Assumptions in using SNPs to estimate trait heritability among ‘unrelated’ individuals

Matthew Keller

Teresa de Candia

University of Colorado at Boulder

Outline

- Overview of estimating genetic variance tagged by SNPs
  - how it works: HE-regression example
  - how to interpret SNP h^2
  - assumptions using GREML approach
  - assortative mating & GREML/HE biases

Regression estimates of V_A

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

product of centered scores (here, z-scores)

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

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

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

(the slope of the regression is an estimate of standardized V_A (i.e., h^2)

Regression estimates of V_A

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

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

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

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

(the slope of the regression is an estimate of standardized V_A (i.e., h^2)

average correlation between 2 genotypes across ALL MEASURED SNPS:

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

Regression estimates of V_A

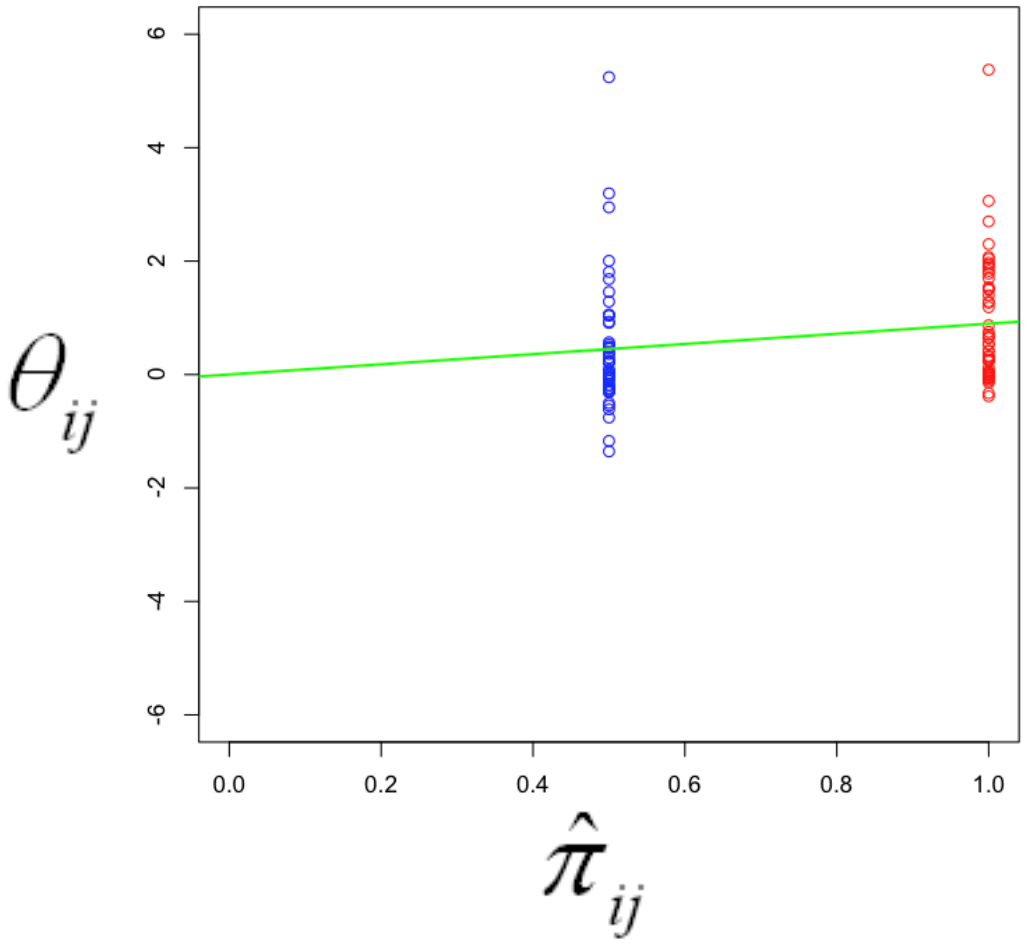
$$\theta_{ij} = Z_i Z_j$$

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

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

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

(the slope of the regression is an estimate of standardized V_A (i.e., h^2)



