Next Generation Statistical Methods for Genome-wide Association Studies

Lecture 4: Heritability, Functional Enrichment and Polygenic scores

Presented by Haoyu Zhang

Oct 25th

Genetic Architecture

- A model for understanding overall nature of genetic associations underlying a trait
  - In contrast to focusing on individual associations

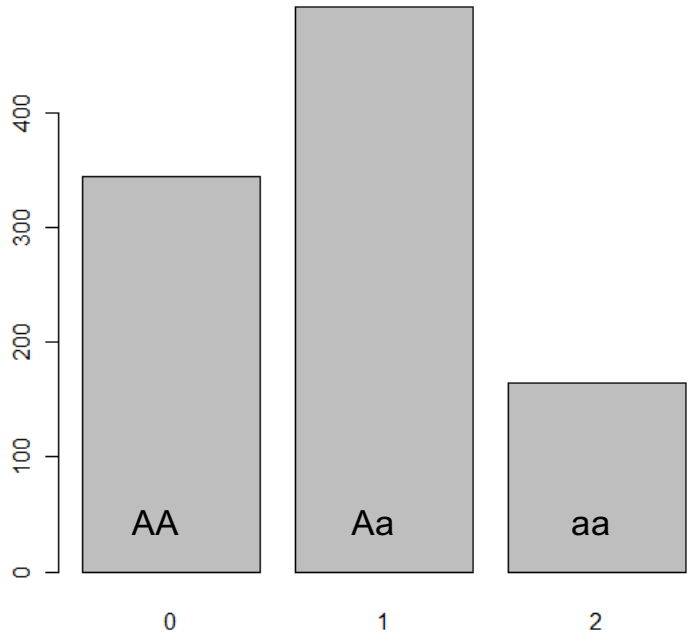
- Can be useful for understanding
  - Potential for genetic association studies
    - Discovery and genetic risk prediction
  - Biology
    - Functional enrichment
  - Population genetic signatures
    - E.g effect of selection

Genetic Architecture: Heritability

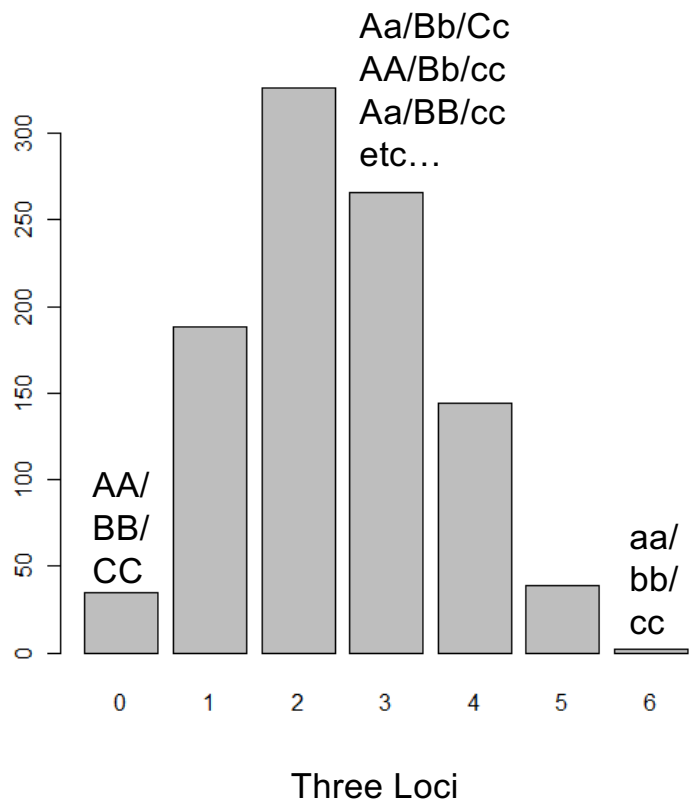
- Quantification of how much of the “variation” of a phenotype is explained by heritable factors

- Starting from the work of Galton in the late 19th century, family studies have been used to assess and characterize heritability in terms of
  - Additive and dominant genetic effects
  - Epistasis
  - Gene-environment interactions
  - Shared environment

1000 individuals,
MAF=0.4



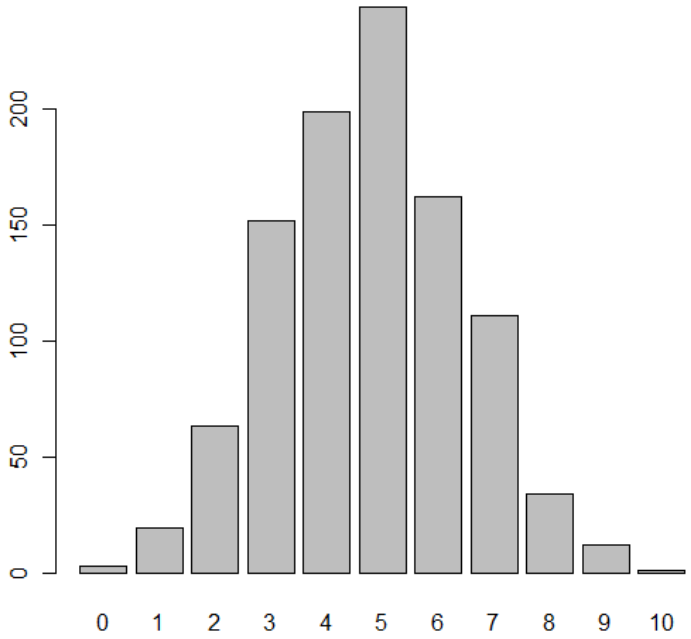
One Locus



1000 individuals, MAF=0.4

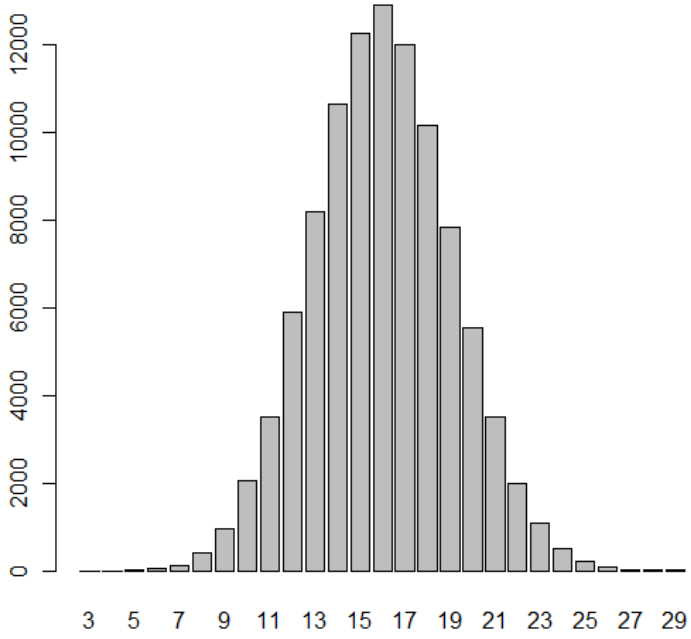
X-axis = Sum of the Minor Alleles across the three loci

