work shop 2019

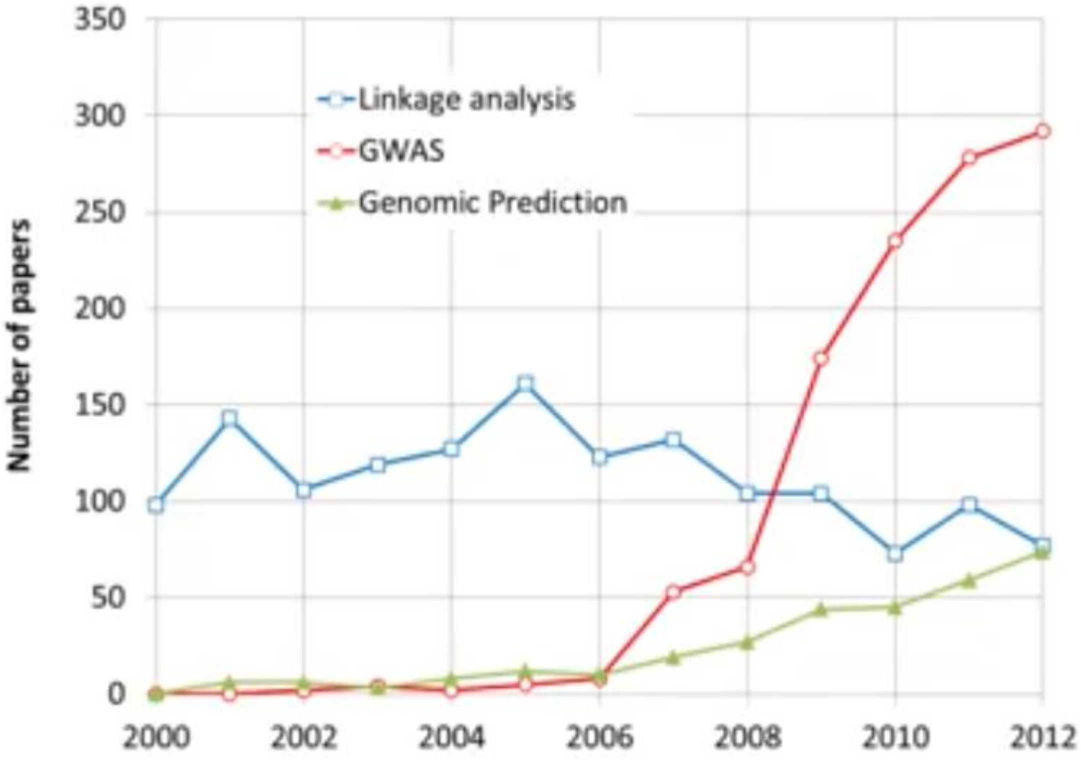
GWAS Introduction

contents

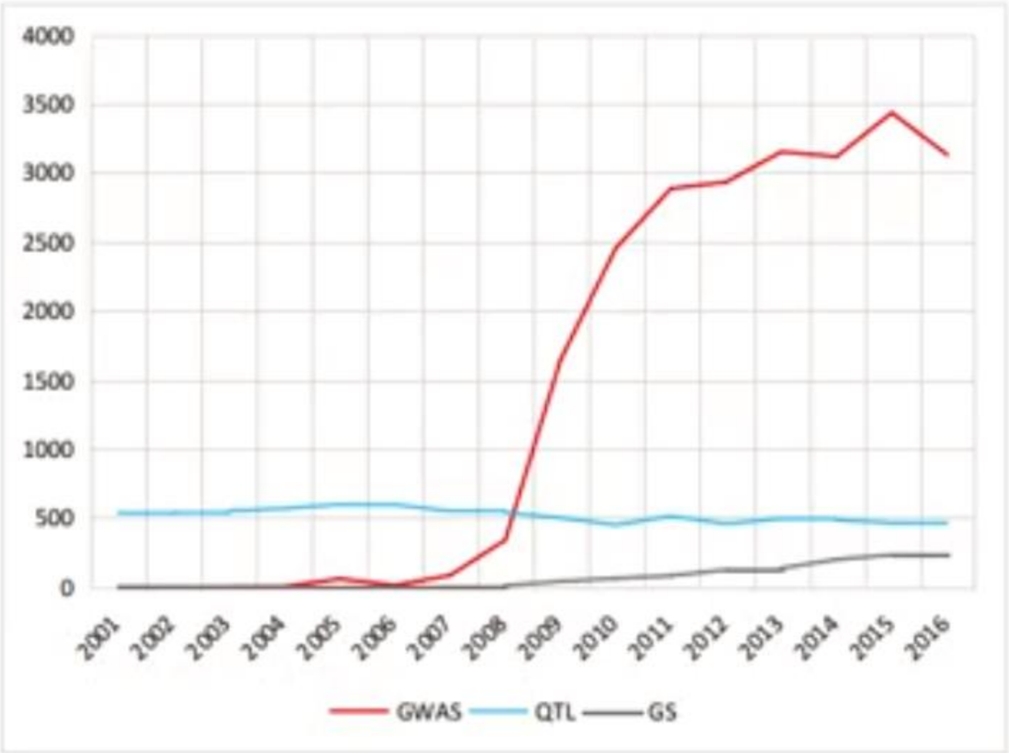
01 Introduction to genome-wide association analysis

02 Introduction of the principle

03 Population structure and its relationship

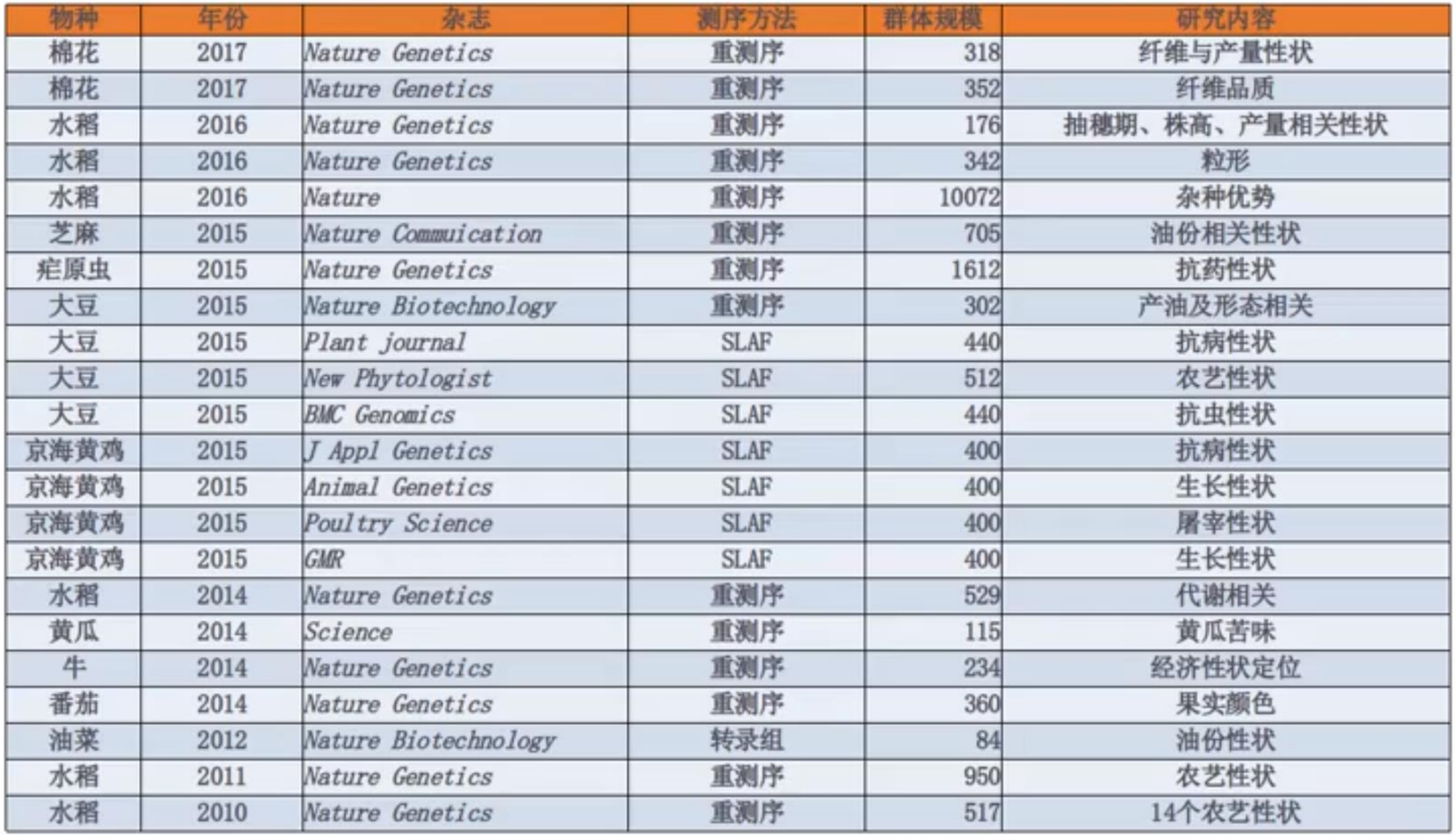


FROM ZHANG'S PPT 201607 WUHAN



PubMed search result
GWAS："GWAS" or "Genome wide association"
QTL："QTL mapping" or "Linkage analysis"
GS："Genomic selection" or "Genomic prediction"

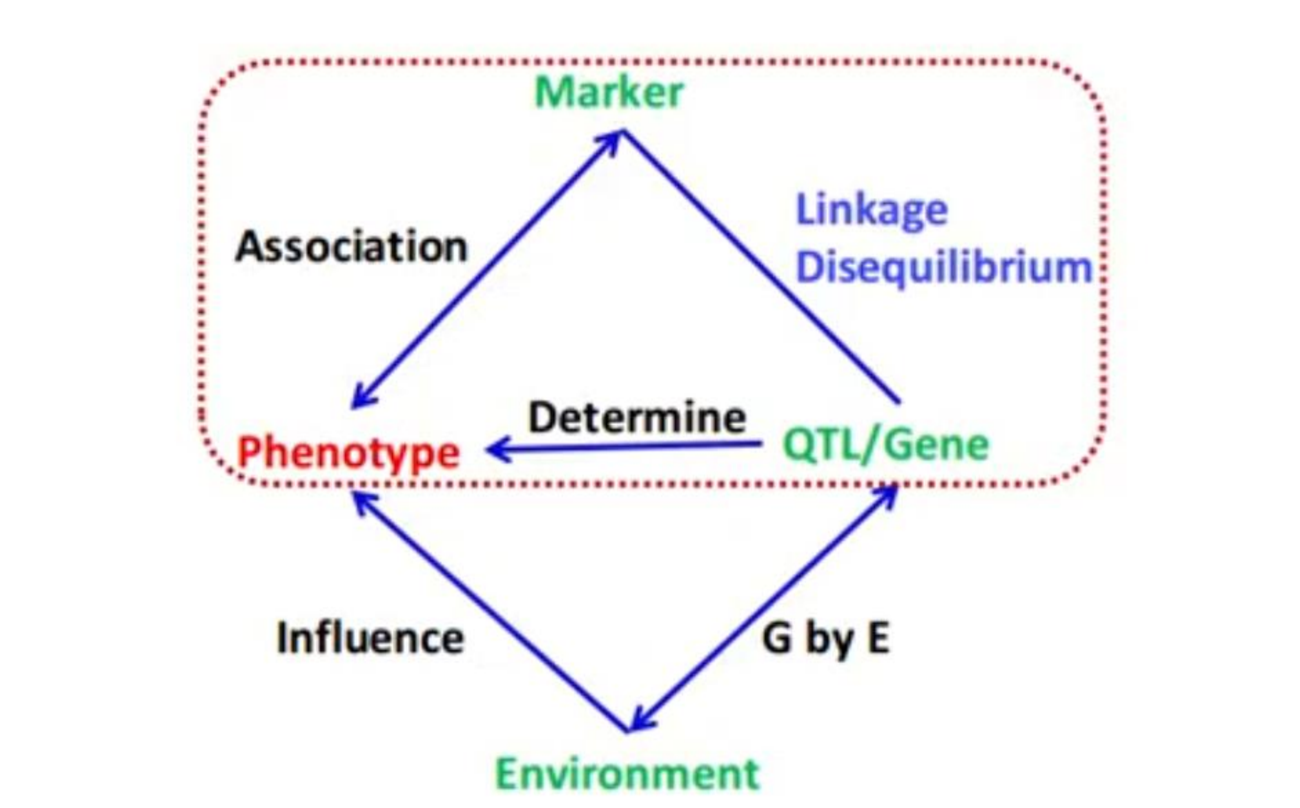
Some Gwas articles in recent years



物种	年份	杂志	测序方法	群体规模	研究内容
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、产量相关性状
水稻	2016	*Nature Genetics*	重测序	342	粒形
水稻	2016	*Nature*	重测序	10072	杂种优势
芝麻	2015	*Nature Commuication*	重测序	705	油份相关性状
疟原虫	2015	*Nature Genetics*	重测序	1612	抗药性状
大豆	2015	*Nature Biotechnology*	重测序	302	产油及形态相关
大豆	2015	*Plant journal*	SLAF	440	抗病性状
大豆	2015	*New Phytologist*	SLAF	512	农艺性状
大豆	2015	*BMC Genomics*	SLAF	440	抗虫性状
京海黄鸡	2015	*J Appl Genetics*	SLAF	400	抗病性状
京海黄鸡	2015	*Animal Genetics*	SLAF	400	生长性状
京海黄鸡	2015	*Poultry Science*	SLAF	400	屠宰性状
京海黄鸡	2015	*GMR*	SLAF	400	生长性状
水稻	2014	*Nature Genetics*	重测序	529	代谢相关
黄瓜	2014	*Science*	重测序	115	黄瓜苦味
牛	2014	*Nature Genetics*	重测序	234	经济性状定位
番茄	2014	*Nature Genetics*	重测序	360	果实颜色
油菜	2012	*Nature Biotechnology*	转录组	84	油份性状
水稻	2011	*Nature Genetics*	重测序	950	农艺性状
水稻	2010	*Nature Genetics*	重测序	517	14个农艺性状

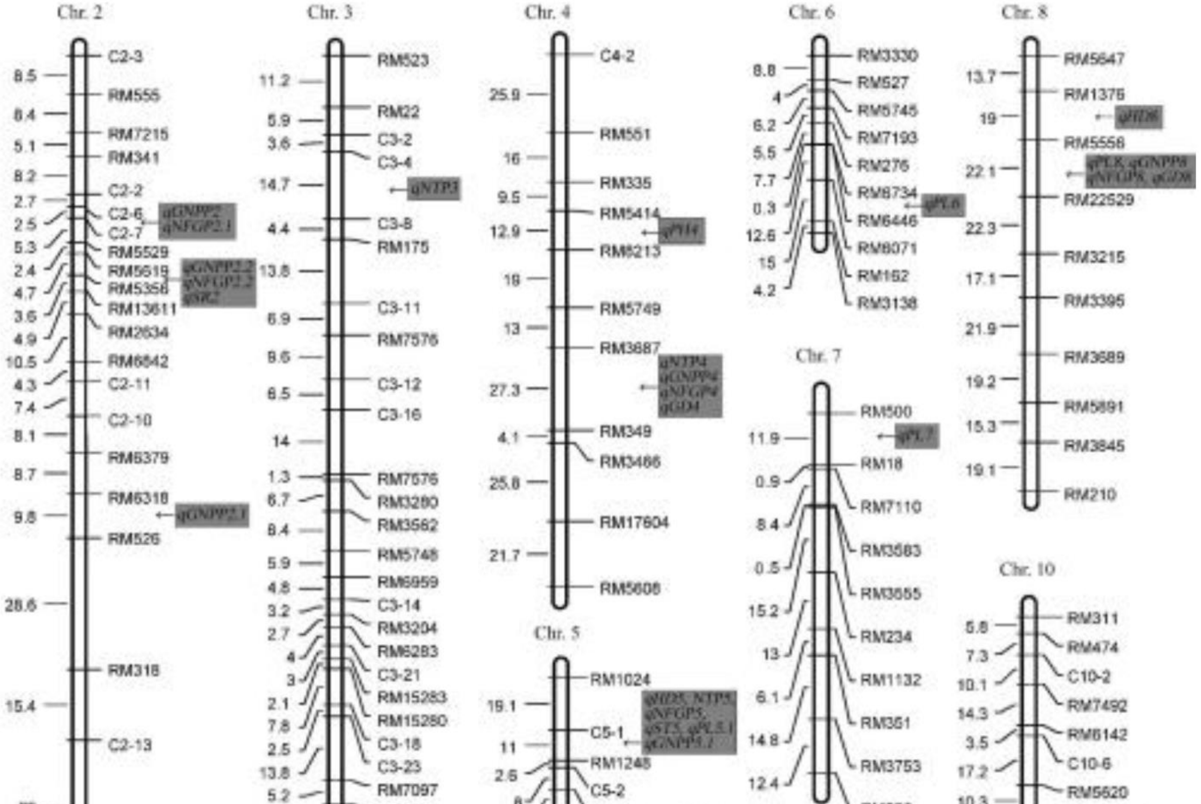
Genome wide association

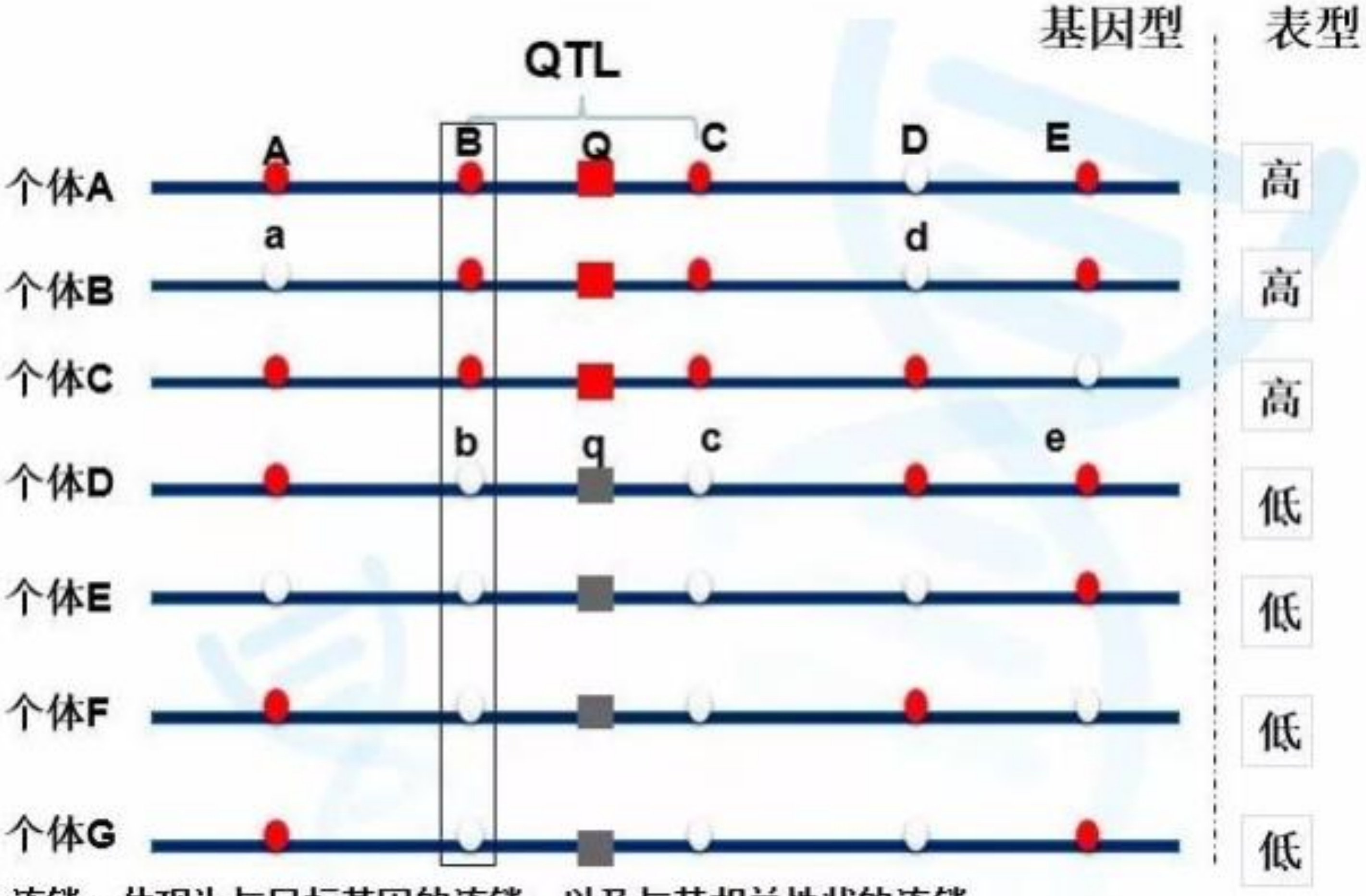
Genome wide association study (GWAS) is a genome-wide genetic variation (marker) polymorphism in multiple individuals to obtain genotypes, and then genotypes and observable traits, ie phenotypes For statistical analysis at the population level, the genetic variation (marker) most likely to affect the trait is screened based on statistics or significant p-values, and genes associated with trait variation are mined.



