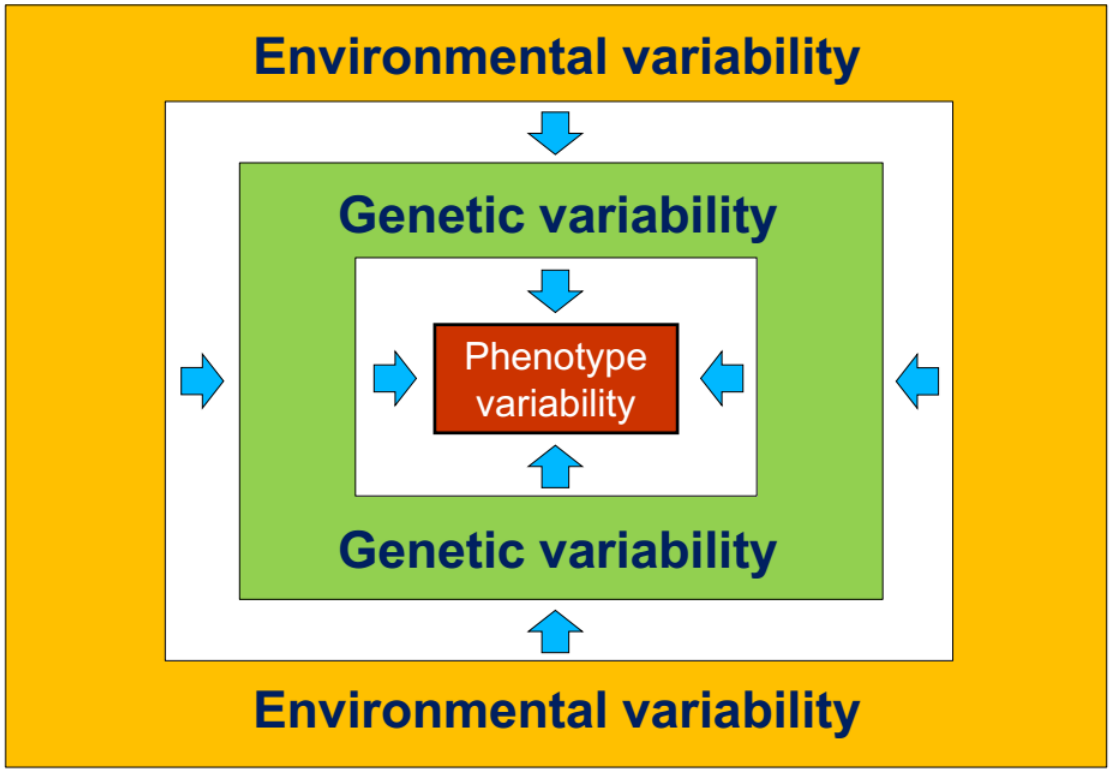
博學至精 明德至善

全基因组关联研究

Genome-wide association study

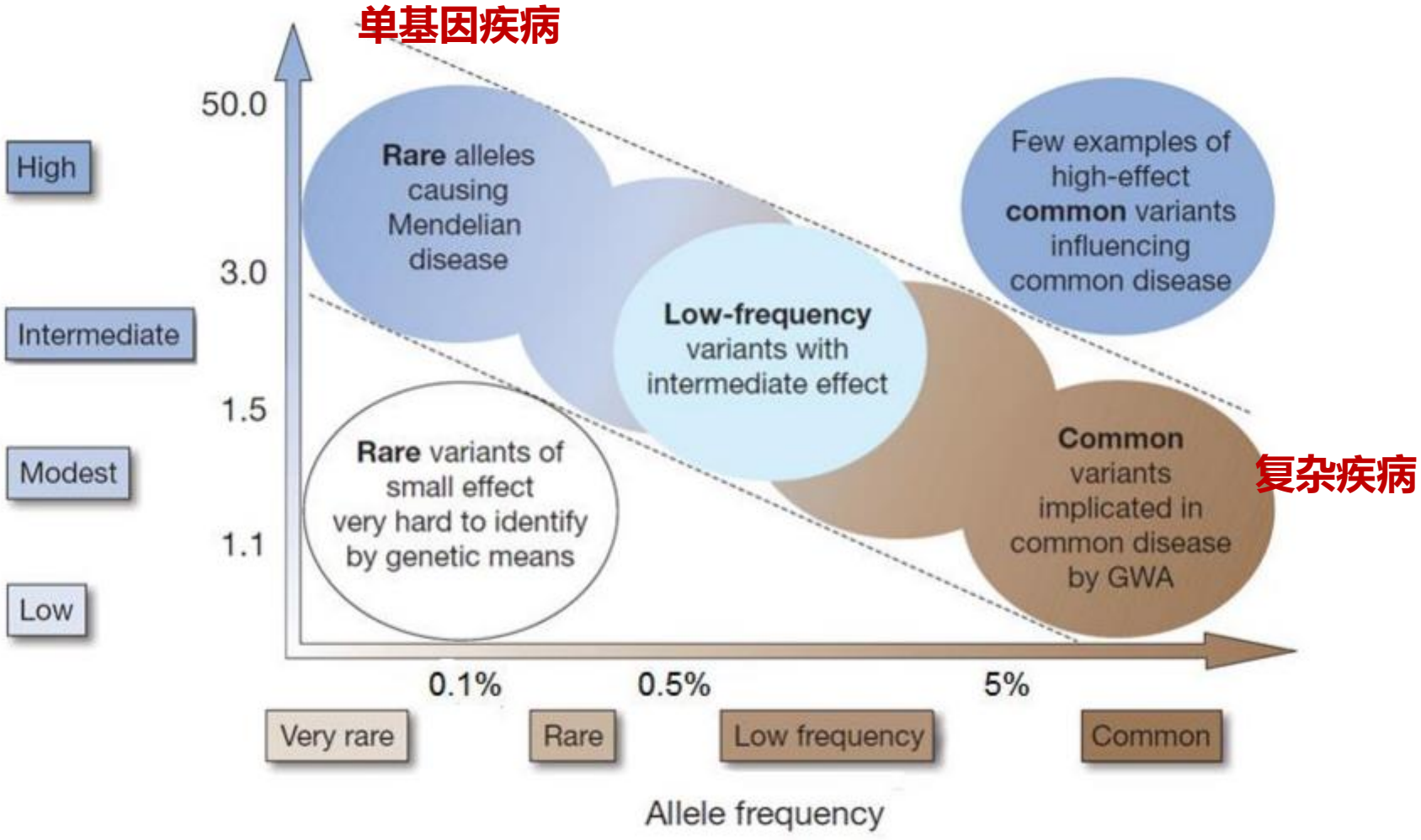
生物信息学系 段巍巍
passion@njmu.edu.cn

疾病的致病因素



哪些遗传变异可以影响疾病的发生?