1000 individuals,
MAF=0.4

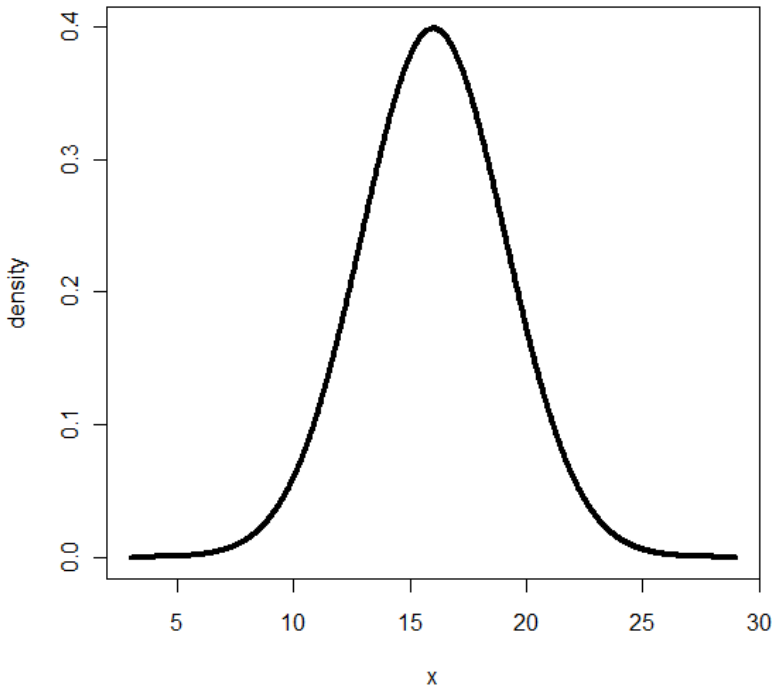


X-axis = Sum of
the Minor Alleles
across the six loci

100,000 individuals,
MAF=0.4



X-axis = Sum of the Minor Alleles across the twenty loci

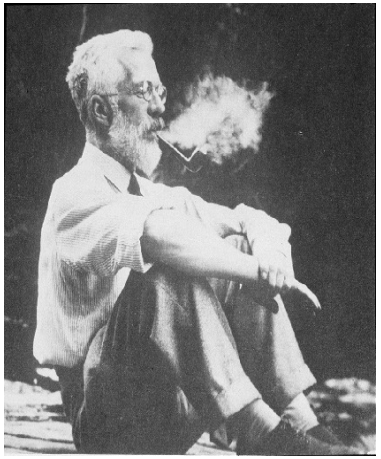


The Sum of the Minor Alleles across the twenty loci is approximately Normally Distributed with mean 16 and variance $40 \times 0.6 \times 0.4$

Key Point

Result from probability theory: the sum of many independent random variables is approximately normally distributed

The polygenic model



RA Fisher (1918)—many many loci each with small effects ⇒ normally distributed "polygenes" with noted covariance structure

Reconciled Mendelism and Biometry

XV.—**The Correlation between Relatives on the Supposition of Mendelian Inheritance.** By **R. A. Fisher**, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

CONTENTS.

	PAGE
1. The superposition of factors distributed independently	402
2. Phase frequency in each array	402
3. Parental regression	403
4. Dominance deviations	403
5. Correlation for parent; genetic correlations	404
6. Fraternal correlation	405
7. Correlations for other relatives	406
8. Epistacy	408
9. Assortative mating	410
10. Frequency of phases	410
11. Association of factors	411
12. Conditions of equilibrium	412
13. Nature of association	413
14. Multiple allelomorphism	415
15. Homogamy and multiple allelomorphism	416
16. Coupling	418
17. Theories of marital correlation; ancestral correlations	419
18. Ancestral correlations (second and third theories)	421
19. Numerical values of association	421
20. Fraternal correlation	422
21. Numerical values for environment and dominance ratios; analysis of variance	423
22. Other relatives	424
23. Numerical values (third theory)	425
24. Comparison of results	427
25. Interpretation of dominance ratio (diagrams)	428
26. Summary	432

Several attempts have already been made to interpret the well-established results of biometry in accordance with the Mendelian scheme of inheritance. It is here attempted to ascertain the biometrical properties of a population of a more general type than has hitherto been examined, inheritance in which follows this scheme. It is hoped that in this way it will be possible to make a more exact analysis of the causes of human variability. The great body of available statistics show us that the deviations of a human measurement from its mean follow very closely the Normal Law of Errors, and, therefore, that the variability may be uniformly measured by the standard deviation corresponding to the square root of the mean square error. When there are two independent causes of variability capable of producing in an otherwise uniform population distributions with standard deviations σ_1 and σ_2, it is found that the distribution, when both causes act together, has a standard deviation $\sqrt{\sigma_1^2+\sigma_2^2}$. It is therefore desirable in analysing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity the Variance of the normal population to which it refers, and we may now ascribe to the constituent causes fractions or percentages of the total variance which they together produce. It is desirable on the one hand that the elementary ideas at the basis of the calculus of correlations should be clearly understood, and easily expressed in ordinary language, and on the other that loose phrases about the "percentage of causation,"

A General Model for Heritability

- $Y = G + E$
- Heritability: $h^2 = var(G)/var(Y)$
  - Heritability is context-specific (different E -> different h^2)
  - Relative measures (functional enrichment, genetic correlation) may be more stable
- If there is only one genetic variant contributing to the effect as

$$Y = G\beta + \epsilon$$

- Heritability: $h^2 = 2f(1-f)\beta^2/var(Y)$
- Many genetic variants contributing the effects to outcomes
  - Random effect model

Missing Heritability Problem

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Estimating Heritability from GWAS Data

A Model for Joint-Effect of GWAS SNPs

- Assume genotypes and phenotypes are standardized to have mean zero and variance as: $G_k = \frac{G_k - 2f_k}{2f_k(1-f_k)}$, $Y = \frac{Y-\mu}{\sigma_Y}$

- Assume a linear additive model (in terms of causal variants)

$$Y_j = \sum_{m=1}^{M} G_{mj}\beta_m + \varepsilon_j = g_j + \varepsilon_j$$

$$\beta_m \sim N(0, h^2/M) \text{ (iid)}, Var(g_j) = M \times h^2/M = h^2$$

Yang et al., Nature Genetics, 2010

Variance-Covariance and Relationship Matrices

- The variance-covariance matrix for vector of phenotype in the sample can be written as