Regression estimates of V_A

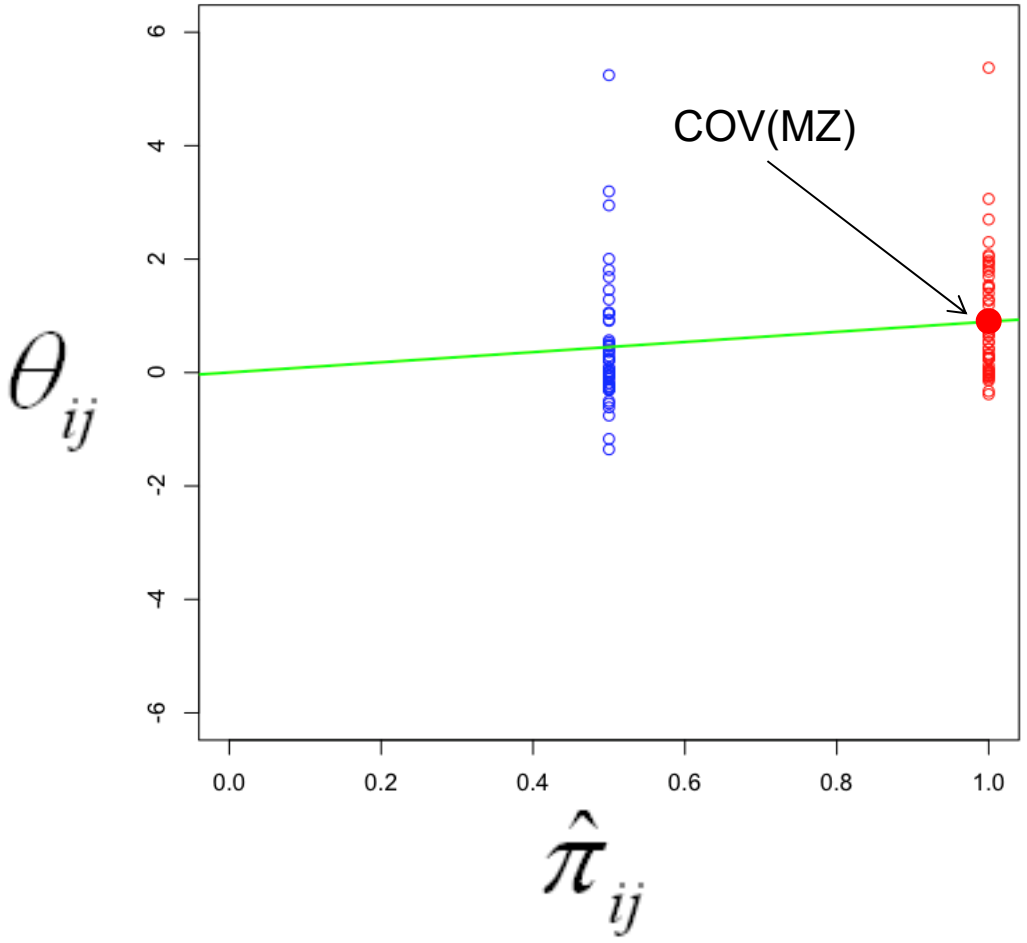
$$\theta_{ij} = Z_i Z_j$$

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

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

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

(the slope of the regression is an estimate of standardized V_A (i.e., h^2)



