中国科学院北京基因组研究所（国家生物信息中心）
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

国家基因组科学数据中心
National Genomics Data Center

第11章 基因型与表型关联分析

章张 博士

内容提纲

- 全基因组关联分析（GWAS）
  - 基本概念
  - 研究方式
  - 统计分析方法
  - GWAS实例
- WAS系列
  - 表观组EWAS
  - 转录组TWAS
  - 蛋白质组PWAS
  - 表型组PheWAS
- 关联知识库

见一叶落，而知岁之将暮；睹瓶中之冰，而知天下之寒。——《淮南子·说山训》

全基因组关联分析GWAS

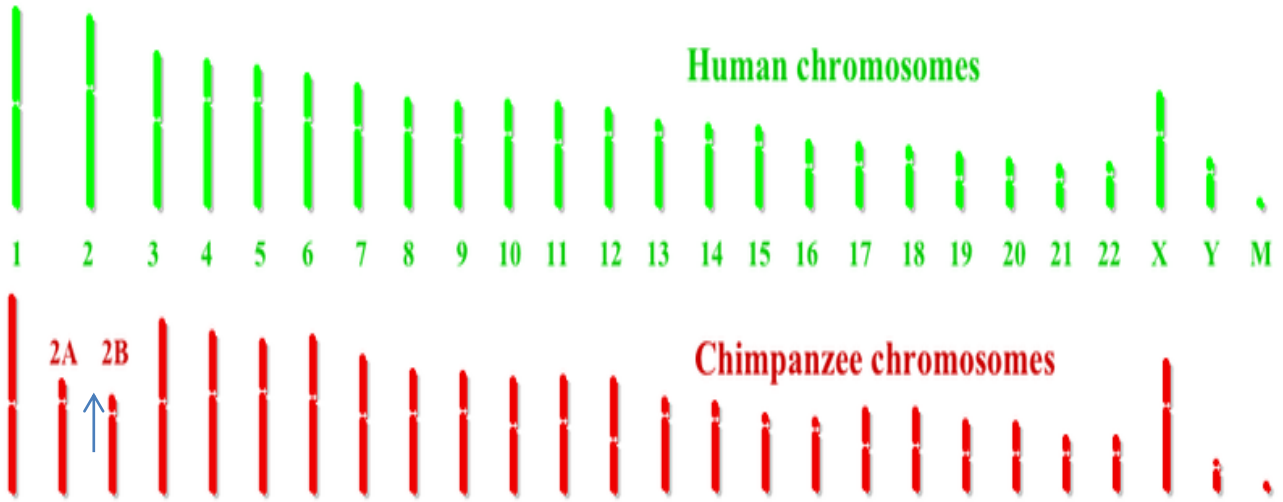
Courtesy of Xiang-feng Wang@CAU

个体差异

Gene	SNP	Allele	No. of Alleles
MC1R	rs86insA	A	0 1 2 NA
MC1R	rs11547464	A	0 1 2 NA
MC1R	rs885479	T	0 1 2 NA
MC1R	rs1805008	T	0 1 2 NA
MC1R	rs1805005	T	0 1 2 NA
MC1R	rs1805006	A	0 1 2 NA
MC1R	rs1805007	T	0 1 2 NA
MC1R	rs1805009	C	0 1 2 NA
MC1R	Y152OCH	A	0 1 2 NA
MC1R	rs2228479	A	0 1 2 NA
MC1R	rs1110400	C	0 1 2 NA
SLC45A2	rs28777	C	0 1 2 NA
SLC45A2	rs16891982	C	0 1 2 NA
KITLG	rs12821256	G	0 1 2 NA
EXOC2	rs4959270	A	0 1 2 NA
IRF4	rs12203592	T	0 1 2 NA
TYR	rs1042602	T	0 1 2 NA
OCA2	rs1800407	A	0 1 2 NA
SLC24A4	rs2402130	G	0 1 2 NA
HERC2	rs12913832	T	0 1 2 NA
ASIP	rs2378249	C	0 1 2 NA
SLC24A4	rs12896399	T	0 1 2 NA
TYR	rs1393350	T	0 1 2 NA
TYRP1	rs683	G	0 1 2 NA

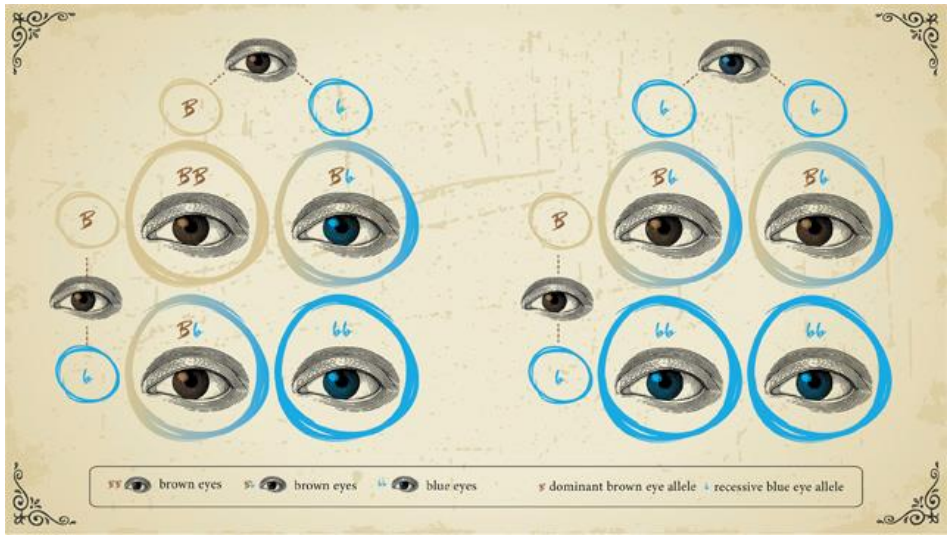
- 为什么有人抽烟喝酒活80多，有的人小心翼翼却得癌症?
- 为什么欧洲人是蓝眼睛，亚洲人是黑眼睛?
- 如何从一滴血判断犯罪分子体貌特征?
- 肥胖基因是好基因还是坏基因?
- 喝酒为什么脸红?

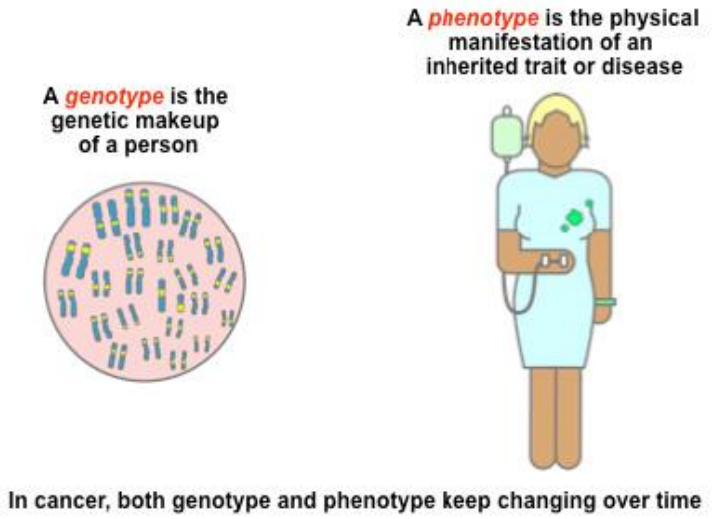
一切答案都写在了你的DNA里面。



基因型和表型

- **基因型（Genotype）**：是指某一生物体全部基因信息的总称，反映生物体的遗传构成，遗传学中具体使用的基因型，往往是指某一性状的基因型。
- **表型（Phenotype）**：具有特定基因型的个体，在一定环境条件下，所表现出来的性状特征（形态、结构、生理、生化、行为等）或疾病状态。
- **表型 = 基因型 + 环境**，表型主要受生物基因型和环境影响





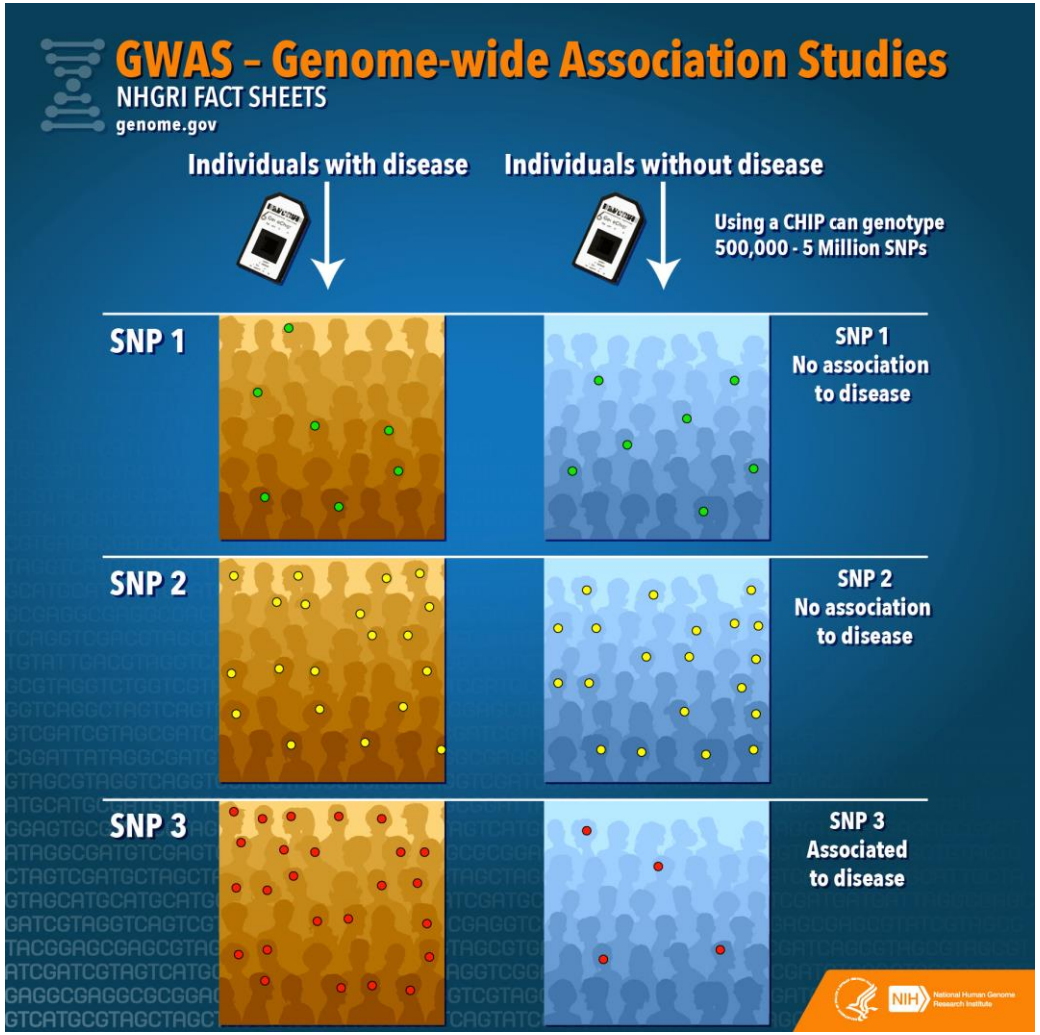
https://cisncancer.org/research/what_we_know/biology/language_of_biology.html

GWAS基本概念

全基因组关联分析

Genome-Wide Association Study，简称GWAS

研究复杂表型性状关联基因变异的方法，通过对照分析或相关性分析，在全基因组范围内找出存在的序列变异，从中筛选出与复杂表型性状相关的基因变异

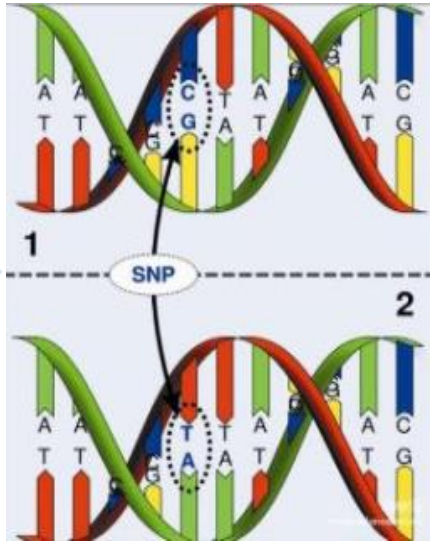


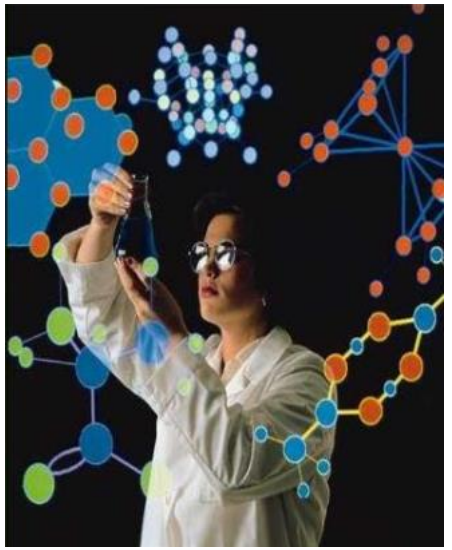
https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet

GWAS发展

GWAS是利用全基因组范围内的分子标记（SNP）对群体进行扫描，分析分子标记与性状之间关联关系的方法

- 1996年，Risch最早提出了GWAS的设想，认为未来人类复杂疾病的研究不再需要候选基因的预测，能够在全基因组水平检测
- 2001年，Hansen等最早应用GWAS在植物中对Sea beet（海甜菜）的生长习性进行了分析
- 2002年，Ozaki等发现2个SNP与人类心肌梗塞密切相关





身高差的遗传基础是?

传统单基因遗传

利用家系连锁分析的定位克隆方法，发现了大量单基因疾病（如囊性纤维化病、亨廷顿病性痴呆等）和单基因遗传性状


亨廷顿病性痴呆

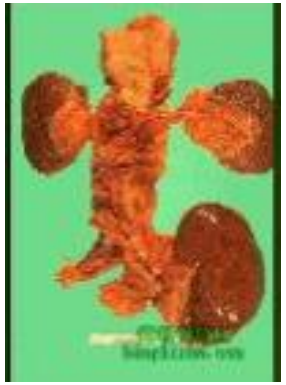

囊性纤维化病


圖8-14 上眼瞼有無皺褶
1.雙眼皮　2.單眼皮


圖8-15 臉頰有無酒窩
1.有酒窩　2.沒有酒窩

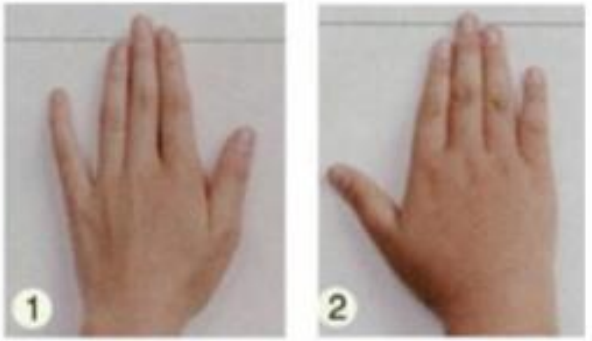


圖8-12 食指長短
1.食指比無名指長　2.食指比無名指短

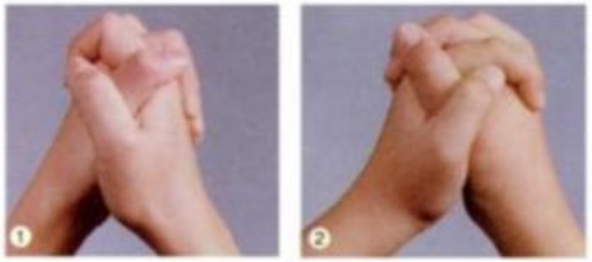


圖8-13 雙手手指嵌合
1.右手拇指在上　2.左手拇指在上

GWAS优势

- 全基因组关联分析
  - 对多个个体在全基因组范围的遗传变异（标记）多态性进行检测，获得基因型
  - 将基因型与可观测的性状，即表型，进行群体水平的统计学分析
  - 根据统计量或显著性P值筛选出最有可能影响该性状的遗传变异（标记），挖掘与性状变异相关的基因
- GWAS优势：
  - 分辨率高（单碱基水平）
  - 研究材料来源广泛，可捕获的变异丰富
  - 节省时间

GWAS技术

基因分型技术和遗传信息学的发展

基因型检测----芯片与测序

Affymetrix Genome-Wide Human SNP Arrays

- **基因芯片技术**：高通量分型

农业基因组芯片

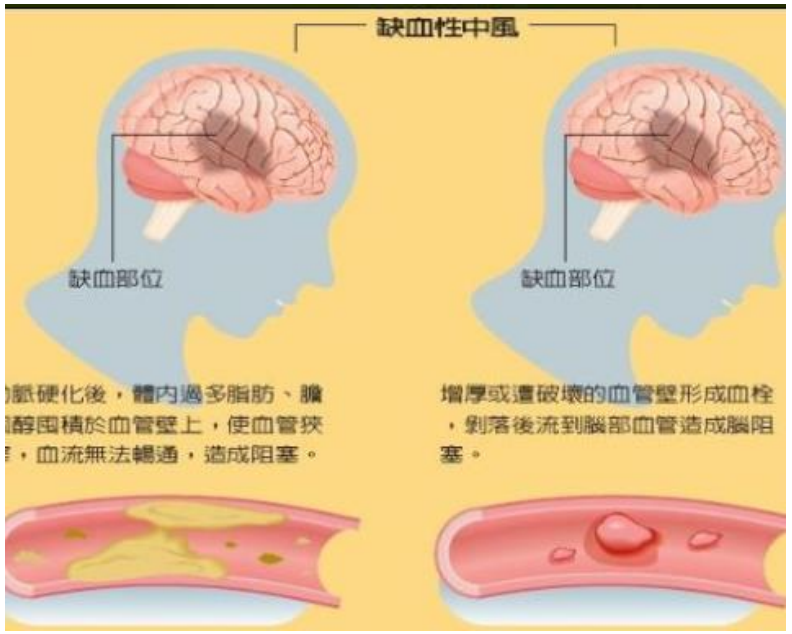
- **全基因组测序技术**：高通量低成本

GWAS jargon

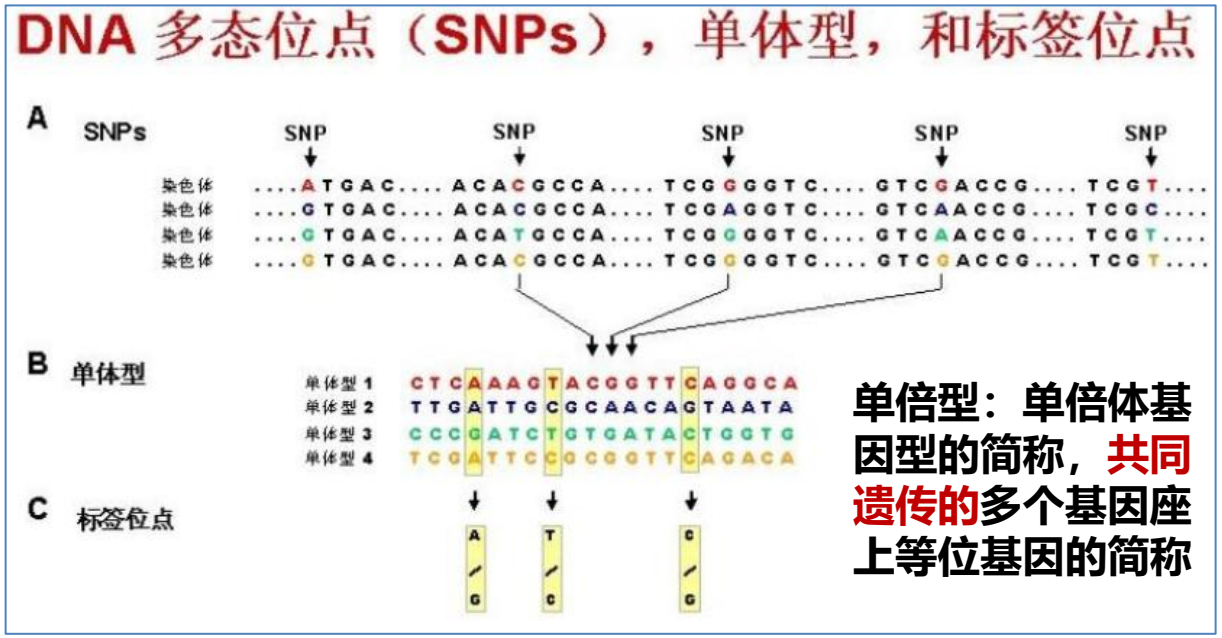
- Locus - genetic position on a chromosome, and a single base pair position in the context of SNPs
- SNP - a locus (single base pair) that exhibits variation (polymorphism) in a population
- Allele (in the context of SNPs) - the alternative forms of a nucleotide at a particular locus
- Genotype - the pair of alleles at a locus, one paternal and one maternal
- Haplotype - a group of SNPs that are inherited jointly from a parent
- Linkage disequilibrium - alleles at multiple loci that exhibit a dependence (nonrandom association)

GWAS研究基础

- 定义复杂性状的表型：尽可能选择可定量反映疾病危险程度的指标、可用于分析疾病临床亚型的特征、或可用于诊断鉴别疾病的表型特征
- 高密度分子遗传标记：包括SNP、indel、CNV等



如缺血性脑卒可能涉及血栓脱落或脑动脉粥样硬化等



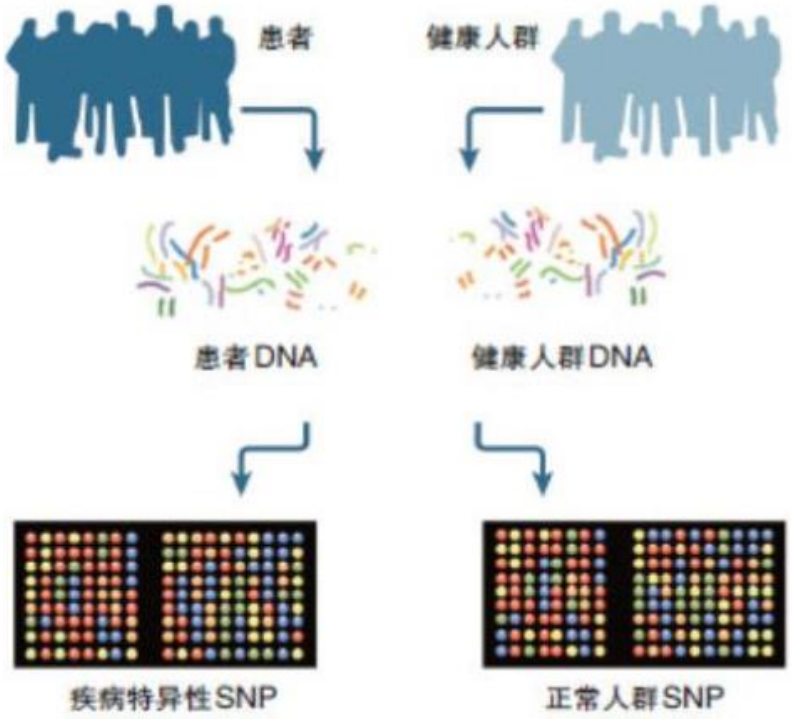
基于单倍型图谱，可选择覆盖一定数量（50w-100w）的全基因组变异位点进行GWAS

GWAS适用数量性状和质量性状

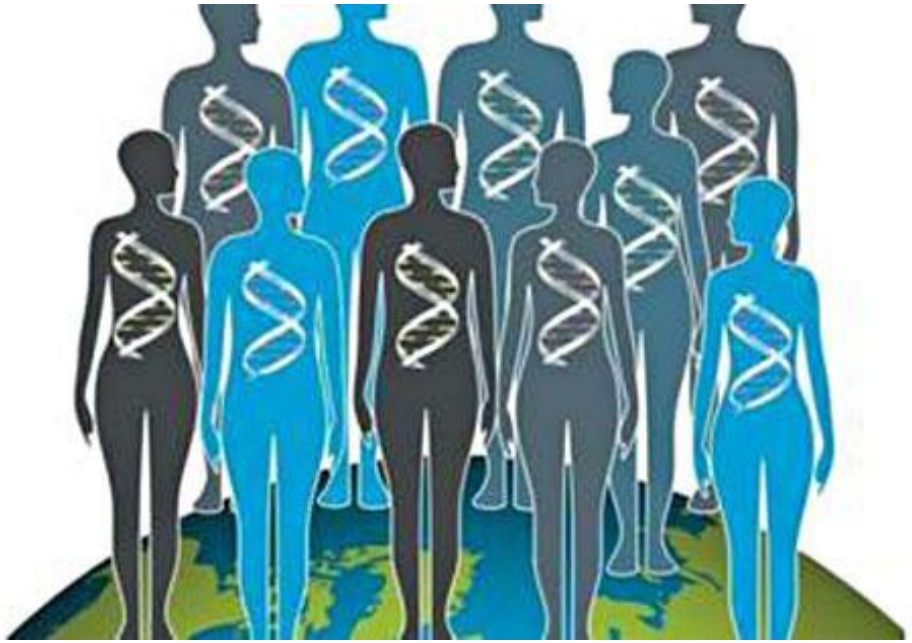
- **数量性状**：多基因控制，能够测量得到具体数值，符合正态分布；考虑到数量性状受环境影响大，建议将所有材料在同一环境下培育或养殖，或者用多年多点的数据分开分析后综合结果或取BLUP (best linear unbiased prediction)值作为性状值进行关联分析。
- **质量性状**：单基因控制，无法用具体数值衡量，可转换成0、1等表示，需注意每个群体选取近似的样本。
- **分级性状**：表型分布类似质量性状，但实际受多基因控制（数量性状），如抗性性状，因此需要提供每一个个体精确的测量数据。
- **多指标性状**：有多个指标可以同时度量时，找出代表原表型数据变异的主成分因子，作为关联分析的表型数据。

GWAS研究方式

GWAS技术通过比较：（1）患者和正常人群；（2）随机人群队列的基因组多态位点，找出疾病特异性的遗传标记，从而识别出某种疾病或表型的发病机制或相关位点。理论上，检测到的多态位点越多，识别关键位点的可能性就越大。



患者和正常



随机人群队列

GWAS研究方式

- 单阶段和多阶段
- 单阶段：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和疾病/性状的关联，早期的GWAS研究中应用较多。
- 多阶段：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；第二阶段采用更大的样本集对第一阶段筛选出的阳性SNP进行分析。

DNA Pooling: a tool for large-scale association studies. *Nature Review Genetics* (2002)
https://doi.org/10.1038/nrg930

GWAS统计分析方法

- Each SNP is an independent test
- Associations are tested by comparing the frequency of each allele in cases and controls
- Odds ratio: strength of association
- P-Value: chi-square test

Table 3. Association of Alleles and Genotypes of rs6983267 on Chromosome 8q24 With Colorectal Cancer[a]

	Number and Frequency of rs6983267 Alleles in Colorectal Cancer					Number and Frequency of rs6983267 Genotypes in Colorectal Cancer						
	C	T	χ^2 (1*df*)	*P* Value	OR	CC	CT	TT	χ^2 (2*df*)	*P* Value	OR	OR
Cases	875 (56.5)	675 (43.5)	24.8	6.3×10^{-7}	1.35[b]	250 (32.3)	375 (48.4)	150 (19.4)	24.5	4.7×10^{-6}	1.33[c]	1.81[d]
Controls	1860 (48.9)	1940 (51.1)				460 (24.2)	940 (49.4)	500 (26.3)				

Abbreviation: OR, odds ratio.
[a] Data are hypothetical; adapted from Tomlinson et al.[56]
[b] Denotes allelic odds ratio.
[c] Denotes heterozygote odds ratio.
[d] Denotes homozygote odds ratio.

