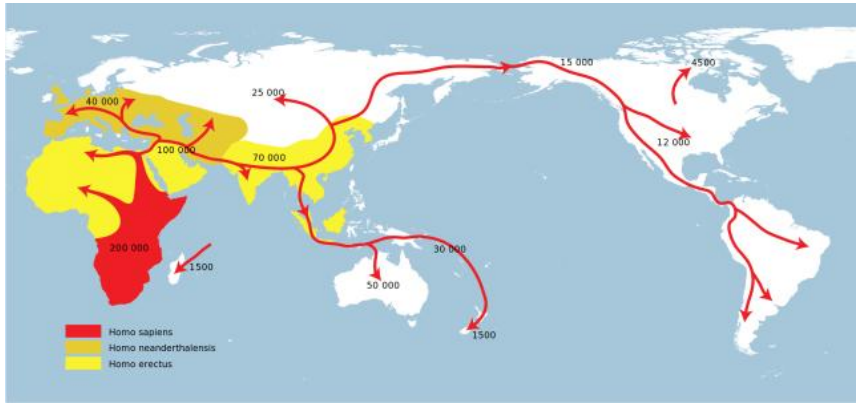
群体结构对GWAS的影响

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

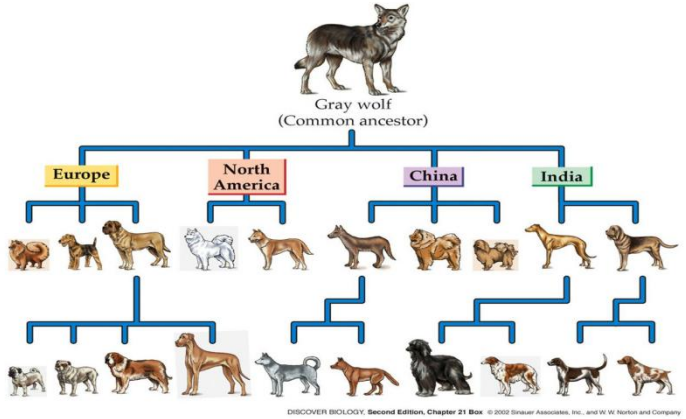
群体结构的来源

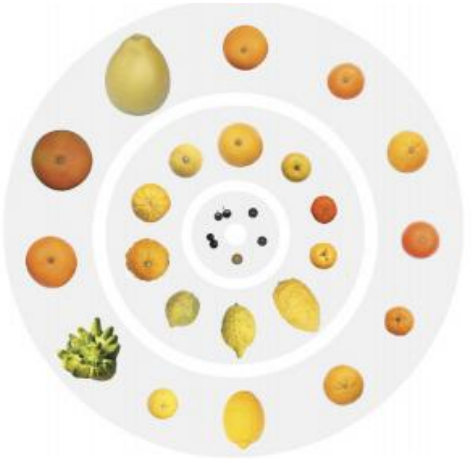
群体结构：又称群体分层，指所研究的群体中存在基因频率不同的亚群。

物种迁徙（比如人类走出非洲）

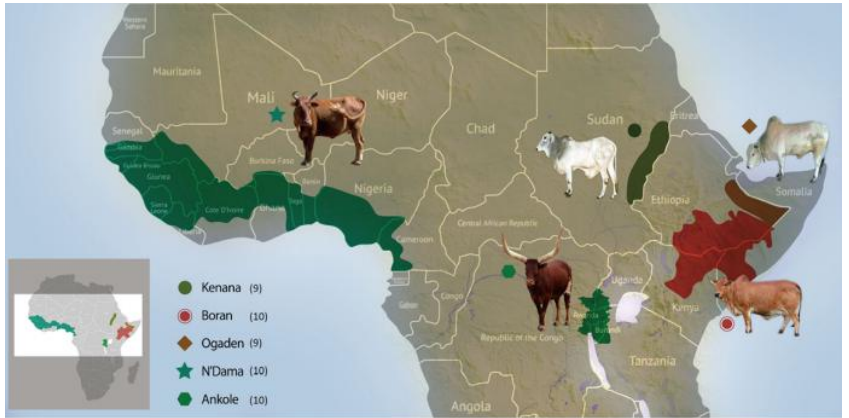


驯化/改良





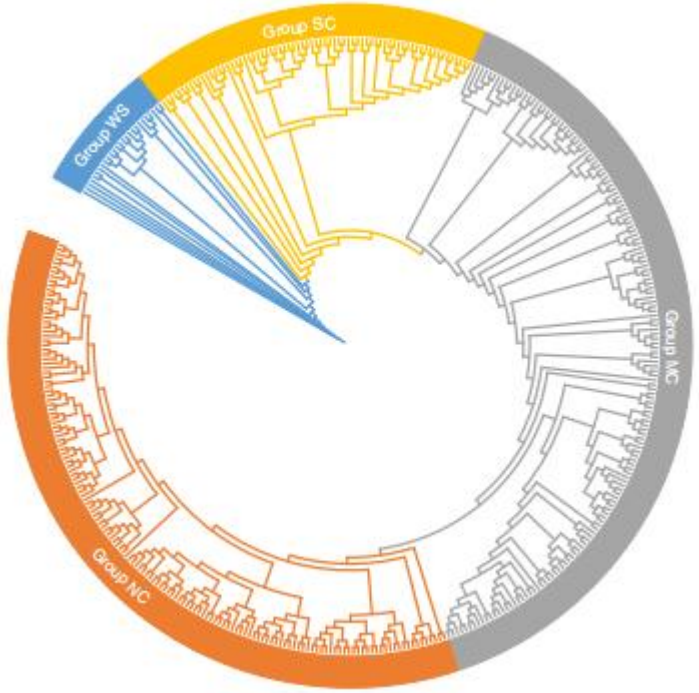
地理隔离，环境适应



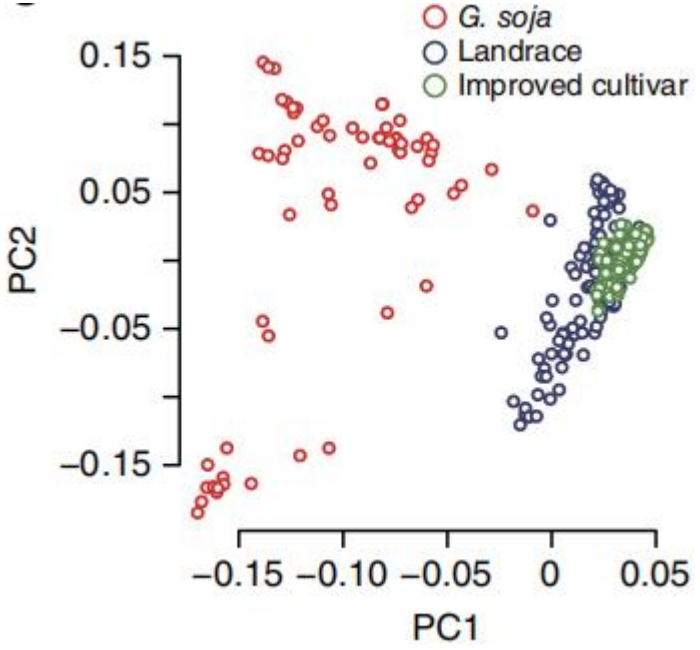
群体结构对GWAS的影响

云南省生物大数据重点实验室
生信帮

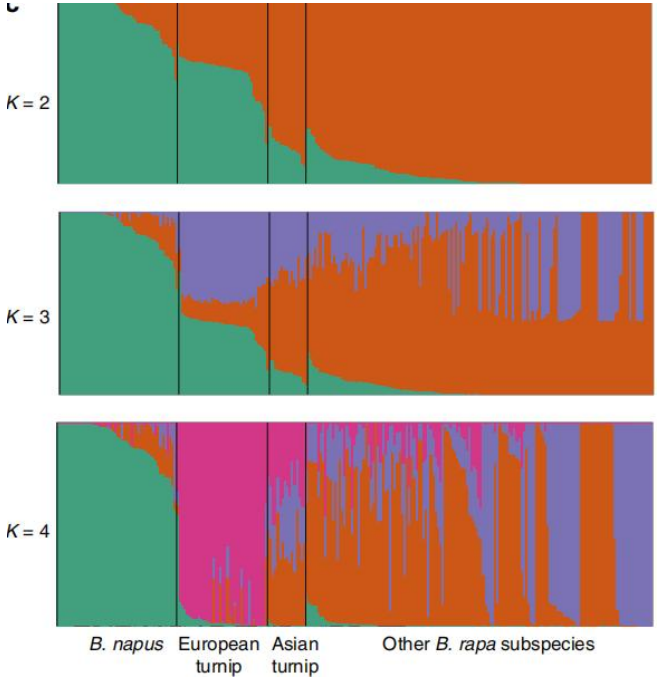
群体结构评估



系统发育树



主成分分析

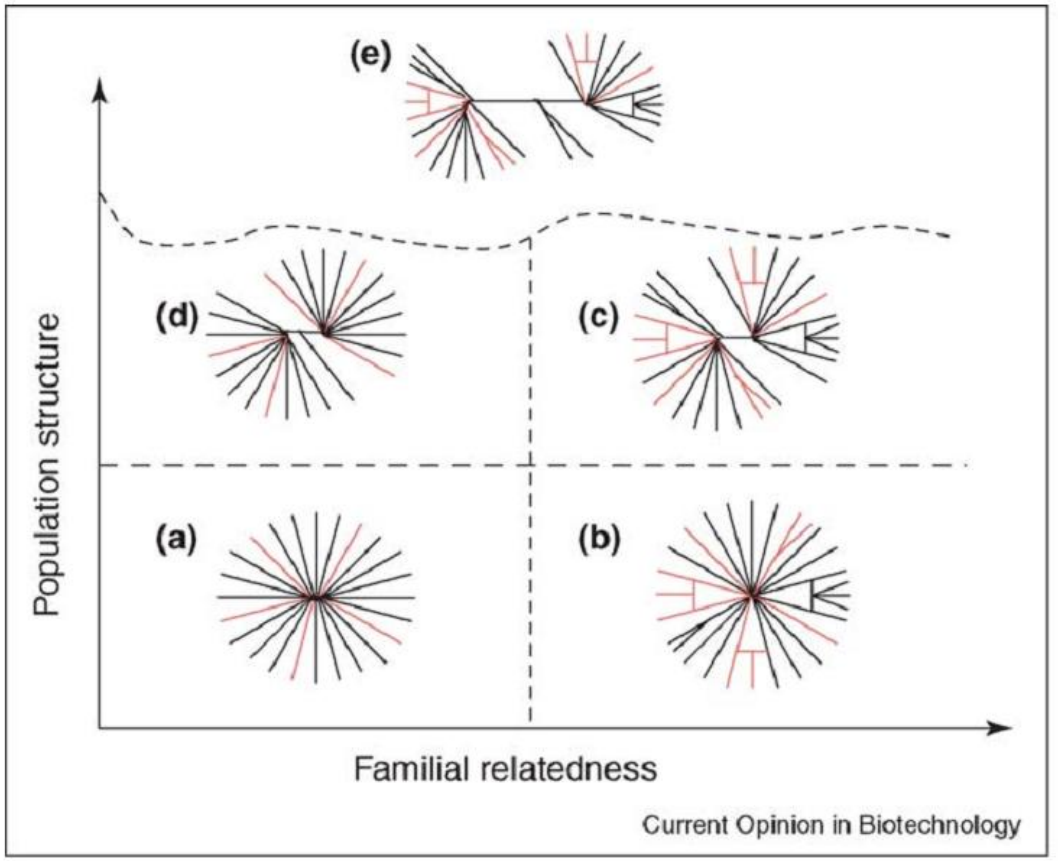


群体结构

群体结构对GWAS的影响

群体类型

- 理想群体（a）
- 多家系群体（b）
- 具有群体结构群体（d）
- 即具有群体结构又具有亲缘关系群体（c）
- 具有高度群体结构和高度亲缘关系的群体（e）



雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

群体结构对GWAS的影响

祖先血统信息（Ancestry）可以导致混淆变量（confounding variable）

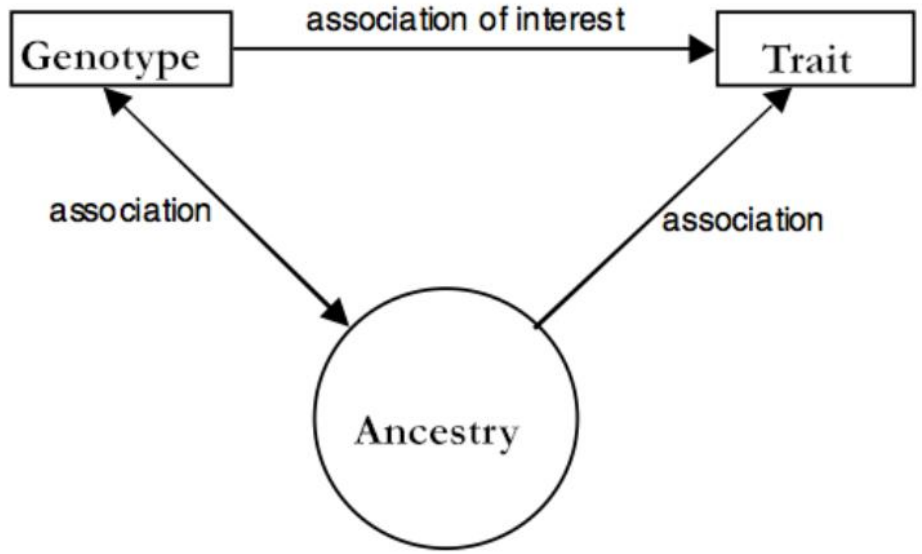
混淆变量：统计学中，混淆变量是指既与因变量相关又与自变量相关的无关变量。在这里群体结构变量其实就是混淆变量，是潜在的第三者变量。