疾病的致病因素


单基因疾病
50.0
3.0
1.5
1.1
High
Intermediate
Modest
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
复杂疾病
0.1%
0.5%
5%
Very rare
Rare
Low frequency
Common
Allele frequency

遗传关联分析(genetic association study)

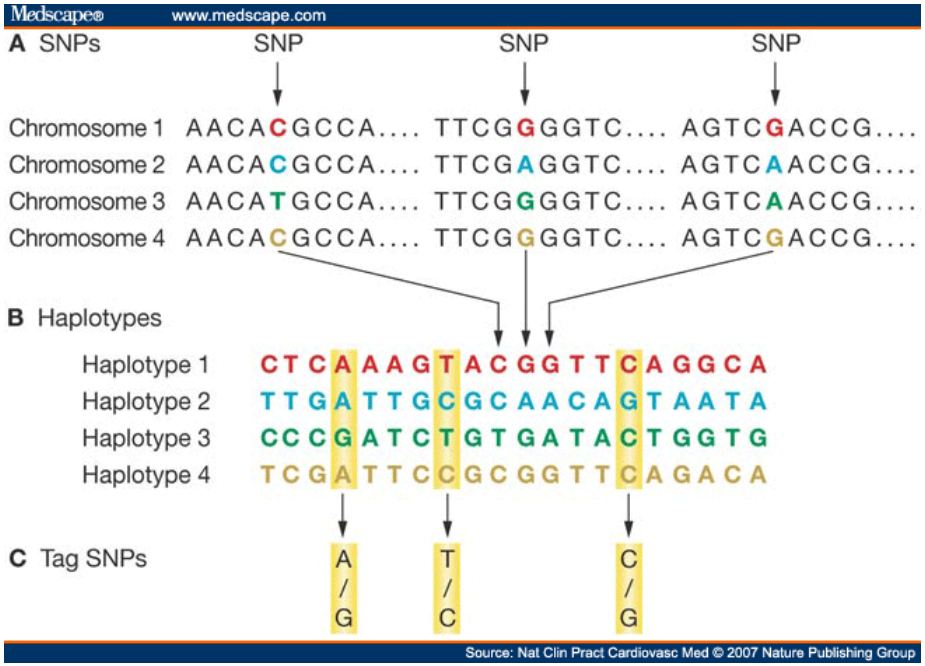
关联分析面临的两个问题：

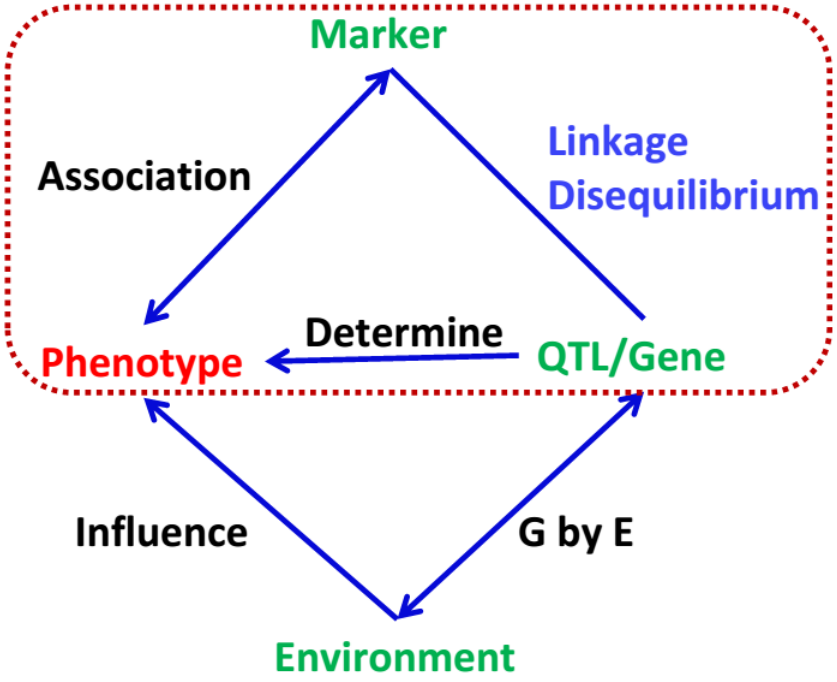
- 合适的遗传标志物　　**1000w个SNPs，怎么选?**
- 经济有效的研究/检测策略

单体型(haplotype)：相邻SNPs的等位位点倾向于以一个整体遗传给后代; 位于染色体上某一区域的一组相关联的SNP等位位点

标签SNP(tagSNP)：一个染色体区域可以有很多SNP位点，能代表其他位点信息的SNP位点称为标签SNP；用少数几个标签SNPs，就能够提供该区域内大多数的遗传多态模式；50万个较常见的SNP，基本上代表了1000万个SNP

遗传关联分析(genetic association study)



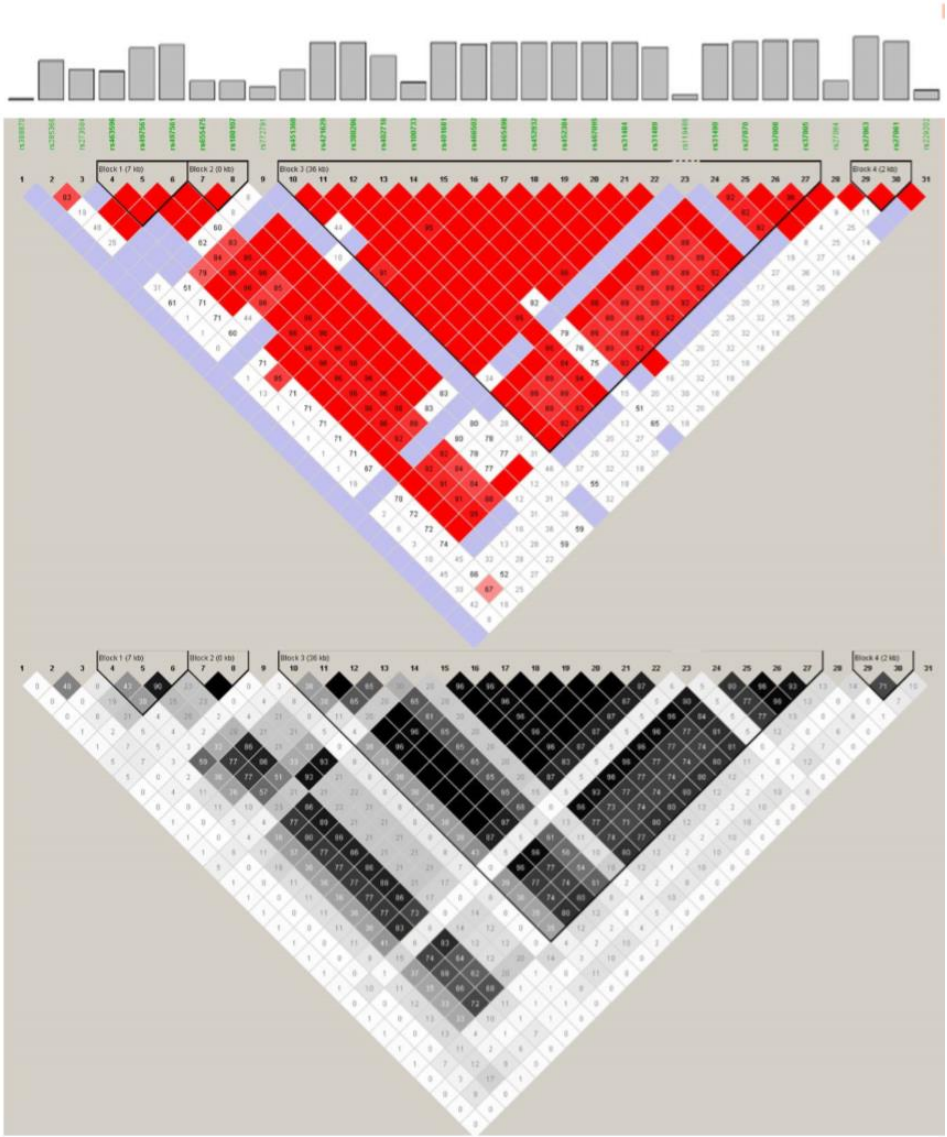


遗传关联分析(genetic association study)