How to Interpret a Genome-wide Association Study. *JAMA* 2008

Odds ratio (OR)

A measure of effect size, or strength of association

- odds = P / (1 - P)
- If probability of winning is 50%, odds is 0.5 / (1-0.5) = 1
- If probably of winning is 75%, odds is 0.75 / (1-0.75) = 3

	Cases	Controls
Allele T	A	B
Allele G	C	D

- OR=1, no association
- OR>1, allele T increases risk
- OR<1, allele G increases risk

$$OR = \frac{\text{odds of case with T allele}}{\text{odds of case with G allele}}$$

$$= (A/B) / (C/D) = (AD) / (BC)$$

95% CI of ln(OR) =
$\ln(OR) \pm 1.96(1/A + 1/B + 1/C + 1/D)^{0.5}$

Odds ratio (OR)

	Cancer	No cancer
Smoker	700	200
Non-smoker	300	800

$$OR = (700/200) / (300/800) \approx 9.33$$

	治疗有效	治疗无效
干预组	500	500
对照组	250	750

$$OR = (500/500) / (250/750) = 3$$

Chi-square test for significance level

	GG	GT	TT	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Observed allele counts

	G	T	Total
Cases	$2r_0+r_1$	r_1+2r_2	$2R$
Controls	$2s_0+s_1$	s_1+2s_2	$2S$
Total	$2n_0+n_1$	n_1+2n_2	$2N$

Expected allele counts

G	T
$2R(2n_0+n_1)/(2N)$	$2R(n_1+2n_2)/(2N)$
$2S(2n_0+n_1)/(2N)$	$2S(n_1+2n_2)/(2N)$

Chi-square test $\sum \frac{(Obs-Exp)}{Exp} \sim \lambda^2$

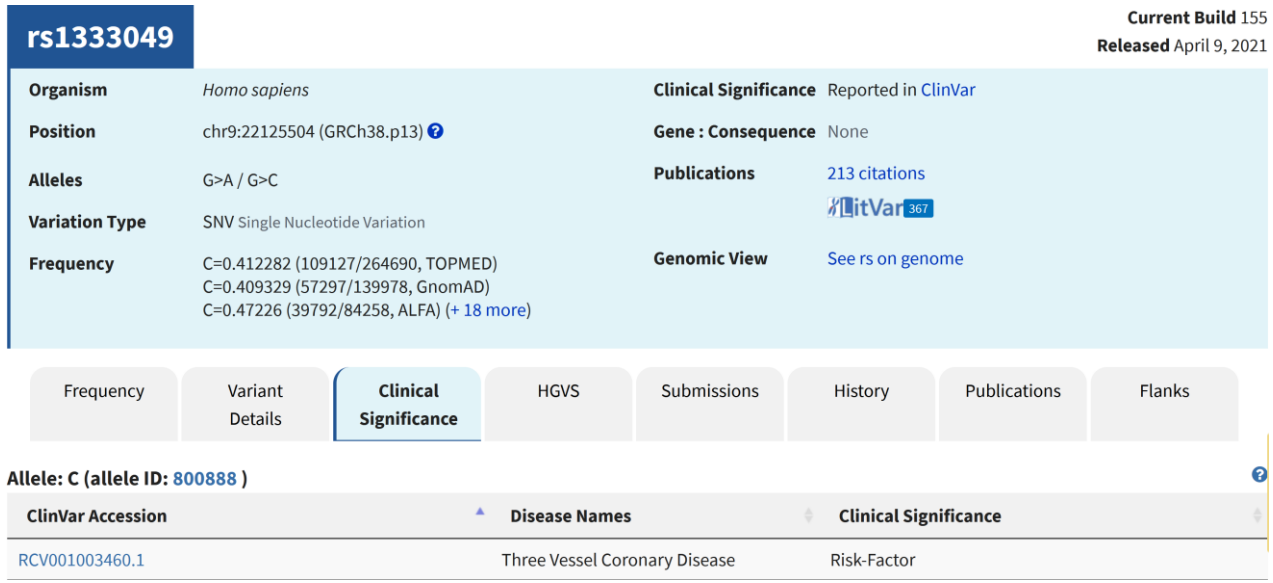
P-value

rs1333049: Association with Coronary Artery Disease (冠状动脉疾病)

	C (%)	G (%)	χ^2 (1df)	P-value
Cases	2,132 (55.4)	1,716 (44.6)	55.1	1.2×10^{-13}
Controls	2,783 (47.4)	3,089 (52.6)		
Allelic Odds Ratio = 1.38				

- OR = 1, no disease association
- OR > 1, allele C increase risk
- OR < 1, allele C decrease risk

Genomewide association analysis of coronary artery disease.
N Engl J Med 2007



https://www.ncbi.nlm.nih.gov/snp/rs1333049

Multiple Hypothesis Testing
多重检验校正

- **Correction for multiple testing**: Statistical hypothesis testing is based on rejecting the null hypothesis if the likelihood of the observed data under the null hypotheses is low. If multiple hypotheses are tested, the chance of observing a rare event increases, and therefore, the likelihood of incorrectly rejecting a null hypothesis increases.
- The **false discovery rate (FDR)** is a method of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons.
  - Benjamini-Hochberg, Holm–Bonferroni
- The **Bonferroni correction** compensates for that increase by testing each individual hypothesis at a significance level of α/m, where α is the desired overall alpha level and m is the number of hypotheses.
  - testing m=10 hypotheses with a desired α = 0.05, then the Bonferroni correction would test each individual hypothesis at α = 0.05/10 = 0.005.

Multiple Hypothesis Testing

When conducting multiple tests on the same samples or gene sets, it needs to adjust the overall false positive rate through correcting/adjusting *p*-values.

Rank	*p*-value	adjusted *p*-value		
		Bonferroni	**Benjamini-Hochberg**	Holm or Holm–Bonferroni
1	**0.001**	**0.01**	**0.0100**	**0.010**
2	**0.003**	**0.03**	**0.0150**	**0.027**
3	**0.006**	0.06	**0.0200**	**0.048**
4	**0.009**	0.09	**0.0225**	0.063
5	**0.016**	0.16	**0.0320**	0.096
6	**0.023**	0.23	**0.0383**	0.115
7	**0.038**	0.38	0.0543	0.152
8	**0.049**	0.49	0.0613	0.147
9	0.051	0.51	0.0567	0.102
10	0.080	0.80	0.0800	0.080

Assume:
n = number of tests
r = rank of p-value
Significance level at 0.05

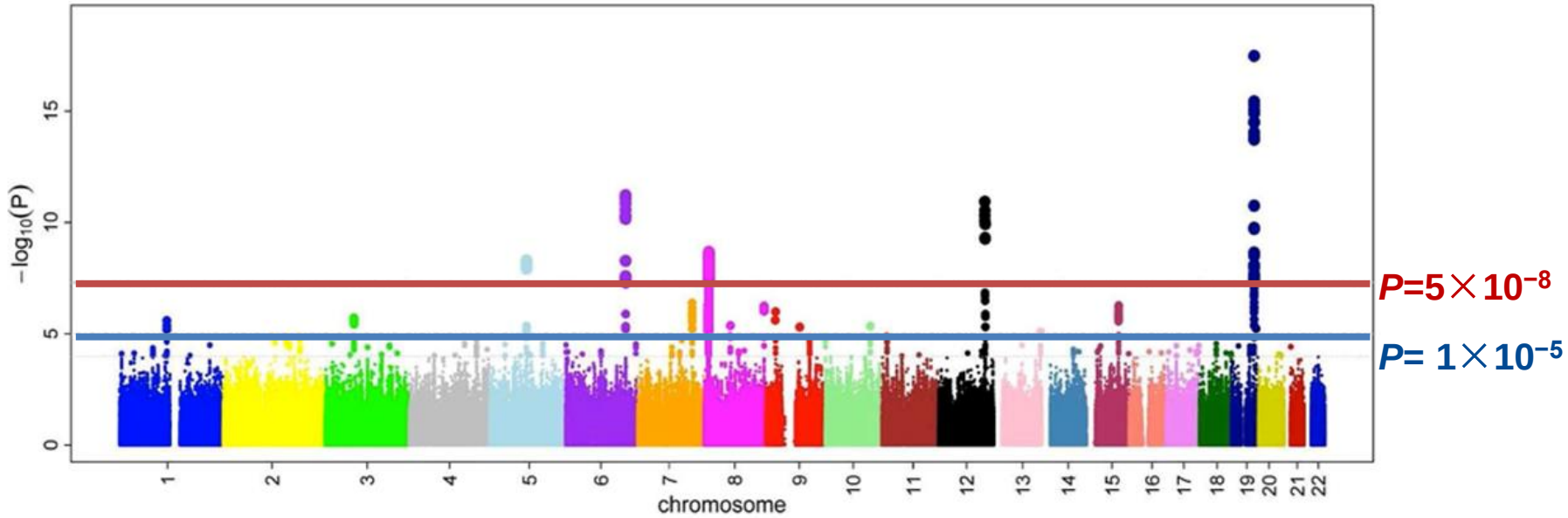
Bonferroni: $p * n$

Benjamini–Hochberg:
- rank *p*-values in ascending order
- calculate each adjusted *p*-value using $p * n / r$

Holm–Bonferroni:
- rank *p*-values in ascending order
- calculate each adjusted *p*-value using $p * (n - r + 1)$

Manhattan Plot

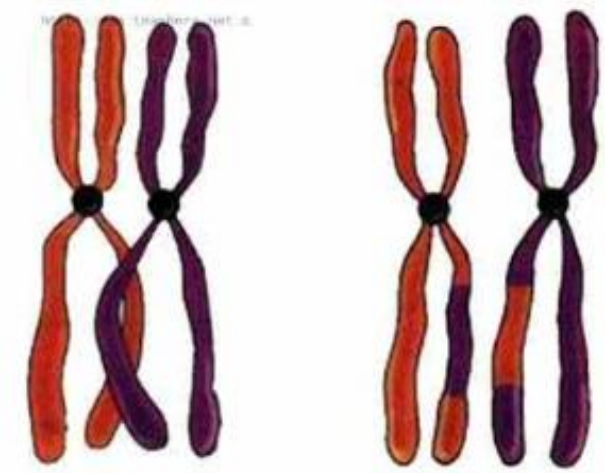
A Manhattan plot is **a type of scatter plot**, commonly used in genome-wide association studies (GWAS) to **display significant SNPs**.



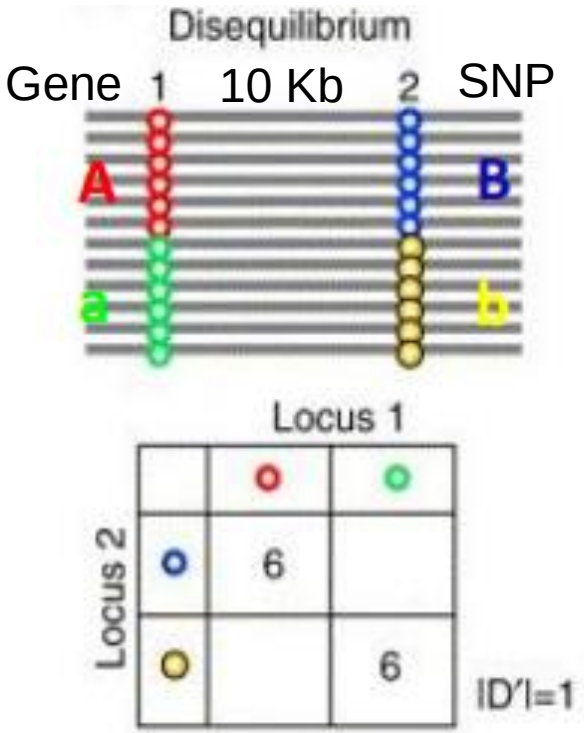
- In GWAS Manhattan plots, **genomic coordinates** are displayed along the X-axis, with the **negative logarithm of the association *P*-value** for each single nucleotide polymorphism (SNP) displayed on the Y-axis.
- The $-\log_{10}P$-values for association are **plotted for each SNP** according to its chromosomal position (10^{-15}, its negative logarithm is 15).

连锁不平衡 (Linkage disequilibrium)

群体内不同位点等位基因间的非随机性组合的关系，即当位于同一条染色体的两个基因座（A，B）基因型同时存在的概率，大于群体中因随机分布而同时出现的概率时，就称这两个点处于LD状态。

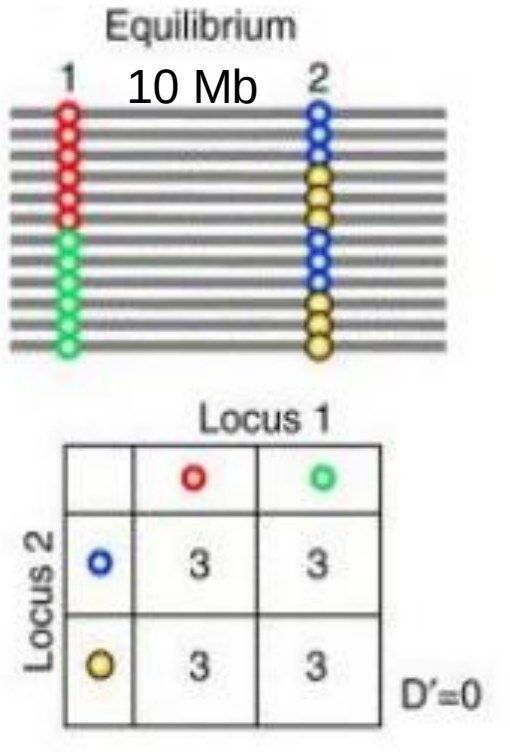


连锁不平衡



A和B连锁遗传
距离很近，重组共同交换
选1个SNP代表2个基因

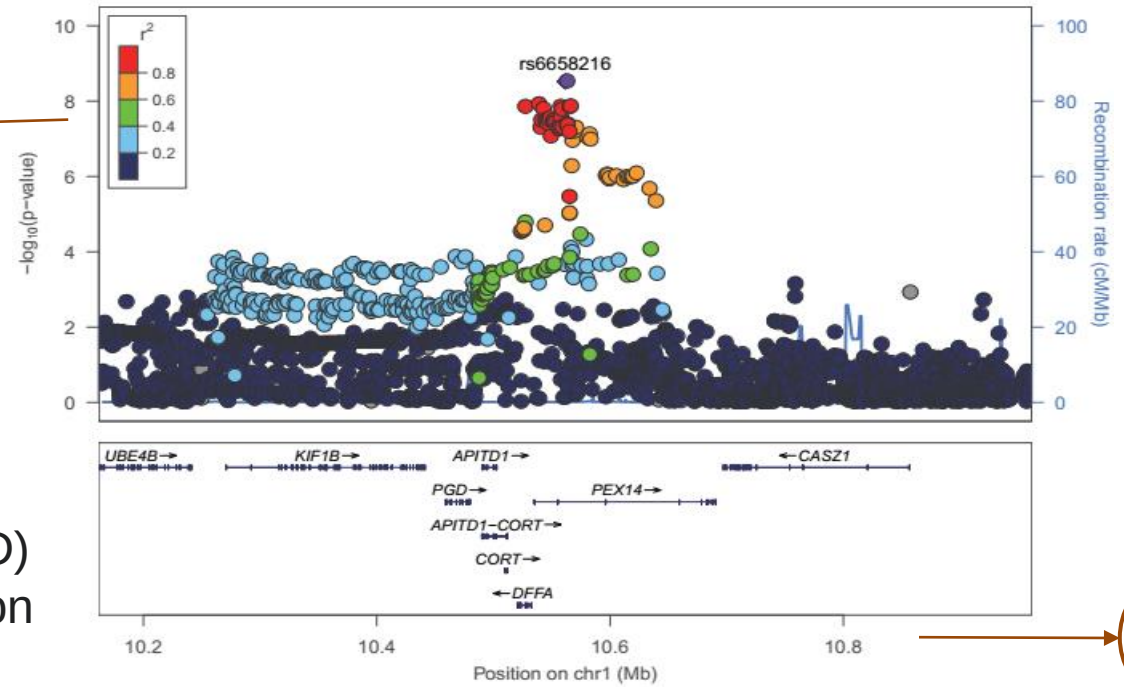
连锁平衡



A和B分离遗传
距离很远，重组分别交换
需要2个SNP代表2个基因

如果A和B存在显著的连锁不平衡，那么B可以作为A的标记

Regional Manhattan Plot & Haploview



LD heatmap (r^2) patterns

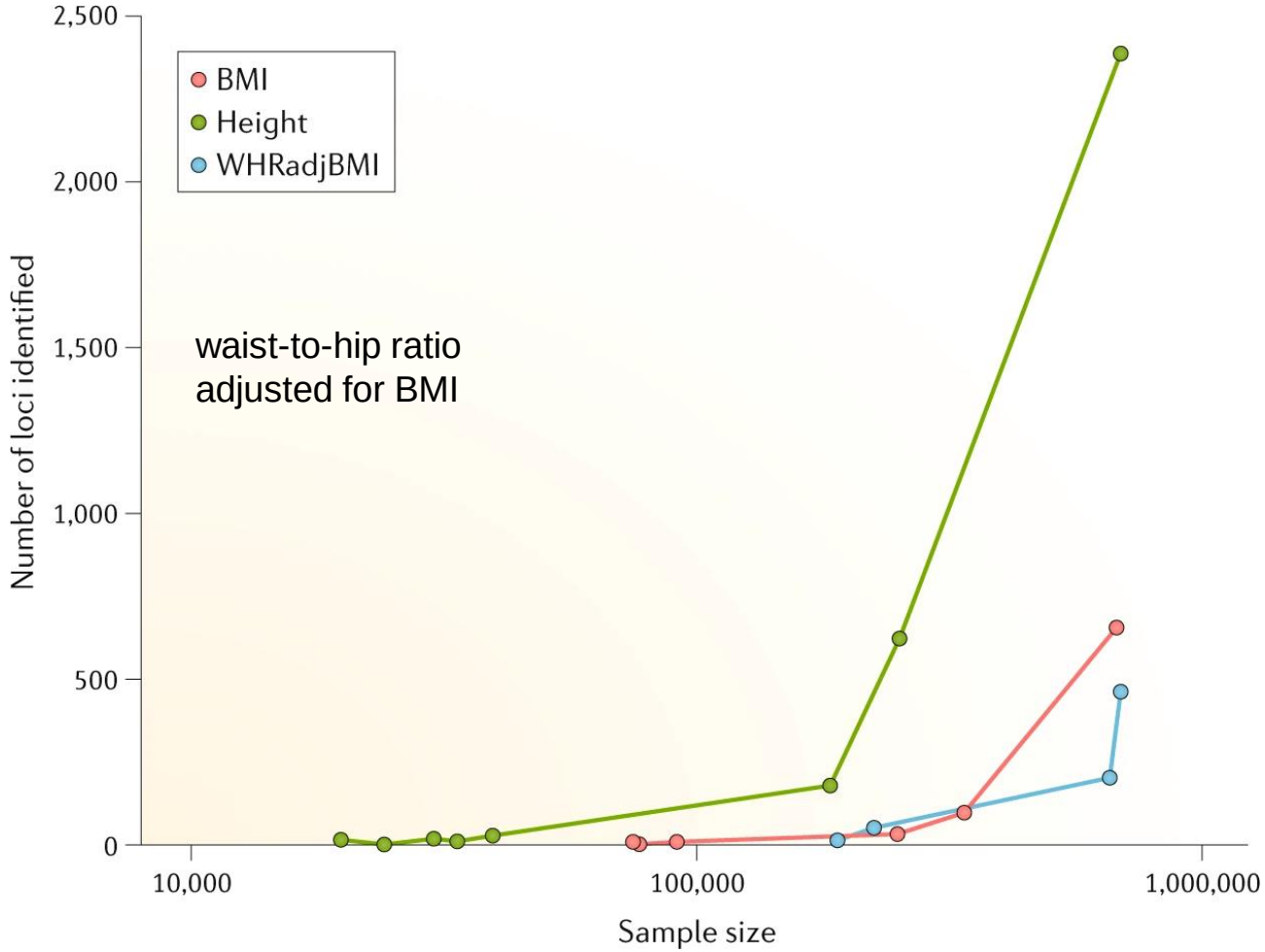
Physical position

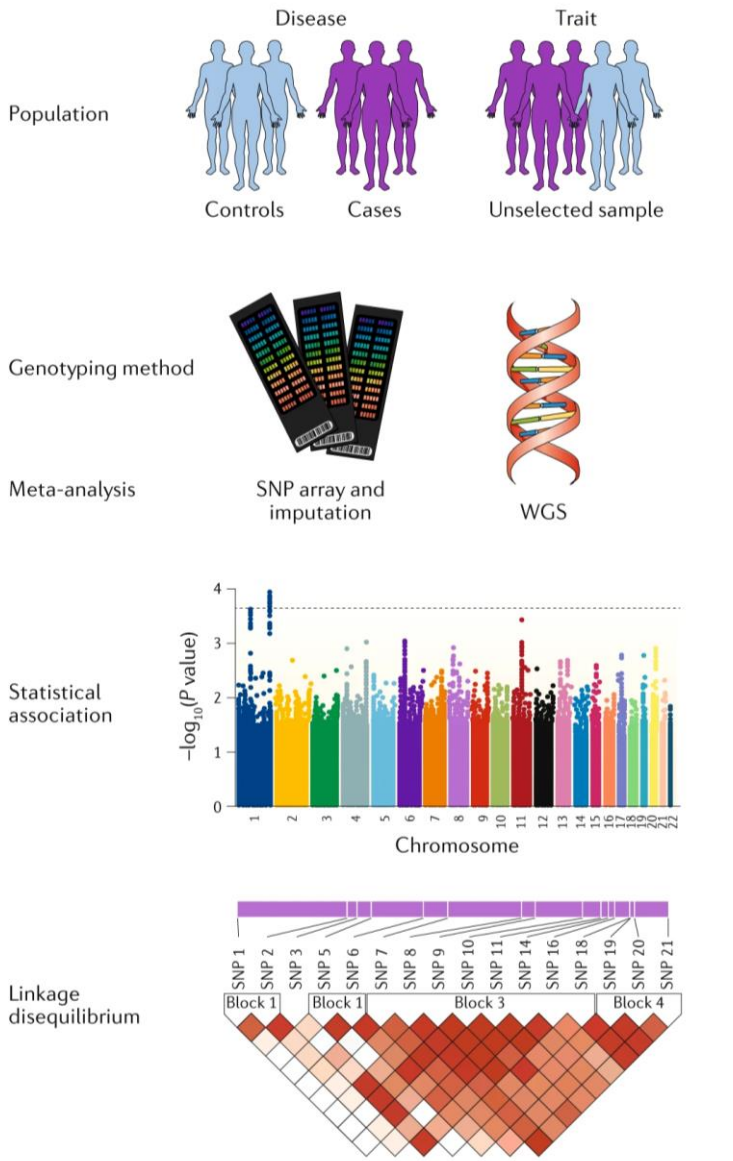
LD pattern using Haploview

Linkage Disequilibrium (LD) is the non-random association of alleles at different loci in a given population.

