

群体进化分析

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20日

目录



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

第一章

群体进化简介

第二章

群体进化的研究技术

第三章

系统发育树(Tree)分析

第四章

主成分(PCA)分析

第五章

群体结构(Structure) 分析

目录



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

第六章

遗传多样性 (π) 计算

第七章

遗传距离 (f_{st}) 计算

第八章

选择性清除分析

第九章

连锁不平衡衰减分析 (LDdecay)

群体进化简介



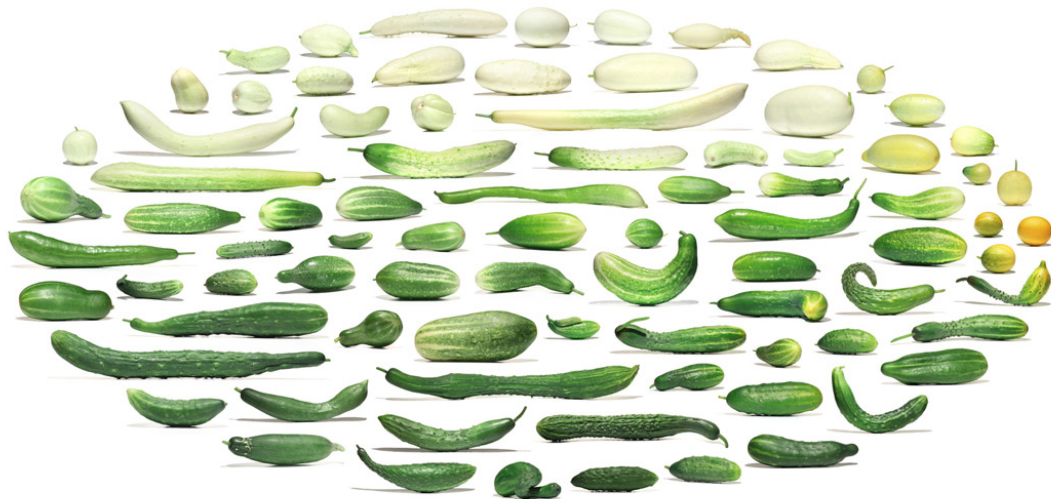
雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

群体进化 (Population Evolution) 是指在群体水平上发生的遗传变异、选择和演化过程。与个体进化 (Individual Evolution) 关注单个个体遗传信息的变化和选择不同，群体进化侧重于研究整个群体的遗传变异和选择。



群体进化简介



雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

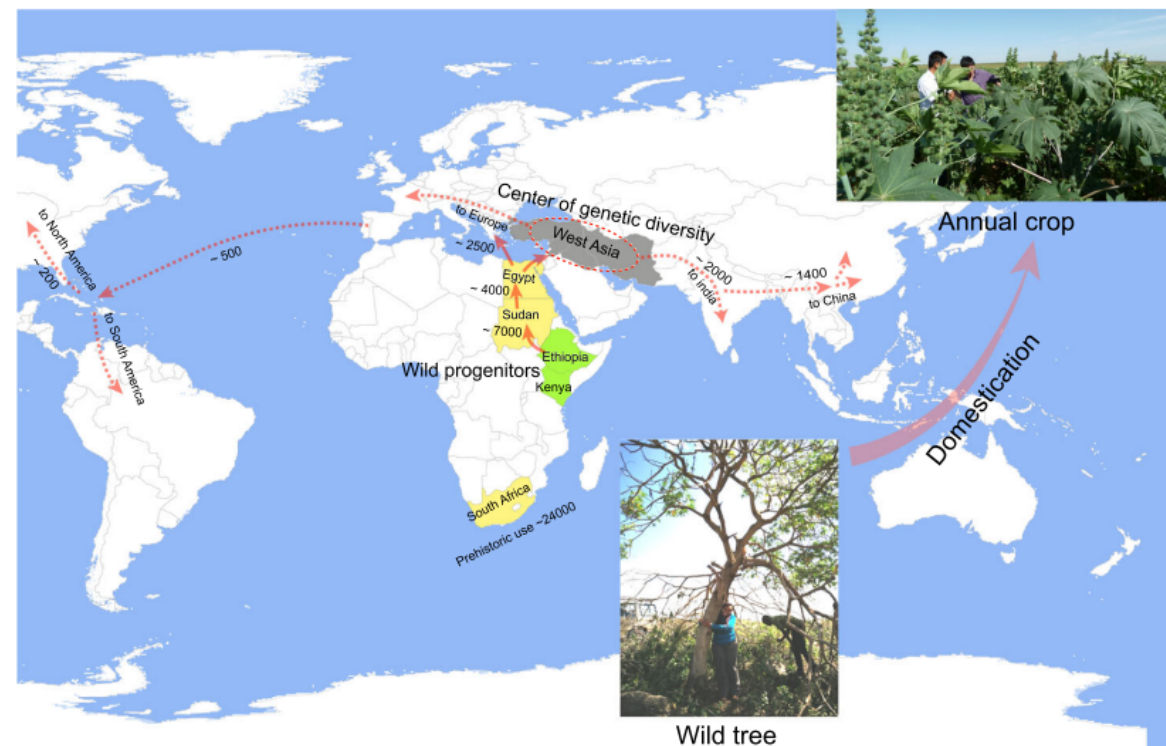
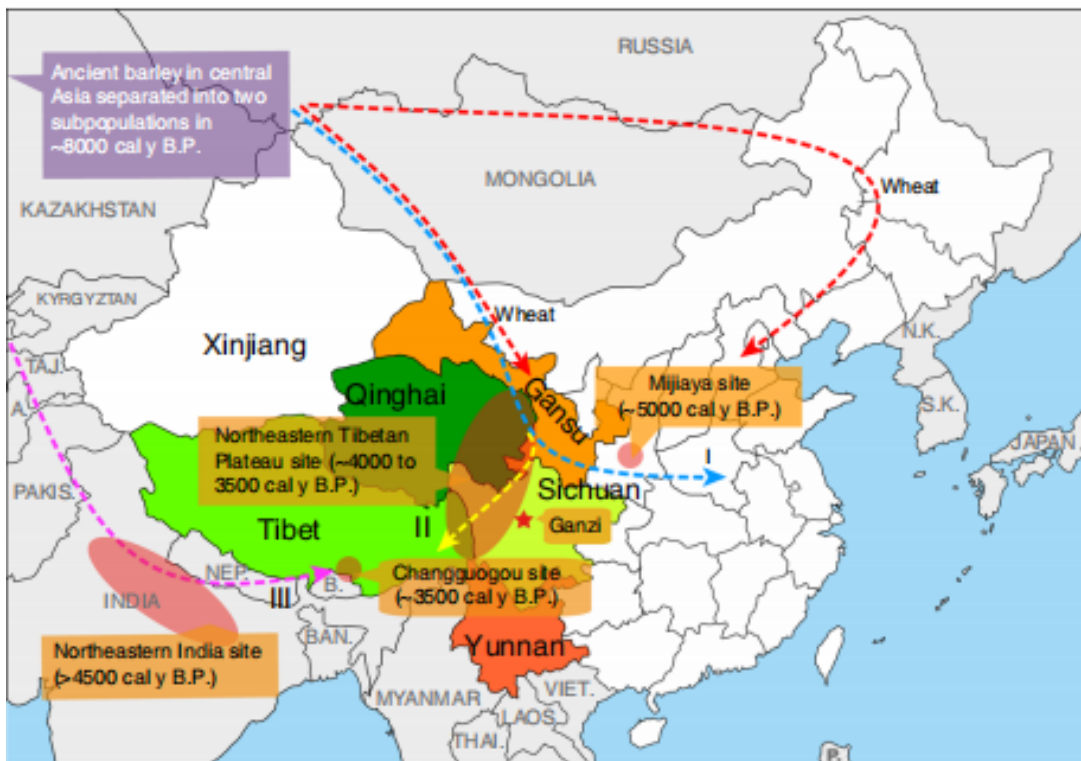
群体进化涉及到群体内的个体之间的相互作用和选择过程，这些相互作用和选择会对群体的遗传构成产生影响。群体内的个体可能会表现出合作、竞争、协调等不同的行为，这些行为的进化会影响整个群体的适应性和生存能力。