QTL positioning principle

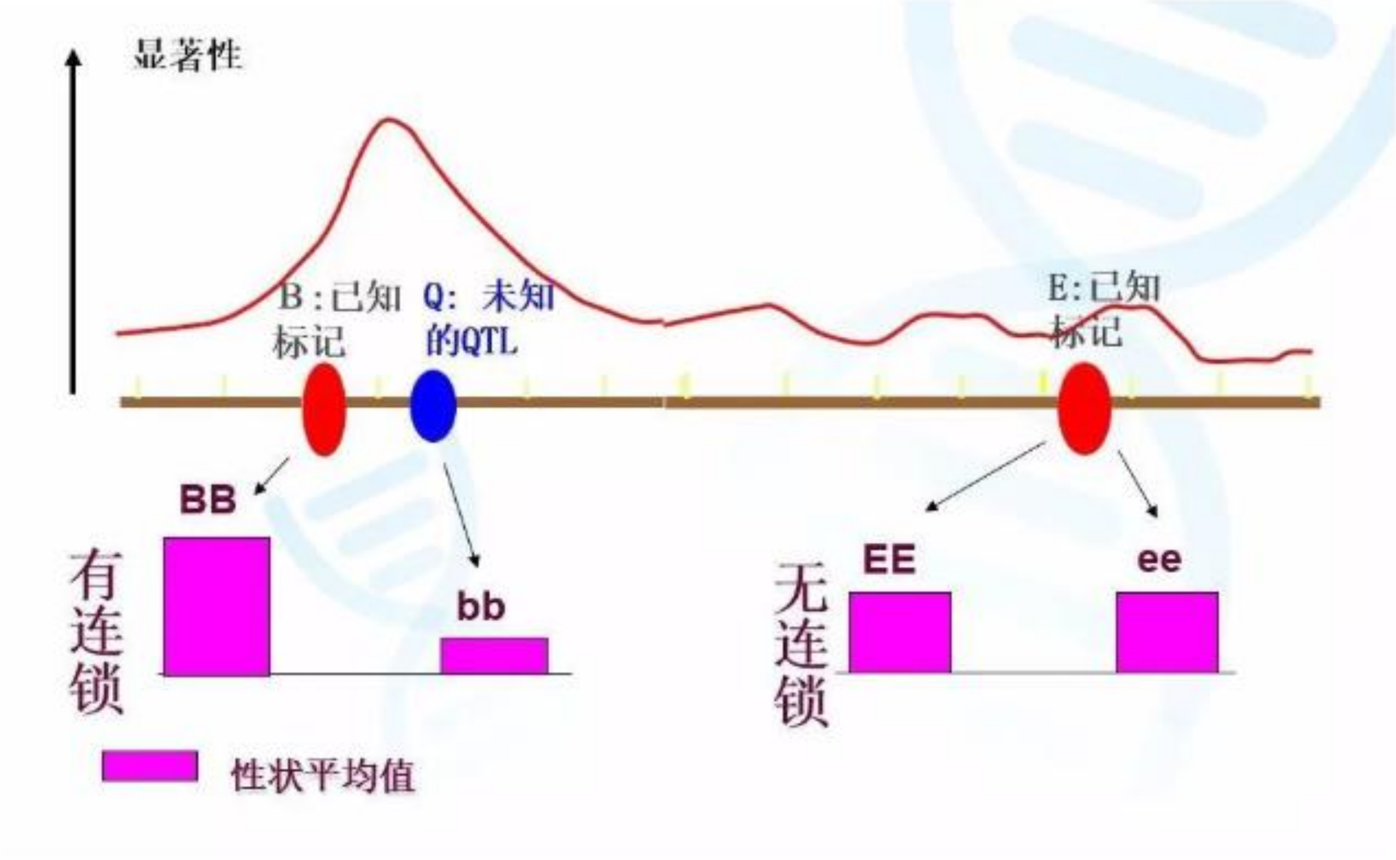
- The linkage analysis, which is called "linkage analysis", is based on the linkage and recombination between functional genes and molecular markers to achieve the location of functional genes.





连锁：体现为与目标基因的连锁，以及与其相关性状的连锁

Single-label analysis using analysis of variance



height= u+A*GT_A+B*GT_B+C*GT_C+D*GT_D+ E*GT_E

- u is the population mean (that is, the intercept of the equation), coefficient A is the genetic effect of the A locus, GT_A is the genotype of the Aa locus, which may be aa, Aa, AA, of course, 0,1 can be used mathematically. 2 replacement. Among them, the coefficients A, B, C, D, E are all variables to be solved。
- If we solve this multiple linear system of equations, we will find that A, D, and E are all 0 (effect is 0), while B and C are significantly greater than 0, then the Bb and Cc loci are inferred to contribute to height. So why do they contribute to height? Because they are linked to functional genes, we know the initial location of functional genes. This is the linear regression model in QTL positioning.

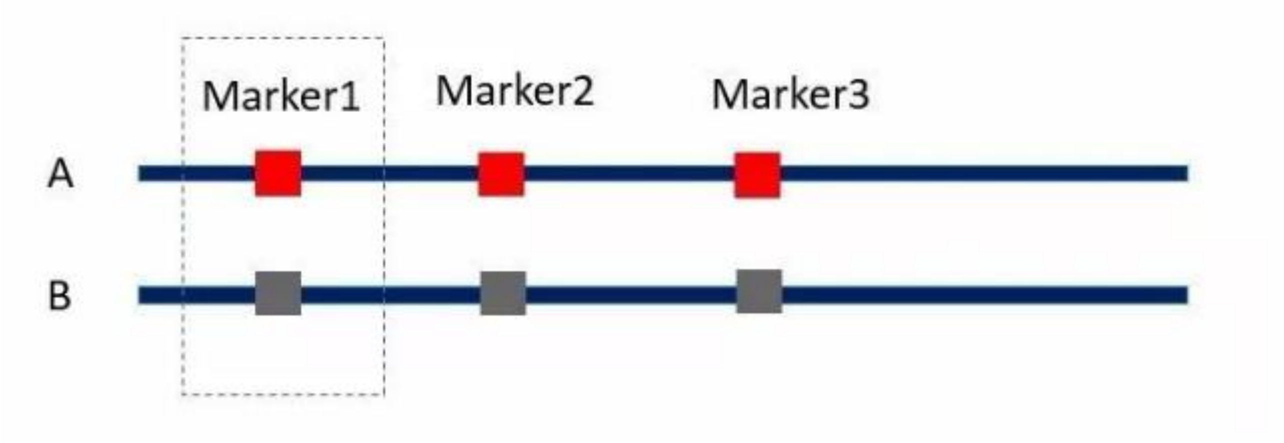
Simple linear regression model

In the actual case, the number of independent variables (number of markers) may be greater than the dependent variable (number of samples), so this equation is not accurate enough to obtain a unique solution. Therefore, multiple linear regression equations are usually reduced to one-dimensional linear regression equations. For example, for the Aa locus, we can construct a system of equations as follows:

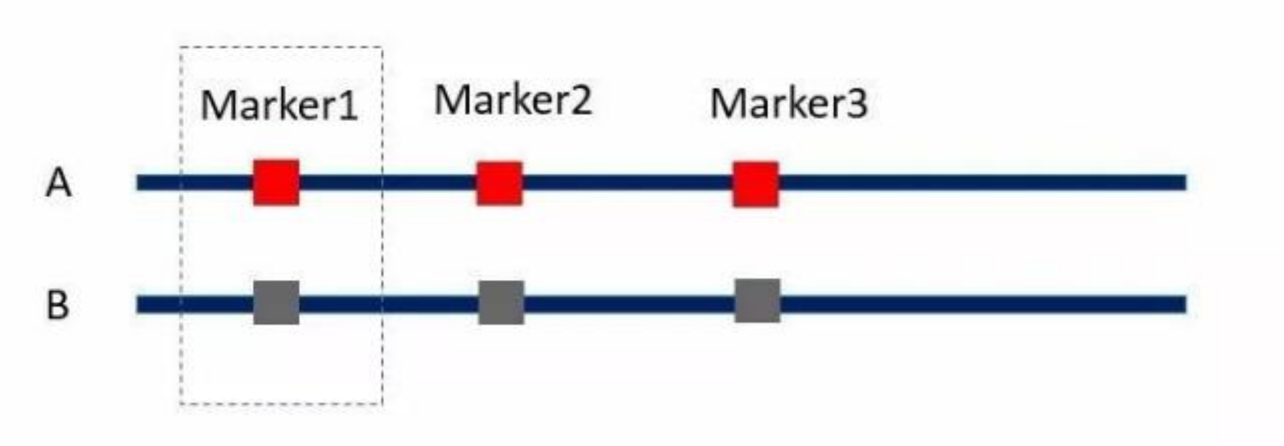
height = u+A*GT_A+e

The most widely used linear regression model

For example, in the figure below, individuals A and B have differences in three QTL loci. It is assumed that the red genotype can increase the height of the individual by 10 cm compared to the brown genotype. Now I want to calculate the effect of Marker1. If we only consider the effect of a single marker Marker1 (using Equation 2), the result of our calculation is that the height advantage of A 30 cm is derived from the difference of Marker1, and the effect meter of Marker1 is mistaken. It is 30 cm (overpriced).



But if we use multiple linear regression analysis, and combine Marker2 and Marker3 into the equations, and consider their effects in the equations, then the estimation of the Marker1 effect will be more accurate (the three marker effects are 10 cm).



However, the current high-density genetic map has hundreds or thousands of markers. As mentioned above, if each marker effect is incorporated into the equation, this equation can not be solved using the standard method (Equation 1). Therefore, in the classic composite interval mapping, a compromise is adopted. The general steps are as follows:

a) Screening several (eg, 10) most potent markers from the entire genome using single-labeled regression and stepwise regression.

b) When calculating a marker (interval) effect, integrate those markers with the strongest regional effects into the equations, such as the following equation:

$$\text{height} = u+A*GT\_A+[\,B*GT\_B+\ldots\;\ldots+K*GT\_K]+e$$

height = u+A*GT_A+[B*GT_B+… …+ K*GT_K]+ e

A is target mark
B~K is the most powerful marker in other regions of the genome.

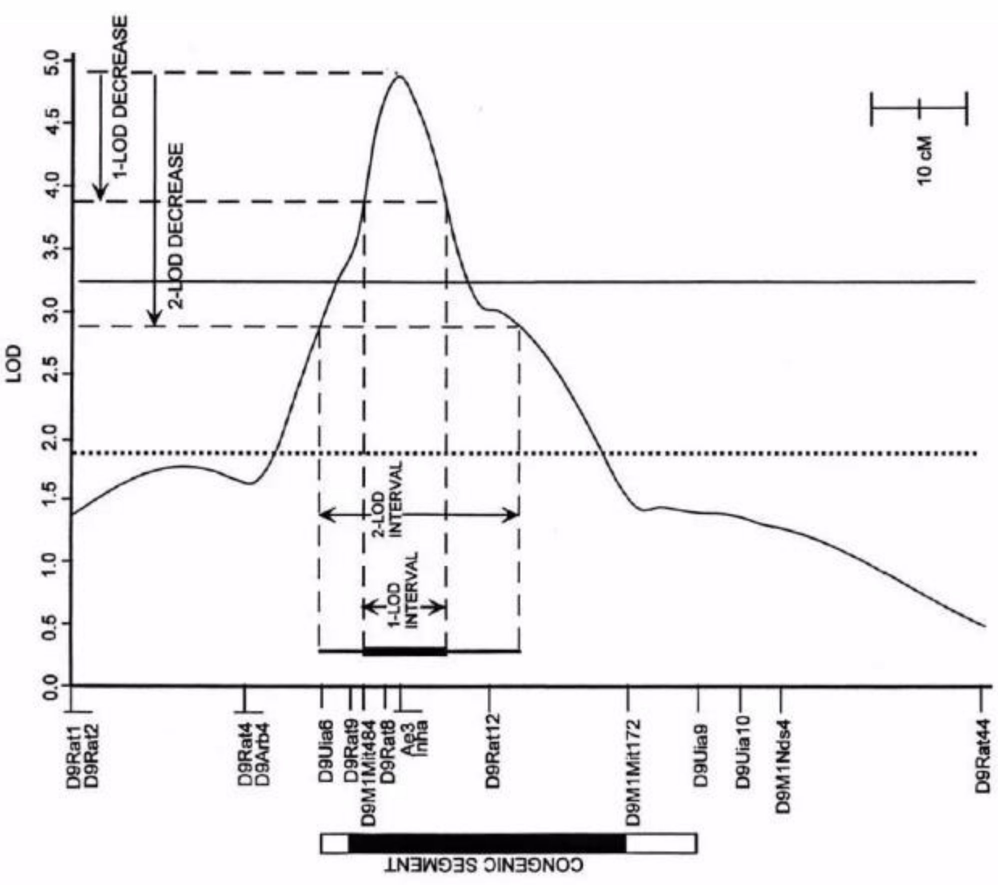
In the equation, there are 11 unknown variables (A~K-labeled effects), which can be solved as long as the individual is sufficiently large. The target mark is A (we expect to calculate their effects). B~K is the most powerful marker in other regions of the genome. Although we don't care about their specific effects for the time being, introducing them into the equation will make us estimate the effect of A more accurately. We mark B~K as not a direct concern, but like the independent variable (A mark), the same mark that affects the dependent variable (height) is called a covariant.

LOD vaule

- A calculation of genetic linkage, defined as the 10-based logarithm (lg) of the ratio of the likelihood data for a linked gene to the likelihood data for a non-linked gene. It is generally assumed that the LOD value of the gene linkage should be 3.0, which is a ratio of 1000:1.
- LOD=log10(L1/L0), where L1 is the probability that this site has a QTL, and L0 is the probability that this site has no QTL. If LOD=3, it means that the probability of this site having QLT is 1000 times that of QTL-free.

QTL positioning result diagram

2-LOD Confidence Interval: The result of QTL mapping is a waveform of a LOD value that changes on the chromosome (as shown below). The LOD value of the QTL region forms a signal peak. The functional gene is theoretically located near the peak of the strongest signal (the largest LOD value). But functional genes are usually only located in this interval, not necessarily at the peak. The farther away from the peak tip distance, the lower the LOD value and the lower the probability that the functional gene is located at that position.



Linkage Disequilibrium

- Linkage Disequilibrium (LD) is a non-random association between different loci within a population, including non-random associations between two markers or between two genes/QTLs or between a gene/QTL and a marker locus.
- It refers to the probability that alleles belonging to two or more gene loci appear on one chromosome at the same time, which is higher than the frequency of random occurrence. Simply , as long as the two genes are not completely independently inherited, they will show some degree of linkage. This situation is called linkage disequilibrium. The linkage disequilibrium can be different regions on the same chromosome or on different chromosomes.

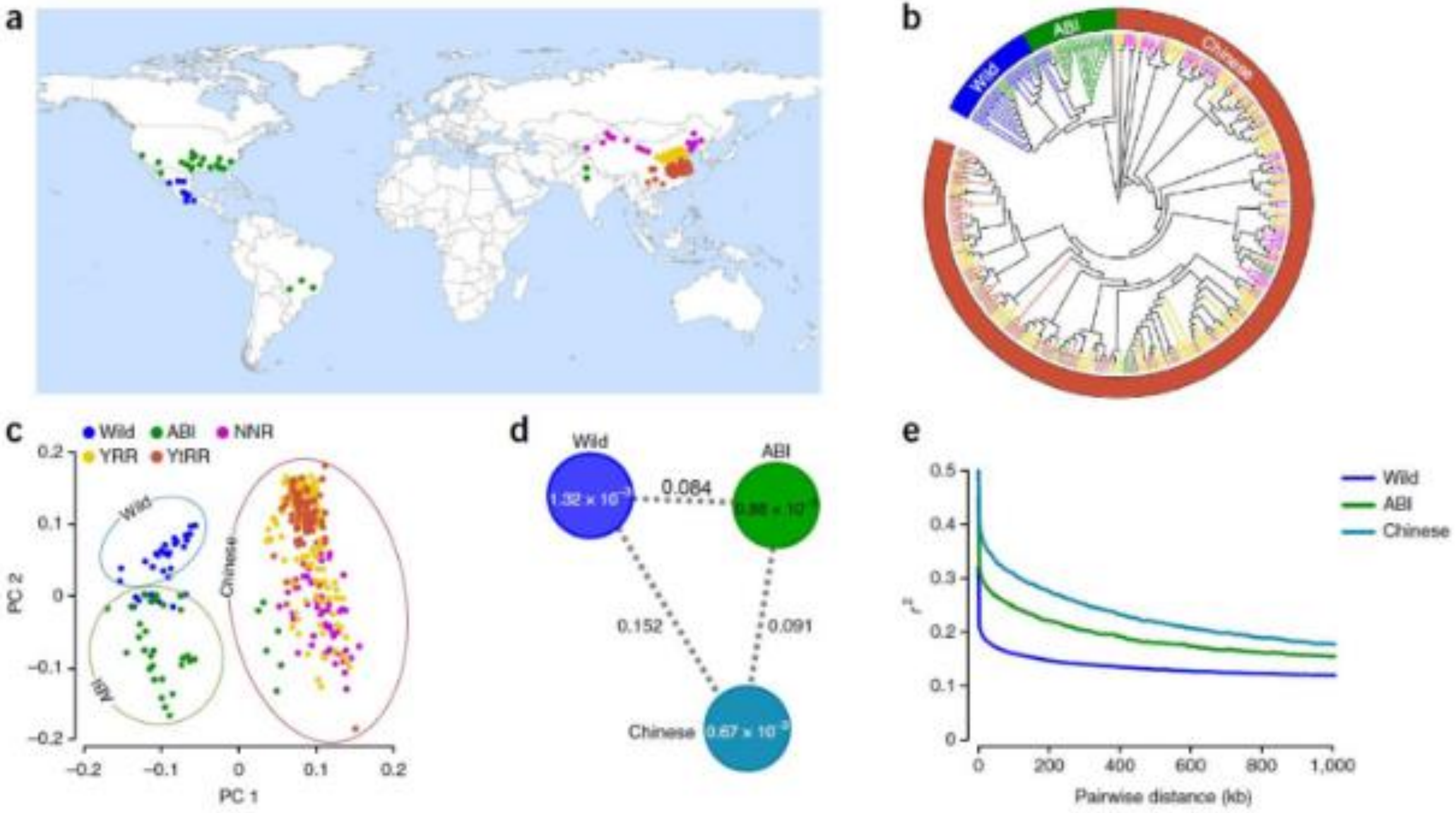
LD counts the difference between the actually observed haplotype frequency and the expected frequency of the haplotype at random separation. Usually, we use the formula :

$$D_{ab} = (\pi_{AB} - \pi_{A}\pi_{B})$$

- For example, two adjacent genes A and B, their respective alleles are a and b. Assuming AB is independent of each other, the probability of P(AB) appearing in the haploid genotype AB observed in the progeny population is P (A) * P(B)
- The probability of simultaneous emergence of the haploid genotype AB in the population was P(AB). If the two pairs of alleles are non-randomly bound, then P(AB)≠P(A)*P(B). The way to calculate this imbalance is:

D = P（AB）－ P(A) * P（B）

Therefore, four haplotypes AB, aB, Ab, and ab may be formed.