A和B两个位点上的等位基因不是独立关联的，这两个位点是连锁不平衡

GWAS study design





Number of loci identified as a function of GWAS sample size

Benefits and Limitations of Genome-Wide Association Studies. *Nature Review Genetics* 2019

First GWAS: myocardial infarction

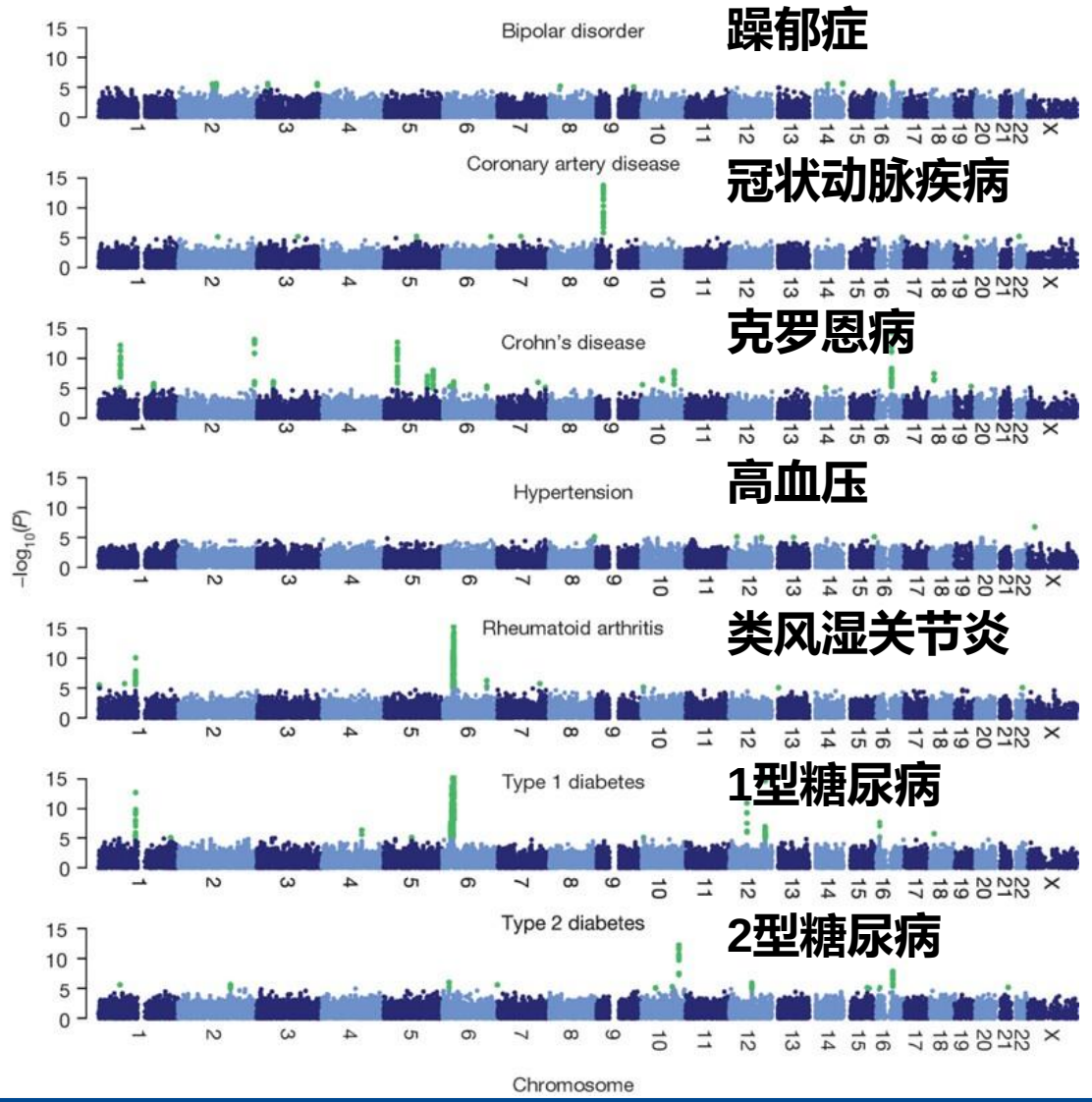
心肌梗塞

Table 3 • Association between myocardial infarction and SNPs in *LTA*, *NFKBIL1* and *BAT1*

Genotype	Myocardial infarction 62.5 ± 11.3* (*n* = 1,133)	Control 1 38.5 ± 20.3* (*n* = 1,006)	Control 2 64.3 ± 11.3* (*n* = 872)	χ^2 (*P* value) versus control 1	χ^2 (*P* value) versus control 2	Odds ratio (95% c.i.) versus control 1	Odds ratio (95% c.i.) versus control 2
LTA exon1 10G→A							
GG	416 (36.7%)	378 (37.6%)	344 (39.4%)	AA versus GG+GA			
GA	504 (44.5%)	512 (50.9%)	428 (49.1%)	21.6	20.2	1.78	1.79
AA	213 (18.8%)	116 (11.5%)	100 (11.5%)	(0.0000033)	(0.0000069)	(1.39–2.27)	(1.38–2.31)
LTA intron1 252A→G							
AA	413 (36.5%)	371 (36.9%)	346 (39.7%)	GG versus AA+AG			
AG	511 (45.1%)	516 (51.3%)	426 (48.8%)	18	17.4	1.69	1.75
GG	209 (18.4%)	119 (11.8%)	100 (11.5%)	(0.000022)	(0.000018)	(1.32–2.15)	(1.35–2.26)
LTA exon3 804C→A, T26N							
CC	414 (36.5%)	376 (37.4%)	343 (39.3%)	AA versus CC+CA			
CA	506 (44.7%)	514 (51.1%)	429 (49.2%)	21.6	20.1	1.78	1.79
AA	213 (18.8%)	116 (11.5%)	100 (11.5%)	(0.0000033)	(0.0000073)	(1.39–2.27)	(1.38–2.31)
NFKBIL1 promoter –63T→A							
TT	406 (35.8%)	374 (37.2%)	340 (39.0%)	AA versus TT+TA			
TA	521 (46.0%)	509 (50.6%)	424 (48.6%)	14.5	12.5	1.6	1.57
AA	206 (18.2%)	123 (12.2%)	108 (12.4%)	(0.00013)	(0.00040)	(1.25–2.03)	(1.22–2.02)

Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics* 2002

Genome-wide scan for seven diseases



Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium

Nature **447**, 661–678 (2007) | Cite this article

102k Accesses | **7185** Citations | **182** Altmetric | Metrics

GWAS represents a powerful approach to identify genes involved in human diseases.

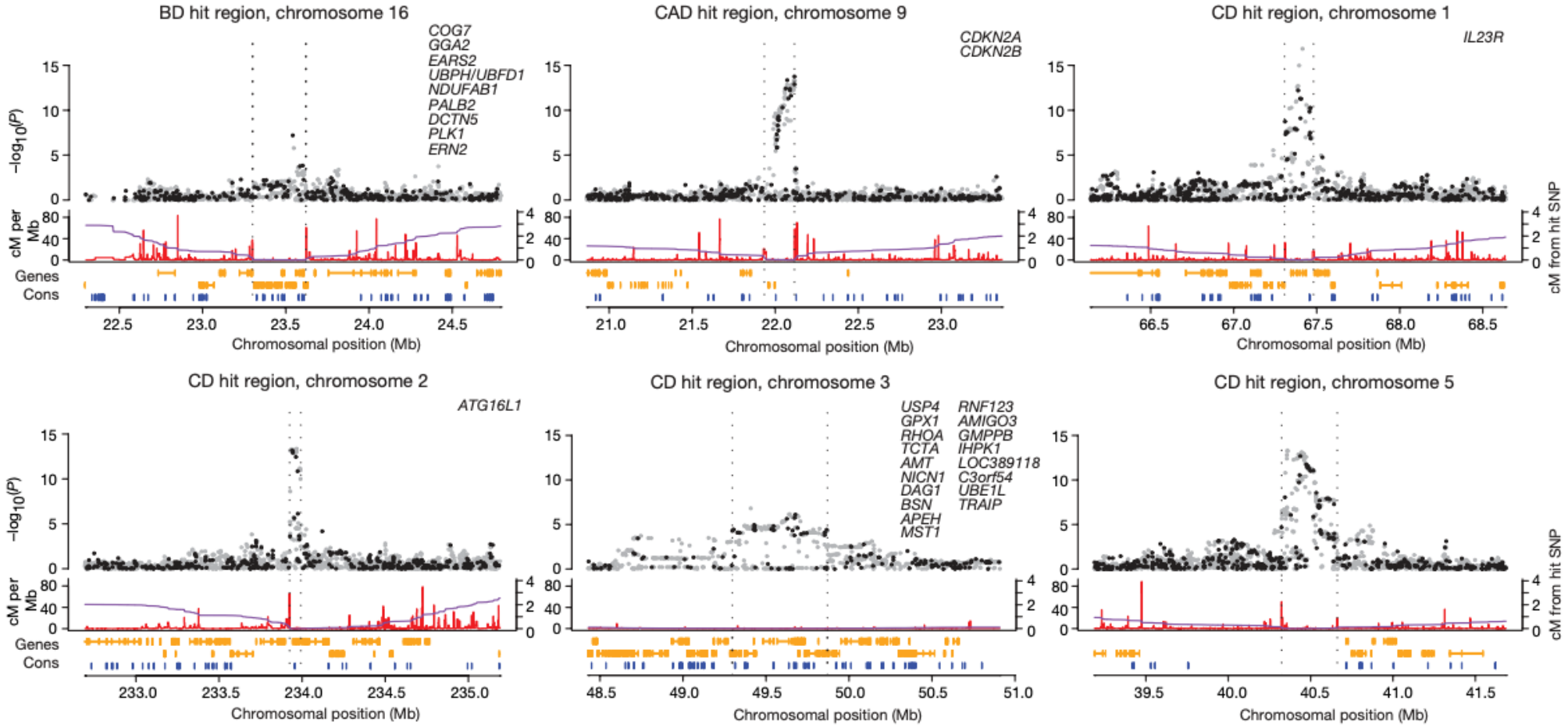
2,000 cases for each disease and 3,000 shared controls

GeneChip 500K Mapping Array Set, comprising 500,568 SNPs

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007

Strong GWAS regions

躁郁症 冠状动脉疾病 克罗恩病



Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007

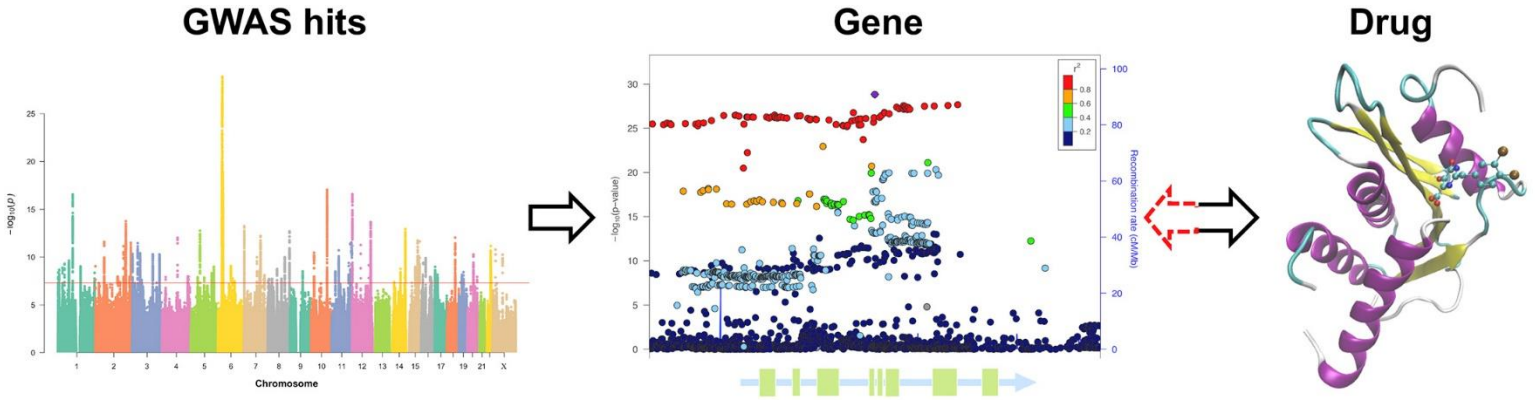
Strong GWAS signals

Collection	Chromosome	Region (Mb)	SNP	Trend *P* value	Genotypic *P* value	$\log_{10}$(BF), additive	$\log_{10}$(BF), general	Risk allele	Minor allele	Heterozygote odds ratio	Homozygote odds ratio	Control MAF	Case MAF
Standard analysis													
BD	16p12	23.3–23.62	rs420259	2.19×10^{-04}	6.29×10^{-08}	1.96	4.79	A	G	2.08 (1.60–2.71)	2.07 (1.6–2.69)	0.282	0.248
CAD	9p21	21.93–22.12	rs1333049	1.79×10^{-14}	1.16×10^{-13}	11.66	11.19	C	C	1.47 (1.27–1.70)	1.9 (1.61–2.24)	0.474	0.554
CD	1p31	67.3–67.48	rs11805303	6.45×10^{-13}	5.85×10^{-12}	10.07	9.41	T	T	1.39 (1.22–1.58)	1.86 (1.54–2.24)	0.317	0.391
CD	2q37	233.92–234	rs10210302	7.10×10^{-14}	5.26×10^{-14}	11.11	11.28	T	C	1.19 (1.01–1.41)	1.85 (1.56–2.21)	0.481	0.402
CD	3p21	49.3–49.87	rs9858542	7.71×10^{-07}	3.58×10^{-08}	4.24	5.22	A	A	1.09 (0.96–1.24)	1.84 (1.49–2.26)	0.282	0.331
CD	5p13	40.32–40.66	rs17234657	2.13×10^{-13}	1.99×10^{-12}	10.41	9.89	G	G	1.54 (1.34–1.76)	2.32 (1.59–3.39)	0.125	0.181
CD	5q33	150.15–150.31	rs1000113	5.10×10^{-08}	3.15×10^{-07}	5.36	5.01	T	T	1.54 (1.31–1.82)	1.92 (0.92–4.00)	0.067	0.098
CD	10q21	64.06–64.31	rs10761659	2.68×10^{-07}	1.75×10^{-06}	4.69	4.13	G	A	1.23 (1.05–1.45)	1.55 (1.3–1.84)	0.461	0.406
CD	10q24	101.26–101.32	rs10883365	1.41×10^{-08}	5.82×10^{-08}	5.91	5.48	G	G	1.2 (1.03–1.39)	1.62 (1.37–1.92)	0.477	0.537
CD	16q12	49.02–49.4	rs17221417	9.36×10^{-12}	3.98×10^{-11}	8.93	8.47	G	G	1.29 (1.13–1.46)	1.92 (1.58–2.34)	0.287	0.356
CD	18p11	12.76–12.91	rs2542151	4.56×10^{-08}	2.03×10^{-07}	5.42	5.00	G	G	1.3 (1.14–1.48)	2.01 (1.46–2.76)	0.163	0.208
RA	1p13	113.54–114.16	rs6679677	4.90×10^{-26}	5.55×10^{-25}	22.36	21.99	A	A	1.98 (1.72–2.27)	3.32 (1.93–5.69)	0.096	0.168
RA	6	MHC	rs6457617*	3.44×10^{-76}	5.18×10^{-75}	74.84	73.18	T	T	2.36 (1.97–2.84)	5.21 (4.31–6.30)	0.489	0.685
T1D	1p13	113.54–114.16	rs6679677	1.17×10^{-26}	5.43×10^{-26}	23.07	22.83	A	A	1.82 (1.59–2.09)	5.19 (3.15–8.55)	0.096	0.169
T1D	6	MHC	rs9272346*	2.42×10^{-134}	5.47×10^{-134}	141.9	142.2	A	G	5.49 (4.83–6.24)	18.52 (27.03–12.69)	0.387	0.150
T1D	12q13	54.64–55.09	rs11171739	1.14×10^{-11}	9.71×10^{-11}	8.89	8.24	C	C	1.34 (1.17–1.54)	1.75 (1.48–2.06)	0.423	0.493
T1D	12q24	109.82–111.49	rs17696736	2.17×10^{-15}	1.51×10^{-14}	12.53	11.88	G	G	1.34 (1.16–1.53)	1.94 (1.65–2.29)	0.424	0.506
T1D	16p13	10.93–11.37	rs12708716	9.24×10^{-08}	4.92×10^{-07}	5.15	4.70	A	G	1.19 (0.97–1.45)	1.55 (1.27–1.89)	0.350	0.297
T2D	6p22	20.63–20.84	rs9465871	1.02×10^{-06}	3.34×10^{-07}	4.15	3.98	C	C	1.18 (1.04–1.34)	2.17 (1.6–2.95)	0.178	0.218
T2D	10q25	114.71–114.81	rs4506565	5.68×10^{-13}	5.05×10^{-12}	10.14	9.43	T	T	1.36 (1.2–1.54)	1.88 (1.56–2.27)	0.324	0.395
T2D	16q12	52.36–52.41	rs9939609	5.24×10^{-08}	1.91×10^{-07}	5.35	5.05	A	A	1.34 (1.17–1.52)	1.55 (1.3–1.84)	0.398	0.453
Multi-locus analysis													
T1D	4q27	123.26–123.92	rs6534347	4.48×10^{-07}	1.83×10^{-06}	5.15	4.69	A	A	1.30 (1.10–1.55)	1.49 (1.25–1.78)	0.351	0.402
T1D	12p13	9.71–9.86	rs3764021	7.19×10^{-05}	5.08×10^{-08}	2.12	4.55	C	T	1.57 (1.38–1.79)	1.48 (1.25–1.75)	0.467	0.426
Sex differentiated analysis													
RA	7q32	130.80–130.84	rs11761231	3.91×10^{-07}	1.37×10^{-06}	-	-	G	A	1.44 (1.19–1.75)	1.64 (1.35–1.99)	0.375	0.327
Combined cases													
RA+T1D	10p15	6.07–6.17	rs2104286	5.92×10^{-08}	2.52×10^{-07}	5.26	4.45	T	C	1.35 (1.11–1.65)	1.62 (1.34–1.97)	0.286	0.245

Regions with at least one SNP with a *P* value of less than 5×10^{-7} for our primary analyses. The $\log_{10}$ value of the Bayes factor (BF) for the bayesian analysis corresponding to the trend and genotypic tests is also given. Region marks the boundaries of signal defined by recombination and return of test statistics to background levels. The minor allele is defined in the controls and its frequency in that group as well as the case sample is reported. MAF, minor allele frequency. Cluster plots for each SNP have been inspected visually, and are shown in Supplementary Fig. 10. Positions are in NCBI build-35 coordinates *Multiple SNPs in the MHC region are significant, we report the most extreme.

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007

From GWAS Discoveries to Drugs



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

10 Years of GWAS Discovery: Biology, Function, and Translation. *AJHG* 2017

GWAS in crops

Published: 24 October 2010

Genome-wide association studies of 14 agronomic traits in rice landraces

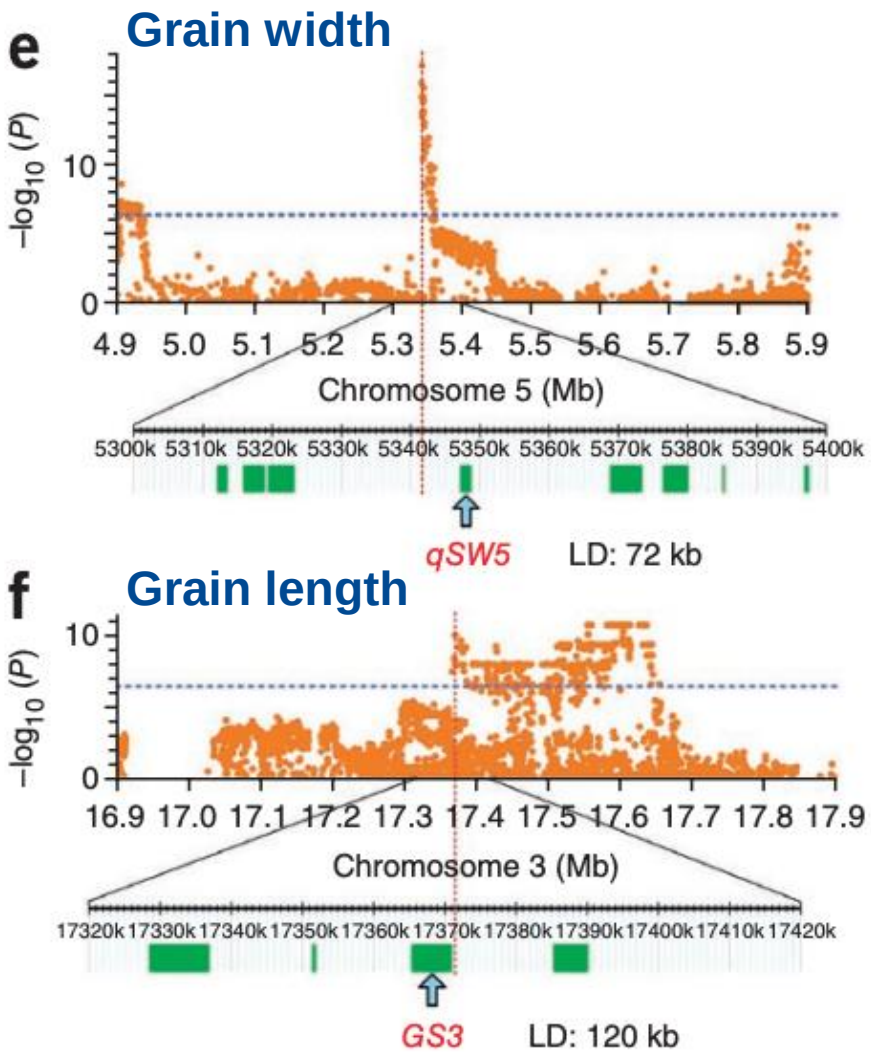
Xuehui Huang, Xinghua Wei, [...] Bin Han

Nature Genetics **42**, 961–967(2010) | Cite this article

~3.6 million SNPs by sequencing 517 rice landraces

Table 1 Genome-wide significant association signals of agronomic traits using the compressed MLM