起源驯化

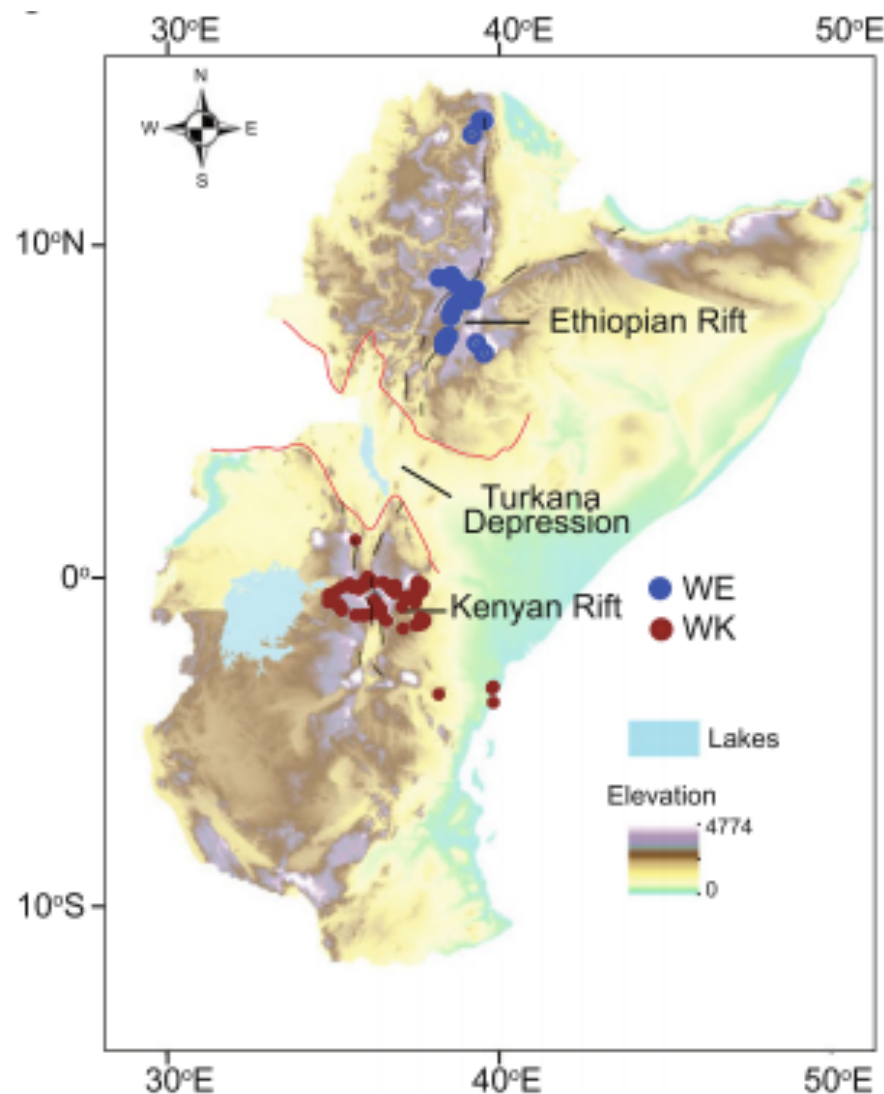
收集野生种、驯化种或栽培种等材料，通过研究不同群体的群体结构、遗传多样性、遗传距离、种群动态变化，基因渗入、遗传漂变等，来推断物种的起源地及驯化过程；





环境适应性

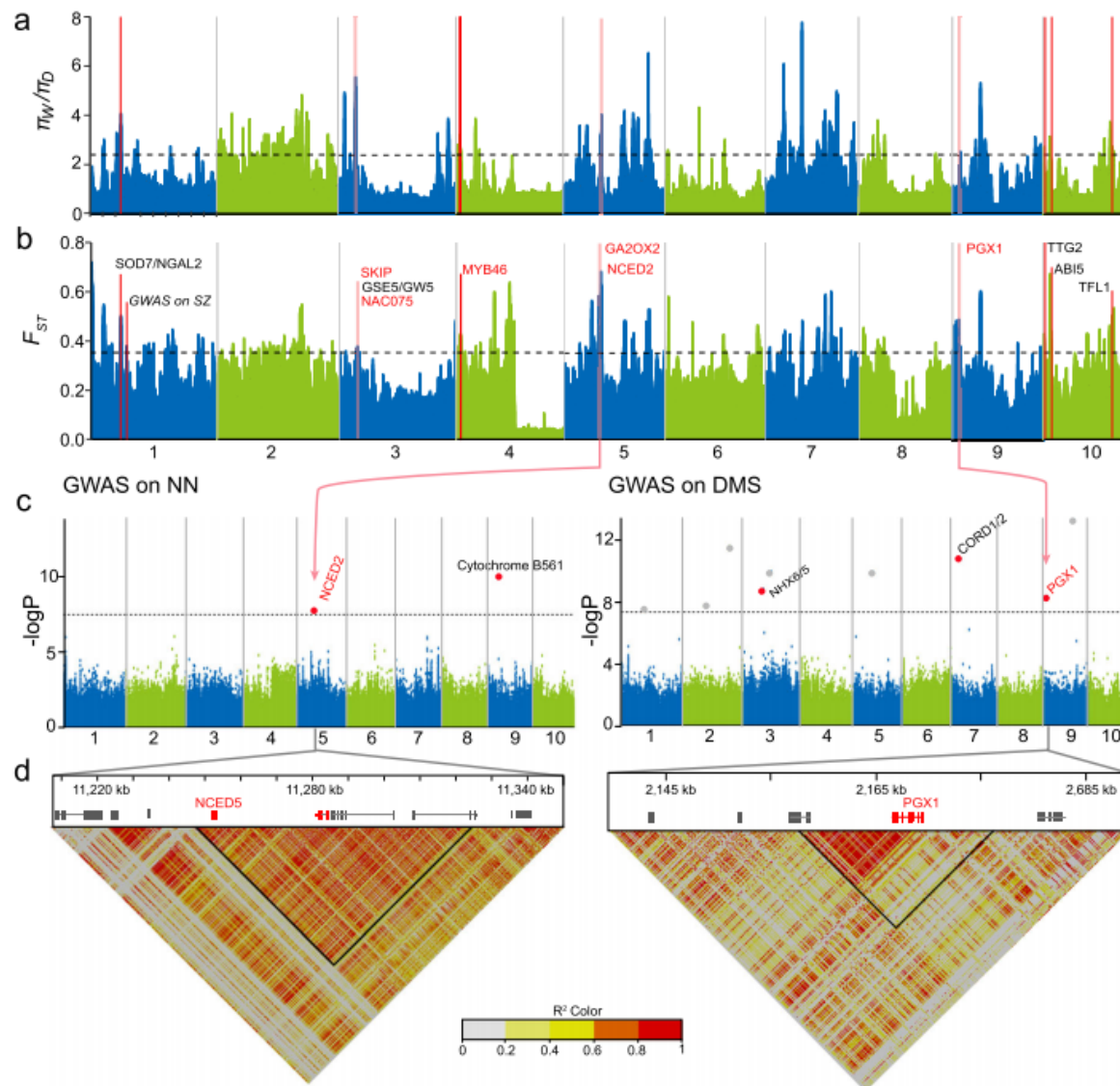
收集不同地区或不同环境的材料，通过研究不同群体间的群体结构、选择压力、连锁不平衡衰减、生态位模拟等，来解释特定群体对特定环境的适应性；





基因挖掘

收集不同地区或不同环境的材料，通过研究不同群体间的选择压力，挖掘受选择的基因，可结合全基因组关联分析（GWAS）进行基因深入挖掘和定位；



群体进化的研究技术



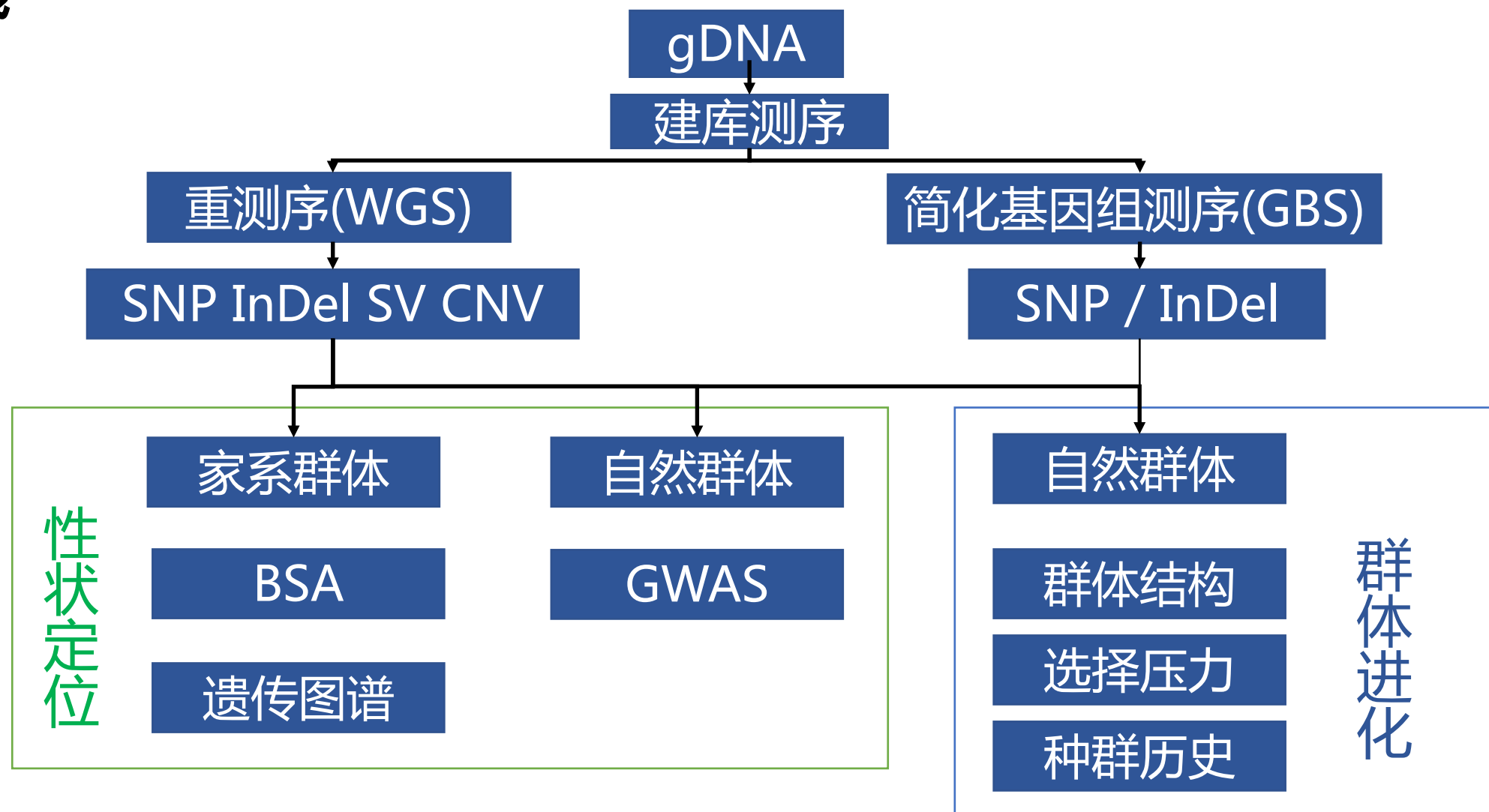
雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

