博學至精 明德至善

全基因组关联研究（2）

Genome-wide association study

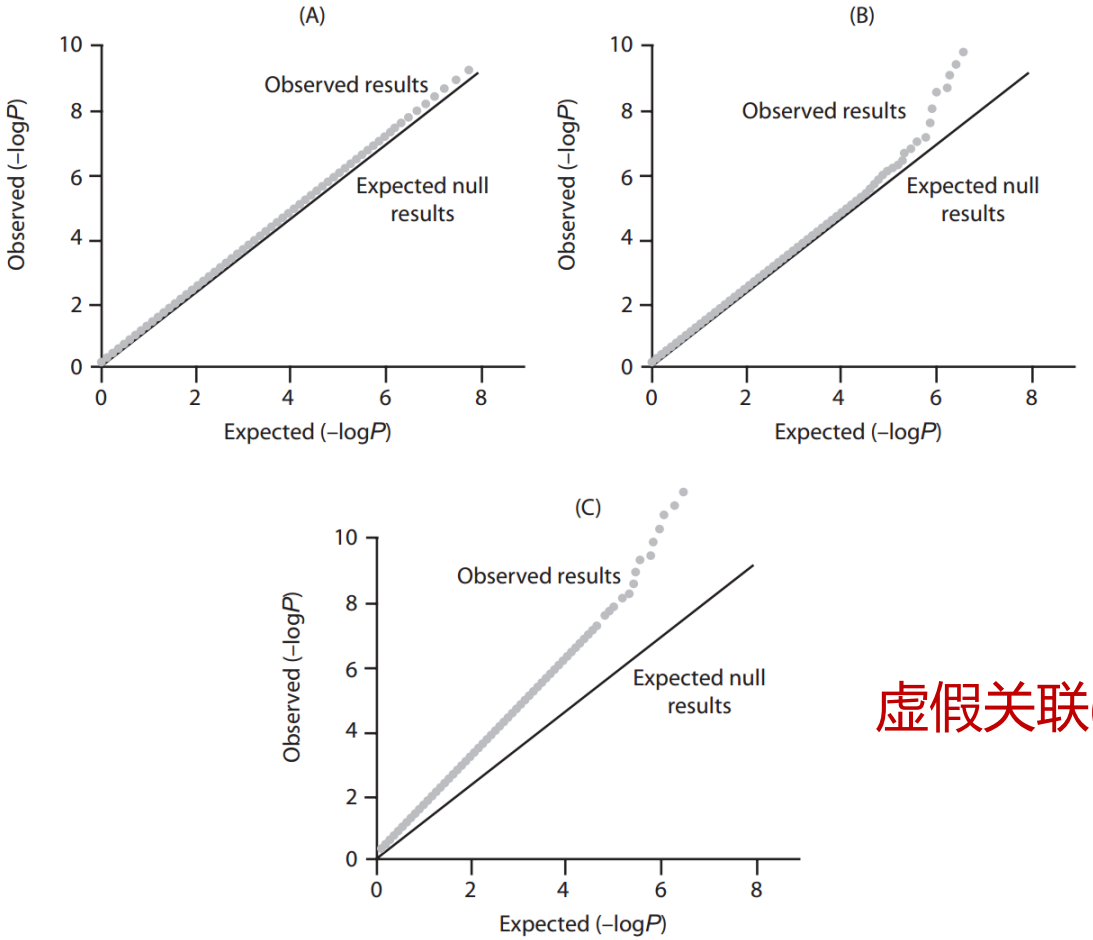
生物信息学系 段巍巍
passion@njmu.edu.cn

GWAS vs. 候选基因研究

Features	Candidate gene study	GWAS
Hypothesis-driven	Yes (based on linkage results or biology)	Hypothesis-free (agnostic)
Sample size	Generally relatively small (hundreds)	Generally large (thousands)
Coverage	Selected gene(s)	Whole genome
SNPs	Fewer than 1000	Up to 5 million
CNVs	May be included	Included in most
STRs	May be included	Not included
Genotyping	Low-throughput methods or custom microarray	Microarray (GWAS chip)
Quality control tests	Basic	Extensive
Analysis	Less stringent	Very stringent (uses genome-wide statistical significance threshold)
Adjustment for ancestry	Not routinely done	Always incorporated
False-positivity rate	Very high	Very low
False-negativity rate	Low	Potentially high
Replication	Required, but usually in a later study	Always incorporated

人群结构(population structure)

遗传关联研究的结果通常会受到人群结构的影响



期望$P=i/(p+1)$

虚假关联(spurious associations)

人群结构(population structure)