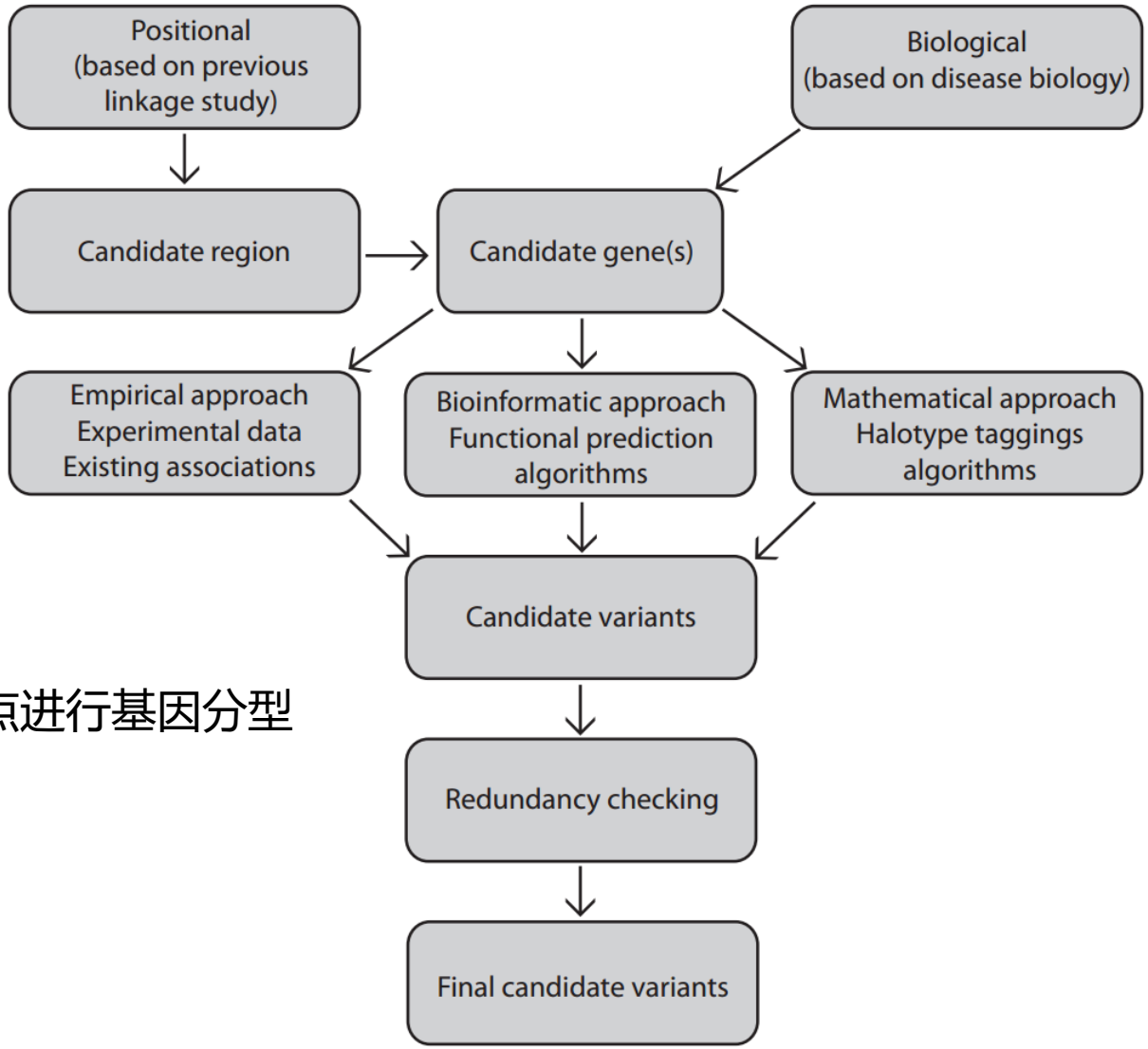
连锁不平衡

CLPTM1L ±20kb

Chr 5： 1,371,007
1,398,002



常用研究策略：候选基因策略



对感兴趣的少数位点进行基因分型

常用研究策略：全基因组策略

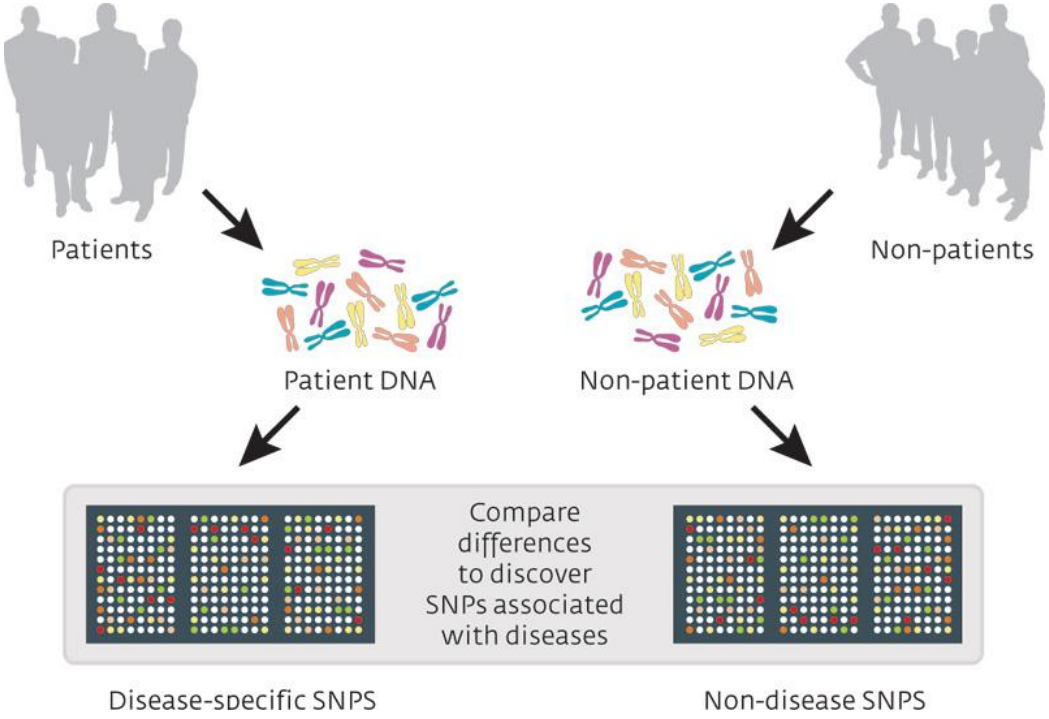
近年来，基因分型技术不断进步，分型成本显著降低，以基因芯片技术为代表的超高通量分型技术更是得到了飞速的发展；

全基因组测序商业化和公司之间的竞争使得基因组测序成本越来越低；

全基因组关联研究的方法学（如研究设计、统计分析、结果的解释）也取得了极大的进步。

全基因组关联研究(GWAS)

全基因组关联研究（Genome-Wide Association Studies，GWAS）是指在全基因组层面上，开展多中心、大样本、反复验证的基因与疾病的关联研究，是通过对大规模的群体DNA样本进行全基因组高密度遗传标记（如SNP或CNV等）分型，从而寻找与复杂疾病相关的遗传因素的研究方法，全面揭示疾病发生、发展与治疗相关的遗传基因。