Regression estimates of V_A

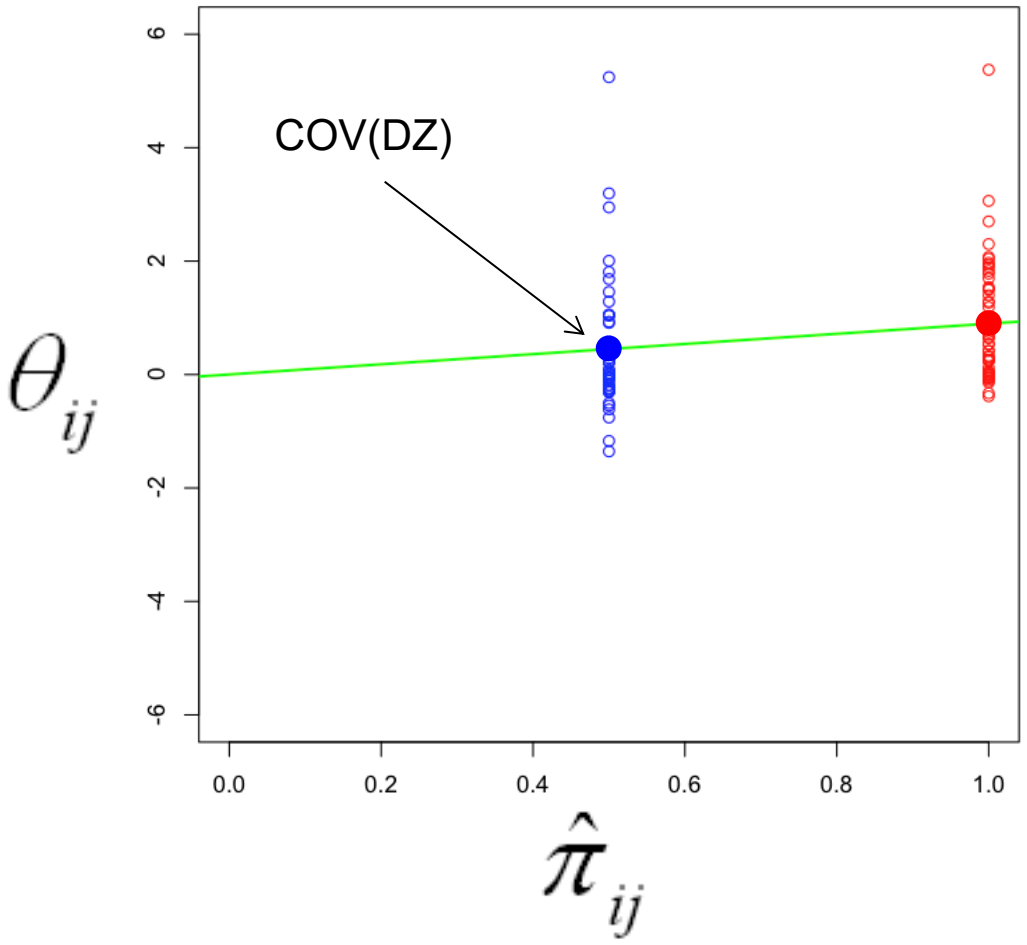
$$\theta_{ij} = Z_i Z_j$$

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

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

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

(the slope of the regression is an estimate of standardized V_A (i.e., h^2)