每个亚群的样本共享一种生长习惯和生活方式，导致许多目标性状直接与某亚群或世系相关。

表型：各个亚群间的表型本身在各个群体间就存在差异

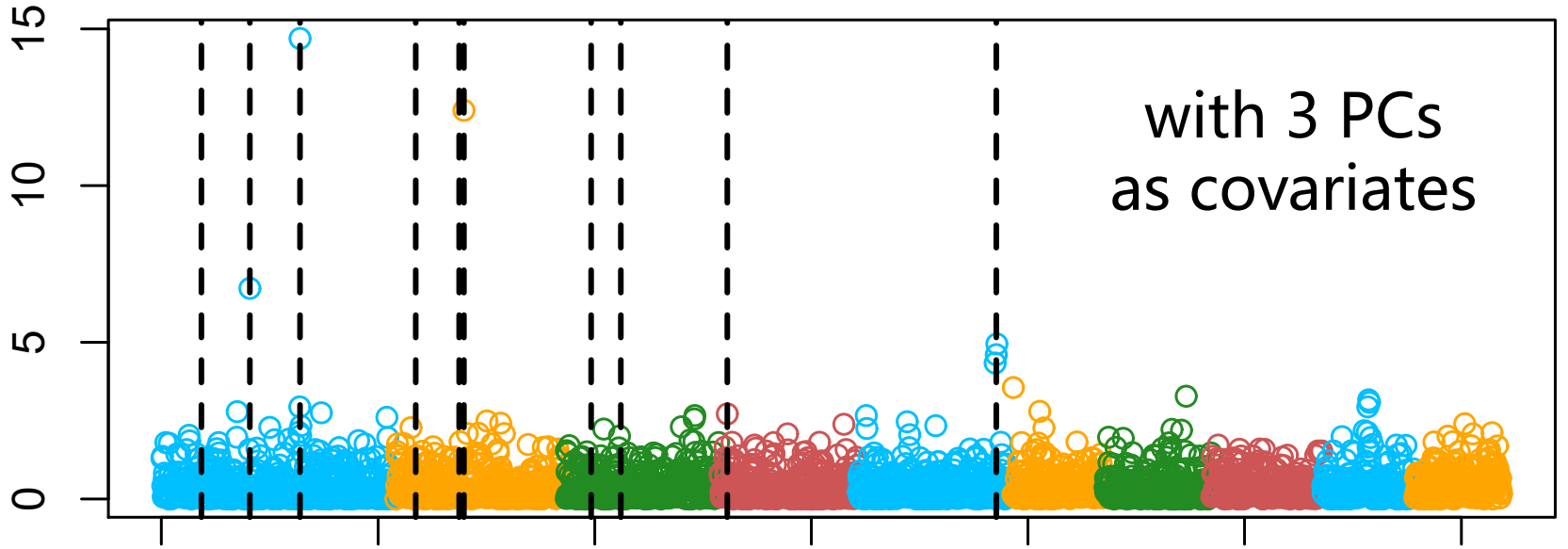
基因型：存在群体特异的位点（各个等位基因频率差异很大），

这些位点会与表型关联，导致大量的假阳性。

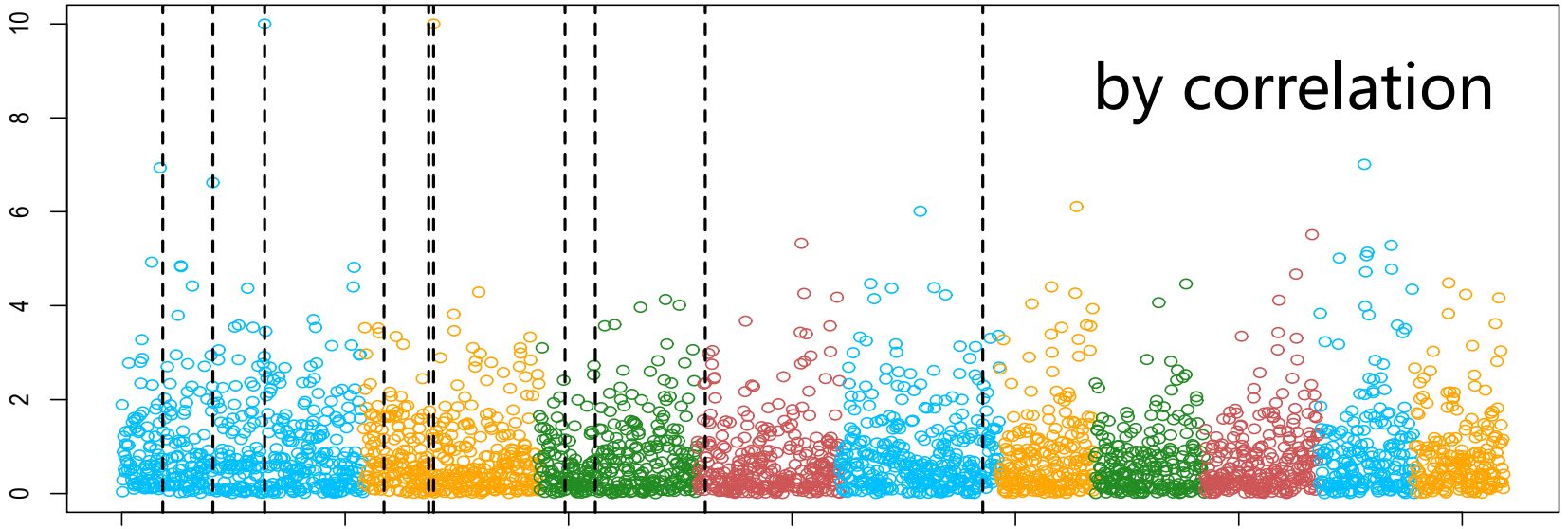


群体结构对GWAS的影响

群体结构对GWAS的影响



假阳性 ⟶



亲缘关系对GWAS的影响

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

亲缘关系（Kinship）矩阵

亲缘关系矩阵（Kinship Matrix）是一种用于描述样本之间亲缘程度的矩阵。它可以用来评估样本间的遗传相似度。

PLOS | GENETICS

Browse Publish About

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Lessons from *Dwarf8* on the Strengths and Weaknesses of Structured Association Mapping

Sara J. Larsson, Alexander E. Lipka, Edward S. Buckler

Published: February 21, 2013 • DOI: 10.1371/journal.pgen.1003246

文章的结论：只考虑群体结构是不够的，亲缘关系也会导致假阳性

亲缘关系对GWAS的影响

亲缘关系（Kinship）的主要来源

1、Blood relationship

1.1、Family ties

1.2、Blood ties

2、Common Ancestry

3、Sharing of characteristics or origins.

