

Lecture 6

Exploring Heritability: From Early Twin Studies to Modern SNP-Based Approaches and Challenges

by Dr. Mustafa İsmail Özkaraca

Contents

1. Understanding Heritability
2. Early Methods: Twin Studies
3. Modern Approaches: SNP-Based Heritability
4. Challenges in Heritability Studies

Understanding Heritability

Nature vs Nurture

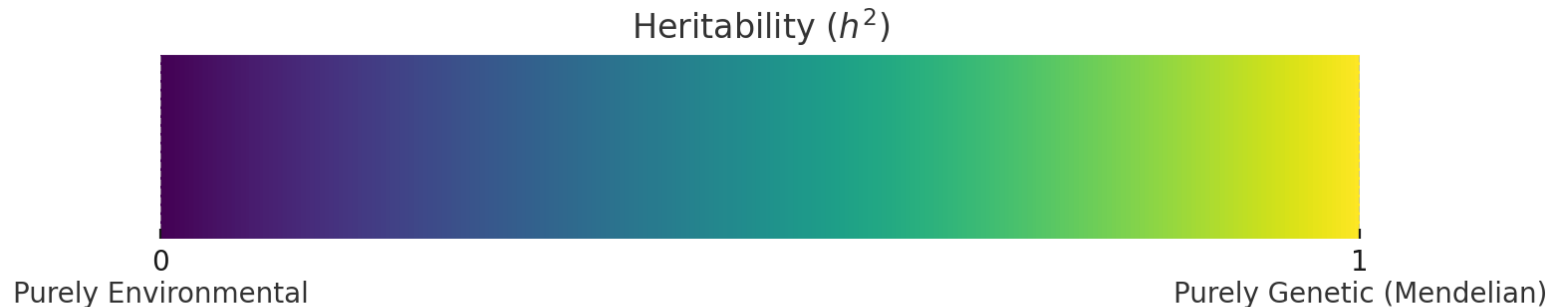
What causes variation in traits?

A. Genetics (e.g. DNA differences)

B. Environment (e.g. nutrition, lifestyle)

Heritability

Measures the proportion of trait variation due to genetics.



Early Methods: Twin Studies

Compare Monozygotic (**MZ**) twins with Dizygotic (**DZ**) twins (**ACE Model**)