a
b
Wild
ABI
Chinese
c
Wild
ABI
NNR
YRR
YtRR
Wild
ABI
Chinese
0.2
0.1
0
−0.1
−0.2
PC 2
−0.2
−0.1
0
0.1
0.2
PC 1
d
Wild
ABI
1.32 × 10^{-3}
0.084
0.88 × 10^{-3}
0.152
0.091
Chinese
0.67 × 10^{-3}
e
Wild
ABI
Chinese
0.5
0.4
0.3
0.2
0.1
0
r^2
0
200
400
600
800
1,000
Pairwise distance (kb)

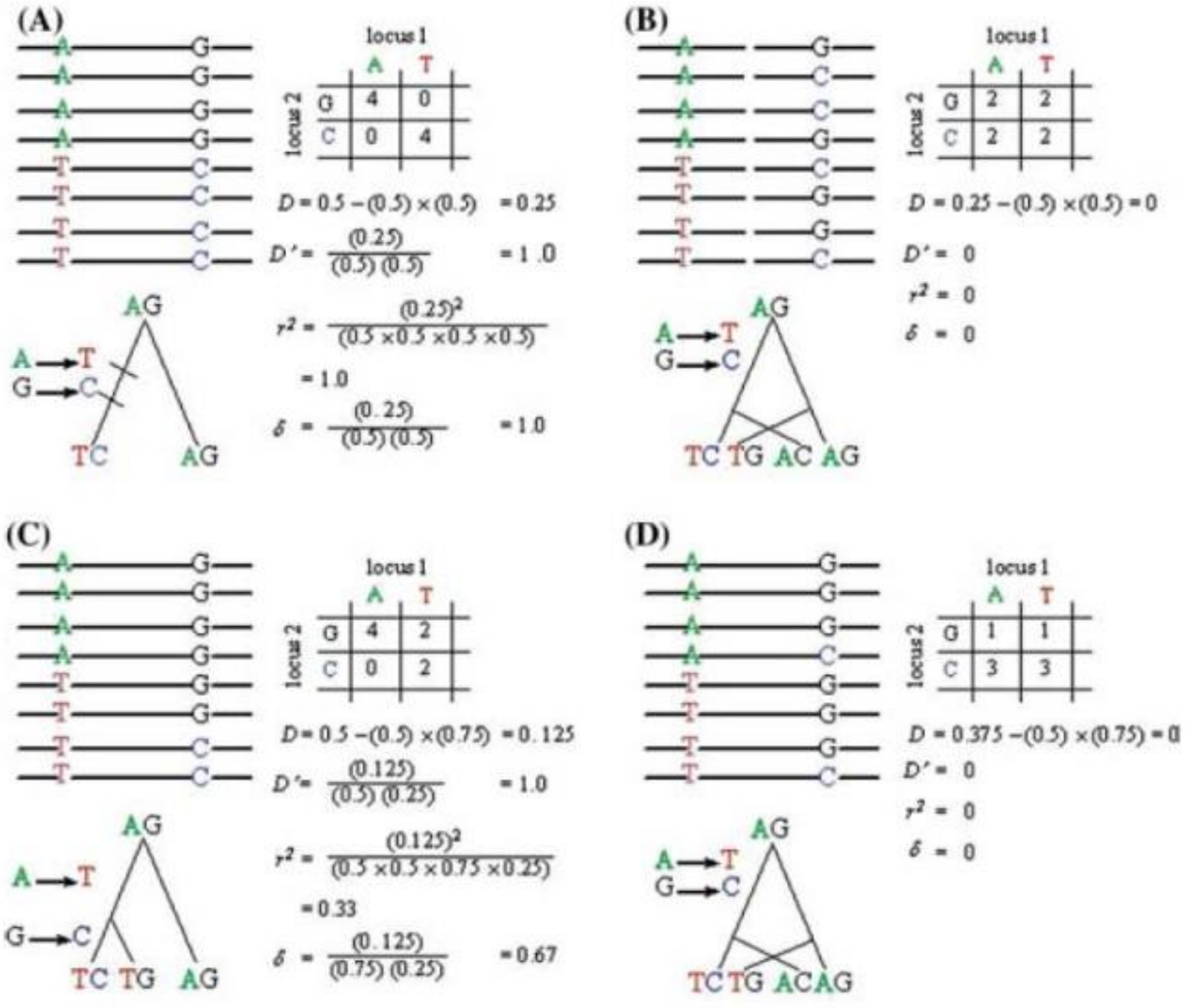
However, for a locus with only two alleles, such as a SNP, r2 and D' are usually used to measure the LD level between the two loci.

$$|D'| = \frac{(D_{ab})^2}{\min(\pi_A \pi_b, \pi_a \pi_B)} \text{ for } D_{ab} < 0$$

$$|D'| = \frac{(D_{ab})^2}{\min(\pi_A \pi_B, \pi_a \pi_b)} \text{ for } D_{ab} > 0$$

$$r^2 = \frac{(D_{ab})^2}{\pi_A \pi_a \pi_B \pi_b}$$

R2 and D' reflect different aspects of LD. R2 includes recombination and mutation, while D' only includes a history of recombination. D' can estimate the difference in recombination more accurately, but the probability of a combination of low-frequency four alleles is greatly reduced when the sample is small, so D' is not suitable for small sample studies. R2 is usually used in the LD plot to represent the LD level of the population.



(A)

locus 2 \ locus 1	A	T
G	4	0
C	0	4

$D = 0.5 - (0.5) \times (0.5) = 0.25$

$D' = \frac{(0.25)}{(0.5)(0.5)} = 1.0$

$r^2 = \frac{(0.25)^2}{(0.5 \times 0.5 \times 0.5 \times 0.5)} = 1.0$

$\delta = \frac{(0.25)}{(0.5)(0.5)} = 1.0$

(B)

locus 2 \ locus 1	A	T
G	2	2
C	2	2

$D = 0.25 - (0.5) \times (0.5) = 0$

$D' = 0$

$r^2 = 0$

$\delta = 0$

(C)

locus 2 \ locus 1	A	T
G	4	2
C	0	2

$D = 0.5 - (0.5) \times (0.75) = 0.125$

$D' = \frac{(0.125)}{(0.5)(0.25)} = 1.0$

$r^2 = \frac{(0.125)^2}{(0.5 \times 0.5 \times 0.75 \times 0.25)} = 0.33$

$\delta = \frac{(0.125)}{(0.75)(0.25)} = 0.67$

(D)

locus 2 \ locus 1	A	T
G	1	1
C	3	3

$D = 0.375 - (0.5) \times (0.75) = 0$

$D' = 0$

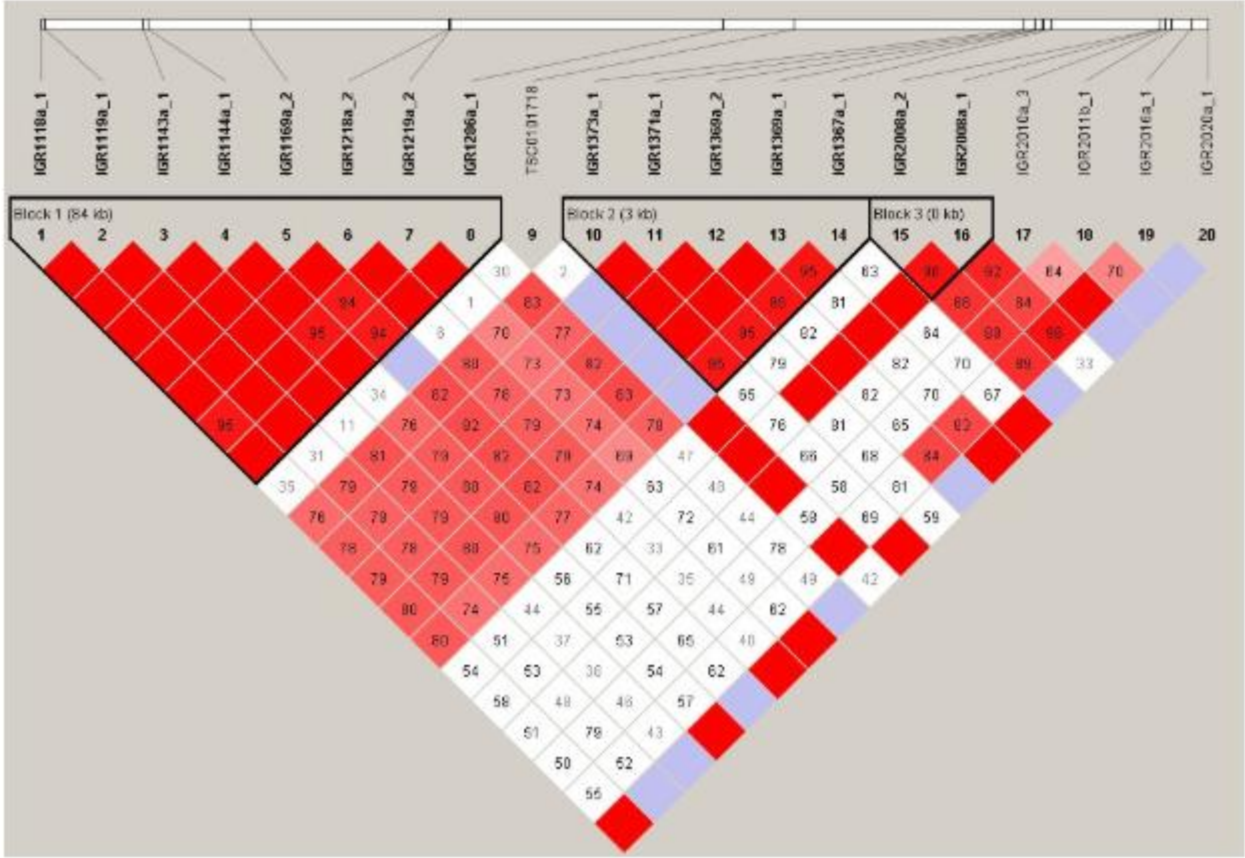
$r^2 = 0$

$\delta = 0$

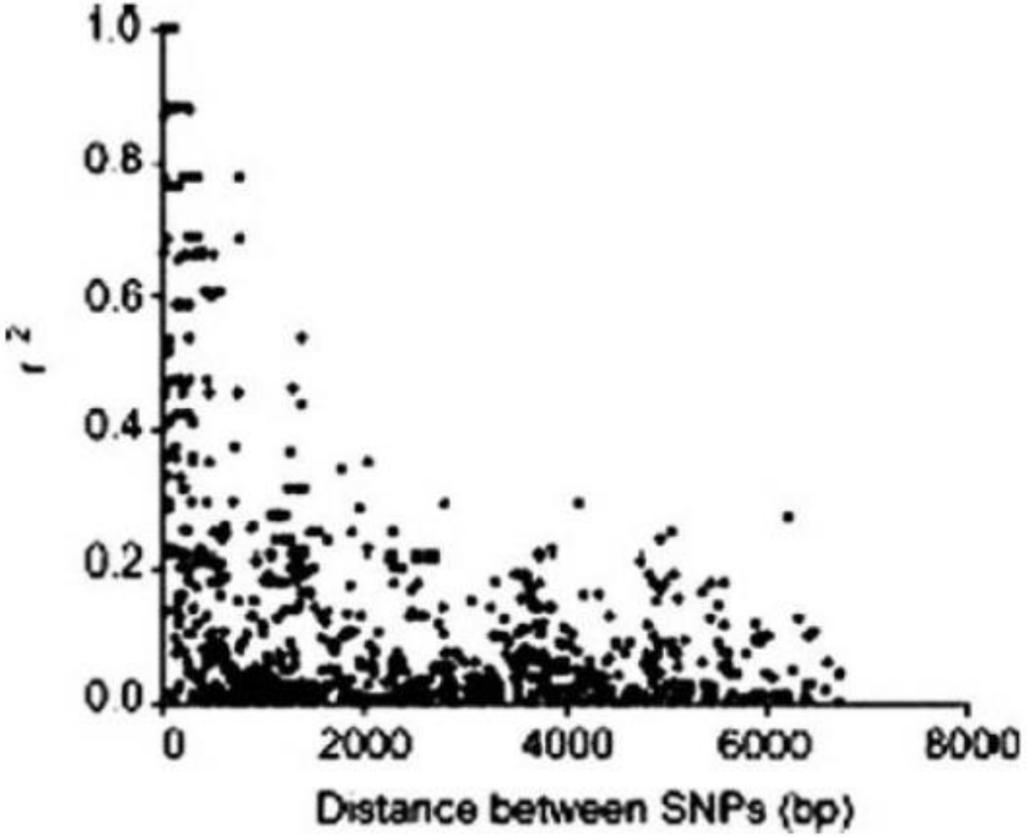
(A) No recombination(mutations at two linked loci not separated in time);
(B)Independent assortment(mutations at two loci not separated in time);
(C) No recombination (onlymutations separated in time);
(D) Low recombination (mutations at two loci notseparated in time)

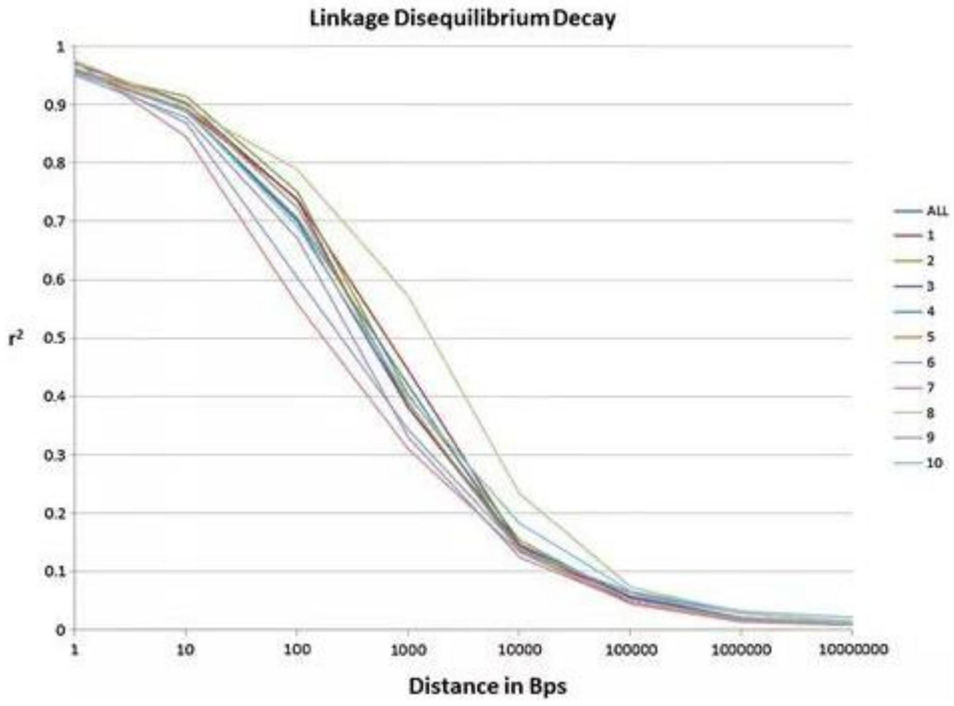
Result display --HEATMAP

In the actual analysis, we usually get the genotyping file of the sample. From this file, we can easily calculate the frequency of allel, but the frequency of the haplotype cannot be directly calculated. The probability of a haplotype is calculated and then calculated. For the calculation of linkage disequilibrium, there are a lot of software available, the most commonly used are plink and haploview, of course, there are many R packages that can be calculated.

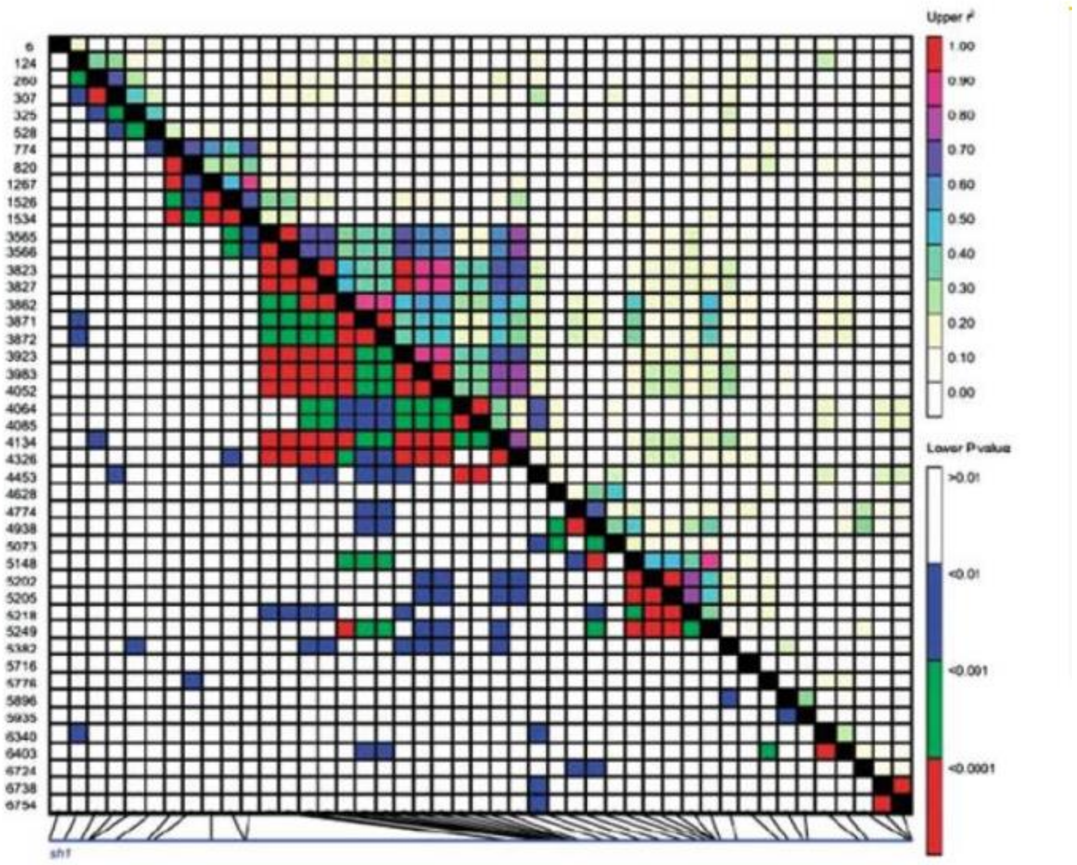


Genome-wide average LD decay



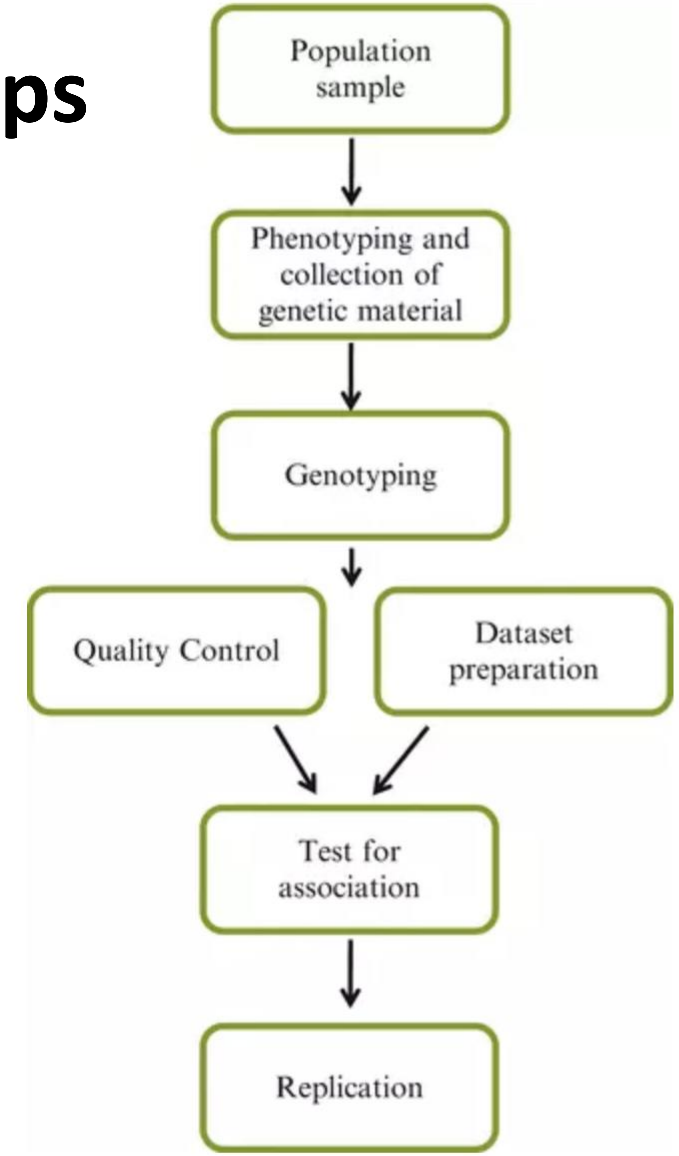


LD matrix for polymorphic sites.