亲缘关系对GWAS的影响

亲缘关系（Kinship）的推断方法 — 基因型推断法

Average proportion of shared alleles across markers

Marker	1	2	3	4	5	Average
Individual 1	AA	AA	AA	BB	AB	
Individual 2	AA	AB	BB	BB	AB	
Similarity	1	0.5	0	1	0.5	0.6

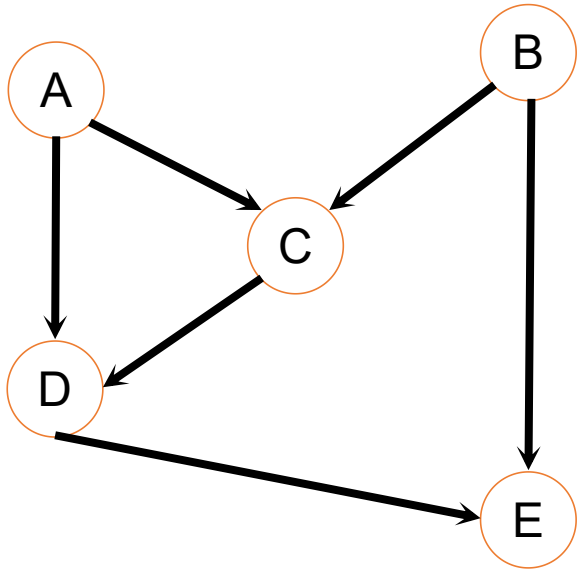
Maximum similarity: 1

亲缘关系对GWAS的影响

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室

生信帮

亲缘关系（Kinship）的推断方法 — 系谱推断法



Individual	Father	Mother
A		
B		
C	A	B
D	A	C
E	D	B

	A	B	C	D	E
A	1	0	0.5	0.75	0.375
B	0	1	0.5	0.25	0.625
C	0.5	0.5	1	0.75	0.625
D	0.75	0.25	0.75	1.25	0.75
E	0.375	0.725	0.625	0.75	1.125

Diagonals=1+F

GWAS分析模型与软件

云南农業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

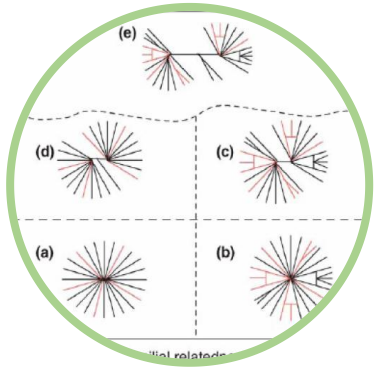
GWAS分析框架

材料选择

1、表型稳定

2、材料丰富

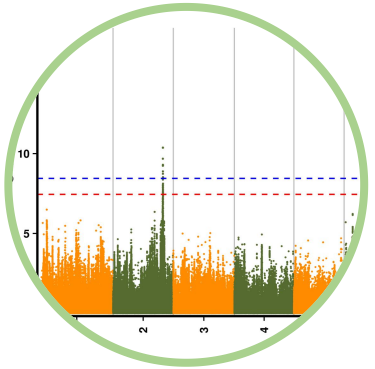
3、染色体倍性相同



群体分层

1、群体结构

2、亲缘关系



关联分析

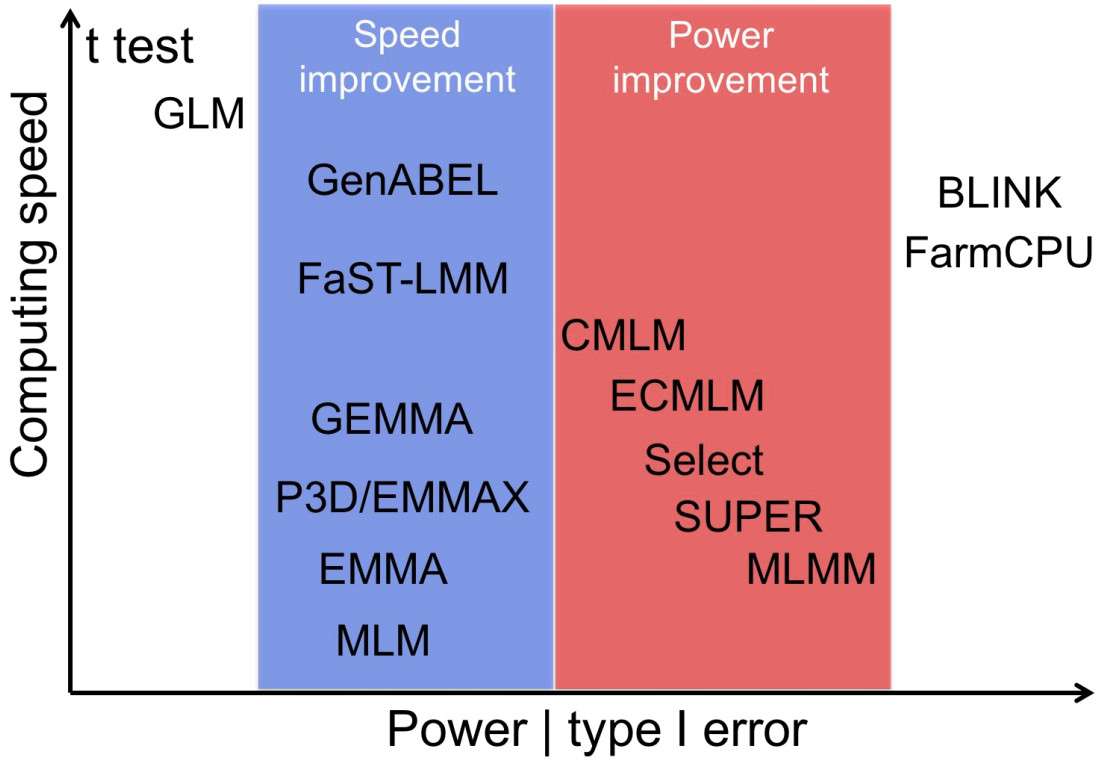
1、筛选候选位点

2、关联结果注释

GWAS分析模型与软件

云南省生物大数据重点实验室
生信帮

GWAS分析软件



Year	Method	Positive semidefinite matrix requirement[a]	Strategy for increasing computational speed			Computational speed	Statistical power[d]
			Approximate/ Two-step approach[b]	Matrix optimization[c]	Low-rank matrix		
2006	Standard MLM					Low	High
2007	GRAMMAR		+			Very fast	Intermediate
2008	EMMA	+		+		Intermediate	High
2010	EMMAX	+	+	+		Fast	High/Intermediate
2010	P3D and CMLM		+		+	Fast	High/Intermediate
2011	FaST-LMM	+		+		Fast	High
2012	GEMMA	+		+		Fast	High
2012	FaST-LMM-Select	+		+	+	Very fast	High
2014	ECMLM		+		+	Intermediate	High/Intermediate
2014	SUPER	+		+	+	Fast	High

