# Genomics: Introduction of GWAS and Complex Diseases II

PEILIN JIA

BEIJING INSTITUTE OF GENOMICS

# Outline

**Statistical inference**

Power in GWAS

Relatedness and population structure

Summary statistics

Meta-analysis

Heritability and the missing heritability

# GWAS Statistics $\hat{\beta}$ and SE

Assuming additive model, $\beta$ is the difference in mean phenotype between genotype classes 0 and 1, and it is also the difference between classes 1 and 2

- For QTs the difference is measured on phenotypic scale, often in units of standard deviation of the phenotype
- For disease traits, the difference is measured on the scale of logarithm of odds of disease
- We never know the ”true” $\beta$ but can only get an estimate $\hat{\beta}$ from the data with some uncertainty

Assuming reasonable sample sizes (say MAF>1% and N > 100), standard error (SE) of $\hat{\beta}$ describes the uncertainty of the estimate

- 95% confidence interval for $\hat{\beta}$ is achieved by putting ~2 SEs around the estimate
- Technically, SE is an estimate of the standard deviation of the sampling distribution of $\hat{\beta}$

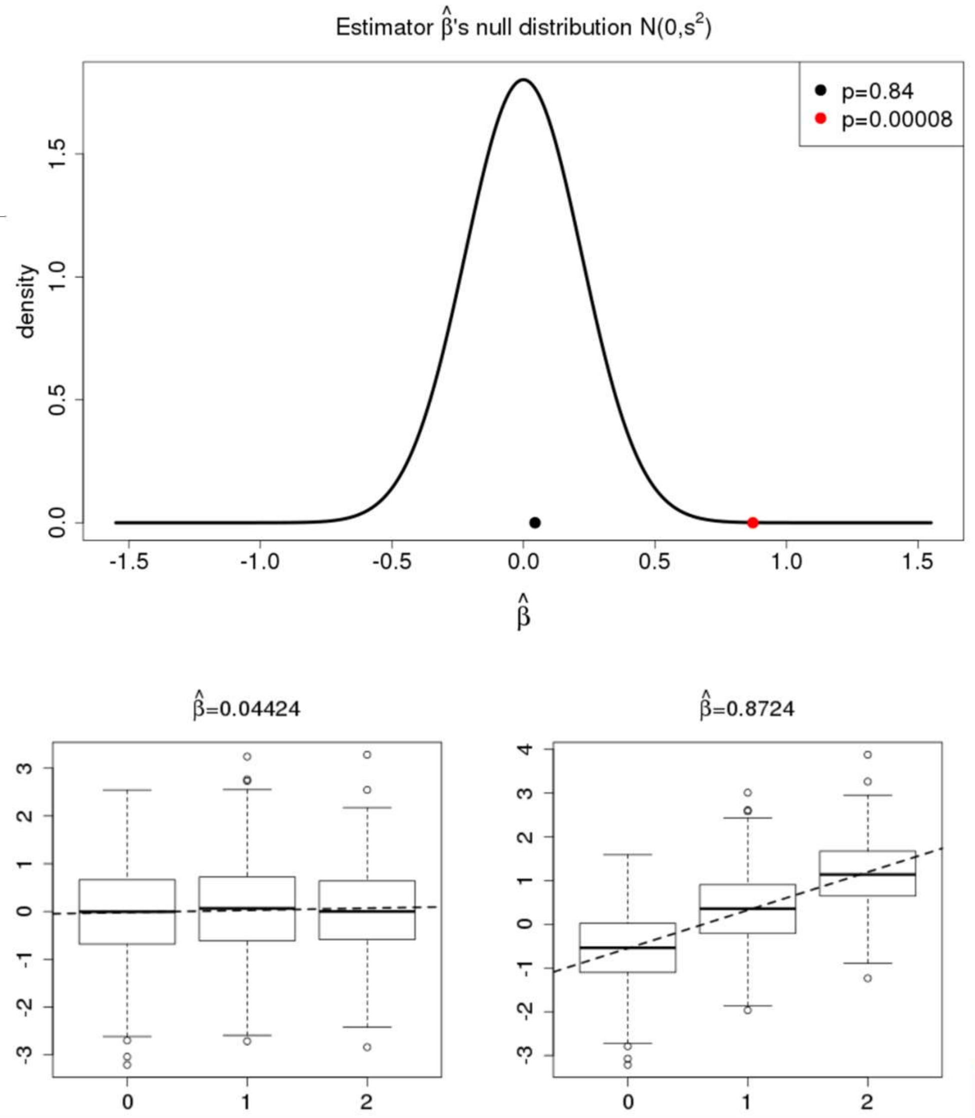
# P-Value

Is the observed slope plausible if true slope = 0 ?

P-value: Probability that “by chance” we get at least as extreme value as we have observed, if true slope = 0

P = 0.84: No evidence for deviation from null

P = $8\times10^{-5}$: Unlikely under the null → maybe not null



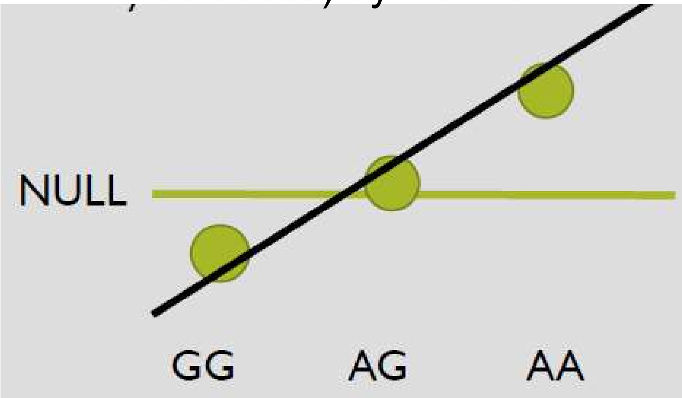
# Why Two-sided *P*-values

What is “at least as extreme data set as what we have observed”?

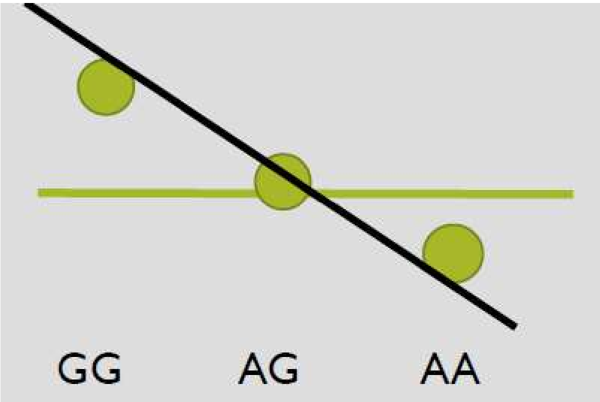
Depends on our Null hypothesis

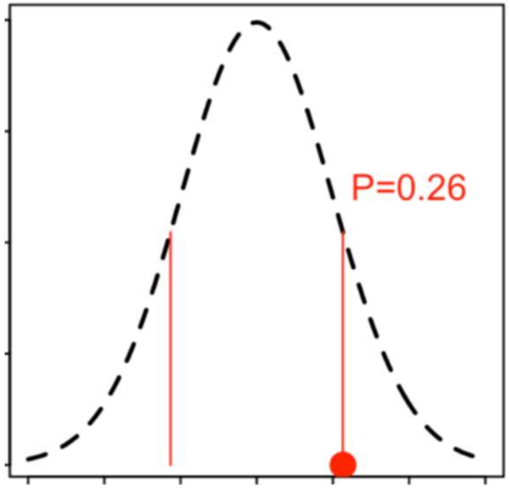
Typically, NULL is that slope $\beta = 0$, and then allele A increasing (and G decreasing) phenotype by 2 units is equally ”extreme” as A decreasing (and G increasing) by 2 units

A increases (and G decreases) by 2 units



A decreases (and G increases) by 2 units





2-sided P-value = sum of the two tail probabilities

# Multiple Testing Correction

Perform 1,000,000 (or “multiple testing”) analyses

- Multiple testing correction
- Bonferroni correction
- False Discovery Rate (FDR)
- Permutation

Significance threshold

- Significance threshold $\alpha$ = Probability that a null variant will reach P-value below $\alpha$
- $\alpha$ is called the statistical power of the significance test

Genome-wide significance

- According to Bonferroni correction: $0.05/1000000 = 5\times10^{-8}$

# Power of GWAS

H0 (NULL HYPOTHESIS): Variant has no effect on phenotype

H1 (ALTERNATIVE HYPOTHESIS): Variant has a non-zero effect on the phenotype

Significance level $\alpha$: “Reject H0” and “accept H1” if P-value (calculated assuming H0) is $< \alpha$

- If $\alpha$ is defined before the experiment, then the proportion of false rejection of H0 would be $\alpha$ in repeated experiments
- By making $\alpha$ small (say 5e-8) we can protect from false positive findings (Type I errors) but increase false negative findings (Type II errors)
- By keeping $\alpha$ larger (say 0.05) we have more statistical power to reject H0 but we are more likely to make a false positive finding (Type I error)

# Power of GWAS

Factors that influence the power of a GWAS

Sample size: N increases power

MAF: MAF increases power

Effect size: larger effect increases power

Case frequency and control frequency: $N\ \phi(1-\phi)$ is the **effective sample size (where** $\phi$ is the proportion of cases in the total samples).

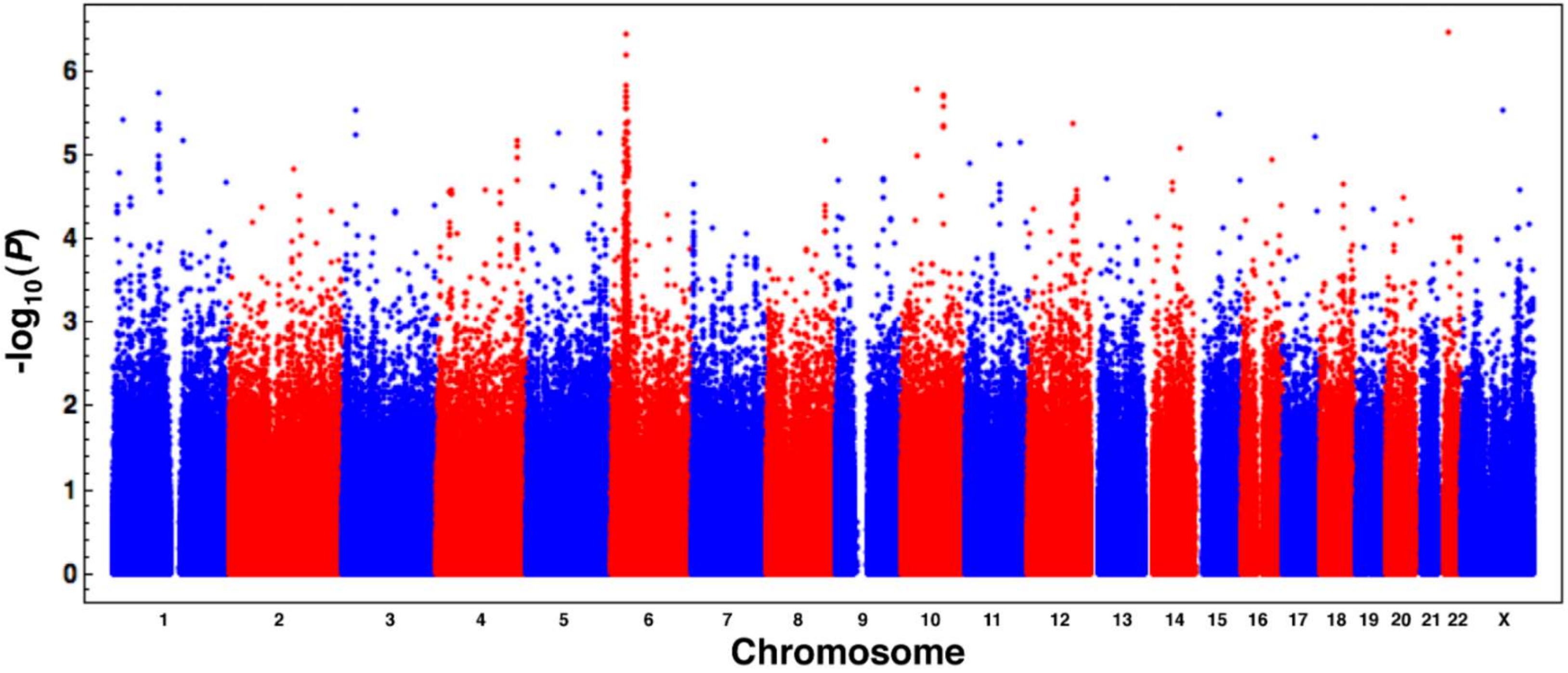
- To increase power, we should have large N and $\phi$ close to 0.5