技术路线



群体进化的研究技术



雲南農業大學
Yunnan Agricultural University



生信帮

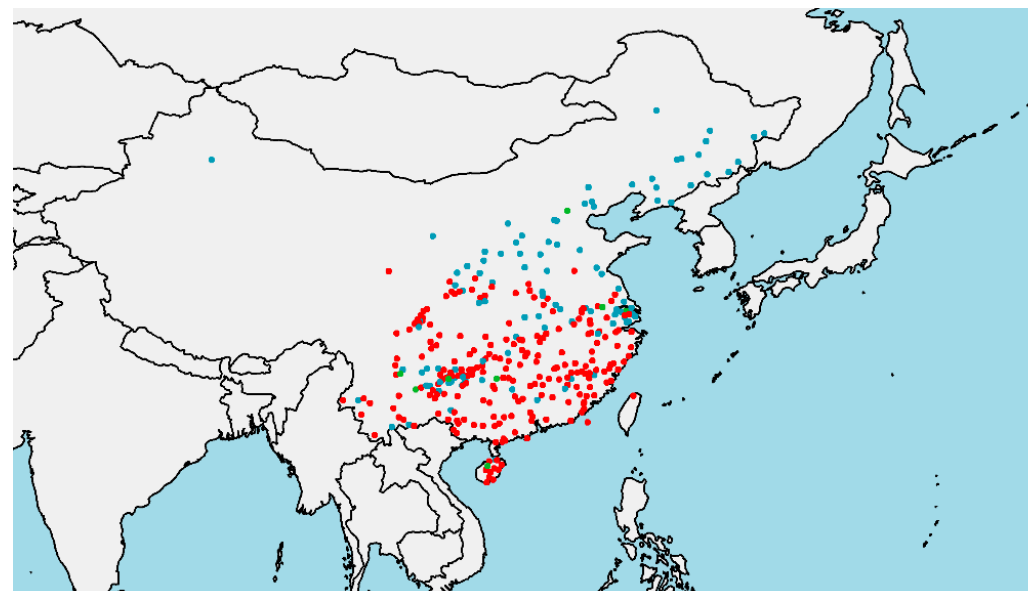
云南省生物大数据重点实验室

样本采集

从地理分布可以获得的信息

1. 是否和山川河流有关?
2. 是否和海拔有关?
3. 是否和南北/东西有关?
4. 是否和沙漠盆地等有关?

* 充分采样





实验设计

➤ 取多少样

最少3个亚群，每个亚群最少10个样本

➤ 预算

样本数 $\times$ 测序数据量 = 测序成本

优先保证样本数，测序深度可适当降低

➤ 测多少数据量

二倍体：最低10X

异源多倍体：最低10X

同源多倍体：最低30X

如果想关注CNV和SV变异，深度需要再高1倍。

➤ 样本不够怎么办

可以考虑下载已发表文章中的样本数据进行分析



实验设计

➤ 用重测序技术还是简化技术？

使用简化技术的情况：

- 目前无参
- 超大基因组
 - 大麦5G基因组
 - 小麦17G基因组

重测序能获得更加全面的变异信息

系统发育树分析



雲南農業大學
Yunnan Agricultural University



生信帮

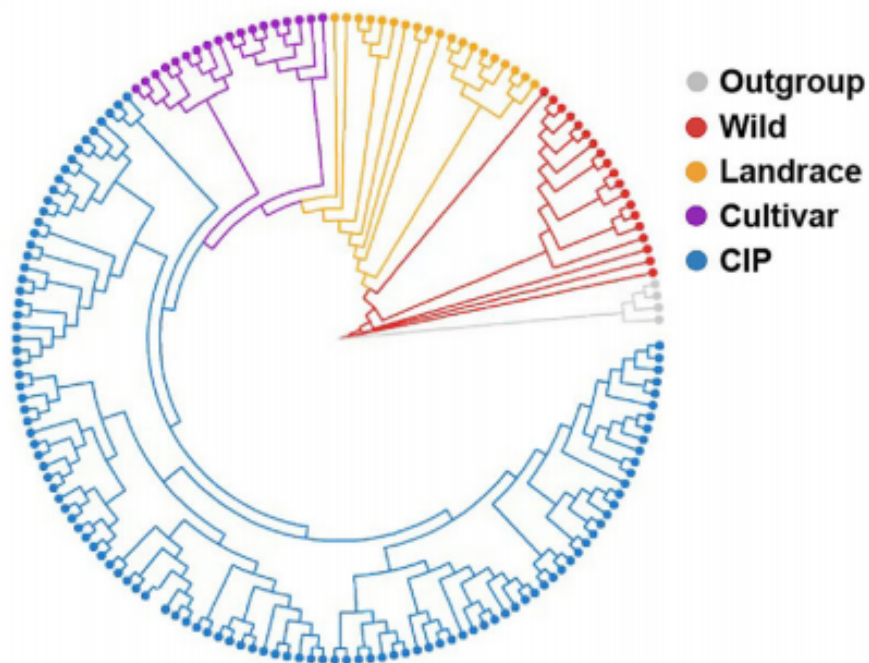
云南省生物大数据重点实验室

系统发育树（Phylogenetic Tree），也称进化树或类系树，是一种用于表示生物物种之间进化关系的图形结构。它反映了物种的演化历史和亲缘关系，展示了不同物种之间的进化关系和共同祖先。

在系统发育树中，物种通过节点（node）和分支（branch）连接在一起。节点表示共同祖先，分支表示进化路径。树的根节点代表最早的共同祖先，而叶节点（又称为末端节点）代表现存的物种。

系统发育树通过对比物种的形态特征、遗传信息（如DNA序列）或其他分子标记等数据来推断物种之间的关系；

A



1. 群体分层
2. 亚群间的进化关系
3. 有外类群的情况下，确定祖先群

系统发育树分析



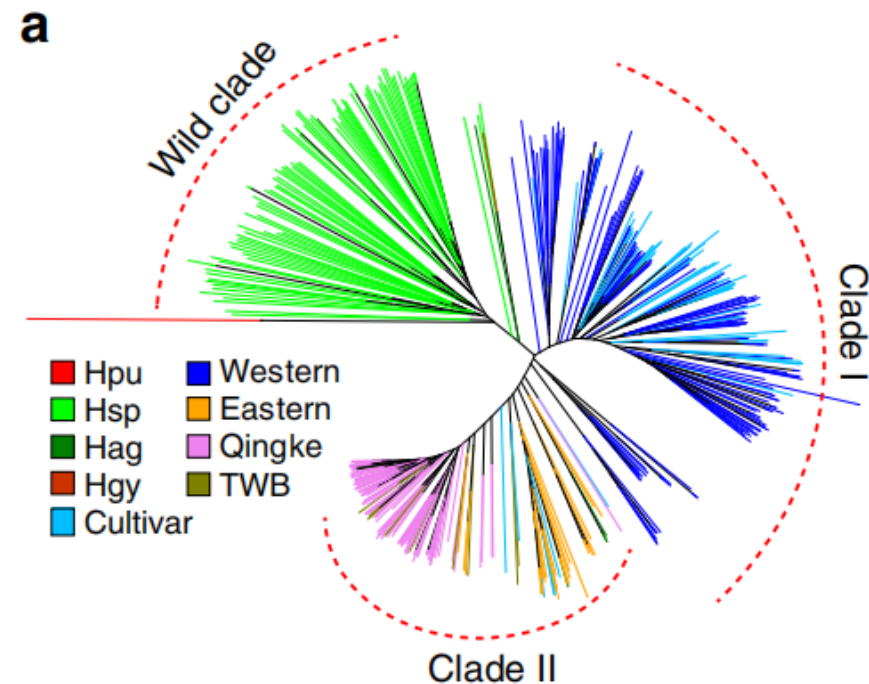
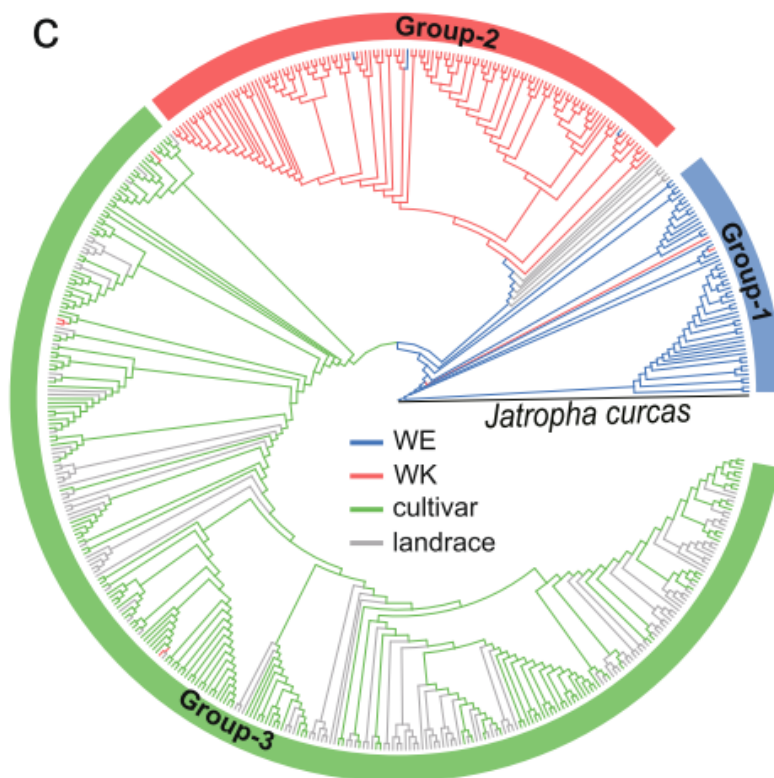
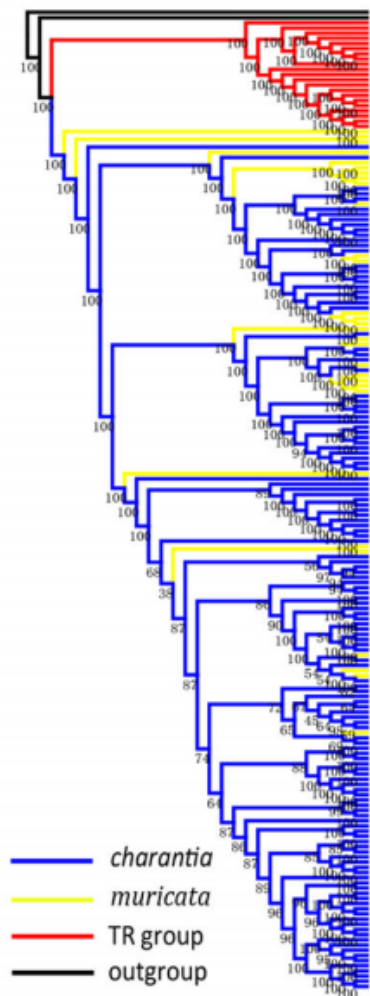
雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