ACE Model

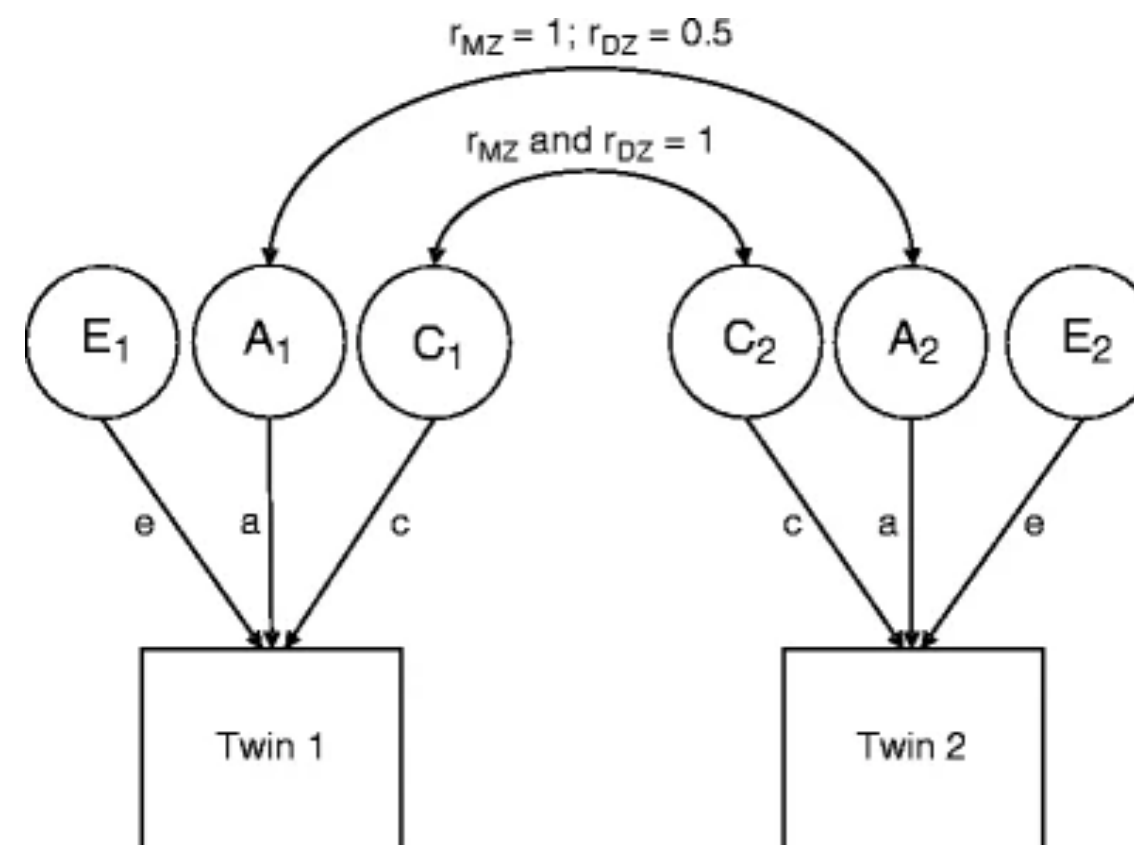
$$V_y = V_A + V_C + V_E$$

V_y : Total phenotypic variance ($V_y = 1$),

V_A : Additive genetic variance (**heritability** - because $V_y = 1$),

V_C : Shared environmental variance,

V_E : Unique environmental variance.



$$r_{MZ} = V_A + V_C$$

$$r_{DZ} = \frac{1}{2}V_A + V_C$$

$$V_A = 2(r_{MZ} - r_{DZ}) \text{ and } V_C = 2r_{DZ} - r_{MZ}$$

Figure is from:

Verweij, K.J.H., Mosing, M.A., Zietsch, B.P., Medland, S.E. (2012). Estimating Heritability from Twin Studies. In: Elston, R., Satagopan, J., Sun, S. (eds) Statistical Human Genetics. Methods in Molecular Biology, vol 850. Humana Press.

Early Methods: Twin Studies

Issues in Study Design

- ▶ Equal Environment Assumption
- ▶ Shared Environment Influence
- ▶ Generalisability
- ▶ Sample Size Limitations

Solution: Genome-wide Approach!

Modern Approaches: SNP-Based Heritability

Method 1: Haseman–Elston (HE) regression [1,2]

Resemblance between related individuals

$$y_i \cdot y_j = a + h^2 \pi_{i,j} + \epsilon_{i,j}$$

y_k : Standardised phenotype of k -th individual

$\pi_{k,l}$: Genetic relatedness between individuals k and l .

a : Intercept

$\epsilon_{i,j}$: Random noise

Why this model? [3]

$$y_i \cdot y_j = a + m \pi_{i,j} + \epsilon_{i,j}$$

Solve the model for m (under the assumption $y = g + e$ with $\text{Cov}(g, e) = 0$ and $g \sim \mathcal{N}(0, \sigma_G^2 \pi)$):

$$m = \frac{\text{cov}(\pi_{i,j}, y_i y_j)}{\text{var}(\pi_{i,j})} \stackrel{\text{[A]}}{=} \frac{\text{cov}(\pi_{i,j}, g_i g_j)}{\text{var}(\pi_{i,j})} \stackrel{\text{[B]}}{=} \frac{E(\pi_{i,j} g_i g_j) - \sigma_G^2 R_{i,j}^2}{\text{var}(\pi_{i,j})} \stackrel{\text{[C]}}{=} \frac{\sigma_G^2 E(\pi_{i,j}^2) - \sigma_G^2 R_{i,j}^2}{\text{var}(\pi_{i,j})} = \frac{\sigma_G^2 \text{var}(\pi_{i,j})}{\text{var}(\pi_{i,j})} = \sigma_G^2$$

where $R_{i,j} = E(\pi_{i,j})$

[A] $\text{Cov}(g, e) = 0$.