# Example: Schizophrenia GWAS (2009)

3,332 SZ cases and 3,587 controls at 1M SNPs

No genome-wide significant findings

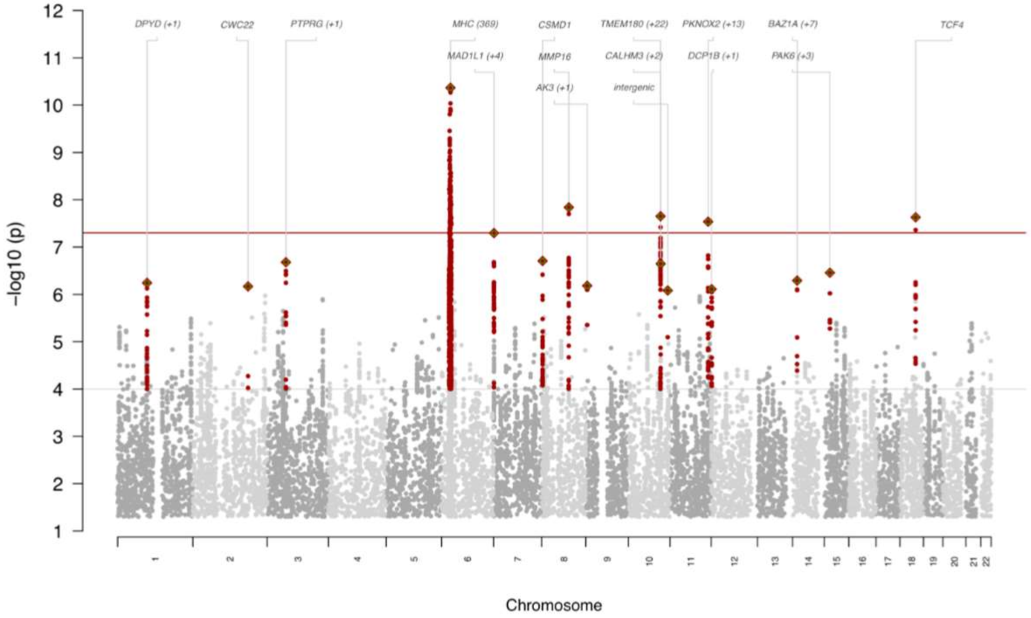
Suggestive evidence for HLA-region on chr 6



# Example: Schizophrenia GWAS (2011)

9,394 SZ cases and 12,462 controls at 1M SNPs

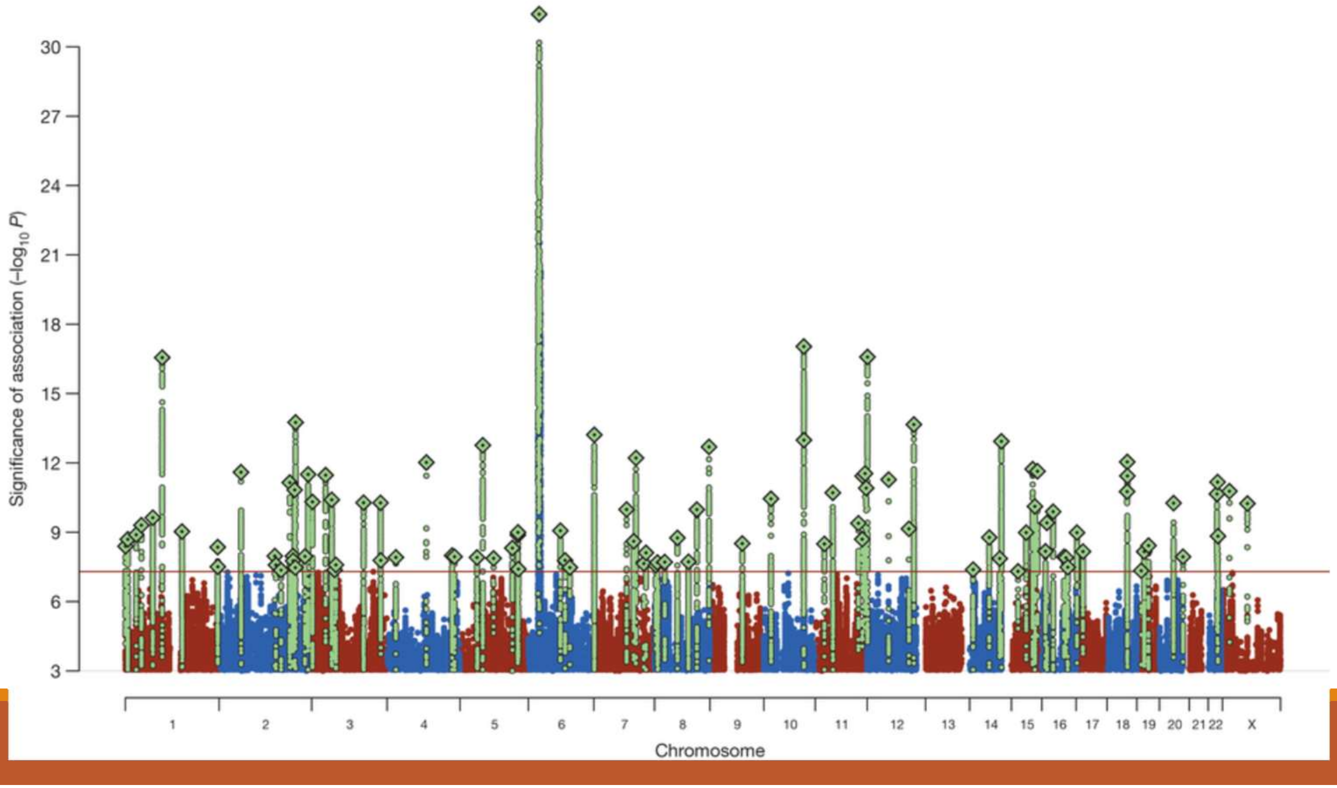
5 GWS loci



# Example: Schizophrenia GWAS (2014)

34,000 SZ cases and 45,600 controls at 9.5M SNPs

108 loci



# Relatedness and Popul

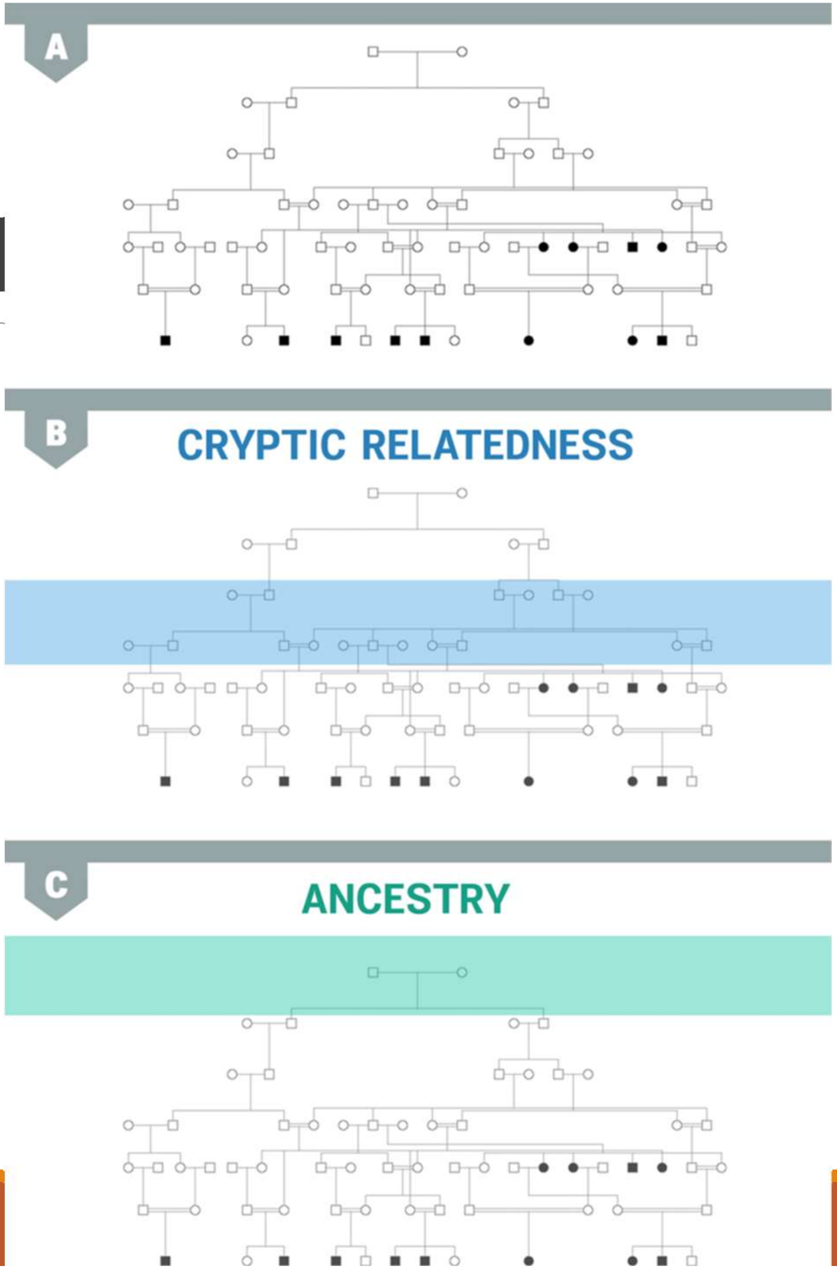
LEVELS OF RELATEDNESS

(A) All of the individuals in a genetic study are related through a large pedigree or family tree. Different parts of the tree induce different types of relatedness.

(B) **Cryptic relatedness** refers to relatively recent genetic relationships which are not otherwise reported in the data except by genetic analysis.

(C) Relatedness due to ancestry refers to relatedness caused by ancestors originating from the same region, i.e., **population structure**.

The boxes in (B) and (C) represent the level of the pedigree that causes that type of relatedness in each case.



Sul et al. 2018 PLoS Gen https://doi.org/10.1371/journal.pgen.1007309

# Relatedness and Population Structure

Intuition:

Close relatives share more genetic ancestors in more recent past than distant relatives.

If we average such information across the genome we have a relatedness estimate for genomes, and if we average over two genomes of individuals we have a relatedness estimate for individuals.