Trait	Chromosome	Position (IRGSP 4)	Major allele	Minor allele	Minor allele freq.	*P* value (compressed MLM)[a]	Known loci[b]
Tiller number 分蘖数	4	3,760,194	A	T	0.20	3.2×10^{-7}	
	9	23,332,559	A	G	0.34	1.5×10^{-7}	
	10	15,239,407	T	A	0.10	4.1×10^{-7}	
Grain width	5	4,907,158	C	G	0.21	2.7×10^{-9}	
	5	5,341,575	G	A	0.17	7.2×10^{-18}	*qSW5* (ref. 29)
Grain length	3	17,371,398	G	C	0.06	1.3×10^{-10}	*GS3* (ref. 30)
	3	17,637,475	C	A	0.08	2.7×10^{-11}	
	3	23,349,781	A	C	0.13	3.3×10^{-7}	
	5	5,343,949	A	G	0.20	1.7×10^{-7}	
	11	3,072,370	C	T	0.11	3.8×10^{-7}	
Spikelet number 小穗粒数	7	18,005,615	C	T	0.44	7.1×10^{-8}	
	10	5,976,140	C	T	0.06	1.3×10^{-7}	

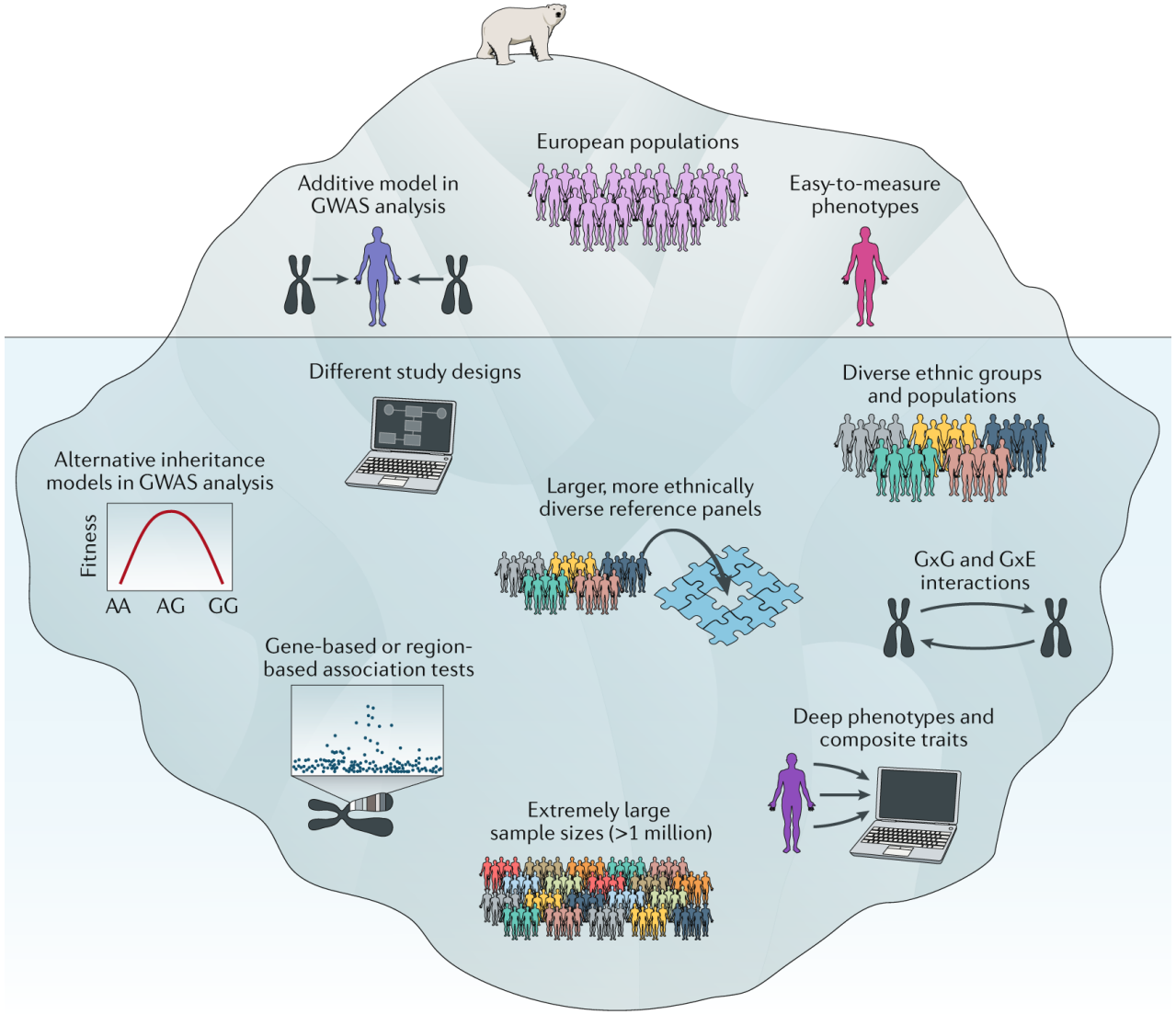


GWAS增长趋势

- 自2007年至今，陆续报道和公布了关于人类身高、体重、血压等主要性状，以及视网膜黄斑、乳腺癌、前列腺癌、白血病、冠心病、肥胖症、糖尿病、精神分裂症等几十种疾病GWAS结果。
- 累计发表论文上万篇，确定了一系列疾病发病的致病基因、相关基因、易感区域和单核苷酸多态位点。

GWAS文章增长趋势图

GWAS发现的仍是冰山一角

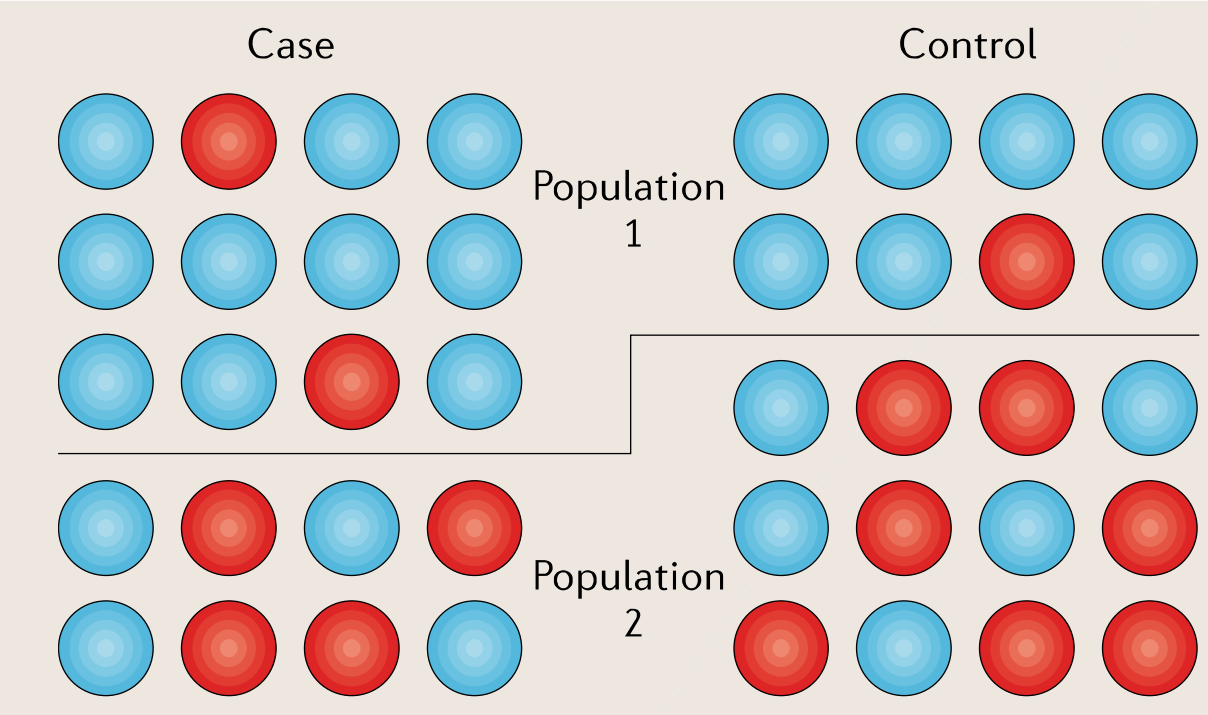


GWAS的研究发现仍然是冰山一角。迄今为止，使用GWAS可以发现的（水面以上的部分）主要是利用易于测量的表型、大部分是欧洲人群、基于加性遗传模型。

冰山下被淹没部分表示通过将当前的GWAS范式扩展到包括更广泛的表型、更大的样本量、更多样的人群和种族群体，以及不同的研究设计和分析。

Confounding variables

- Population stratification: geographic and ethnic background
- Sex, Age



Nature Reviews Genetics 2006

人群分层是导致许多大样本研究出现假阳性或假阴性结果的一个主要原因。

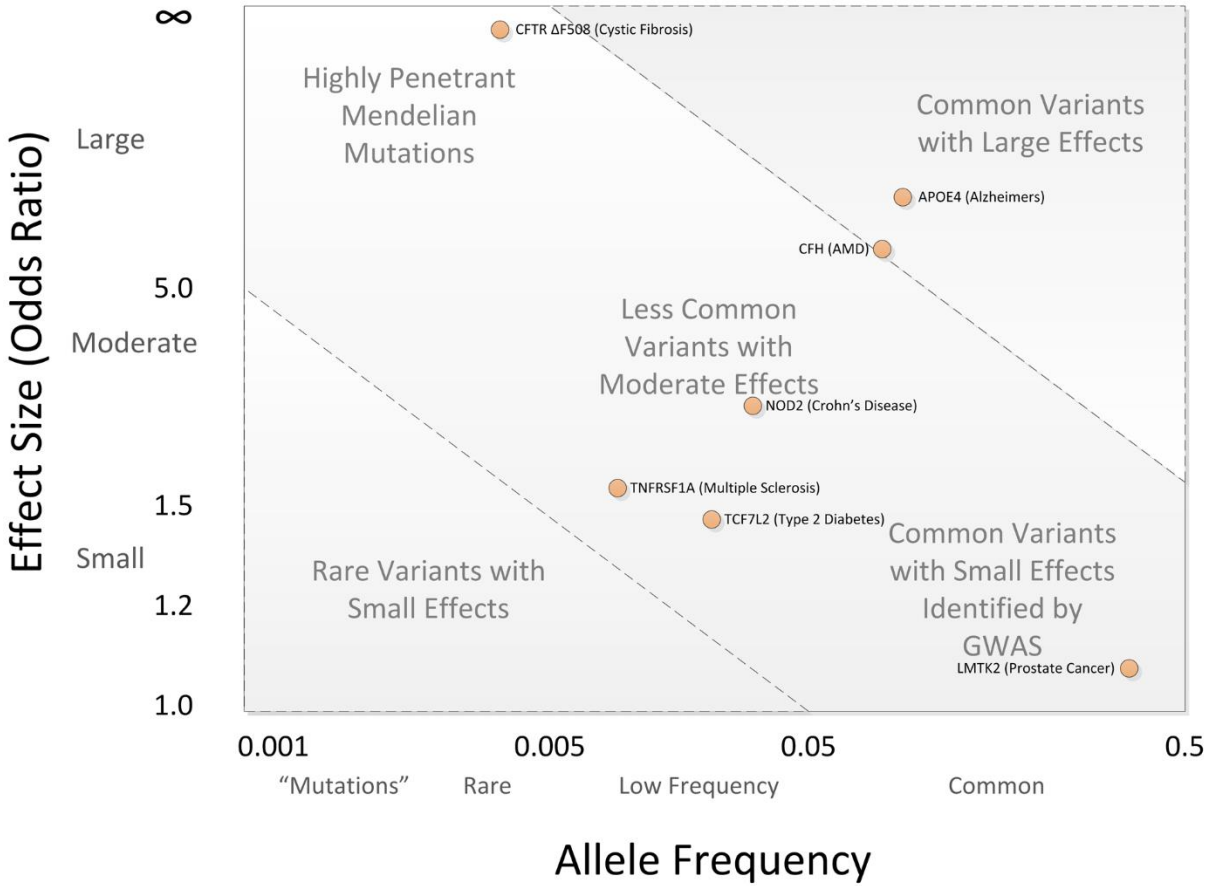
如：Campbell等（2015）采用欧裔美国人研究与身高表型乳糖酶基因型的关联，其结果在其他人群中难以重复即是受研究对象在不同地域存在极大差异引起的人群分层影响。

GWAS局限性

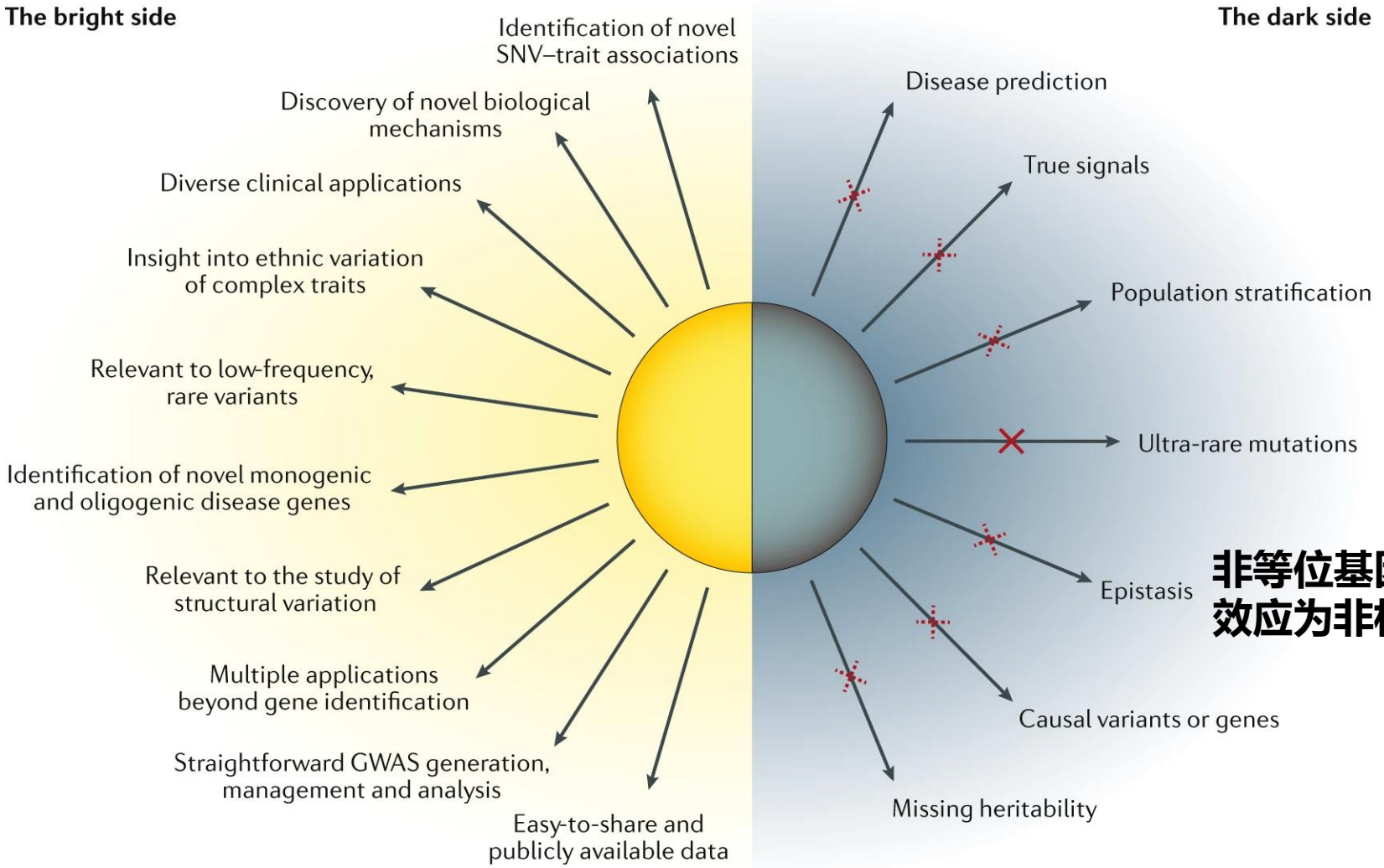
GWAS是一种发现符合“常见疾病-常见变异假说”（common disease-common variant hypothesis）相关位点的方法。

其可以确定相关位点但不能直接确定基因本身，且在任何特定人群中GWAS都不能方便地识别罕见的风险等位基因位点

Disease associations are often conceptualized in two dimensions: **allele frequency and effect size**. Highly penetrant alleles for Mendelian disorders are extremely rare with large effect sizes (upper left), while most GWAS findings are associations of common SNPs with small effect sizes (lower right). **The bulk of discovered genetic associations lie on the diagonal denoted by the dashed lines.**



全基因组关联分析-优缺点



非等位基因的遗传效应为非相加性

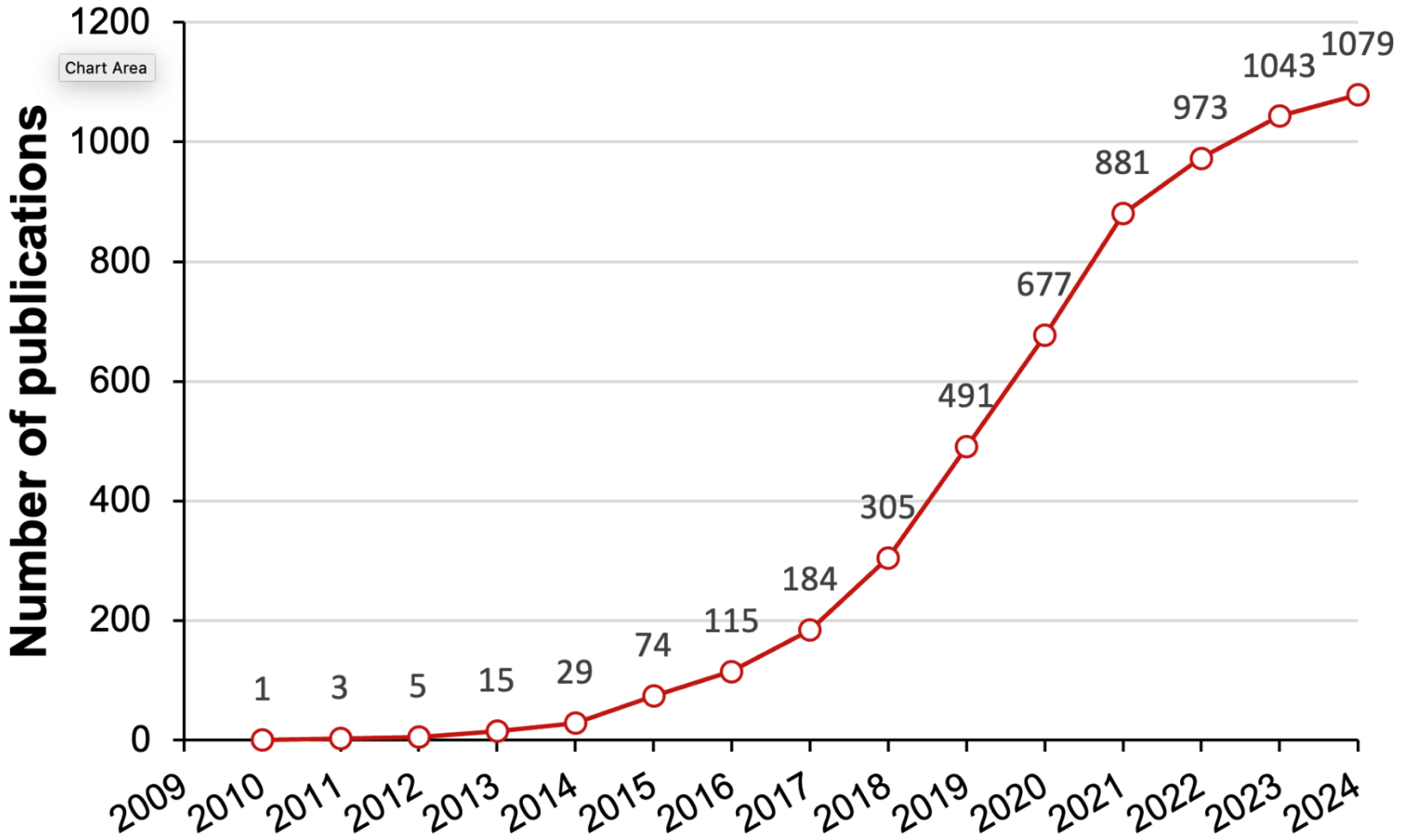
Benefits and Limitations of Genome-Wide Association Studies. *Nature Review Genetics* 2019

EWAS, TWAS, PWAS, PheWAS

表观基因组关联分析

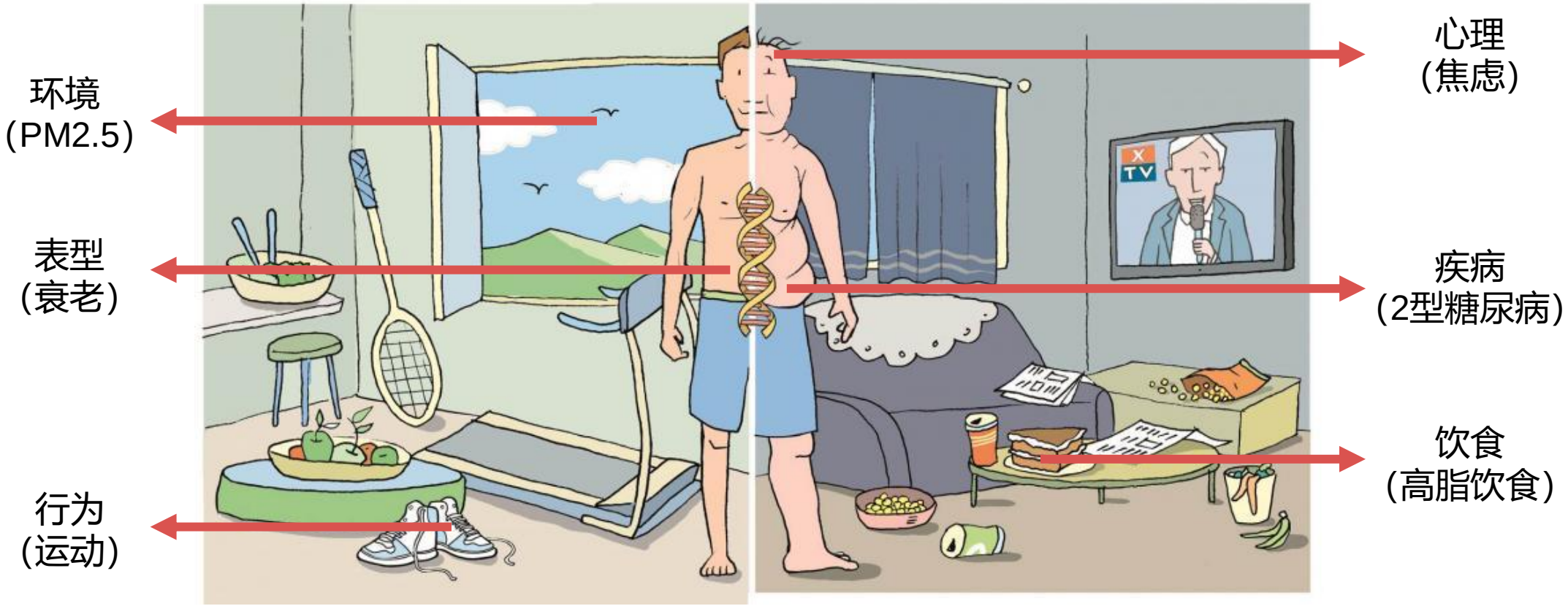
- EWAS（Epigenome-wide association study，表观基因组关联分析）
- 与GWAS形成互补，将表观变异和复杂疾病或性状进行关联，在表观基因组学层面对复杂疾病的致病原因或性状关联进行解读，找到与致病原因相关或复杂性状相关的表观遗传学变异位点。
- 通过检测整个基因组成千上万特异表观修饰差异（比如DNA核苷酸上甲基分布），来鉴别出疾病中常见的表观突变或与复杂性状密切相关的表观修饰。