[B] $g \sim \mathcal{N}(0, \sigma_G^2 \pi)$.

[C] $E(\pi_{i,j} g_i g_j) = E \left(E(\pi_{i,j} g_i g_j \mid \pi_{i,j}) \right) = E(\pi_{i,j} \cdot E(g_i g_j \mid \pi_{i,j})) = E(\pi_{i,j} \cdot \sigma_G^2 \pi_{i,j}) = \sigma_G^2 E(\pi_{i,j}^2)$.

↑
Law of total expectation

[1] Haseman, J.K., Elston, R.C. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2, 3–19 (1972).

[2] Elston, Robert C., et al. "Haseman and Elston revisited." Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 19.1 (2000): 1-17.

[3] Visscher, Peter M., et al. "Statistical power to detect genetic (co) variance of complex traits using SNP data in unrelated samples." PLoS genetics 10.4 (2014): e1004269.



Modern Approaches: SNP-Based Heritability

Method 2: REstricted-Maximum Likelihood (REML) methods [4]

Estimate covariance matrices

$$y = X_c\beta + G + E$$

X_c :Matrix of covariates,

β :Fixed effects,

G :Total genetic effects, $G \sim \mathcal{N}(0, \pi\sigma_G^2)$, where π is the GRM,

E :Residuals, $E \sim \mathcal{N}(0, I\sigma_E^2)$.

REML Steps

1. Adjusts for fixed effects ($X_c\beta$)
2. Maximum Likelihood for estimating σ_G^2 and σ_E^2 .

$$3. h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}.$$

[4] Yang, Jian, et al. "GCTA: a tool for genome-wide complex trait analysis." *The American Journal of Human Genetics* 88.1 (2011): 76-82.