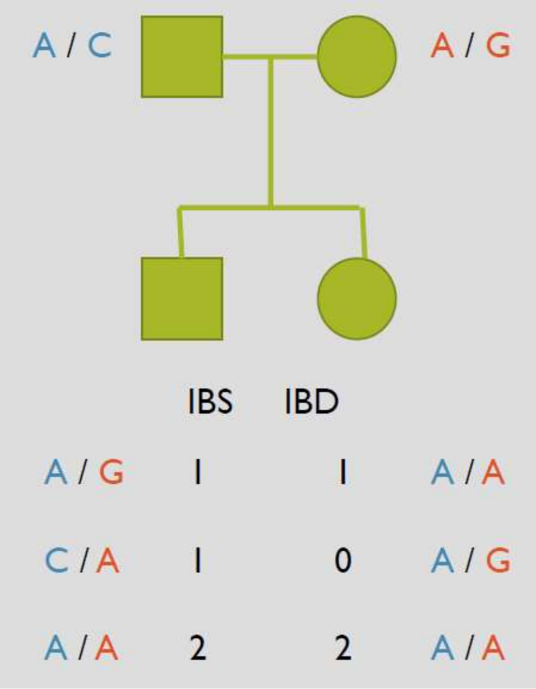
# IBS VS IBD

IBS (identical-by-state) means that DNA sequence is identical

IBD (identical-by-descent) means that DNA sequence is inherited from a common ancestor (within a given timeframe)

- For pedigrees the timeframe is founder generation
- For population data, we don’t have an exact timeframe and then IBD is measured by how much more the pair shares IBS than would be expected from a random pair from the population
- Most accurate IBD estimation methods look for sharing of longer segments not just individual loci, but here we consider sharing at independent loci

IBD implies IBS but not vice versa: there can be IBS sharing without there being a common ancestor **within a given timeframe** (such as known pedigree structure)



| | IBS | IBD | |
|---|---|---|---|
| A / G | 1 | 1 | A / A |
| C / A | 1 | 0 | A / G |
| A / A | 2 | 2 | A / A |

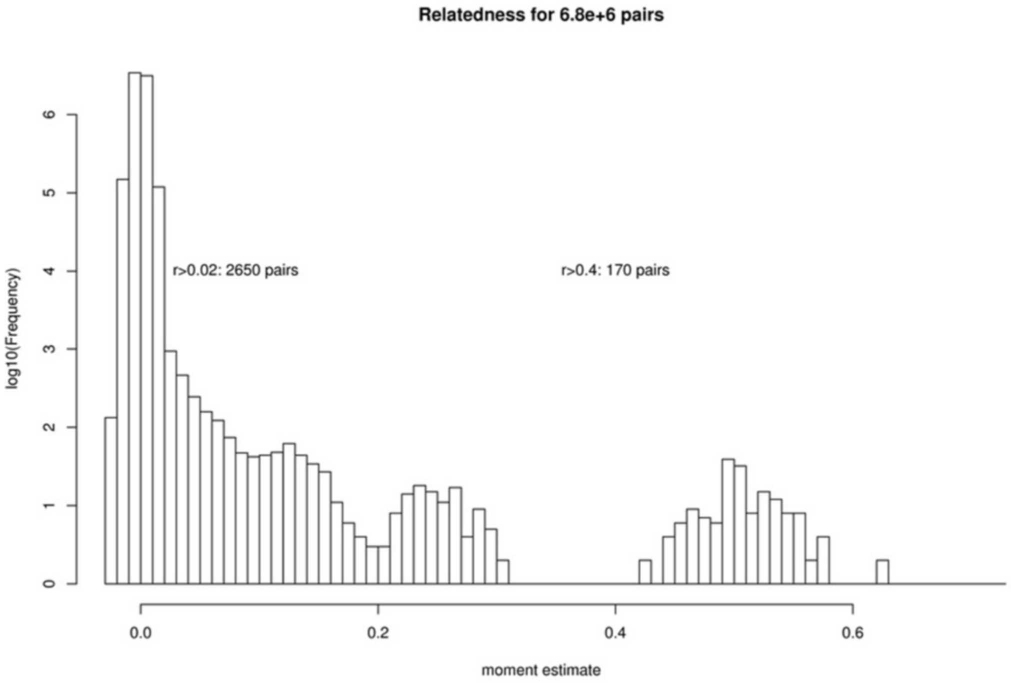
# Relatedness as a Quality Control Step

High relatedness to many individuals can imply contamination

When DNA from many samples are mixed the result has excessive heterozygosity

Correlation estimator was used as an estimate of rij

- 50% (e.g. Full sibs)
- 25% (e.g. Half sibs)
- 12.5% (e.g. First cousins)



# Population Structure

Populations = “groups of individuals who **on average** are more closely related to the other members of the same group than to the members of another group”

Asian population vs European population

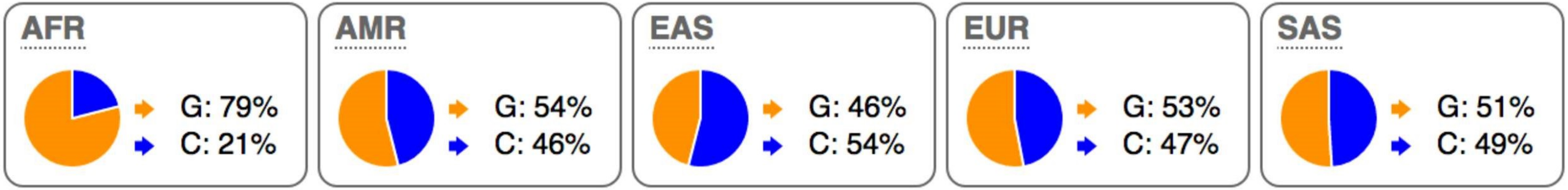
Finnish population vs Swedish population

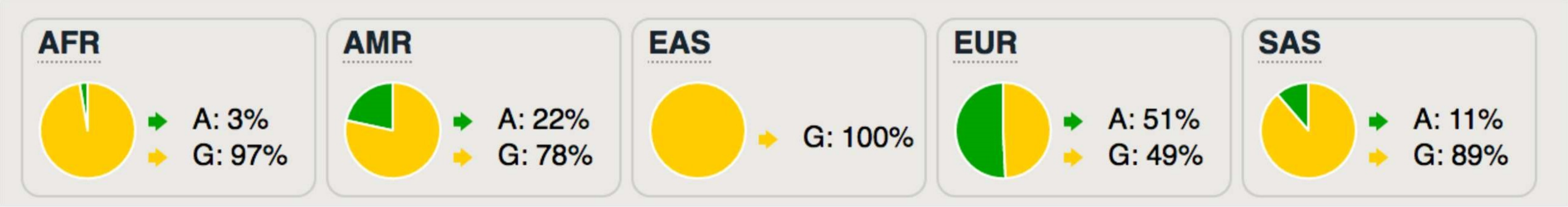
Eastern Finns vs Western Finns

Population membership is not an objectively defined characteristic but depends on the level of detail.

# Global Allele Frequency Differences

Frequencies can change due to genetic drift, selection, admixture or all of them.





rs4988235 (Lactase tolerance in Europeans, selection has had effect)

Three levels of structure as revealed by PC analysis:

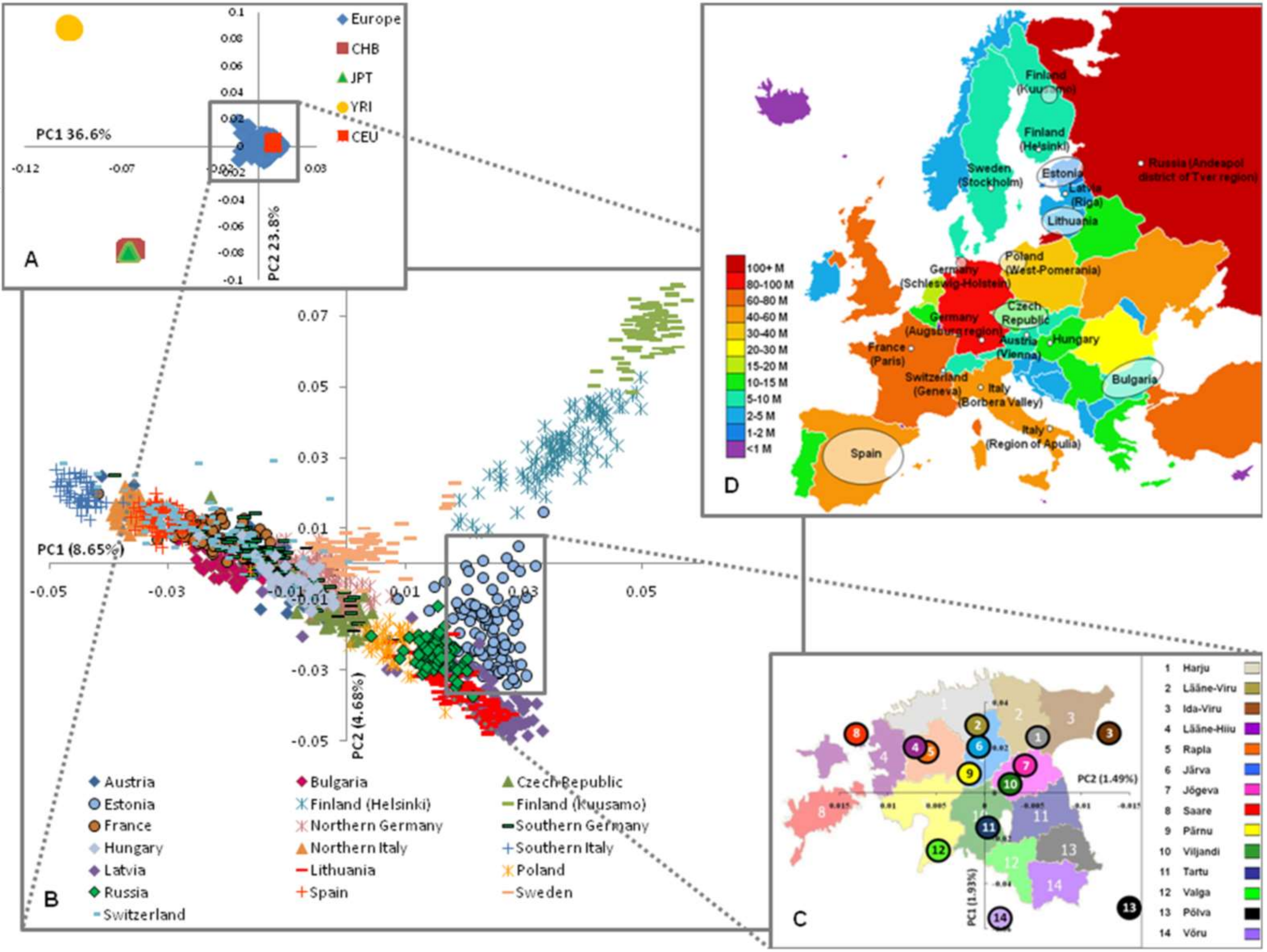
- A) inter-continental;
- B) intra-continental; and
- C) inside Estonia, where median values of the PC1&2 are shown.
- D) European map illustrating the origin of sample and population size.

CEU - Utah residents with ancestry from Northern and Western Europe,

CHB – Han Chinese from Beijing,

JPT - Japanese from Tokyo

YRI -Yoruba from Ibadan, Nigeria.