$$Var(\vec{Y}) = h^2 R + \sigma_\varepsilon^2 I_{N\times N}, \text{ where}$$

$$R = \vec{G}\vec{G}^T, \quad R_{ij} = 1/M \sum_{m=1}^{M} G_{im} G_{jm}.$$

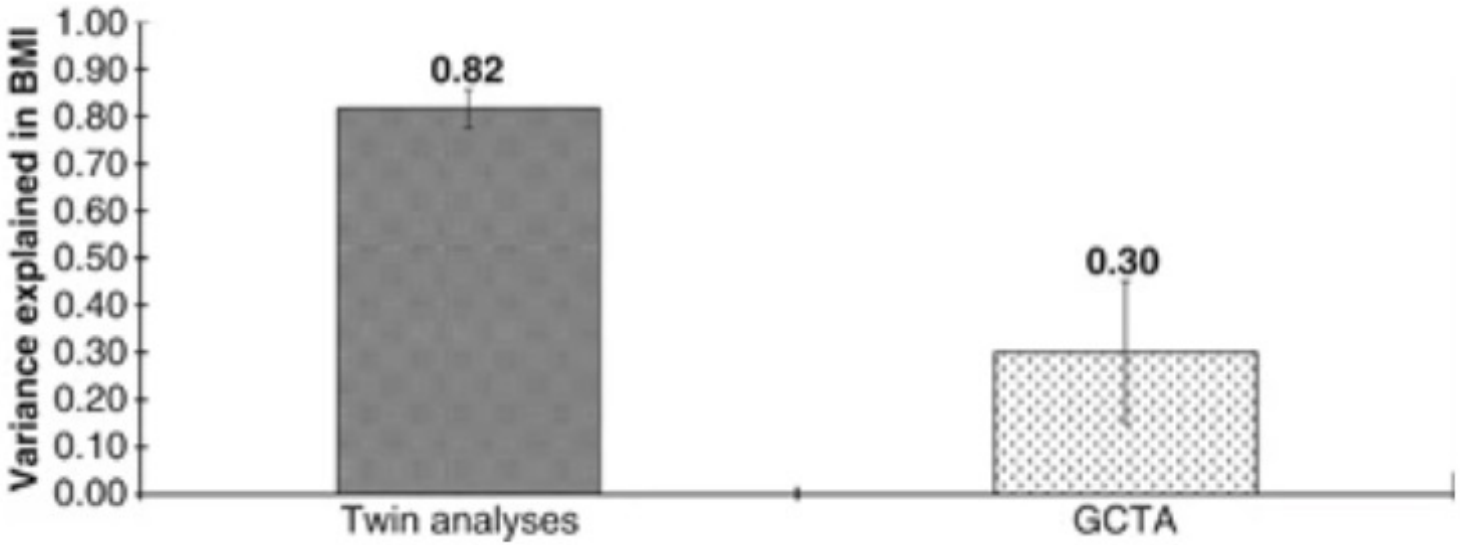
- Interestingly, the following "similarity" regression hold

$$E((Y_i - Y_j)^2 \,|\, \vec{G}_i, \vec{G}_j) = 2 - 2h^2 R_{ij}$$

Yang et al., Nature Genetics, 2010

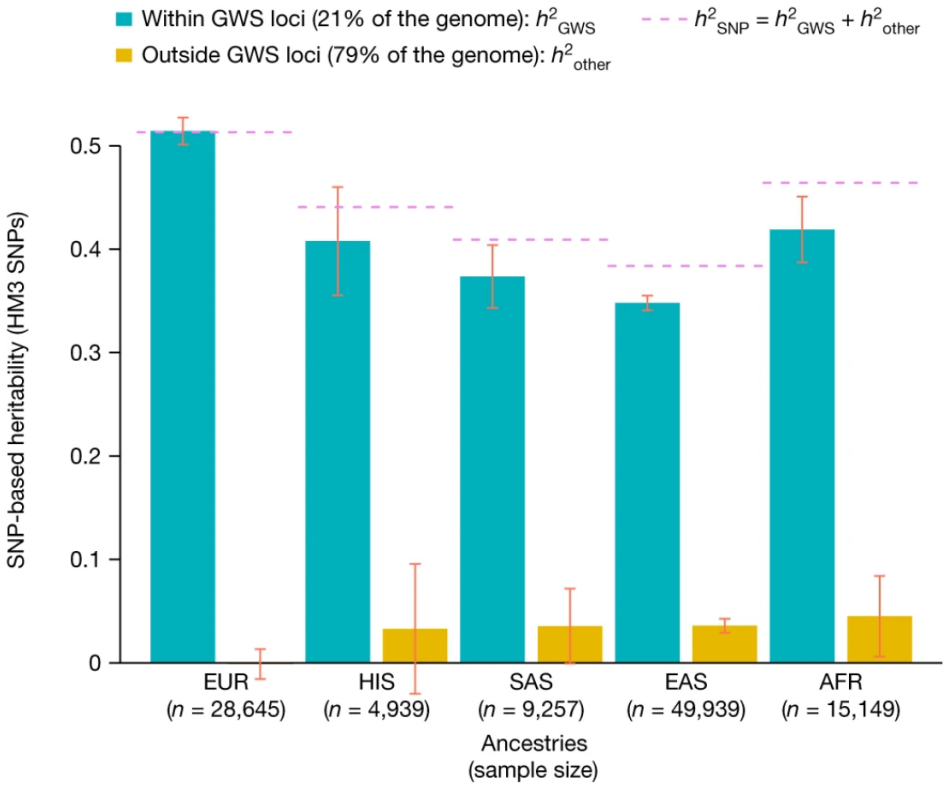
Comparison of Analyses Results between Twin Studies and GCTA for BMI

Genome-wide significant SNPs explain < 2% of heritability for BMI in 2013



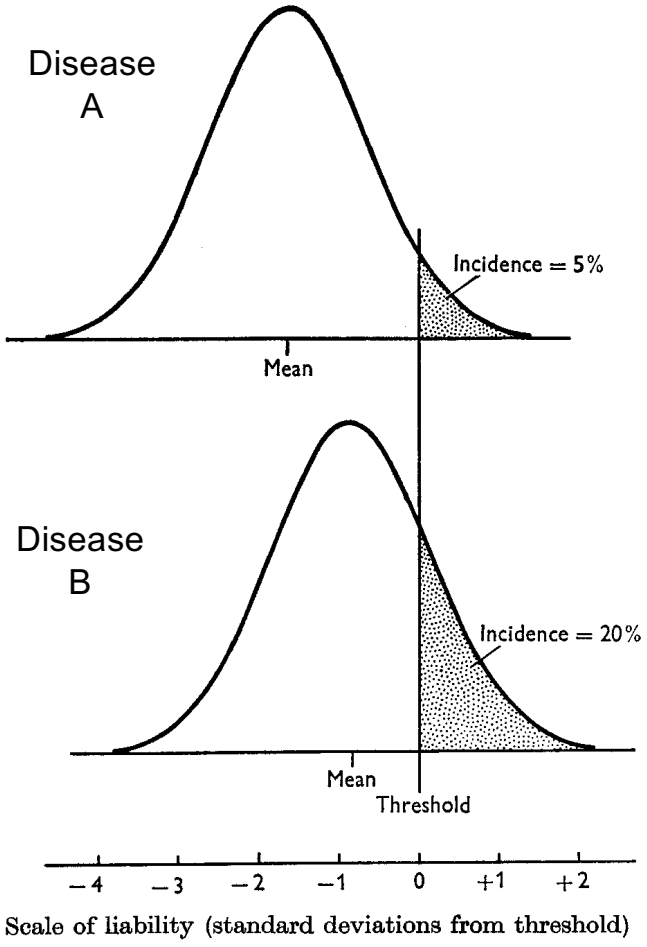
Liewellyn, International Journal of Obesity, 2013

Missing Heritability Found for Height



A combined analysis of 281 genome-wide association studies finds 12,111 common DNA variants associated with a person's height — and shows that larger studies will not yield more variants in populations of European ancestry.
(Yengo, et al. Nature, 2022)

Estimating Heritability for Binary Traits



Basic idea: everybody has some unobserved normally-distributed continuous trait, called a liability, that defines case status. If an individual's liability is above a certain threshold (that depends on incidence), then that individual has disease.