不同树形展示



主成分 (PCA) 分析



雲南農業大學
Yunnan Agricultural University

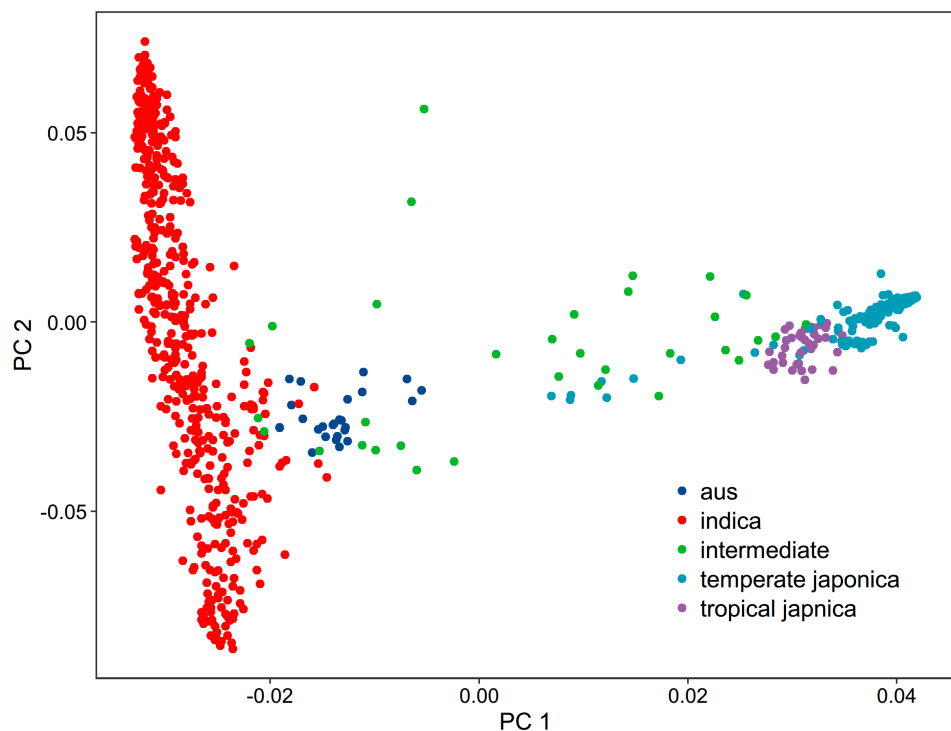


生信帮

云南省生物大数据重点实验室

主成分分析 (Principal Component Analysis, PCA) 是一种常用的数据降维技术和统计分析方法。它能够从高维数据中提取出最重要的特征，将原始数据转换为一组相互正交的低维变量，称为主成分。这些主成分按照重要性排序，第一个主成分包含了原始数据中大部分的变异性，第二个主成分则捕捉到剩余变异性中的最大部分，依次类推；

通过PCA分析，可以实现数据的降维和可视化，发现数据中的变异模式和聚类结构；



1. 群体分层
2. 剔除离群样本
3. 群体间遗传关系直观展示

主成分 (PCA) 分析



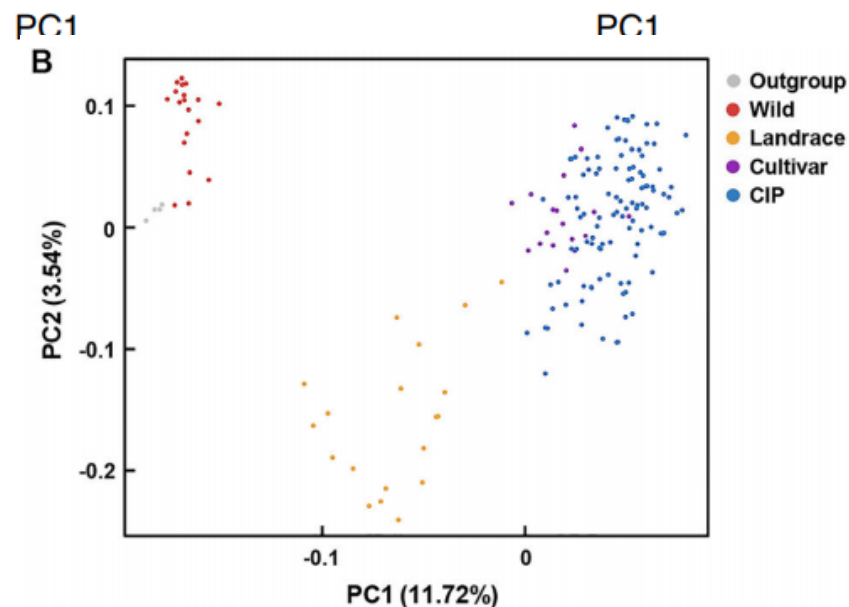
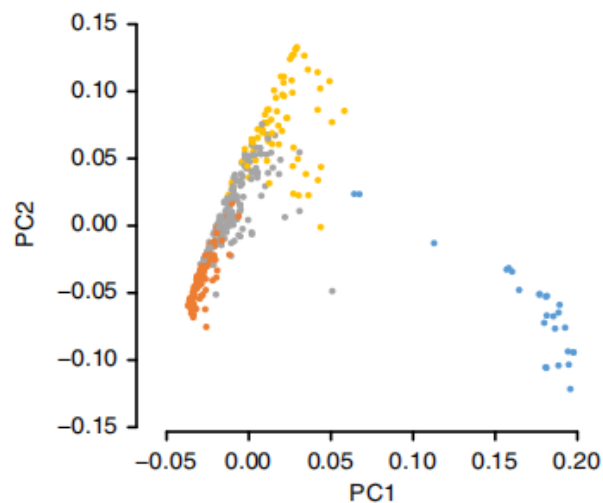
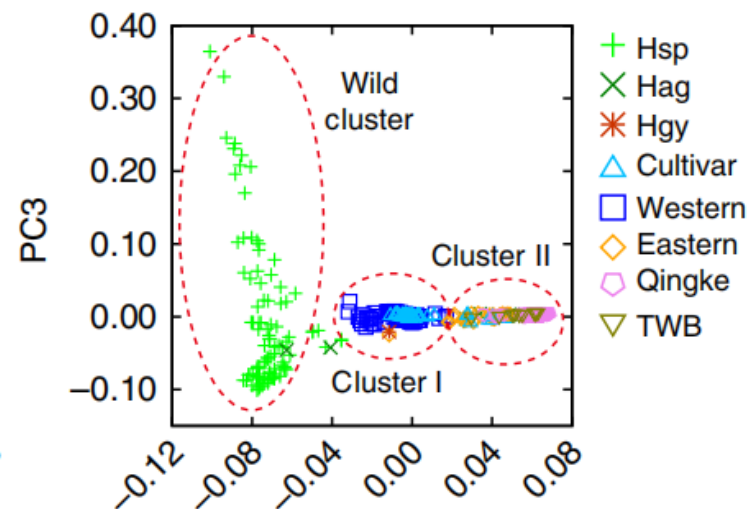
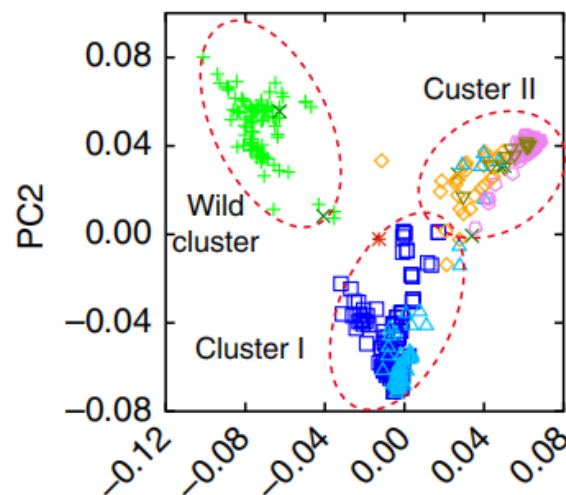
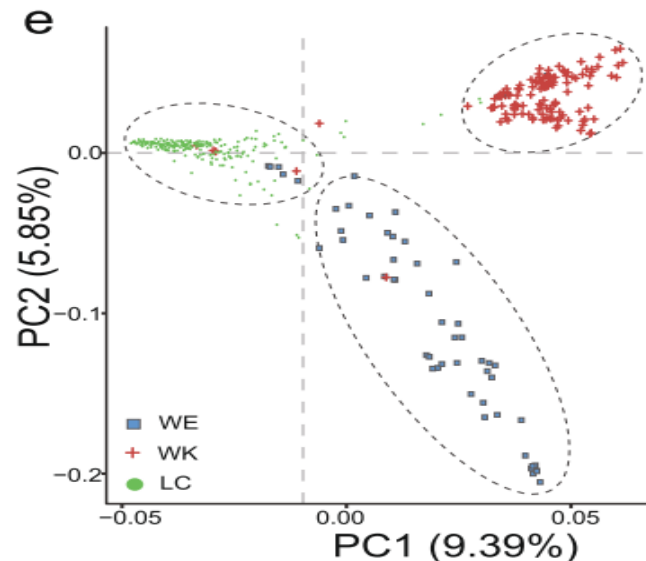
雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

