

Estimation of SNP-heritability

Matthew Keller

Loic Yengo

University of Colorado at Boulder
University of Queensland, Brisbane

Outline

- Motivation
- \hat{p}
- How to estimate h^2_{SNP} - HE-regression example
- Interpretation of h^2_{SNP} (or V_{A_SNP})

Polygenicity in complex traits

- The sum of R^2 of significantly associated SNPs of complex traits typically $< 10\%$, despite twin/family $h^2 \sim .5 \pm .2$. Why?
- One possibility: large number of small-effect (\sim the 'infinitesimal model'; Fisher, 1918) causal variants (CVs) that failed to reach genome-wide significance (many type-II errors)
- Growing consensus: 100s to 1000s of CVs contribute to the genetic variation of traits like schizophrenia, each with small effects ($OR < 1.3$), often in unpredicted loci

Using genetic similarity at SNPs to estimate V_A

- Multiple approaches to derive unbiased estimate of V_A captured by measured (typically common) SNPs (we'll cover 3 today and tomorrow)
- Determine extent to which genetic similarity (\hat{p}) at SNPs is related to phenotypic similarity

pihat

IBD vs. IBS

- IBD – identity by descent – alleles descended from common ancestor
- IBS – identity by state – alleles that look the same but not necessarily from a common ancestor within a given time frame (e.g., since the base population)
- Problem: from coalescent theory, \sim all IBS alleles came from same mutation and are thus IBD (though each IBS allele is IBD from different time in past).
- Reconciliation: IBD estimates should be designed to estimate $P(\text{alleles at unobserved loci are IBS})^*$

$\hat{\pi} = E(\text{IBD})$, usually genome-wide

- $\hat{\pi}$ among close relatives captures long stretches of identical chromosomes, and estimate IBS at both common and rare alleles. Traditionally with close relatives, we know the expectation of this and use this (without variance) for modeling.
- $\hat{\pi}$ among unrelateds (distant relatives) assumes base population is the current sample, and thus its expectation is 0. It is typically measured with SNPs, and so only captures IBS at measured SNPs and unmeasured SNPs in LD with measured SNPs. It can go negative (less related than average).

$\hat{\pi}$ = genome-wide mean correlation of SNP values between a pair of individuals j, k

$$\hat{\pi}_{jk} = \frac{1}{m} \sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}$$

$$\hat{\pi}_{jk} = \frac{1}{m} \sum_i \left(\frac{x_{ij} - 2p_i}{\sqrt{2p_i(1 - p_i)}} \right) \left(\frac{x_{ik} - 2p_i}{\sqrt{2p_i(1 - p_i)}} \right)$$

$$\hat{\pi}_{jk} = \frac{1}{m} \sum_i \left(\frac{x_{ij} - E(x_i)}{S(x_i)} \right) \left(\frac{x_{ik} - E(x_i)}{S(x_i)} \right)$$

$$\hat{\pi}_{jk} = \frac{1}{m} \sum_i (z_{ij})(z_{ik})$$

$$\hat{\pi}_{jk} = \frac{1}{m} \sum_i \text{cor}(x_{ij}, x_{ik})$$

H-E REGRESSION

Regression estimates of h^2

$\theta_{ij} = Z_i Z_j$ ← product of centered scores
(here, z-scores)

$$E[\theta_{ij}] = COV(Z_i, Z_j)$$

$$E[\theta_{ij} | \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is
an estimate of h^2)