# PCA as a Quality Control Step

PCA can be used to inform if data have samples with

- Relative groups
- Different ancestry
- Technical problems (e.g. contamination)

These are typically removed from further analyses

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x + COV1 + COV2 \ldots + \varepsilon$$

$$y = \beta_0 + \beta_1 SNP + COV1 + COV2 \ldots + \varepsilon$$

# Outline

Statistical inference

Power in GWAS

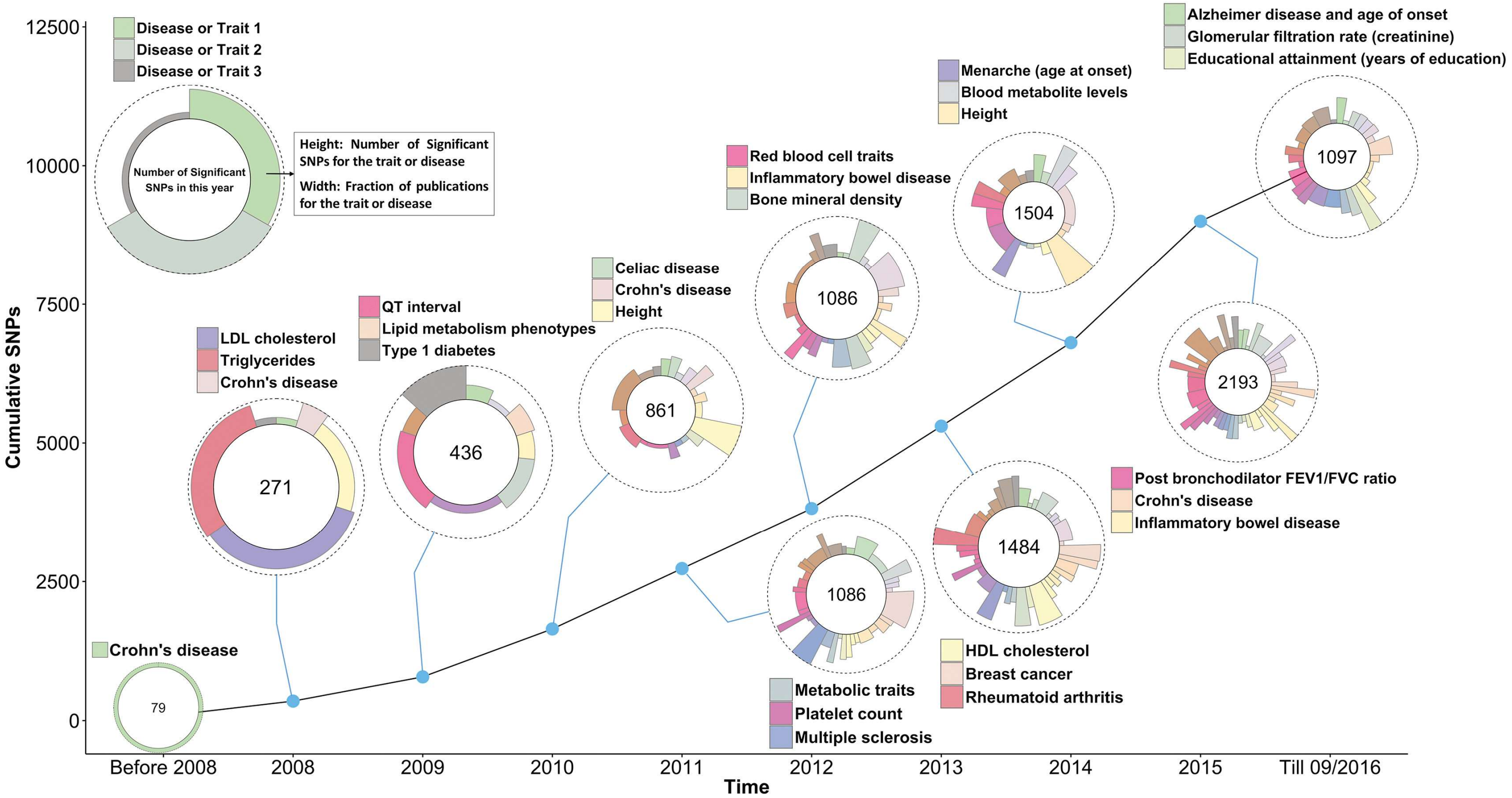
Relatedness and population structure

**Summary statistics**

Meta-analysis

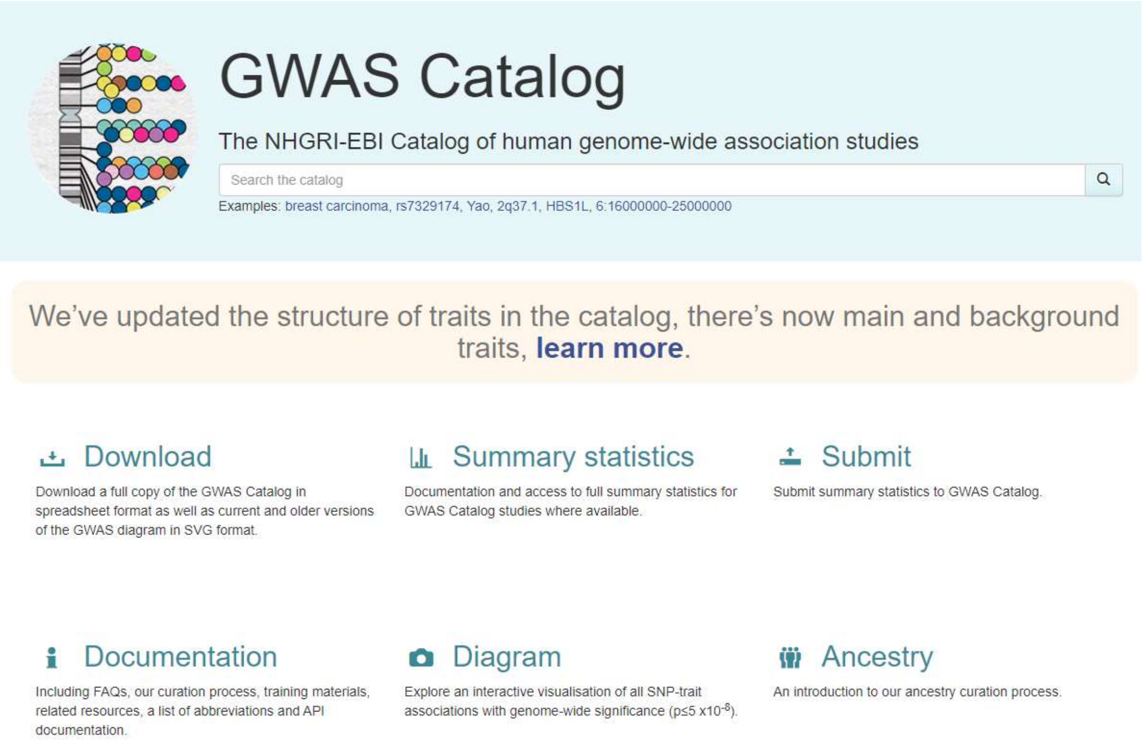
Heritability and the missing heritability

# More than 10 Years of GWAS



# GWAS Catalog

The GWAS Catalog was established with the aim of providing a central repository for variant-trait associations identified through GWASs, serving as a starting point for investigations to identify causal variants, understand disease mechanisms, and establish targets for new therapies



GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

We've updated the structure of traits in the catalog, there's now main and background traits, **learn more**.

Download

Download a full copy of the GWAS Catalog in spreadsheet format as well as current and older versions of the GWAS diagram in SVG format.

Summary statistics

Documentation and access to full summary statistics for GWAS Catalog studies where available.

Submit

Submit summary statistics to GWAS Catalog.

Documentation

Including FAQs, our curation process, training materials, related resources, a list of abbreviations and API documentation.

Diagram

Explore an interactive visualisation of all SNP-trait associations with genome-wide significance ($p \leq 5 \times 10^{-8}$).

Ancestry

An introduction to our ancestry curation process.

# GWAS Data

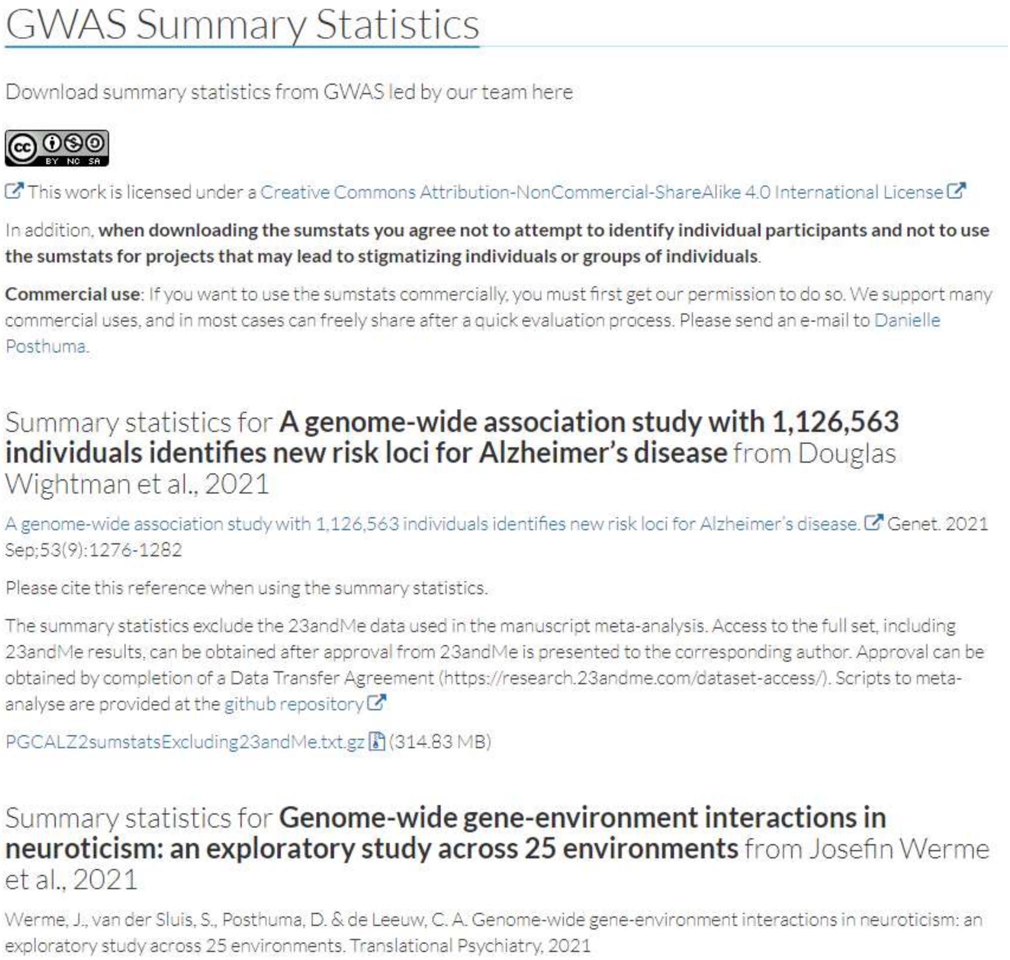
GWAS data are deposited in NCBI dbGAP

Individual genotype data are under controlled access.

GWAS summary statistics (SumStats) are defined as the aggregate p values and association data for every variant analyzed

Hundreds of GWAS summary statistics data sets are made available in the public domain.

UK BioBank summary statistics are available



## GWAS Summary Statistics

Download summary statistics from GWAS led by our team here

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

In addition, **when downloading the sumstats you agree not to attempt to identify individual participants and not to use the sumstats for projects that may lead to stigmatizing individuals or groups of individuals**.

**Commercial use**: If you want to use the sumstats commercially, you must first get our permission to do so. We support many commercial uses, and in most cases can freely share after a quick evaluation process. Please send an e-mail to Danielle Posthuma.

### Summary statistics for **A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease** from Douglas Wightman et al., 2021

A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Genet. 2021 Sep;53(9):1276-1282

Please cite this reference when using the summary statistics.