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

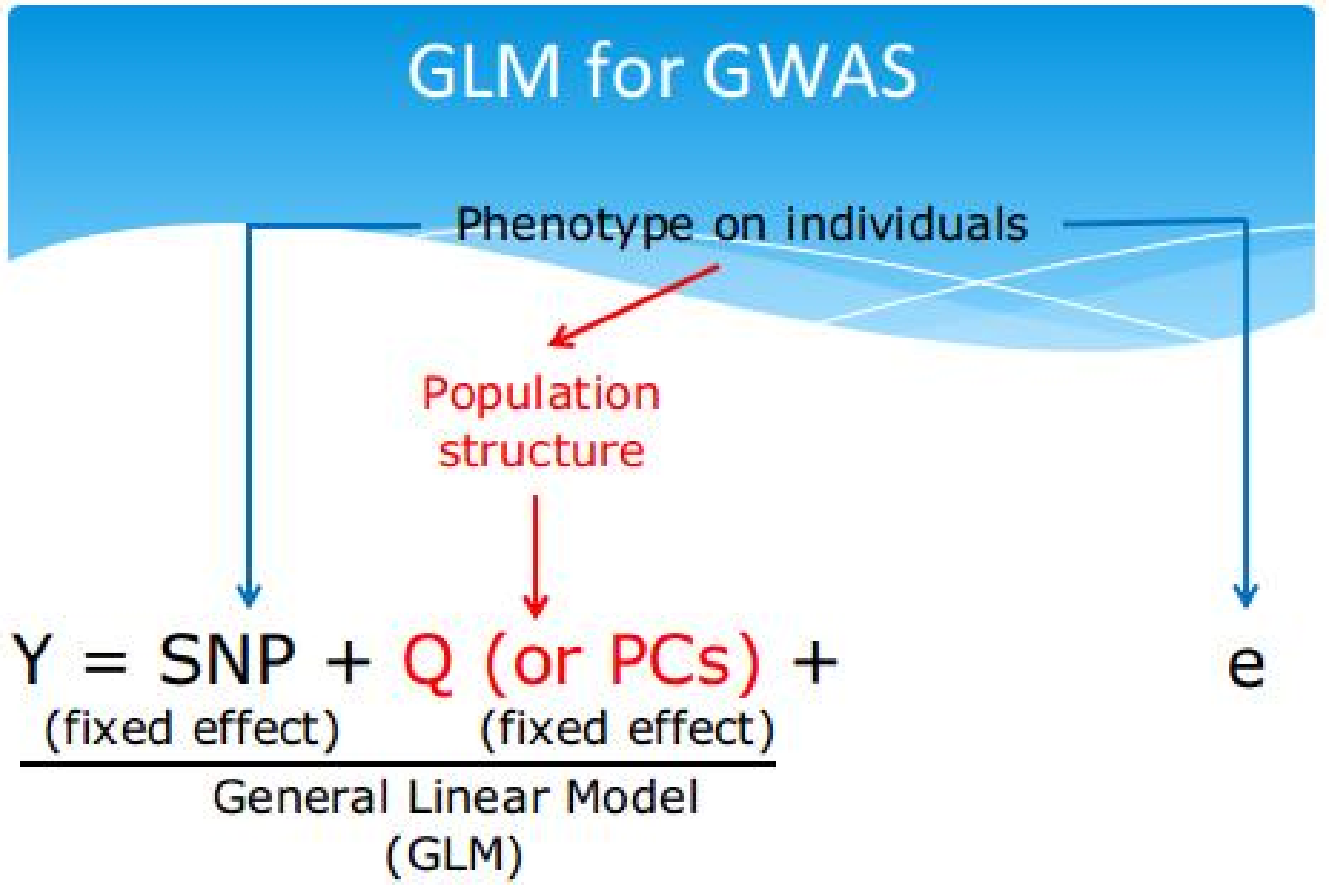
关联分析

选择正确的统计方法进行关联分析

1. 小标记量候选基因关联分析：可以选择比较简单的t-test或者ANOVA；
2. Case/Control-质量性状：卡平方检验，OR检验，逻辑回归，也可以视为数量性状使用较为复杂的线性模型；
3. 数量性状：根据群体结构评估的情况，选用相应的模型，但在实际操作中一般使用多种模型（GLM/MLM/EMMAX/FaST-LMM）同时分析，根据结果进行取舍。

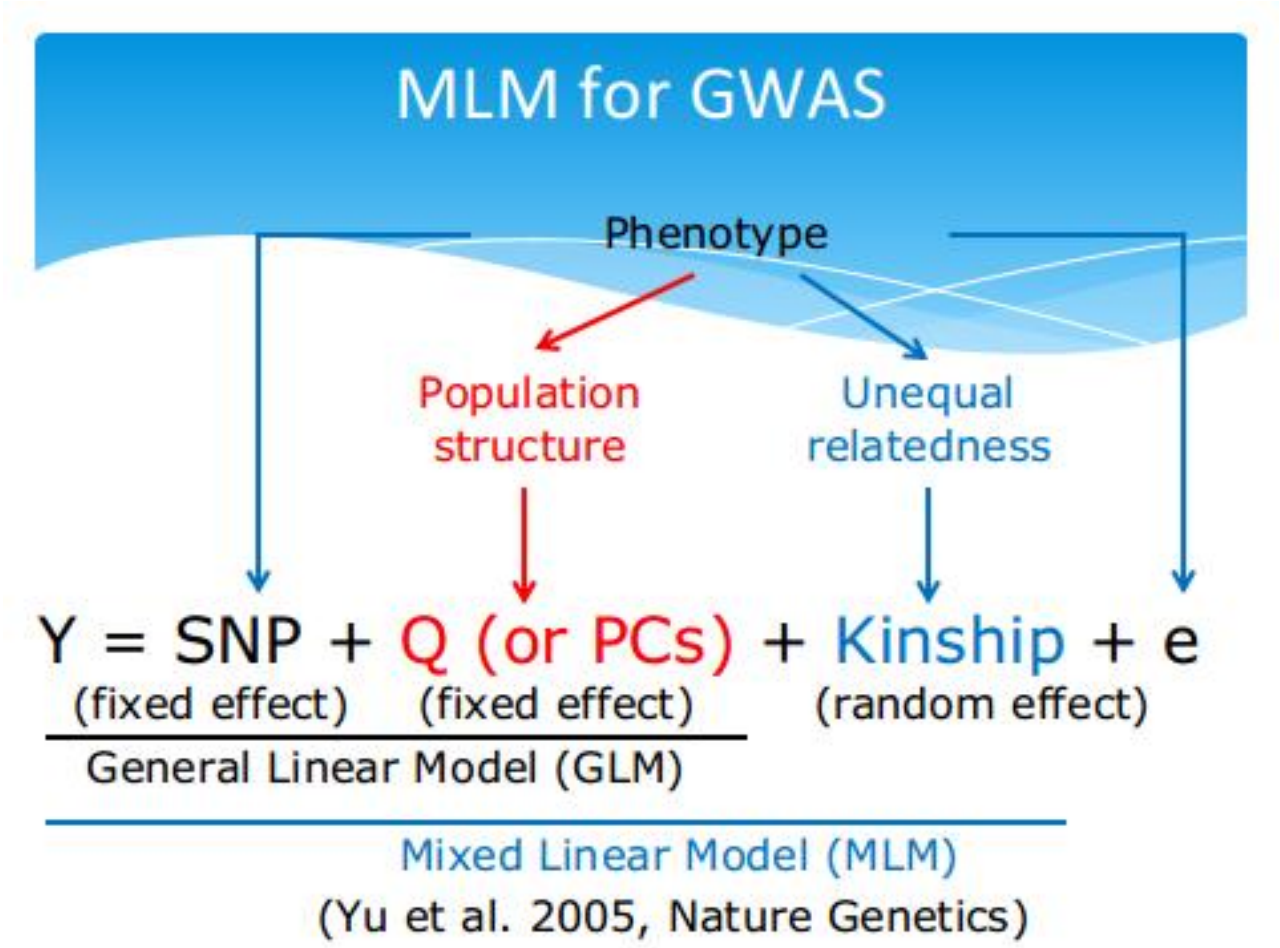
GWAS分析模型与软件

关联分析模型 - 一般线性模型GLM（广义线性回归模型）



GWAS分析模型与软件

关联分析模型 - 混合线性模型MLM



关联分析软件-EMMAX

EMMAX（Efficient Mixed-Model Association eXpedited）是一款用于全基因组关联分析（GWAS）的软件，其底层算法基于线性混合模型（linear mixed model）和EM算法（Expectation-Maximization algorithm）。

在GWAS中，我们希望通过比较基因型和表型数据之间的关联性来识别与表型相关的基因变异。但是，由于基因组具有高度的相关性和复杂性，常规的线性回归方法容易被混淆和误导。为了解决这个问题，EMMAX引入了线性混合模型，其中包括一个固定效应和一个随机效应。固定效应用于基因型与表型之间的显性关系，而随机效应则用于考虑样本之间的相关性。这种方法可以降低基因型与表型之间的冗余和噪声，从而提高了GWAS的准确性。

EMMAX使用EM算法来估计混合模型中的参数。EM算法是一种迭代算法，包括两个步骤：E步骤和M步骤。在E步骤中，根据当前的参数值，计算每个随机效应的条件期望。在M步骤中，根据这些条件期望，重新估计混合模型的参数。通过不断迭代E步骤和M步骤，可以找到最优的混合模型参数。

EMMAX还引入了一个基于PCA的过滤器，用于减少基因型数据中的冗余信息。它通过计算基因型数据的主成分分析（PCA）来降低基因型数据的维数，从而提高了计算效率和降低了计算成本。

总之，EMMAX的底层算法基于线性混合模型和EM算法，通过考虑固定效应和随机效应之间的关系，以及引入基于PCA的过滤器，来提高GWAS的准确性和计算效率

关联分析软件-EMMAX

Kang, H. M., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. Nature Genetics, 42(4), 348-354.

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nature Genetics, 44(7), 821-824.

Zhu, Z., et al. (2017). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics, 48(5), 481-487.

Xiao, R., et al. (2018). Genetic variants associated with gestational hypertriglyceridemia and pancreatitis. Science Advances, 4(2), eaar4478.

关联分析软件-GEMMA

数据准备：首先，需要准备好基因型数据和表型数据。基因型数据通常以单核苷酸多态性（SNP）的形式表示，可以使用PLINK等软件进行预处理和格式转换。表型数据包括感兴趣的性状或表型特征。

遗传关系矩阵（GRM）的计算：GEMMA利用基因型数据估计样本间的遗传关系矩阵（GRM），用于捕获样本之间的相关性。GRM表示样本之间的遗传相似性程度。

线性混合模型（LMM）的拟合：GEMMA使用线性混合模型，将遗传关系矩阵和表型数据结合起来进行建模。线性混合模型考虑了遗传结构和样本间的相关性，并用于对遗传和非遗传因素进行建模和分析。模型中的固定效应表示SNP的影响，而随机效应表示样本间的相关性。

GWAS统计检验：根据线性混合模型的结果，进行统计检验来评估SNP与表型之间的关联。常用的检验方法包括基因型频率的卡方检验（χ^2检验）或线性回归模型等。GEMMA使用Wald检验或样本混合后验概率来评估SNP的显著性。