Regression estimates of h^2

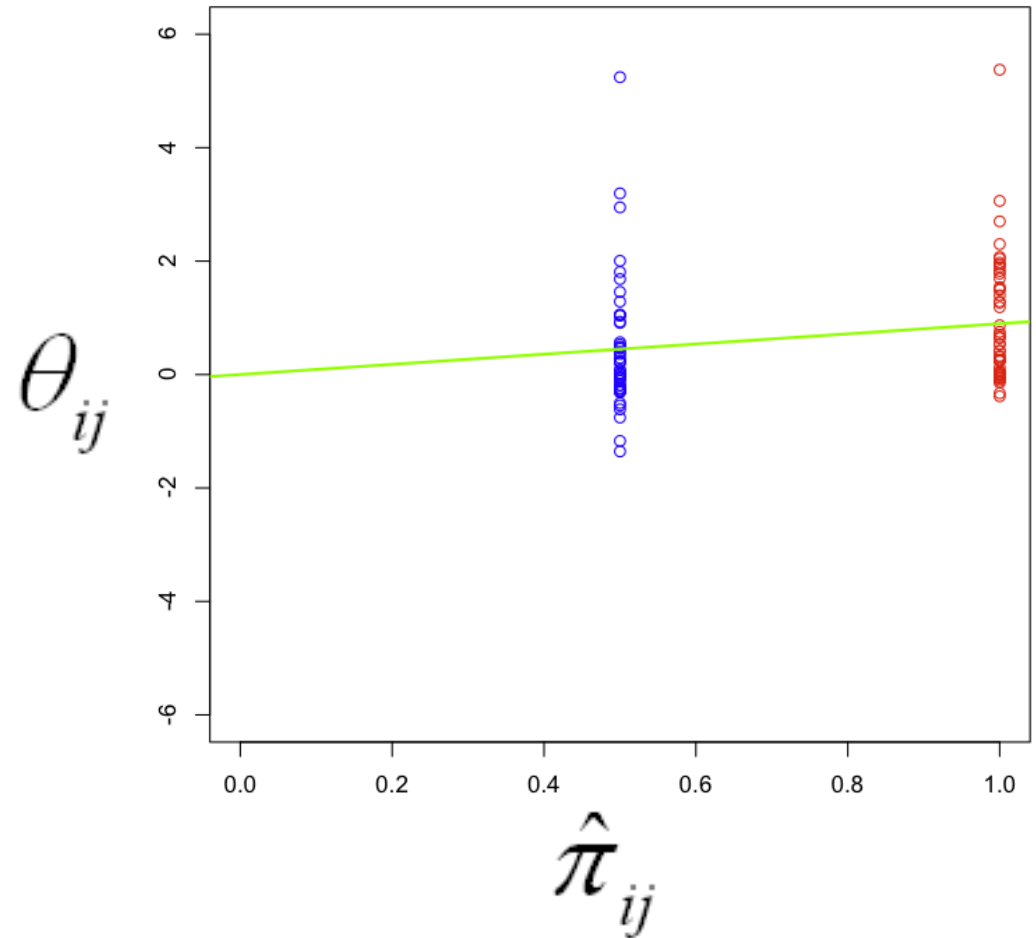
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = \text{COV}(Z_i, Z_j)$$

$$E[\theta_{ij} | \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Regression estimates of h^2