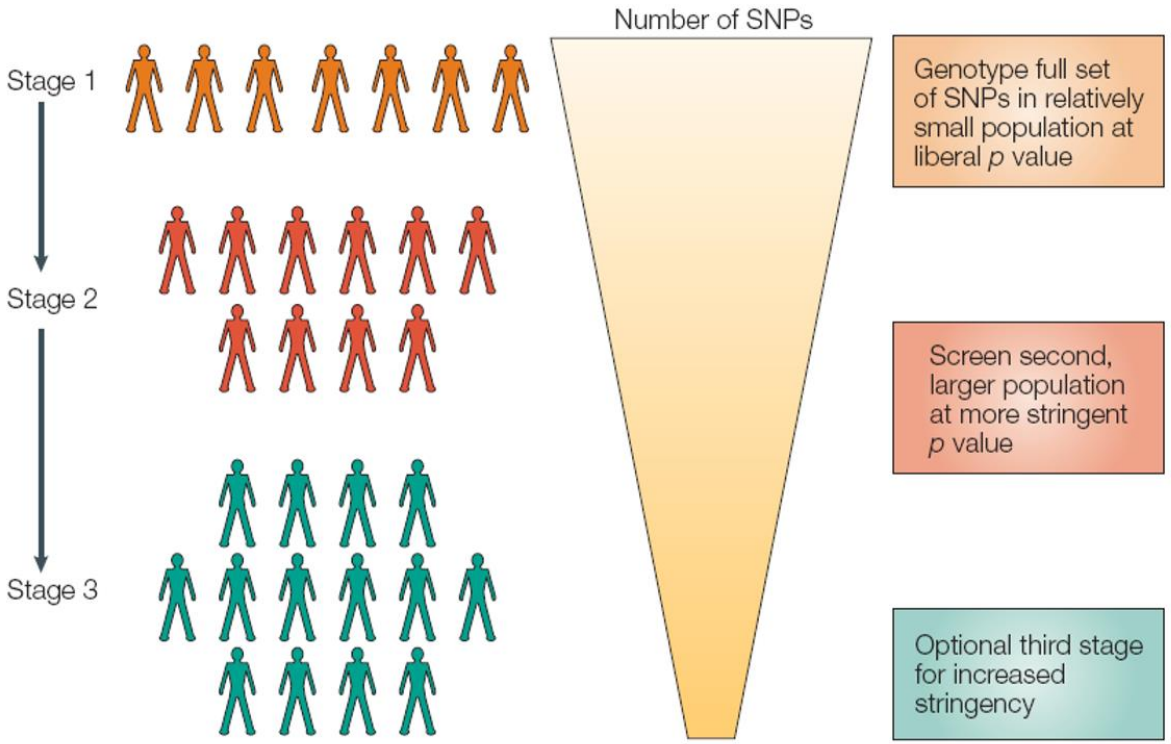
SNPs选择：全基因组tagSNPs + 随机 + 前人结果
分型数目：≥50w

GWAS的研究类型

GWAS目前分为单阶段研究和多阶段研究：

单阶段研究即选择足够的样本，一次性在所有研究对象中对选中的SNP进行基因分型,然后分析每个SNP与疾病的关联，在早期GWAS多使用；

多阶段研究多为两阶段研究。



两阶段GWAS

第一阶段的分析可以是以个体为单位，也可以采用DNA pooling的方法，筛选出较少量的阳性SNP

后者简单，但误差大，其估计的等位基因的频率标准差在1%—4%之间，对检验效能有重要影响

Published: 01 November 2002

DNA Pooling: a tool for large-scale association studies

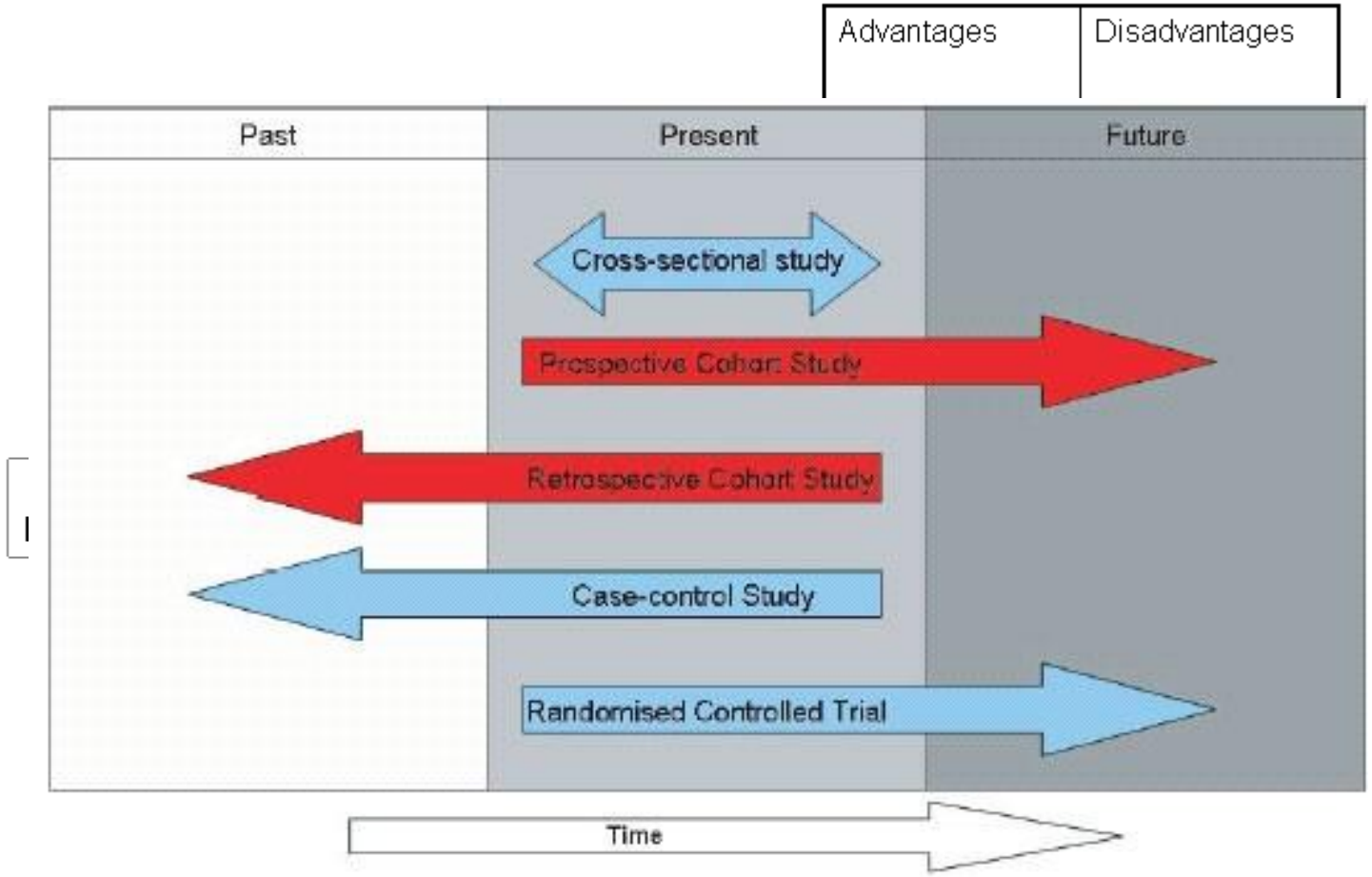
Pak Sham ✉, Joel S. Bader, Ian Craig, Michael O'Donovan & Michael Owen