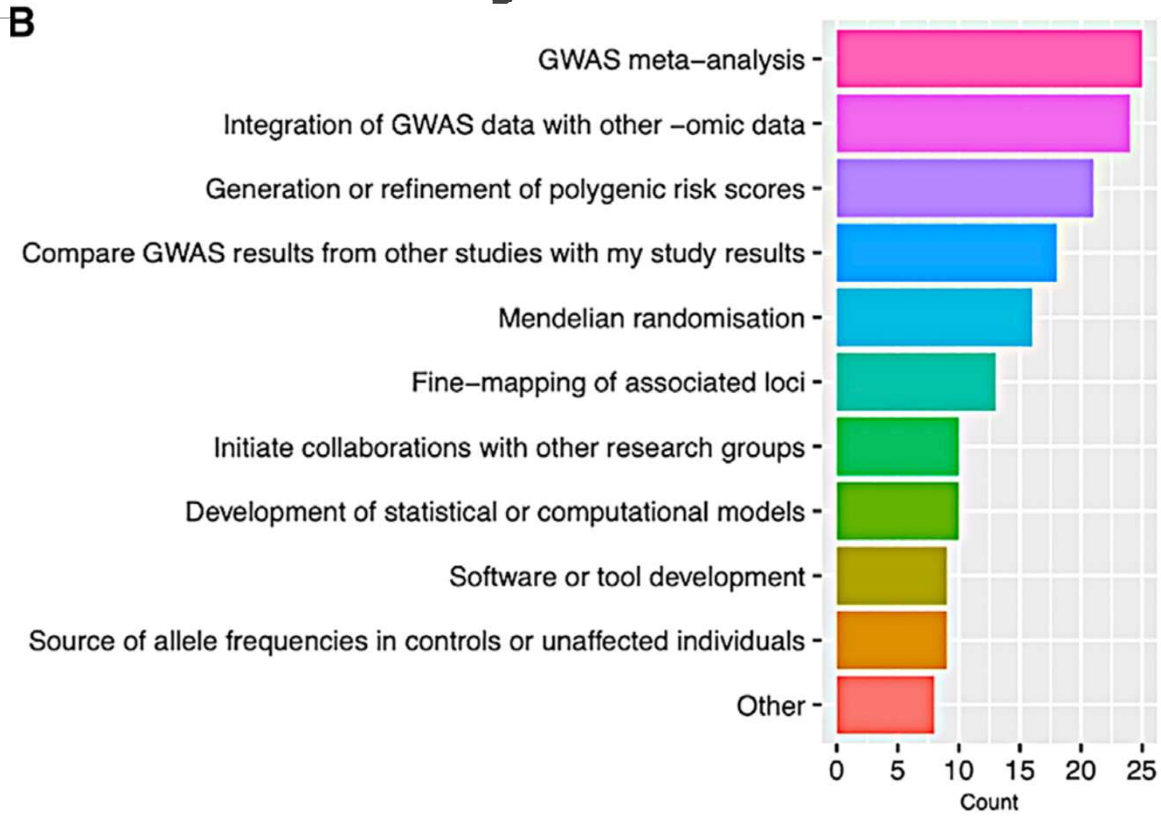
The summary statistics exclude the 23andMe data used in the manuscript meta-analysis. Access to the full set, including 23andMe results, can be obtained after approval from 23andMe is presented to the corresponding author. Approval can be obtained by completion of a Data Transfer Agreement (https://research.23andme.com/dataset-access/). Scripts to meta-analyse are provided at the github repository

PGCALZ2sumstatsExcluding23andMe.txt.gz (314.83 MB)

### Summary statistics for **Genome-wide gene-environment interactions in neuroticism: an exploratory study across 25 environments** from Josefin Werme et al., 2021

Werme, J., van der Sluis, S., Posthuma, D. & de Leeuw, C. A. Genome-wide gene-environment interactions in neuroticism: an exploratory study across 25 environments. Translational Psychiatry, 2021

# GWAS Summary Statistics



# Outline

Statistical inference

Power in GWAS

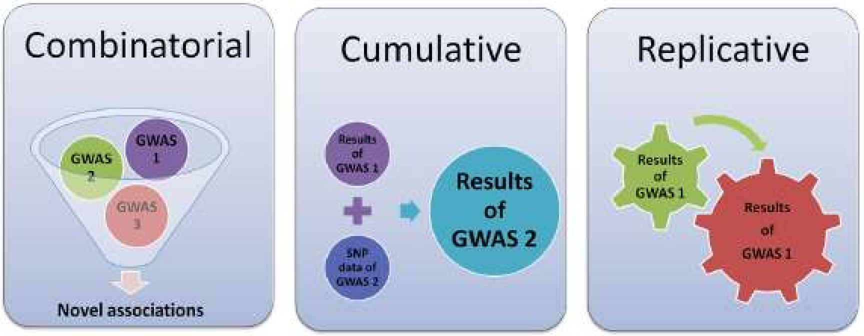
Relatedness and population structure

Summary statistics

**Meta-analysis**

Heritability and the missing heritability

# Meta Analysis



Problem of individual GWAS

- Relatively modest sample and effect sizes
- Limited power to find novel loci underlying complex traits and common diseases

Meta-analysis

- Using a statistical approach to combine the results from multiple studies
- Increasing power and reducing false-positive findings
- Improving the estimation of effect size
- Solving the new problem which is unclear in individual study
- Needing to assess the heterogeneity of association result between all data set

Evangelou E et al 2013

# A Typical Plan for a Meta-analysis

Dealing with heterogeneity

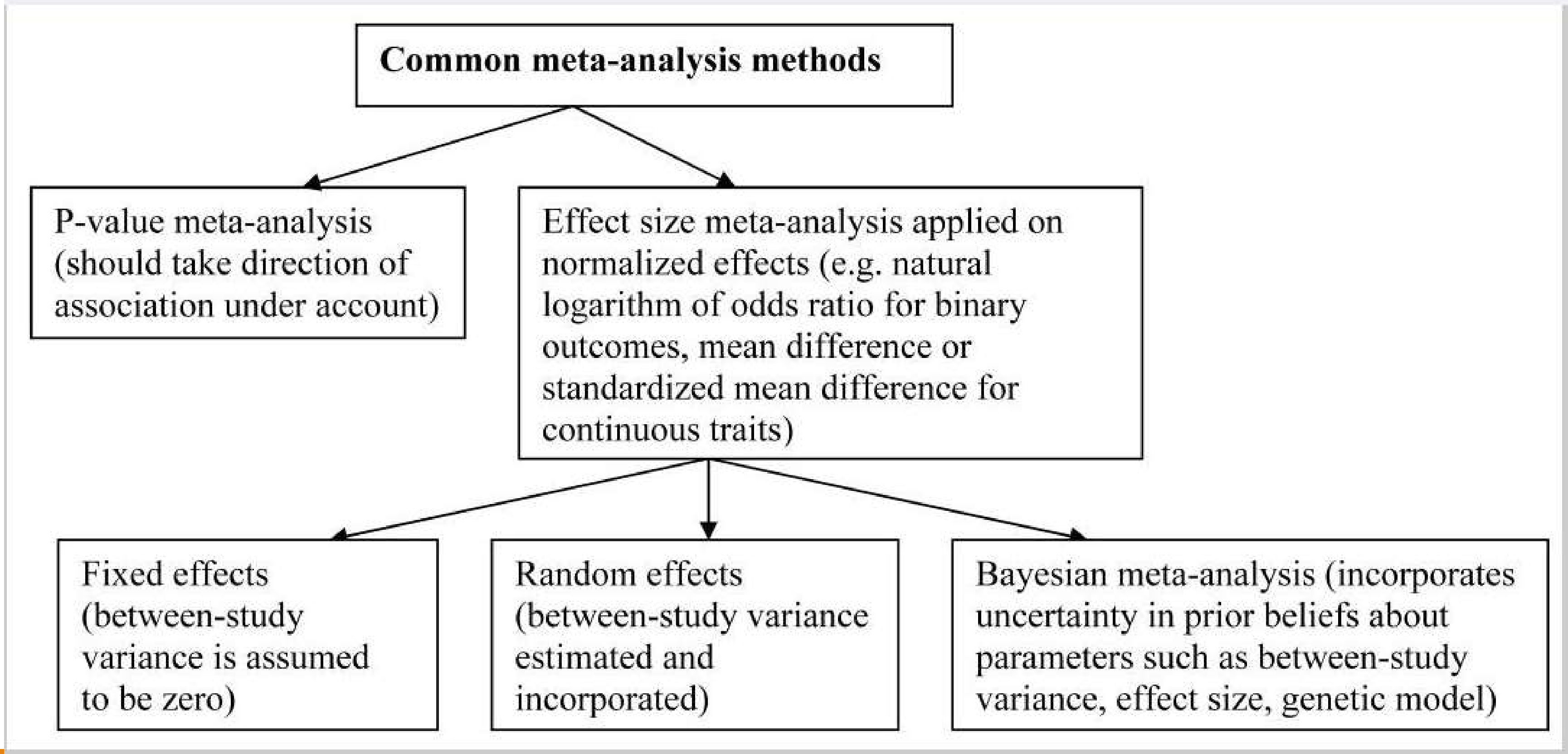
- Standardized phenotype definitions applied to all data sets
- Genotyping platforms
- Imputation metrics
  - Imputation in genetics refers to the statistical inference of unobserved genotypes. The goal is to predict the genotypes at the SNPs that are not directly genotyped in the study sample. These 'in silico' genotypes can then be used to boost the number of SNPs that can be tested for association.
  - Genotype imputation can be carried out across the whole genome or in a more focused region as part of a fine-mapping study.
- Software, reference panel and quality index

Quality control

- Stringent Inclusion and exclusion criteria of subjects and variants in all data sets. Such as popular exclusion thresholds of >5%, Hardy–Weinberg threshold of $p < 10^{-5}$, and the same ancestral groups

Evangelou E et al 2013

# Meta-analysis



Common meta-analysis methods

P-value meta-analysis (should take direction of association under account)

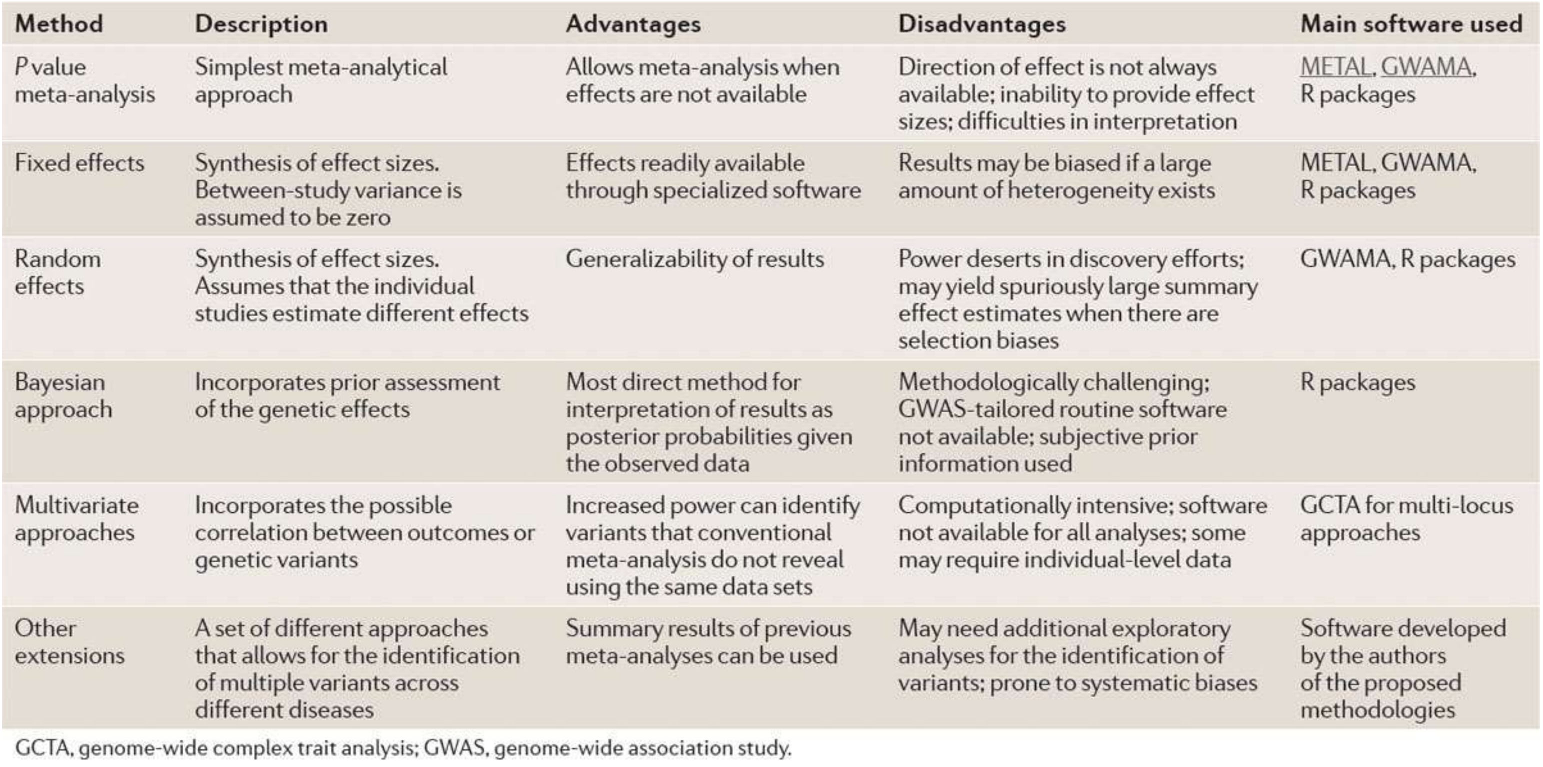
Effect size meta-analysis applied on normalized effects (e.g. natural logarithm of odds ratio for binary outcomes, mean difference or standardized mean difference for continuous traits)