GWAS basic analysis of the content and interpretation of the results

GWAS analysis steps

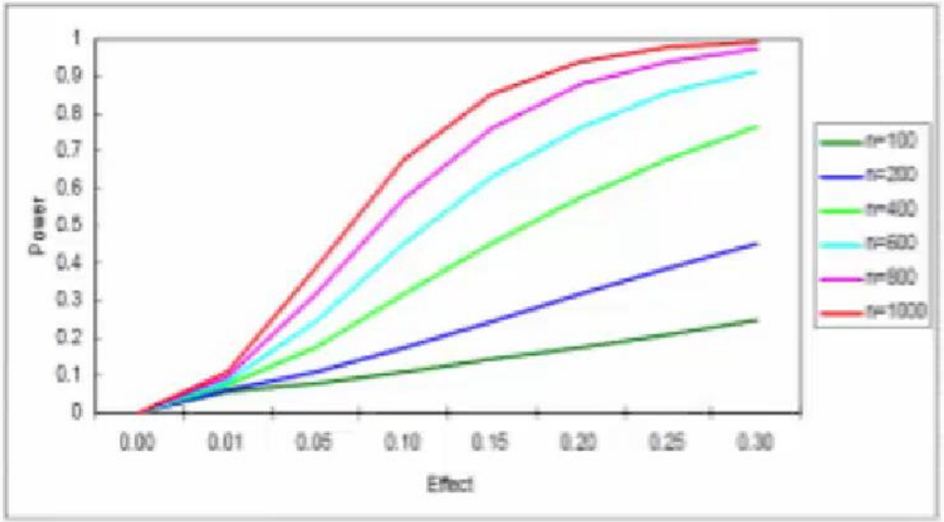


Group material selection

1.Group size

2.group diversity

3.Try to choose the core collection that maximizes the diversity of germplasm resources



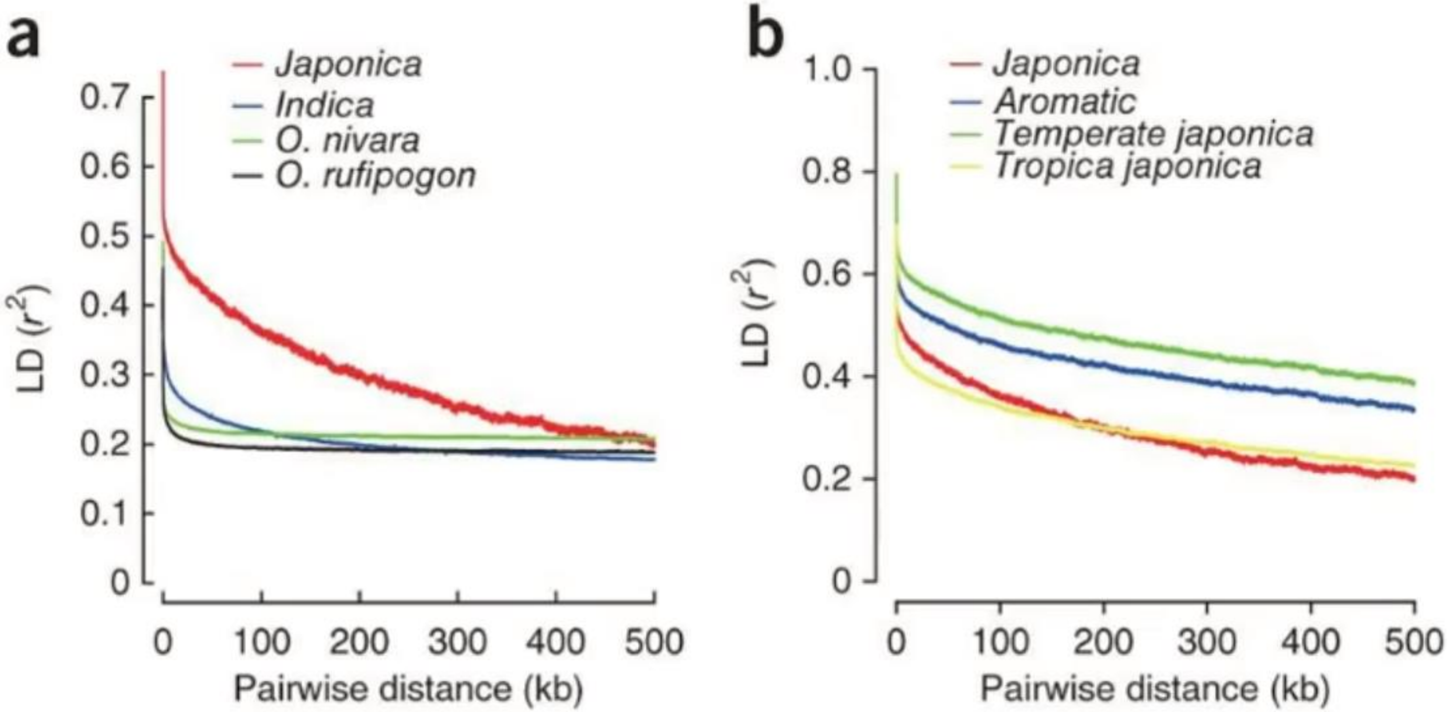
Statistical power of detection in GWAS for variants that explain 1-30% of the variation at type I error =0.05

Genotype data quality control

- 1) Filtering according to the percentage of classification, generally remove the deletion rate of more than 20%, if the amount of data is relatively large, you can relax to 50%.
- 2) Filter by allele frequency to remove the second allele with a frequency less than 5%. If the amount of data is relatively large, it can be relaxed to 1%.
- 3) Filtering of multiple alleles According to the needs of the software, some software does not support multiple alleles.
- 4) Hardy Weinberg balance filtering in human case/control will generally Filters that do not meet the equilibrium of Hardy Weinberg are filtered out, animals and plants do not use this filter.
- 5) Removal of extreme phenotypes

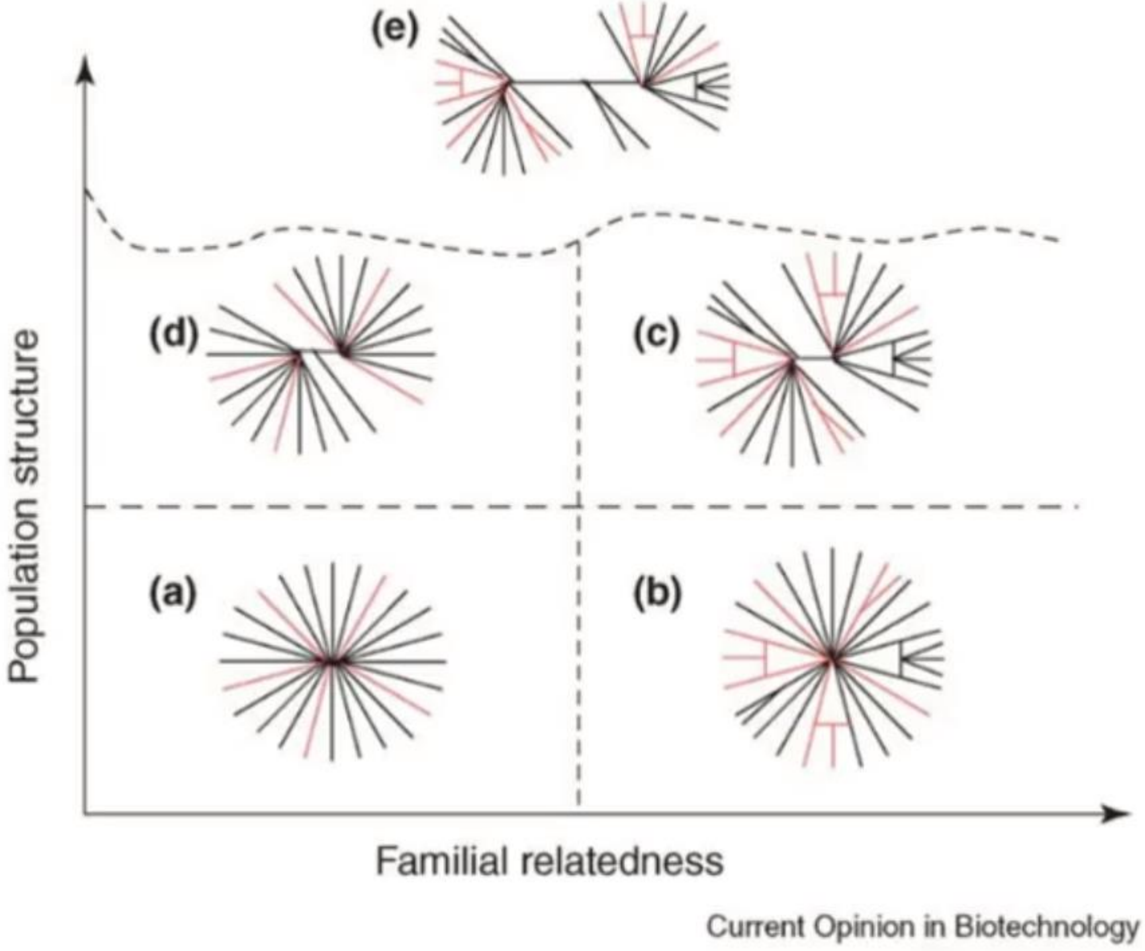
LD attenuation analysis

- **Minimum saturation marker = genomic size / LD attenuation distance**
- **The higher the density, the better: the probability of detecting functional sites increases; the sites in the same block verify each other.**
- **The range of upstream and downstream of the candidate gene can be determined according to the LD attenuation distance.**



Nature Biotechnology 30, 105–111 (2012) doi:10.1038/nbt.2050

Assessment of group structure and kinship



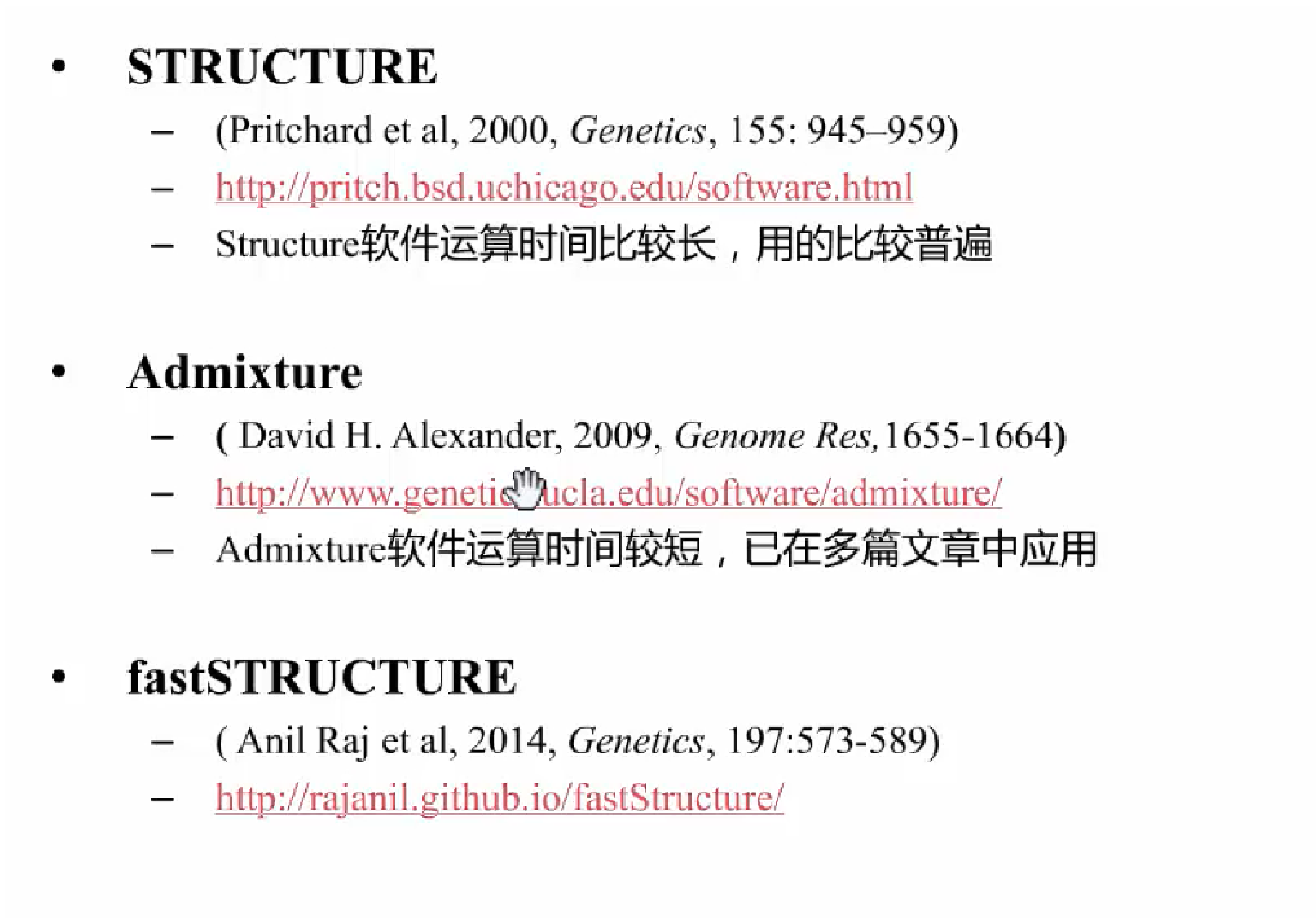
Curr Opin Biotechnol. 2006 Apr;17(2):155-60.

1.Group structure and kinship are the two main factors leading to false positives in association analysis

2.Evaluate group structure and kinship to determine the statistical model used and obtain the corresponding matrix

A--ideal group
B--multiple groups
C--has a group structure group
D--a group with a group structure and a close relationship
E-- a group with a high group structure and a high degree of affinity

Group structure - Q matrix

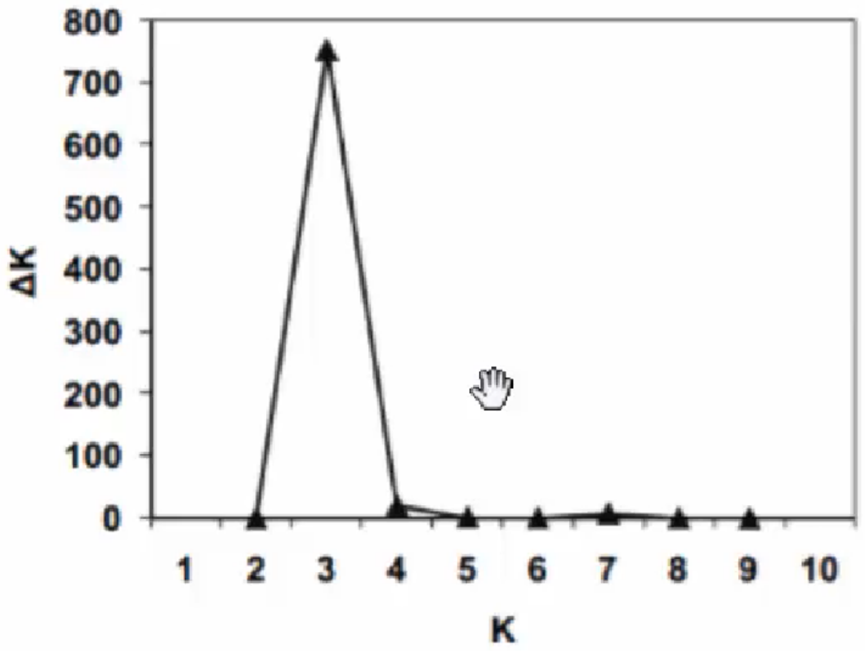


- **STRUCTURE**
  - (Pritchard et al, 2000, *Genetics*, 155: 945–959)
  - http://pritch.bsd.uchicago.edu/software.html
  - Structure软件运算时间比较长，用的比较普遍

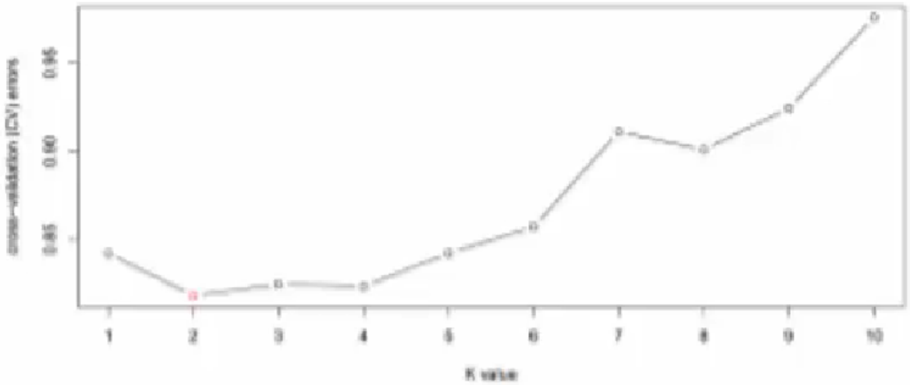
- **Admixture**
  - (David H. Alexander, 2009, *Genome Res*, 1655-1664)
  - http://www.genetics.ucla.edu/software/admixture/
  - Admixture软件运算时间较短，已在多篇文章中应用

- **fastSTRUCTURE**
  - (Anil Raj et al, 2014, *Genetics*, 197:573-589)
  - http://rajanil.github.io/fastStructure/

How to judge the number of subgroups of a group

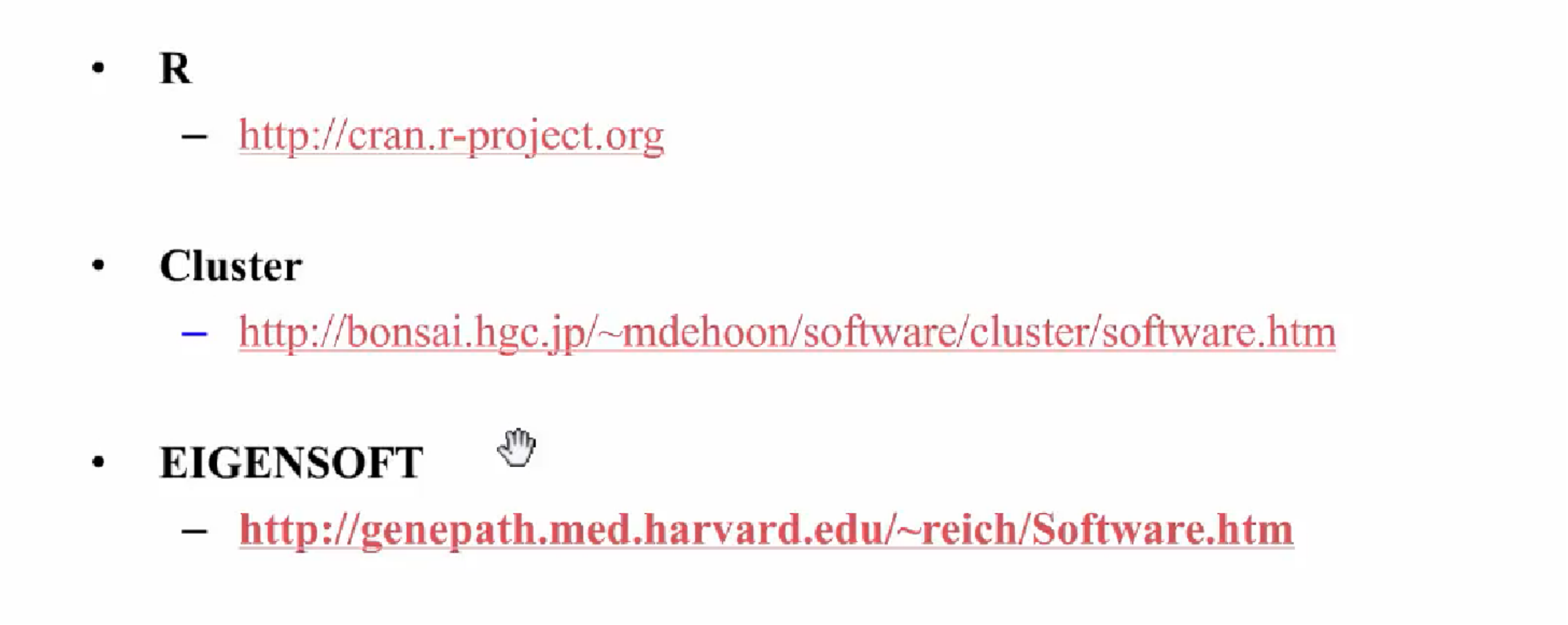


Structure ΔK 计算公式：选最大的ΔK



Admixture CV error 计算公式：选最小的CV

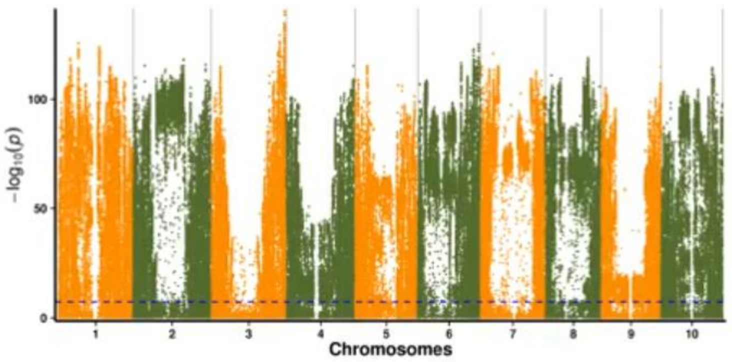
Another way to calculate the population structure - PCA