Nature Reviews Genetics **3**, 862–871(2002) | Cite this article

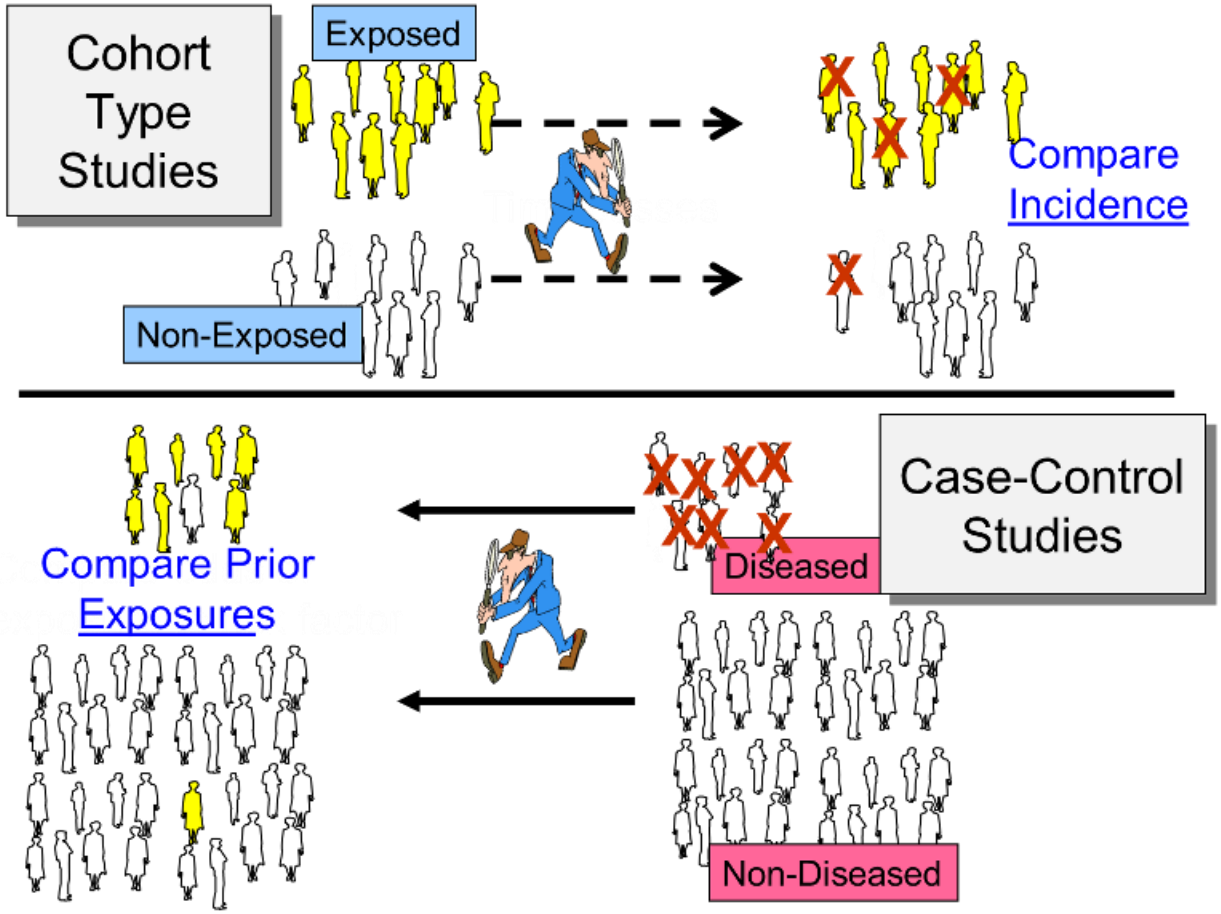
第二阶段采用更大的样本对第一阶段筛选出的阳性SNP进行分析

应用大样本人群甚至在多种人群中进行基因分型验证

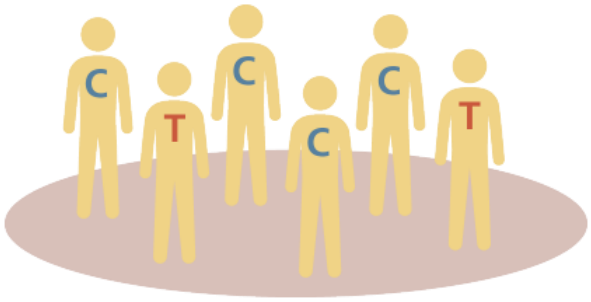
常用研究设计


Advantages
Disadvantages
Past
Present
Future
Cross-sectional study
Prospective Cohort Study
Retrospective Cohort Study
Case-control Study
Randomised Controlled Trial
Time

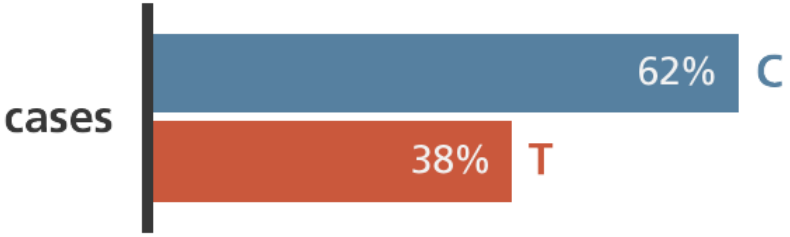
常用研究设计

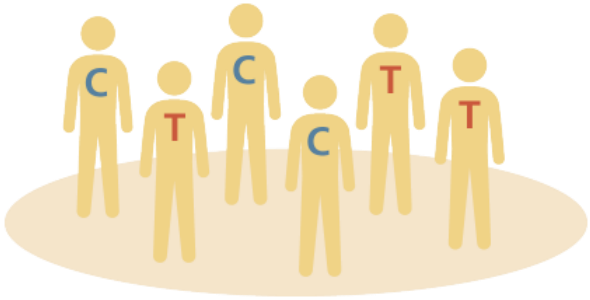

Cohort Type Studies
Exposed
Non-Exposed
Compare Incidence
Case-Control Studies
Diseased
Non-Diseased
Compare Prior Exposures