由于随机漂变，不同人群遗传变异位点通常具有不同的等位基因频率

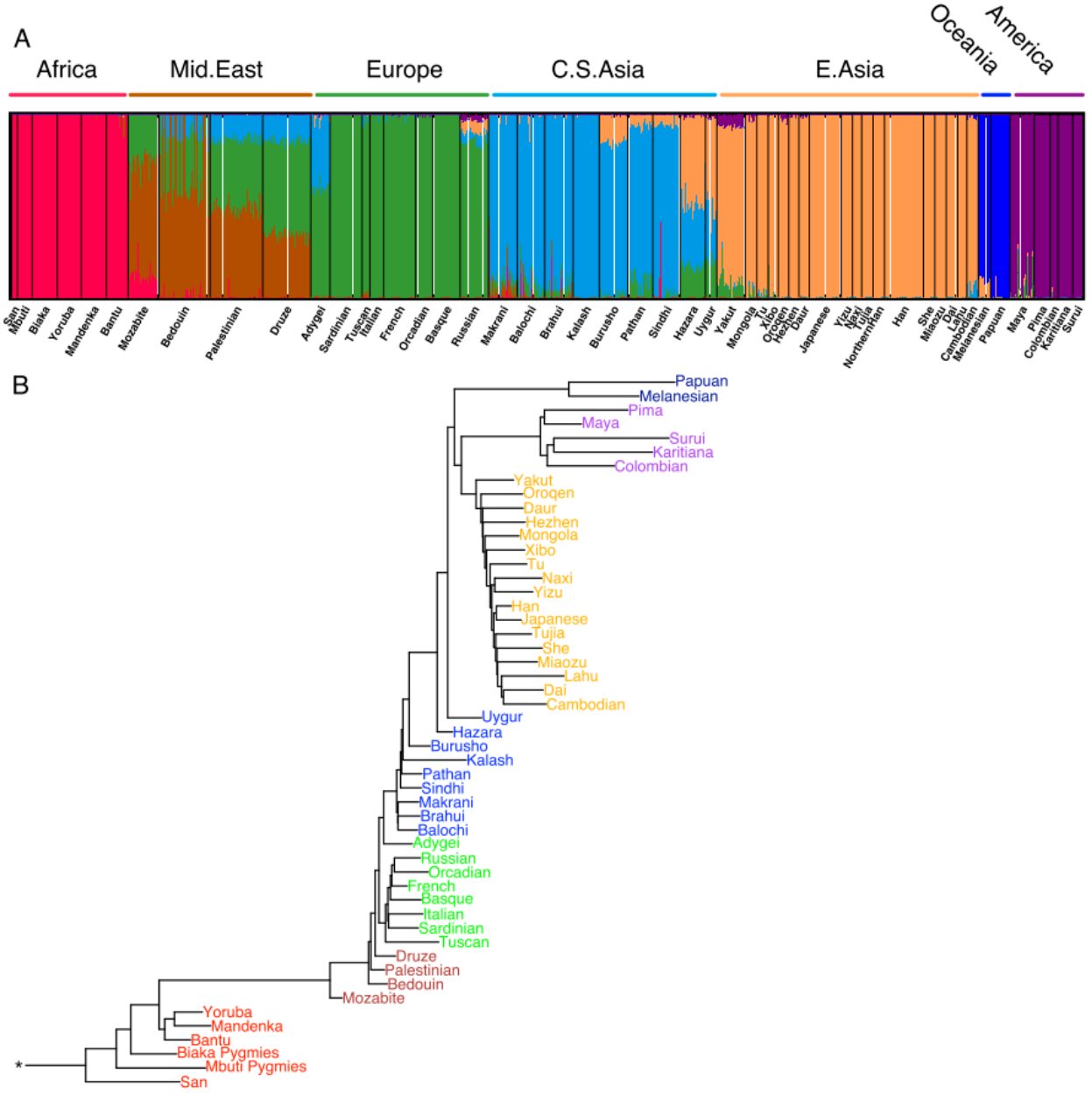
人群结构的两种类型：

- 人群分层(population stratification, PS)
  - remote common ancestry of large groups of individuals

  PS能够影响关联结果的条件：
  - 不同人群的等位基因频率不同
  - 不同人群的疾病患病率/表型分布不同

  混杂因素！
- 隐藏关联(cryptic relatedness, CR)
  - recent common ancestry among smaller groups of individuals

人群结构(population structure)



人群结构(population structure)

GWAS中，人群结构现象更为普遍

检验和校正人群分层

- 使用自报的详细的人种/种族/籍贯信息
- 基因组控制法(genomic control)
- 结构关联(Structured Association)
- 主成分分析(principle components analysis, PCA)
- 混合效应模型

人群结构(population structure)

基因组控制

- 减少人群分层/隐藏关联带来的影响，用于Cochran-Armitage趋势检验下的校正

$$\hat{\lambda} = \frac{median(\chi_1^2, \chi_2^2, \ldots, \chi_m^2)}{0.456}$$

- 如果$\hat{\lambda} > 1$，则提示存在人群结构; $\hat{\lambda} > 1.05$较为严重的人群结构
- $\chi_j^2/\hat{\lambda}$ 即为校正后的趋势卡方统计量，服从自由度为1的卡方分布
- m的选择

人群结构(population structure)

主成分分析

- 个体的遗传背景可以由遗传标志物来表示，而主成分可以综合这些信息
- 相似主成分的个体通常来自于相同亚群

Published: 23 July 2006

Principal components analysis corrects for stratification in genome-wide association studies

Alkes L Price ✉, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick & David Reich

Nature Genetics **38**, 904–909(2006) | Cite this article

人群结构(population structure)