Fixed effects (between-study variance is assumed to be zero)

Random effects (between-study variance estimated and incorporated)

Bayesian meta-analysis (incorporates uncertainty in prior beliefs about parameters such as between-study variance, effect size, genetic model)

# Summary of Methods for Meta-analysis



| Method | Description | Advantages | Disadvantages | Main software used |
|---|---|---|---|---|
| *P* value meta-analysis | Simplest meta-analytical approach | Allows meta-analysis when effects are not available | Direction of effect is not always available; inability to provide effect sizes; difficulties in interpretation | METAL, GWAMA, R packages |
| Fixed effects | Synthesis of effect sizes. Between-study variance is assumed to be zero | Effects readily available through specialized software | Results may be biased if a large amount of heterogeneity exists | METAL, GWAMA, R packages |
| Random effects | Synthesis of effect sizes. Assumes that the individual studies estimate different effects | Generalizability of results | Power deserts in discovery efforts; may yield spuriously large summary effect estimates when there are selection biases | GWAMA, R packages |
| Bayesian approach | Incorporates prior assessment of the genetic effects | Most direct method for interpretation of results as posterior probabilities given the observed data | Methodologically challenging; GWAS-tailored routine software not available; subjective prior information used | R packages |
| Multivariate approaches | Incorporates the possible correlation between outcomes or genetic variants | Increased power can identify variants that conventional meta-analysis do not reveal using the same data sets | Computationally intensive; software not available for all analyses; some may require individual-level data | GCTA for multi-locus approaches |
| Other extensions | A set of different approaches that allows for the identification of multiple variants across different diseases | Summary results of previous meta-analyses can be used | May need additional exploratory analyses for the identification of variants; prone to systematic biases | Software developed by the authors of the proposed methodologies |

GCTA, genome-wide complex trait analysis; GWAS, genome-wide association study.

Evangelou E et al 2013

# P-value Based Meta-analysis

Fisher's method

$X^2 = -2\sum_{i=1}^{k} \log(p_i)$

where $p_i$ is the p-value for the $i^{\text{th}}$ study and *k* is the number of studies.

$X^2$ follows a chi-square distribution with 2*k* degree of freedom

# Common Methods for Meta-analysis

Fixed effects

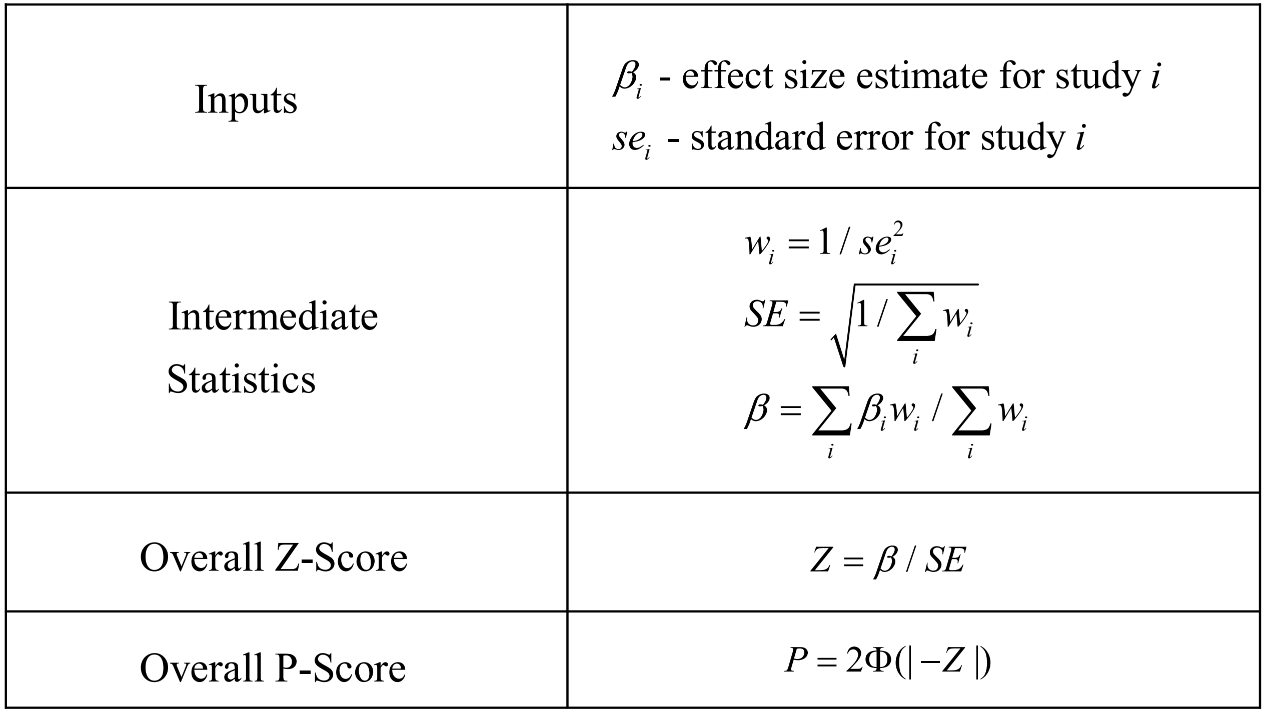
- The most popular approach for synthesizing GWAS data
- The most powerful approach for prioritizing and discovering phenotype-associated SNPs
- Assuming that the true effect of each risk allele is the same in each data set
- Compared to random effects models, fixed effects models have the major advantage of maximizing discovery power

Common models for fixed effects meta-analysis

- Inverse variance based approach (predominantly used)
- Each study is weighted according to the inverse of its squared standard error.
- Cochran–Mantel–Haenszel approach
- It provides almost identical results to the inverse variance weighting method.

# Common Methods for Meta-analysis

Formula of inverse variance based meta-analysis



| | |
|---|---|
| Inputs | $\beta_i$ - effect size estimate for study $i$<br>$se_i$ - standard error for study $i$ |
| Intermediate Statistics | $w_i = 1 / se_i^2$<br>$SE = \sqrt{1 / \sum_i w_i}$<br>$\beta = \sum_i \beta_i w_i / \sum_i w_i$ |
| Overall Z-Score | $Z = \beta / SE$ |
| Overall P-Score | $P = 2\Phi(\lvert -Z \rvert)$ |

# Common Methods for Meta-analysis

Random effects

- Incorporating variance between study as the random effects
- Far more limited power than fixed effects models
- More appropriate than fixed effects models when considering the generalizability of the observed association and estimating the average effect size of the associated variant and its uncertainty across different populations: for example, for predictive purposes
- Formula of random effects

$$\beta = \sum_i w_i \beta_i$$

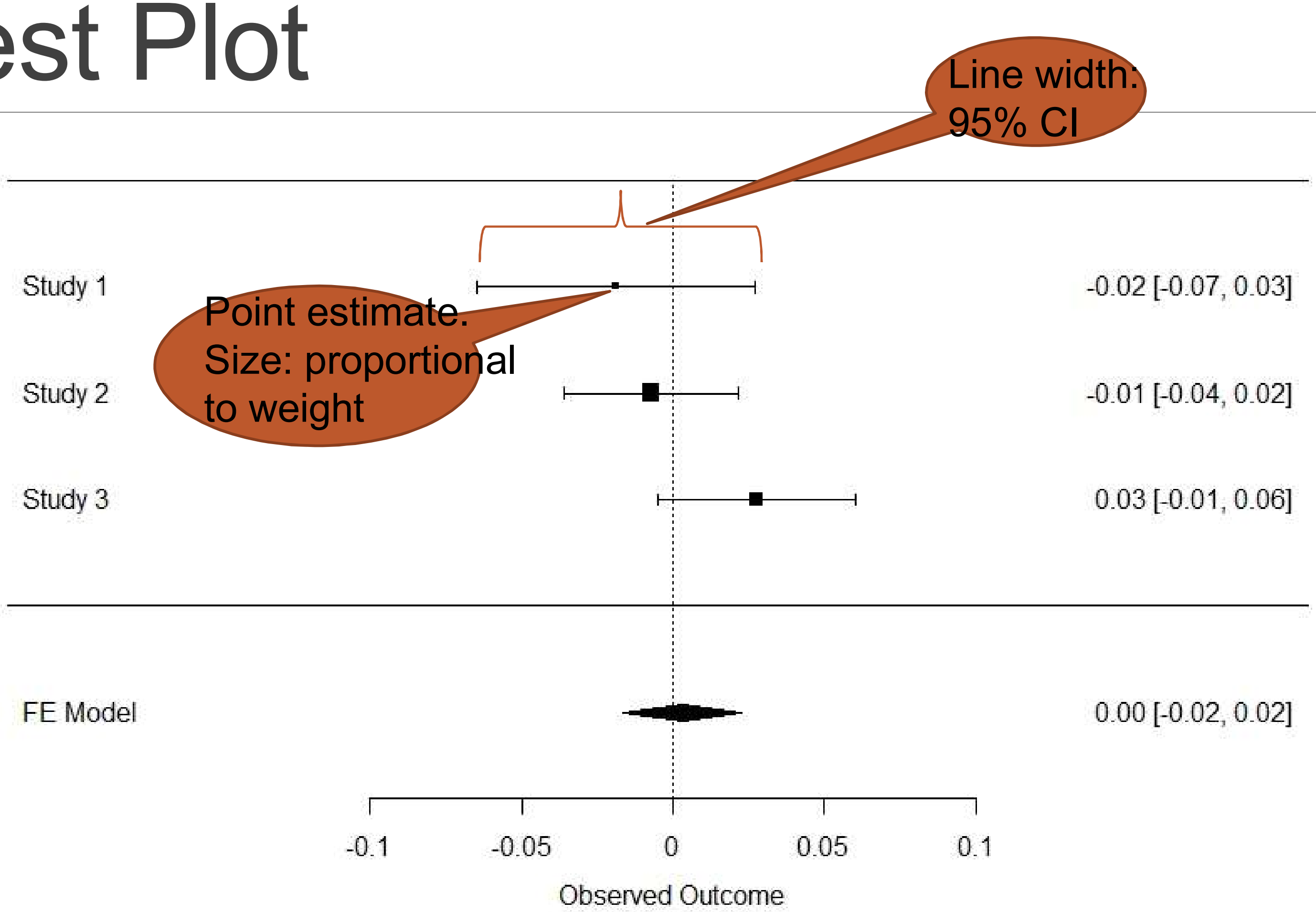
$\beta_i$ : estimation of effect size from study $i$

$w_i$ : weight assigned to study $i$

Fixed effects, $w_i \propto 1/v_i$, $v_i$ is the within-study variance of study $i$

Random effects, $w_i \propto 1/u_i$, $u_i = v_i + \tau^2$, $\tau^2$ is the between-study variance.

# Forest Plot


Line width:
95% CI
Point estimate.
Size: proportional
to weight
Study 1
Study 2
Study 3
FE Model
-0.02 [-0.07, 0.03]
-0.01 [-0.04, 0.02]
0.03 [-0.01, 0.06]
0.00 [-0.02, 0.02]
-0.1
-0.05
0
0.05
0.1
Observed Outcome

# Outline

Statistical inference

Power in GWAS

Relatedness and population structure

Summary statistics

Meta-analysis

**Heritability and the missing heritability**

# Heritability

A statistic used in the fields of breeding and genetics .