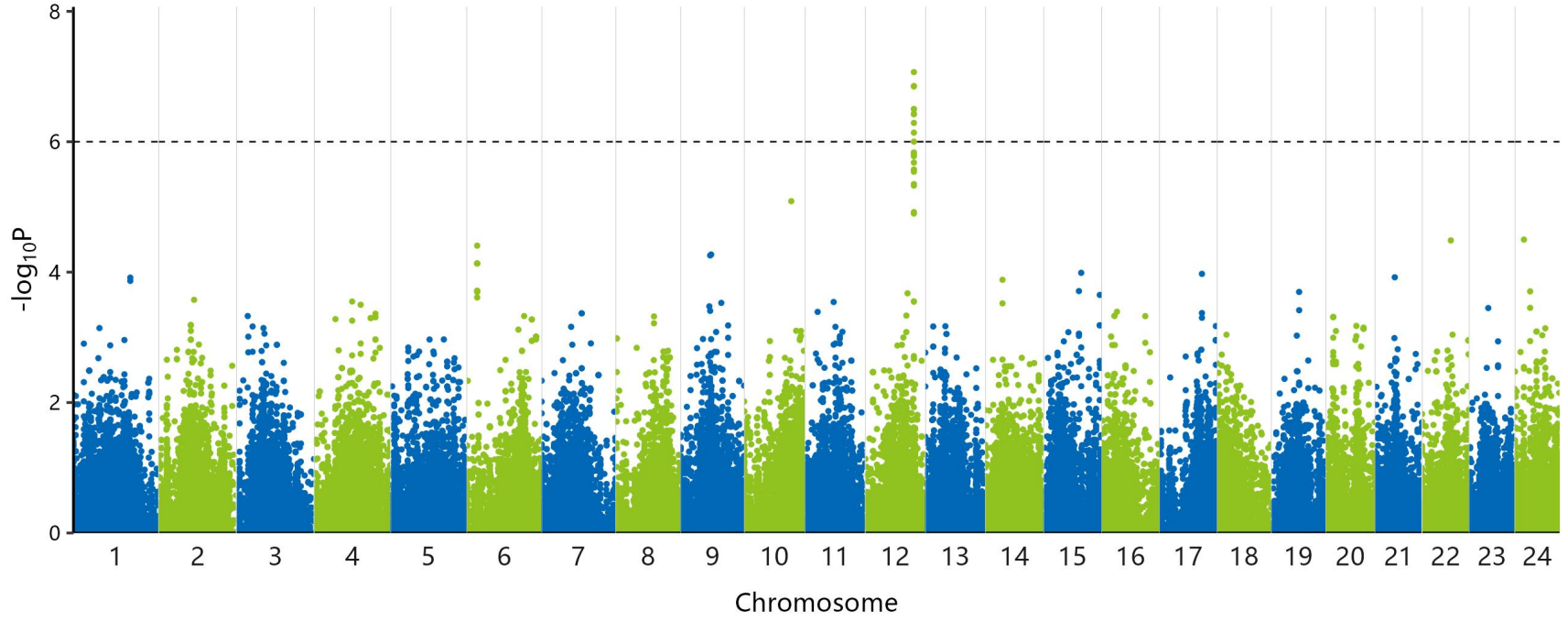
校正与解释：GEMMA考虑了遗传结构和样本间的相关性，通过进行群体结构校正来减少虚假阳性结果。此外，GEMMA还提供了一些额外的输出和统计信息，帮助解释GWAS结果。

关联结果可视化-曼哈顿图和QQ图

曼哈顿图

曼哈顿图本质上就是散点图，通常用于展示具有大量数据点的数据，在全基因组关联分析结果中，该图展示的是所有SNP位点的效应值即p值分布，图中的横坐标为每个snp位点在基因组上的物理位置，纵坐标为每个SNP关联p值的负对数（-log10p），p值越小关联性越强，在图中的表现为纵坐标越高。

如果参考基因组非染色体水平，可考虑挑选最长的30条scaffold的进行可视化绘图。

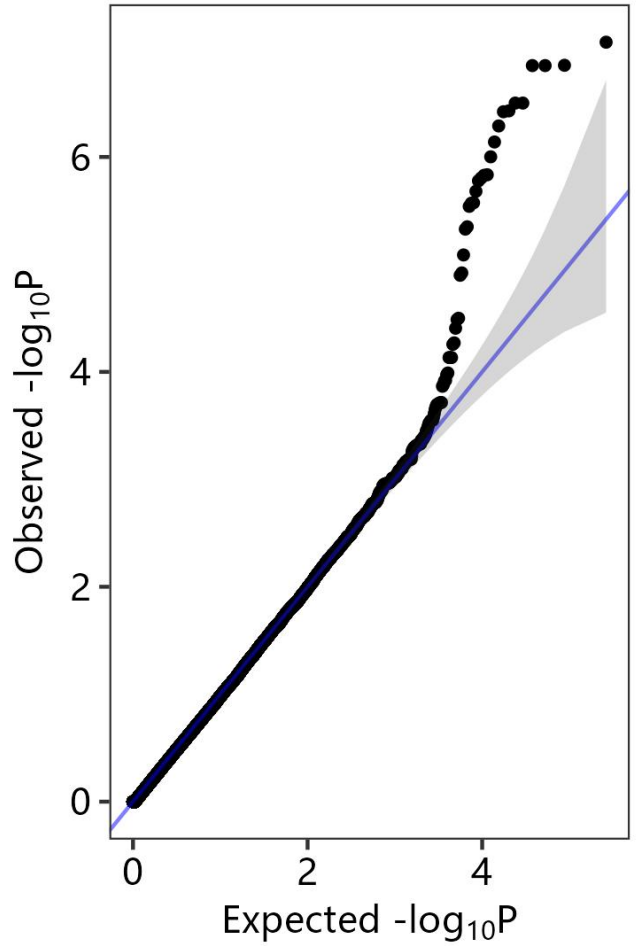


关联结果可视化-曼哈顿图和QQ图

QQ图

QQ plot的全称是Quantile-Quantile plot，它比较的是各SNP的Pvalue的观测值（纵轴）与期望值（横轴）的一致性。图中蓝色实线表示观测值与期望值相等，即零假设下应该出现的情况。如果散点分布与蓝色实线在前端贴合一致，仅在末端翘起，代表SNP假阳性较低，关联分析结果较为可靠。

QQ-plot 用于展示假阳性控制情况



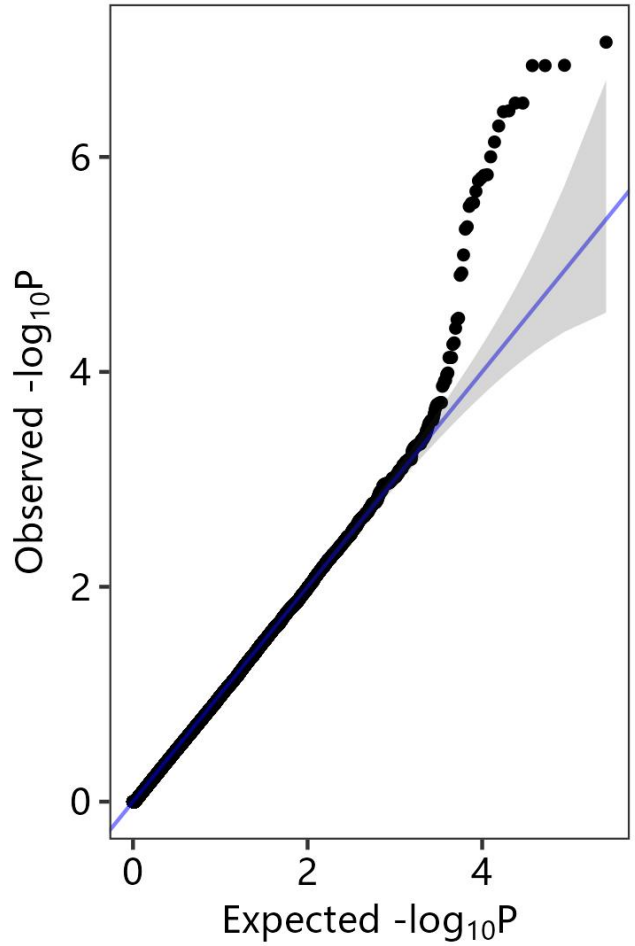
关联结果可视化-曼哈顿图和QQ图

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

QQ图的算法

纵坐标的观测p-value值，就是GWAS分析实际得到的结果（是observed的结果），横坐标则是均匀分布的p-value值（是Expecte的结果），同时换算为-log10。

那横轴这个期望的p-value值是如何计算的呢？实际上，它就是均匀分布的分位数，分位数的个数与GWAS研究的SNP位点数是一一对应的。比如我们研究中使用了5百万个SNP位点，那么分位数的个数就是5百万个，从1/5000000,2/5000000,3/5000000,...一直往下排直到5000000/5000000，当然都是转换为-log10，就是横坐标值。然后与GWAS关联的p-value值一起作图而成。



关联结果可视化-曼哈顿图和QQ图

在GWAS分析中，当我们通过曼哈顿图看到某些SNP和表型性状（或者疾病）有着很强的相关信号（比如，p-value < 10^-6甚至10^-8）时，依然不能直接认为这些位点就与表型显著相关的。这是因为基因组上基因位点的突变通常有两个来源：

第一是自然选择（Selection），这里所说的自然选择不仅指达尔文在《进化论》中所描述的物竞天择，还指所有对物种适应性有影响作用的“力量”，比如高辐射环境、疾病、病毒等，这也是我们在GWAS研究中真正关心的突变；

第二是遗传漂变（genetics drift），它是一种比较随机的基因组突变而且数量也不少，虽然也是物种演化的一种重要力量，但是由于它的突变都比较随机，目前认为它与环境的变迁没有必然联系，但也会在某些时候，有些随机的突变带来了生存优势，便会在种群中显示出它的作用。但绝大多数情况下，对于已经在群体中稳定存在的性状而言，并不认为它们有明显的作用，所以GWAS研究是不关心这一类突变的，要把它们全部排除掉。

关联结果可视化-曼哈顿图和QQ图

得到QQ plot之后，如何通过它来协助我们判断GWAS关联结果到底是好还是坏呢？

严格来讲，这里其实是不应该用好坏来形容，应该用是否与表型性状相关来形容。

判断的秘密在于横轴用的是**均匀分布**算法。这是因为均匀分布恰好可以用来近似描述基因组上的随机漂变现象。如果表型性状并非真的受自然选择所左右，那么应该会看到GWAS p-value的分布和均匀分布的结果将集中在一条直线上，反之，就应该能够看到相互分离的情况，特别是p-value越低的时候分离程度就越高，即QQ-plot的尾部会翘起来（这是因为GWAS的零假设就是与随机突变相比没有区别）。

而且，我们知道基因组上的随机漂变是一定存在的，所以一定会有位点与随机漂变相关，特别是在p-value比较大的位点看起来就应该和随机漂变重叠，这就表现在QQ-plot的前半部分里，位点的分布会和均匀分布重叠！而且，比较好的结果是，当p-value < 10^-3时，GWAS结果开始与均匀分布出现快速分离——也就是说，自然选择的力量明显地显示出来了，使得结果在群体中快速摆脱随机性，最后看到一个高高翘起的QQ-plot。这时基本就可以断定，我们所研究的表型和基因型之间是存在着显著相关的自然选择作用的。