Assume
liability = G + E
and define heritability as before.

Falconer (1965) Ann Hum Genet

Liability Threshold Model for Binary Traits

- Assume each binary trait results from an underlying liability variable (latent) exceeding a threshold

$L \sim N(X\beta, \sigma^2)\ and\ L > c$ <=> $Y = 1$ (Liability Threshold Model)

$\Pr(Y = 1|X) = 1 - \Phi(\{c - X\beta\}/\sigma)$ (Probit model)

- note under this model effects are identifiable in terms of per unit of the variance of L

- assume without loss of generality σ^2=1

GCTA for Binary Traits

- Binary traits
  - Run GCTA (ie LMM) in the observed (y=0/1) scale to estimate heritability as h_0^2
  - Then obtain estimate of heritability in the liability threshold scale by making the transformation
  - $h_l^2 = h_0^2 K(1-K)/z^2,\ K = \Pr(D=1),\ c = \Phi^{-1}(1-K),\ z = \phi(c)$
    - Derived from using properties of moment of the truncated normal distribution $(L|L > c)$
    - $cov(y, l) = z \Longrightarrow u = c + zg$
      - Relationship between risk in liability and observed scale
  - Further adjustment for case-control studies
  - $h_l^2 = h_0^2\{\frac{K(1-K)}{z^2}\} \times \frac{K(1-K)}{P(1-P)},\ P = sample\ prevalence$

Lee et al., AJHG, 2011

Estimating Heritability using GWAS Summary Data

Summary-level Association Statistics are Widely Available

- Results from standard one-SNP-at-a-time GWAS analysis
  - Linear regression coefficients and standard errors
  - Log-odds-ratios and standard errors

- Many associations within same regions simply because of correlation across SNPs
  - Association results for SNPs are not adjusted using a multivariate model

- Can we make inference about an underlying joint model based on summary-level data?

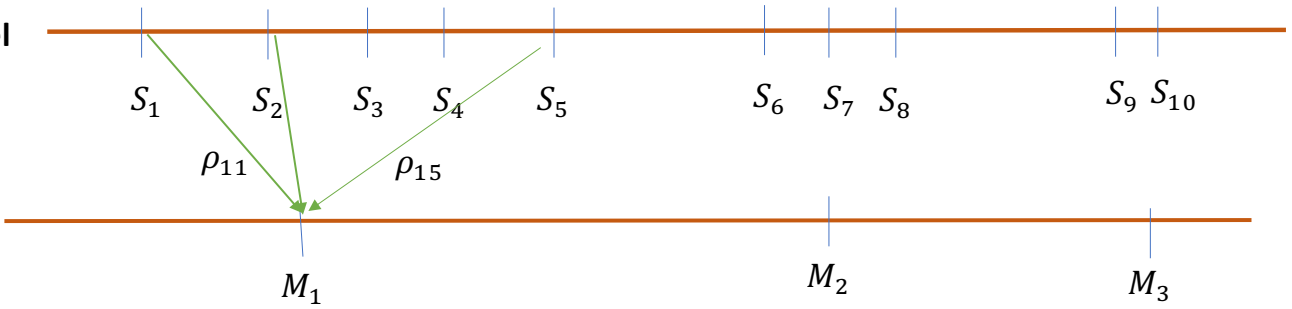
LD-score regression

- A technique for estimating heritability and co-heritability using only summary-level data

- The strength of association statistics for a GWAS marker is expected to be linearly related with degree of its LD-score – the total LD the marker has with all SNPs in the genome
  - Markers that are in high LD with many underlying SNPs are expected to show stronger association under highly polygenic model
  - Strength of the relationship is determined by the heritability of the trait and sample size of the study

Association Parameters in Marginal and Joint Models are Related by Linear Set of Equations Defined by LD-coefficients

$$Y = \sum_k \beta_k S_k + \epsilon, \beta_k \sim \mathrm{N}(0, h_g^2/K)$$



GWAS marginal Association coefficients

$\theta_1 = \sum_{k=1}^{5} \rho_{k1}\beta_k$ $\quad$ $\theta_2 = \sum_{k=6}^{8} \rho_{k2}\beta_k$ $\quad$ $\theta_3 = \sum_{k=9}^{10} \rho_{k3}\beta_k$

Summary-level data

$(\widehat{\theta}_1, s_1)$ $\quad$ $(\widehat{\theta}_2, s_2)$ $\quad$ $(\widehat{\theta}_3, s_1)$

LD-score Regression

- A technique for estimating heritability/co-heritability using summary-level data

$$E(Z_m^2) = N(\frac{h_g^2}{K} l_m + a) + 1, l_m = \sum_k \rho_{mk}^2$$

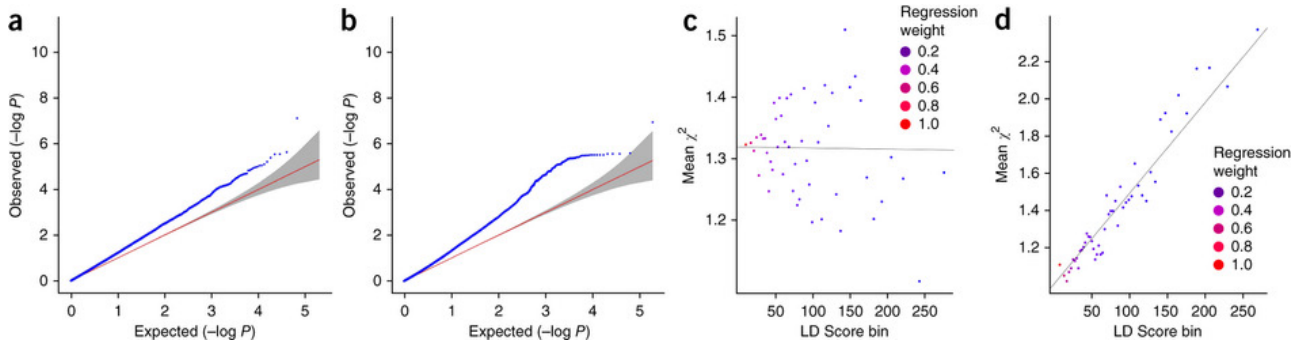
  - $Z_m = \sqrt{N}\hat{\theta}_m$ ($\hat{\theta}_m$ marginal association-statistics after genotype standardization)
  - LD-score (l_m) can be estimated from external reference panel

- Marginal effects of SNPs are driven by sum of PS-effect (not related to LD patterns) and polygenic effects (related to LD patterns)

$$\theta_m = u_m + \sum_k \rho_{mk} \beta_k \Rightarrow Var(\theta_m) = Var(u_m) + \frac{h_g^2}{K} \sum_k \rho_{mk}^2 = a + \frac{h_g^2}{K} l_m$$
$$\hat{\theta}_m | \theta_m \sim N(\theta_m, 1/N)$$

Builick-Sullivan et al., Nature Genetics, 2015

LD-score Regression can Distinguish Population Stratification and Polygenetic Effects



PS effect only

Polygenic effects only

Log-linear Model for Binary Disease Outcome

- $\Pr(D = 1|U) = \exp(\alpha + U)\,, U \sim N(0, \sigma^2)$ (assuming no fixed effects)

- Under an additive model: $\lambda_{sib} = \exp(0.5\sigma_A^2)$ (Pharaoh et al., 2002, Nat. Genet.)