表观组（DNA甲基化）关联研究



https://ngdc.cncb.ac.cn/ewas/statistics

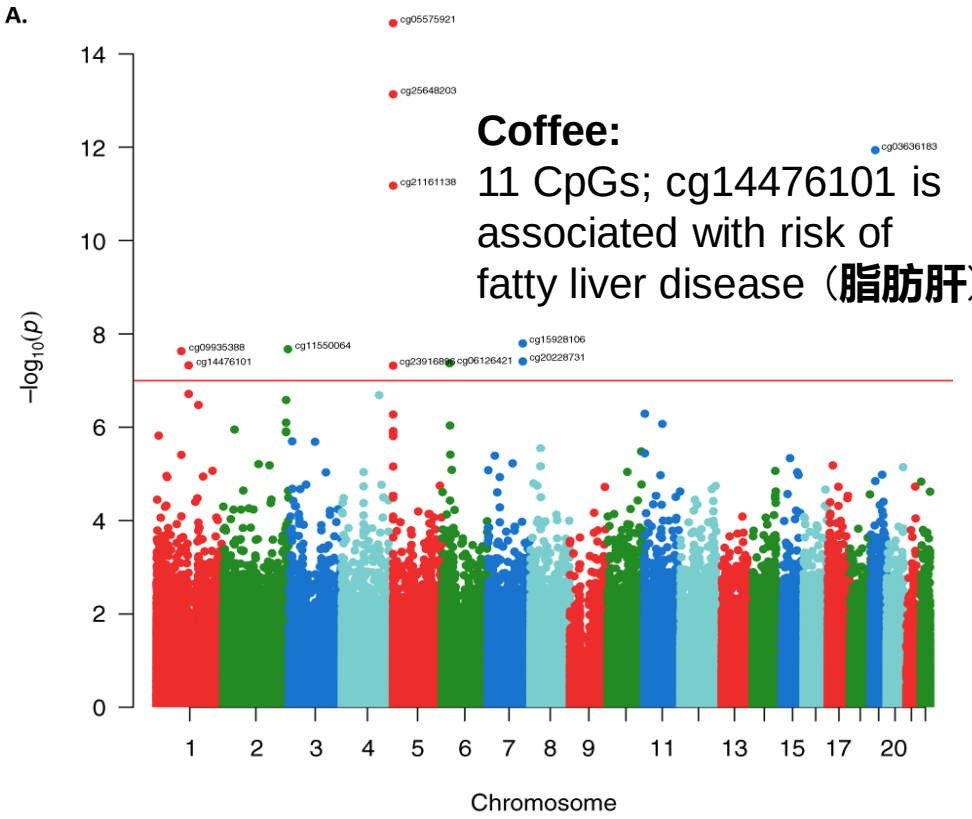
表观修饰与环境、表型、行为



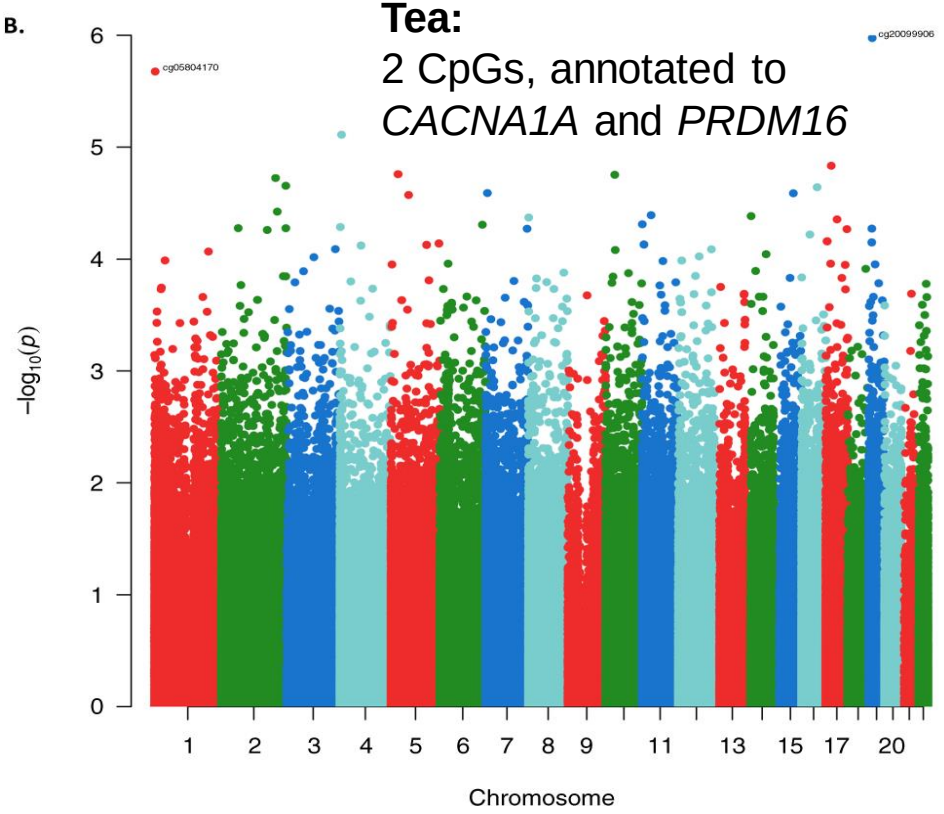
表观修饰是表型、环境和行为的纽带

EWAS on coffee and tea consumption

15,789 participants of European and African-American ancestries from 15 cohorts



Coffee:
11 CpGs; cg14476101 is associated with risk of fatty liver disease（**脂肪肝**）



Tea:
2 CpGs, annotated to *CACNA1A* and *PRDM16*

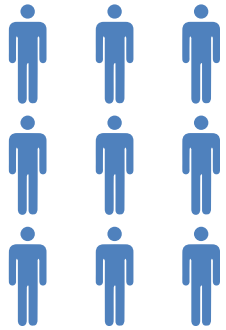
Coffee-associated changes in DNA methylation levels may explain the mechanism of action of coffee consumption in conferring risk of diseases.

Epigenome-wide association meta-analysis of DNA methylation with coffee and tea consumption. *Nature Communications* 2021

DNA甲基化是表观组关联研究的热点

- Dynamic Nature（动态性）
  - 绝大多数环境和行为因素都能影响DNA甲基化
- Quantitative Accuracy（定量准确）
  - DNA甲基化的定量能达到1%分辨率
- Reversibility （可逆性）
  - 受环境或行为影响发生的DNA甲基化改变具有可逆性

表观基因组关联分析



样本

样本信息

表观修饰测序/芯片检测

关联：

- Case-control
- 回归（年龄）

得到相关位点

下游分析

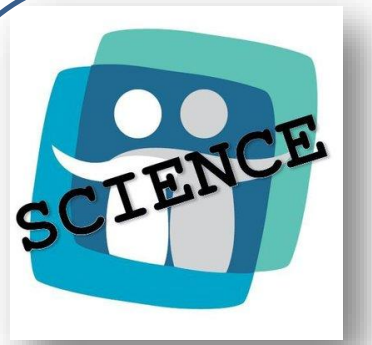
DNA甲基化检测

	DNA甲基化芯片	全基因组甲基化测序
分析时间	两分钟	两天
测序成本	3000元	20000元/30X
捕获区域	~850,000个位点	全基因组
上样量	250 ng DNA	100 ng DNA
测序深度	相当于500X左右	通常不超过100X

芯片适合**大规模快速检测**

全基因组甲基化测序适合**探索性研究**

人群队列（cohort）



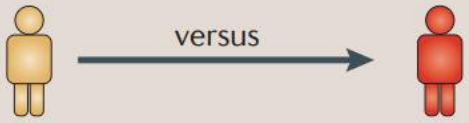
NTR
11,989 人
已追踪32年

TwinsUK
14,274 对双胞胎
已追踪27年

EPIC study
521,000人
已追踪15年

- 大型人群队列研究是医学逐步走向精准，提高治疗效果的必经之路
- 针对不同的问题，需要设计不同的队列

不同类型的队列

	Key advantage	Key disadvantage
Case versus control (singletons) 	Many cohorts exist	Cannot easily control for environmental and genetic confounders
Families 	Can study potential inheritance	Few large cohorts of this type exist
Disease-discordant monozygotic twins 	Can control for genetics	Few large cohorts of this type exist
Prospectively sampled, longitudinal 	Can establish causality	Slow and difficult to establish

Case-control

家族队列

双胞胎队列

长期追踪队列

EWAS研究的难点

- 由于表观修饰容易受环境、行为和表型的影响，因此在选择样本时需要控制这些因素
- 表观修饰水平受细胞比例的影响，研究时需要分析细胞比例是否变化
- 表观修饰的变化不一定发挥功能，对因果关系的解析非常重要

EWAS典型案例：癌症检测

前列腺癌

SEPT9
结直肠癌

肺癌

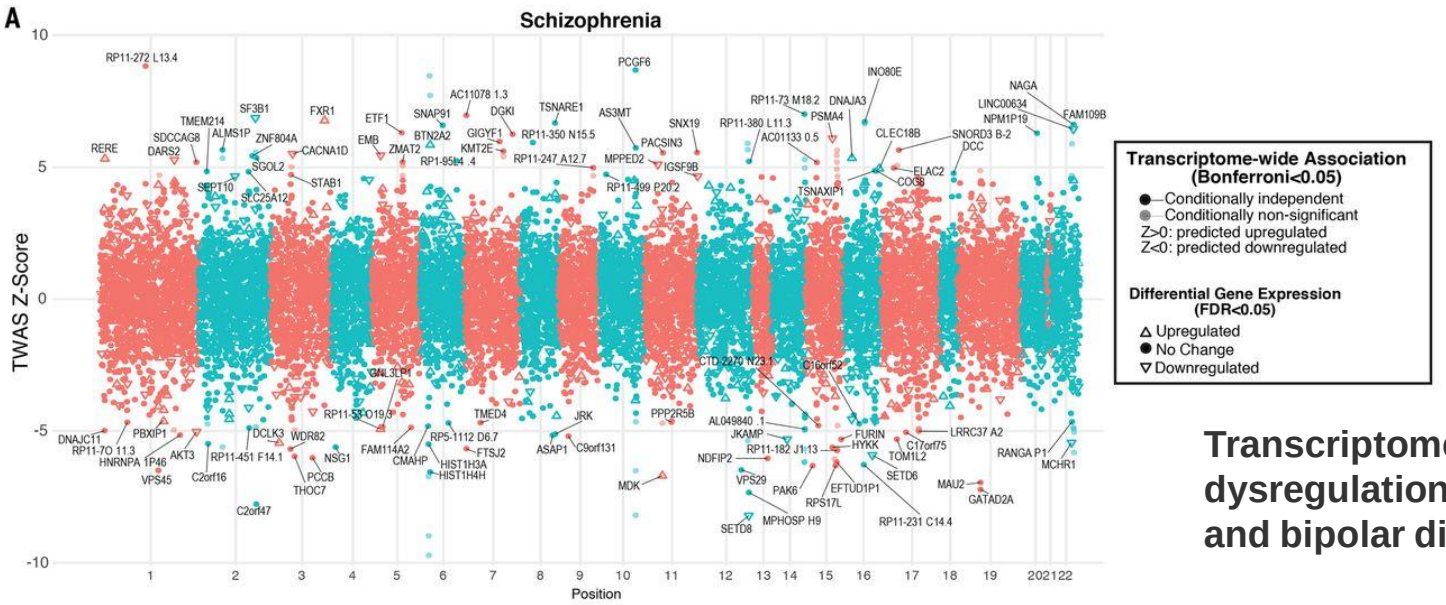
膀胱癌

MGMT
脑胶质瘤

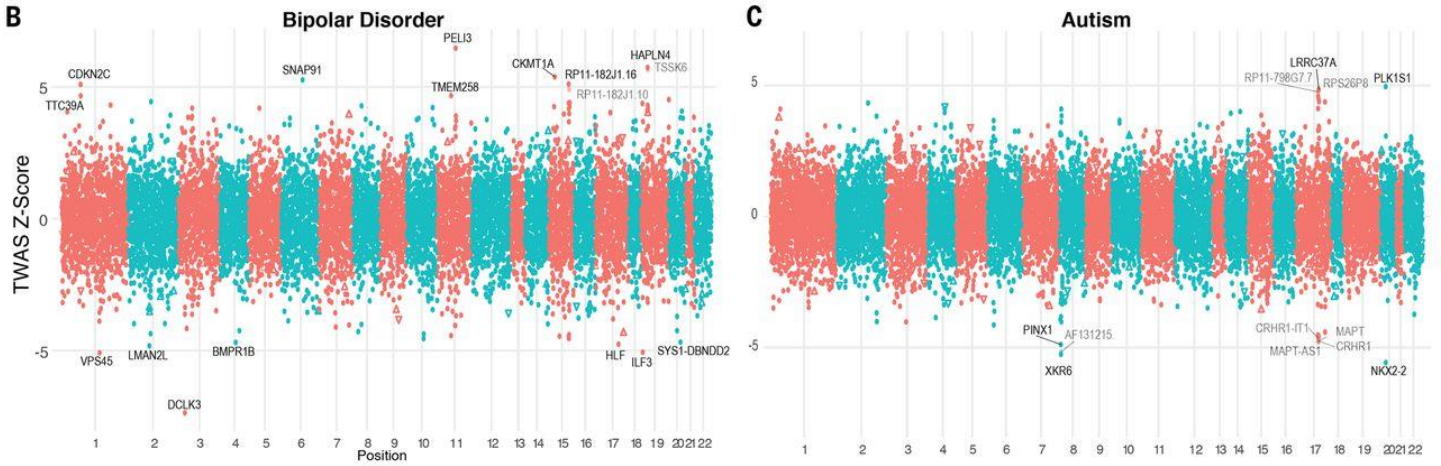
Genes	Indication	Hypermethylated in tumour?	Correlation with gene expression?	Promoter region?	CpG island?	Clinical guideline?	FDA approved?
ConfirmMDx							
GSTP1	Diagnosis and prognosis of prostate cancer	Yes	Yes	Yes	Yes	Yes[2]	No
APC		Yes	No	No	No		
RASSF1		Yes	No	Yes	Yes		
Cologuard							
NDRG4	Early detection of CRC	Yes	Yes	Yes	Yes	Yes[3–5]	Yes
BMP3		Yes	Yes	Yes	Yes		
Epi proColon/ColoVantage (Licensed to Quest Diagnostics and ARUP Laboratories)[a]							
SEPT9[a]	Early detection of CRC	Yes	No	No	Yes	Yes[4,7]	Yes
Epi proColon[a]							
SEPT9[a]	Early detection of CRC	Yes	No	No	Yes	Yes[4,7]	Yes
Epi proLung							
SHOX2	Early detection of lung cancer	Yes	Yes[b]	No	Yes	No	No
AssureMDx							
TWIST1	Risk of bladder cancer in patients with haematuria	Yes	Yes	Yes	Yes	No	No
OTX1		Yes	Yes[b]	No	Yes		
ONECUT2		Yes	Yes[b]	No	Yes		
PredictMDx							
MGMT	Prediction of response to temozolomide in glioblastoma	NA[c]	Yes	Yes	Yes	Yes[6]	No
Colvera							
BCAT1	Early detection of CRC recurrence	Yes	Yes	Yes	Yes	No	No
IKZF1		Yes	Yes	Yes	Yes		

Transcriptome-wide association study

193 genes are prioritized at Bonferroni-corrected $P < 0.05$.



Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* 2018



17 genes

12 genes

Proteome-wide association study

Letter | Published: 28 January 2021

Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis

Aliza P. Wingo, Yue Liu, Ekaterina S. Gerasimov, Jake Gockley, Benjamin A. Logsdon, Duc M. Duong, Eric B. Dammer, Chloe Robins, Thomas G. Beach, Eric M. Reiman, Michael P. Epstein, Philip L. De Jager, James J. Lah, David A. Bennett, Nicholas T. Seyfried, Allan I. Levey & Thomas S. Wingo

Nature Genetics **53**, 143–146 (2021) | Cite this article

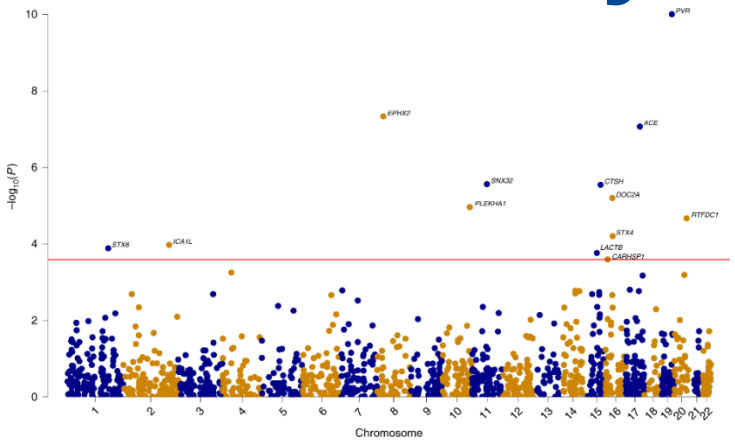


Table 1 | The discovery AD PWAS identified 13 significant genes, of which 10 were found in the confirmation PWAS, and all 10 replicated

			Discovery PWAS			Confirmation PWAS		Evidence for replication
	Gene	Chromosome	PWAS *z*-score	PWAS *P*	PWAS FDR *P*	PWAS *z*-score	PWAS *P*	
1	*ACE*	17	−5.36	8.5×10^{-8}	4.2×10^{-5}	−5.28	1.3×10^{-7}	Yes
2	*EPHX2*	8	5.46	4.7×10^{-8}	3.4×10^{-5}	4.68	2.8×10^{-6}	Yes
3	*SNX32*	11	−4.69	2.8×10^{-6}	8.4×10^{-4}	−4.27	2.0×10^{-5}	Yes
4	*DOC2A*	16	−4.51	6.4×10^{-6}	1.6×10^{-3}	−4.23	2.3×10^{-5}	Yes
5	*LACTB*	15	3.76	1.7×10^{-4}	2.1×10^{-2}	4.08	4.5×10^{-5}	Yes
6	*ICA1L*	2	−3.88	1.1×10^{-4}	1.6×10^{-2}	−3.96	7.5×10^{-5}	Yes
7	*CARHSP1*	16	3.66	2.6×10^{-4}	2.9×10^{-2}	3.48	5.1×10^{-4}	Yes
8	*RTFDC1*	20	4.25	2.1×10^{-5}	3.9×10^{-3}	3.10	2.0×10^{-3}	Yes
9	*STX6*	1	3.83	1.3×10^{-4}	1.7×10^{-2}	2.96	3.1×10^{-3}	Yes
10	*CTSH*	15	4.68	2.9×10^{-6}	8.4×10^{-4}	2.36	1.8×10^{-2}	Yes
11	*PLEKHA1*[a]	10	4.40	1.1×10^{-5}	2.3×10^{-3}	-	-	-
12	*PVR*[b]	19	−10.94	7.1×10^{-28}	1.0×10^{-24}	-	-	-
13	*STX4*[b]	16	4.00	6.2×10^{-5}	1.0×10^{-2}	-	-	-

This table gives the *z*-scores for the AD PWAS associations with their corresponding *P* values and FDR-adjusted *P* values for all significant genes in the AD discovery PWAS. Confirmation AD PWAS *z*-scores and their corresponding unadjusted *P* values are provided for the significant genes in the discovery AD PWAS. [a]Protein not profiled in the confirmation proteomic dataset. [b]Proteins profiled but without significant SNP-based heritability estimates in the confirmation proteomic dataset.

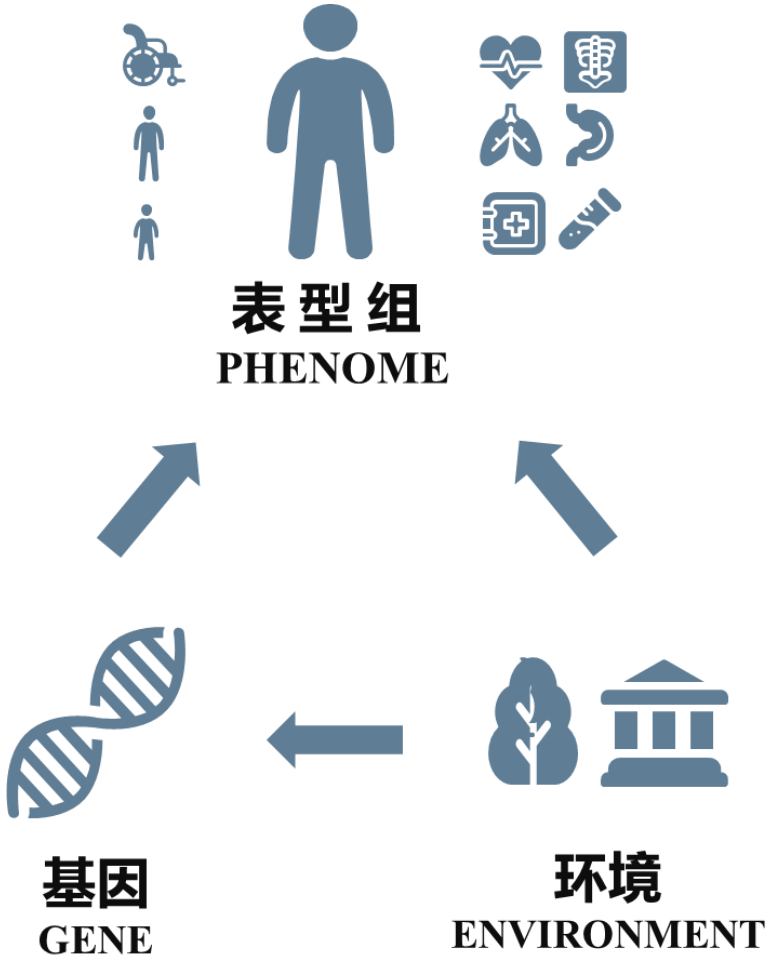
- In the **discovery** phase, we performed a PWAS by integrating AD GWAS results ($n = 455{,}258$) with 376 human brain proteomes profiled from the dorsolateral prefrontal cortex (大脑背外侧前额叶皮层).
- A **confirmation** PWAS was performed using the same AD GWAS and an independent set of 152 human brain proteomes.

Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. *Nature Genetics* 2021

人类表型组

表型（Phenotype）：基因和环境相互作用决定的人体特征

表型组（Phenome）：是生物体从胚胎发育到出生、成长、衰老乃至死亡过程中，形态特征、功能、行为、分子组成规律等所有生物、物理和化学特征的集合



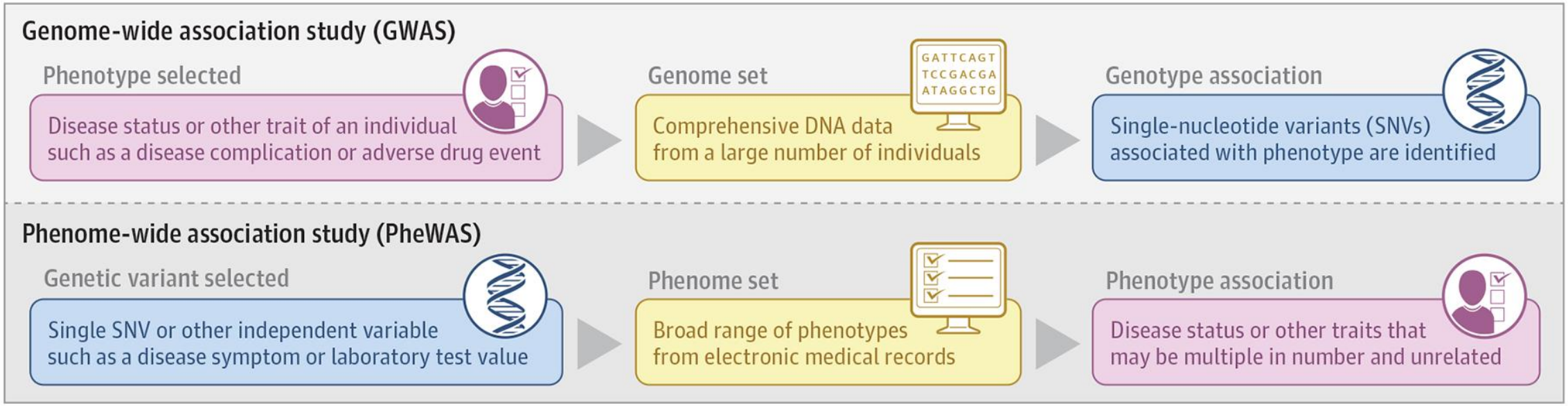
国际人类表型组计划

每人超2万个指标的全景表型测量

- 精密测量人体表型，全景解析人类表型组
- 系统解构表型之间强关联，构建表型网络
- 明确宏观表型与微观表型之间的跨尺度关联

Phenome-wide association study

Phenome-wide association studies (PheWAS) invert the idea of a GWAS by searching for phenotypes associated with specific SNVs across the range of thousands of human phenotypes, or the “phenome”. Analogous to GWAS, PheWAS have shown that specific genetic variations may be associated with multiple phenotypes.