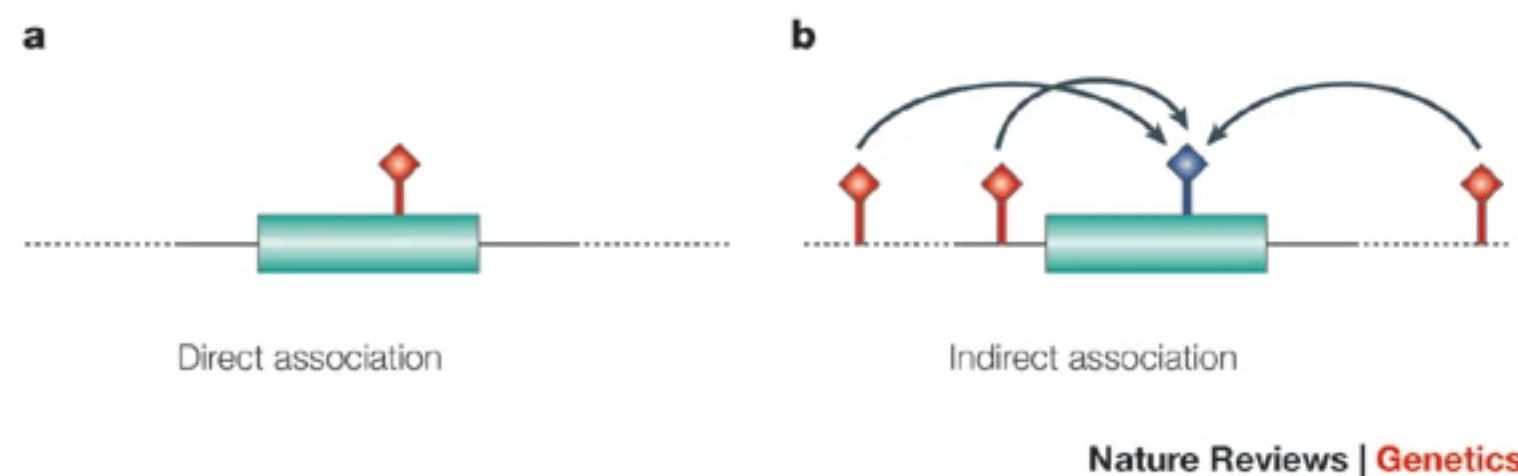


Modern Approaches: SNP-Based Heritability

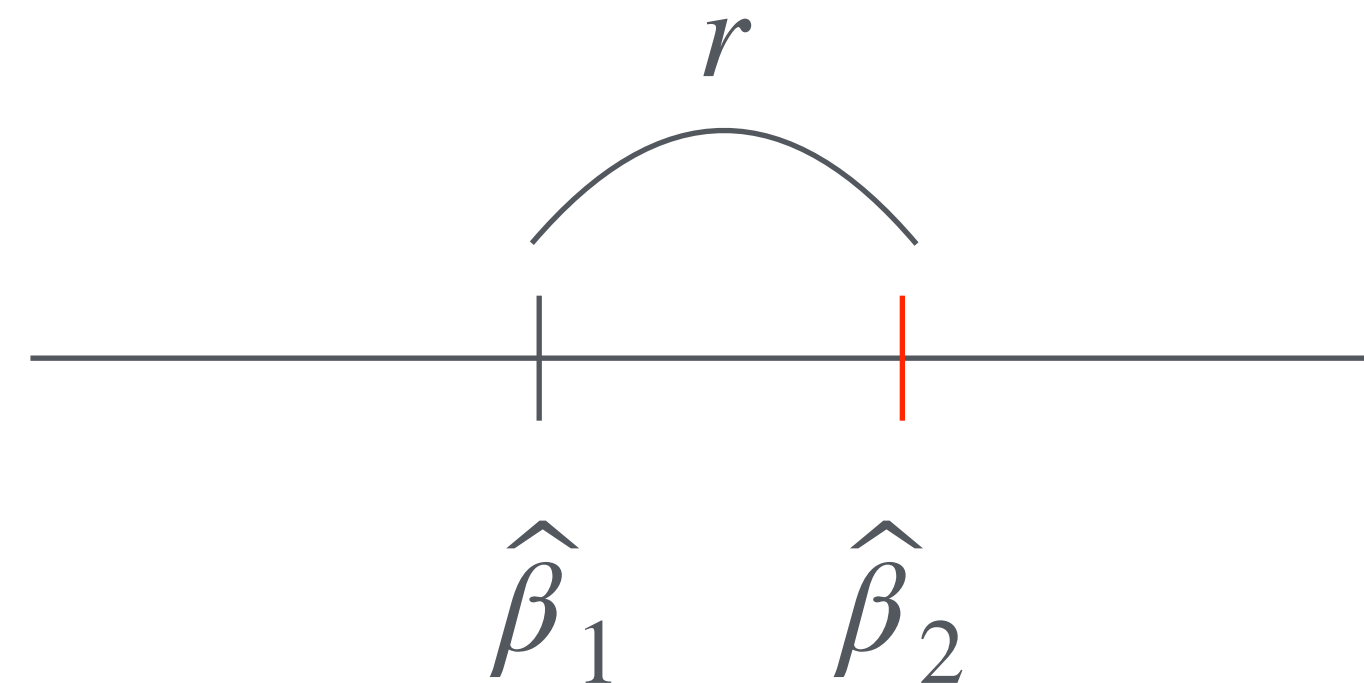
Method 3: LD Score Regression [5]

From GWAS summary statistics

LD Problem



LD block of 2 SNPs



$$y = g_1\beta + \epsilon \quad \text{where } \epsilon \sim \mathcal{N}(0, \text{Var}(\epsilon)).$$

$$g_1 = bg_2 + a$$

$$b = \frac{\text{Cov}(g_1, g_2)}{\text{Var}(g_2)} = \text{Cov}(g_1, g_2) = \text{Corr}(g_1, g_2) = r,$$

$$E(a) = E(g_1) - bE(g_2) = 0.$$

$$y = (rg_2 + a)\beta + \epsilon = g_2(r\beta) + a\beta + \epsilon = g_2(r\beta) + \epsilon' \quad \text{where } \epsilon' \sim \mathcal{N}(0, \text{Var}(\epsilon')).$$

NOTE: GWAS effect size estimates present **marginal effect sizes**.