PCA可视化结果展示



群体结构 (Structure)



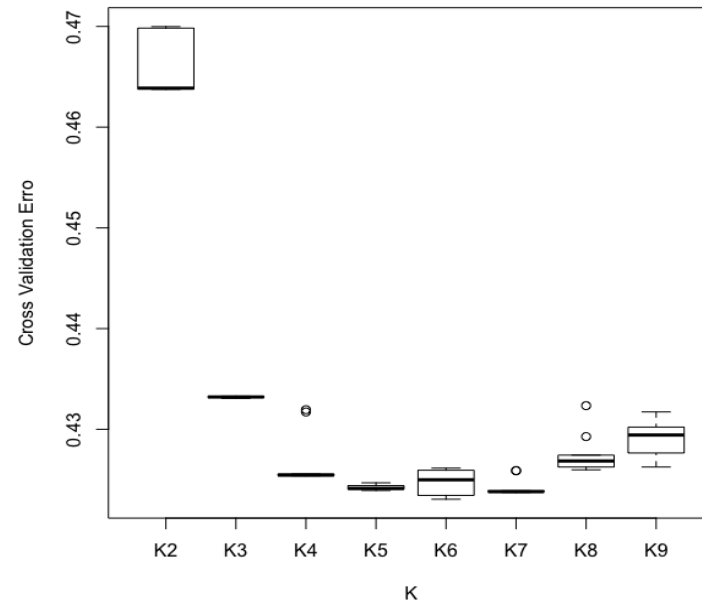
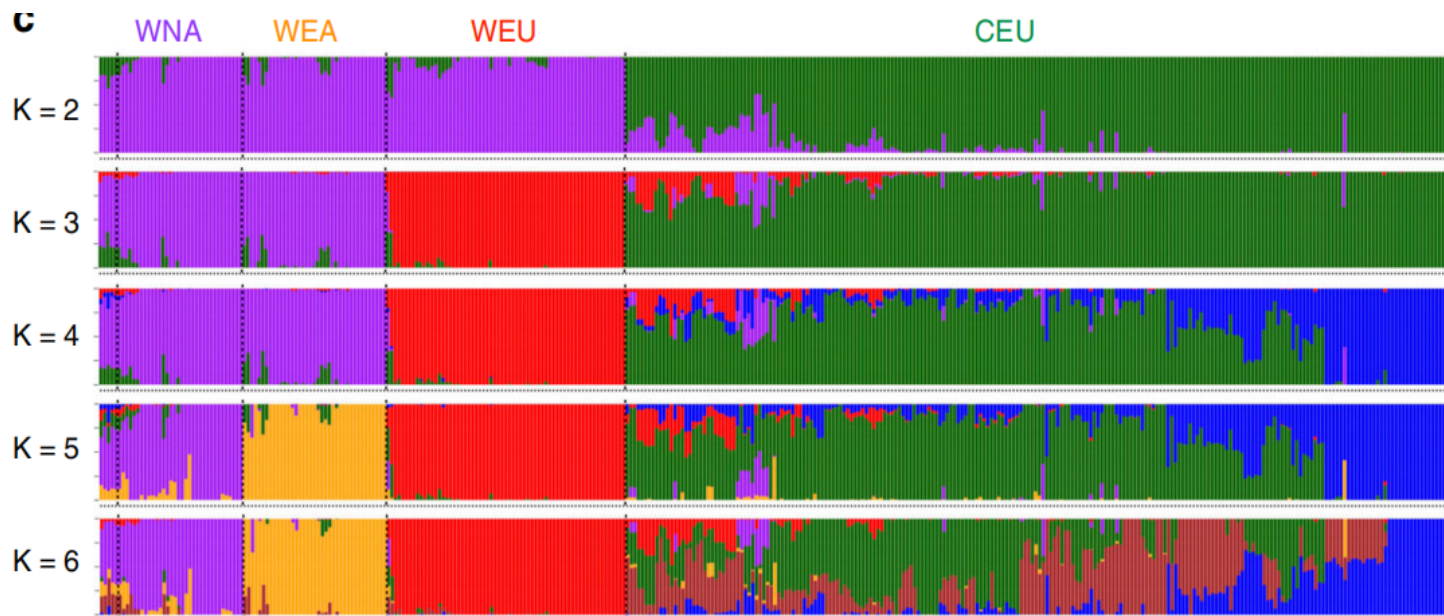
雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

群体结构 (Population structure) 分析是一种通过遗传数据来研究群体间遗传结构和关系的方法。通过遗传变异位点的多态性揭示不同个体或群体之间的遗传差异，从而推断出个体或群体之间的亲缘关系、迁移模式、遗传交流以及群体历史等信息。



1. 个体遗传背景的混杂程度
2. 群体间的进化关系
3. 最优分群数

群体结构 (Structure)



