

Using SNPs to estimate heritability among 'unrelated' individuals

Matthew Keller

University of Colorado at Boulder

Motivating question:

What is the genetic architecture
of complex human traits?

What is Genetic Architecture?

- Overall importance of genetic variation to trait variation (heritability, or h^2)
- The degree of genetic overlap (pleiotropy) with other traits
- An understanding of the properties of causal variants (CVs) that underlie trait h^2 :
 - Their number
 - The distribution of their effect sizes
 - The distribution of their allele freq's (not just MAF)
 - Their mode of action (degree of non-additivity)
 - The relative importance of different annotations

Why care about genetic architecture?

- Guides expectations (e.g., effect sizes & sample sizes required)
- Guides the design of future studies & investment (e.g., sequencing vs. imputation)
- Provides insight into biological pathways relevant to traits (e.g., genes expressed in fetal brain for SZ)
- Helps with nosology and better trait definitions (e.g., high r_g between SZ and BPD)
- Provides insight into evolutionary questions (e.g., purifying selection vs. neutral drift)
- Basic understanding (“Because it’s there” - George Mallory)

Insight into genetic architecture using twin/family studies

- For most of history of behavioral genetics, we could only get at genetic architecture using similarity between twins/family members
- Relatives are more similar than random pairs
- Identical twins are more similar than fraternal twins
 - ACE models: $VA' = 2(CV_{mz} - CV_{dz})$; $VC' = 2CV_{dz} - CV_{mz}$
 - ADE models: $VA' = 4CV_{dz} - CV_{mz}$; $VD' = 2CV_{mz} - 4CV_{dz}$
- Average meta-analysis* $h^2 = .49$, and 69% of twin studies support a purely additive model
- However, estimation of h^2 provides only a small part of what we want to know about genetic architecture.
- Moreover, models that use close relatives require strong assumptions about the causes of similarity within relative pairs. To the degree these assumptions are violated, estimates are systematically biased.

What are the effects of violations of assumptions in twin studies?

- a) Only genetic factors cause MZ twins to be more similar to each other than DZ twins: VA and VD overestimated and VC underestimated
- b) Either VD or VC is zero: VA overestimated and VD & VC underestimated
- c) No epistasis: VD or VA overestimated and VC underestimated
- d) No assortative mating: VA and VD underestimated and VC overestimated
- e) No gene-environment interactions or correlations: $A \times C$: VA overestimated; $A \times E$: VE overestimated; passive $Cov(A,C)$: VC overestimated

An alternative: using similarity at SNPs to estimate h^2_{snp}

- h^2_{snp} - extent to which SNP similarity (\hat{p}) corresponds to phenotypic similarity between “unrelated” individuals
- Several advantages for insight into genetic architecture:
 - Only captures h^2 from CVs in LD with SNPs used in the analysis. This allows for estimates of allelic spectra (distribution of CV MAF)
 - h^2_{snp} should be unbiased by environmental factors that increase close relative similarity. As we continue to capture lower MAF SNPs through imputation or sequencing, estimates of h^2_{snp} approach full narrow-sense h^2 .
 - Can estimate r_g between low prevalence disorders that are impractical to estimate using twins/family designs
 - Estimating h^2_{snp} from binned SNPs allows for estimates of relative importance of different SNP annotations.
 - Certain approaches (e.g., Bayes-R) allow for estimates of # CVs and their effect size distribution

Multiple approaches to estimating h^2_{snp}

- Least Squares Regression (Haseman-Elston)
- Mixed effects models (GREML):
 - Typical approach (GCTA assumptions)
 - Alternative assumptions (LDAK)
 - Multi-GRM approaches
- LD-score regression
- Bayesian approaches

Multiple approaches to estimating h^2_{snp}

- Least Squares Regression (Haseman-Elston)
- Mixed effects models (GREML):
 - Typical approach (GCTA assumptions)
 - Alternative assumptions (LDAK)
 - Multi-GRM approaches
- LD-score regression
- Bayesian approaches