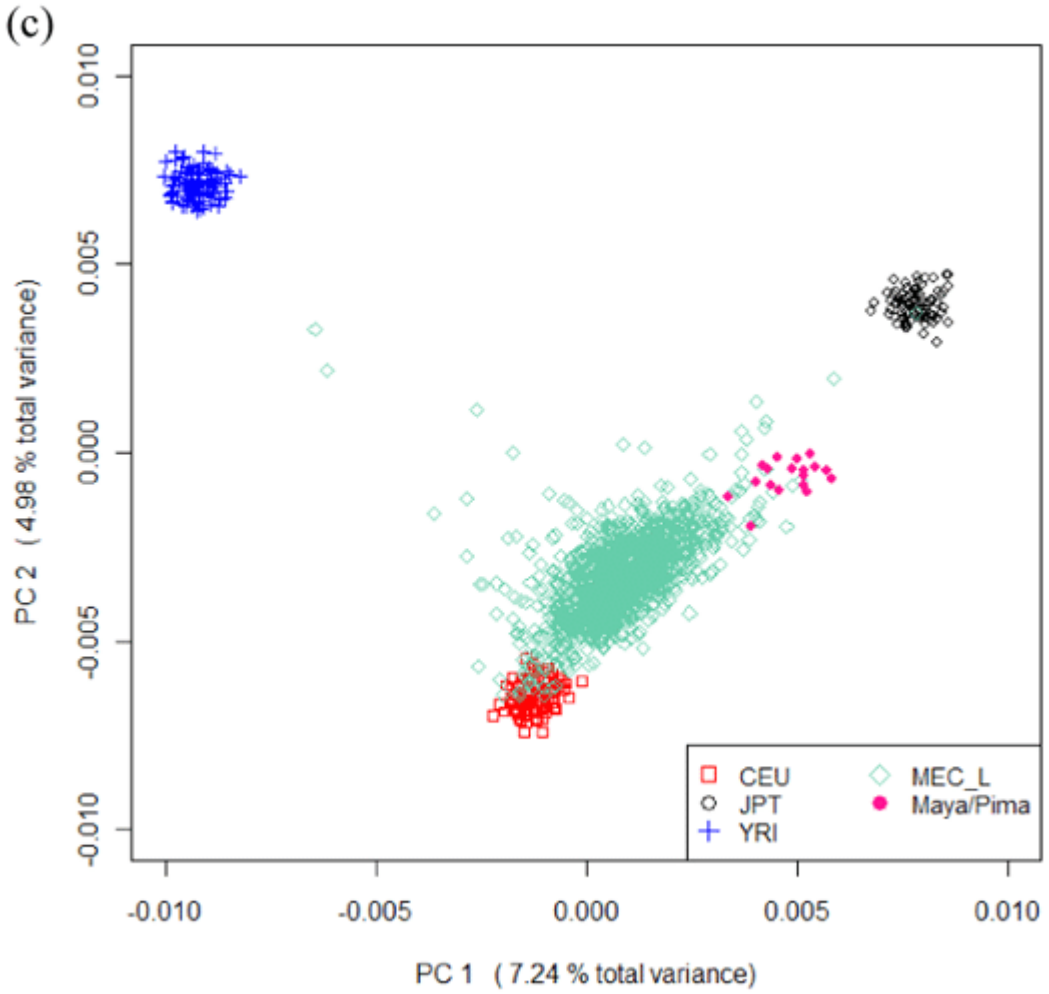
主成分分析

- 设$X_{n\times p}$为标准化后的基因型矩阵

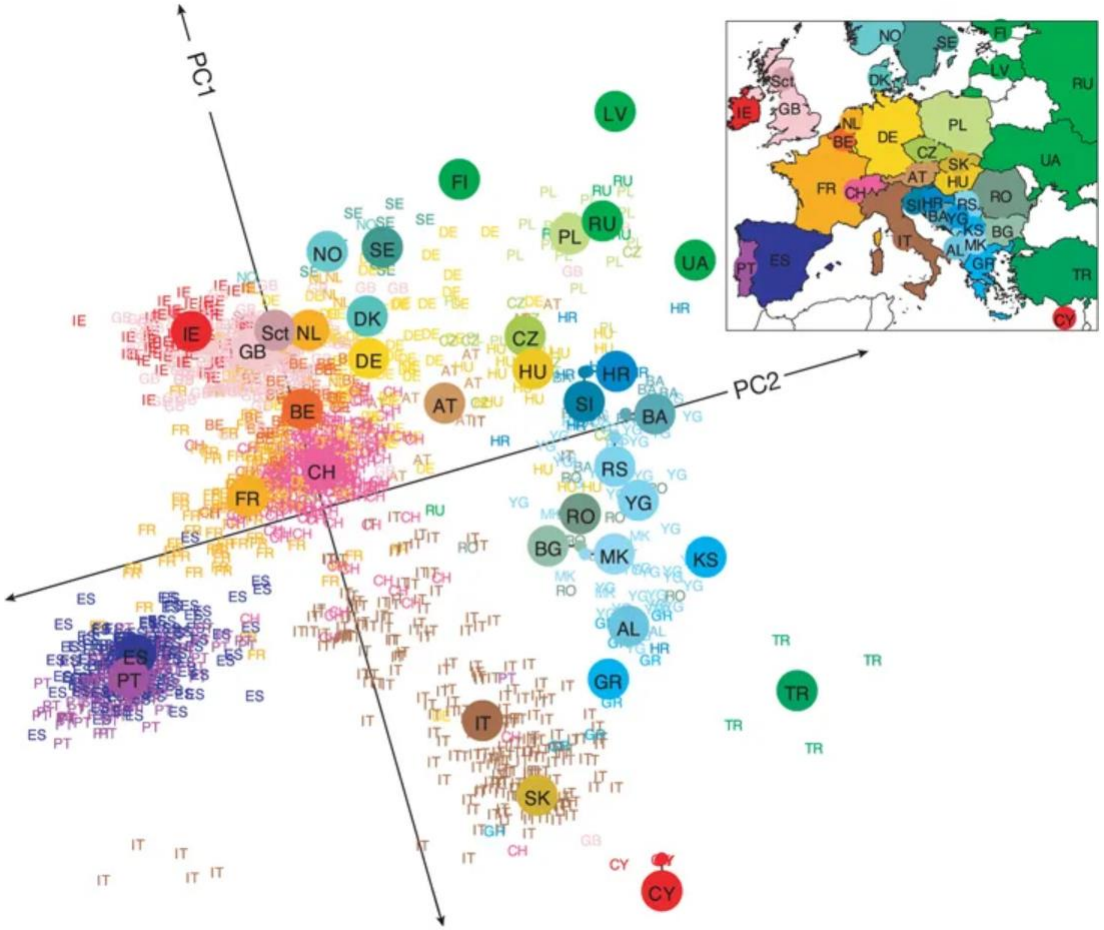
$$\widehat{K} = XX^T/p = \upsilon\Lambda\upsilon^T$$