Heritability is defined as the proportion of phenotypic variation ($V_P$) that is due to variation in genetic values ($V_G$).

A measure of the contribution of genetics to phenotype.

Two traditions now dominate the study of heredity: population genetics and molecular biology.

- The notion of a quantitative measure of the heritability of any given trait comes from population genetics and heritability measures are commonly used in behavioral genetics.
- The idea that what is inherited is a stock of DNA, or the information contained in the DNA sequence, comes from molecular biology.

# Estimating heritability $h^2$ using H-E regression: close relatives have similar values of height

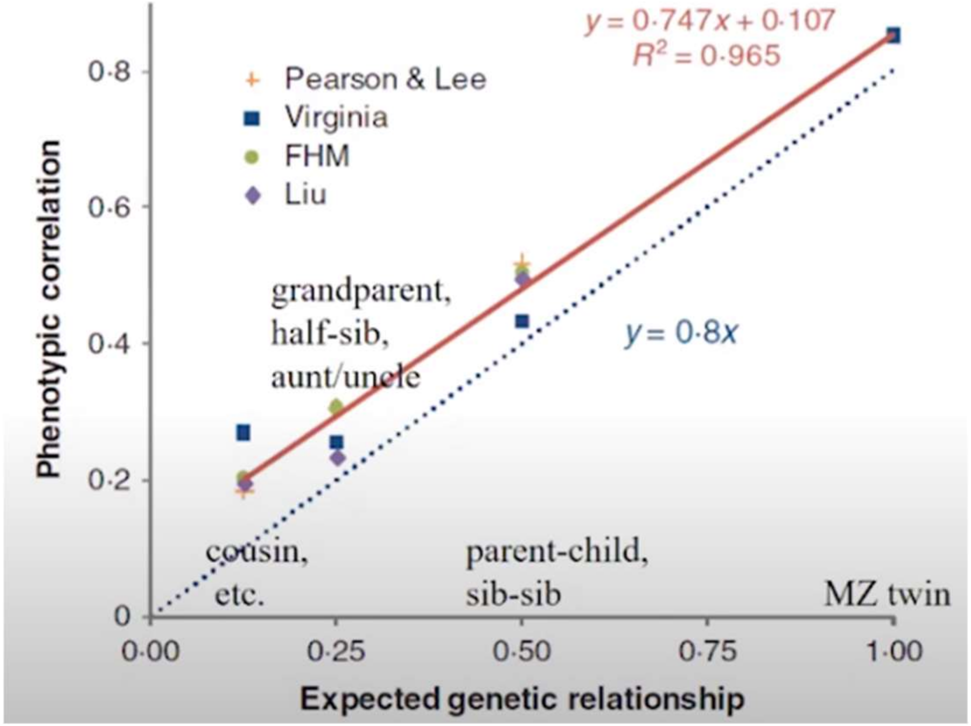
Each point represents one study.

Generally people who are related to each other tend to have very correlated values of height. (The height values have been adjusted for sex.)

People who are a little bit similar to each other tend to have height that's a little bit correlated.

The slope of the line is estimated as 0.77.

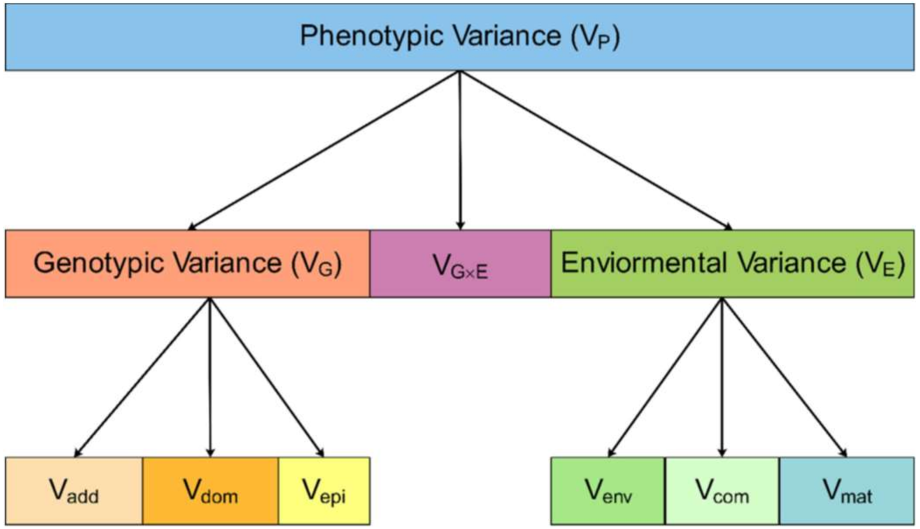
Intercept not equal 0: environmental effect?



# Defining Heritability

The phenotypic variance ($V_p$) in a population is influenced by genetic variance ($V_G$) and environmental sources ($V_E$)

- $V_p = V_G + V_E$


Phenotypic Variance ($V_P$)
Genotypic Variance ($V_G$)
$V_{G \times E}$
Enviormental Variance ($V_E$)
$V_{add}$
$V_{dom}$
$V_{epi}$
$V_{env}$
$V_{com}$
$V_{mat}$

# Broad Sense Heritability H2

- **Broad sense heritability H2** : the ratio of total genetic variance to phenotypic variance, including additive effects, dominant effects, and epistasis effects.

$$\boxed{H^2 = \frac{V_G}{V_P} = \frac{V_A + V_D + V_I}{V_P}}$$

- $H^2$ varies from 0 (all environment) to 1 (all genetic)
- Estimating $H^2$ : correlation between twins reared apart by unrelated adoptive parents

- **Narrow-sense heritability *h*²**: is the proportion of phenotypic variance due to additive genetic effects.

$$\boxed{h^2 = \frac{V_A}{V_P}}$$

# Heritability is not Necessarily Constant

Heritability can change over time because:

- the variance in genetic values can change,
- the variation due to environmental factors can change, or
- the correlation between genes and environment can change.

Genetic variance can change if:

- allele frequencies change (e.g., due to selection or inbreeding),
- new variants come into the population (e.g., by migration or mutation), or
- existing variants only contribute to genetic variance following a change in genetic background or the environment.

The same trait measured over an individual's lifetime may have different genetic and environmental effects influencing it, such that the variances become a function of age. For example,

- variance in weight at birth is influenced by maternal uterine environment,
- variance in weight at weaning depends on maternal milk production,
- but variance of mature adult weight is unlikely to be influenced by maternal factors.

# Misconceptions Regarding Heritability

Heritability is the proportion of a phenotype that is passed on to the next generation

High heritability implies genetic determination

Low heritability implies no additive genetic variance

Heritability is informative about the nature of between-group differences

A large heritability implies genes of large effect

# Methods of Estimating Heritability (in the absence of genetic data)

- Correlation/regression method
  - Parent-offspring regression
  - Close relatives comparison
  - Twin studies
- Analysis of variance method
  - ANOVA
- Linear mixed model
  - Pedigree-based
  - SNP-based

$h_{\text{SNP}}$

# Heritability Estimation (recently)

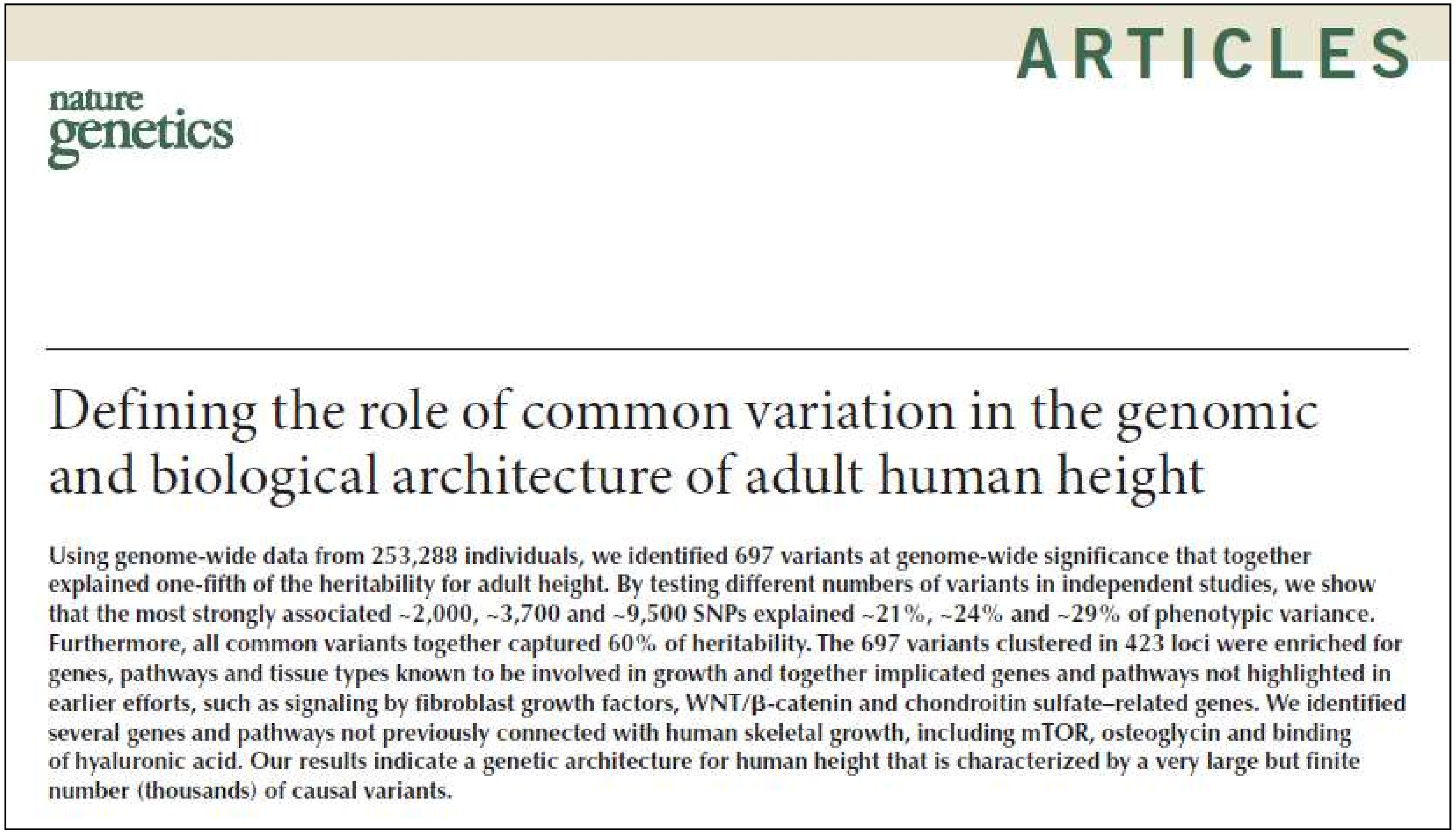
Array-based techniques, and more recently, whole-genome sequencing (WGS) techniques, have enabled accurate measurement of genotypes on millions of SNPs across the entire genome.