常用研究设计



cases (n=1,000)
people with heart disease



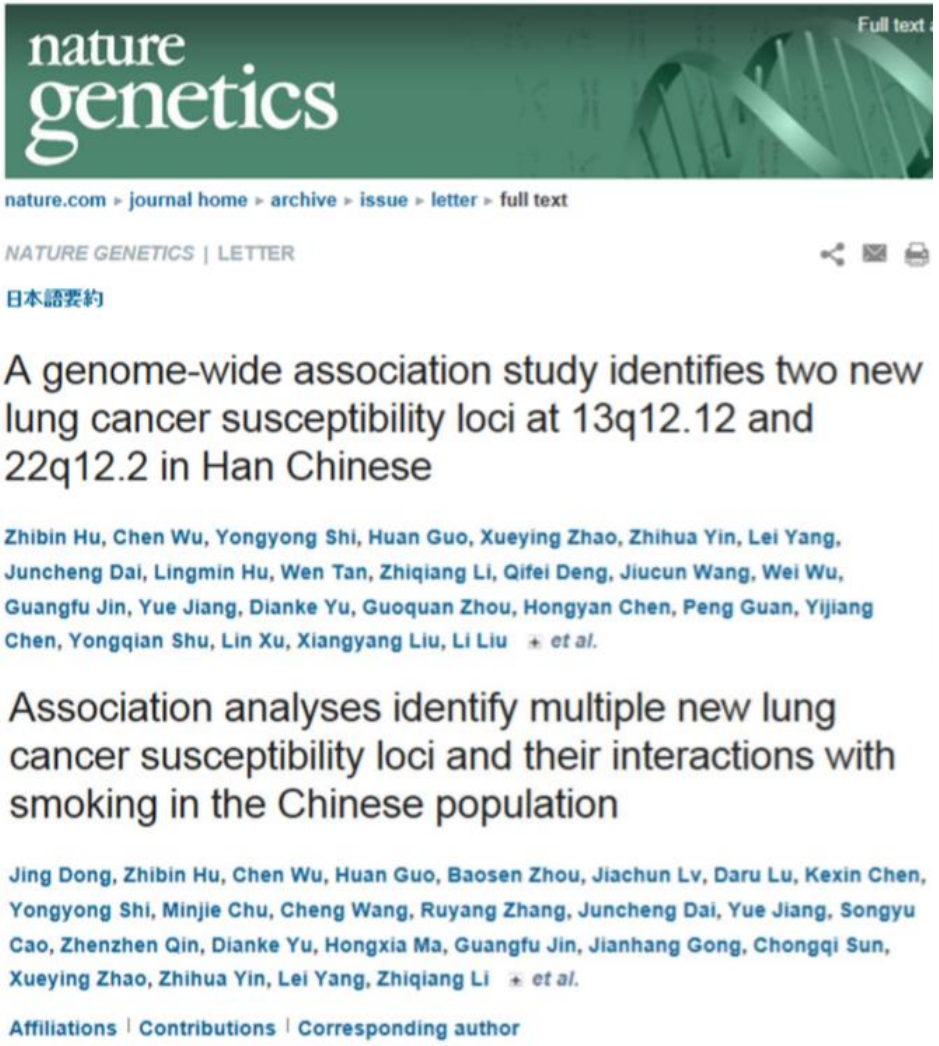


controls (n=1,000)
people without heart disease

常用研究设计

病例–对照设计

“To identify genetic factors that modify the risk of lung cancer in individuals of Chinese ancestry, we performed a genome-wide association scan in 5,408 subjects (2,331 individuals with lung cancer (cases) and 3,077 controls) followed by a two-stage validation among 12,722 subjects (6,313 cases and 6,409 controls). ”



GWAS的分析流程

样本收集： 取得表型和种族数据，采集种族及其他参数匹配的对照样本，准备数据处理和分析系统

样本准备： 高质量的DNA，采用尽盘减少批间差异的DNA抽提方法，将病例和对照随机分配到板上

分型后检测：去掉数据质量较差的芯片，检测基因型和表型性别的符合性，去掉少见SNP，去掉数据质量低的SNP，去掉分型质量低的样本

GWAS的分析流程

质量控制： 在对照中检测Hardy-Weinberg平衡，通过病例对照状态检测系统错误，通过SNP和基因型检测系统错误，检测亲属关系，修正人群分层

关联分析后检测：留意非常显著的P值，使用连锁不平衡的SNP检测关联，检查原始荧光数据

验证：在独立人群样本中检测关联性，和（或）通过测序、组织样本或细胞株或基因敲除（入）动物进行功能提示实验

GWAS的分析流程

质量控制非常重要，数据的异常就会使得后面的分析都丧失意义：

1. 大量SNP位点不能成功分型的个体应该剔除，因为这意味着DNA样本可能存在问题
2. 对照样本中需要计算Hardy Weinberg平衡
3. 病例样本中基因分型丢失比对照中丢失更频繁的系统错误
4. 使用IBD 评价来处理重复样本（或同卵双胞胎样本）或关联个体的全基因组 SNP 数据。

GWAS的分析流程

哈迪-温伯格平衡(Hardy-Weinberg Equation)：在较大、随机婚配的人群中，如无突变、自然选择和迁移等影响，基因频率基本保持不变。

检验哈迪-温伯格平衡来评价位点检测的质量和样本的代表性。

设A(突变)和a(野生)分别表示一个基因的等位基因；$P(A) = p$, $P(a) = q = 1 - p$，若满足HW平衡，则有基因型：

$$P(AA) = p^2,\ P(Aa) = 2pq,\ P(aa) = q^2$$

GWAS的分析流程

HW平衡定律：

若亲代基因型具有HW平衡比例，则在随机婚配下， 子代也具有HW平衡比例；

若亲代基因型不具有HW平衡比例，在随机婚配下， 则子代将具有HW平衡比例;

GWAS的分析流程

HW平衡的检验

Genotype	AA	Aa	aa	Total
Observed(O)	n_{AA}	n_{Aa}	n_{aa}	N
Expected(E)	$N\hat{p}^2$	$2N\hat{p}\hat{q}$	$N\hat{q}^2$	N

$$\hat{p} = \frac{2n_{AA} + n_{Aa}}{2N}, \qquad \hat{q} = \frac{2n_{aa} + n_{Aa}}{2N}$$

则有

$$\chi^2 = \sum \frac{(O-E)^2}{E} \sim \chi^2_{v=1}$$

GWAS的分析流程

高血压病的遗传流行病学调查，197人的DNA样本，ACE位点上三个基因型分别为AA,Aa和aa，人数为26,93,78。

$$\hat{p} = 0.368 \quad \hat{q} = 0.632$$

$$\hat{E}_{AA} = 26.68$$

$$\hat{E}_{Aa} = 91.63$$

$$\hat{E}_{aa} = 78.69$$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 0.0439$$