Regression estimates of V_A

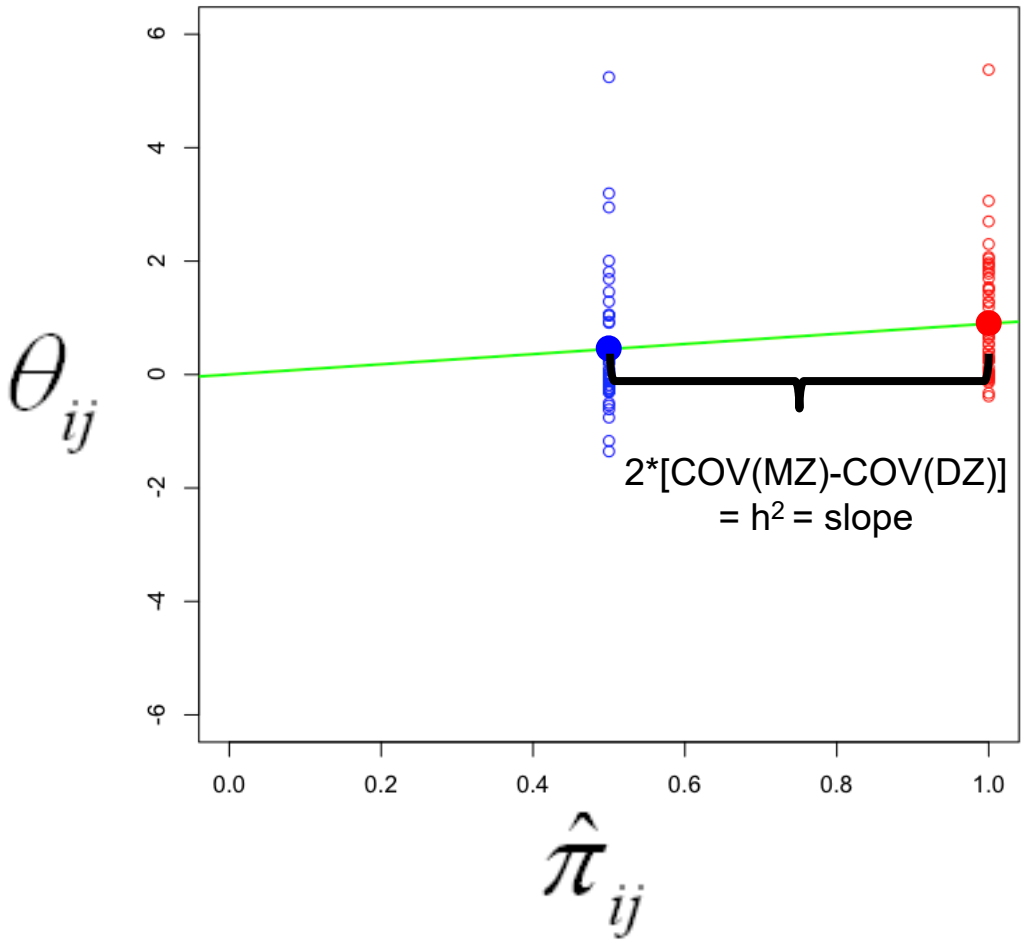
$$\theta_{ij} = Z_i Z_j$$

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

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

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

(the slope of the regression is an estimate of standardized V_A (i.e., h^2)



Regression estimates of V_A

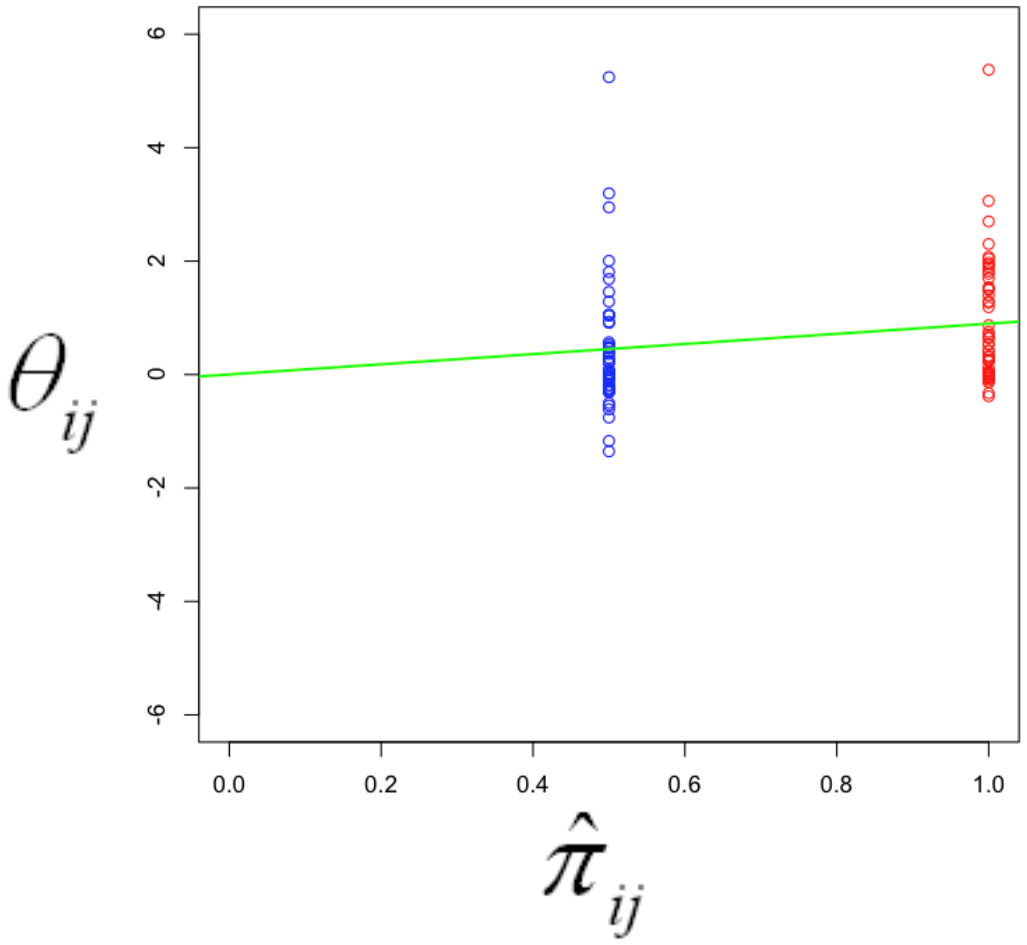
$$\theta_{ij} = Z_i Z_j$$

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

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

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

(the slope of the regression is an estimate of standardized V_A (i.e., h^2)