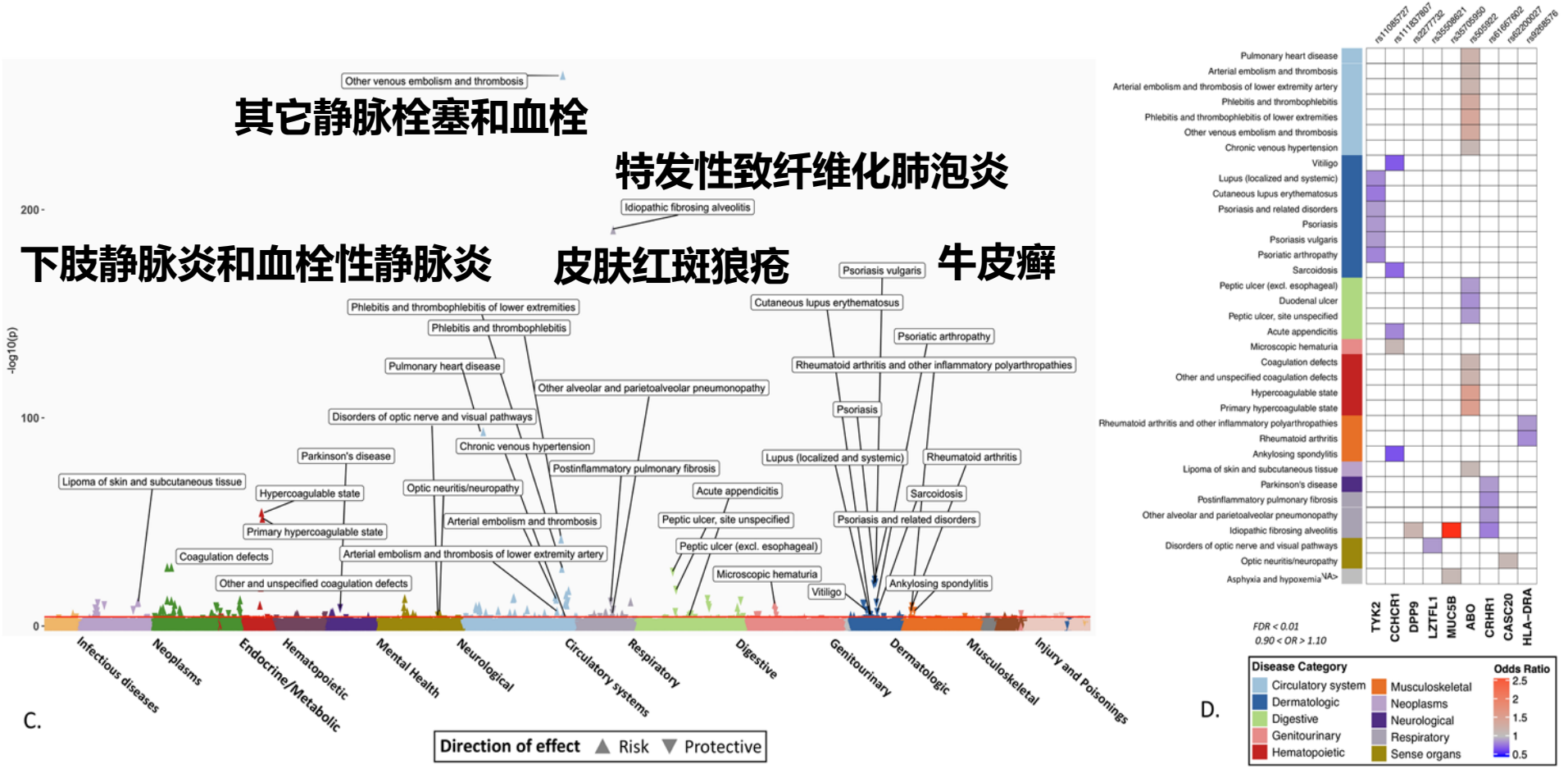


- A GWAS starts with families or populations in which individuals have been assigned affected or unaffected status for a disease or other trait, and searches for associated genetic variants.
- **A PheWAS starts with a genetic variant and searches across a set of curated phenotypes (the “phenome”) to identify associated phenotypes**.

Phenome-Wide Association Studies. *JAMA* 2022 https://dx.doi.org/10.1001/jama.2021.20356

PheWAS of COVID-19 risk alleles

Variants associated with severe COVID-19 were tested for association across **1,559 phenotypes**. Among 67 respiratory conditions tested, 11 had significant associations including MUC5B locus (rs35705950) with increased risk of **idiopathic fibrosing alveolitis**.



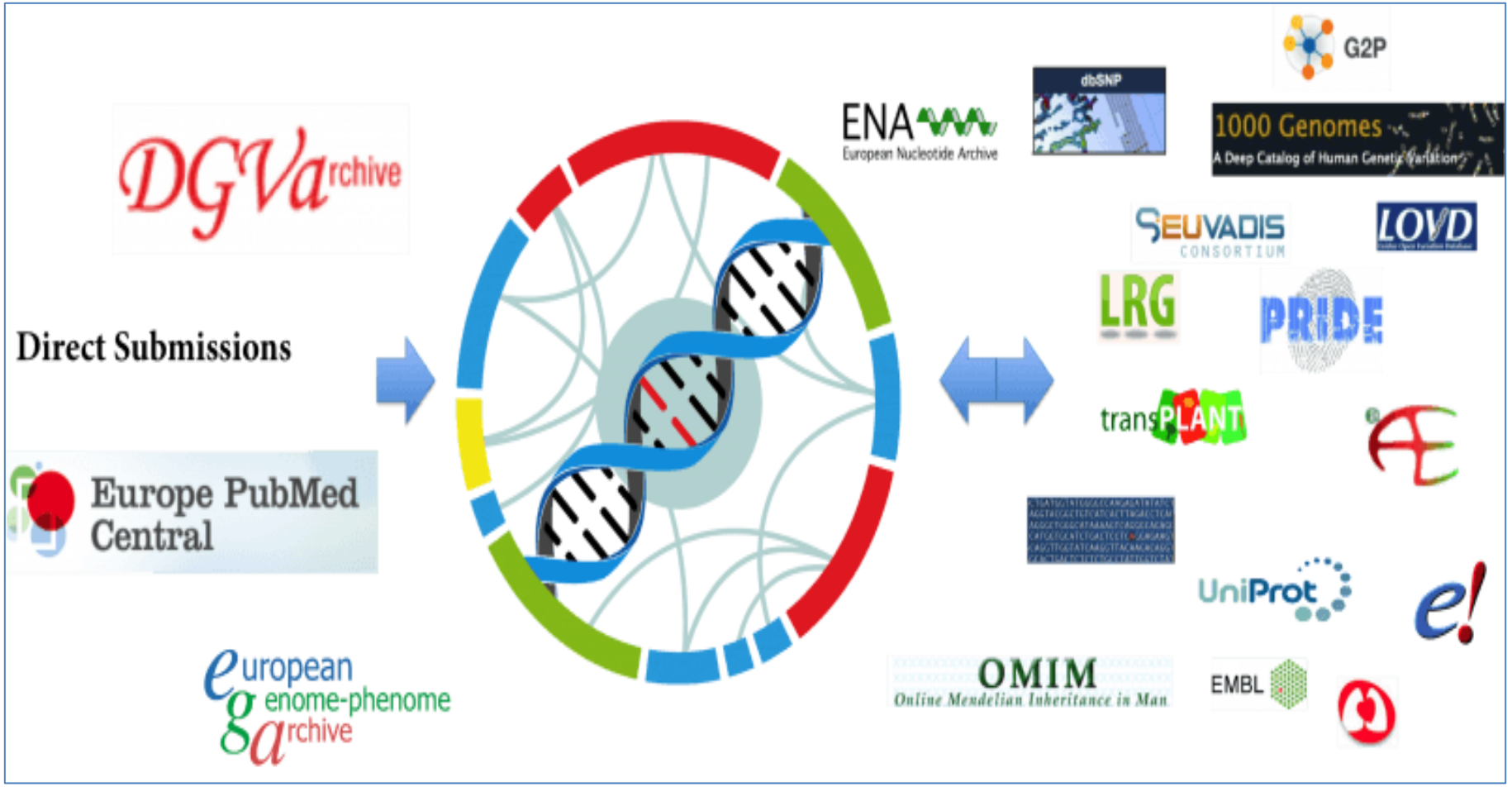
A Phenome-Wide Association Study of genes associated with COVID-19 severity reveals shared genetics with complex diseases in the Million Veteran Program. *PLoS Genetics* 2022

关联知识库

全基因组/表观组关联知识库

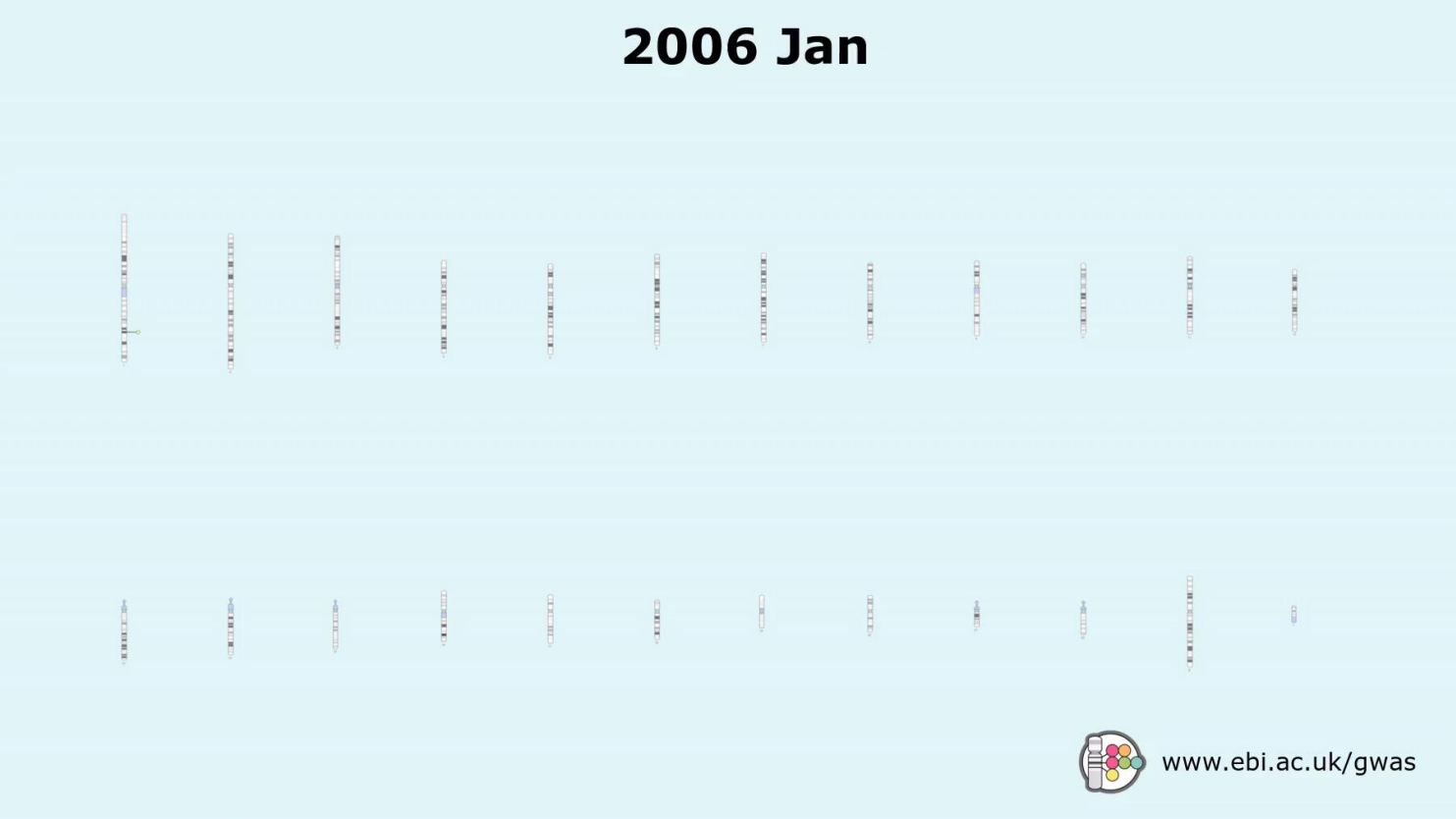
Database & URL	Publications	Associations	Traits	Species	Year
dbGaP https://www.ncbi.nlm.nih.gov/gap		136,527		1	2007
GWAS Catalog http://www.ebi.ac.uk/gwas/	3378	68,000		1	2014
GWASdb http://jjwanglab.org/gwasdb	2479	252,530	1610	1	2012
PhenoScanner http://www.phenoscanner.medschl.cam.ac.uk/phenoscanner		>350000000		1	2016
iPhemap http://www.iphemap.org					2017
AraGWAS Catalog https://aragwas.1001genomes.org		222000	167	1	2018
EWASdb http://www.bioapp.org/ewasdb/					2018
EWAS Atlas http://bigd.big.ac.cn/ewas/	401	329,172	305	1	2018
GWAS Atlas https://bigd.big.ac.cn/gwas/	254	75467	614	9	2019

GWAS Catalog



人类全基因组关联知识数据可免费访问，并广泛应用在多个关键数据库中，可通过常用的数据门户（如：ENSEMBL、UCSC、PheGenI和许多工具）获得

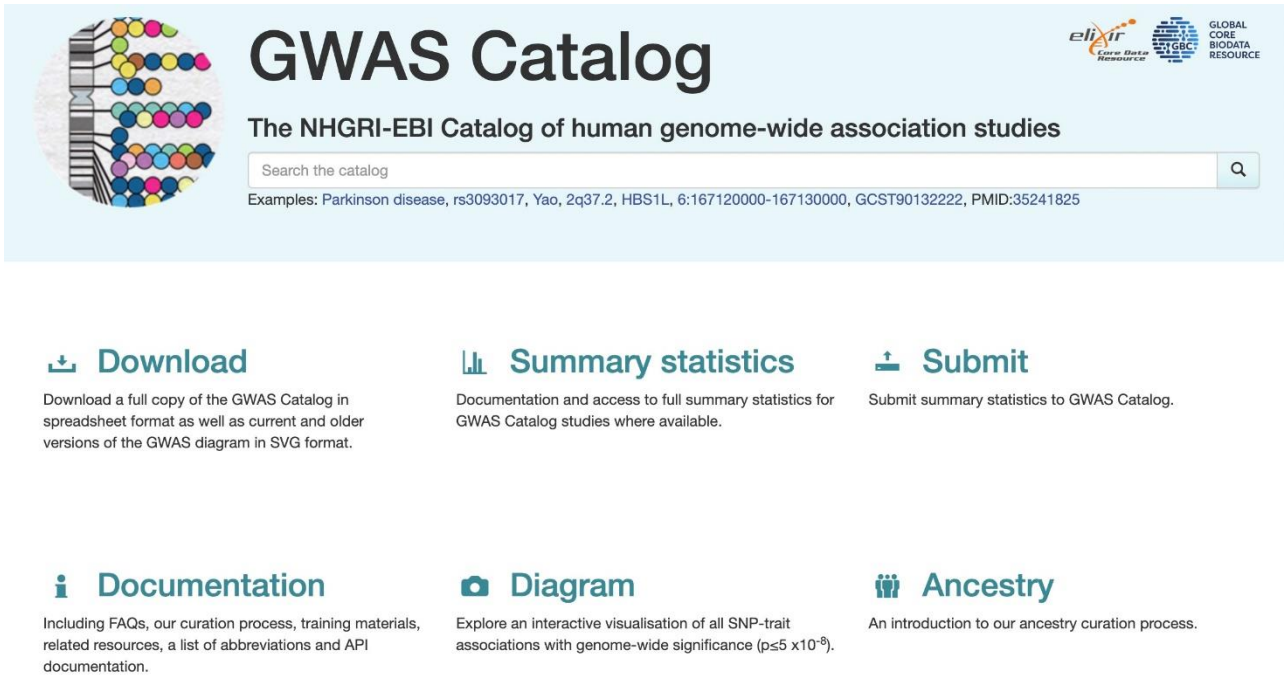
GWAS Catalog

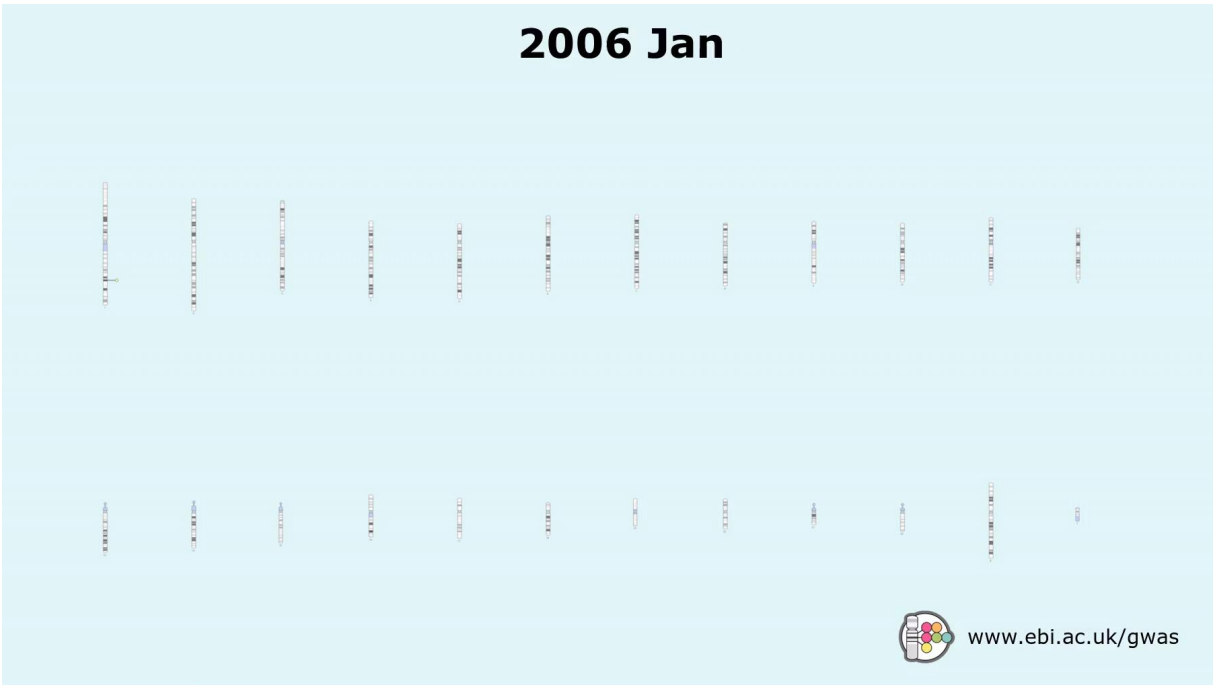


53820 Studies, **6367** Publications, **257947** Associations

GWAS发现的与人类性状或复杂疾病关联SNP位点约6.8万

GWAS Catalog



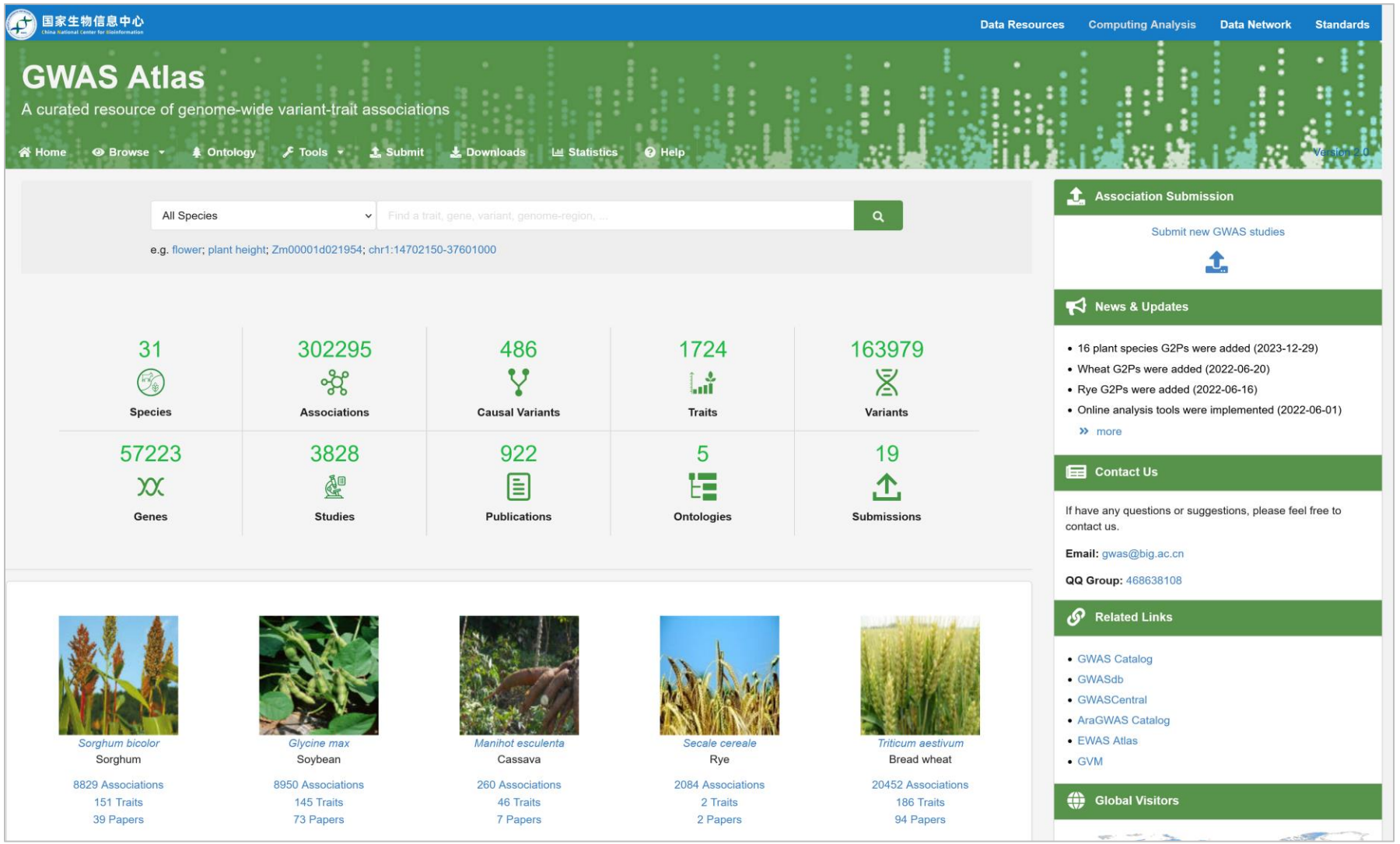


53820 Studies, **6367** Publications, **257947** Associations

GWAS发现的与人类性状或复杂疾病关联SNP位点约6.8万

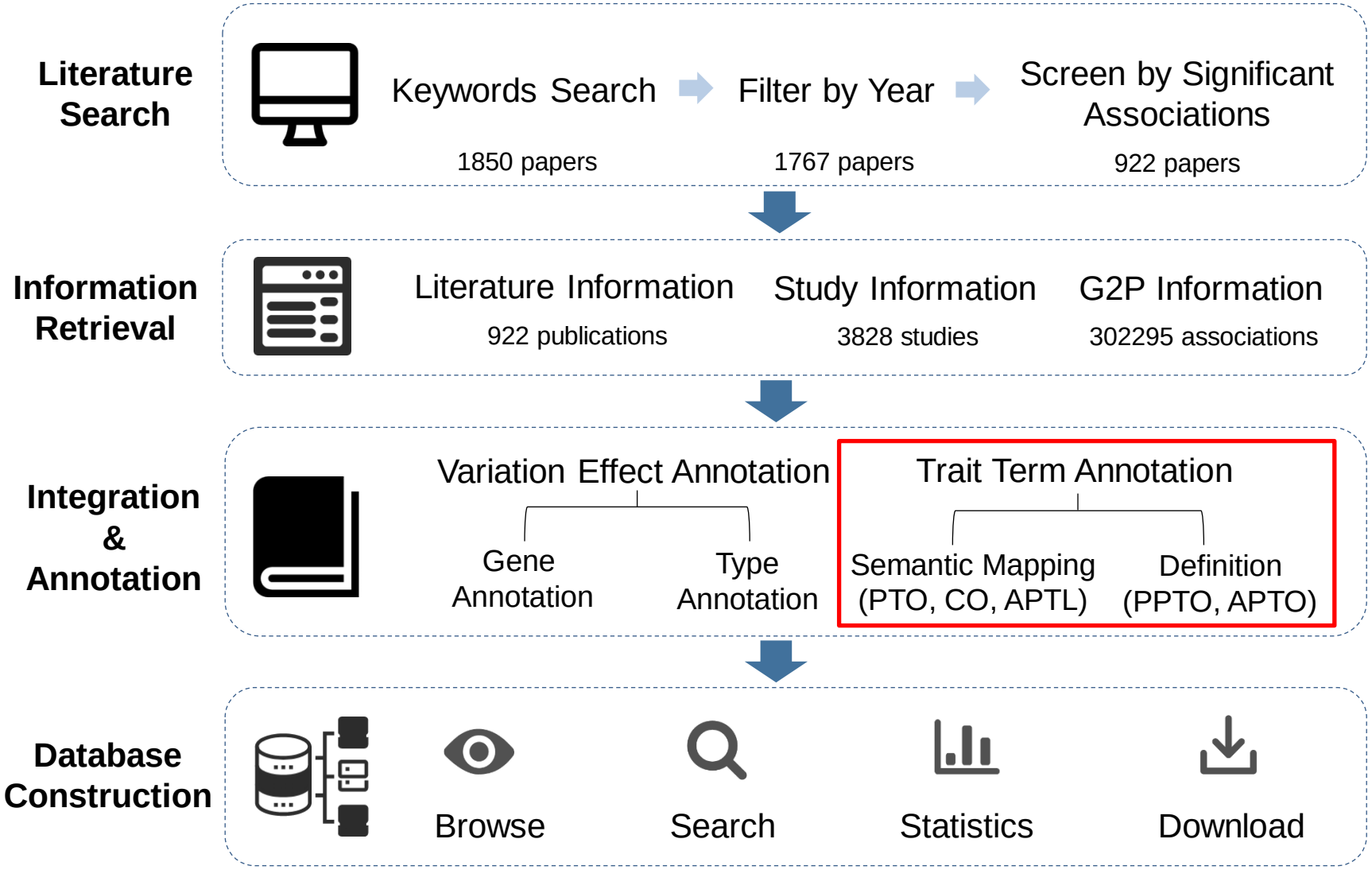
https://www.ebi.ac.uk/gwas/docs/diagram-downloads