[5] Bulik-Sullivan, B., Loh, P.R., Finucane, H. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).

Modern Approaches: SNP-Based Heritability

Method 3: LD Score Regression [5]

From GWAS summary statistics

$$y = \sum_{i=1} g_i \lambda_i + \epsilon \text{ where } \lambda_i \text{ is the } \textit{true effect size}. \quad (\text{Causal Model})$$

$$\widehat{\beta}_j = \text{Cov}(y, g_j) = \text{Cov}\left(\sum_{i=1} g_i \lambda_i + \epsilon, g_j\right) = \sum_{i=1} r_{ij} \lambda_i + \text{Cov}(\epsilon, g_j) = \sum_{i=1} r_{ij} \lambda_i + e_j.$$

Assumptions

$$\begin{aligned} E(\lambda_i) &= 0 & E(r_{ji} \lambda_i) &= 0 & e_j &\sim \mathcal{N}\left(0, \frac{\sigma_e^2}{N}\right) \\ \text{Var}(\lambda_i) &= h^2/M & & & \sigma_e^2 &\approx \text{var}(y) = 1 \\ \text{Cov}(\lambda_i \lambda_j) &= E(\lambda_i \lambda_j) = 0 & & & & \end{aligned}$$

$$\chi_j^2 = \frac{(\widehat{\beta}_j)^2}{\widehat{\text{Var}(e_j)}} \approx N \widehat{\beta}_j^2$$

$$\begin{aligned} \Rightarrow E(\chi_j^2) &\approx N \cdot \left(E(e_j^2) + \sum_i r_{ij}^2 E(\lambda_i^2) \right) = N \cdot \left(\text{Var}(e_j) + \sum_i r_{ij}^2 \text{Var}(\lambda_i) \right) \\ &\approx 1 + \frac{Nh^2}{M} \sum_i r_{ij}^2 = 1 + \frac{Nh^2}{M} \mathcal{L}_j \end{aligned}$$