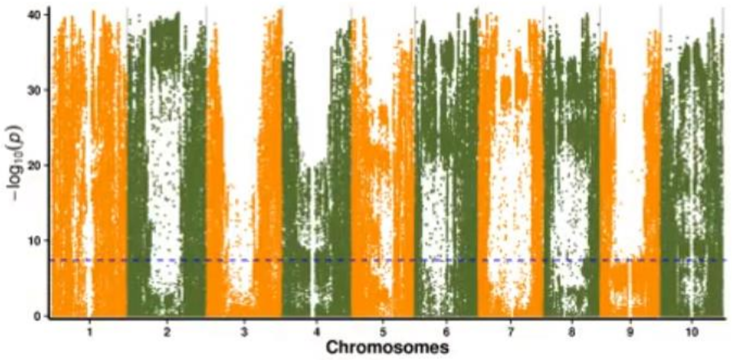
- **R**
  - http://cran.r-project.org
- **Cluster**
  - http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm
- **EIGENSOFT**
  - **http://genepath.med.harvard.edu/~reich/Software.htm**

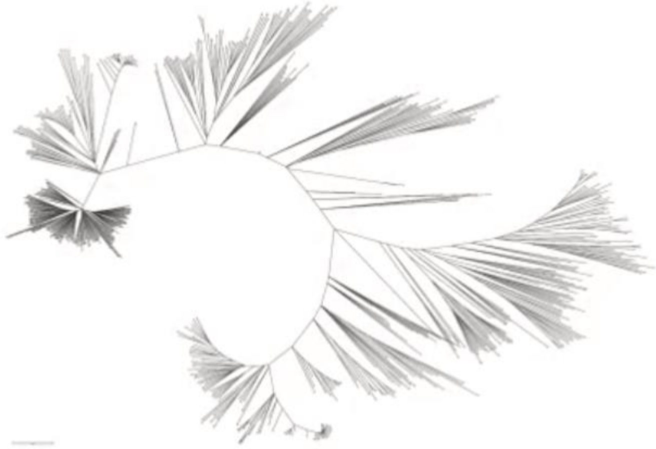
The impact of group structure on GWAS

卡平方检验

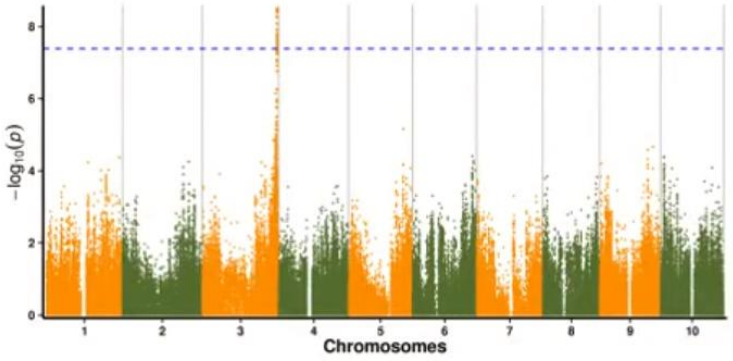


逻辑回归（无群体结构协变量）

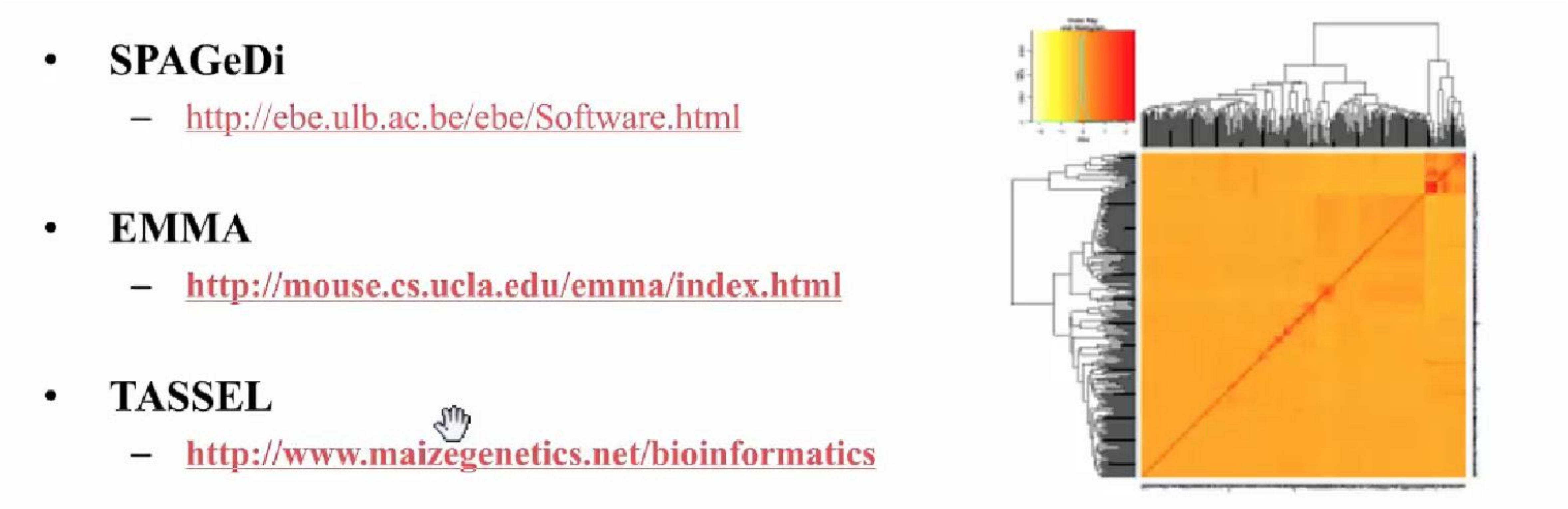




逻辑回归（前5个PC作为协变量）



Inter-individual kinship - K matrix



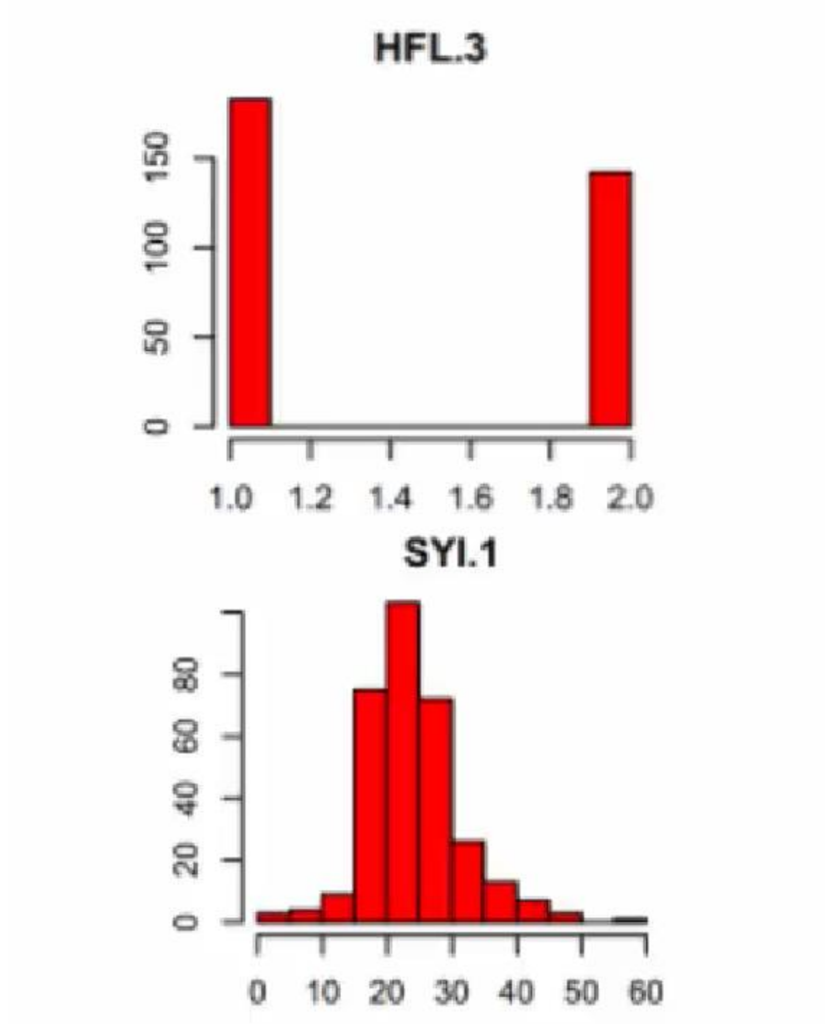
- **SPAGeDi**
  - http://ebe.ulb.ac.be/ebe/Software.html
- **EMMA**
  - **http://mouse.cs.ucla.edu/emma/index.html**
- **TASSEL**
  - **http://www.maizegenetics.net/bioinformatics**

Phenotypic detection

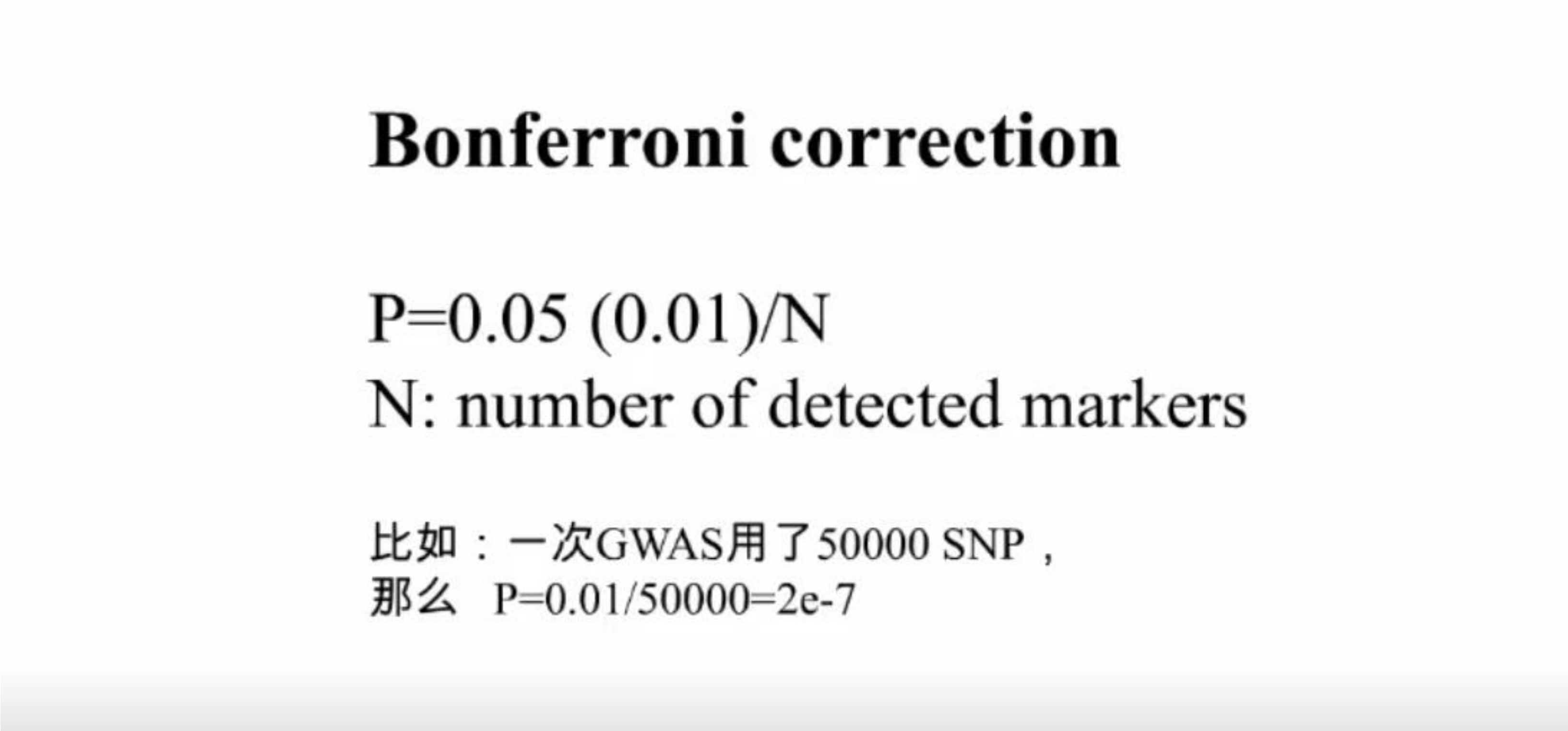
1.Accurate phenotypic testing is a key analysis of correlation analysis

2.Gwas is suitable for both discrete quantitative traits and quality traits

3.When multiple indicators of complex traits can be measured simultaneously, the principal component factors representing the original phenotypic data variation are found as phenotypic data for association analysis.



Screening of GWAS association thresholds



Bonferroni correction

P=0.05 (0.01)/N

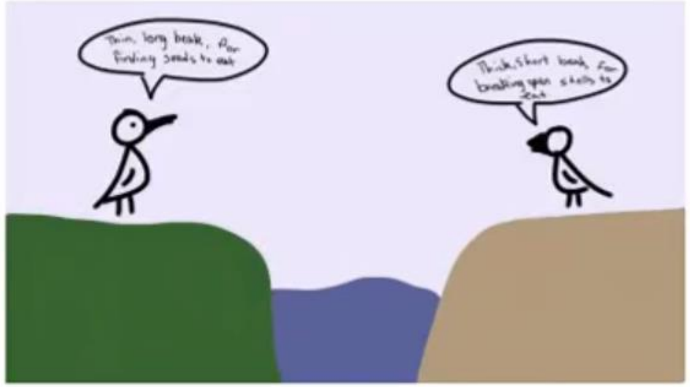
N: number of detected markers

比如：一次GWAS用了50000 SNP，

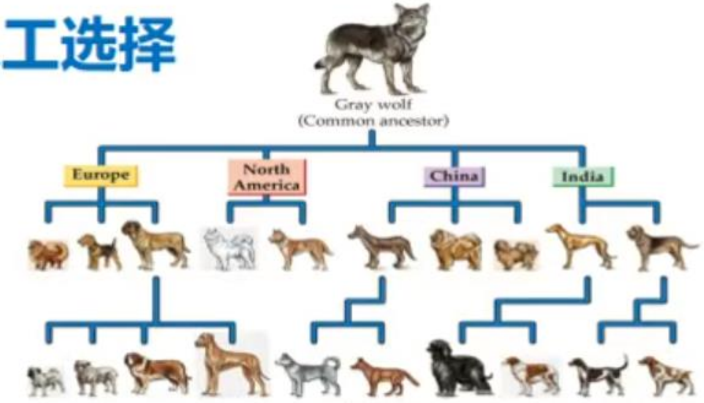
那么 P=0.01/50000=2e-7

Group structure source

- 地理隔离，适应不同的环境

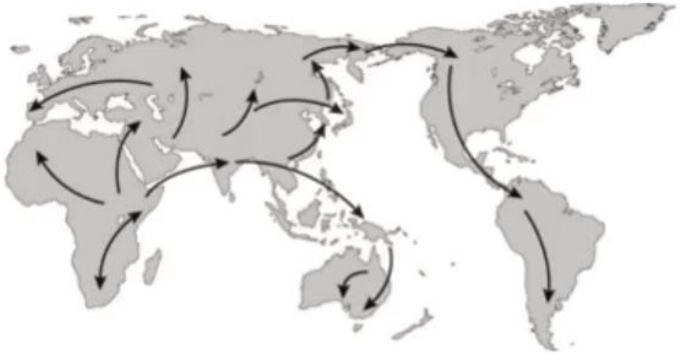

Thin, long beak, for finding seeds to eat
Thick, short beak, for breaking open shells to eat

- 人工选择



DISCOVER BIOLOGY, Second Edition, Chapter 21 Box © 2002 Sinauer Associates, Inc., and W. W. Norton and Company

- Humans originally spread across the world many thousand years ago.
- Migration and genetic drift led to genetic diversity between isolated groups.

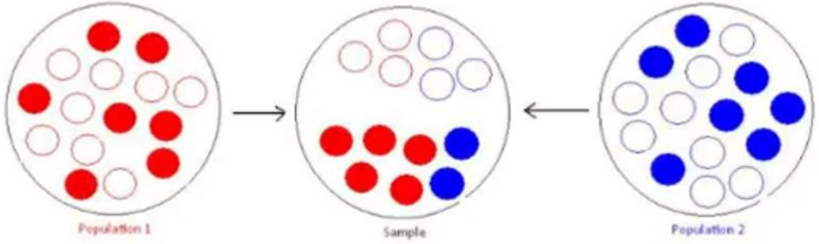


The impact of group structure on GWAS - false positives

Case/control association analysis

1.Compare case/control allele frequency differences

2.At gwas, the proportion of sample cases/controls in each group is out of proportion, resulting in markers associated with group stratification being associated with a large number of false positives.