则P = 0.834，认为满足HW平衡

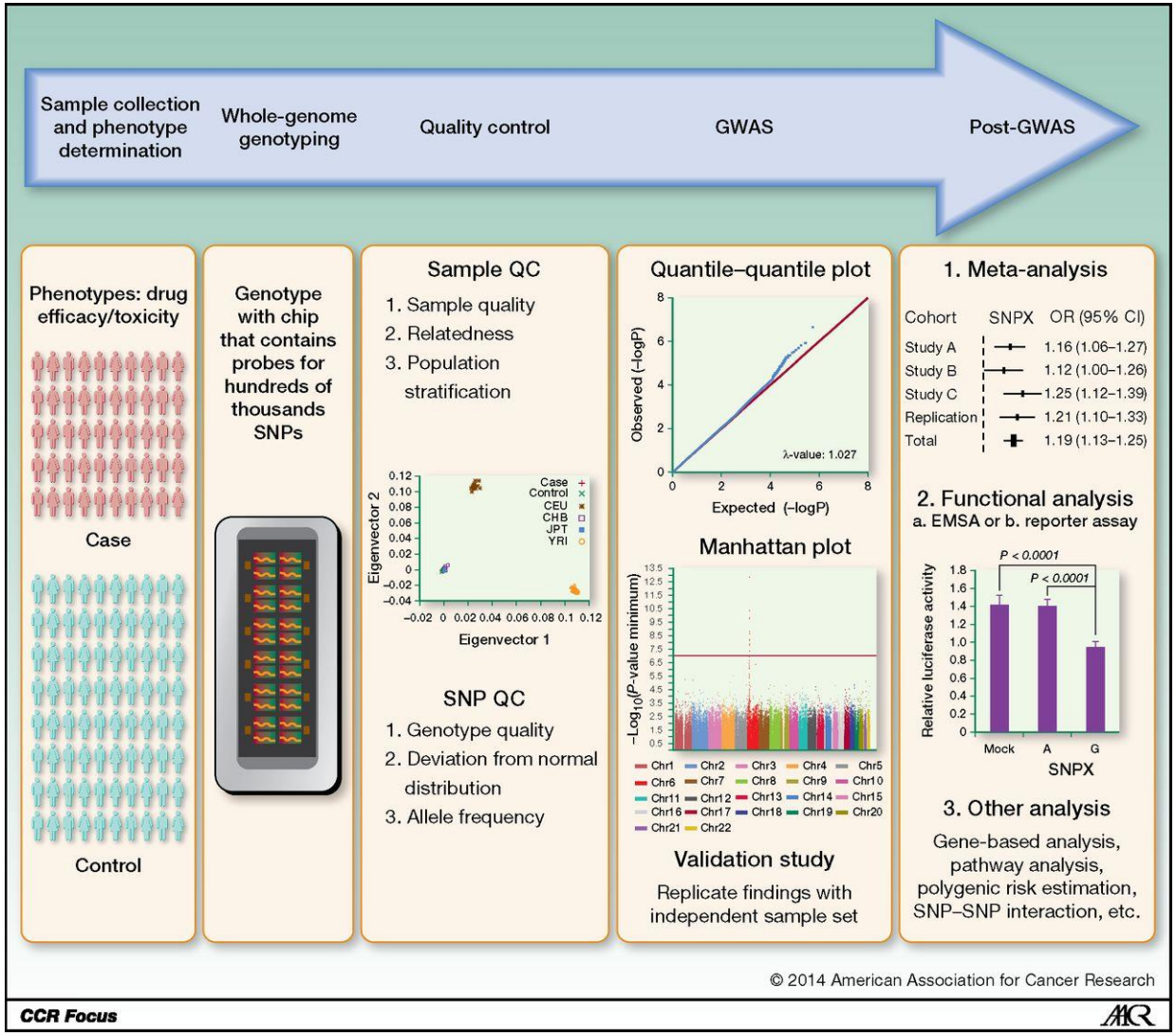
GWAS的分析流程

偏离HW平衡的可能原因：

- 基因分型错误
- 选择性配对，如inbreeding
- 选择偏移
- 群体分层
- 偶然因素

在GWAS中，通常删除HW平衡检验P<1e-6的SNP位点

GWAS的分析流程



GWAS芯片：

Affymetrix公司

“Genome-Wide Human SNP Array 6.0”

Illumina公司

“Human Omni5 BeadChip”

GWAS数据的统计分析

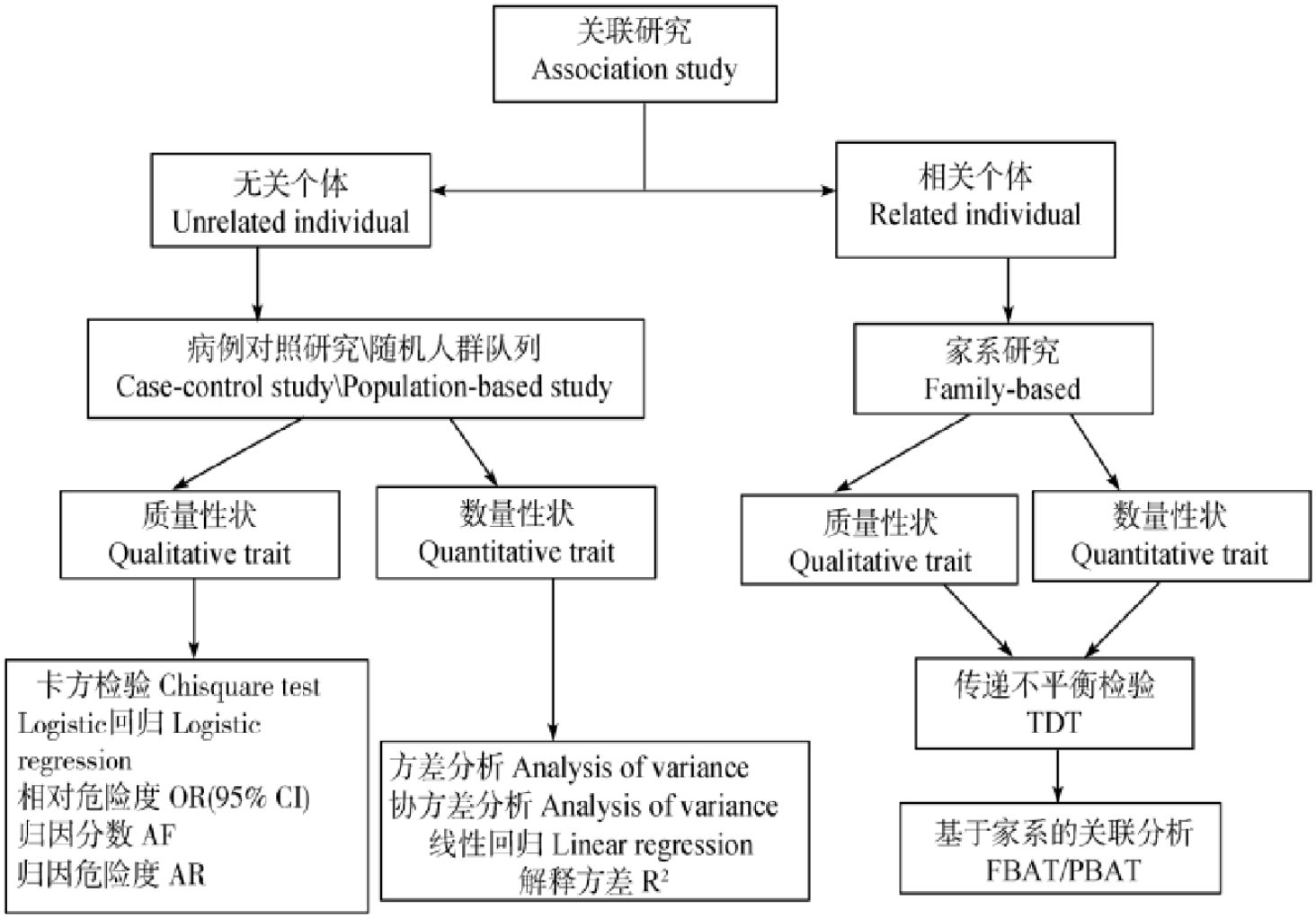
基因型

ID	表型	1	2	…	p
1	case	0	1	…	0
2	control	1	0	…	0
3	case	0	1	…	1
4	case	0	0	…	0
…	…	…	…	…	…
n	control	2	0	…	1

Genotypes are measured at p SNPs, and are coded as 0, 1 or 2 copies of a reference allele

通常有p>>n

GWAS数据的统计分析



GWAS数据的统计分析

单位点分析策略

思考：可以将所有位点放入模型吗?

$$for\ j = 1{:}p,$$

连续表型： $Y = X\alpha + G_j\beta_j + \varepsilon$

二分类表型： $\mathrm{logit}P = X\alpha + G_j\beta_j$

相加模型Additive Model：0, 1, 2
显性模型Dominant Model：2, 1 vs. 0
隐性模型Recessive Model：2 vs. 1, 0
共显性模型Codominant Model：1 vs. 0; 2 vs. 0

如何选择?

GWAS数据的统计分析

曼哈顿图
(Manhattan plot)

GWAS数据的统计分析

多重比较(multiple comparison)校正

狭义多重比较：事后两两比较，如方差分析后两组间比较

广义多重比较：存在于多个假设检验

GWAS中p个位点对应p次独立的假设检验，则至少出现一次I类错误的概率为：$1-(1-\alpha)^p$

为了控制总的I类错误，每次检验的$\alpha = 0.05$是否合适?