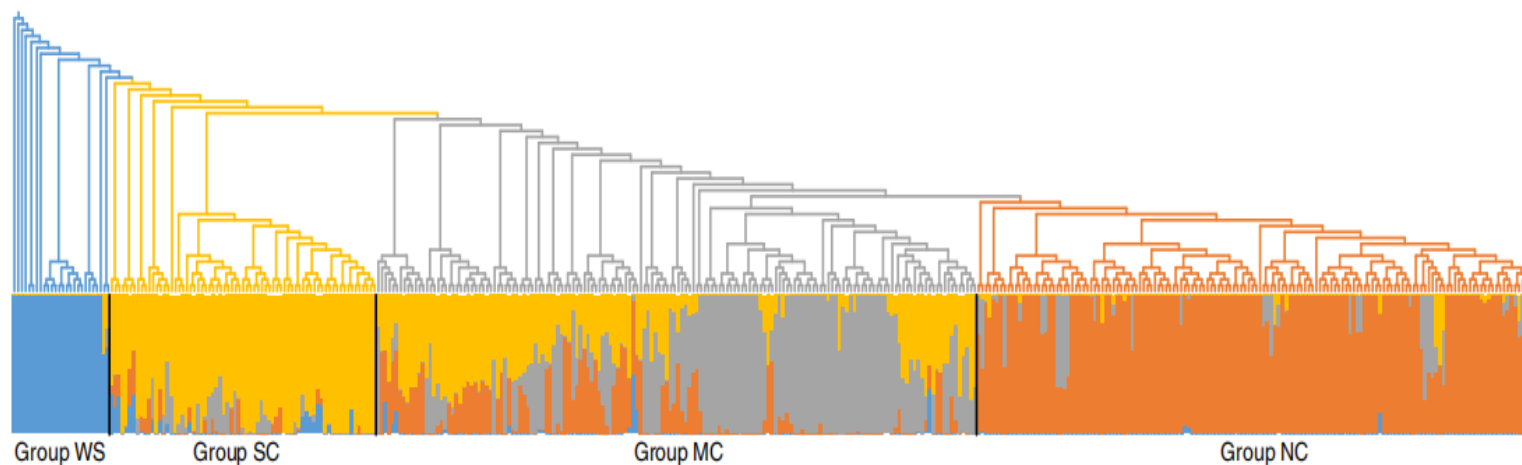
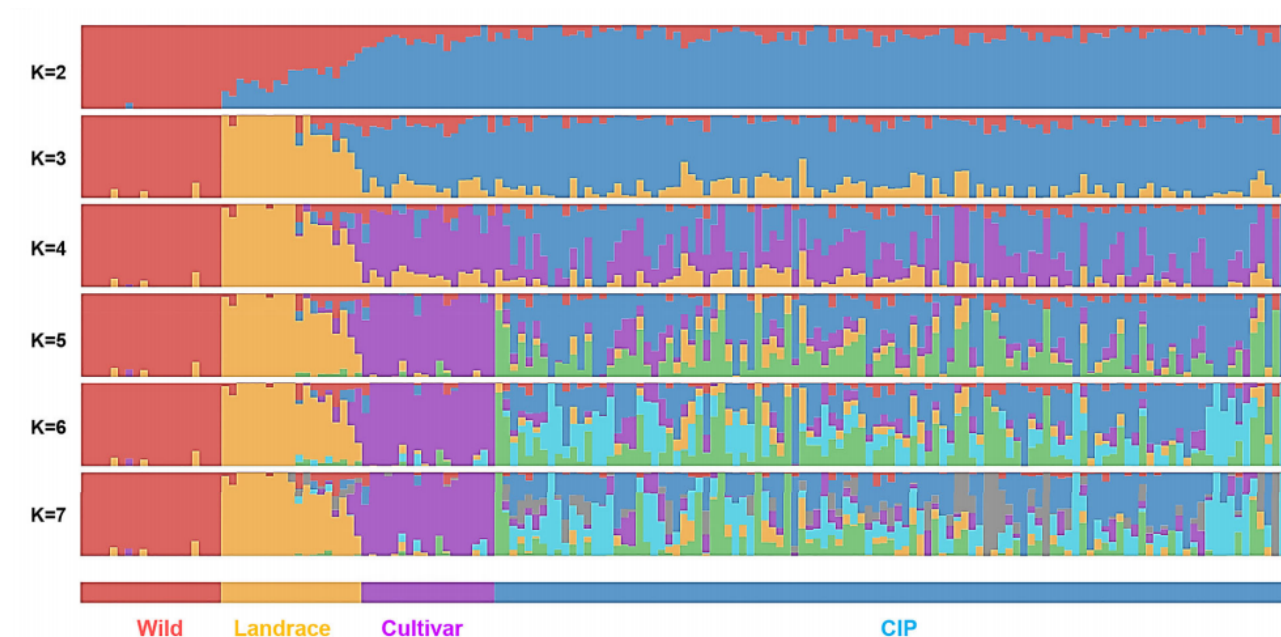
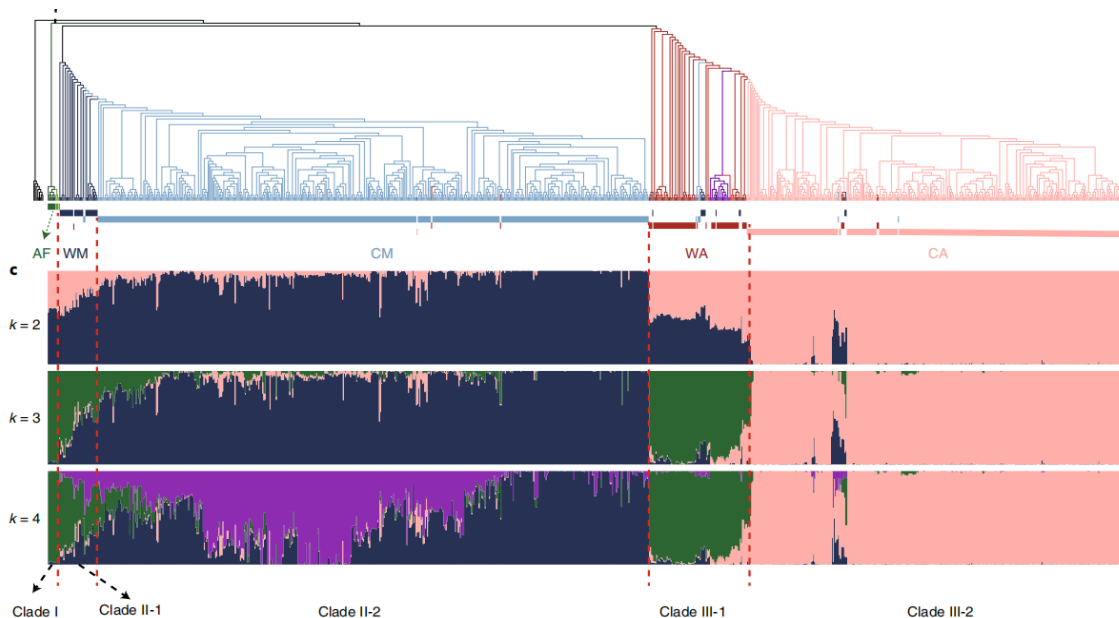
雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

structure可视化结果展示



遗传多样性 (pi) 计算



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

遗传多样性 (Genetic Diversity Pi) 是一种用于衡量**群体内**遗传变异水平的统计指标。用来描述种群内个体之间的遗传差异程度。

公式：

$$\pi = \sum_{i < j}^m \frac{d_{ij}}{c}$$

$$\pi \text{ (per site)} = \pi / n \quad (n : \text{序列的总长度})$$

m : 序列 (个体) 的数目 ; **c** : 序列间比较的次数 (C_m^2) ; **d** : 两序列间差异的碱基数

Pi值越大, 表示种群内个体之间的遗传差异越大, 即种群的遗传多样性越高。反之, **Pi**值越小意味着种群内个体之间的遗传差异越小, 即种群的多样性越低。

通常来说, 受到自然驯化或人工选择的种群, 其遗传多样性会降低, 原始野生群体的多样性往往比较高。

遗传距离 (fst) 计算



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

遗传距离 (Fixation index, F_{st}) 是一种用于衡量不同群体之间遗传差异的统计指标。

F_{st} 是基于杂合度计算的，它衡量了总的遗传变异中由于群体内遗传变异和群体间遗传变异所贡献的比例。

计算公式: $F_{st} = (H_t - H_s) / H_t$

H_t : 总群体的期望杂合度;

H_s : 每个亚群内随机交配的平均期望杂合度;

F_{st} 值介于0和1之间，其中0表示群体间没有遗传差异，1表示群体间没有基因交流完全隔离。

遗传距离 (fst) 计算



雲南農業大學
Yunnan Agricultural University



生信帮

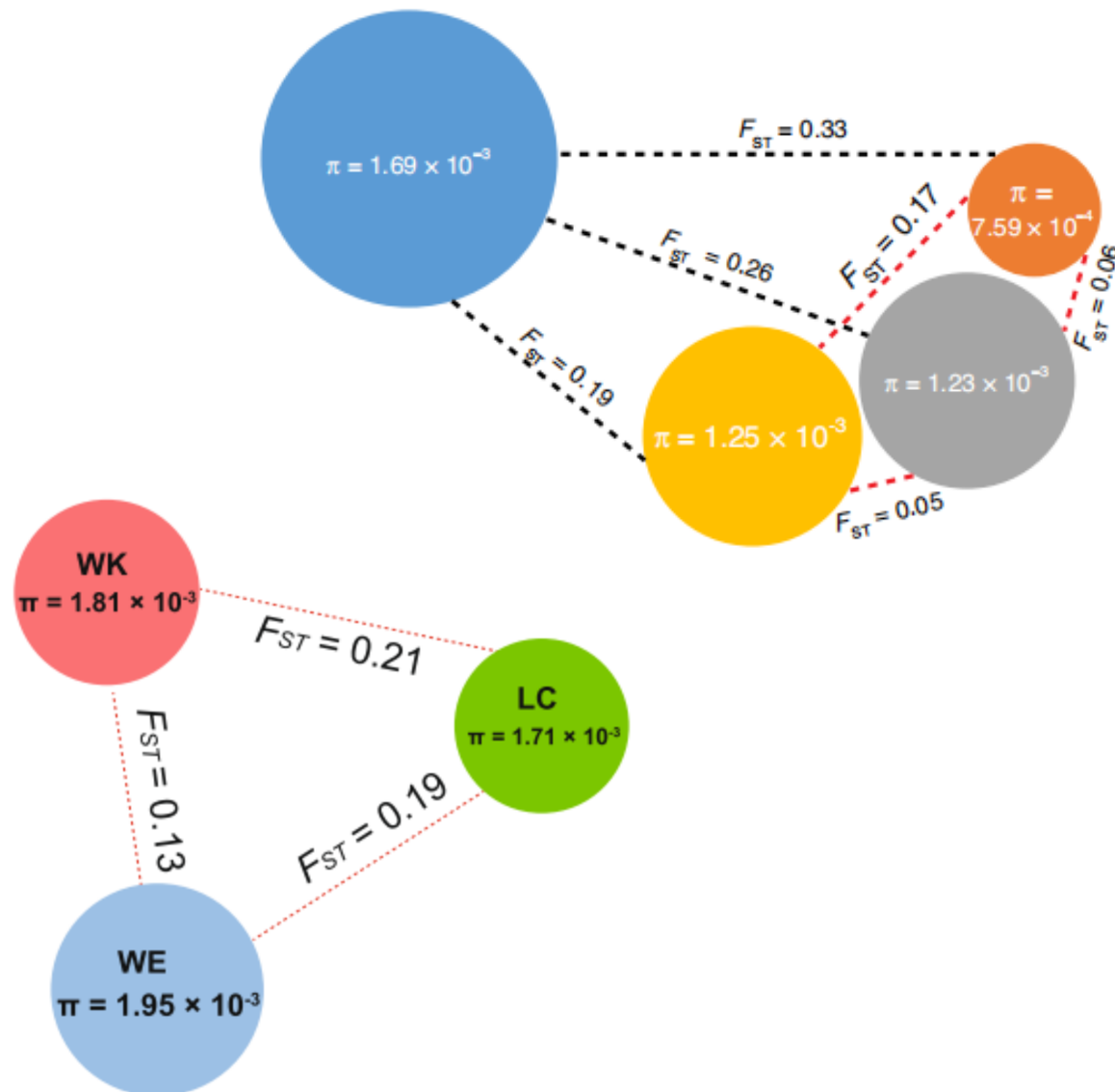
云南省生物大数据重点实验室

遗传统计量展示

Table 1 Genetic diversity and Tajima's *D* in barley groups

	Overlapped SNPs data				WGS SNPs data	
	Wild	Western	Eastern	Qingke	Western cultivars	Qingke
π	3.03	2.06	1.75	1.10	2.77	1.49
θ_W	5.39	2.07	1.57	1.00	2.79	1.47
H_E	2.85	1.94	1.52	0.99	2.51	1.36
ρ	2.85	0.23	0.07	0.04	0.32	0.06
Tajima's <i>D</i>	-1.16	0.40	1.23	0.86	0.39	0.53