- Identified SNPs heritability: $\sum \mathbf{2}\, \boldsymbol{p_i}(\mathbf{1} - \boldsymbol{p_i})\{\widehat{\boldsymbol{\beta}^2} - \boldsymbol{var}(\widehat{\boldsymbol{\beta}})\}$

- Logistic model doesn't require the population prevalence which is usually hard to estimate

Example Results for Breast Cancer Subtypes

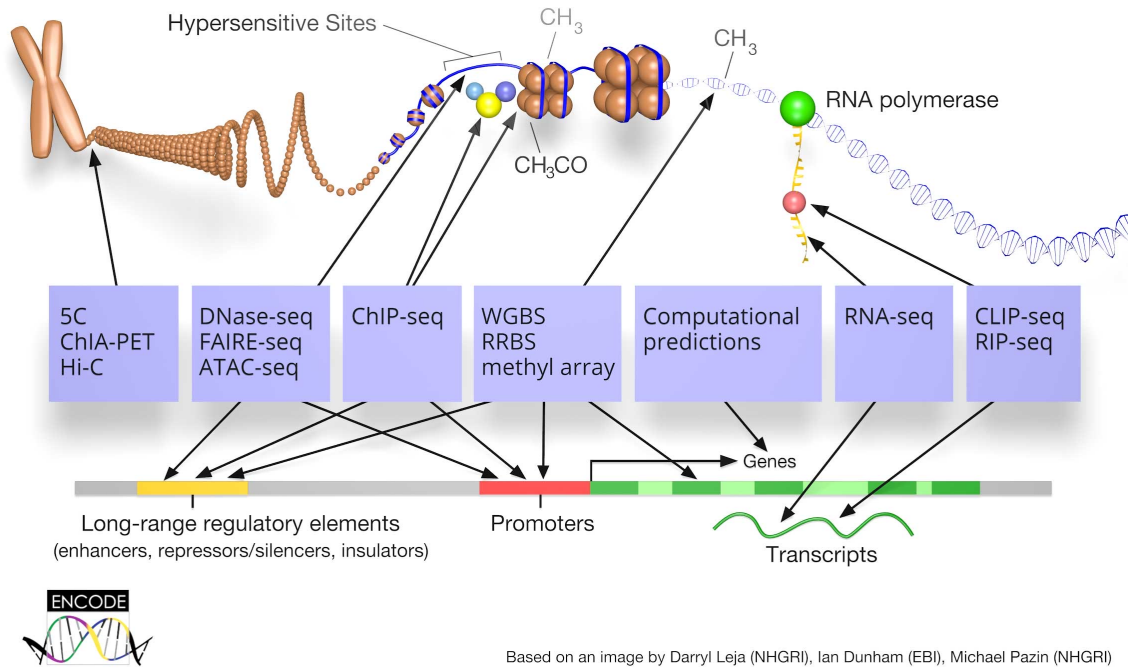
Phenotype	Heritability (all identified 210 SNPs) $\sum 2\,p_i(1-p_i)\{\widehat{\beta^2} - var(\widehat{\beta})\}$	Heritability (se) (all GWAS variants)	**Proportion**
Luminal A	0.336	0.620 (0.056)	54.2%
Luminal B	0.270	0.740 (0.093)	36.5%
Luminal_B_HER2Neg	0.233	0.597 (0.077)	38.9%
HER2 enriched	0.200	0.689 (0.154)	29.1%
Triple negative	0.185	0.492 (0.072)	37.6%
CIMBA *BRCA1*	0.083	0.3094 (0.0813)	26.9%

Proportion of heritability explained by identified SNPs over all GWAS variants

Enrichment Analyses Using LD-score Regression

Stratified LD-score Regression for Estimation of Enrichment of Associations by Genome Annotations

ENCODE: Large Scale Genome-wide Functional Genomics Study using Cell Lines and Tissue Samples



Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Transcription factor Binding

Epigenetic meachanisms

Chromatin accessibility

Role of promoter and enhancer sequences

Alternative splicing

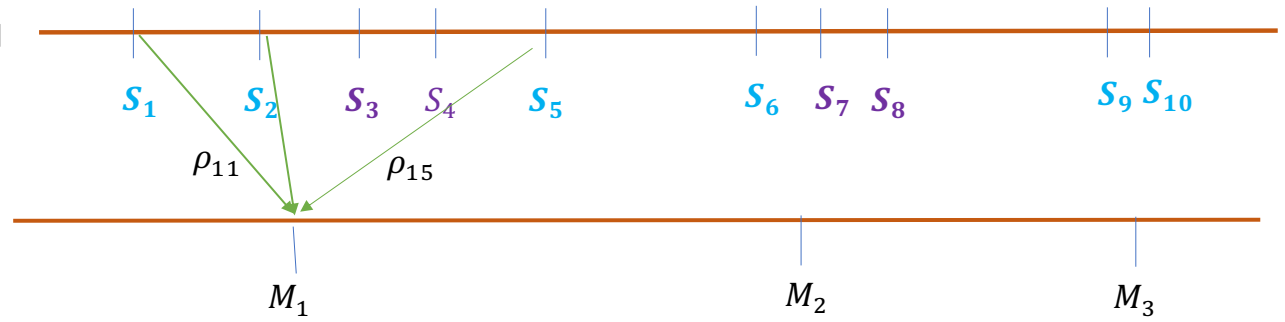
Chromosomal folding and interaction

Tissue and cell-type specificity

Stratified LD-score Regression for Estimation of Enrichment of Associations by Genome Annotation

$$Y = \sum_k \beta_k S_k + \varepsilon, \beta_k \sim N(0, \sum_{C:k \in C} \tau_C)$$

$C = sets\ corresponding\ to\ genome\ annotation$



GWAS markers

GWAS marginal Association coefficients $\theta_1 = \sum_{k=1}^{5} \rho_{k1}\beta_k$ $\theta_2 = \sum_{k=6}^{8} \rho_{k2}\beta_k$ $\theta_3 = \sum_{m=9}^{10} \rho_{k3}\beta_k$