GWAS Atlas



GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals. *Nucleic Acids Res* 2020, 2023 https://ngdc.cncb.ac.cn/gwas/

GWAS Atlas – 审编流程



GWAS Atlas – 数据统计信息

Species	# Publication	# Study	# Trait	# Association	# Variant	# Gene
Cotton (Gossypium hirsutum)	4	23	24	21955	6115	991
Japanese apricot (Prunus mume)	2	3	10	1865	1556	625
Maize (Zea mays)	153	485	314	38127	30245	9280
Rapeseed (Brassica napus)	26	88	67	2395	1708	1791
Rice (Oryza sativa)	219	1184	461	163479	74455	22412
Sorghum (Sorghum bicolor)	39	176	151	8829	6805	5490
Soybean (Glycine max)	73	324	145	8950	6148	5123
Goat (Capra hircus)	15	19	51	971	857	59
Pig (Sus scrofa)	67	87	146	1767	1494	509
Sheep (Ovis aries)	35	62	70	800	706	314
Chicken (Gallus gallus)	64	103	130	3782	2701	1096
Cattle (Bos taurus)	19	36	50	1651	1481	256

https://ngdc.cncb.ac.cn/gwas/browse/species

GWAS Atlas – 关联信息

Home > Associations

Show filters

Cotton (21955) | Japanese apricot (1865) | Maize (38127) | Oilseed rape (3137) | **Rice (163479)** | Sorghum (8829) | Soybean (8950) | Cassava (260) | Rye (2084)

Bread wheat (20452) | Field mustard (197) | Cowpea (2195) | Poplor (2315) | Sweet potato (56) | Mung bean (4667) | Common bean (2065) | Tartary Buckwheat (117)

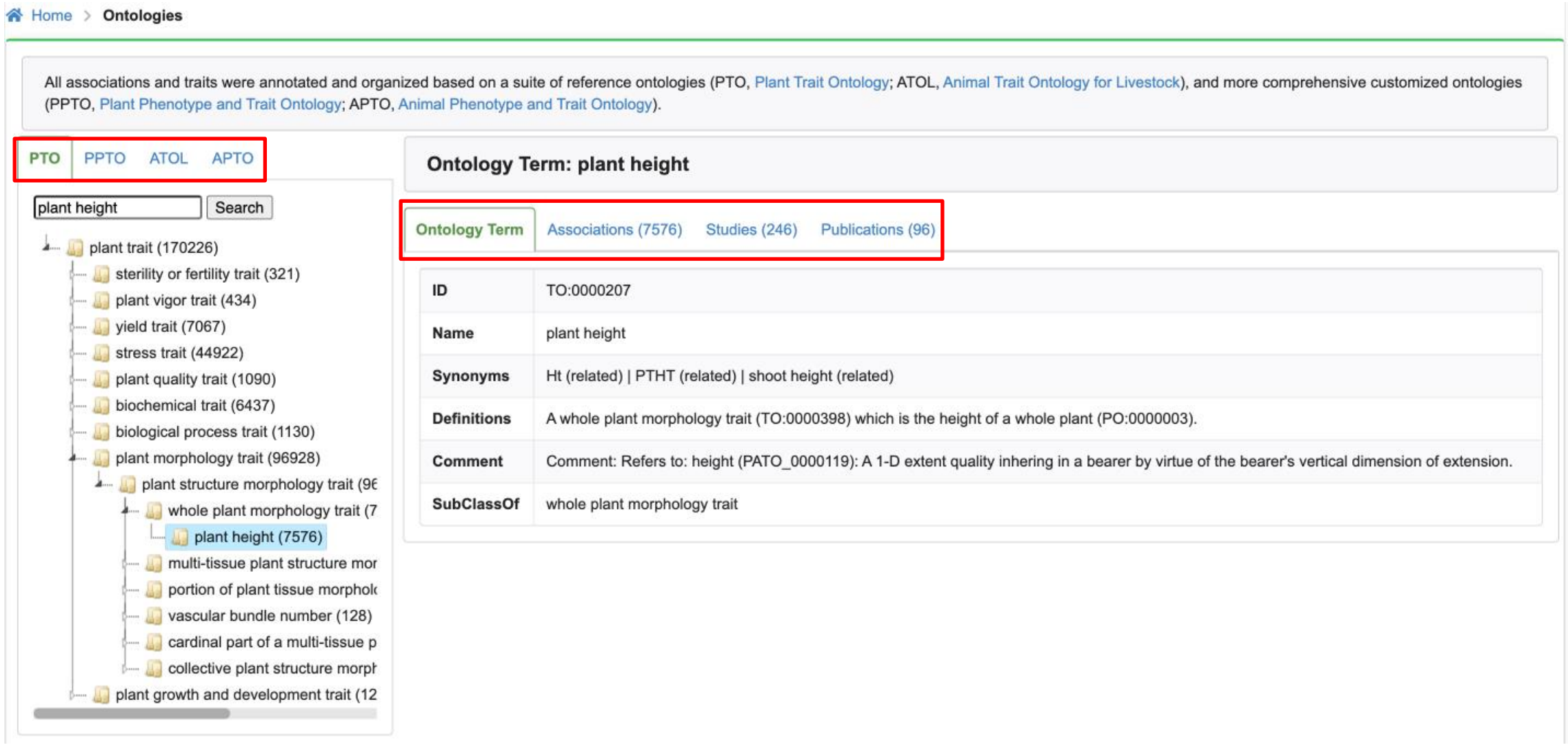
Foxtail millet (699) | Chickpea (1989) | Tomato (623) | Cucumber (775) | Potato (650) | Barley (227) | Grape (1587) | Apple (4898) | Muskmelon (1126)

Cattle (1651) | Chicken (3782) | Goat (971) | Pig (1767) | Sheep (800)

Show 10 entries Search:

Variant ID	Genomic Location	Traits	*P*-values	R^2(%)	Genes (Consequence Types)	Lead SNPs	PMID
osa10000416	chr6:27887714	percentage of head rice recovery	1.94E-4	-	-	-	30512035
osa10001374	chr6:27906467	straighthead disorder severity rate	5.13E-8	-	Os06g0672400 (upstream_gene_variant)	-	35371188
osa10001374	chr6:27906467	straighthead disorder severity rate	1.45E-6	-	Os06g0672400 (upstream_gene_variant)	-	35371188
osa10001652	chr6:27910816	tiller number	3.38E-6	-	-	-	34804103
osa10001700	chr6:27911777	tiller number	5.34E-6	-	Os06g0672550 (downstream_gene_variant)	-	34804103
osa10001888	chr6:27914235	tiller number	4.95E-6	-	Os06g0672550 (downstream_gene_variant)	-	34804103
osa10002046	chr6:27916496	the xylem area of large vascular bundle	4.1E-13	5.07	Os06g0672550 (upstream_gene_variant)	Leader	33288021

GWAS Atlas – 性状本体



GWAS Atlas – 基因

Home > Genes

Show filters

Cotton (991) | Japanese apricot (625) | Maize (9280) | Oilseed rape (2135) | **Rice (22412)** | Sorghum (5490) | Soybean (5123) | Cassava (259) | Rye (23)

Bread wheat (6464) | Field mustard (210) | Cowpea (6) | Poplor (462) | Sweet potato (55) | Common bean (1) | Tartary Buckwheat (36) | Foxtail millet (672)

Tomato (484) | Grape (261) | Cattle (256) | Chicken (1096) | Goat (59) | Pig (509) | Sheep (314)

Show 10 entries Search:

Gene ID	Gene Symbols	Genomic Location	Description	Variants	# Traits	Traits	# Associations	# Studies	# Publications
Os01g0100100	-	chr1:2983-10815	RabGAP/TBC domain containing protein.	osa398	1	secondary internode length	2	2	1
Os01g0100200	-	chr1:11218-12435	Conserved hypothetical protein.	osa398	1	secondary internode length	2	2	1
Os01g0100300	-	chr1:11372-12284	Cytochrome P450 domain containing protein.	osa398	1	secondary internode length	2	2	1

根据基因组位点的基因注释信息，汇总基因与表型的关联信息，方便用户查询浏览

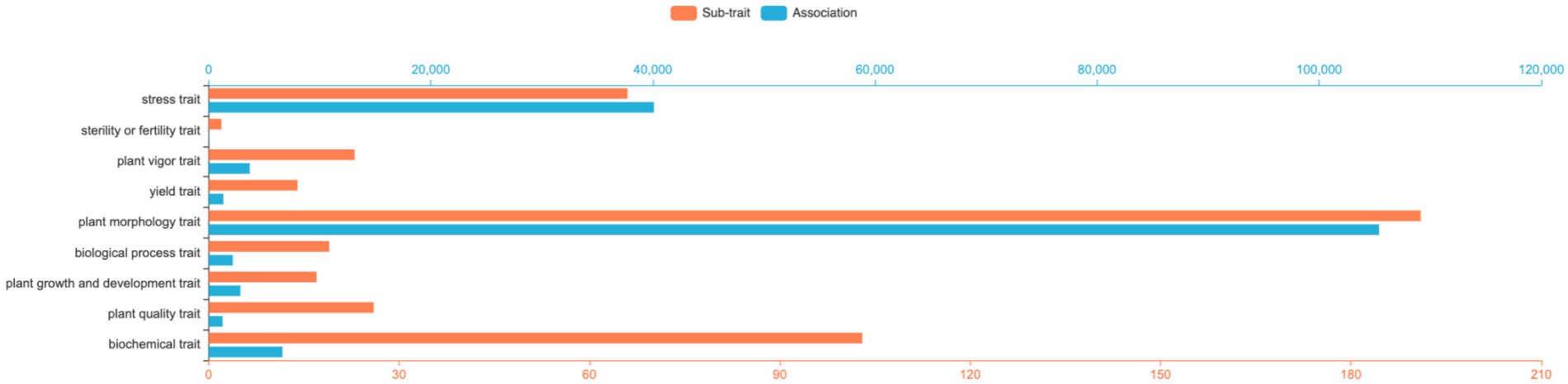
GWAS Atlas – 性状

Home > Traits

Cotton (24) | Japanese apricot (10) | Maize (314) | Oilseed rape (74) | **Rice (461)** | Sorghum (151) | Soybean (145) | Cassava (46) | Rye (2) | Bread wheat (186) | Field mustard (6) | Cowpea (15)

Poplor (74) | Sweet potato (6) | Mung bean (59) | Common bean (67) | Tartary Buckwheat (10) | Foxtail millet (60) | Chickpea (37) | Tomato (47) | Cucumber (13) | Potato (39) | Barley (8) | Grape (8)

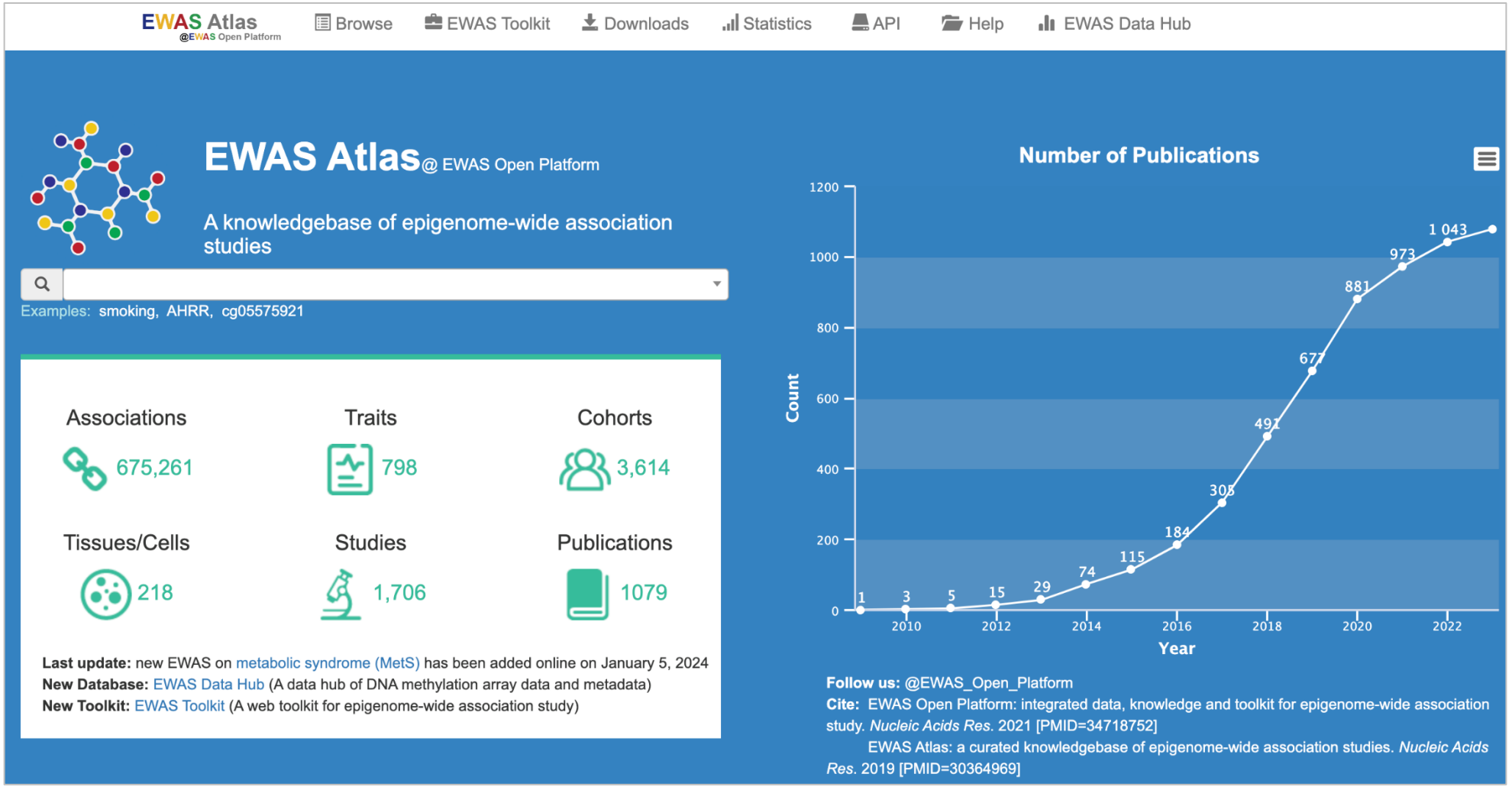
Apple (20) | Muskmelon (3) | Cattle (50) | Chicken (130) | Goat (51) | Pig (146) | Sheep (70)



Show 10 entries Search:

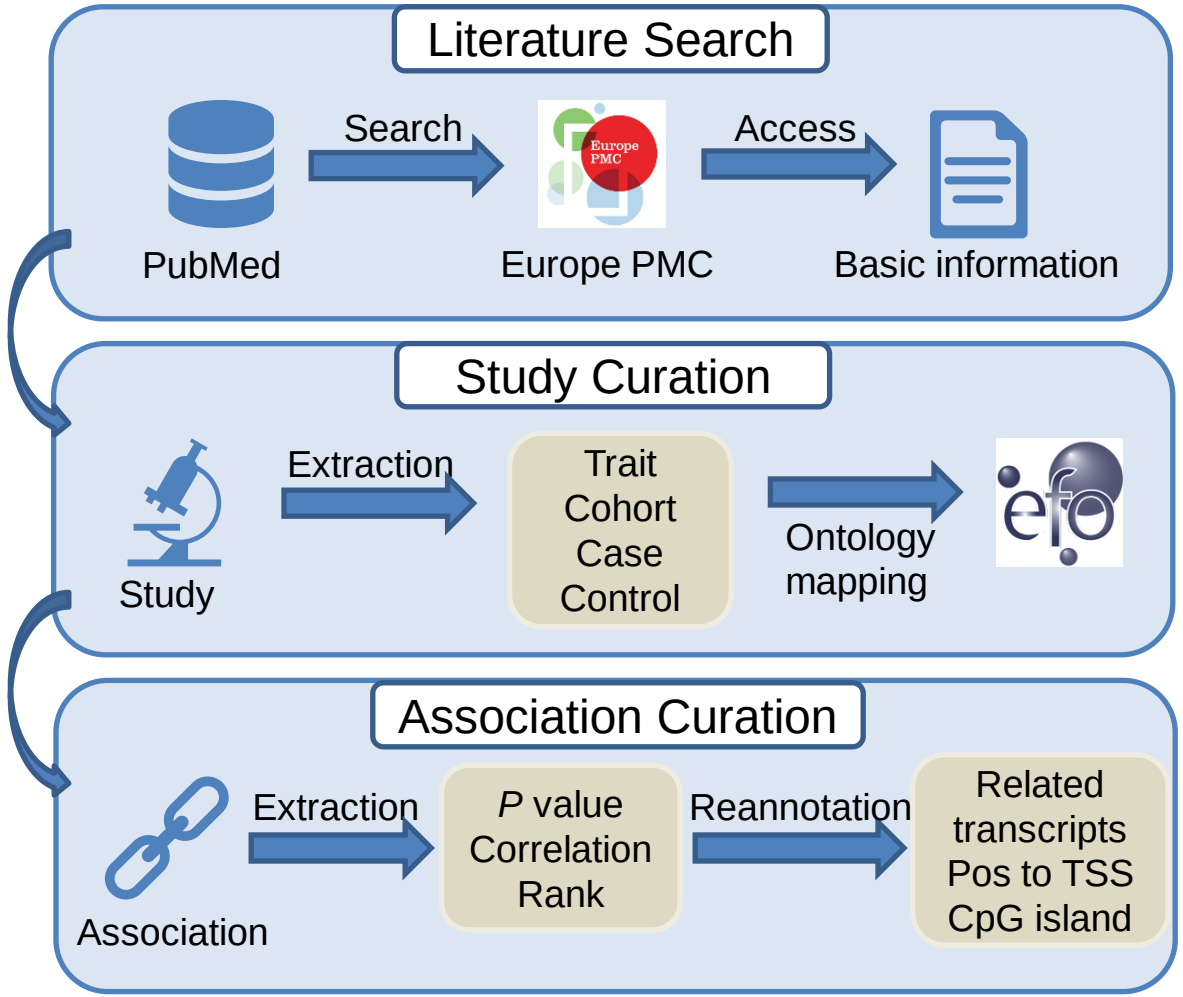
Trait	Trait ID	Mapped Terms	Description	# Associations	# Studies	# Publications
100-dehulled grain weight	PPTO:0000234	TO:0000591	A dehulled grain weight (TO:0000590) which is the weight of a sample of 100 dehulled grains (caryopsis fruit; PO:0030104).	8	1	1
100-grain weight	PPTO:0000117	TO:0000269	A grain weight (TO:0000919) which is the weight of a sample of 100 grains (caryopsis fruit; PO:0030104), with the caryopsis hull (PO:0006000).	44	10	4
1000-grain weight	PPTO:0000305	TO:0000592 TO:0000382 TO:0000595	Weight of the 1000-dehulled grains having pericarp (seed coat).	1615	48	18

EWAS Atlas

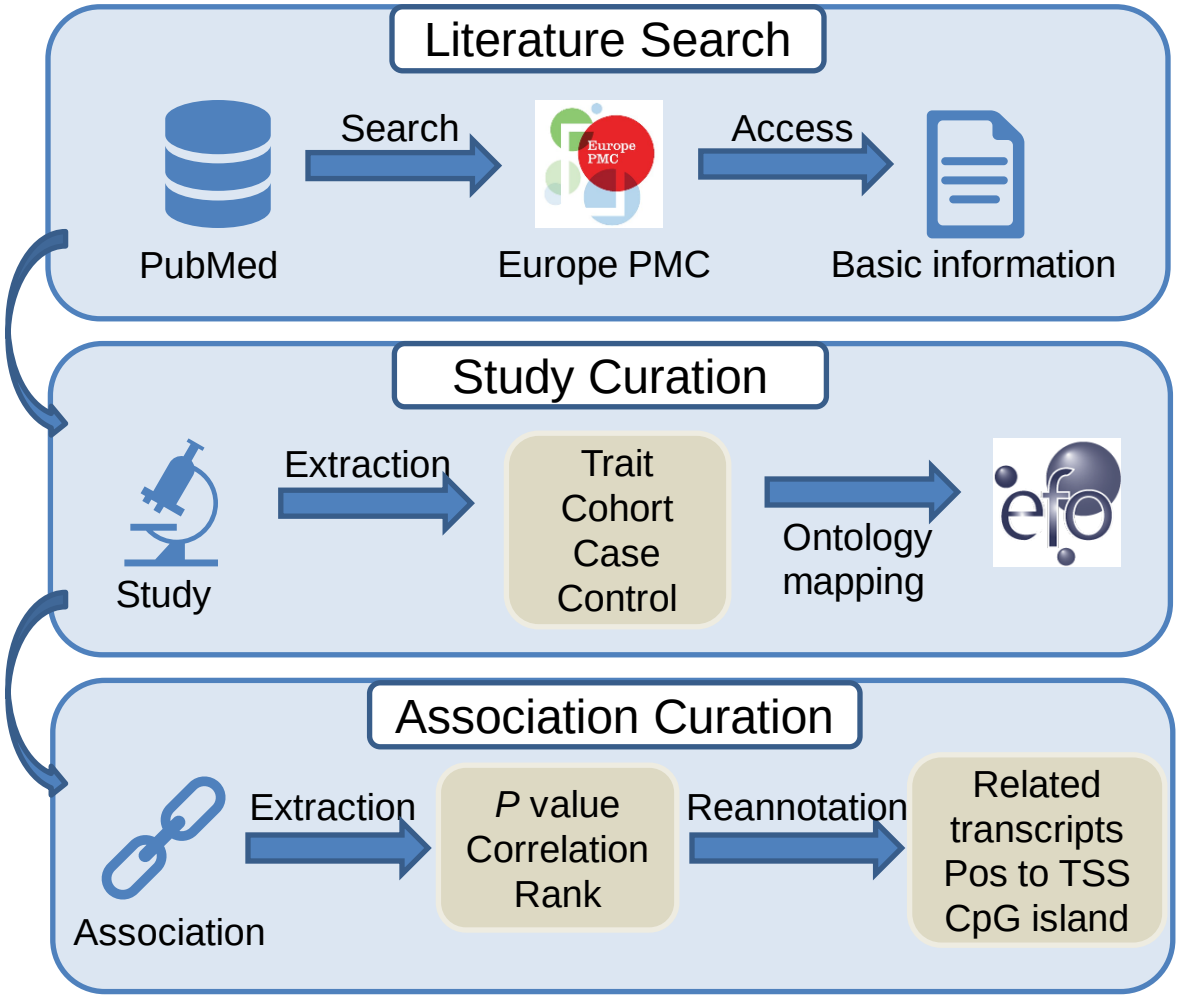


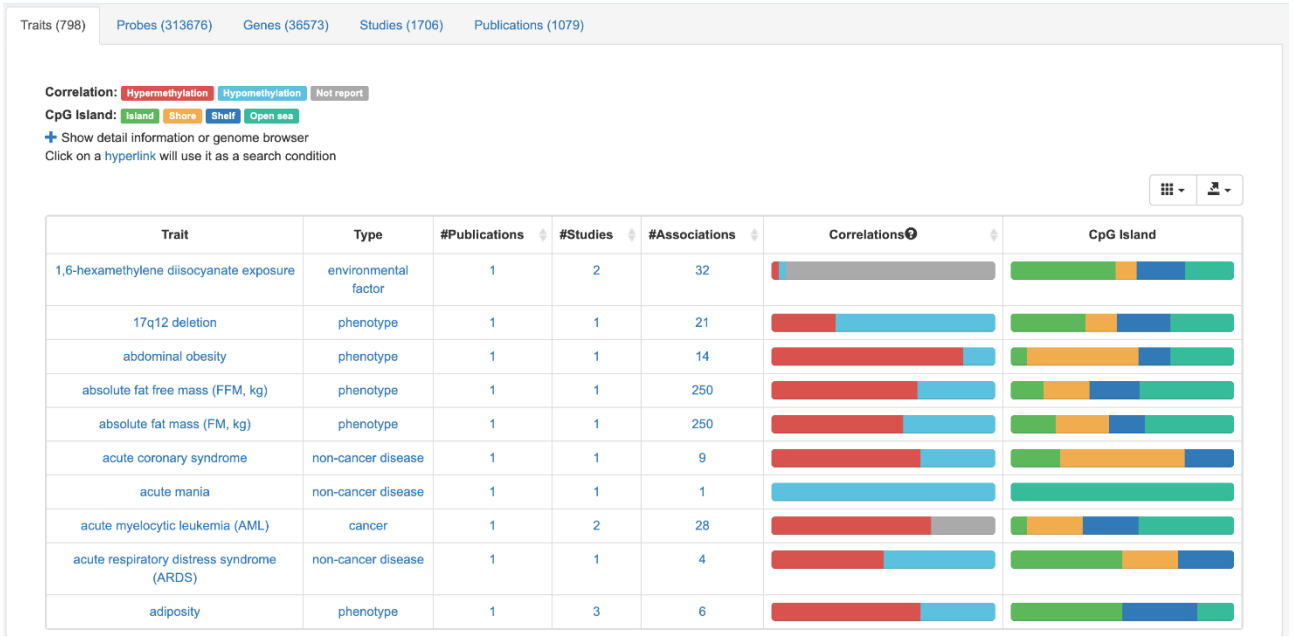
EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res* 2019. https://ngdc.cncb.ac.cn/ewas