Diversity (1×10^{-3}) is described by nucleotide diversity (π), Watterson's estimator (θ_W), gene diversity/heterozygosity (H_E), and recombination rate (ρ) and reported per bp. Source data are provided as a Source Data file



选择性清除分析



云南农业大学
Yunnan Agricultural University

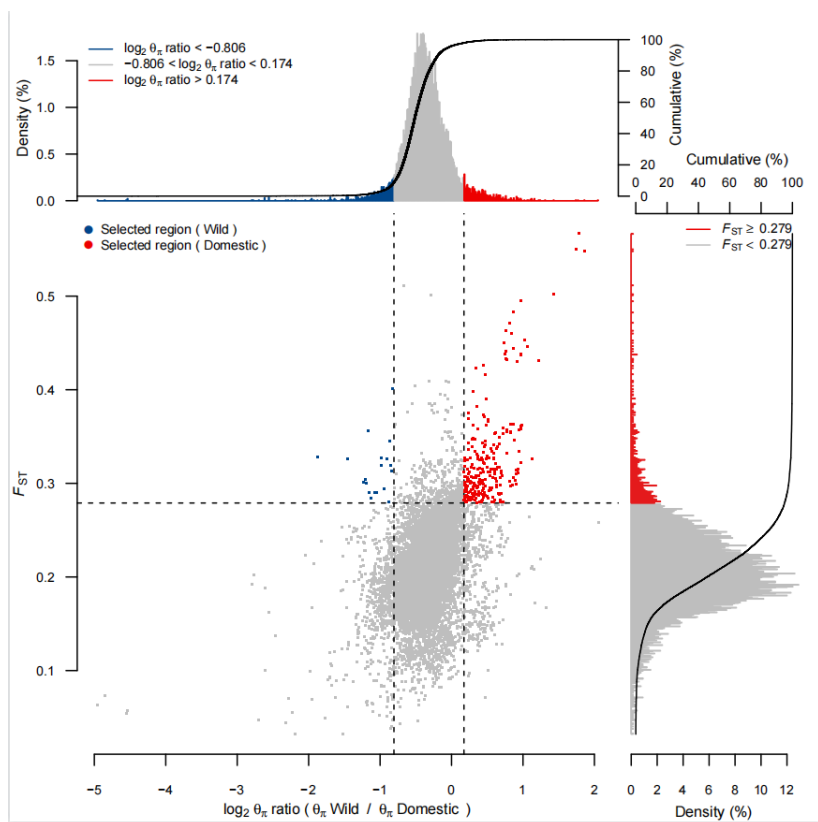


生信帮

云南省生物大数据重点实验室

选择性清除分析 (Selective Sweep Analysis) 是群体遗传学研究中的一种分析方法，用于检测在自然选择作用下导致某些基因或基因区域频率发生显著变化的信号。

由于在自然选择过程中，某些特定基因型或等位基因受到有利的选择压力，会导致其频率在群体中迅速增加。选择性清除分析就是从大量的遗传标记数据中检测这种选择信号，并确定哪些基因或基因区域经历了选择性清除事件。



如果一个基因受到了正选择，那基因组上会发生什么？

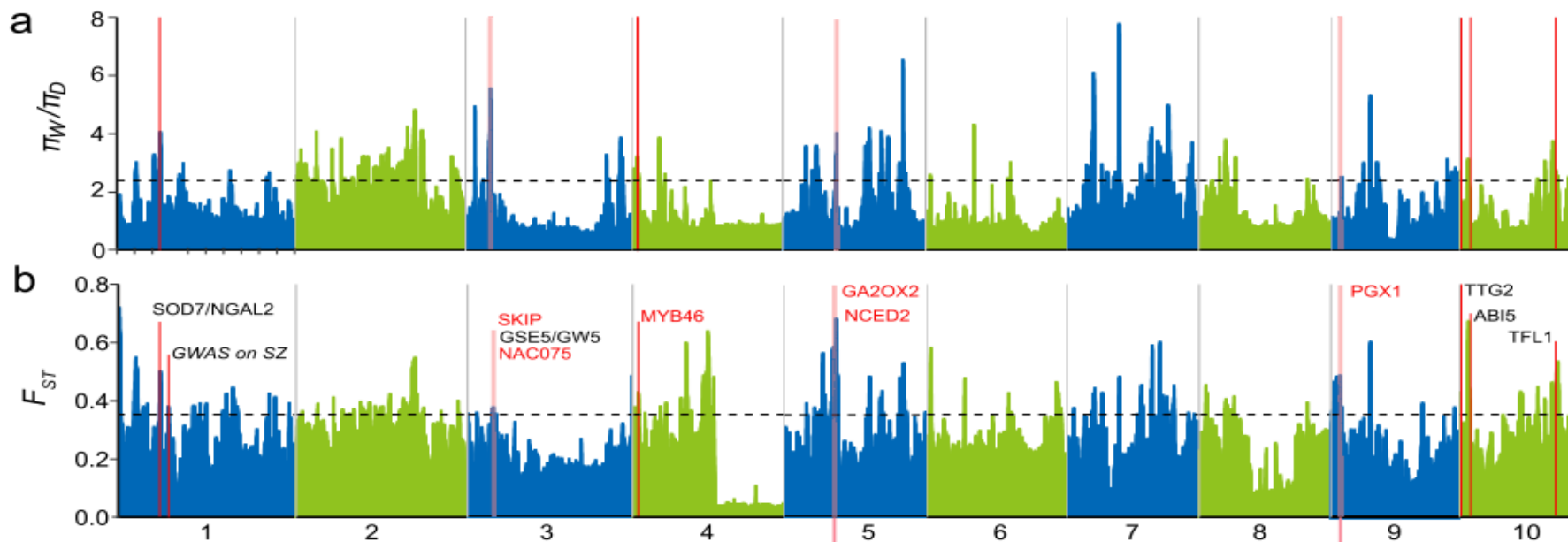
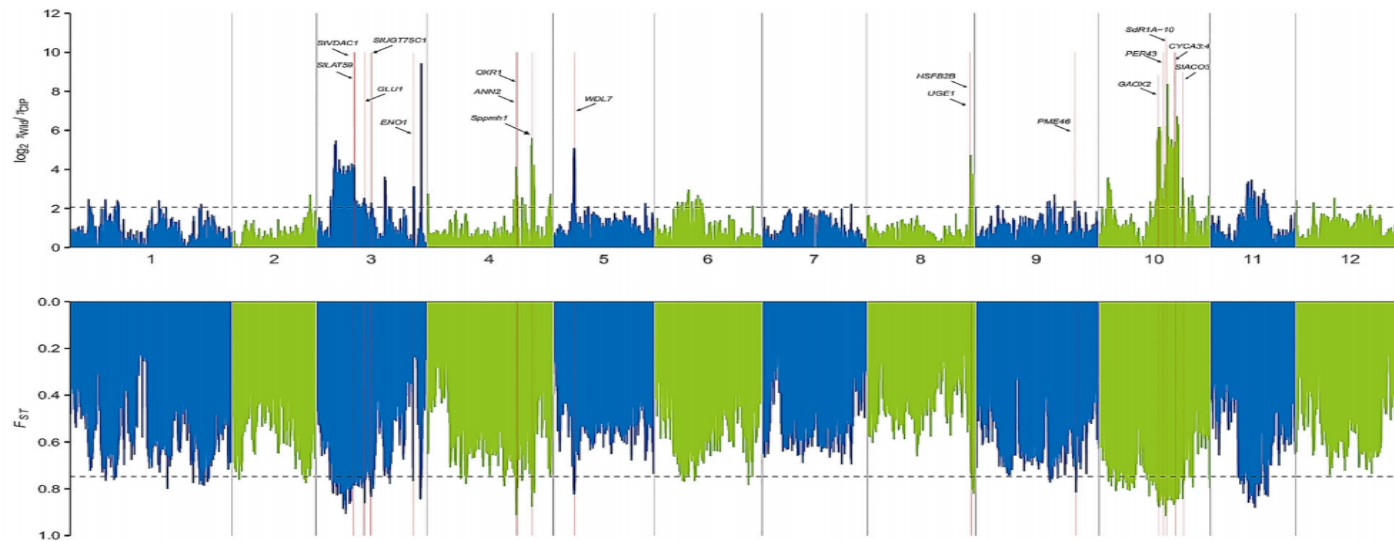
1. 相应等位基因的频率增加
2. 周围连锁不平衡增加，即连锁程度增加
3. 漂变增加，多样性减少
4. 与未受选择群体相较，同一个位点的差异增加（即遗传距离增加）

风险提示：简化数据由于snp密度太低，做该分析获得的受选择区域和基因可能不全面，即假阴性较高。

云南省生物大数据重点实验室



生信帮



连锁不平衡衰减 (LDdecay) 分析



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

连锁不平衡 (linkage disequilibrium, LD) 是指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不等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)。

连锁不平衡衰减 (LDdecay) 分析是一种用于研究遗传标记之间连锁不平衡的程度随着物理距离的增加而逐渐减弱的方法。它主要应用于分析基因组中的遗传连锁 (genetic linkage) 和连锁不平衡 (linkage disequilibrium, LD)。