Summary-level data $(\hat{\theta}_1, s_1)$ $(\hat{\theta}_2, s_2)$ $(\hat{\theta}_3, s_1)$

Stratified LD-score Regression

- Key Equation
  - $E(Z_m^2) = \mathrm{N} \times \{\sum_C \tau_C \, l_{m,C} + a\} + 1$
  - $l_{m,C} = \sum_{k \in C} \rho_{mk}^2$ (stratified LD-score)

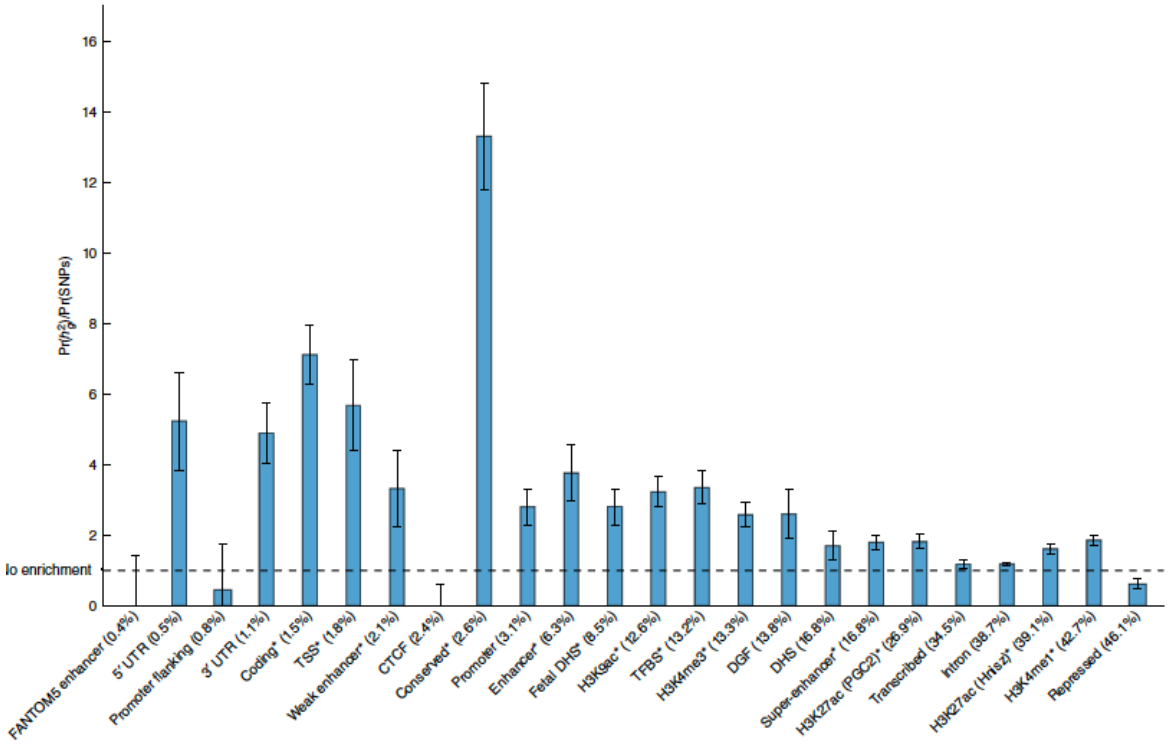
- Annotation can partition genome to overlapping/non-overlapping regions
  - Same SNPs can contribute to multiple categories
  - τ_C= per-SNP heritability of category C
  - $\sum_{C:j \in C} \tau_C$= per-SNP heritability of j-th SNP

Finucane et al., Nature Genetics, 2015

Data Analysis: Setting Up the Design Matrix for SNPs

- Base model defined by 53 annotation categories (overlapping) not specific to cell types
  - Promoter, enhancer, DHSs,……
- Individual cell-type specific annotation are further incorporated in addition to the base model
  - Allows assessment of whether cell-type specificity leads to additional enrichment
- Sample sizes recommendations
  - Continuous traits more than 5,000 people
  - Binary traits: more than 10,000 controls and 10,000 cases

Stratified LD-score Regression Allows Assessment of Enrichment of Polygenic Signals by Functional SNP Categories over Nine Traits



Enrichment Analyses Results for Specific Traits

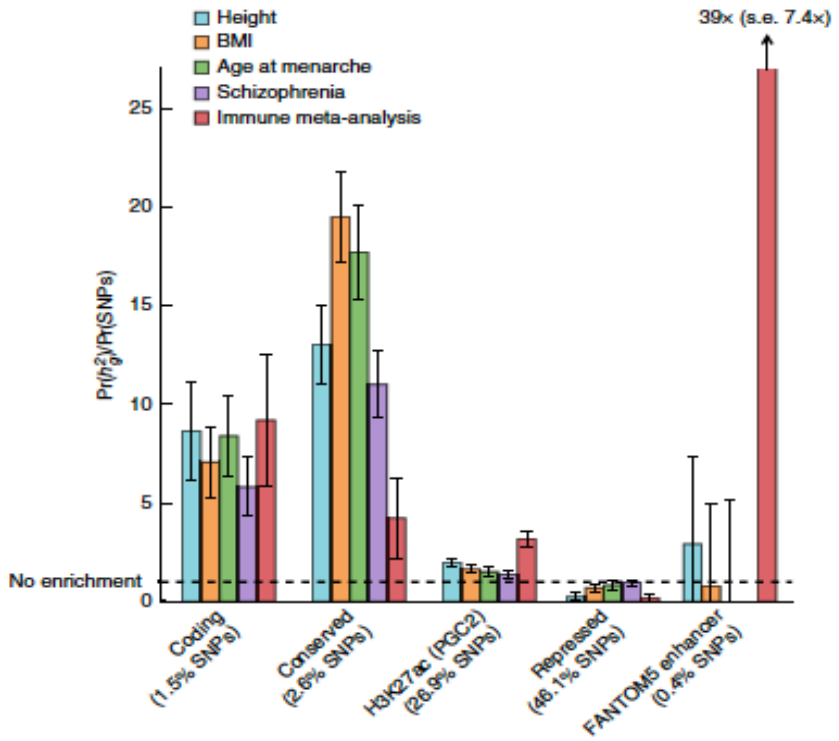


Figure 5 Enrichment estimates for selected annotations and traits. Error bars represent jackknife standard errors (s.e.) around the estimates of enrichment.