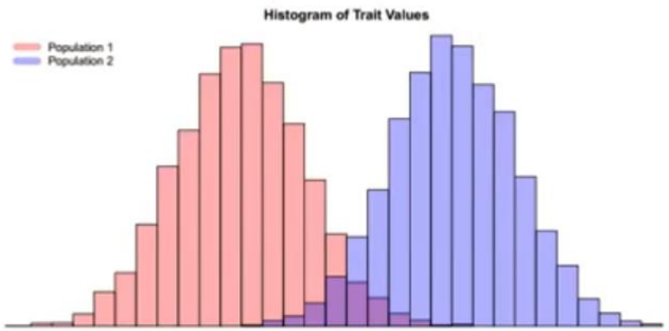
实心Case，空心Control；红色Pop1，蓝色Pop2

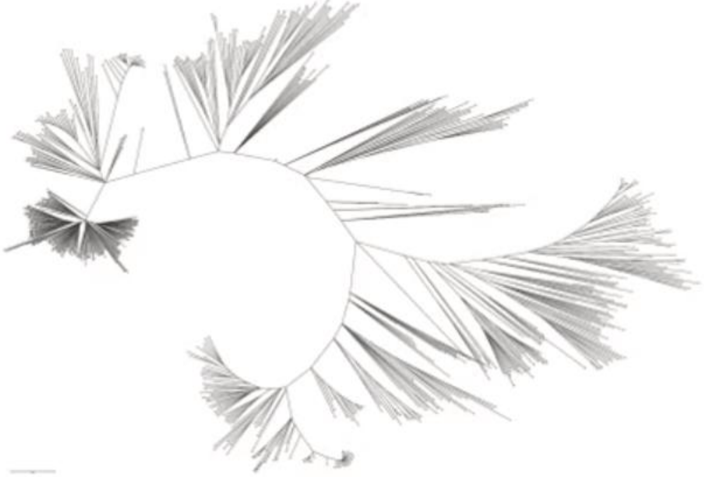
Quantitative trait association analysis

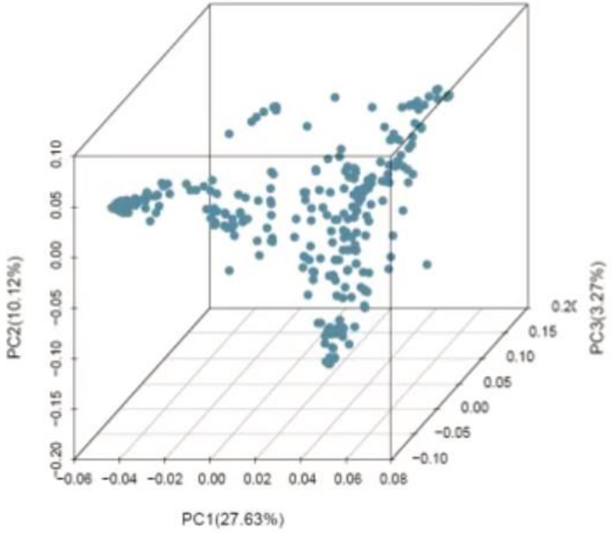
1.Verify the correlation between phenotype and genotype

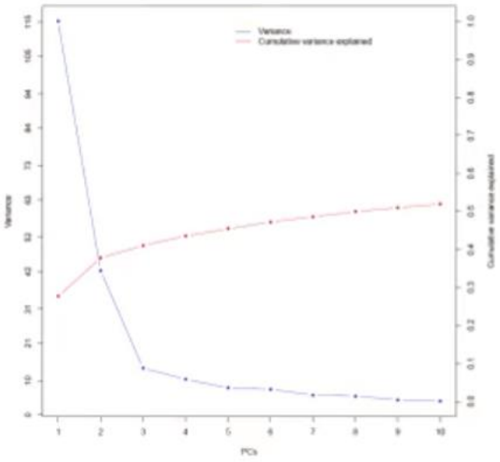
2.Phenotype： The phenotype between subgroups varies from group to group.

3.Genotype: There are population-specific loci that are associated with phenotypes, resulting in a large number of false positives.

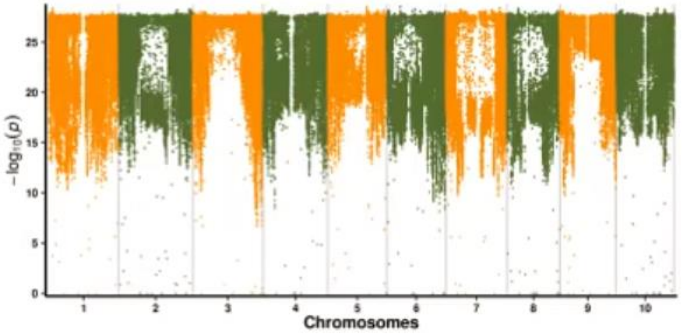




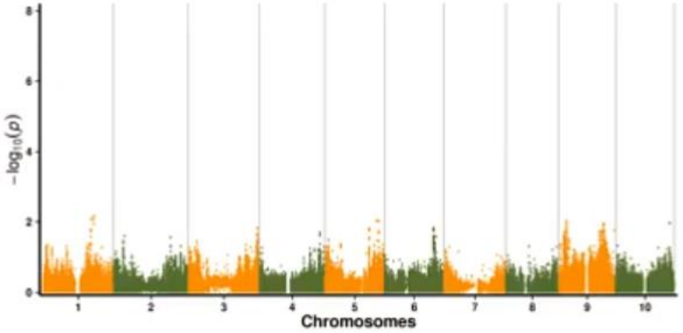




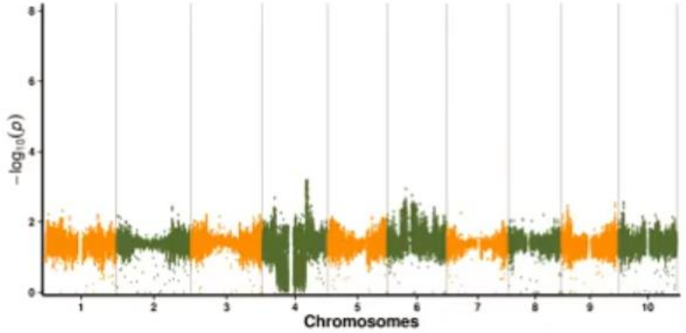
genotype～PC1



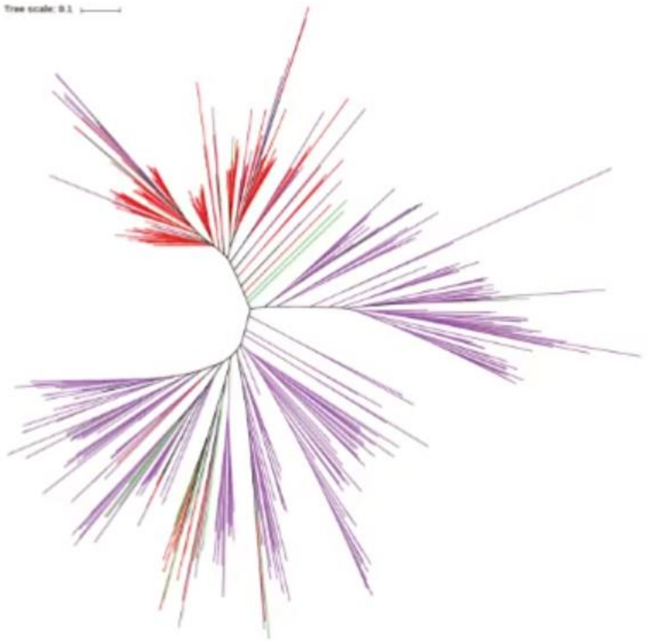
genotype～PC2



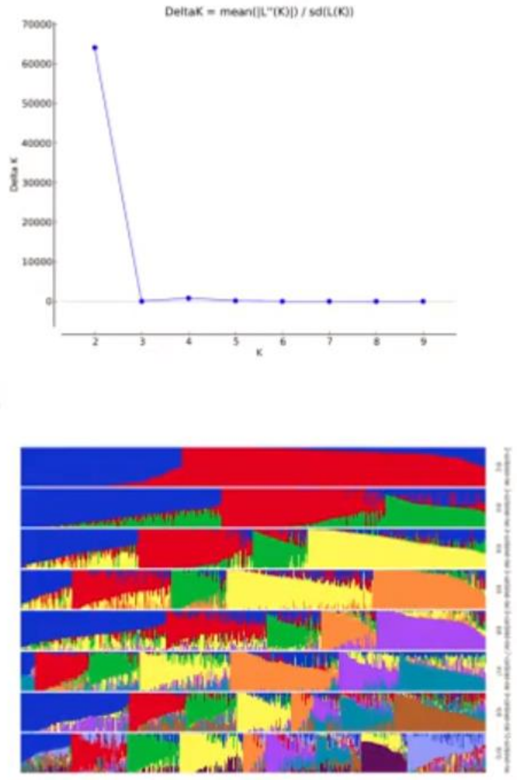
genotype～PC3



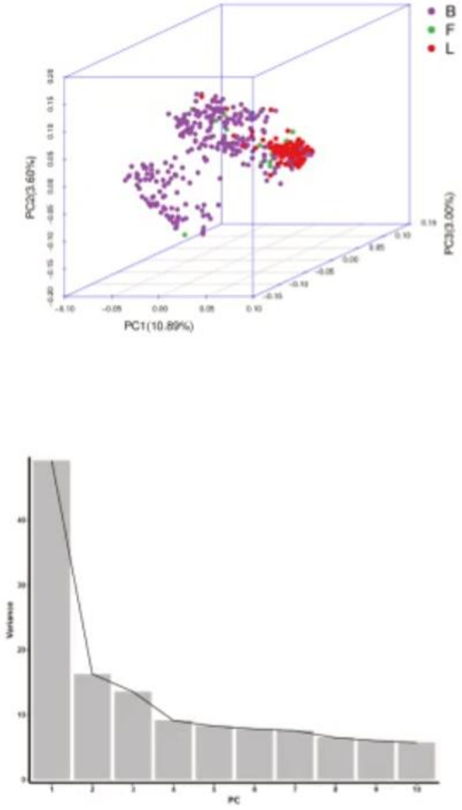
Population structure assessment



Building a phylogenetic tree



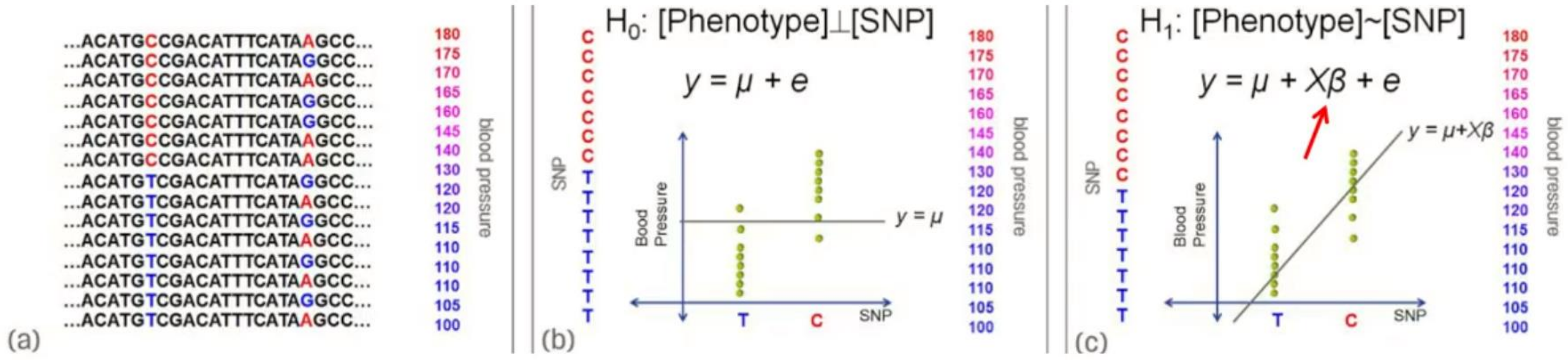
Group structure analysis



PCA analysis

Introduction to commonly used GWAS statistical methods and models

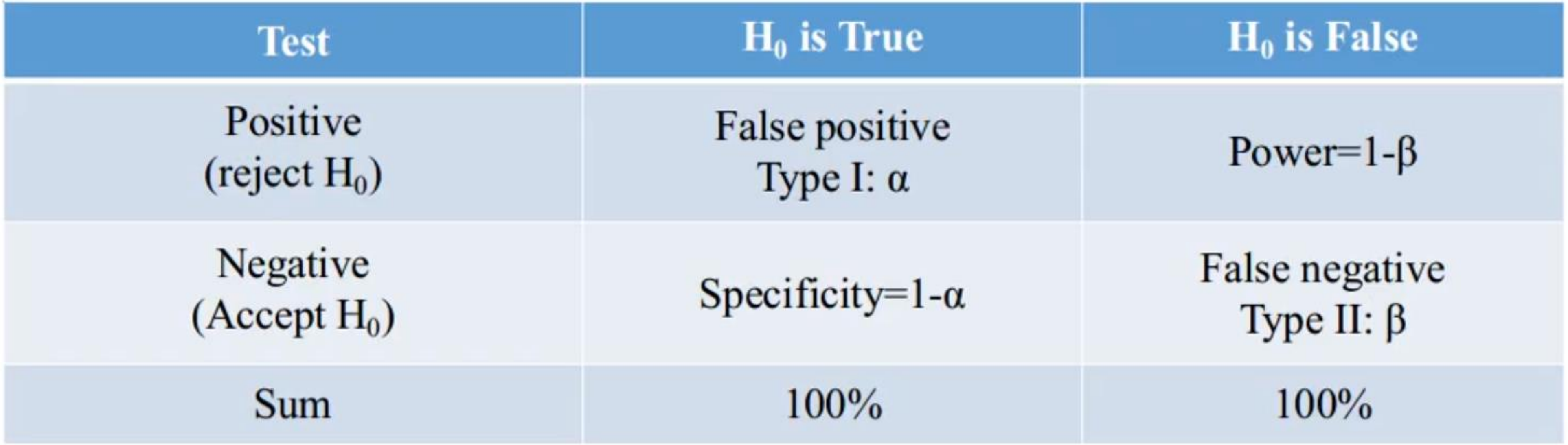
- H0 (null hypothesis): The null hypothesis, which is a pre-established hypothesis when performing statistical tests, generally a hypothesis that wishes to prove its error. H0 in GWAS is zero with a regression coefficient of the marker, and SNP has no effect on the phenotype.
- Alternative hypothesis (H1, also called Alternative Hypothesis): A hypothesis against the null hypothesis that H1 in GWAS means that the regression coefficient of the marker is not zero, and the SNP is related to the phenotype.


...ACATGCCGACATTTCATAAGCC...
...ACATGCCGACATTTCATAGGCC...
...ACATGCCGACATTTCATAAGCC...
...ACATGCCGACATTTCATAGGCC...
...ACATGCCGACATTTCATAGGCC...
...ACATGCCGACATTTCATAAGCC...
...ACATGCCGACATTTCATAAGCC...
...ACATGTCGACATTTCATAGGCC...
...ACATGTCGACATTTCATAAGCC...
...ACATGTCGACATTTCATAGGCC...
...ACATGTCGACATTTCATAAGCC...
...ACATGTCGACATTTCATAGGCC...
...ACATGTCGACATTTCATAAGCC...
...ACATGTCGACATTTCATAGGCC...
...ACATGTCGACATTTCATAAGCC...
180 175 170 165 160 145 140 130 120 120 115 110 110 110 105 100
blood pressure
(a)
H_0: [Phenotype]⊥[SNP]
$y = \mu + e$
SNP
C C C C C C C T T T T T T T T
Bood Pressure
$y = \mu$
T
C
SNP
180 175 170 165 160 145 140 130 120 120 115 110 110 110 105 100
blood pressure
(b)
H_1: [Phenotype]~[SNP]
$y = \mu + X\beta + e$
$y = \mu + X\beta$
SNP
C C C C C C C T T T T T T T T
Blood Pressure
T
C
SNP
180 175 170 165 160 145 140 130 120 120 115 110 110 110 105 100
blood pressure
(c)

http://dx.doi.org/10.1101/092106

Two types of errors and statistical power

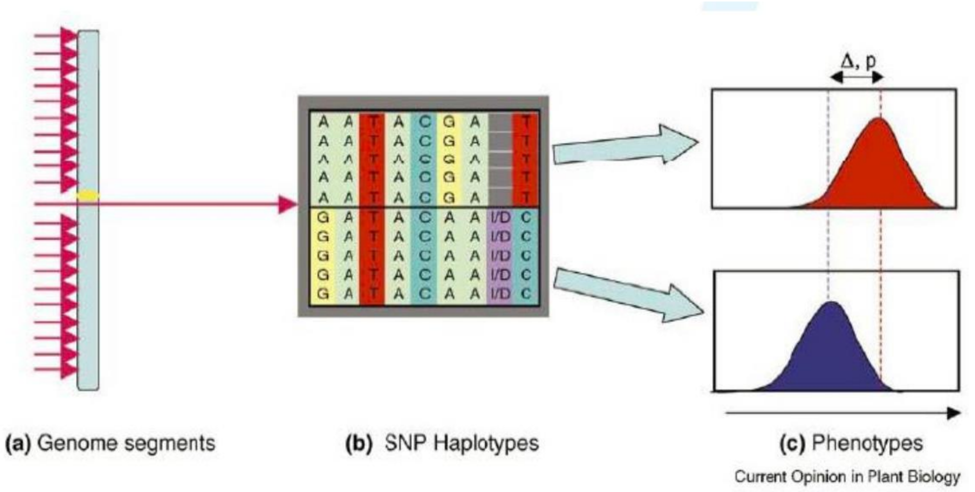
Type I error: rejects the true H0, which is a false positive, and the probability α is the level of significance;
Type II error: Accepts the wrong H0, which is a false negative with a probability of β;
Power: The probability of rejecting the error H0 1-β



Test	H_0 is True	H_0 is False
Positive (reject H_0)	False positive Type I: α	Power=1-β
Negative (Accept H_0)	Specificity=1-α	False negative Type II: β
Sum	100%	100%

The simplest model - analysis of variance

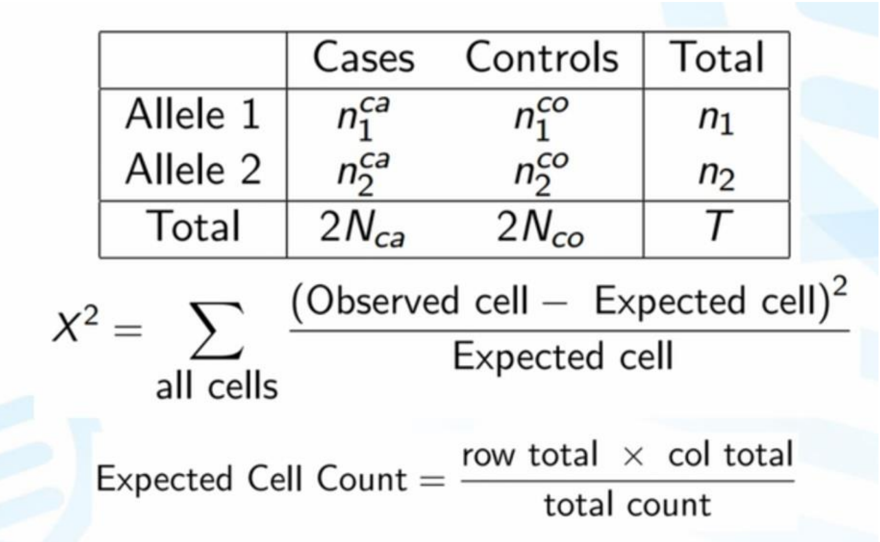
Single site
Association



A	A	T	A	C	G	A		T
A	A	T	A	C	G	A		T
A	A	T	A	C	G	A		T
A	A	T	A	C	G	A		T
A	A	T	A	C	G	A		T
G	A	T	A	C	A	A	I/D	C
G	A	T	A	C	A	A	I/D	C
G	A	T	A	C	A	A	I/D	C
G	A	T	A	C	A	A	I/D	C
G	A	T	A	C	A	A	I/D	C

(a) Genome segments **(b)** SNP Haplotypes **(c)** Phenotypes

Current Opinion in Plant Biology

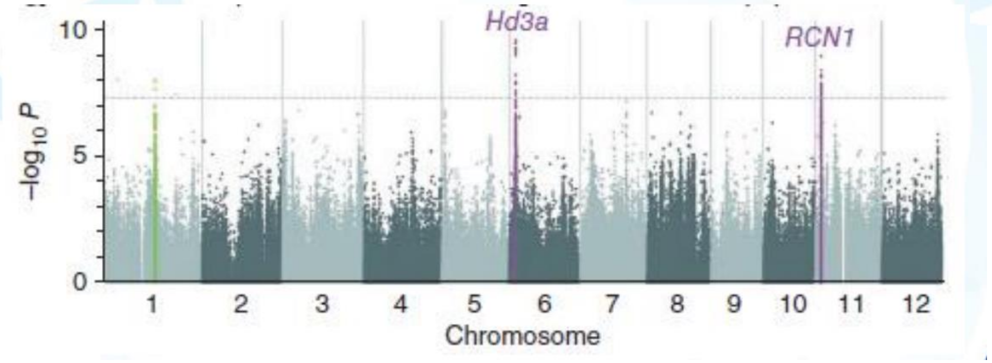


	Cases	Controls	Total
Allele 1	n_1^{ca}	n_1^{co}	n_1
Allele 2	n_2^{ca}	n_2^{co}	n_2
Total	$2N_{ca}$	$2N_{co}$	T