- 矩阵υ的列即为主成分
- 通常需要事先排除掉高LD的SNP，及有亲缘关系的个体
- 选取前m个PC作为协变量，放入回归模型中调整，m通常取2~15

人群结构(population structure)



人群结构(population structure)



人群结构(population structure)

线性混合效应模型(linear mixed model, LMM) OR 方差成分模型(variance component model)

- 对于每个SNP, $y = \mu + G_j\beta_j + \mathrm{u} + \varepsilon$

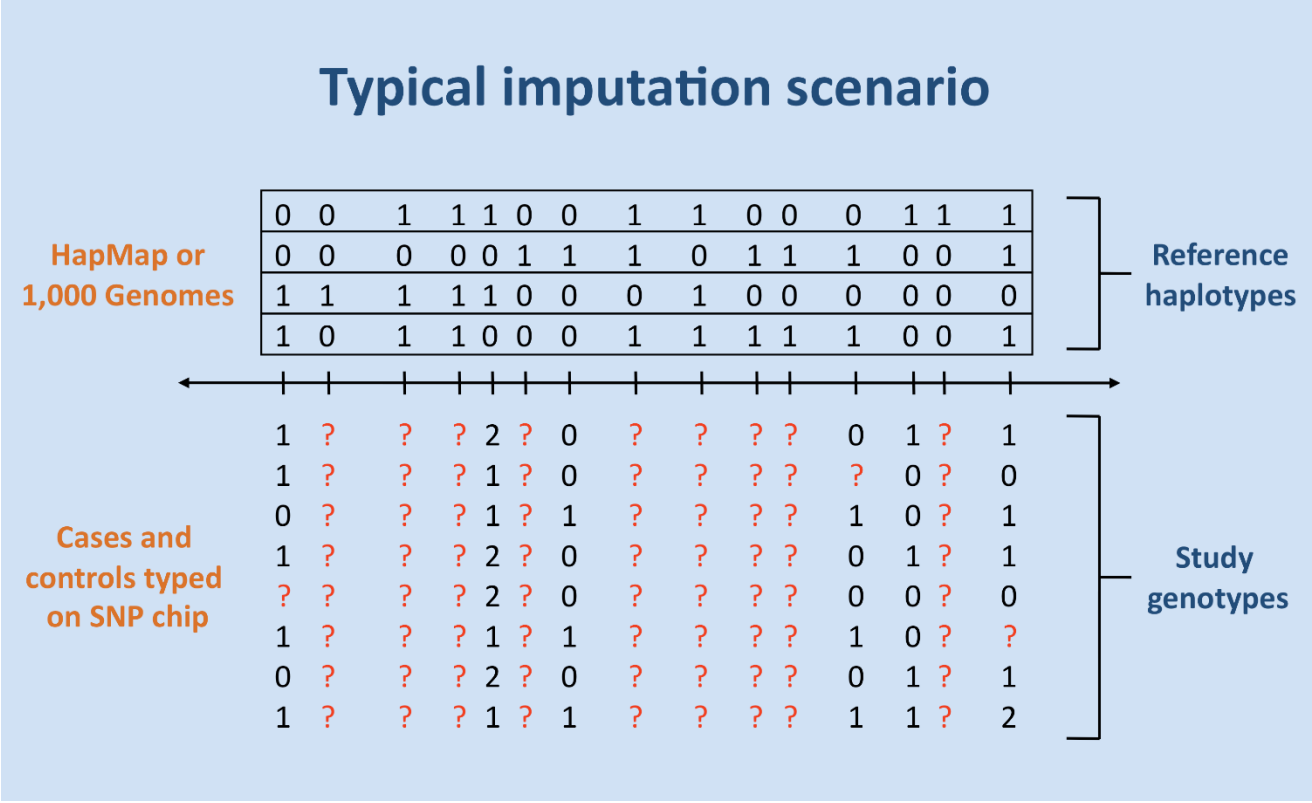
$$\mathrm{u} \sim \mathrm{MVN}_n\left(0, \sigma_g^2 K\right)$$

$$\varepsilon \sim \mathrm{MVN}_n\left(0, \sigma_e^2 I\right)$$

- 似然比检验(LRT)：H_0: β_j=0
- 同时校正人群分层和隐藏关联

基因型填补(genotype imputation)