Finucane et al., Nature Genetics, 2015

Genetic Correlation Estimation Using LD-score Regression

LD-score regression can be used to estimate genetic correlation across traits

- Mixed model for genetic correlation

$$Y_1 = \underset{\substack{GWAS \\ SNPs}}{X}\ \beta_1 + \varepsilon_1,\ Y_2 = \underset{\substack{GWAS \\ SNPs}}{X}\ \beta_2 + \varepsilon_2,$$

Genetic Correlation explained by GWAS SNPs

$$cov(\beta_1, \beta_2) = \begin{pmatrix} h_{1G}^2 & \rho_G h_{1G} h_{2G} \\ \rho_G h_{1G} h_{2G} & h_{2G}^2 \end{pmatrix}$$

Total Phenotypic Correlation

$$cov(Y_1, Y_2) = \begin{pmatrix} \sigma_{1P}^2 & \rho_P \sigma_{1P} \sigma_{2P} \\ \rho_G \sigma_{1P} \sigma_{2P} & \sigma_{2P}^2 \end{pmatrix}$$

The LD-score equation for genetic correlation Analysis

$$E(Z_{1m}Z_{2m} \mid l_m) = \frac{N_s}{\sqrt{N_1 \times N_2}}\rho + \sqrt{N_1 \times N_2}\rho_g l_m$$

Phenotypic correlation

Genetic correlation

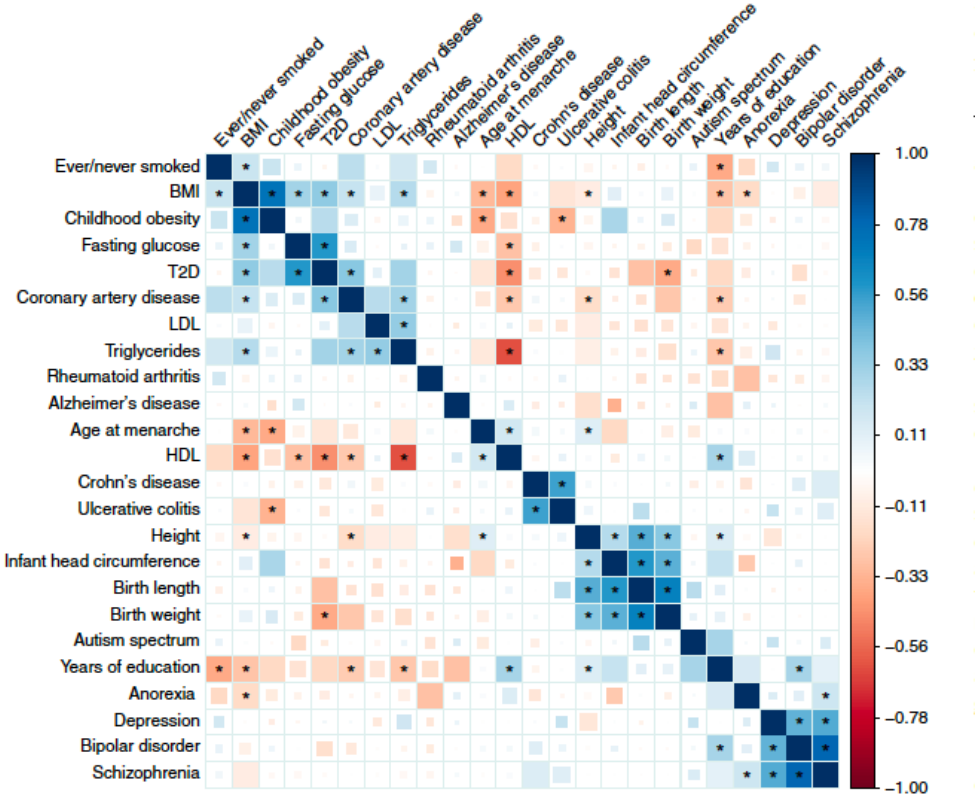
$N_1 = Sample\ size\ for\ GWAS\ of\ trait\ 1$

$N_2 = Sample\ size\ for\ GWAS\ of\ trait\ 2$

$N_s = Overlapping\ sample\ size$

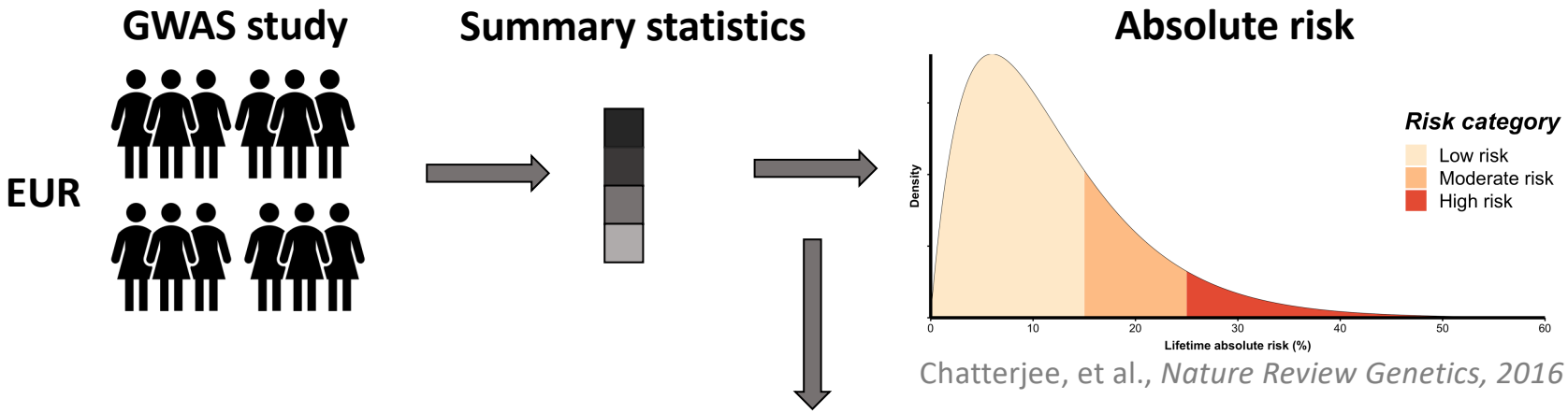
Bulick-Sullivan et al., 2015, Nature Genetics