Regression estimates of V_A

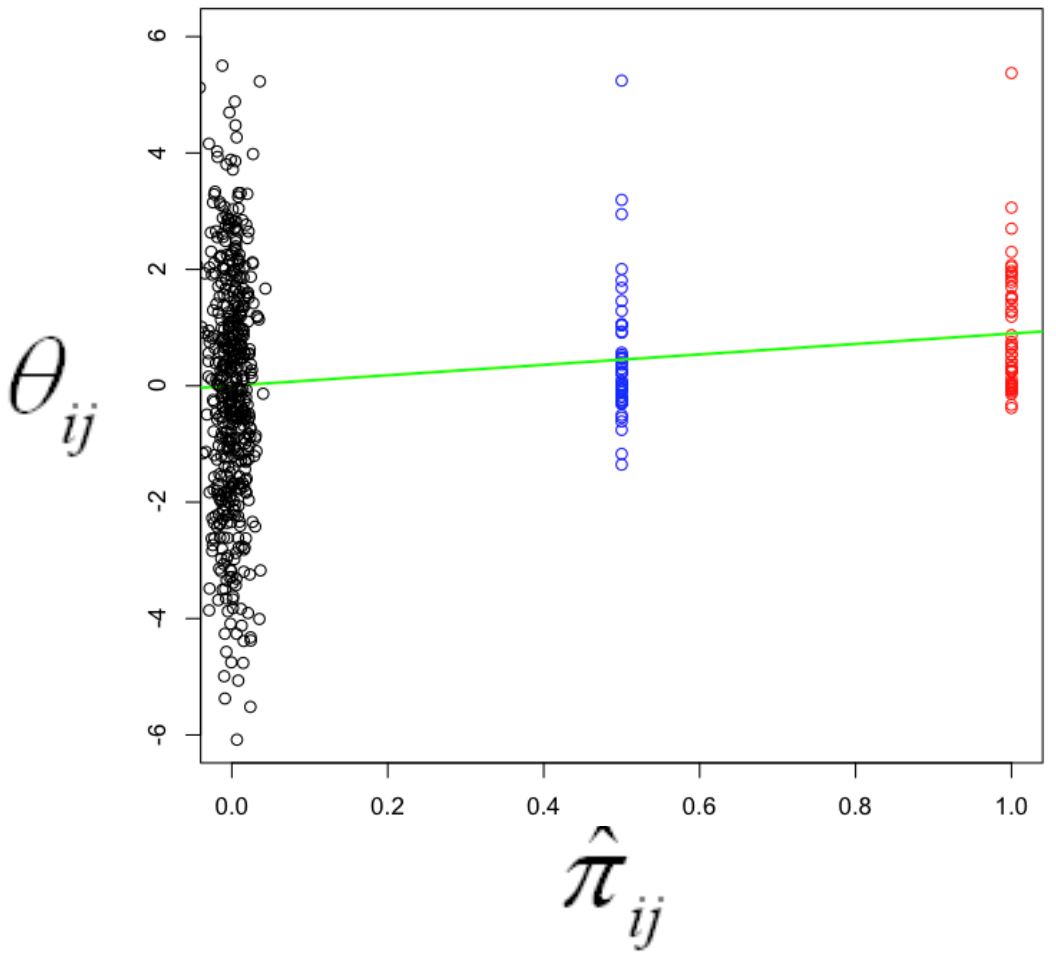
$$\theta_{ij} = Z_i Z_j$$

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

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

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

(the slope of the regression is an estimate of standardized V_A (i.e., h^2)



