# Pro a problem

version 0.0.2 How to understand Figure 10 A?

A

| Groups | Promoter | | | | | | | | | | | ORF | | | | | | | | | | | No. of accessions | | | | | | No. of cvs. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HG Hap | -1148 | -1139 | -1082 | -1060 | -1059 | -901 | -718 | -354 | -310 | -309 | -(214-216) | 36 | 72 | 194 | 740 | 771 | 844 | 932 | 1088 | 1162 | 1407 | 1550 | AUS | BAS | IND | WILD | TEJ | TRJ | |
| 1 | A | G | T | C | T | T | C | G | A | C | CCT | C | A | A | A | A | G | A | A | A | C | C | 4 | 52 | 205 | 147 | 226 | 313 | 947 |
| 2 | G | - | T | CG | A | C | A | A | - | G | --- | A | A | G | G | G | A | C | G | A | T | C | - | 1 | 697 | 48 | 1 | 2 | 749 |
| 3 | A | G | T | C | T | T | C | G | A | C | CCT | C | A | A | A | A | G | A | A | G | C | C | 5 | - | 1 | - | 34 | 104 | 144 |
| 4 | G | - | T | CG | A | C | A | A | - | G | CCT | A | A | G | G | G | A | C | G | A | T | C | - | - | 114 | - | - | 1 | 115 |
| 5 | G | - | T | CG | A | C | A | A | A | C | --- | A | A | G | G | G | A | C | G | A | T | C | 1 | - | 105 | - | - | - | 106 |
| 6 | G | - | T | CG | A | C | A | A | - | G | --- | A | A | G | G | G | A | C | G | A | C | T | 1 | - | 96 | 4 | - | - | 101 |
| 7 | G | G | C | CG | A | C | C | A | - | G | --- | A | G | G | G | G | A | C | G | A | C | C | 79 | 3 | - | 3 | - | - | 85 |
| 8 | G | - | T | CG | A | C | A | A | - | G | --- | C | A | G | G | G | A | C | G | A | T | C | - | - | 71 | 6 | 1 | - | 78 |

> The Hapoltype and evolutionary analysis of OsNCED3. (A) Haplotype analysis in 5152 rice germplasm accessions. Orange indicates the reference (Nipponbare) allele sequence, dark blue indicates the alternative allele sequence

> Haplotype

## 过程

> Based on variations in the OsNCED3 promoter and coding *sequences*, we identified 8 haplotypes with highest frequency variations across the OsNCED3 genome sequence (Figure 10A) 原文说我们依据的是测序数据，将它分为8个单倍型。那么很容易想到，过程是啥呢？且看下文

### Haplotype and nucleotide diversity analysis

**Haplotype and nucleotide diversity analysis**

The haplotypes of *OsNCED3* were determined using the genotype data of of 5152 rice accessions (Table S3 and S4) downloaded from 3K database (https://snp-seek.irri.org/). The phylogenetic tree of haplotypes was constructed in MEGA X using the neighbor-joining method with P-distance statistical model. The nucleotide diversity (π) for each rice group were calculated using VCFtools with a 10 kb window and a 5 kb step size (Danecek et al., 2011), respectively. 8 haplotypes with the highest frequency to construct a phylogenetic tree using IQ-TREE with the default parameters (Minh et a., 2020). iTOL was used to visualize the phylogenetic tree (Letunic et al., 2021).

1. snp-seek数据库下载数据看看 1.1 里面有Phenotyped Data，数据量很大， 1.2 原文的table s3 | table s4也看了下

> README 3K Rice Genomes Permissive license This SNP , indel, and large SV dataset is released using the permissive license stated in The Toronto Statement from the Toronto International Data Release Workshop Authors (Nature 461, 168-170(10 September 2009), doi:10.1038/461168a By downloading and using this dataset in part or in its entirety, you agree to abide to the spirit of the Toronto Statement.