An Atlas of Genetic Correlation



Polygenic Risk Scores

Polygenic Risk Scores (PRS) efficiently stratify populations by disease risk

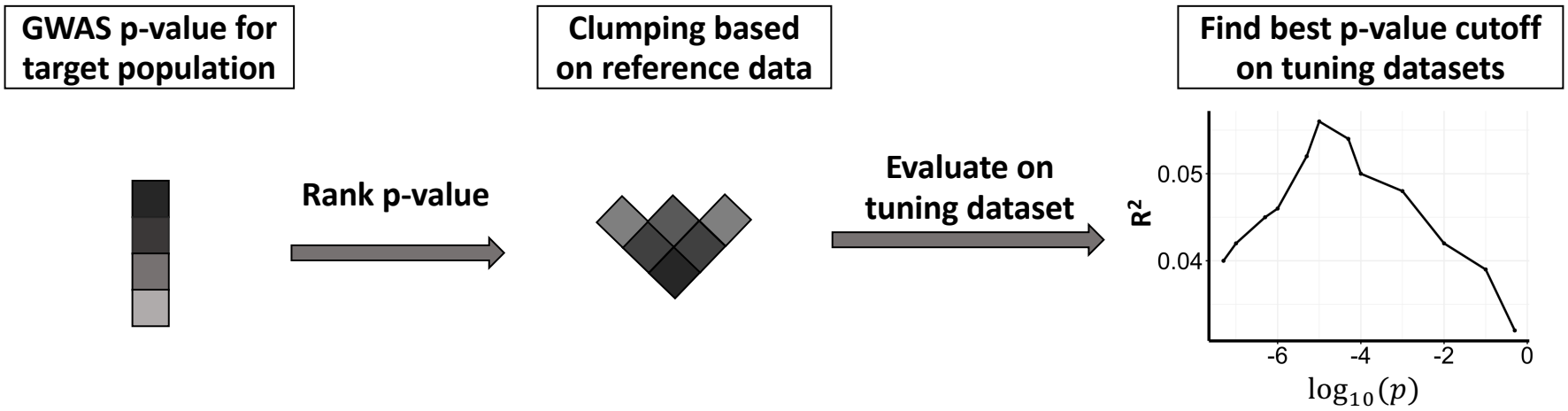


Chatterjee, et al., *Nature Review Genetics, 2016*

$$PRS = \beta_1 SNP_1 + \beta_2 SNP_2 \ldots + \beta_p SNP_p$$

Effect size **Number of risk alleles** **Number of SNPs**

Review of single ethnic clumping and thresholding approach



- Selected SNPs are relatively independent
- Prediction performance will first increase and then decrease as p-value cutoff increase
- For highly polygenic traits, the p-value cutoff tend to be large

PRS building Based on Effect-Size Distribution: LDPred

- Assumes spike and slab type of model for effect-size

$$pr(\beta^J) \sim \pi_0 \delta_0 + (1 - \pi_0) N(0, \sigma^2)$$

- Optimal PRS should have the form

$$- PRS = \sum_{m=1}^{M} E\{ \beta_m^{(J)} | \text{Data} \} G_m$$

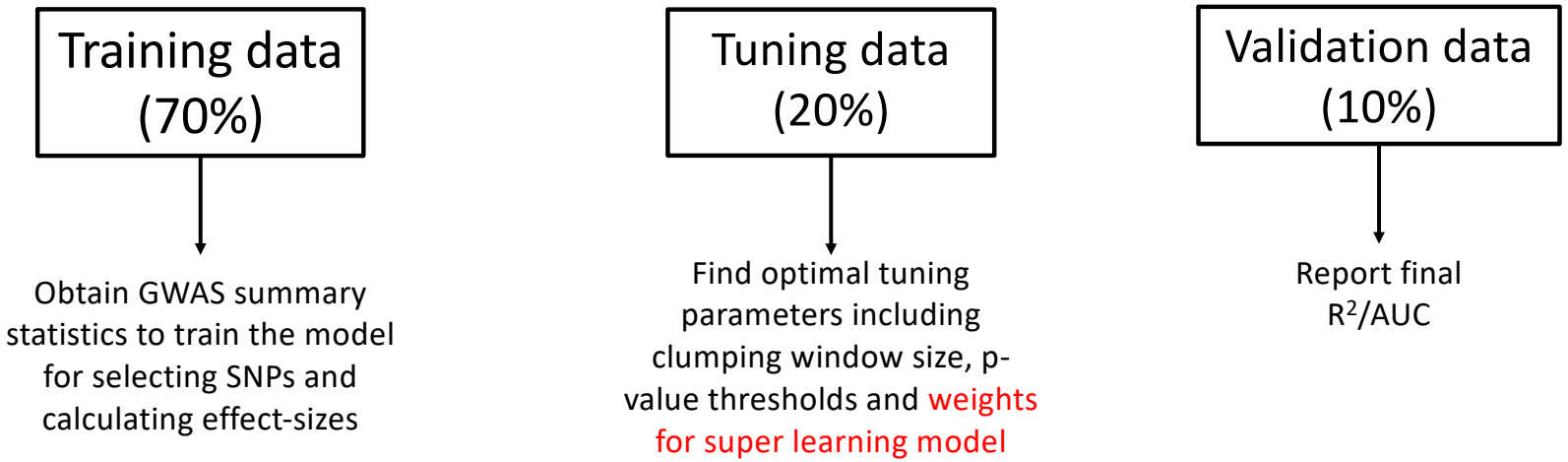
Bjarni Vilhjálmsson, et al., American Journal of Human Genetics, 2015

Summary of PRS Methods

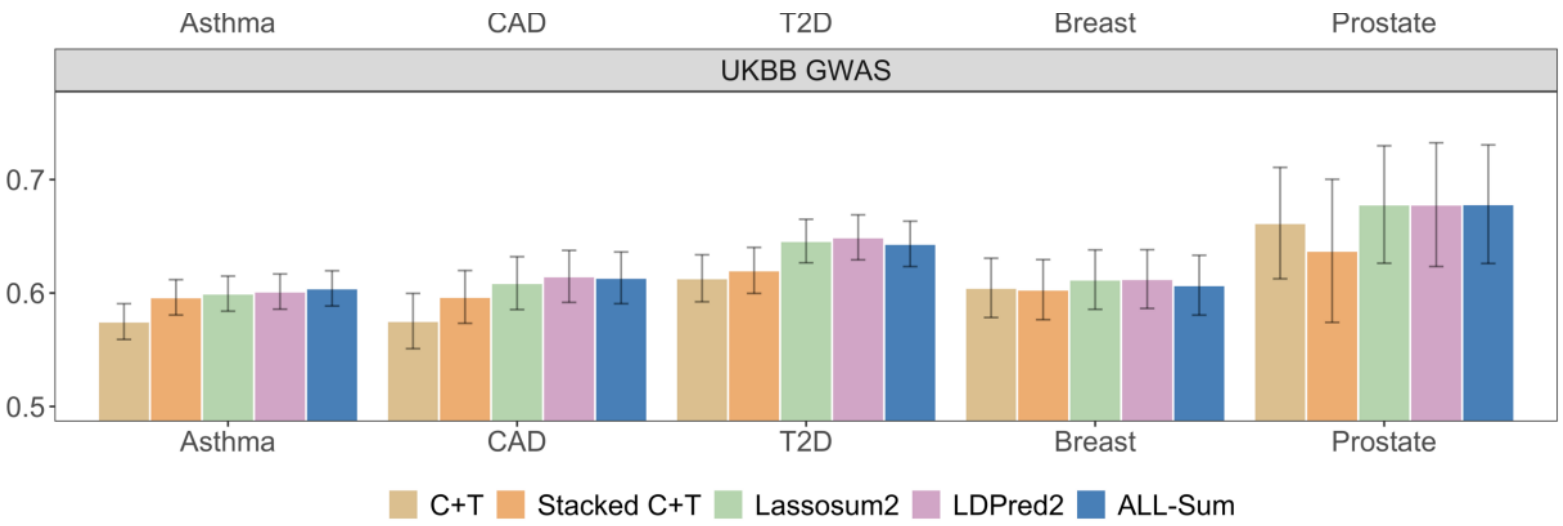
- Input
  - individual level data OR summary stats
  - LD reference
  - Algorithm
- Trade-off between complexity of the PRS and prediction power
- Summary Statistics + LD -> high power
- Mismatch of reference panel and potential in-sample LD

Design for simulations and real data analysis

- Data sets are split into training data, tuning data and validation dataset



Examples of PRS Prediction



Tony Chen, et al., biorxiv, 2023