加密芯片：省钱 + 识别causal位点



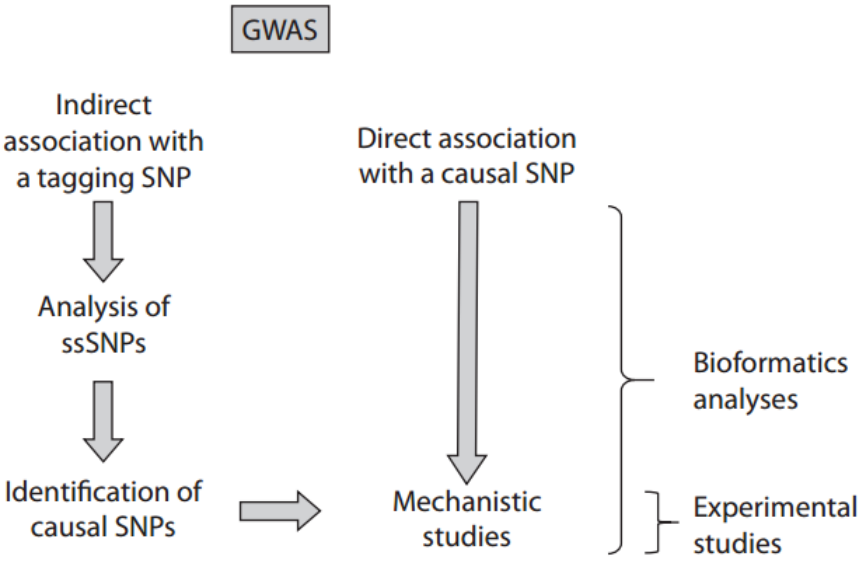
GWAS结果解释

功能注释(functional annotation)

- ✓ 这些位点的功能是什么?

精细定位(fine mapping)

- ✓ causal位点在哪?



统计学关联≠生物学意义

GWAS发现的位点多位于非编码区

GWAS结果解释

Table 10.3 Bioinformatics resources for the assessment of individual effects of SNPs

Resource	Type of variant/input	Function tested	Features
PolyPhen2 (Polymorphism Phenotyping)	Missense/rs ID	Deleterious effect on protein function	Based on sequence conservation in a multiple sequence alignment, structure-to-model position of amino acid substitution, and SWISS-PROT annotation
SIFT (Sorting Intolerant From Tolerant)	Missense/rs ID	Deleterious effect on protein function	Based on sequence conservation
SNPs3D	Missense/rs ID	Deleterious effect on protein function	Based on structure- and sequence-based support vector machines Provides a score for the effect of the SNP on protein function. Also includes gene and gene interactions based on text mining
MutPred	Protein sequence variation	Classifies the protein mutation as disease-causing or neutral	Requires a protein sequence in FASTA format and a list of amino acid substitutions for analysis
dbSTEP (Database of Splice Translational Efficiency Polymorphisms)	Splice translational efficiency polymorphisms (STEPs) that alter splicing of the 5′ UTR of pre-mRNAs (potentially affecting protein quantity but not quality)	Effect on protein quantity	Allows searching by gene, SNP, or genomic coordinates and provides a detailed report. It cross-matches to known SNP associations in GWASs
AASsites (Automatic Analysis of SNP Sites)	The input is the DNA sequence; the program identifies the SNP	Effect on protein sequence (splicing pattern)	Identifies the splice regulatory site and predicts its effect on protein function
GTEx (Gene and Tissue Expression)	Any SNP/rs ID	Effect on tissue-specific gene expression level	Currently the most comprehensive database for eQTLs and with tissue specificity; uses an extreme statistical significance threshold (no eQTL for most genes)
SCAN	Any SNP/rs ID	Effect on tissue-specific	Samples are from the original

后GWAS时代(post-GWAS era)

- **丢失的遗传度(missing heritability)**
  - ✓ 提升统计学效能：方法学/策略改进 + 样本量
  - ✓ 不仅仅常见变异：罕见变异关联研究
  - ✓ 其它效应：基因-基因交互
- **疾病风险预测**
- **易感人群识别：基于遗传风险评分**