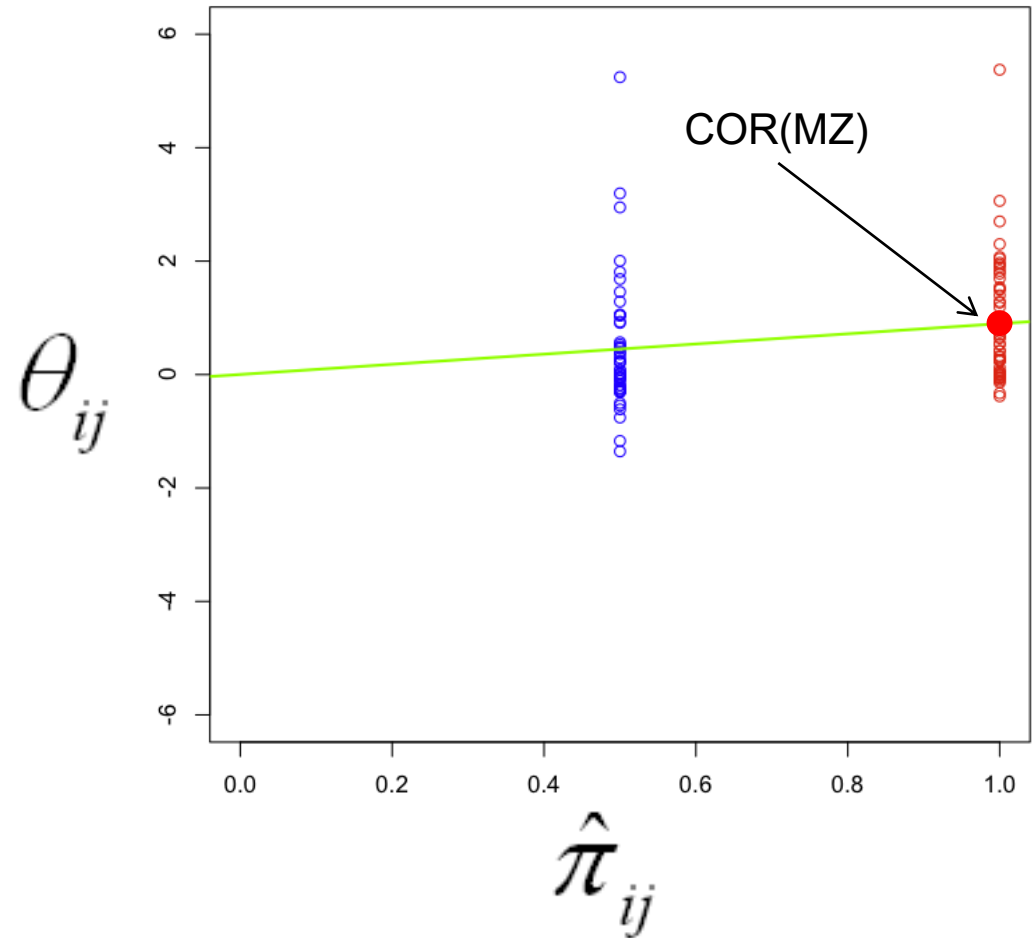
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = COV(Z_i, Z_j)$$

$$E[\theta_{ij} | \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Regression estimates of h^2

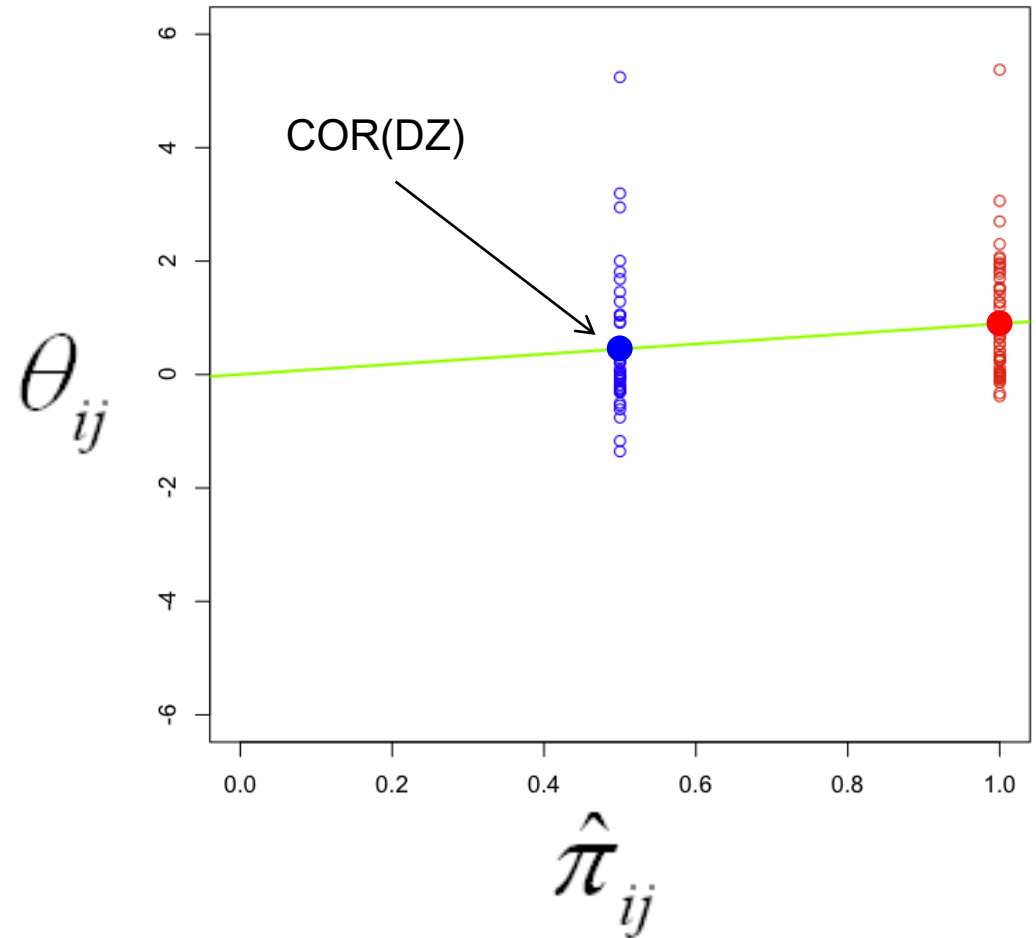
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = \text{COV}(Z_i, Z_j)$$