Regression estimates of V_{A_snp}

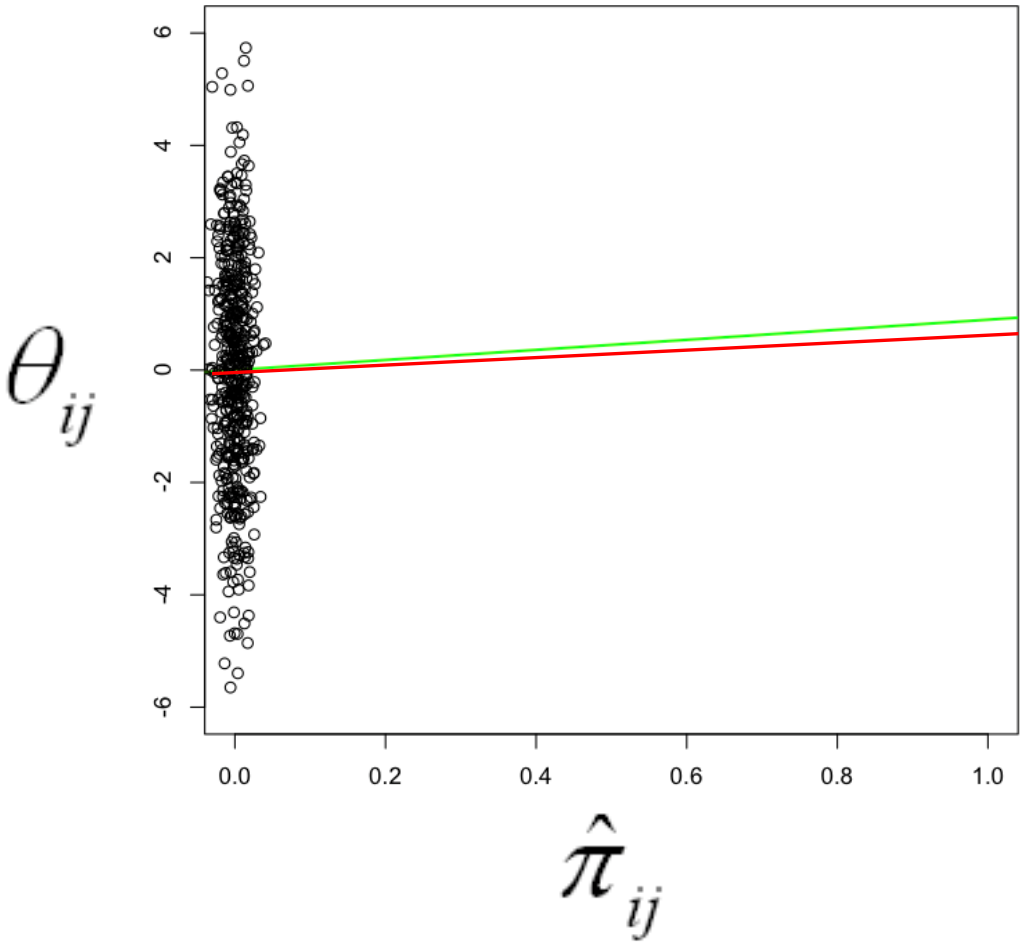
$$\theta_{ij} = Z_i Z_j$$

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

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

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

(the slope of the regression is an estimate of standardized V_{A_SNP} (i.e., $h^2_{\_SNP}$)



But how to interpret this "SNP" h^2 (h^2_{snp})

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

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

average correlation between 2 genotypes across ALL (k=1…m) MEASURED SNPS:

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

To estimate total narrow h^2, we'd need the average correlation between 2 genotypes across ALL (k=1…q) CAUSAL VARIANTS:

$$\hat{\pi}^*_{ij} = \frac{1}{q} \sum_k \frac{(x^*_{ik} - 2p_k)(x^*_{jk} - 2p_k)}{2p_k(1-p_k)}$$

But how to interpret this "SNP" h^2 (h^2_{snp})

$$E[\theta_{ij} \mid \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij} \longrightarrow \hat{\beta}_1 = \hat{h}^2_{SNP}$$

$$E[\theta_{ij} \mid \hat{\pi}^*_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}^*_{ij} \longrightarrow \hat{\beta}_1 = \hat{h}^2$$

To the degree $\hat{\pi}^*_{ij}$ isn't predicted well by $\hat{\pi}_{ij}$, $\hat{h}^2_{SNP} < \hat{h}^2$

WHEN is $\hat{\pi}^*_{ij}$ not well predicted by $\hat{\pi}_{ij}$? This is actually the usual case, and it is not a bad thing! It happens to the degree that (typically unmeasured) causal variants aren't well tagged by (in high LD with) measured SNPs E.g., rare causal variants.