丢失的遗传度

遗传度/遗传力：表型变异中归因于遗传因素的比例

$$P = G + E$$

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$$

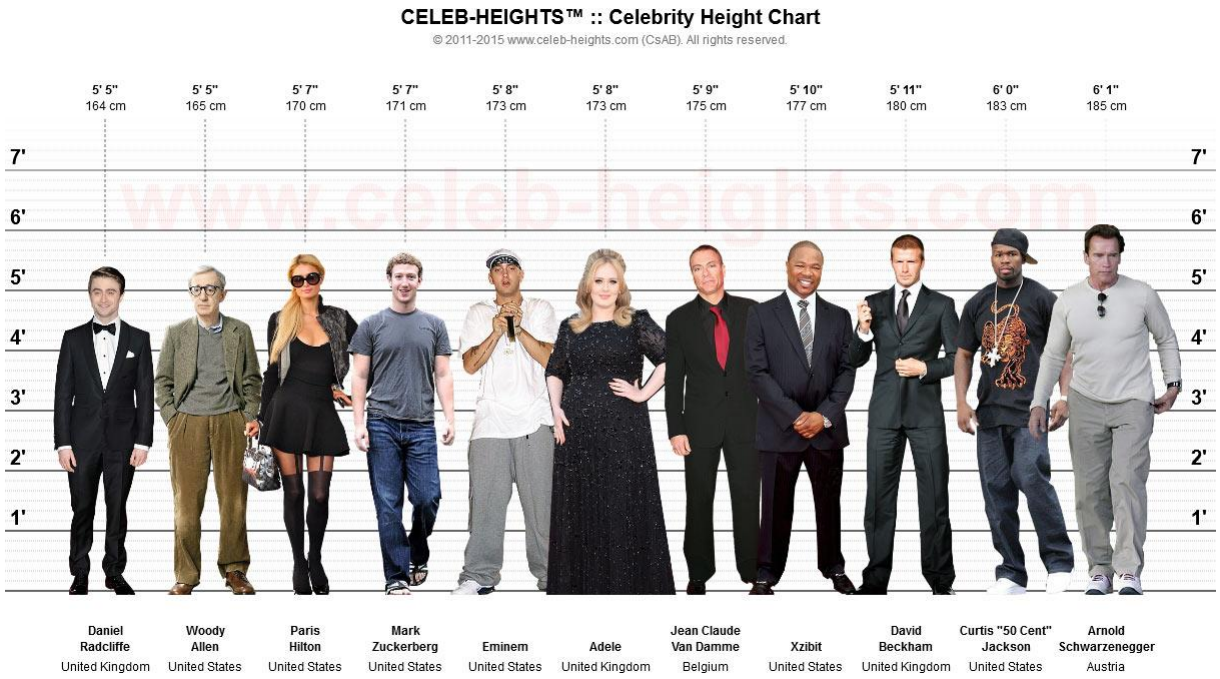
广义遗传度：$H^2 = \sigma_G^2 / \sigma_P^2$

狭义遗传度：$h^2 = \sigma_A^2 / \sigma_P^2$

丢失的遗传度

丢失的遗传度：GWAS发现的位点仅能解释表型很小一部分变异，远低于“真实”的遗传度

- ~80%, ~5%



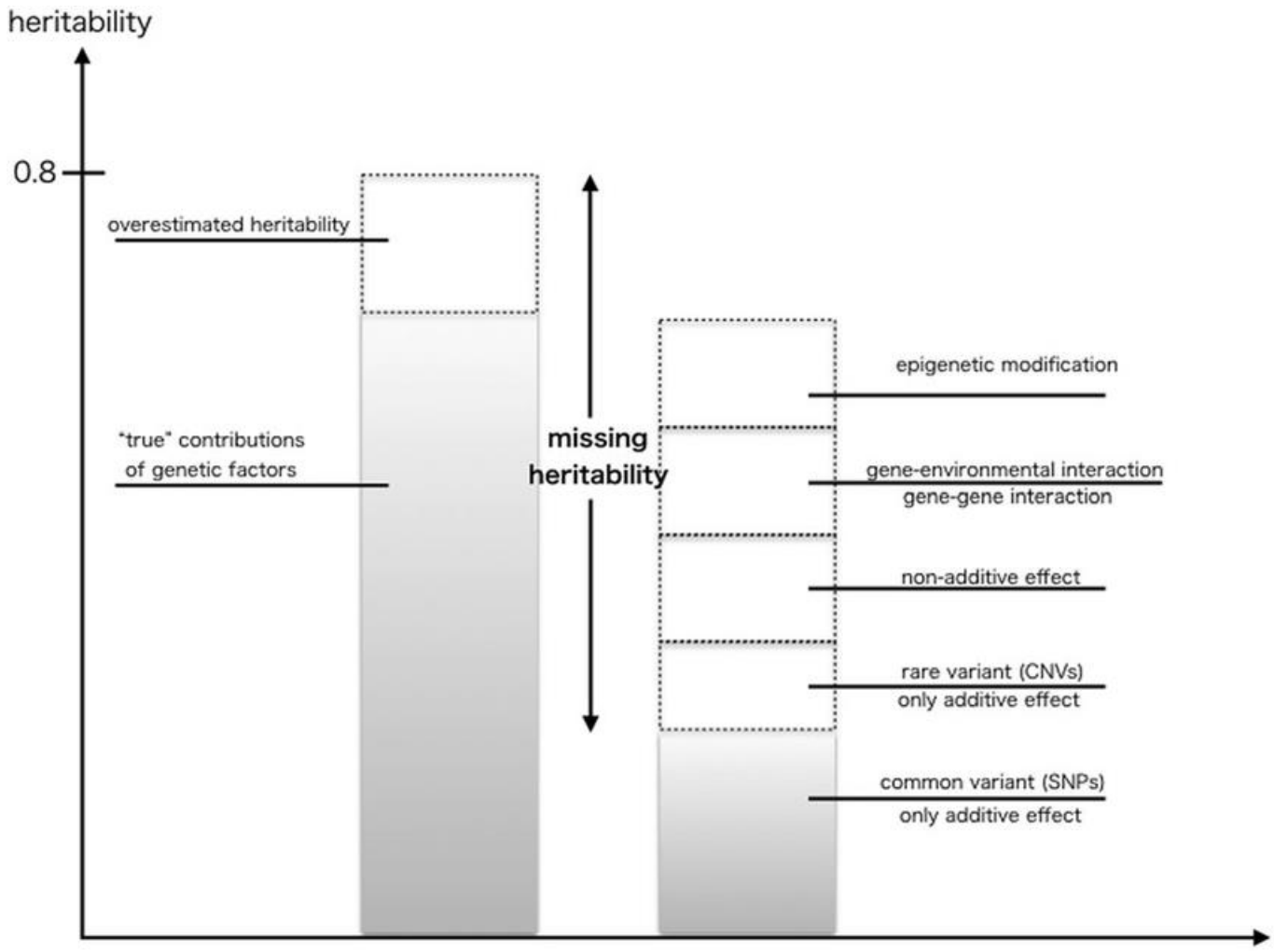
丢失的遗传度

可能原因：

- ✓ 多基因效应(polygenic effect): 多、低
- ✓ 罕见或低频变异(rare/low-frequent variants): MAF<0.05
- ✓ 结构变异(structural variants)
- ✓ 基因交互(gene-gene interaction)

- ✓ 过高估计“真实”的遗传度(due to shared environment)
- ✓ 暗物质(dark matter): 只闻其声，不见其人

丢失的遗传度


heritability
0.8
overestimated heritability
"true" contributions
of genetic factors
missing
heritability
epigenetic modification
gene-environmental interaction
gene-gene interaction
non-additive effect
rare variant (CNVs)
only additive effect
common variant (SNPs)
only additive effect

丢失的遗传度

解决办法1：针对多基因效应、低频变异，增加统计学效能，控制假阳性

- ✓ 增加样本量：多个研究合并 or 基于summary data的meta分析，但会引入人群差异

- ✓ 改进统计学方法/分析策略：
  - 多位点模型：SNP-set / Gene-based / Pathway-based / Haplotype-Based
  - 整合功能注释
  - 存在即合理：混合效应模型使用全部SNP位点，提升SNP-based heritability估计，捕获marker与causal的imperfect LD；身高遗传度提升至45%

丢失的遗传度

解决办法2：针对罕见变异(MAF<0.5%)或结构变异

- ✓ 技术/样本量：定制芯片；全外显子组 or 全基因组测序；大样本；家系研究
- ✓ 方法：统计学方法改进(如SKAT方法)，结构变异calling算法

解决办法3：针对基因-基因交互作用

- ✓ 方法：检验海量的交互作用项，机器学习、基于核函数的方法；家系研究

丢失的遗传度

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

AJHG

Volume 89, Issue 1, 15 July 2011, Pages 82-93

Article

Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test

Michael C. Wu [1, 5], Seunggeun Lee [2, 5], Tianxi Cai [2], Yun Li [1, 3], Michael Boehnke [4], Xihong Lin [2]

AJHG

Volume 86, Issue 6, 11 June 2010, Pages 929-942

Article

Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies

Michael C. Wu [1], Peter Kraft [2, 3], Michael P. Epstein [4], Deanne M. Taylor [2], Stephen J. Chanock [5], David J. Hunter [3], Xihong Lin [2]

Show more

Eur J Hum Genet. 2010 Jan;18(1):111-7. doi: 10.1038/ejhg.2009.115.

Gene and pathway-based second-wave analysis of genome-wide association studies

Gang Peng [1], Li Luo, Hoicheong Siu, Yun Zhu, Pengfei Hu, Shengjun Hong, Jinying Zhao, Xiaodong Zhou, John D Reveille, Li Jin, Christopher I Amos, Momiao Xiong

Ann. Appl. Stat.
Volume 8, Number 1 (2014), 352-376.

← Previous article | TOC | Next

Joint analysis of SNP and gene expression data in genetic association studies of complex diseases

Yen-Tsung Huang, Tyler J. VanderWeele, and Xihong Lin

ns + expand

9584899 PMCID: PMC2987176 DOI: 10.1038/ejhg.2009.115

C article

Published: 18 March 2012

Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits

Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Pamela A F Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael N Weedon, Ruth J Loos, Timothy M Frayling, Mark I McCarthy, Joel N Hirschhorn, Michael E Goddard & Peter M Visscher

Nature Genetics **44**, 369–375(2012) | Cite this article

4654 Accesses | **470** Citations | **22** Altmetric | Metrics

Published: 20 June 2010

Common SNPs explain a large proportion of the heritability for human height

Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard & Peter M Visscher

Nature Genetics **42**, 565–569(2010) | Cite this article

9141 Accesses | **2199** Citations | **164** Altmetric | Metrics

遗传风险评分与疾病风险预测

基于GWAS结果，衡量个体患病的遗传风险大小

$$GRS = \sum_{i}^{m} w_i G_i$$

THE LANCET
Oncology
Volume 21, Issue 10, October 2020, Pages 1378-1386

Articles

Genetic risk, incident gastric cancer, and healthy lifestyle: a meta-analysis of genome-wide association studies and prospective cohort study