Goals today

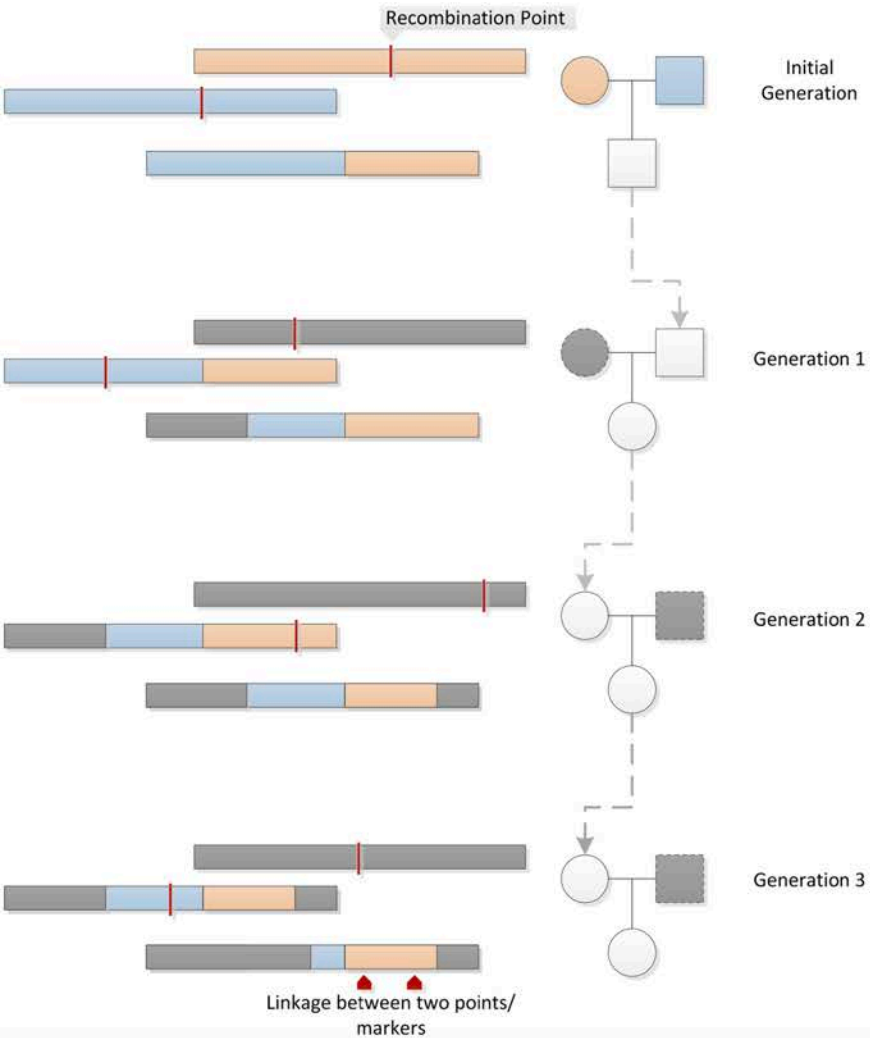
- Understand various ways to estimate h^2_{snp}
- Compare the performance of these procedures to estimate h^2_{snp}
- Understand the factors that bias estimates of h^2_{snp}
- We focus on univariate estimates (VG_{snp} or h^2_{snp}). In general (exceptions noted), the same factors influence bivariate estimates (COV_{snp} or r_{snp}).

LD

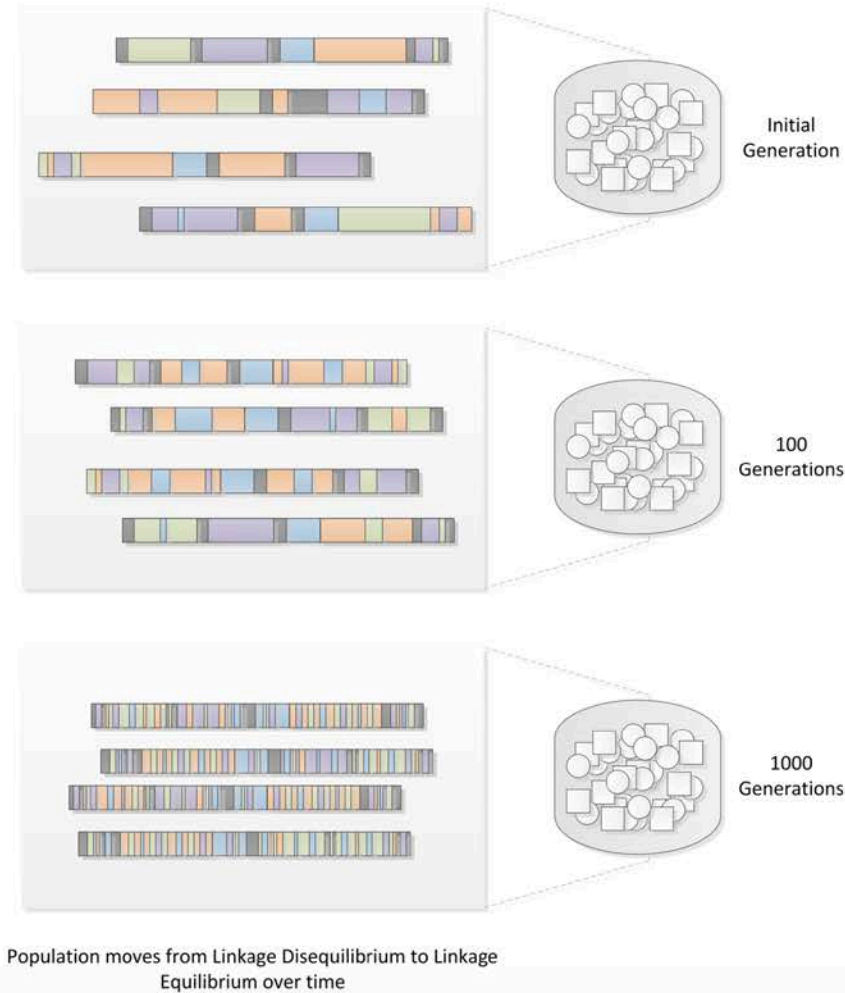
Linkage disequilibrium (LD)