$$E[\theta_{ij} | \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Regression estimates of h^2

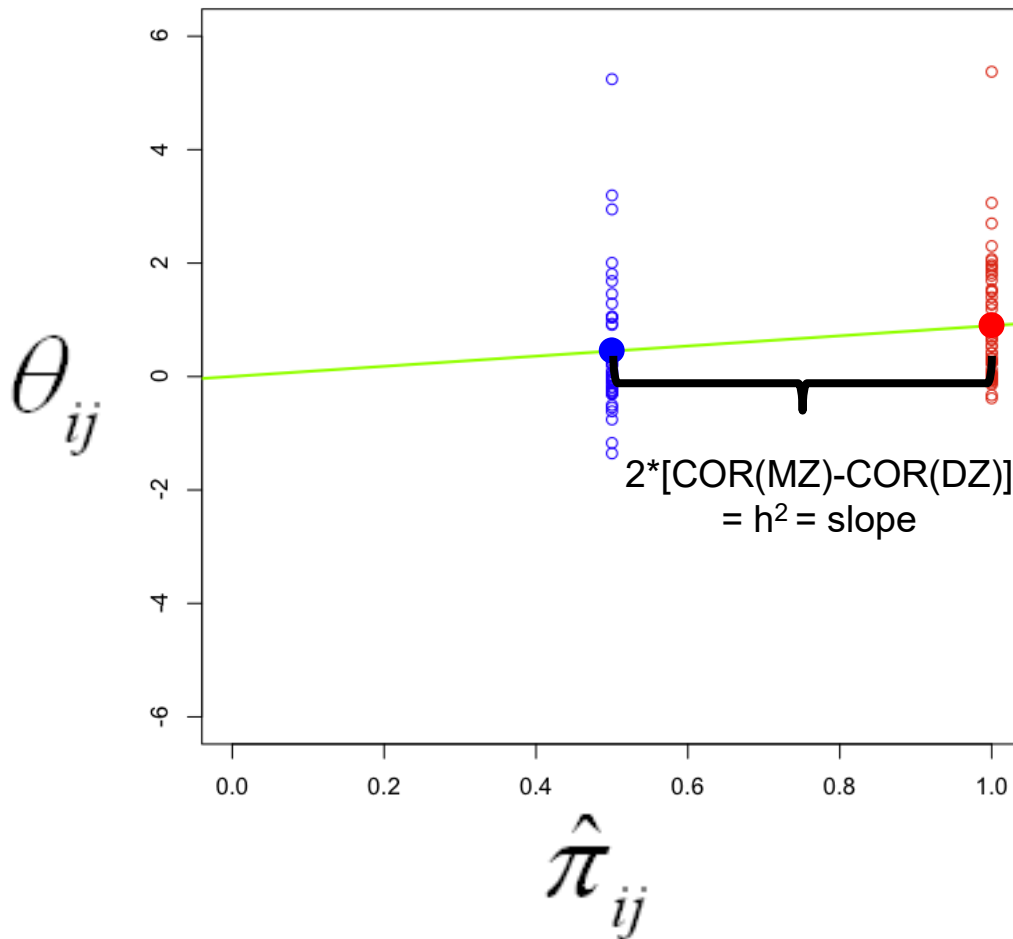
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = COV(Z_i, Z_j)$$

$$E[\theta_{ij} | \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Regression estimates of h^2