这也是评估一个GWAS研究的假阳性情况的最基本的判断。

关联结果可视化-曼哈顿图和QQ图

阈值的确定

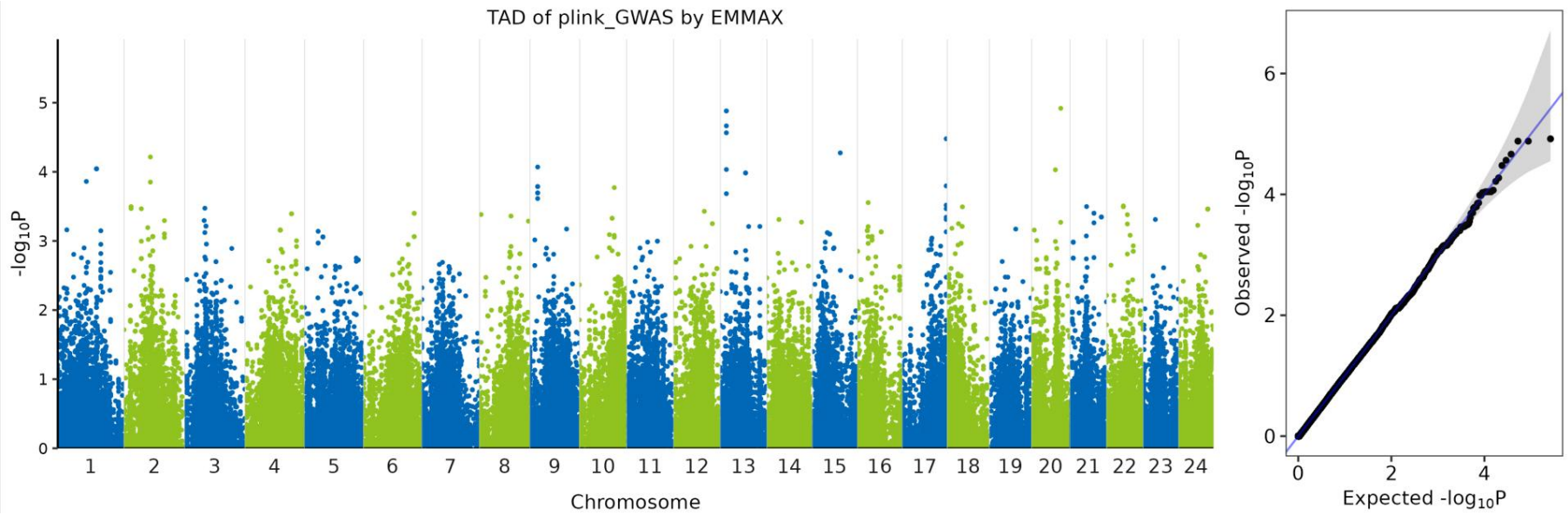
- 置换检验（permutation test)：科学可靠，但是计算量大
- **Bonferroni threshold**

  公式：显著水平（0.1，0.01，0.05）/检验次数，

  在GWAS中，检验次数就是SNP位点的个数，0.05比较常用。特点：严格但是比较灵活
- FDR 校验： 使用FDR矫正后，再确定一个显著水平筛选显著位点
- 经验值： 10^{-5}， 10^{-6}， $5x10^{-8}$等

关联结果可视化-曼哈顿图和QQ图

表型关联不显著的结果



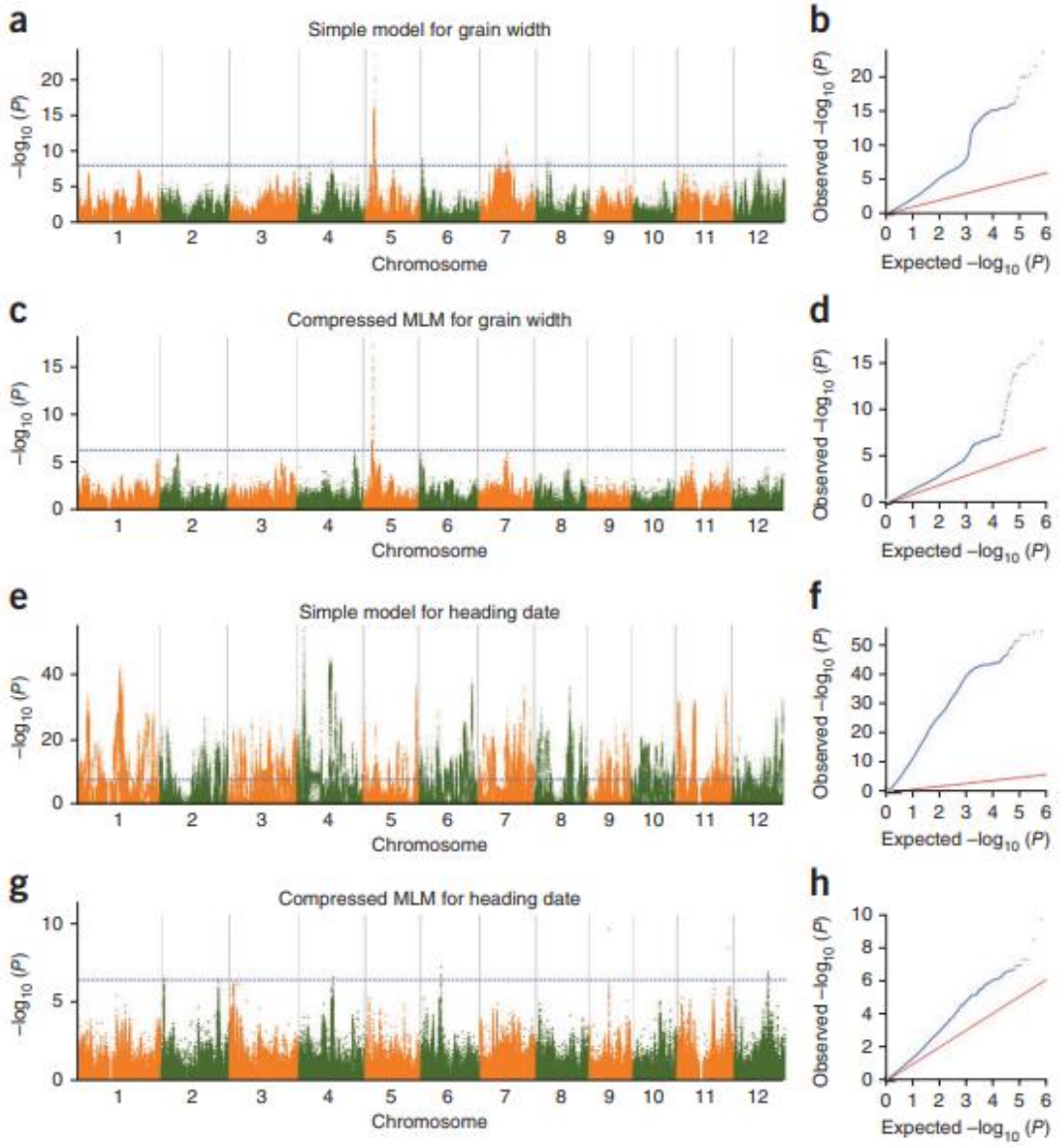
可能的原因：（1）性状考察不准确；（2）性状受环境影响比较大，性状由多个微效位点控制；（3）模型检测效力问题；（4）标记密度不够；（5）性状的变异是由表观修饰引起的，与表型无关。

挽救方法：
1）提供准确的表型数据重新进行关联，考虑改进表型检测方法；
2）多年多点的重复表型记录；
3）增加样本量（解决效应太小），或增加标记量；
4）更换关联分析模型

关联结果可视化-曼哈顿图和QQ图

假阳性结果

假阳性过高，可以更换分析模型，示例结果中亚群分化比较明显，用简单模型会有很多假阳性信号，更换CMLM模型后效果好很多，或者可以每个亚群单独进行关联分析。



关联结果可视化-曼哈顿图和QQ图

后续生信分析——获得显著位点后，如何锁定候选基因

1）对显著位点附近LD block分析，确定候选区间的范围；

2）对候选区间的基因进行结构注释；

编码区：非同义突变>同义突变。此外，启动子区域，内含子区域，UTR区域也要关注。

3）对候选区间的基因进行结构注释功能注释（包括GO,KEGG,NR)等，进一步筛选候选基因；

4）结合转录组数据或者qRT-PCR，分析候选基因表达量差异，进一步缩小候选基因范围；

5）同源基因分析，结合其他物种已发表的同源基因功能推测候选基因功能；

6）群体遗传-选择清楚分析结果的比较；

7）功能验证；

LD Block分析

LD: 当位于某一座位的特定等位基因与另一座位的某一等位基因同时出现的概率大于群体中因随机分布的两个等位基因同时出现的概率时，就称这两个座位处于连锁不平衡状态（linkage disequilibrium）

LD block也称单倍型块（Haplotype block)，是指同一条染色体上处于连锁不平衡状态的一段连锁区域

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

LD程度的评估—D

LD 计算的是实际观测的单倍型频率与随机分布时单倍型的期望频率之间的差异。一般用**D**，**D'**和 r^2 来表示LD的程度。

D的计算：

1、假设有两个位点（A、B），等位基因分别是A、a、B、b，在群体中对应的频率分别为P(A)、P(a)、P(B)和P(b)

2、两个位点共有四种单倍型AB、Ab、aB、ab，对应频率P(AB)、P(Ab)、P(aB)和P(ab)

3、计算：D_{AB}=P(AB)-P(A)*P(B)