GWAS heritability ($h^2_{GWAS}$)

- Often estimated in a multi-SNP model to account for linkage disequilibrium among SNPs
- Significantly associated SNPs (genome-wide significance)

SNP-based heritability ($h^2_{SNP}$)

- The proportion of phenotypic variance explained by all SNPs on a genotyping array (hence dependent of the number of SNPs on a SNP array)
- The variance explained by any set of SNPs



ARTICLES

nature genetics

# Defining the role of common variation in the genomic and biological architecture of adult human height

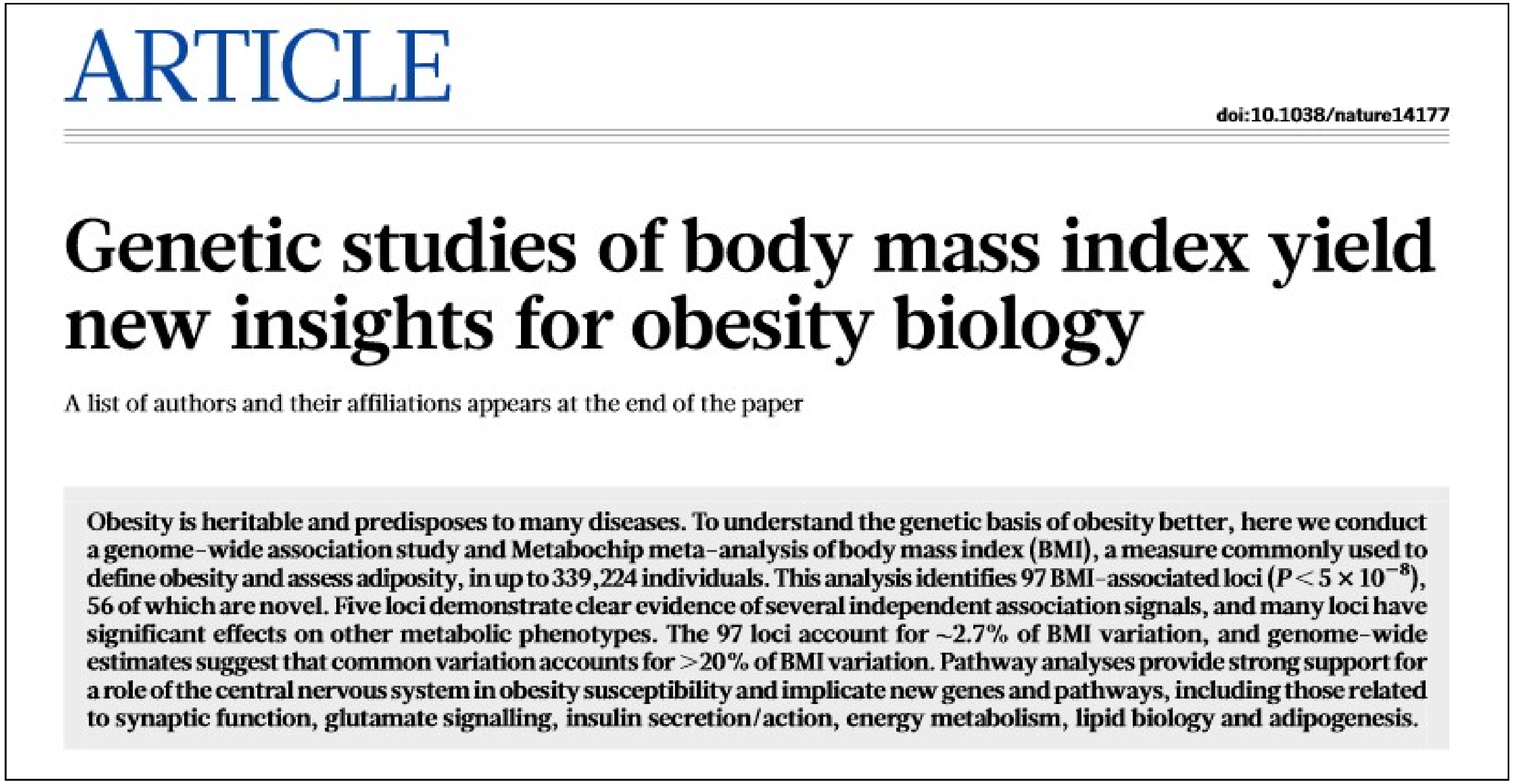
Using genome-wide data from 253,288 individuals, we identified 697 variants at genome-wide significance that together explained one-fifth of the heritability for adult height. By testing different numbers of variants in independent studies, we show that the most strongly associated ~2,000, ~3,700 and ~9,500 SNPs explained ~21%, ~24% and ~29% of phenotypic variance. Furthermore, all common variants together captured 60% of heritability. The 697 variants clustered in 423 loci were enriched for genes, pathways and tissue types known to be involved in growth and together implicated genes and pathways not highlighted in earlier efforts, such as signaling by fibroblast growth factors, WNT/β-catenin and chondroitin sulfate–related genes. We identified several genes and pathways not previously connected with human skeletal growth, including mTOR, osteoglycin and binding of hyaluronic acid. Our results indicate a genetic architecture for human height that is characterized by a very large but finite number (thousands) of causal variants.

**697 SNPs explain 20% of height**

**~2,000 SNPs explain 21% of height**

**~3,700 SNPs explain 24% of height**

**~9,500 SNPs explain 29% of height**



ARTICLE

doi:10.1038/nature14177

# Genetic studies of body mass index yield new insights for obesity biology

A list of authors and their affiliations appears at the end of the paper

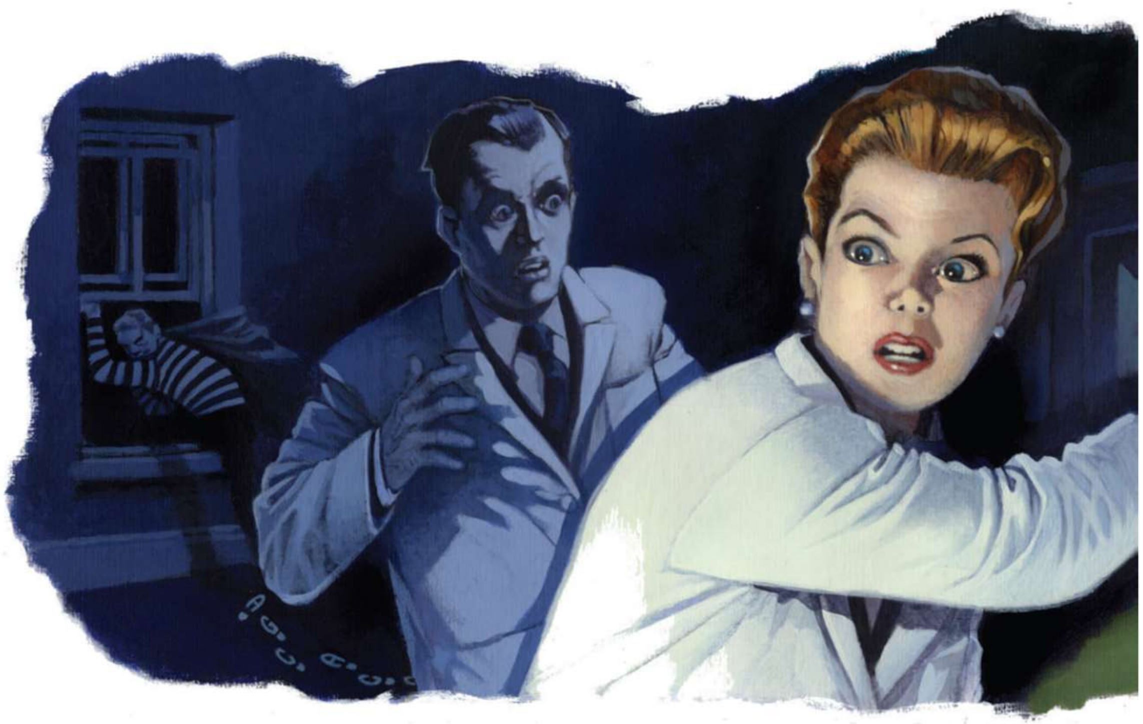
Obesity is heritable and predisposes to many diseases. To understand the genetic basis of obesity better, here we conduct a genome-wide association study and Metabochip meta-analysis of body mass index (BMI), a measure commonly used to define obesity and assess adiposity, in up to 339,224 individuals. This analysis identifies 97 BMI-associated loci ($P < 5 \times 10^{-8}$), 56 of which are novel. Five loci demonstrate clear evidence of several independent association signals, and many loci have significant effects on other metabolic phenotypes. The 97 loci account for ~2.7% of BMI variation, and genome-wide estimates suggest that common variation accounts for >20% of BMI variation. Pathway analyses provide strong support for a role of the central nervous system in obesity susceptibility and implicate new genes and pathways, including those related to synaptic function, glutamate signalling, insulin secretion/action, energy metabolism, lipid biology and adipogenesis.

**97 SNPs explain 2.7% of BMI**
**All common SNPs may explain 20% of BMI**

**Do we give up on GWAS, fine map everything, or think differently?**

# The missing heritability problem

The heritability for height explained by significantly associated SNPs is only 10%, while that explained by all measured SNPs is 45% -- still much smaller than a frequently quoted $h^2$ of 80% from family or twin studies.



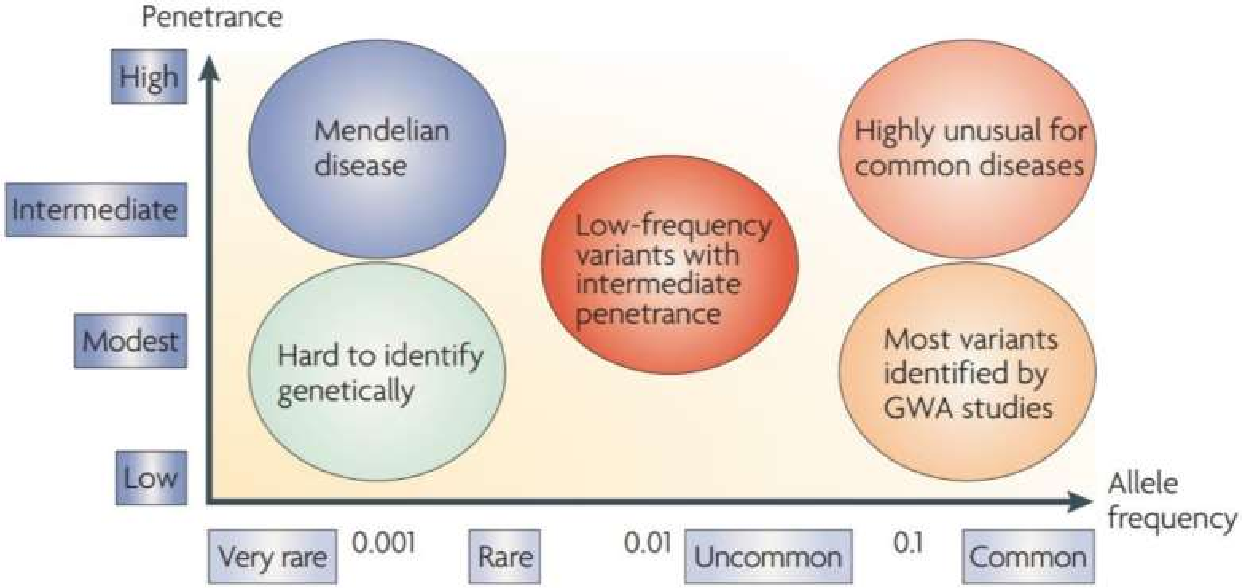
# The missing heritability problem

The presence of a large number of common variants of small effect yet to be discovered

Rare variants of large effect not tagged by common SNPs on genotyping arrays

Inflation in pedigree-based h2 due to shared environmental effects, non-additive genetic variation and/or epigenetic factors

# Genetic architecture


Penetrance
High
Intermediate
Modest
Low
Mendelian disease
Hard to identify genetically
Low-frequency variants with intermediate penetrance
Highly unusual for common diseases
Most variants identified by GWA studies
Allele frequency
Very rare
0.001
Rare
0.01
Uncommon
0.1
Common

McCarthy, M., Abecasis, G., Cardon, L. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9,** 356–369 (2008).

# CDCV/RAME/infinitesimal/Broad-sense-heritability

**Common Disease Common Variant (CDCV)**

- Complex disease is largely attributable to a moderate number of common variants, each of which explains several per cent of the risk in a population.

**The rare alleles of major effect (RAME) model**

- a large number of large-effect rare variants

**The infinitesimal model**

- A large number of small-effect common variants across the entire allele frequency spectrum

**Broad sense heritability model**

- Non-additive G×G and G×E interactions and epigenetic effects

# Thank you for your attention!