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

- If close relatives included (e.g., sibs), $h^2_{snp} \cong h^2$ estimated from a family-based method. Thus, interpret h^2_{snp} as you would h^2 from AE model (including all biases!).
- If use only distant 'relatives' (the usual approach; e.g., pihat< .05):
  - h^2_{snp} = proportion of V_P due to V_A captured by common SNPs. Upper bound of % V_P GWAS can detect
  - Gives idea of the aggregate importance of causal variants tagged by SNPs (mostly common ones b/c rare ones poorly tagged by common SNPs)
  - By not using relatives who also share environmental effects: (a) V_A estimate 'uncontaminated' by V_C & $V_{NA;}$ (b) does not rely on assumptions required in family studies (e.g., r(MZ) > r(DZ) for only genetic reasons)

Big picture: using SNPs to estimate h^2

- It is an independent approach to estimating heritability
  - Its assumptions are different from family-based models. It takes increasingly tortuous reasoning to suggest trait X isn't heritable because methodological flaws in estimating h^2
- It provides a downwardly biased estimate of h^2 – and this is *good!*
  - h^2_{snp} helps elucidate the genetic architecture of complex traits
  - The "still missing" h2 ($h^2_{family} - h^2_{snp}$) provides insight into the importance of rare variants, non-additive genetics, or over-estimation of h^2_{twin}.
- But it is *not* a panacea!
  - Many issues still need to be worked out. As with twin designs, we can be misled under certain scenarios (e.g., assortative mating, maternal effects, etc.). Is an active area of research.

Unimportant assumptions in estimating h^2_{snp}

See excellent paper by Speed et al, 2012, who investigated the following and found that these violations had little effect on estimates:

1. Genetic and error effects are normally distributed
2. effect sizes of causal variants is proportionate to 1/2pq (low MAF variants have bigger effects)
3. All SNPs have an association with the phenotype

Important assumptions in estimating h^2_{snp}

1. There is no correlation between environmental similarity and SNP_pihats. A common way this can occur is in stratified samples:
   - E.g., estimating h^2_{snp} of diabetes in a sample that incudes Native Americans and Caucasians would get a serious inflation in h^2_{snp}. This be dealt with by controlling for principal components of GRM
   - Another way: cases and controls genotyped separately
2. The avg. LD between causal variants (CVs) & SNPs is the same as the average LD between SNPs and other SNPs.
   - If E[r(CV,SNP)] > E[r(SNP,SNP)], you overestimate h^2_{snp} (and vice-versa)
3. The average LD between CVs is same as between SNPs.
   - If E[r(CV,CV)] > E[r(SNP,SNP)], you overestimate h^2_{snp} (and vice-versa).

Important assumptions in estimating h^2_{snp}

1. There is no correlation between environmental similarity and SNP_pihats. A common way this can occur is in stratified samples:
   - E.g., estimating h^2_{snp} of diabetes in a sample that incudes Native Americans and Caucasians would get a serious inflation in h^2_{snp}. This be dealt with by controlling for principal components of GRM
   - Another way: cases and controls genotyped separately
2. The avg. LD between causal variants (CVs) & SNPs is the same as the average LD between SNPs and other SNPs.
   - If E[r(CV,SNP)] > E[r(SNP,SNP)], you overestimate h^2_{snp} (and vice-versa)
3. The average LD between CVs is same as between SNPs.
   - If E[r(CV,CV)] > E[r(SNP,SNP)], you overestimate h^2_{snp} (and vice-versa).

Assortative mating & estimating h^2_{snp}

- As discussed Thursday, AM biases estimates of V_A down and V_C up in twin studies. What kind of effect does it have on SNP-heritability estimates?
- Two years ago at this workshop, there was an interesting debate (and bet – no-one won) about effects AM would have on GCTA estimates.
- AM leads to very long-range LD between CVs themselves, but not between other parts of the genome (SNPs that don't tag CVs)
- After much work by de Candia, Eaves, Carey, Evans, and myself:
  - AM leads to an <u>upward</u> bias in estimates of equilibrium h^2_{snp}
  - This occurs because AM creates covariances between CVs and these are correctly reflected in phenotypic covariances between individuals but poorly reflected in pihat matrix. Thus, variance of pihats is underestimated. Underestimating variance in a predictor leads to overestimates of the coefficients associated with that predictor.