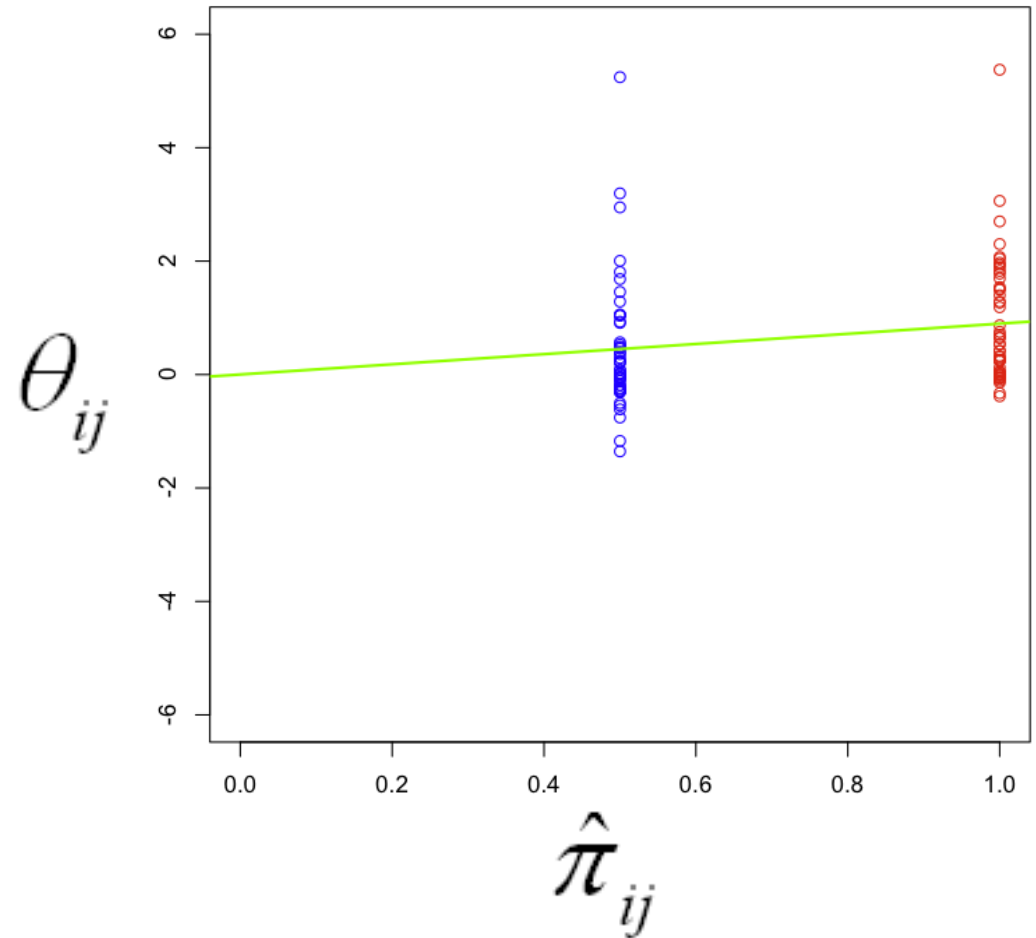
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = \text{COV}(Z_i, Z_j)$$

$$E[\theta_{ij} | \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Regression estimates of h^2