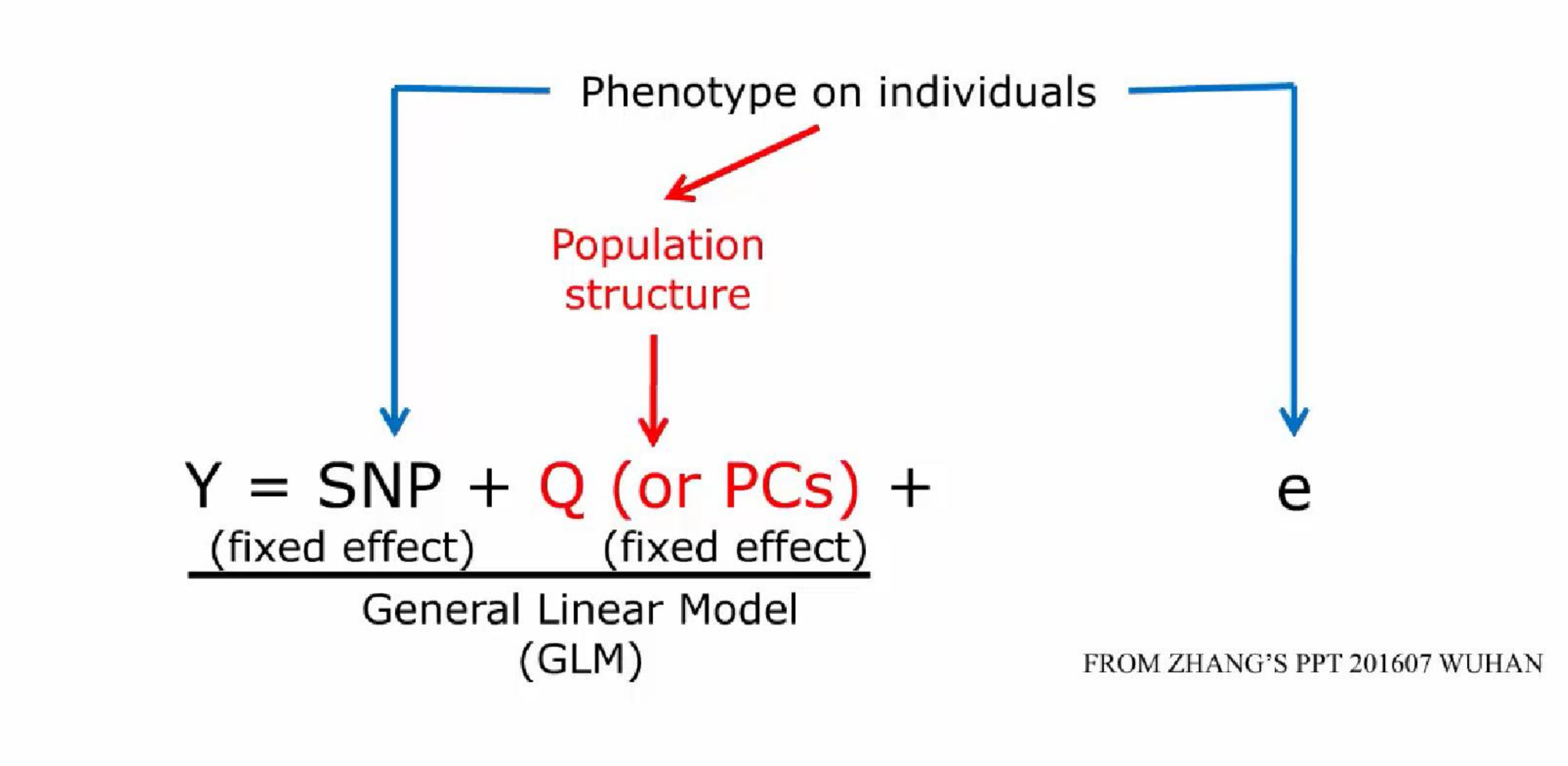
$$X^2 = \sum_{\text{all cells}} \frac{(\text{Observed cell} - \text{Expected cell})^2}{\text{Expected cell}}$$

$$\text{Expected Cell Count} = \frac{\text{row total} \times \text{col total}}{\text{total count}}$$

Manhattan map:
Whole genome
There is a show of places
Show

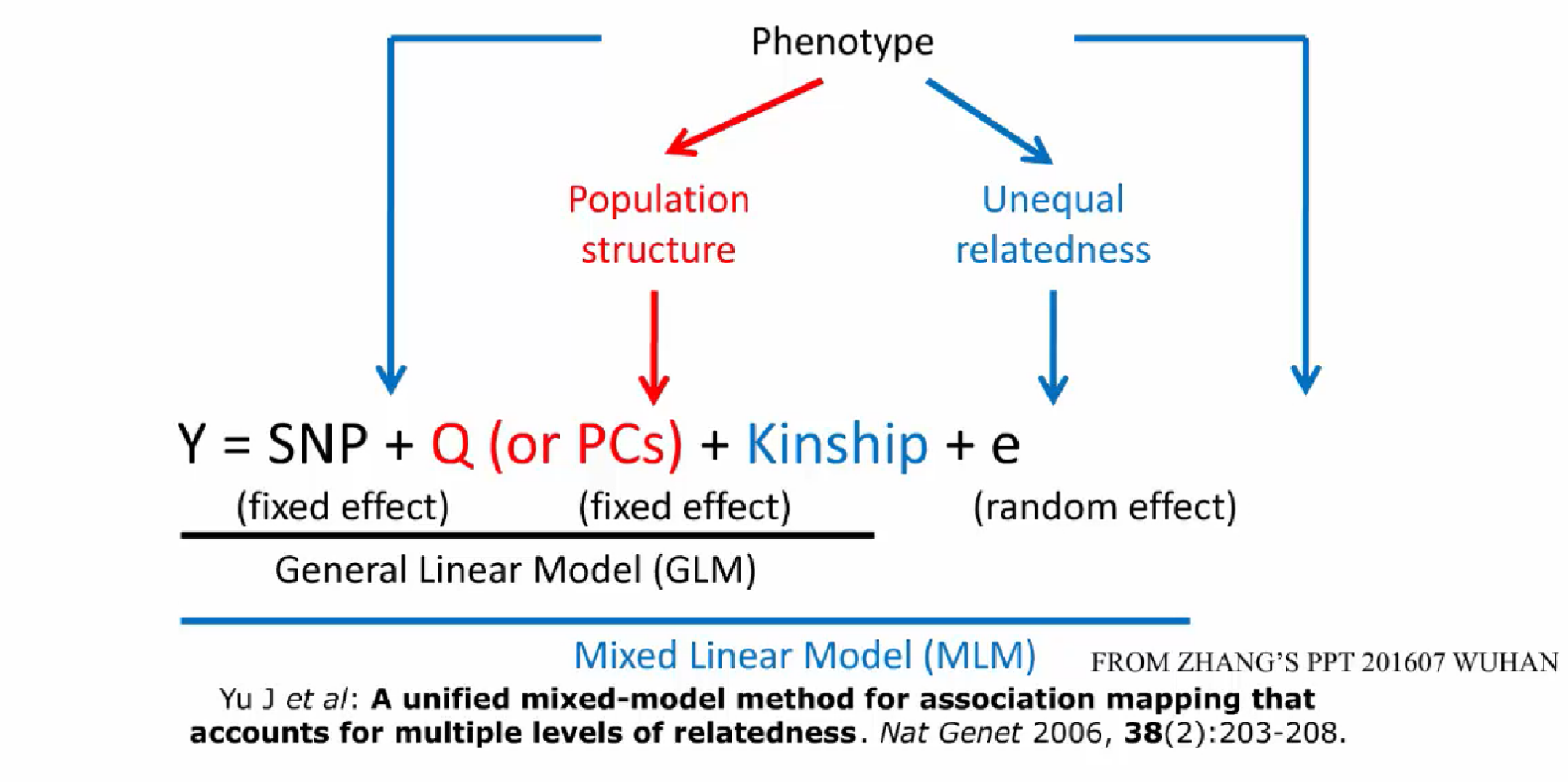


Logistic regression： General Linear Analysis Model (GLM)



FROM ZHANG'S PPT 201607 WUHAN

Logistic regression：Mixed linear model MLM

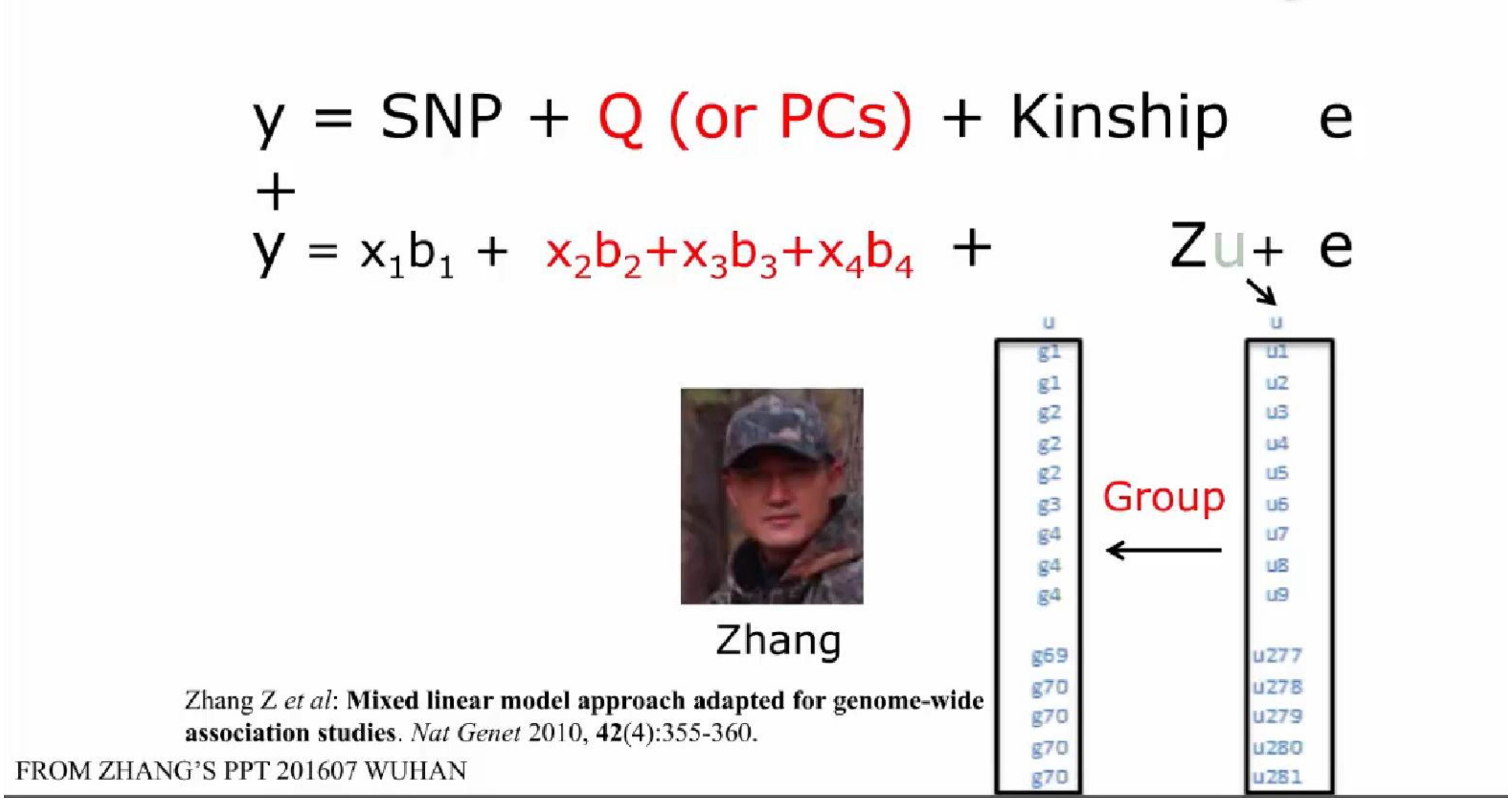


Y = SNP + Q (or PCs) + Kinship + e

FROM ZHANG'S PPT 201607 WUHAN

Yu J *et al*: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness**. *Nat Genet* 2006, **38**(2):203-208.

CMLM： Compressed mixed linear model



$$y = \text{SNP} + Q\ (\text{or PCs}) + \text{Kinship} \quad e$$

$+$

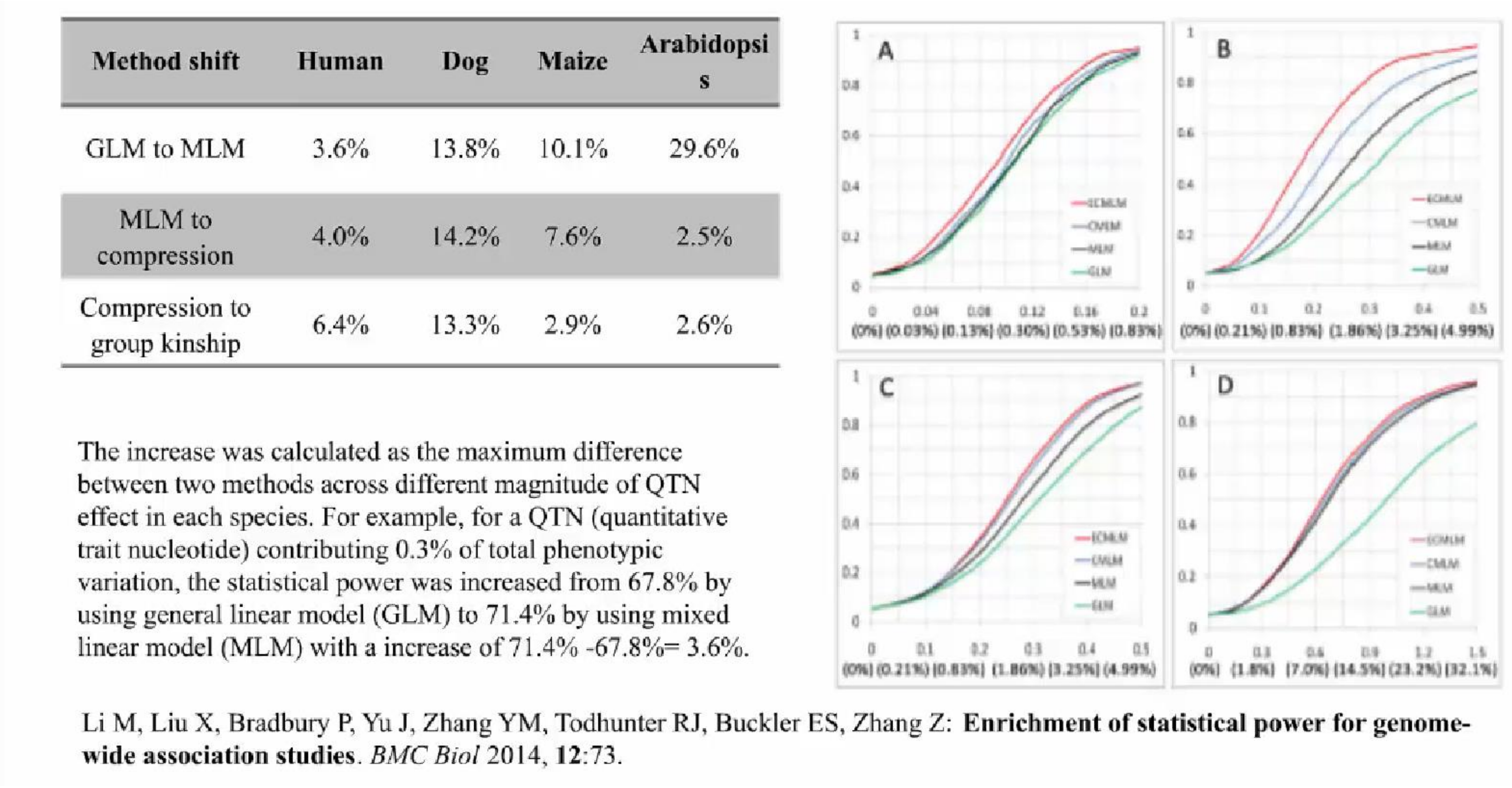
$$y = x_1b_1 + x_2b_2+x_3b_3+x_4b_4 + Zu+ e$$

Zhang

Zhang Z *et al*: **Mixed linear model approach adapted for genome-wide association studies**. *Nat Genet* 2010, **42**(4):355-360.

FROM ZHANG’S PPT 201607 WUHAN

Comparison of different models

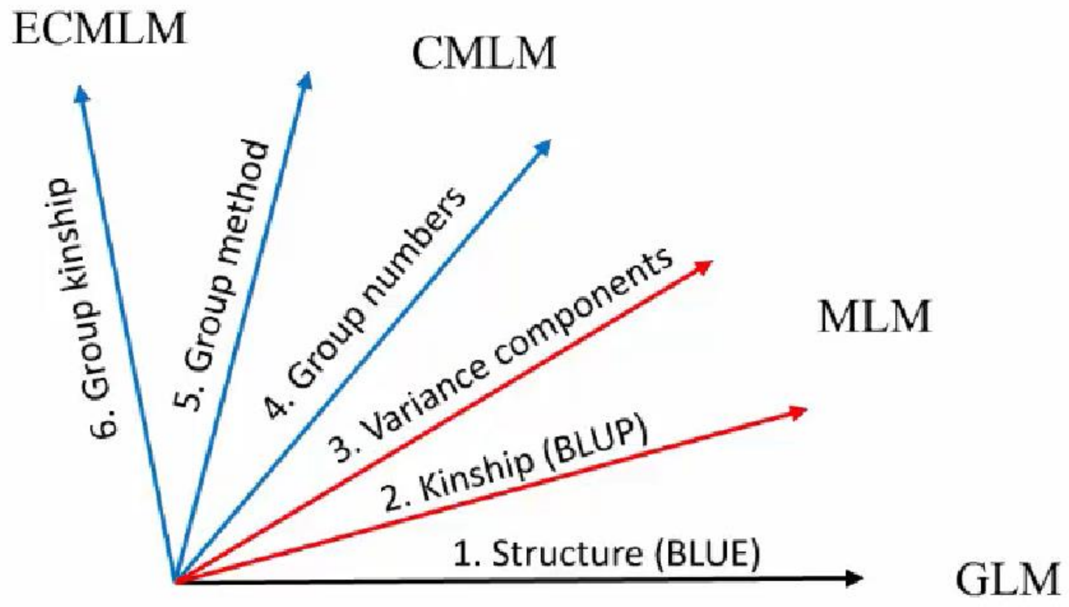


Method shift	Human	Dog	Maize	Arabidopsi s
GLM to MLM	3.6%	13.8%	10.1%	29.6%
MLM to compression	4.0%	14.2%	7.6%	2.5%
Compression to group kinship	6.4%	13.3%	2.9%	2.6%

The increase was calculated as the maximum difference between two methods across different magnitude of QTN effect in each species. For example, for a QTN (quantitative trait nucleotide) contributing 0.3% of total phenotypic variation, the statistical power was increased from 67.8% by using general linear model (GLM) to 71.4% by using mixed linear model (MLM) with a increase of 71.4% -67.8%= 3.6%.