Background and Previous Results

- Increases genetic variance in population by creating correlations among increasing alleles within individuals.
- Previously we had simulated very simplistic "genomes" of multivariate normally distributed causal variants.
- We analytically predicted and observed that, under AM, HE regression produces overestimates of equilibrium genetic variance
- We also observed, but did not analytically prove, that REML estimates are unstable across increasing sample size (decreasing) & increasing number of markers (increasing)

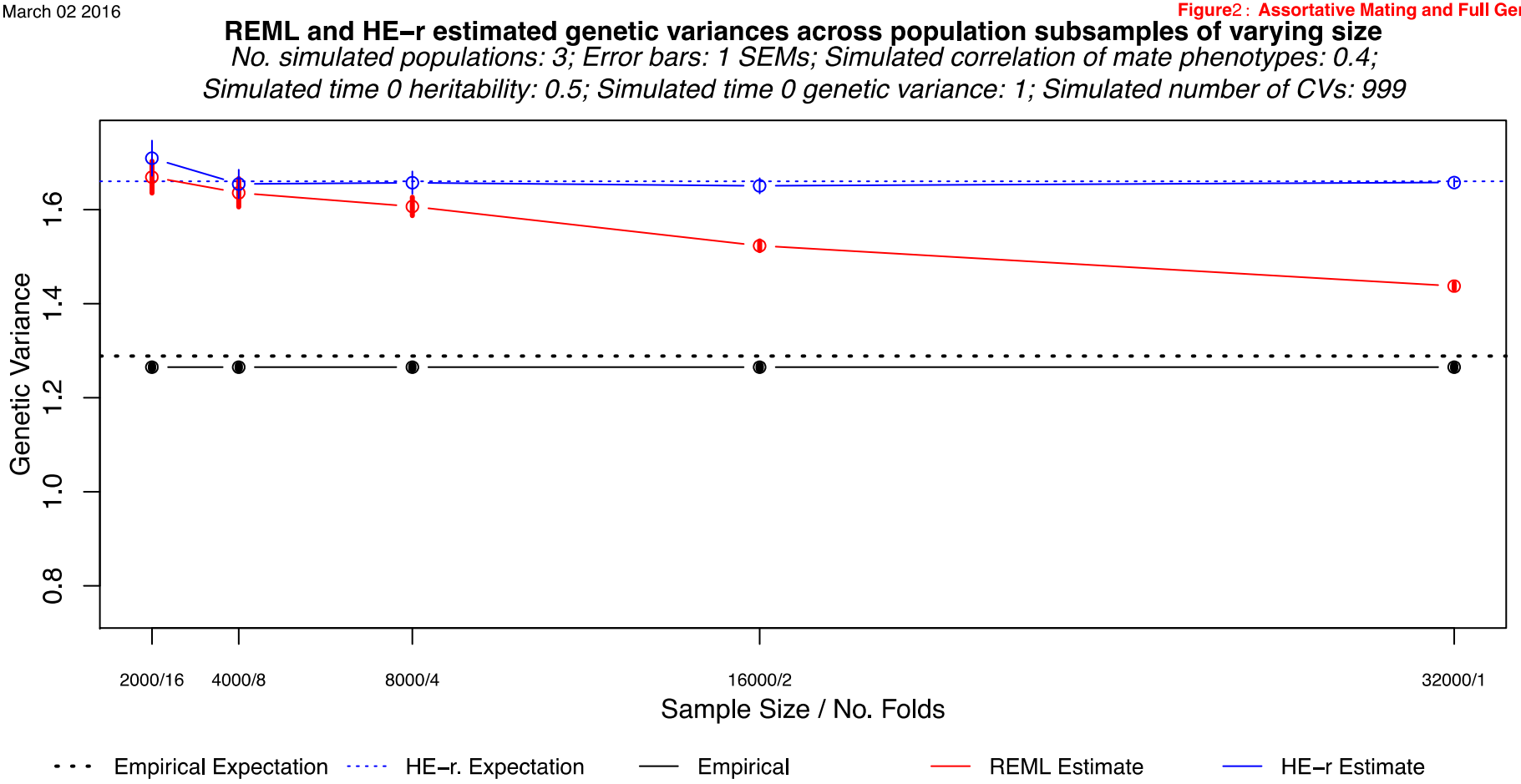
Current Results - Simulations

- Simulated populations over 20 generations of mating under varying levels of AM for a single trait using GeneEvolve (Rasool Tahmasbi).
- No. populations: 3
- CVs: 1000
- Heritability: 0.5
- Relative pruning: >.05
- Spousal phenotypic correlation: .4

Current Results - Simulations

March 02 2016

Figure2 : **Assortative Mating and Full Genome**



Current Results – Height, Raw Phenotypic Variance: 85

March 07 2016

Figure1 : Unresidualized Height

REML and HE–r estimated genetic variances in Unresidualized Height across population subsamples of varying size

Error bars: 1 SEMs; GRM pruned for relatedness > .05