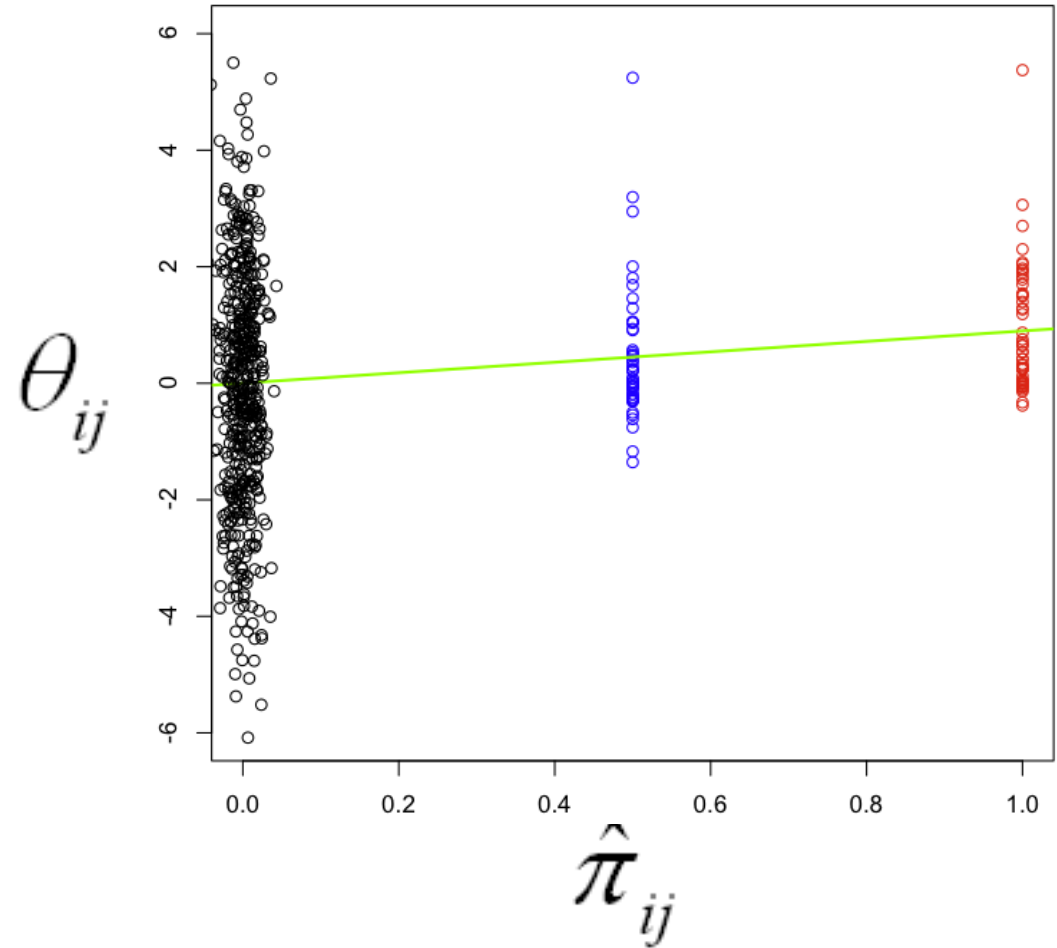
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = \text{COV}(Z_i, Z_j)$$

$$E[\theta_{ij} | \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2$$

(the slope of the regression is an estimate of h^2)