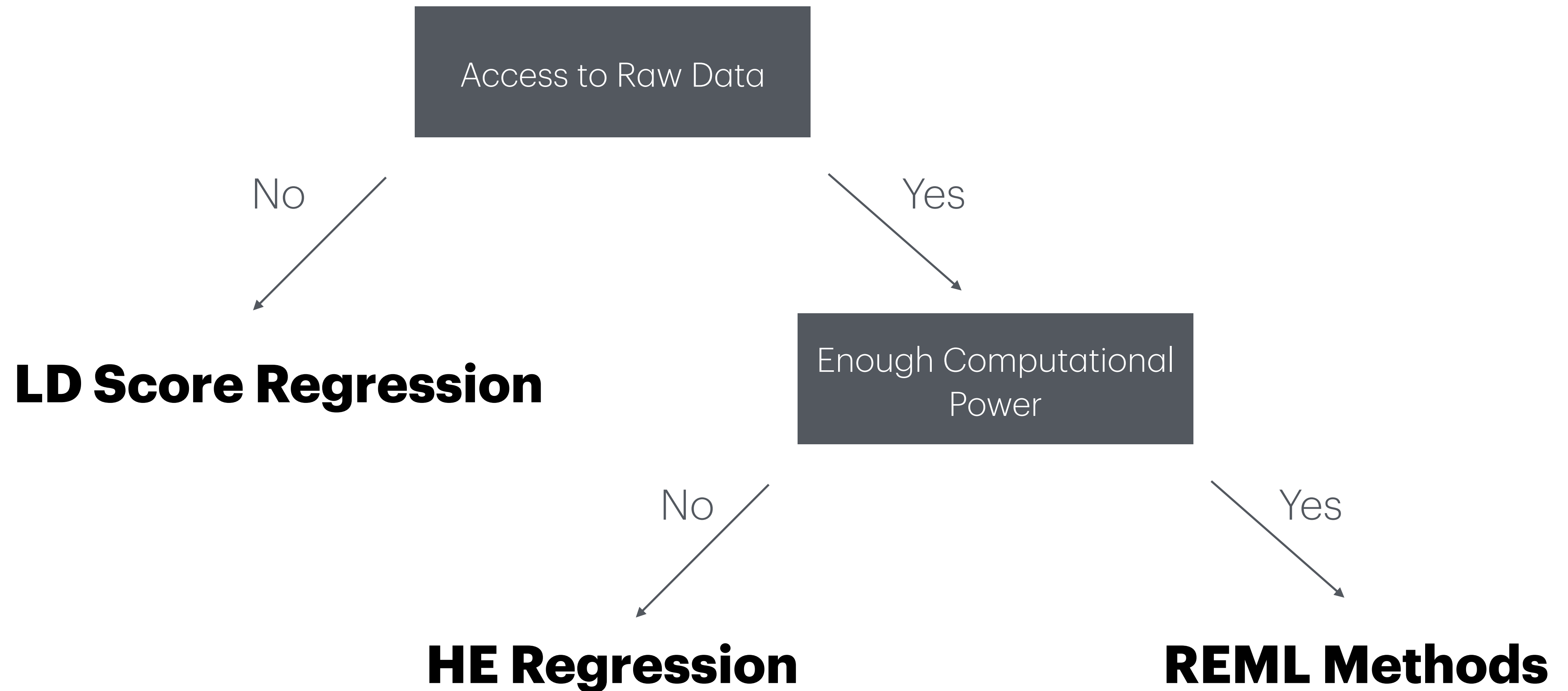
where $\mathcal{L}_j = \sum_i r_{ij}^2$ is referred to as the **LD score of g_j** .

[5] Bulik-Sullivan, B., Loh, P.R., Finucane, H. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).



Modern Approaches: SNP-Based Heritability

Choosing the “*Right*” Method



Challenges in Heritability Studies

Heritability of Diseases [6,7]

LMM Model

$$y_o = \mu \cdot 1 + g_o + e_o$$

$$h_o^2 = \frac{\text{Var}(g_o)}{K(1 - K)}$$

where $K = \text{Prevalence}$

$\text{Var}(g_o)$ and h_o^2 can be estimated by REML.

Liability Model (no ascertainment)