通过LDdecay分析，可以获得遗传标记之间的连锁不平衡衰减模式，了解基因组中的遗传连锁和演化过程。

通常来说，受到自然驯化或人工选择的群体，其基因组上存在较强连锁的区域会变长，即连锁不平衡衰减的距离会变长，即连锁不平衡衰减的速度会变慢。

连锁不平衡衰减 (LDdecay) 分析



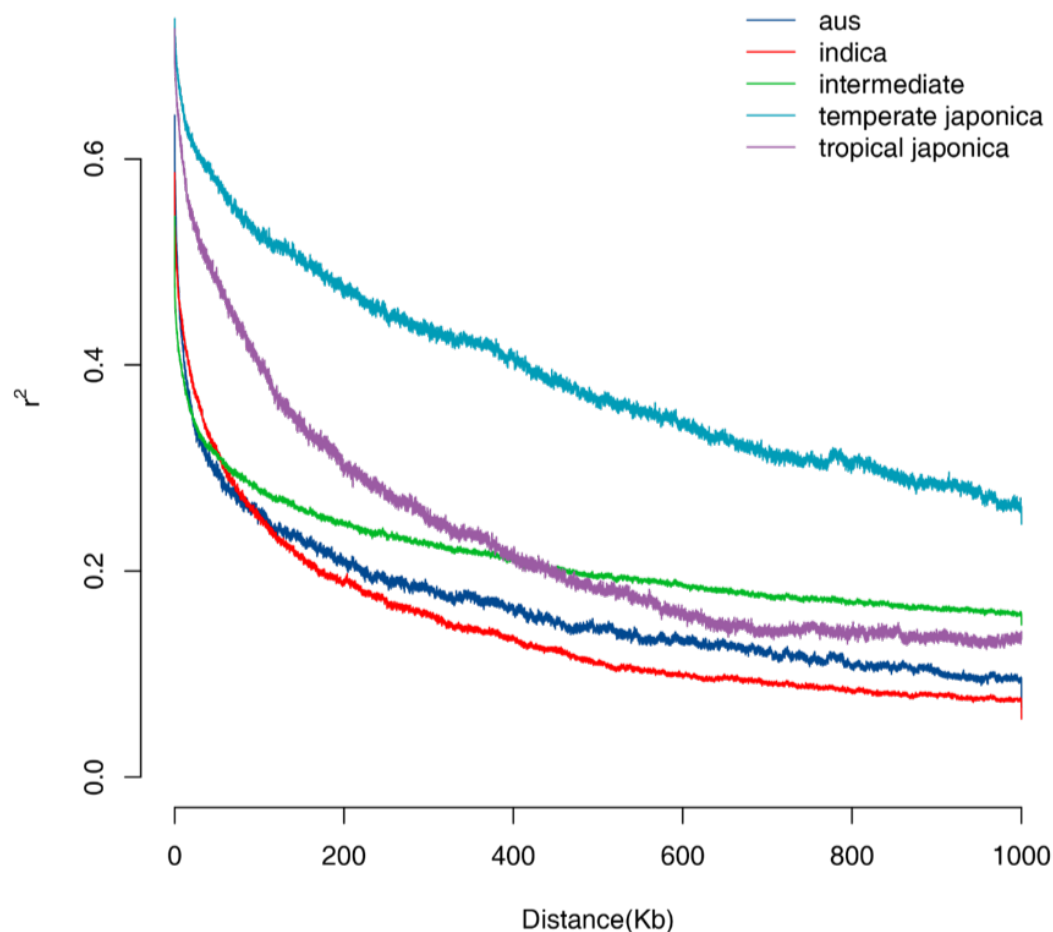
雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

LD decay



连锁程度高表示群体总体上基因组多样性较低；
一般来说，衰减速度最快的亚群更接近原始野生种；

不同亚群分析结果的比较：

- (1) 每个亚群下降到同一连锁强度时，对应的衰减距离；
- (2) 每个亚群下降到各自最高强度的一半时，所对应的衰减距离；

风险提示：简化数据做该分析得到的每个亚群的基因组LD衰减距离是不准确的，仅能定性地比较不同亚群LD衰减的快慢。

连锁不平衡衰减 (LDdecay) 分析



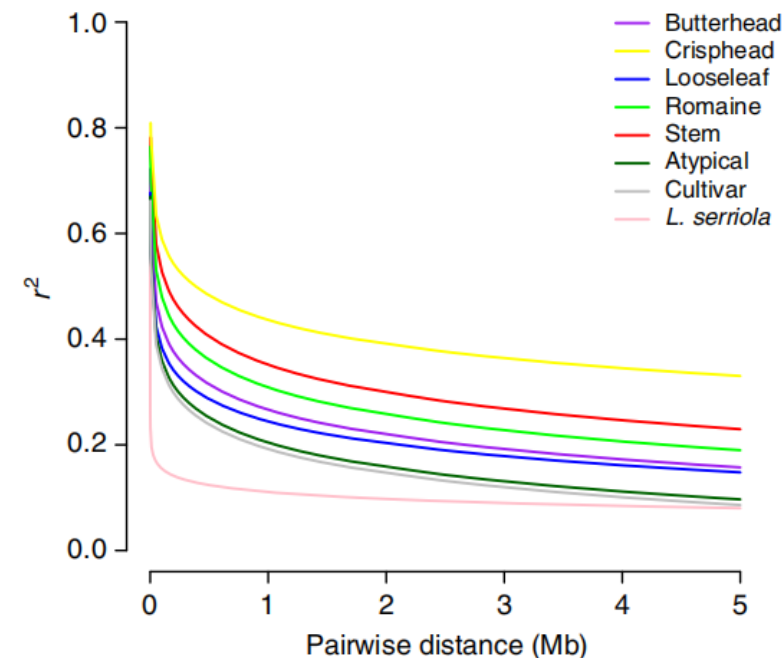
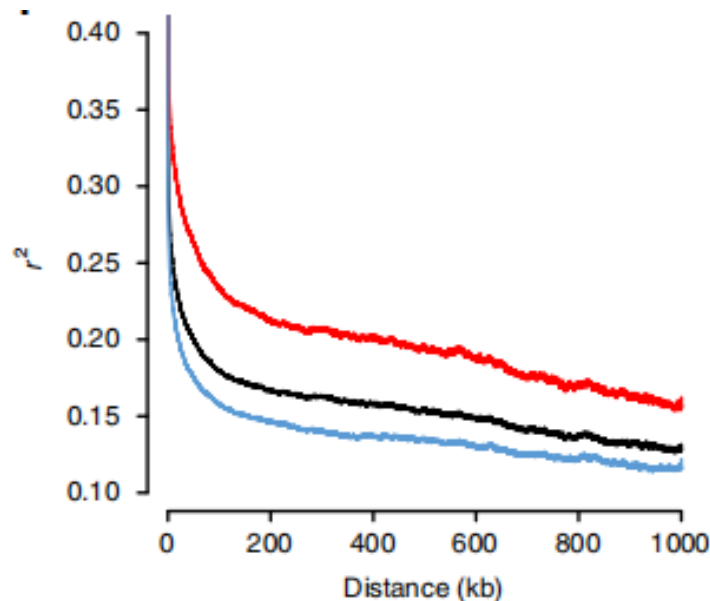
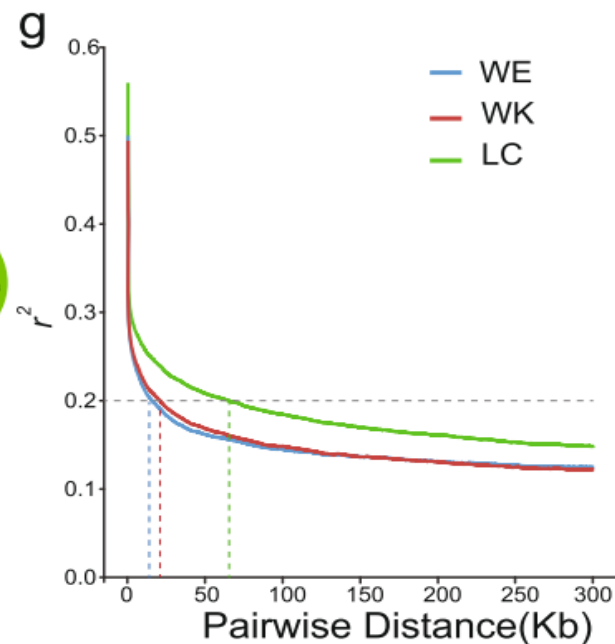
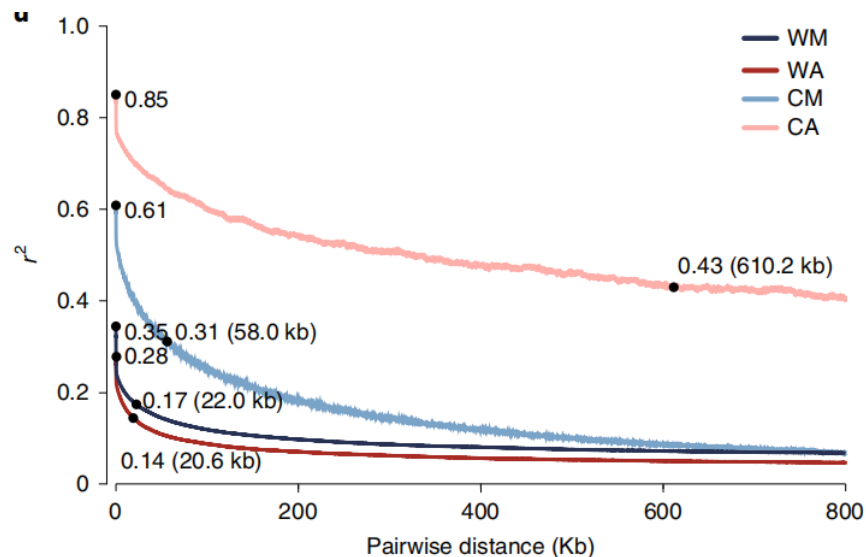
雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

可视化图展示





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，答你所问