EWAS Atlas审编流程



EWAS Atlas审编流程



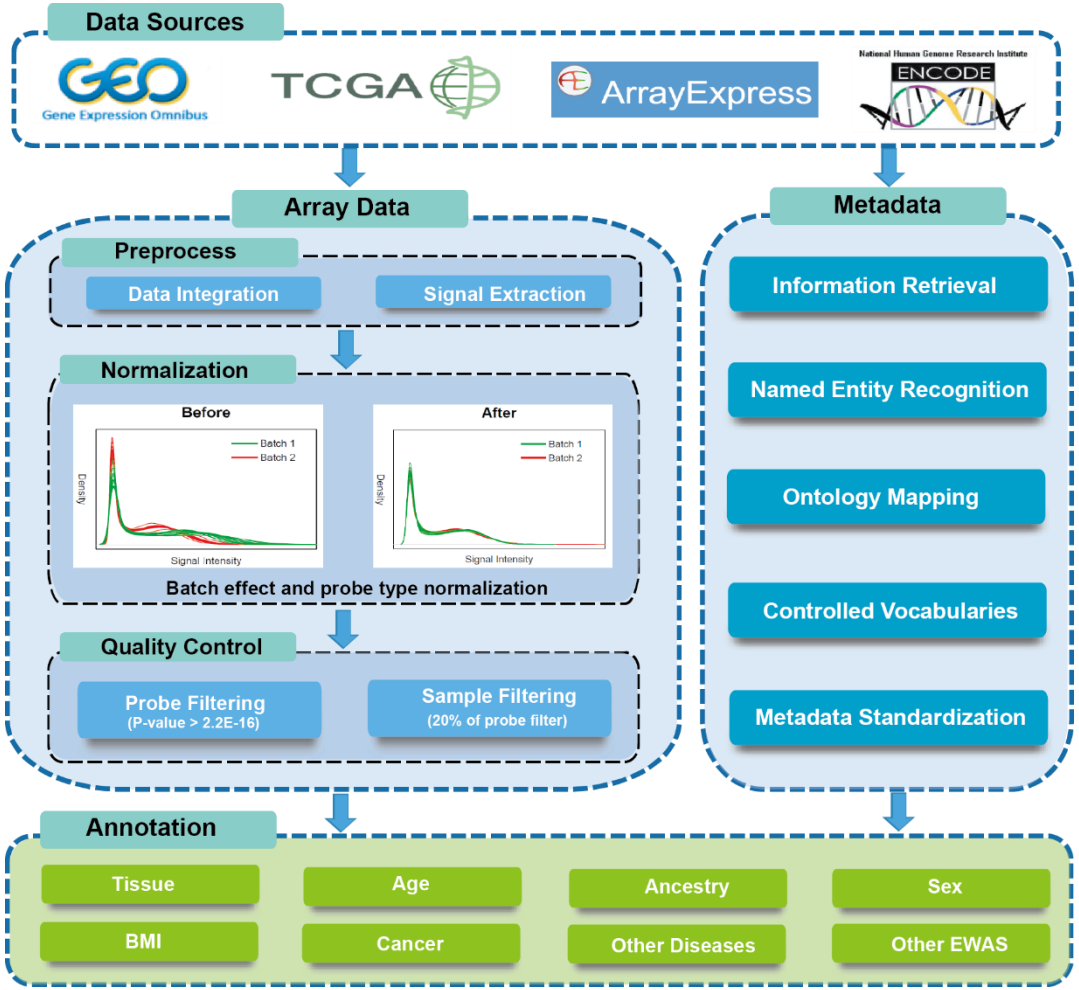


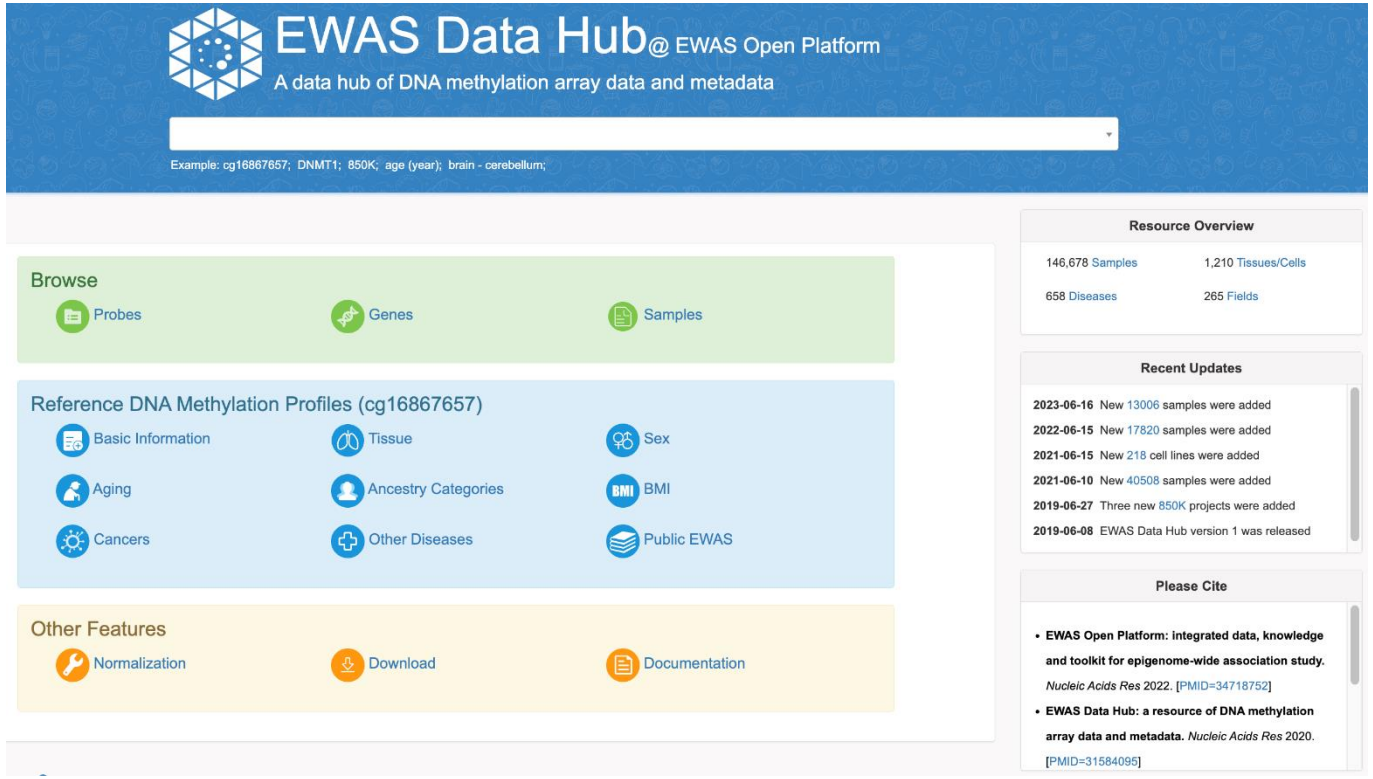
Trait	Type	#Publications	#Studies	#Associations	Correlations	CpG Island
1,6-hexamethylene diisocyanate exposure	environmental factor	1	2	32		
17q12 deletion	phenotype	1	1	21		
abdominal obesity	phenotype	1	1	14		
absolute fat free mass (FFM, kg)	phenotype	1	1	250		
absolute fat mass (FM, kg)	phenotype	1	1	250		
acute coronary syndrome	non-cancer disease	1	1	9		
acute mania	non-cancer disease	1	1	1		
acute myelocytic leukemia (AML)	cancer	1	2	28		
acute respiratory distress syndrome (ARDS)	non-cancer disease	1	1	4		
adiposity	phenotype	1	3	6		

798个性状；313676个甲基化探针；
36573个基因

https://ngdc.cncb.ac.cn/ewas/browse/

EWAS Data Hub

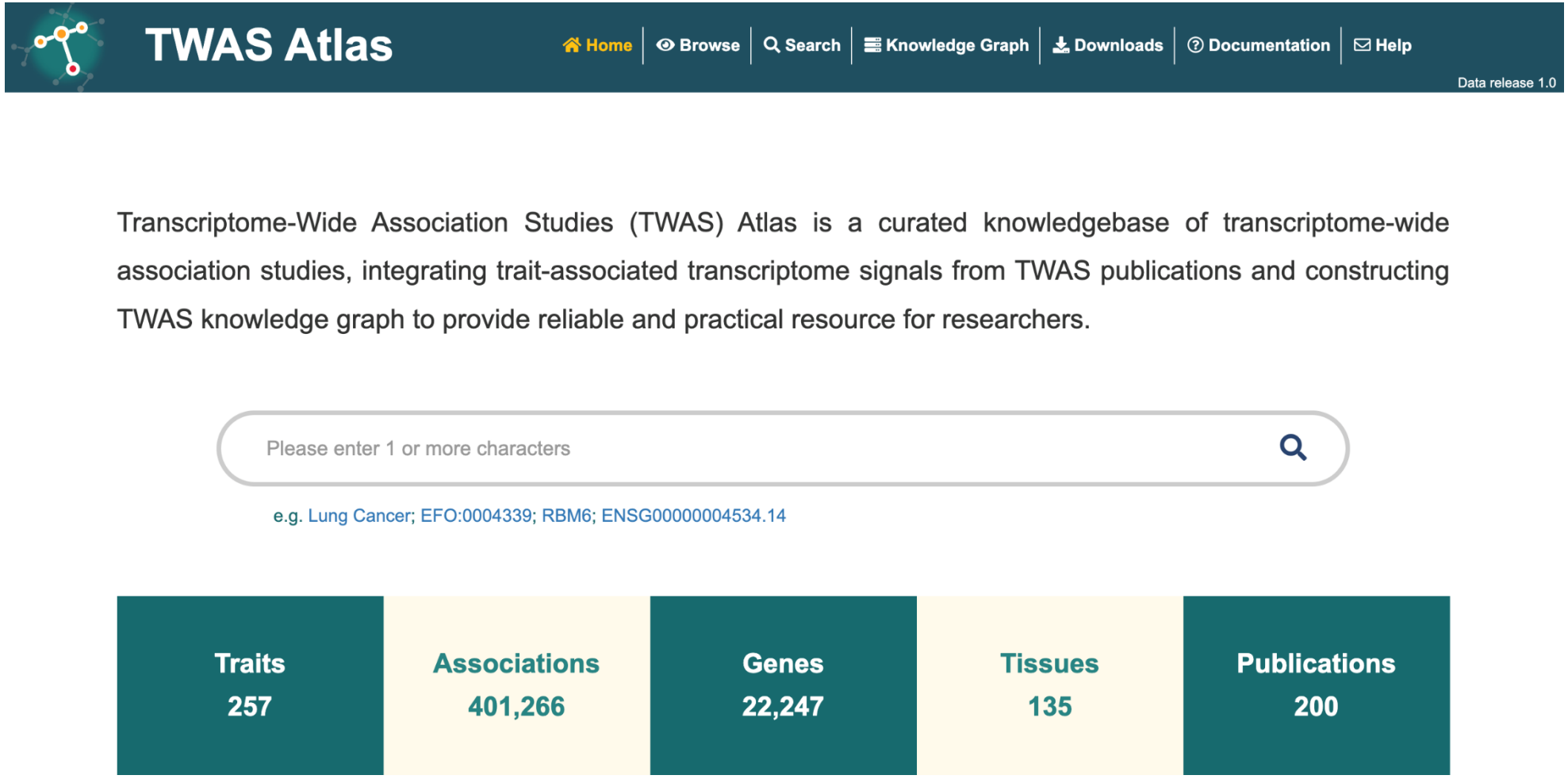




146,678个样本；1210种组织；658种疾病

https://ngdc.cncb.ac.cn/ewas/datahub

TWAS Atlas

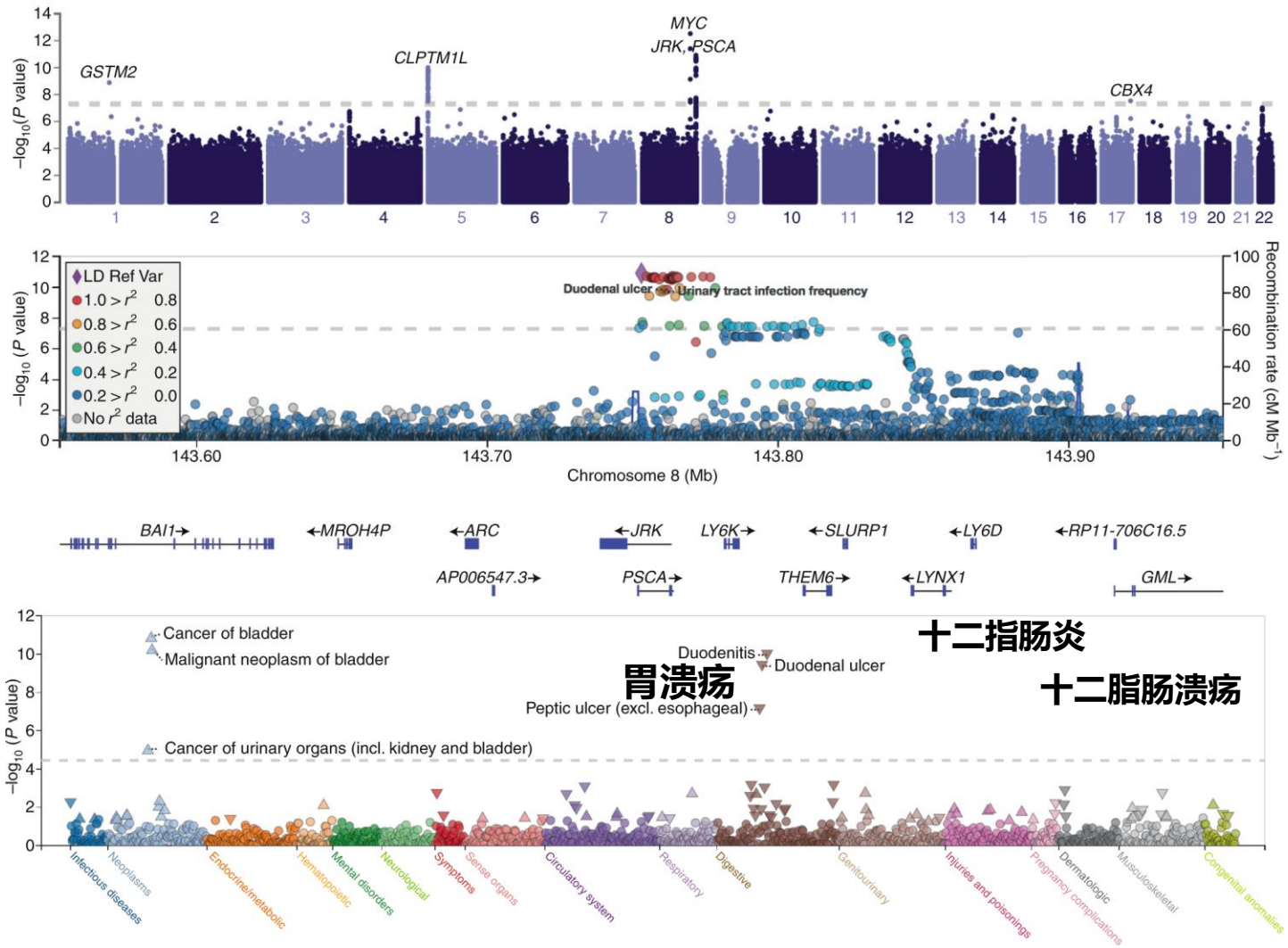


TWAS Atlas: a curated knowledgebase of transcriptome-wide association studies. *Nucleic Acids Res* 2023. https://ngdc.cncb.ac.cn/twas

其它相关数据库

- EWASdb：通过数据分析构建的表观知识库
- MRC-IEU EWAS Catalog：布里斯托大学开发的表观关联知识库
- ENCODE：综合的表观数据库，整合多组学数据
- NIH Roadmap Epigenomics：当前较为全面的表观基因组数据库
- IHEC Data Portal ： 不同组织的表观基因组参考数据集

PheWeb for PheWAS



Bladder cancer: 2,427 cases and 404,796 controls

Regional view for **rs2976384** (purple diamond)

Phenome-wide view for variant **rs2976384**

Triangles:

- upward-facing (positive effect)
- downward-facing (negative effect)

Exploring and visualizing large-scale genetic associations by using PheWeb. *Nature Genetics* 2020

PheWeb: exploring and visualizing genetic and phenotypic associations

278.1: Obesity

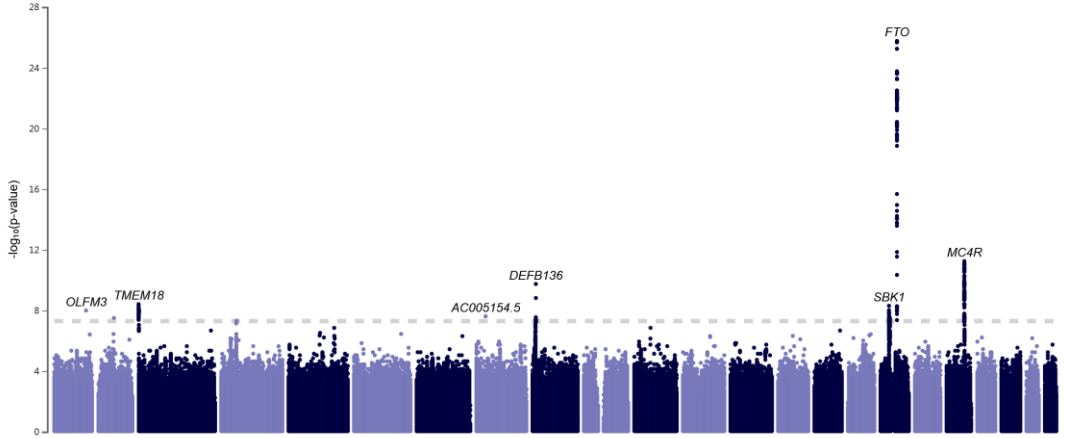
10799 cases, **397993** controls

Category: **endocrine/metabolic**

38 total variants

Variant	Nearest Gene(s)	MAF	P-value	Effect Size (se)
16:53,802,494 C / T (rs11642015)	*FTO*	0.40	1.7e-26	0.15 (0.014)
18:57,852,587 T / C (rs476828)	*MC4R*	0.24	5.6e-12	0.12 (0.017)
8:11,830,150 C / T (rs11998678)	*DEFB136*	0.46	1.8e-10	0.091 (0.014)
2:600,575 G / T (rs2683992)	*TMEM18*	0.18	3.8e-9	0.11 (0.019)
16:28,340,945 A / G (rs12598357)	*SBK1*	0.42	4.9e-9	0.084 (0.014)
1:102,731,430 G / A (rs547940655)	*OLFM3*	0.0018	1.0e-8	1.3 (0.22)
7:30,554,781 T / C (rs190685589)	*AC005154.5*	0.0079	2.4e-8	0.50 (0.089)
16:30,878,863 C / T (rs4889634)	*BCL7C*	0.38	2.9e-8	-0.081 (0.015)
1:191,536,872 T / C (rs143404628)	*RGS18*	0.0048	3.1e-8	0.61 (0.11)
3:50,207,075 T / C (rs2518796)	*SEMA3F*	0.43	4.7e-8	0.078 (0.014)

https://pheweb.org/UKB-SAIGE/pheno/278.1



Correlated phenotypes for 278.1: Obesity

Genetically correlated traits are calculated using genome-wide association summary statistics via LD Score regression (LDSC).

Trait	r_g	SE	Z	P-value
278: Overweight, obesity and other hyperalimentation	0.9992	0.0004	2645.1766	0
716.9: Arthropathy NOS	0.6846	0.0347	19.7451	8.8371e-87
716: Other arthropathies	0.6808	0.0347	19.6134	1.188e-85
401: Hypertension	0.5786	0.0315	18.3876	1.6526e-75
401.1: Essential hypertension	0.5791	0.0316	18.3207	5.661e-75
318: Tobacco use disorder	0.6564	0.0396	16.5596	1.3647e-61
250.2: Type 2 diabetes	0.668	0.0418	15.9972	1.3358e-57
250: Diabetes mellitus	0.6647	0.0423	15.6979	1.5623e-55
411.3: Angina pectoris	0.6222	0.0408	15.2309	2.2041e-52
411: Ischemic Heart Disease	0.5759	0.0379	15.1909	4.063e-52

关节病

心绞痛

Summary

- 基因型与表型
- GWAS技术原理：芯片 vs 测序，各自优缺点
- GWAS统计分析：
  - Odds Ratio
  - Chi-square test, p-value
  - 多重检测校正：FDR、Bonferroni、BH
- GWAS：Manhattan Plot、LD
- EWAS、TWAS、PWAS、PheWAS
- 关联知识库：GWAS Catalog、GWAS Atlas、TWAS Atlas、EWAS Atlas、PheWeb

Further Reading

- **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007
- **How to Interpret a Genome-wide Association Study**. *JAMA* 2008
- **Genome-wide association studies of 14 agronomic traits in rice landraces**. *Nature Genetics* 2010
- **Chapter 11: Genome-Wide Association Studies**. *PLoS Comput Biology* 2012
- **Natural Variations and Genome-Wide Association Studies in Crop Plants**. *Annual Review of Plant Biology* 2014
- **Benefits and Limitations of Genome-Wide Association Studies.** *Nature Review Genetics* 2019
- **Genome-Wide Association Studies.** *JAMA* 2019
- **Identification of type 2 diabetes loci in 433,540 East Asian individuals**. *Nature* 2020
- **Exploring and visualizing large-scale genetic associations by using PheWeb**. *Nature Genetics* 2020
- **Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis**. *Nature Genetics* 2021
- **Maturation and application of phenome-wide association studies**. *Trends in Genetics 2022*
- **Phenome-Wide Association Studies.** *JAMA* 2022