$$y_l = g_l + e_l \text{ where } y_l \sim \mathcal{N}(0,1)$$

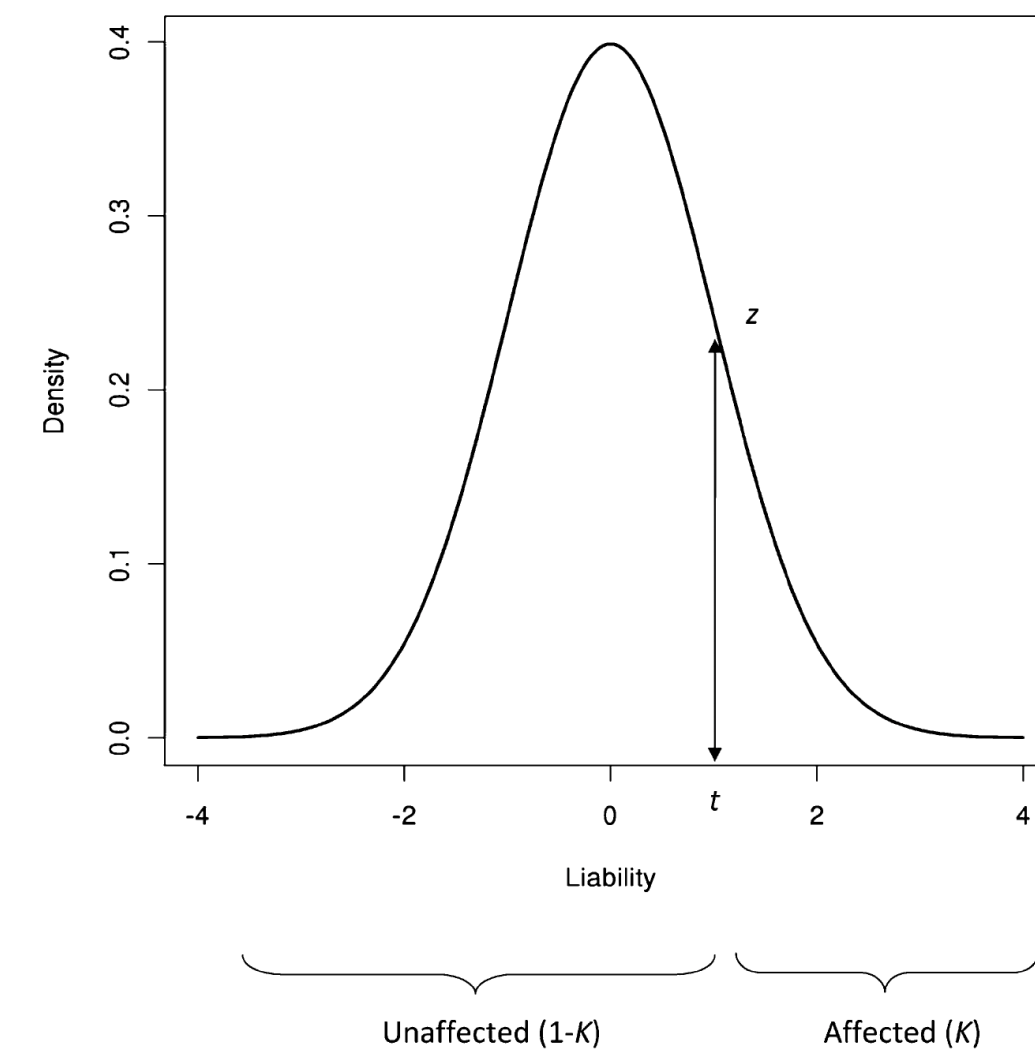


Figure 1. The Liability Threshold Model for a Disease Prevalence of K
An underlying continuous random variable determines disease status. If liability exceeds the threshold t , then individuals are affected.

Figure is from:

Lee, Sang Hong, et al. "Estimating missing heritability for disease from genome-wide association studies." *The American Journal of Human Genetics* 88.3 (2011): 294-305.

$$g_o = a + b \cdot g_l$$

Solve for b : $b \stackrel{[7]}{=} z$

$$h_o^2 = \frac{\text{Var}(g_o)}{K(1 - K)} = \frac{z^2 \text{Var}(g_l)}{K(1 - K)} = \frac{z^2 h_l^2}{K(1 - K)}$$

$$\Rightarrow h_l^2 = h_o^2 \frac{(1 - K)K}{z^2}$$

[6] Dempster, Everett R., and I. Michael Lerner. "Heritability of threshold characters." *Genetics* 35.2 (1950): 212. (APPENDIX BY ALAN ROBERTSON)

[7] Lee, Sang Hong, et al. "Estimating missing heritability for disease from genome-wide association studies." *The American Journal of Human Genetics* 88.3 (2011): 294-305..



Challenges in Heritability Studies

Missing Heritability

$$h^2_{GWAS-Hits} \ll h^2_{Pedigree}$$

(e.g., height 0.1 vs 0.8)