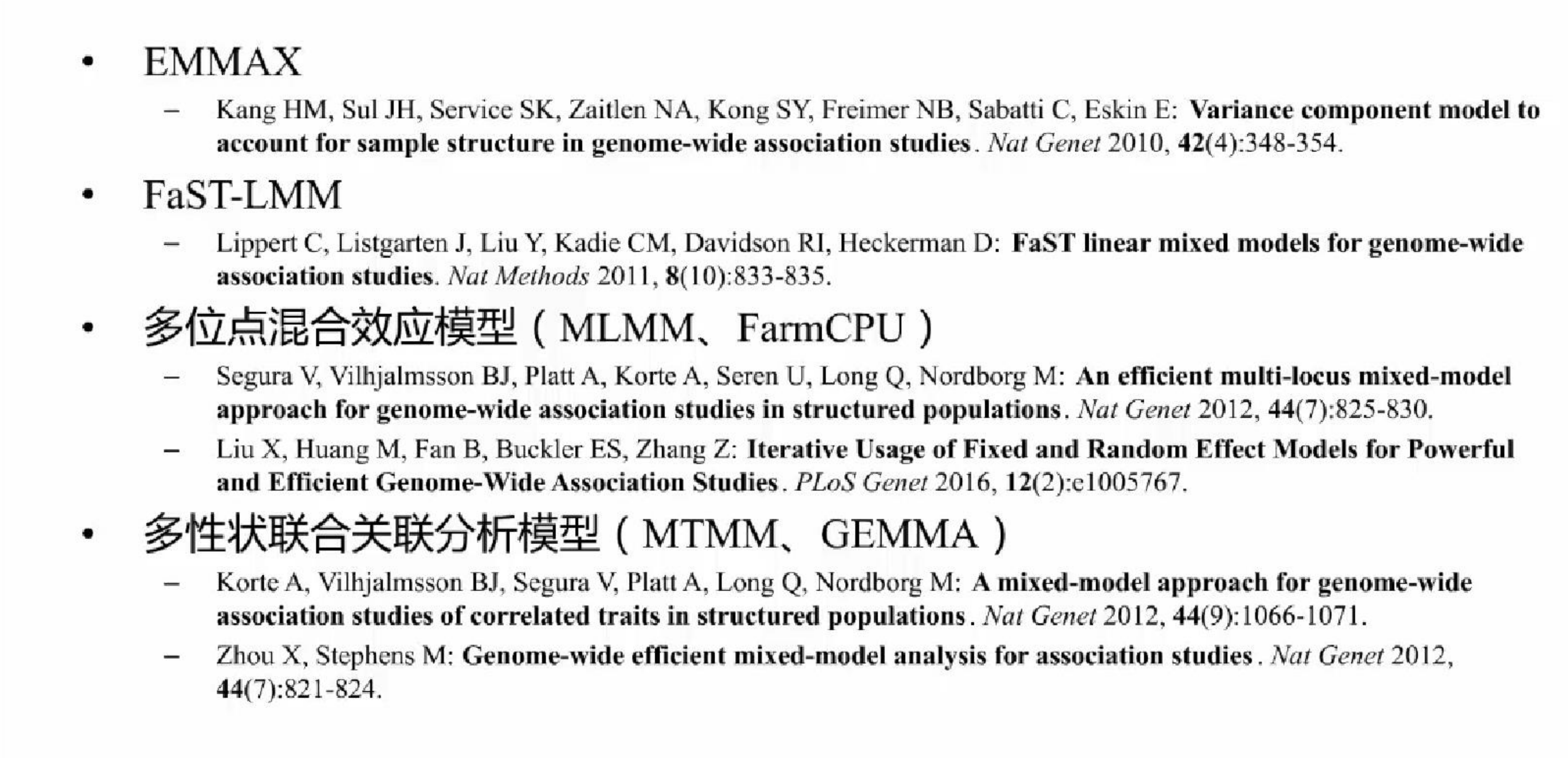
Li M, Liu X, Bradbury P, Yu J, Zhang YM, Todhunter RJ, Buckler ES, Zhang Z: **Enrichment of statistical power for genome-wide association studies**. *BMC Biol* 2014, **12**:73.

Comparison of different models



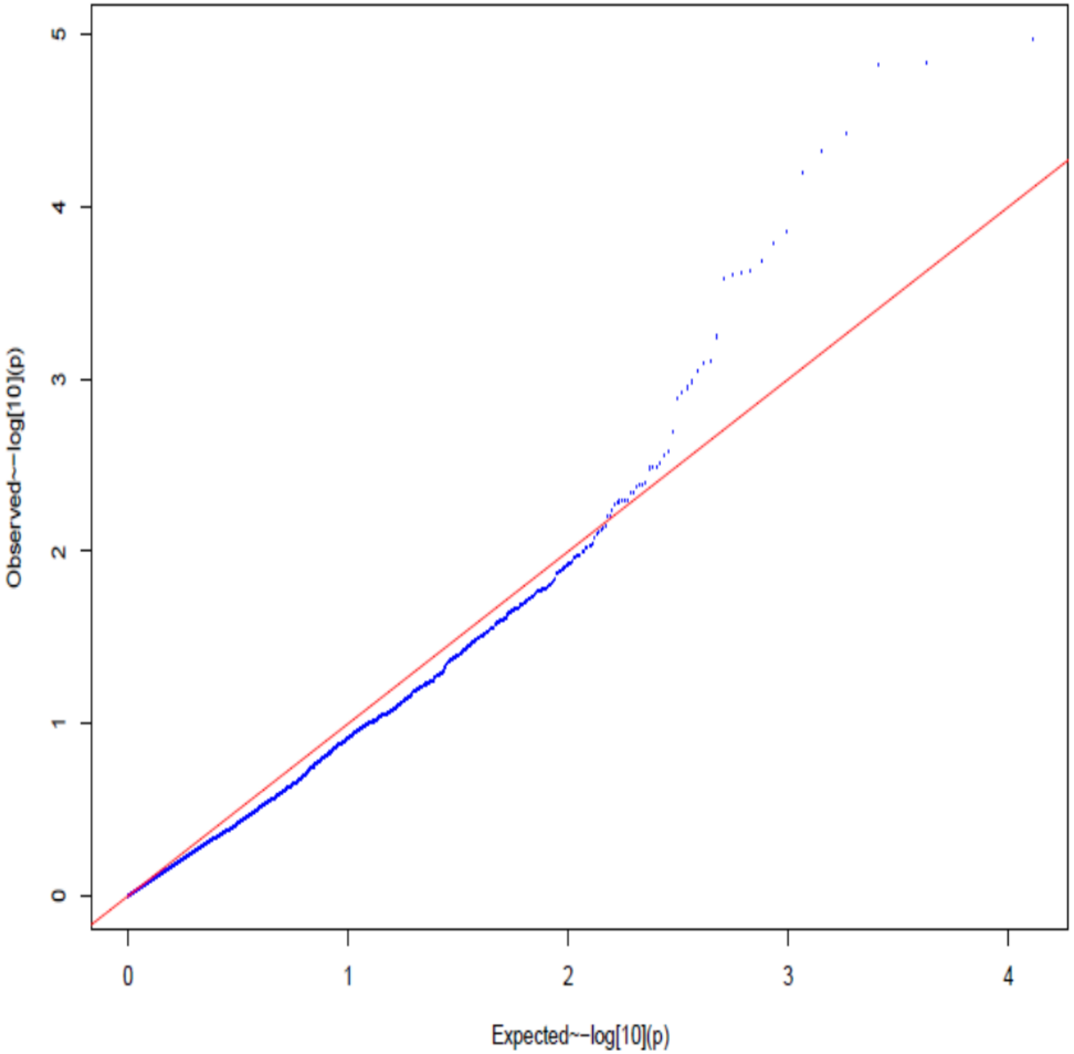
Li M, Liu X, Bradbury P, Yu J, Zhang YM, Todhunter RJ, Buckler ES, Zhang Z: **Enrichment of statistical power for genome-wide association studies**. *BMC Biol* 2014, **12**:73.

Other association analysis models



- EMMAX
  - Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E: **Variance component model to account for sample structure in genome-wide association studies**. *Nat Genet* 2010, **42**(4):348-354.
- FaST-LMM
  - Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D: **FaST linear mixed models for genome-wide association studies**. *Nat Methods* 2011, **8**(10):833-835.
- 多位点混合效应模型（MLMM、FarmCPU）
  - Segura V, Vilhjalmsson BJ, Platt A, Korte A, Seren U, Long Q, Nordborg M: **An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations**. *Nat Genet* 2012, **44**(7):825-830.
  - Liu X, Huang M, Fan B, Buckler ES, Zhang Z: **Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies**. *PLoS Genet* 2016, **12**(2):e1005767.
- 多性状联合关联分析模型（MTMM、GEMMA）
  - Korte A, Vilhjalmsson BJ, Segura V, Platt A, Long Q, Nordborg M: **A mixed-model approach for genome-wide association studies of correlated traits in structured populations**. *Nat Genet* 2012, **44**(9):1066-1071.
  - Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies**. *Nat Genet* 2012, **44**(7):821-824.

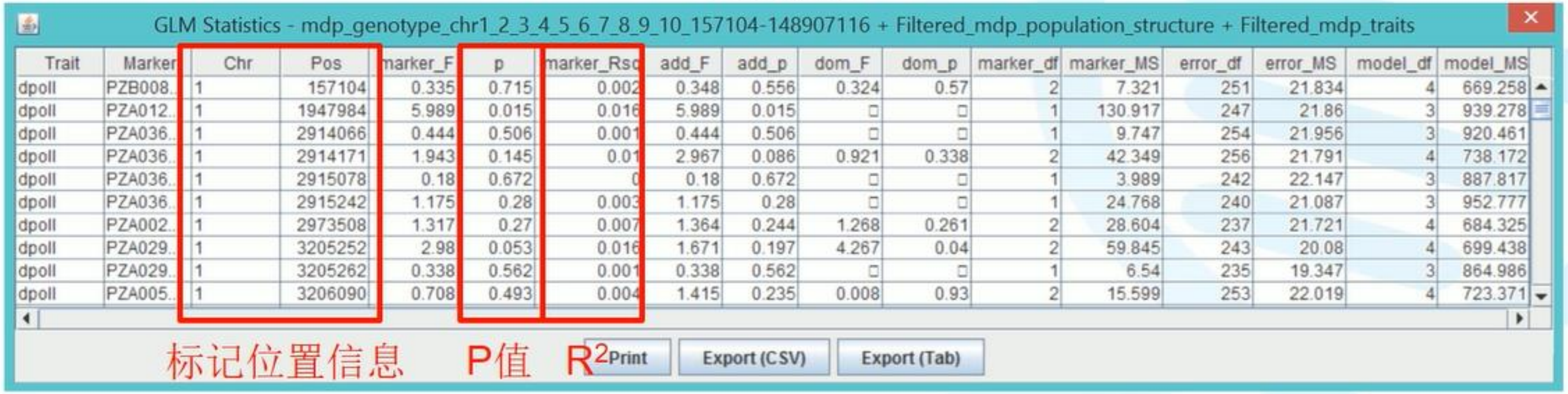
Judging the rationality of the model - QQplot



Good mode: early stage consistent, late rise

RESULT-DATA

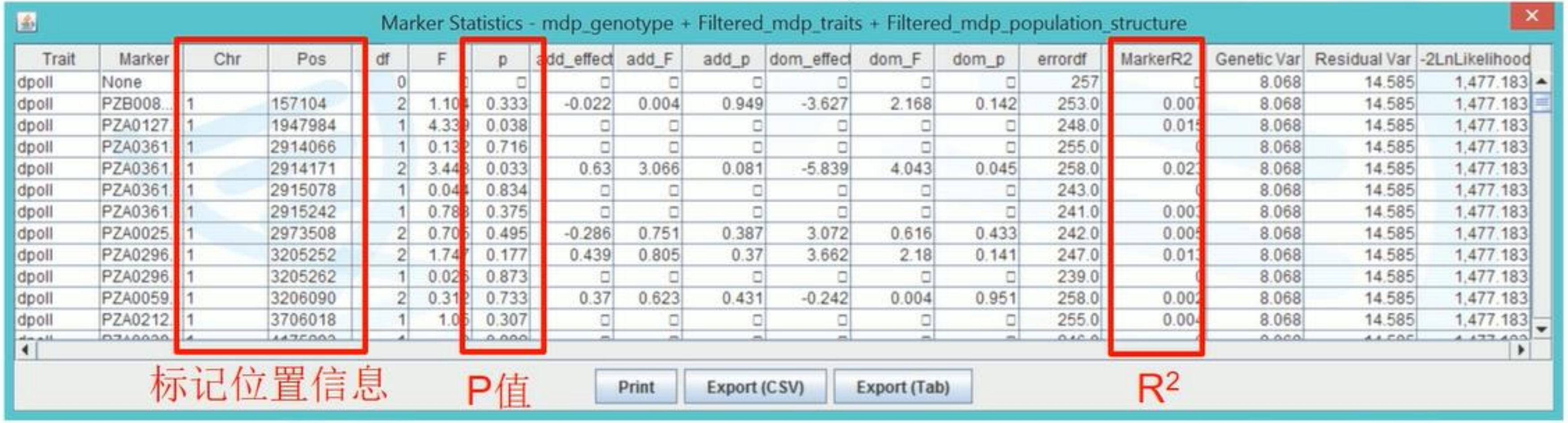
GLM结果



GLM Statistics - mdp_genotype_chr1_2_3_4_5_6_7_8_9_10_157104-148907116 + Filtered_mdp_population_structure + Filtered_mdp_traits

Trait	Marker	Chr	Pos	marker_F	p	marker_Rsq	add_F	add_p	dom_F	dom_p	marker_df	marker_MS	error_df	error_MS	model_df	model_MS
dpoll	PZB008...	1	157104	0.335	0.715	0.002	0.348	0.556	0.324	0.57	2	7.321	251	21.834	4	669.258
dpoll	PZA012...	1	1947984	5.989	0.015	0.016	5.989	0.015	□	□	1	130.917	247	21.86	3	939.278
dpoll	PZA036...	1	2914066	0.444	0.506	0.001	0.444	0.506	□	□	1	9.747	254	21.956	3	920.461
dpoll	PZA036...	1	2914171	1.943	0.145	0.01	2.967	0.086	0.921	0.338	2	42.349	256	21.791	4	738.172
dpoll	PZA036...	1	2915078	0.18	0.672	0	0.18	0.672	□	□	1	3.989	242	22.147	3	887.817
dpoll	PZA036...	1	2915242	1.175	0.28	0.003	1.175	0.28	□	□	1	24.768	240	21.087	3	952.777
dpoll	PZA002...	1	2973508	1.317	0.27	0.007	1.364	0.244	1.268	0.261	2	28.604	237	21.721	4	684.325
dpoll	PZA029...	1	3205252	2.98	0.053	0.016	1.671	0.197	4.267	0.04	2	59.845	243	20.08	4	699.438
dpoll	PZA029...	1	3205262	0.338	0.562	0.001	0.338	0.562	□	□	1	6.54	235	19.347	3	864.986
dpoll	PZA005...	1	3206090	0.708	0.493	0.004	1.415	0.235	0.008	0.93	2	15.599	253	22.019	4	723.371

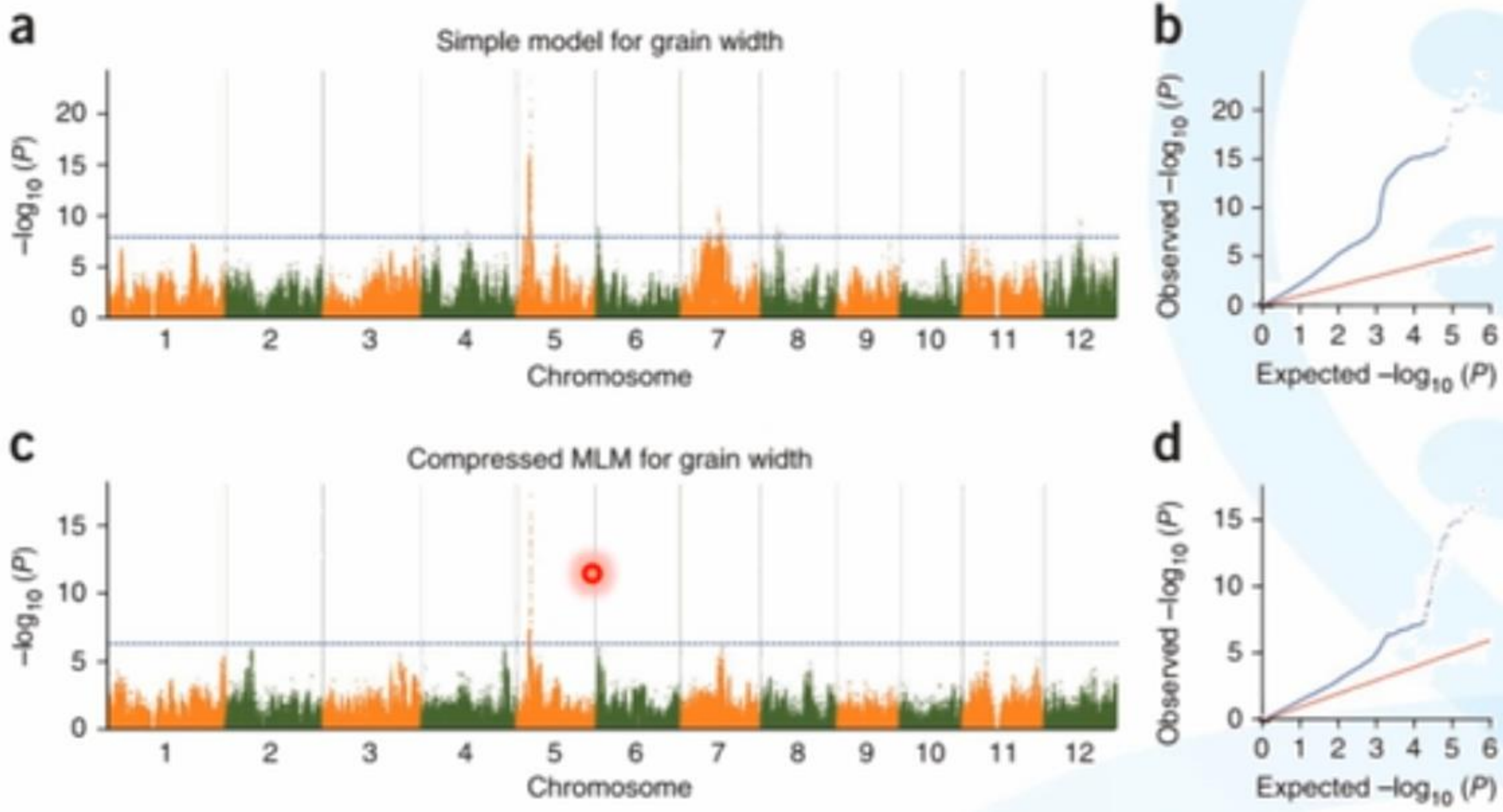
MLM结果



Marker Statistics - mdp_genotype + Filtered_mdp_traits + Filtered_mdp_population_structure

Trait	Marker	Chr	Pos	df	F	p	add_effect	add_F	add_p	dom_effect	dom_F	dom_p	errordf	MarkerR2	Genetic Var	Residual Var	-2LnLikelihood
dpoll	None			0	□	□	□	□	□	□	□	□	257	□	8.068	14.585	1,477.183
dpoll	PZB008...	1	157104	2	1.10	0.333	-0.022	0.004	0.949	-3.627	2.168	0.142	253.0	0.007	8.068	14.585	1,477.183
dpoll	PZA0127	1	1947984	1	4.33	0.038	□	□	□	□	□	□	248.0	0.015	8.068	14.585	1,477.183
dpoll	PZA0361	1	2914066	1	0.13	0.716	□	□	□	□	□	□	255.0	0	8.068	14.585	1,477.183
dpoll	PZA0361	1	2914171	2	3.44	0.033	0.63	3.066	0.081	-5.839	4.043	0.045	258.0	0.02	8.068	14.585	1,477.183
dpoll	PZA0361	1	2915078	1	0.04	0.834	□	□	□	□	□	□	243.0	0	8.068	14.585	1,477.183
dpoll	PZA0361	1	2915242	1	0.78	0.375	□	□	□	□	□	□	241.0	0.00	8.068	14.585	1,477.183
dpoll	PZA0025	1	2973508	2	0.70	0.495	-0.286	0.751	0.387	3.072	0.616	0.433	242.0	0.005	8.068	14.585	1,477.183
dpoll	PZA0296	1	3205252	2	1.74	0.177	0.439	0.805	0.37	3.662	2.18	0.141	247.0	0.01	8.068	14.585	1,477.183
dpoll	PZA0296	1	3205262	1	0.02	0.873	□	□	□	□	□	□	239.0	0	8.068	14.585	1,477.183
dpoll	PZA0059	1	3206090	2	0.31	0.733	0.37	0.623	0.431	-0.242	0.004	0.951	258.0	0.00	8.068	14.585	1,477.183
dpoll	PZA0212	1	3706018	1	1.0	0.307	□	□	□	□	□	□	255.0	0.00	8.068	14.585	1,477.183

Manhattan map



GWAS fine positioning

- The SNP is only a marker. The results of GWAS are statistically significant but not necessarily biologically significant. Therefore, after finding some sites that have passed the correction line, it is necessary to see which regions the sites fall in and extract the genes from these regions. Further filtering to determine candidate genes, where the region is determined, is mainly two methods: 1. A certain interval of up and down 0.2. The LD Block in which it is located. Filtering genes is to see if functional annotations and other things are related to your traits. If they are irrelevant, they can be filtered out.

Thank you