Regression estimates of h^2_{snp}

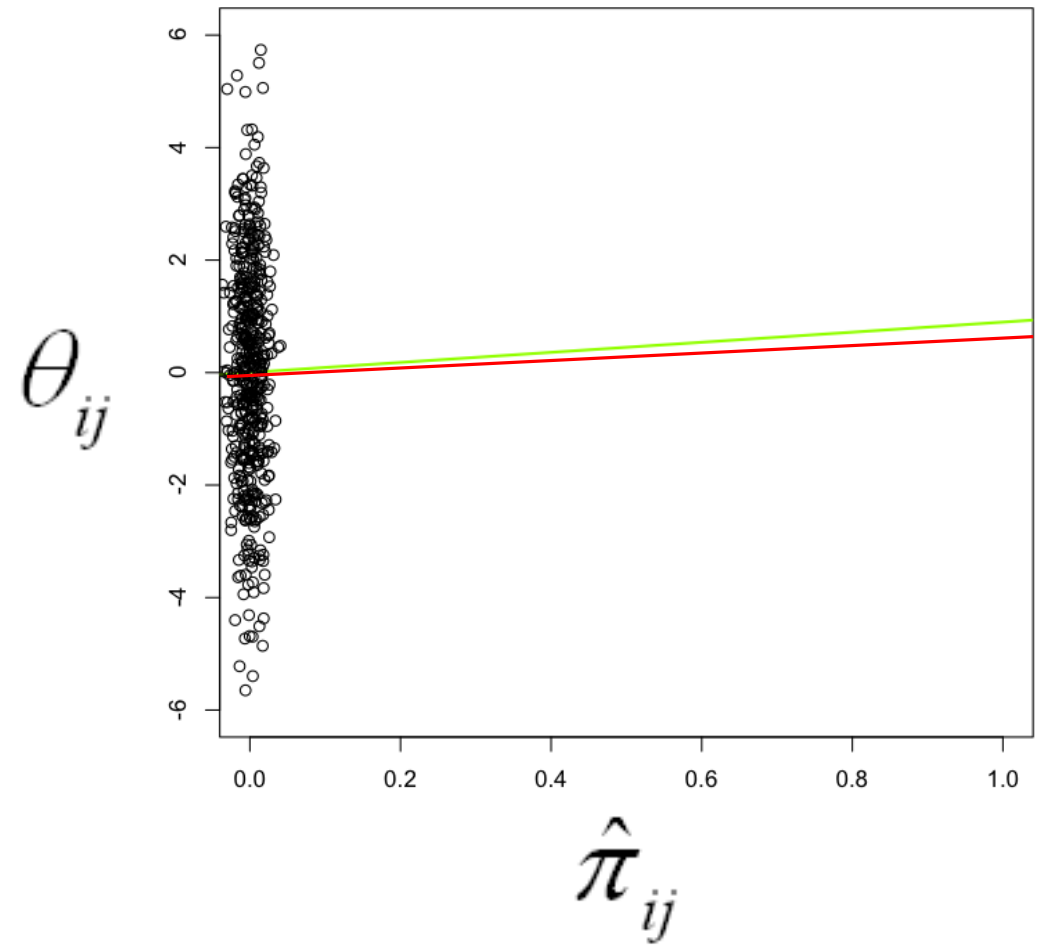
$$\theta_{ij} = Z_i Z_j$$

$$E[\theta_{ij}] = \text{COV}(Z_i, Z_j)$$

$$E[\theta_{ij} | \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\hat{\beta}_1 = \hat{h}^2_{\text{snp}}$$

(the slope of the regression is an estimate of h^2_{snp})



Interpreting h^2 estimated from SNPs (h^2_{snp})

- If close relatives included (e.g., sibs), $h^2_{\text{snp}} \cong h^2$ estimated from a family-based method, because of great influence of extreme pihats. Interpret h^2_{snp} as from these designs.
- If use 'unrelateds' (e.g., $p_{\text{ihat}} < .05$):
 - h^2 estimate 'uncontaminated' by shared environment and non-additive genetic effects
 - Does not rely on family/twin study assumptions
 - Evidence for h^2_{snp} to degree similarity at SNPs corresponds to phenotypic similarity. Thus, $h^2_{\text{snp}} =$ proportion of V_P due to CVs tagged by (in LD with) SNPs used in the GRM.
 - Typically, $h^2_{\text{snp}} < h^2$. It is the max r^2 possible from a PRS using those SNPs.