当$_{DAB}$=0时，处于连锁平衡状态，即这两个位点不连锁；

当D_{AB}≠0时，处于连锁不平衡状态，即这两个位点存在连锁。

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

LD程度的评估—D′

因为D的取值强烈地依赖于人为决定的等位基因频率，所以它不利于LD程度的比较。标准化后的不平衡系数**D'**能够避免这种对等位基因频率的依赖。

$$D'=D/Dmax$$

当D < 0, 群体中P (AB)或者P(ab) 为0，所以Dmax = min{P(A)P(b),P(a)f(B) }

当D > 0, 群体中P(Ab)或者P(aB) 为0，Dmax = min{ P(A)P(B),P(a)P(b) }

-1 <= D′ <=1

当|D′ | = 1，表示**完全连锁不平衡**，没有重组打破；

当 D′ = 0，表示**完全连锁平衡**，随机组合；

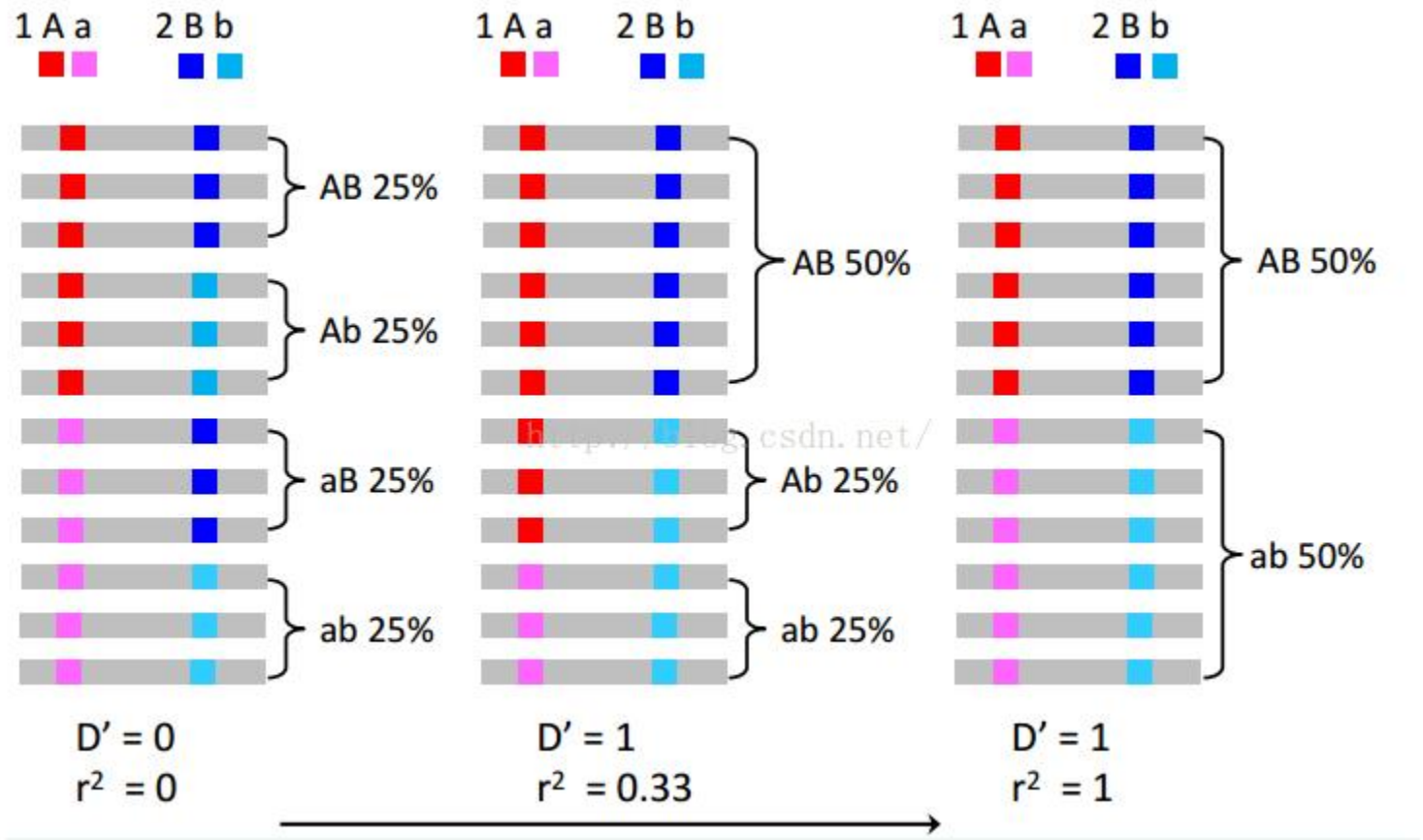
当0 <|D′ | < 1时，表示存在连锁不平衡，但是D′ 的数值不能表示连锁不平衡的程度，因此，引进 r^2 来表示LD。

LD Block分析

LD程度的评估—r^2

公式：
$r^2=D^2/(P(A) * P(a) * P(B) * P(b))$

当$r^2=1$，表示连锁完全不平衡，没有重组
当$r^2=0$，表示连锁完全平衡，随机组合



LD Block分析

R^2和D’ 的应用

- r^2和D’ 反映了LD的不同方面。r^2包括了重组和突变，而D’ 只包括重组史。D’ 能更准确地估测重组差异，但当样本量较小时，低频率等位基因组合可能无法观测到，导致LD强度被高估，所以D’ 不适合小样本群体研究；
- LD衰减作图中通常采用r^2来表示群体的LD水平；
- Haplotype Block中通常采用D’ 来定义Block；

LD Block分析

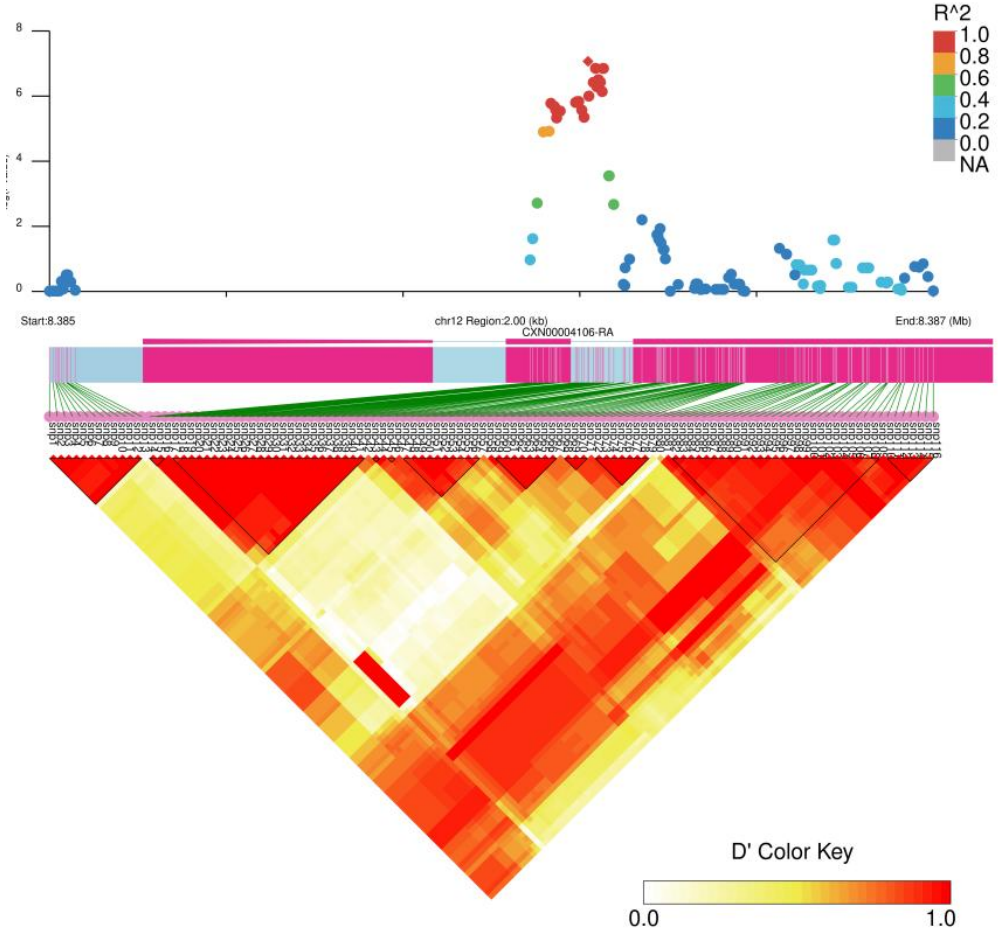
LD Block分析结果解读

上图横坐标：染色体的位置坐标；
上图纵坐标：GWAS结果的-logp值；
上图中的每个圆点代表一个snp位点，圆点的颜色表示该snp与目标snp（红色菱形的点）的连锁程度，连锁程度越强，R^2越大，即颜色越红；

中间的蓝色矩形上方：位于该区间内的基因名及基因结构；
中间的蓝色矩形也表示染色体的位置坐标，矩形上的每条线代表一个snp位点；
中间的蓝色矩形下方：每条绿线也代表一个snp位点；

最下方的倒三角形：表示该区间内所有snp位点两两间的连锁程度，连锁程度越强，D’越大，即颜色越红。

该图的重点是看下方的倒三角形中有没有强连锁（很红）的一段block（黑色边框标注），且该block区域内包含目标snp。



LD Block分析

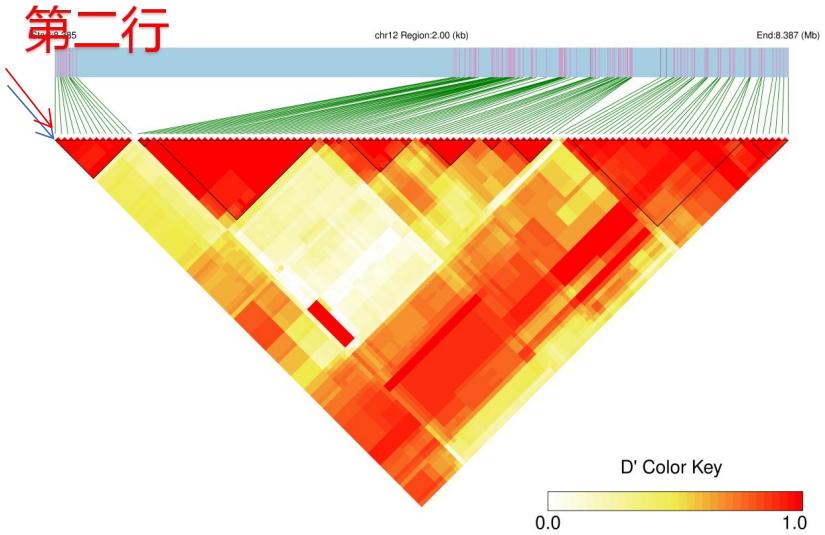
云南农业大学 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