GWAS数据的统计分析

Bonferroni方法

- 如果$P_j \leq \alpha/p$, 则拒绝第j个SNP位点对应的H_0假设
- 严格、保守

FDR(false discovery rate)方法

- 对p个位点的P值，即$P_1, P_2, \ldots, Pp$从小到大进行排序：$P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(p)}$
- 调整后$\widetilde{P}_j = \min\limits_{k=j,\ldots,p} \left\{ min\left(\frac{p}{k} P_{(k)}, 1\right)\right\}$，若$\widetilde{P}_j \leq \alpha$，则拒绝该位点的$H_0$假设

Permutation方法

GWAS数据的统计分析

- 如果采用较为宽松的多重假设检验方法就可能导致 I 类错误的膨胀, 出现大量的假阳性关联 ;
- 但是如果采用最为严格 Bonferroni校正, 则又可能导致过度校正 ,结果使假阴性概率增加 , 而与疾病真正关联的 SNP难以发现

单位点分析：效应弱，多重校正严格(几十万次)

其它策略：多位点分析(Gene-based or SNP set)，再精细定位

GWAS数据的统计分析

事实上，大多数GWAS研究为多阶段设计(即重复验证)，因此第一阶段的初筛并不会严格控制总I类错误，因为串联的设计本身就可以控制最终发现位点的假阳性率

因此：GWAS不能仅凭 *P* 值判断某个 SNP 是否与疾病真正关联，多种族、多群体、大样本的重复验证研究(replication)才是提高检验效能、确保发现真正疾病关联SNP的关键 。

GWAS的检验效能和样本量

此步骤应该在研究设计时确定！！

- 大多数基因在复杂疾病中对疾病风险的增加作用甚微，要获得高的效能需要大样本

- 例如对一个10%的频率的SNP，乘法模型下杂合体相对风险为1.3，如果在没有进行多种检验校正的条件下达到0.05的显著性水平，至少需要1146 个实验和对照个体才可以获得80%的效能

- 多个研究的Meta分析

Published: 01 September 2014

Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations

Chen Wu, Zhaoming Wang, [...] Stephen J Chanock

Nature Genetics **46**, 1001–1006(2014) | Cite this article

GWAS的检验效能和样本量

About GAS Power Calculator

This Genetic Association Study (GAS) Power Calculator is a simple interface that can be used to compute statistical power for large one-stage genetic association studies. The underlying method is derived from the CaTS power calculator for two-stage association studies (2006).

Inputs

Sample Size

Cases/Controls = 1.000

Cases		1000
Controls		1000

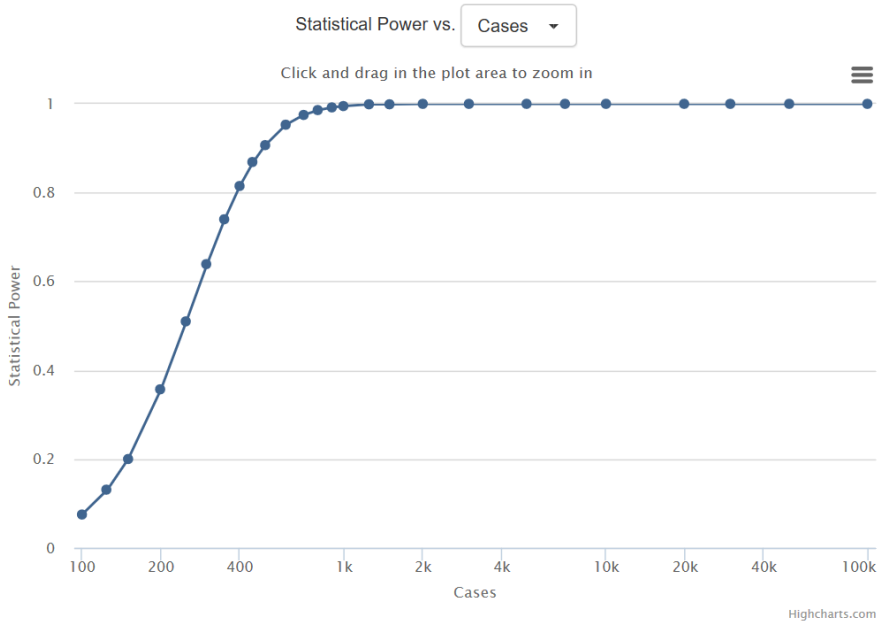
Study Design

Significance Level		0.0000070

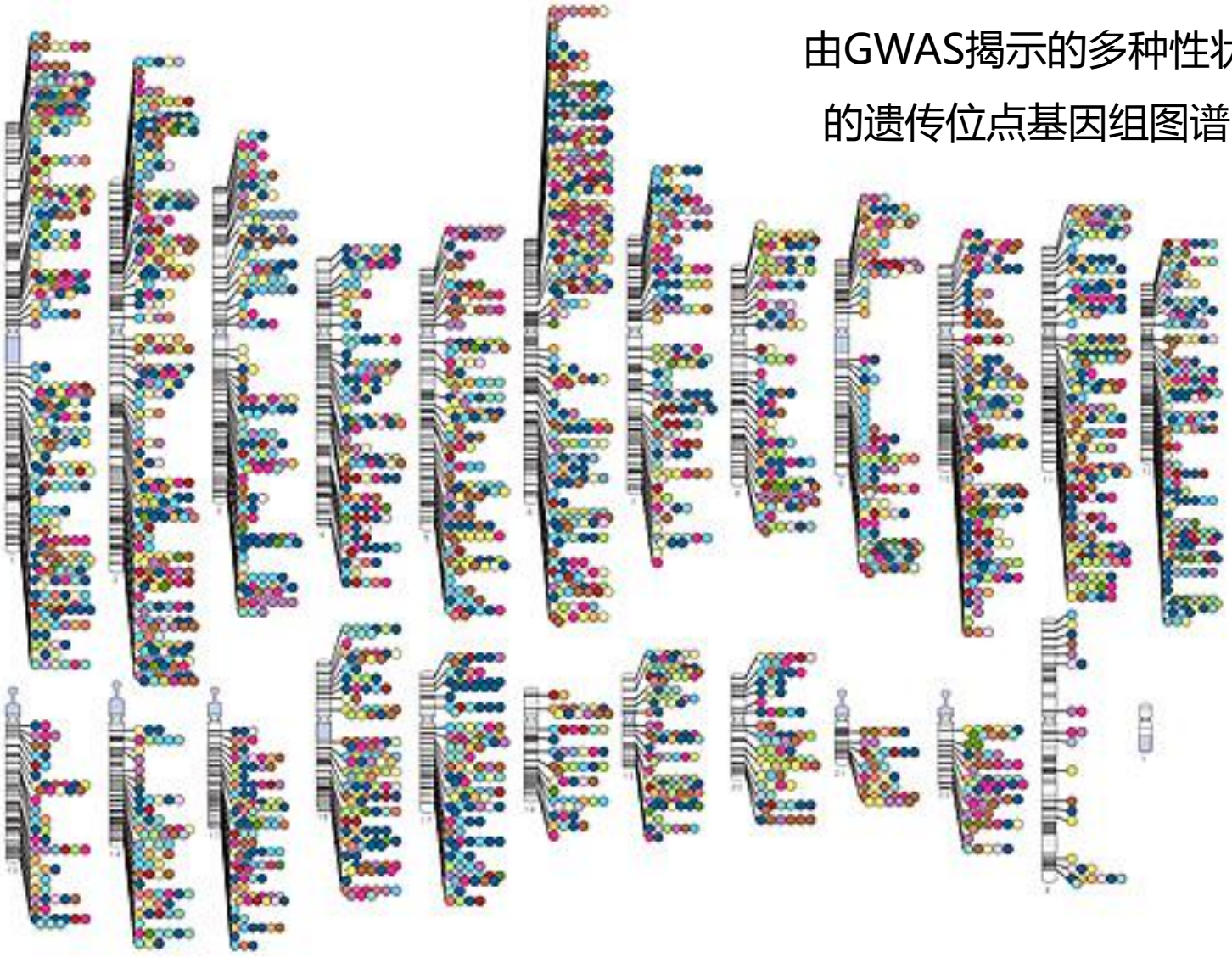
Disease Model: Multiplicative

Prevalence		0.1000
Disease Allele Frequency		0.5000
Genotype Relative Risk		1.5000

Graph



全基因组关联研究的结果



由GWAS揭示的多种性状的遗传位点基因组图谱