Why $h^2_{\text{snp}} < h^2$ (usually)

- Because we only estimate genetic variance from CVs in LD with the SNPs used in the analysis. Common CVs are in high LD with array/imputed SNPs, but this is less the case with rare CVs.
- In particular*:

$$\hat{h}^2_{\text{snp}} \cong h^2 \frac{\bar{r}^2_{MQ}}{\bar{r}^2_{MM}}$$

where

\bar{r}^2_{MQ} is the average r^2 between CVs and SNPs

\bar{r}^2_{MM} is the average r^2 between SNPs and SNPs

Big picture: Using SNPs to estimate h^2

- Independent approach to estimating h^2
 - Different assumptions than family models. Increasingly tortuous reasoning to suggest traits aren't heritable because methodological flaws
- When using SNPs with same allele frequency distribution as CVs, provides unbiased estimate of h^2
- When using common (array) SNPs to estimate relatedness, generally provides downwardly biased estimate of h^2
 - “Still missing” h^2 ($h^2_{\text{family}} - h^2_{\text{snp}}$) provides insight into the importance of rare variants, non-additive, or biased h^2_{family} .
- But not a panacea. Biases still exist. Issues need to be worked out (e.g., assortative mating, etc.).

Time permitting: QC & various
ways to estimate h^2_{SNP}

SNP QC

- Poor SNP calls can inflate SE and cause downward bias in h^2_{snp}
- Clean data for
 - SNPs missing $> \sim .05$
 - HWE $p < 10e-6$
 - MAF $< \sim .01$
 - Plate effects:
 - Remove plates with extreme average inbreeding coefficients or high average missingness


Individual QC

- Remove individuals missing $> \sim .02$
- Remove close relatives (e.g., `--grm-cutoff 0.05`)
 - Correlation between pi-hats and shared environment can inflate h^2_{snp} estimates
- Control for stratification (usually 5 to 20 PCs)
 - Different prevalence rates (or ascertainment) between populations can show up as h^2_{snp}
- Control for plates and other technical artifacts
 - Be careful if cases & controls are not randomly placed on plates (can create upward bias in h^2_{snp})



Comparison of approaches for estimating h^2_{snp}

APPROACH (METHOD)	ADVANTAGES	DISADVANTAGES
HE-regression	Fast. Point estimates usually unbiased	Large SEs (~30% larger than REML). SE estimates biased. Limited model building.



Comparison of approaches for estimating h^2_{snp}

APPROACH (METHOD)	ADVANTAGES	DISADVANTAGES
HE-regression	Fast. Point estimates usually unbiased	Large SEs (~30% larger than REML). SE estimates biased. Limited model building.
GREML (e.g., GCTA) 	Point estimates & SEs usually unbiased. Well maintained & easy to use.	Limited model-building.




Comparison of approaches for estimating h^2_{snp}

APPROACH (METHOD)	ADVANTAGES	DISADVANTAGES
HE-regression	Fast. Point estimates usually unbiased	Large SEs (~30% larger than REML). SE estimates biased. Limited model building.
GREML (e.g., GCTA) 	Point estimates & SEs usually unbiased. Well maintained & easy to use.	Limited model-building.
LD-score regression 	Requires only summary statistics; mostly robust to stratification/relatedness	Limited model building. Does not give good estimates of variance due to rare CVs

Comparison of approaches for estimating h^2_{snp}

APPROACH (METHOD)	ADVANTAGES	DISADVANTAGES
HE-regression	Fast. Point estimates usually unbiased	Large SEs (~30% larger than REML). SE estimates biased. Limited model building.
GREML (e.g., GCTA) 	Point estimates & SEs usually unbiased. Well maintained & easy to use.	Limited model-building.
LD-score regression 	Requires only summary statistics; mostly robust to stratification/relatedness	Limited model building. Does not give good estimates of variance due to rare CVs

Comparison of approaches for estimating h^2_{snp}

APPROACH (METHOD)	ADVANTAGES	DISADVANTAGES
HE-regression	Fast. Point estimates usually unbiased	Large SEs (~30% larger than REML). SE estimates biased. Limited model building.
GREML (e.g., GCTA) 	Point estimates & SEs usually unbiased. Well maintained & easy to use.	Limited model-building (e.g., no nonlinear constraints).
LD-score regression 	Requires only summary statistics; mostly robust to stratification/relatedness	Limited model building. Does not give good estimates of variance due to rare CVs
GSEM 	Flexible. Ability to build complex models.	Uses summary output from LDSC or GREML