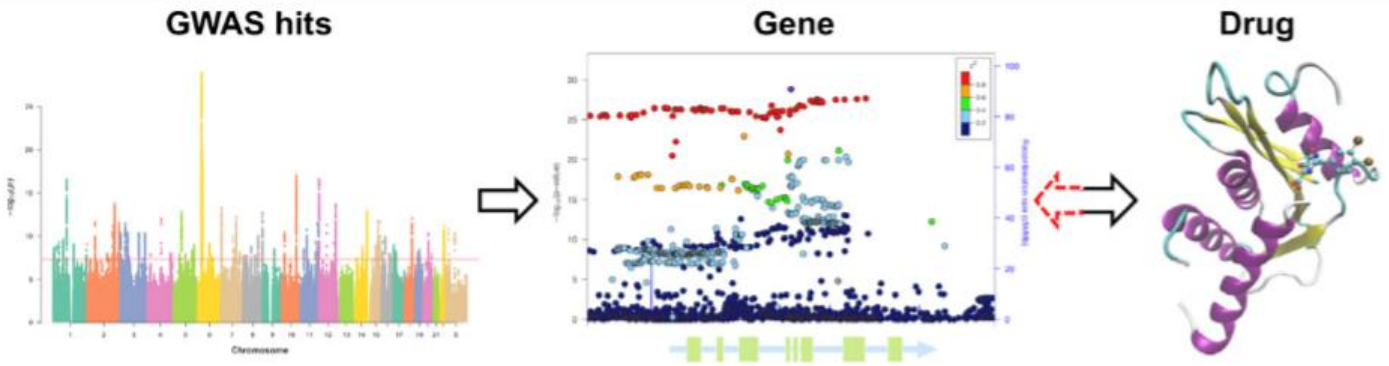
我国是胃癌发病率最高的国家之一，全球超过40%的胃癌发病和死亡发生在我国。遗传易感体质、幽门螺杆菌感染（Helicobacter pylori，H. pylori）以及不健康的生活方式可能共同导致了我国胃癌的巨大疾病负担。近十年来，国内外学者先后开展了多项胃癌全基因组关联研究（Genome-wide association study, GWAS），先后鉴定了10余个胃癌易感区域，然而这些易感位点是否可以预测胃癌的发病风险并识别高危人群，目前尚缺乏前瞻性人群研究证据的支持。本研究在前期研究的基础上，进一步扩大全基因组关联研究覆盖人群，开展了迄今为止最大的中国人群胃癌全基因组关联研究（累计10254例胃癌病例和10914例无癌对照），并基于遗传关联研究结果构建了中国人群胃癌多基因遗传风险评分（Polygenic Risk Score，PRS）模型；在此基础上，应用中国慢性病前瞻性研究队列（China Kadoorie Biobank，CKB）的10余万例研究对象前瞻性评估了PRS与胃癌发病风险的关系，并系统探讨了健康生活方式在不同遗传负荷的情况下对胃癌的保护效果。

研究发现中国人群胃癌多基因遗传风险评分PRS-112与胃癌发病风险与胃癌发病风险呈线性关联。根据PRS-112的五分位间距将研究对象分为Q1-Q5五类，随着遗传负荷的增加，个体罹患胃癌的风险显著增高，并呈剂量-反应关系（**图1A**）。当以Q1、Q2-Q4、Q5分别定义低、中、高遗传风险人群时，我们发现三个人群胃癌累计标化发病率存在显著差异，高风险人群胃癌发病率是低风险人群的2.08倍（HR=2.08, 95% CI: 1.61-2.69）（**图1B**）。

遗传风险评分与疾病风险预测

研究	疾病	方法	位点数	预测效果
Speliotes, et al.	动脉粥样硬化	GRS	32	AUC=0.52
Li, et al.	肺癌	GRS	4	AUC=0.55
David, et al.	高血压	GeRSI	whole	AUC=0.59
	躁郁症	GeRSI	whole	AUC=0.62
Lello, et al.	2型糖尿病	LASSO	4168	AUC=0.64
	高血压	LASSO	9674	AUC=0.67
	乳腺癌	LASSO	480	AUC=0.58

GWAS成果转化



Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*RANKL/ESR1*	Denosumab/Raloxifene and HRT
Schizophrenia	*DRD2*	Anti-psychotics
LDL cholesterol	*HMGCR*	Pravastatin
AS, Ps, Psoriatic Arthritis	*IL12B*	Ustekinumab

总结

Table 1. The Role of GWAS SNP Arrays in Human Genetic Discoveries

Analysis	Purpose	Discoveries
GWAS	detecting trait-SNP associations	~10,000 robust associations with diseases and disorders, quantitative traits, and genomic traits
Genome-wide CNV analysis	detecting trait-CNV associations	hundreds of associations with diseases and disorders
Genome-wide assessment of LD	quantifying genome architecture	large variation in LD in the genome
Estimation of SNP heritability[a]	genetic architecture	large proportion of genetic variation captured by common SNPs
Estimation of genetic correlation[a]	detecting and quantifying pleiotropy	pleiotropy is ubiquitous
Polygenic risk scores[a]	detecting pleiotropy; validating GWAS discoveries	out-of-sample prediction works as expected; detection of novel trait associations
Mendelian randomization[a]	testing causal relationships	replication of known causal relationships; empirical evidence of observational associations that are not causal
Population differences in allele frequencies	reconstructing human population history; detecting selection	genetic structure can mimic geographical structure; evidence of natural selection
Trait GWAS with -omics GWAS[a]	fine-mapping; detecting target genes; function	two-thirds of GWAS-associated loci implicate a gene that is not the nearest gene to the most associated SNP

[a]These analyses can be performed with GWAS summary statistics.

A lot of work to do…