2. 下载文件，用vcftools处理，需要筛选，
3. 处理后的文件，依据表型文件，依据the highest frequency .

比如这里的`references`的中例子:通过vcf文件对基因进行单倍型分析

通过例子，那么知道怎么来的，至于8个，就随它吧。

ps:它的species是7个，然而这个没关系

- O. barthii
- O.Glaberrima
- O. longistaminata
- O. nivara
- O. meridionalis
- O. rufipogon
- O. sativa

# references

> 3K水稻SNP数据集的简单利用

> Haplotype Analysis on RFGB

> download from snp-seek

> 基于Vcf文件进行基因单倍型分析

> 单倍型分析技术研究进展

> GWAS分析-说人话（20）-单倍体关联分析

> GWAS分析-说人话（14）- 如何查看SNP的基因型

# 基础概念

由于，西方科学建立在逻辑思维上面，逻辑思维基于各种概念。所以，趁这个机会，学习一下。

- A haplotype (haploid genotype) is a group of alleles in an organism that are inherited together from a single parent. (from wiki)
- 群体遗传结构（population genetic structure):指基因型在空间和时间上的分布形式，它包括群内的遗传遗传变异和种群的遗传分化。群体结构遗传是经过长时间的进化而形成的，很多物种的遗传结构反映了其进化历史中的特殊事件
- 群体结构中亚群的分类方法分为两种：基于距离和基于模型的方法。
    - 基于距离，通常会计算群体中每对个体的距离，以成对矩阵表示，并通过树形图或多维比例图呈现，但很难纳入其他信息（地理、表型信息）进行精确统计，人为确定亚群分类。
    - 基于模型的方法，通过假设每个亚群的观测值（如：基因频率）来自某个参数模型的随机抽样，使用统计（最大似然法或贝叶斯统计）对每个亚群成员和亚群模型参数进行推断，可以对群体进行精确聚类，检测群体内个体基因交流及混合程度。软件：STRUCTURE(Prithchard et al, 2000)
        - 可用FSTAT估计亚群分化固定系数、群体内分化系数（Fst)和群体遗传多样度(Hs)。对于大数据量的分析，可使用VCFtools、PopGenome等进行全基因组水平上群体分化相关参数计算。
        - 例如：水稻及其野化群体分析(引自Qiu et al, 2020)
        - Aus等就是这么分类的
- 自然选择的统计检验：

- 自然选择：：正选择（positive selection)、负选择(净化选择negative selection)和平衡选择(balacing selection)，平衡选择在多数情况下会倾向维持群体的遗传多样性
- 中性检验是判断群体进化一个方法：
  - 以中性进化学说作为零假设，通过统计检验的方法，检测一个群体的遗传参数是否符合中性进化模型。
  - 大体分为两类:基于种类多态性（intraspecific polymorphism)的检验方法、
- 基于位点变异频率分布
  - Tajima's D检验
  - Fu & Li's D和F检验
  - Fay & Wu's H检验
- 基于连锁不平衡
  - 在一段DNA序列中，**位点与位点之间存在连锁**的关系。不同位点间的连锁构成了单倍型。随着，重组的积累，特定的单倍型会被稍弱而逐渐消失。由于重组率与连锁距离有关，所以连锁不平衡范围会逐渐缩短。对于新产生的一个单倍型，由于重组来不及破坏位点之间的连锁，所以它们之间的连锁不平衡的区域往往会比较大。在中性假设下，如果某个单倍型是较新产生的，那么它的频率往往比较低，而频率较高的单倍型，需要经历很长一段时间才可能因为受到遗传漂变的影响达到较高的频率。如果群体经历了正向选择，那么有利于位点连锁的周围位点会由于搭载效应频率很快提升。所以，包含有利位点的单倍体型，一方面有着较高的频率，另一方面由于经历的时间不长,存在较大的连锁不平衡范围。基于这个特征可以用来检测群体内是否发生了正向选择。
    - LRH 检验
    - HS检验
    - LDD检验
    - IBD检验
- 基于群体分化
- 基于溯祖树