$$h^2_{GWAS-Hits} \overset{[8]}{\ll} h^2_{GWAS-All} < h^2_{Pedigree}$$

Analysis | Published: 20 June 2010

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

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

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



Table 1. Population Variation Explained by GWAS for a Selected Number of Complex Traits			
Trait or Disease	h ² Pedigree Studies	h ² GWAS Hits ^a	h ² All GWAS SNPs ^b
Type 1 diabetes	0.9 ⁹⁸	0.6 ^{99, c}	0.3 ¹²
Type 2 diabetes	0.3–0.6 ¹⁰⁰	0.05–0.10 ³⁴	
Obesity (BMI)	0.4–0.6 ^{101, 102}	0.01–0.02 ³⁶	0.2 ¹⁴
Crohn’s disease	0.6–0.8 ¹⁰³	0.1 ¹¹	0.4 ¹²
Ulcerative colitis	0.5 ¹⁰³	0.05 ¹²	
Multiple sclerosis	0.3–0.8 ¹⁰⁴	0.1 ⁴⁵	
Ankylosing spondylitis	>0.90 ¹⁰⁵	0.2 ¹⁰⁶	
Rheumatoid arthritis	0.6 ¹⁰⁷		
Schizophrenia	0.7–0.8 ¹⁰⁸	0.01 ⁷⁹	0.3 ¹⁰⁹
Bipolar disorder	0.6–0.7 ¹⁰⁸	0.02 ⁷⁹	0.4 ¹²
Breast cancer	0.3 ¹¹⁰	0.08 ¹¹¹	
Von Willebrand factor	0.66–0.75 ^{112, 113}	0.13 ¹¹⁴	0.25 ¹⁴
Height	0.8 ^{115, 116}	0.1 ¹³	0.5 ^{13, 14}
Bone mineral density	0.6–0.8 ¹¹⁷	0.05 ¹¹⁸	
QT interval	0.37–0.60 ^{119, 120}	0.07 ¹²¹	0.2 ¹⁴
HDL cholesterol	0.5 ¹²²	0.1 ⁵⁷	
Platelet count	0.8 ¹²³	0.05–0.1 ⁵⁸	

^a Proportion of phenotypic variance or variance in liability explained by genome-wide-significant and validated SNPs. For a number of diseases, other parameters were reported, and these were converted and approximated to the scale of total variation explained. Blank cells indicate that these parameters have not been reported in the literature.
^b Proportion of phenotypic variance or variance in liability explained when all GWAS SNPs are considered simultaneously. Blank cell indicate that these parameters have not been reported in the literature.
^c Includes pre-GWAS loci with large effects.

Table is from:
Visscher, Peter M., et al. "Five years of GWAS discovery." *The American Journal of Human Genetics* 90.1 (2012): 7-24.

Article | Published: 07 March 2022

Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data

[Pierrick Wainschtein](#) , [Deepti Jain](#), [Zhili Zheng](#), [TOPMed Anthropometry Working Group](#), [NHLBI Trans-Omics for Precision Medicine \(TOPMed\) Consortium](#), [L. Adrienne Cupples](#), [Aladdin H. Shadyab](#), [Barbara McKnight](#), [Benjamin M. Shoemaker](#), [Braxton D. Mitchell](#), [Bruce M. Psaty](#), [Charles Kooperberg](#), [Ching-Ti Liu](#), [Christine M. Albert](#), [Dan Roden](#), [Daniel I. Chasman](#), [Dawood Darbar](#), [Donald M. Lloyd-Jones](#), [Donna K. Arnett](#), [Elizabeth A. Regan](#), [Eric Boerwinkle](#), [Jerome I. Rotter](#), [Jeffrey R. O’Connell](#), [Lisa R. Yanek](#), ... [Peter M. Visscher](#)  [+ Show authors](#)

[Nature Genetics](#) **54**, 263–273 (2022) | [Cite this article](#)

"We found ~ 0.70 ($SE = 0.09$) for **height** and ~ 0.29 ($SE = 0.09$) for **BMI** (Supplementary Fig. 12). The estimates for height are close to the pedigree estimates of **0.7 – 0.8**, whereas this is not the case for BMI at **0.4** and **0.6**, respectively." [9]

[8] Yang, Jian, et al. "Common SNPs explain a large proportion of the heritability for human height." *Nature genetics* 42.7 (2010): 565-569.
[9] Wainschtein, P., Jain, D., Zheng, Z. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet* **54**, 263–273 (2022).

What's Next

1. Introduction to Mendelian Randomisation (MR)
2. Core Logic and Theory of MR
3. Applications and Key Challenges