LD Block分析结果解读

文件名：*.TriangleV.gz (D')

1	#Region	PairWise D'														
2	1.000	1.000	1.000	1.000	1.000	0.970	0.969	0.968	0.968	0.968	1.000	1.000	0.194	0.417	0.422	0.439	0.505
3	1.000	1.000	1.000	1.000	0.970	0.969	0.968	0.968	0.968	1.000	1.000	0.194	0.417	0.422	0.439	0.505	0.457
4	1.000	1.000	1.000	0.970	0.969	0.968	0.968	0.968	1.000	1.000	0.194	0.417	0.422	0.439	0.505	0.457	0.455
5	1.000	1.000	0.970	0.969	0.968	0.968	0.968	1.000	1.000	0.194	0.417	0.422	0.439	0.505	0.457	0.455	0.462
6	1.000	0.970	0.969	0.968	0.968	0.968	1.000	1.000	0.194	0.417	0.422	0.439	0.505	0.457	0.455	0.462	0.462
7	0.970	0.969	0.968	0.968	0.968	1.000	1.000	0.194	0.417	0.422	0.439	0.505	0.457	0.455	0.462	0.462	0.455
8	1.000	1.000	1.000	1.000	1.000	1.000	0.227	0.441	0.447	0.463	0.453	0.410	0.409	0.416	0.417	0.409	0.466
9	0.969	0.969	0.969	1.000	1.000	0.235	0.447	0.453	0.469	0.459	0.417	0.416	0.422	0.424	0.409	0.472	0.477
10	1.000	1.000	1.000	1.000	0.257	0.330	0.258	0.357	0.345	0.310	0.310	0.310	0.312	0.304	0.382	0.387	0.382
11	1.000	1.000	1.000	0.257	0.330	0.258	0.357	0.345	0.310	0.310	0.310	0.312	0.304	0.382	0.387	0.382	0.374
12	1.000	1.000	0.257	0.330	0.258	0.357	0.345	0.310	0.310	0.310	0.312	0.304	0.382	0.387	0.382	0.374	0.377
13	1.000	0.286	0.356	0.287	0.321	0.310	0.278	0.280	0.280	0.284	0.267	0.312	0.317	0.312	0.304	0.307	0.306
14	0.022	0.178	0.166	0.270	0.260	0.237	0.234	0.226	0.234	0.226	0.272	0.283	0.280	0.272	0.276	0.279	0.272
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

第一行



第一行代表的是第一个SNP与其他SNP位点的D’值
第二行代表的是第二个SNP与其他SNP位点的D’值
……

LD Block分析

LD Block分析结果解读

文件名：*.TriangleB.gz(R^2)

1	#Region	PairWise	R^2															
2	1.000	1.000	1.000	1.000	1.000	0.840	0.816	0.744	0.744	0.744	0.703	0.485	0.003	0.023	0.031	0.061	0.080	0.069	0.071
3	1.000	1.000	1.000	1.000	0.840	0.816	0.744	0.744	0.744	0.703	0.485	0.003	0.023	0.031	0.061	0.080	0.069	0.071	0.076
4	1.000	1.000	1.000	0.840	0.816	0.744	0.744	0.744	0.703	0.485	0.003	0.023	0.031	0.061	0.080	0.069	0.071	0.076	0.081
5	1.000	1.000	0.840	0.816	0.744	0.744	0.744	0.703	0.485	0.003	0.023	0.031	0.061	0.080	0.069	0.071	0.076	0.081	0.076
6	1.000	0.840	0.816	0.744	0.744	0.744	0.703	0.485	0.003	0.023	0.031	0.061	0.080	0.069	0.071	0.076	0.081	0.076	0.125
7	0.840	0.816	0.744	0.744	0.744	0.703	0.485	0.003	0.023	0.031	0.061	0.080	0.069	0.071	0.076	0.081	0.076	0.125	0.121
8	0.972	0.890	0.890	0.890	0.788	0.543	0.005	0.030	0.039	0.076	0.070	0.061	0.063	0.068	0.074	0.069	0.120	0.123	0.120
9	0.860	0.860	0.860	0.811	0.559	0.006	0.031	0.042	0.081	0.075	0.065	0.068	0.072	0.079	0.069	0.128	0.131	0.128	0.120
10	1.000	1.000	0.885	0.610	0.008	0.019	0.015	0.051	0.046	0.039	0.041	0.041	0.046	0.042	0.090	0.093	0.090	0.083	0.087
11	1.000	0.885	0.610	0.008	0.019	0.015	0.051	0.046	0.039	0.041	0.041	0.046	0.042	0.090	0.093	0.090	0.083	0.087	0.082
12	0.885	0.610	0.008	0.019	0.015	0.051	0.046	0.039	0.041	0.041	0.046	0.042	0.090	0.093	0.090	0.083	0.087	0.082	0.083
13	0.689	0.011	0.025	0.021	0.047	0.043	0.036	0.039	0.039	0.044	0.036	0.070	0.072	0.070	0.064	0.068	0.062	0.058	0.050
14	0.000	0.009	0.010	0.049	0.043	0.038	0.038	0.034	0.041	0.037	0.071	0.080	0.078	0.071	0.076	0.075	0.071	0.055	0.070
15	0.711	0.546	0.309	0.308	0.293	0.291	0.291	0.288	0.288	0.226	0.227	0.226	0.226	0.225	0.228	0.226	0.242	0.240	0.240
16	0.768	0.434	0.433	0.411	0.409	0.409	0.406	0.406	0.318	0.319	0.318	0.318	0.317	0.320	0.318	0.339	0.337	0.337	0.277
17	0.566	0.565	0.536	0.534	0.534	0.530	0.530	0.416	0.417	0.416	0.416	0.414	0.417	0.416	0.442	0.439	0.439	0.381	0.381
18	0.898	0.950	0.950	0.950	0.949	0.949	0.745	0.746	0.745	0.745	0.744	0.746	0.745	0.784	0.782	0.782	0.732	0.732	0.730

LD Block分析

LD Block分析结果解读

文件名：*.blocks.gz

CHR	BP1	BP2	KB	NSNPS	SNPS
12	8384734	8384792	0.059	13	8384734|8384742|8384748|8384749|8384753|8384756|8384761|8384765|8384772|8384773|8384775|8384782|8384792
12	8385821	8385837	0.017	3	8385821|8385827|8385837
12	8385851	8386010	0.16	26	8385851|8385864|8385868|8385876|8385879|8385881|8385889|8385924|8385927|8385931|8385938|8385943|8385952|8385954|8385969|8385972|8385975|8385976|8385980|8385984|8385986|8385987|8385999|8386000|8386010
12	8386032	8386035	0.004	2	8386032|8386035
12	8386036	8386046	0.011	2	8386036|8386046
12	8386074	8386156	0.083	11	8386074|8386108|8386110|8386112|8386115|8386117|8386122|8386124|8386127|8386138|8386156
12	8386196	8386259	0.064	9	8386196|8386200|8386207|8386208|8386218|8386241|8386247|8386255|8386259
12	8386270	8386276	0.007	4	8386270|8386271|8386272|8386276
12	8386283	8386307	0.025	8	8386283|8386290|8386294|8386295|8386302|8386303|8386305|8386307
12	8386420	8386659	0.24	28	8386420|8386423|8386431|8386439|8386442|8386451|8386458|8386471|8386477|8386478|8386506|8386510|8386513|8386514|8386542|8386548|8386549|8386573|8386575|8386583|8386587|8386615|8386627|8386629|8386630|8386653|8386659
12	8386661	8386733	0.073	7	8386661|8386668|8386691|8386701|8386710|8386721|8386733

文件名：out.site.gz

chr12	8384734
chr12	8384742
chr12	8384748
chr12	8384749
chr12	8384753
chr12	8384756

LD Block分析

LD Block分析结果解读

文件名：*.blocks.gz

CHR	BP1	BP2	KB	NSNPS	SNPS
12	8384734	8384792	0.059	13	8384734|8384742|8384748|8384749|8384753|8384756|8384761|8384765|8384772|8384773|8384775|8384782|8384792
12	8385821	8385837	0.017	3	8385821|8385827|8385837
12	8385851	8386010	0.16	26	8385851|8385864|8385868|8385876|8385879|8385881|8385889|8385924|8385927|8385931|8385938|8385943|8385952|8385954|8385969|8385972|8385975|8385976|8385980|8385984|8385986|8385987|8385999|8386000|8386010
12	8386032	8386035	0.004	2	8386032|8386035
12	8386036	8386046	0.011	2	8386036|8386046
12	8386074	8386156	0.083	11	8386074|8386108|8386110|8386112|8386115|8386117|8386122|8386124|8386127|8386138|8386156
12	8386196	8386259	0.064	9	8386196|8386200|8386207|8386208|8386218|8386241|8386247|8386255|8386259
12	8386270	8386276	0.007	4	8386270|8386271|8386272|8386276
12	8386283	8386307	0.025	8	8386283|8386290|8386294|8386295|8386302|8386303|8386305|8386307
12	8386420	8386659	0.24	28	8386420|8386423|8386431|8386439|8386442|8386451|8386458|8386471|8386477|8386478|8386506|8386510|8386513|8386514|8386542|8386548|8386549|8386573|8386575|8386583|8386587|8386615|8386627|8386629|8386630|8386653|8386659
12	8386661	8386733	0.073	7	8386661|8386668|8386691|8386701|8386710|8386721|8386733

文件名：out.site.gz

chr12	8384734
chr12	8384742
chr12	8384748
chr12	8384749
chr12	8384753
chr12	8384756

LD Block分析

LD Block分析分析后我们还能做什么?

如何确定候选基因?

1. 根据tag SNP 处于同一个LD block中的SNP位点的功能基因，也可以扩大范围

2.单倍型分析结合一代测序明确变异位点

3. 结合转录组数据锁定候选基因

生信帮，答你所问

雲南農業大學

Yunnan Agricultural University

云南省生物大数据重点实验室