# 单倍型估计方法(Haplotype_estimation)

按照 Grant 等(1998)提出的标准，单倍型多样性以 0.5 为临界值，核苷酸多样性以 0.005 为临界值，二者的值越大，群体的多样性程度越高。

已经提出了许多统计方法来估计单倍型。一些最早的方法使用简单的多项式模型，其中与样本一致的每个可能的单倍型都被赋予一个未知的频率参数，并且这些参数是用期望最大化算法估计的。这些方法一次只能处理少量站点，尽管后来开发了顺序版本，特别是 SNPHAP 方法。

最准确和广泛使用的单倍型估计方法利用某种形式的隐马尔可夫模型(HMM) 进行推理。长期以来，PHASE是最准确的方法。PHASE 是第一个利用聚结理论中关于单倍型联合分布的想法的方法。该方法使用Gibbs 抽样方法，其中每个个体的单倍型都根据来自所有其他样本的单倍型的当前估计值进行更新。以一组其他单倍型为条件的单倍型分布的近似值被用于吉布斯采样器的

条件分布。PHASE 用于估计来自HapMap 项目的单倍型. PHASE 受到速度的限制，不适用于全基因组关联研究的数据集。

fastPHASE 和 BEAGLE 方法引入了适用于GWAS大小的数据集的单倍型集群模型。随后引入了与 PHASE 方法相似但速度更快的 IMPUTE2 和 MaCH 方法。这些方法迭代地更新每个样本的单倍型估计，条件是其他样本的 K 个单倍型估计的子集。IMPUTE2 引入了仔细选择单倍型子集以提高准确性的想法。精度随 K 而增加，但随二次方O(K^{2})计算复杂度。

SHAPEIT1 方法通过引入线性O(K)仅在与个体基因型一致的单倍型空间上运行的复杂性方法。HAPI-UR 方法随后提出了一种非常相似的方法。SHAPEIT2 结合了 SHAPEIT1 和 IMPUTE2 的最佳特性，以提高效率和准确性。

- 间接推断法根据研究对象的不同又可分为两类：
  - 群体推断法和家族推断法。
  - 群体推断法是通过构建一些关联群体的基因池并用统计学方法对预测结果进行分析推断样本的单倍型。如果群体中存在一些突变频率较低的个体，它受连锁不平衡程度的影响往往会被遗漏而无法获得其单倍型信息。
  - 家族推断法是根据同一家族众多个体的基因型信息对待测样本进行推断获得其单倍型信息，在使用前要确保同一家族中这些样本基因型信息的可靠性。家族推断法在遗传疾病的研究中有非常重要的作用。研究者对一个家庭的父母及其子女四口人进行全基因组测序，经过序列分析可以得知子代基因组精确的重组位点和一些稀有的单核苷酸变异位点。更重要的是，发现该家庭两子女含有米勒综合症和原发性纤毛运动障碍性疾病两个隐性致病基因，对寻找致病基因和疾病治疗方法有重要作用。

# 实战

我还没做，

> 3K水稻SNP数据集的简单利用

> 通过vcf文件对基因进行单倍型分析

# 引用

> Haplotype

> Haplotype_estimation

> 单倍型组装与推断

> 杨少滢,缪立生,肖成,刘宇,沈宏旭,吴健,高青山,赵玉民,曹阳.沃金黑牛FGF14基因单倍型与生长性状的相关性分析[J/OL].中国畜牧杂志:1-10[2022-11-13].DOI:10.19556/j.0258-7033.20220630-02.

> 单倍型分析技术研究进展

> Grant WAS, Bowen BW. Shallow population histories in deep evolutionary lineages of marine fishes: Insights from sardines and anchovies and lessons for conservation. Journal of Heredity, 1998, 89(5): 415－426

> haploview 单倍型分析

> Haploview软件使用-连锁不平衡分析

> 【求助】单体型(haplotype)分析