- Statistical association (e.g., r^2) between two SNPs
- Typically arises from a mutation that occurs on a haplotype. It will co-segregate with nearby SNPs. As it rises in frequency, so too will nearby SNPs.
- It decays as a function of number of recombination events that break the two SNPs apart, which is itself a function of:
 - Time (# generations) since the mutational event
 - Distance (cM) between the two SNPs
- SNPs can only predict SNPs that are similar in MAF. Rare-rare or common-common. Rare-common is not possible.

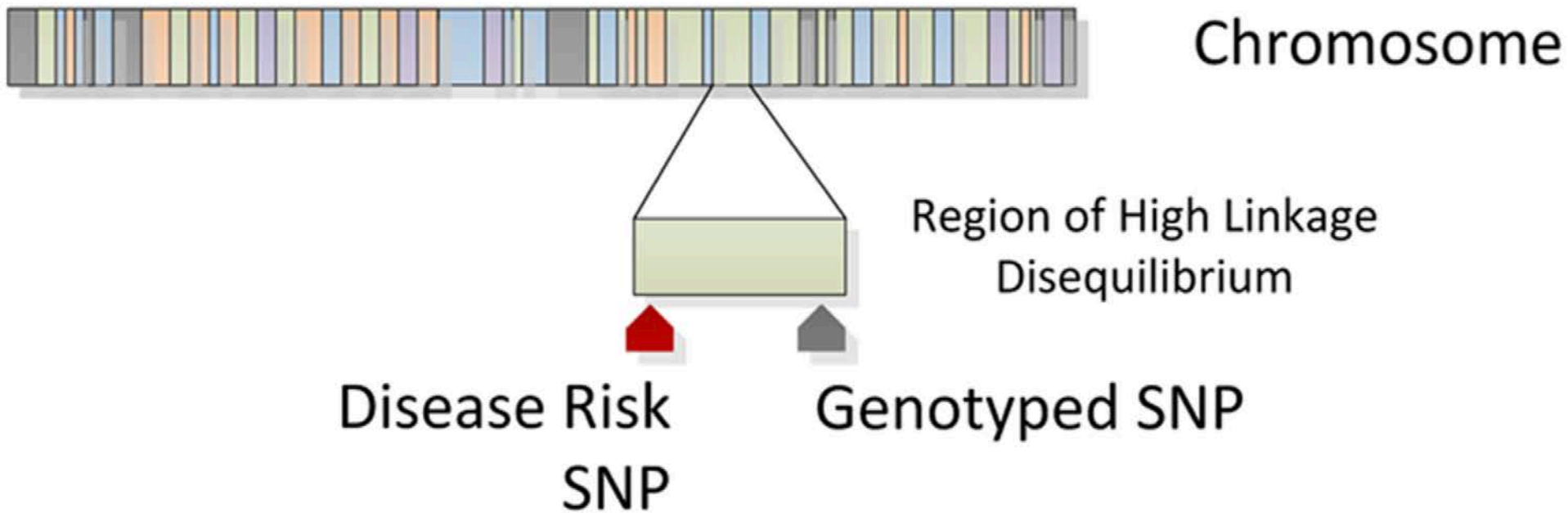
How LD arises & decays



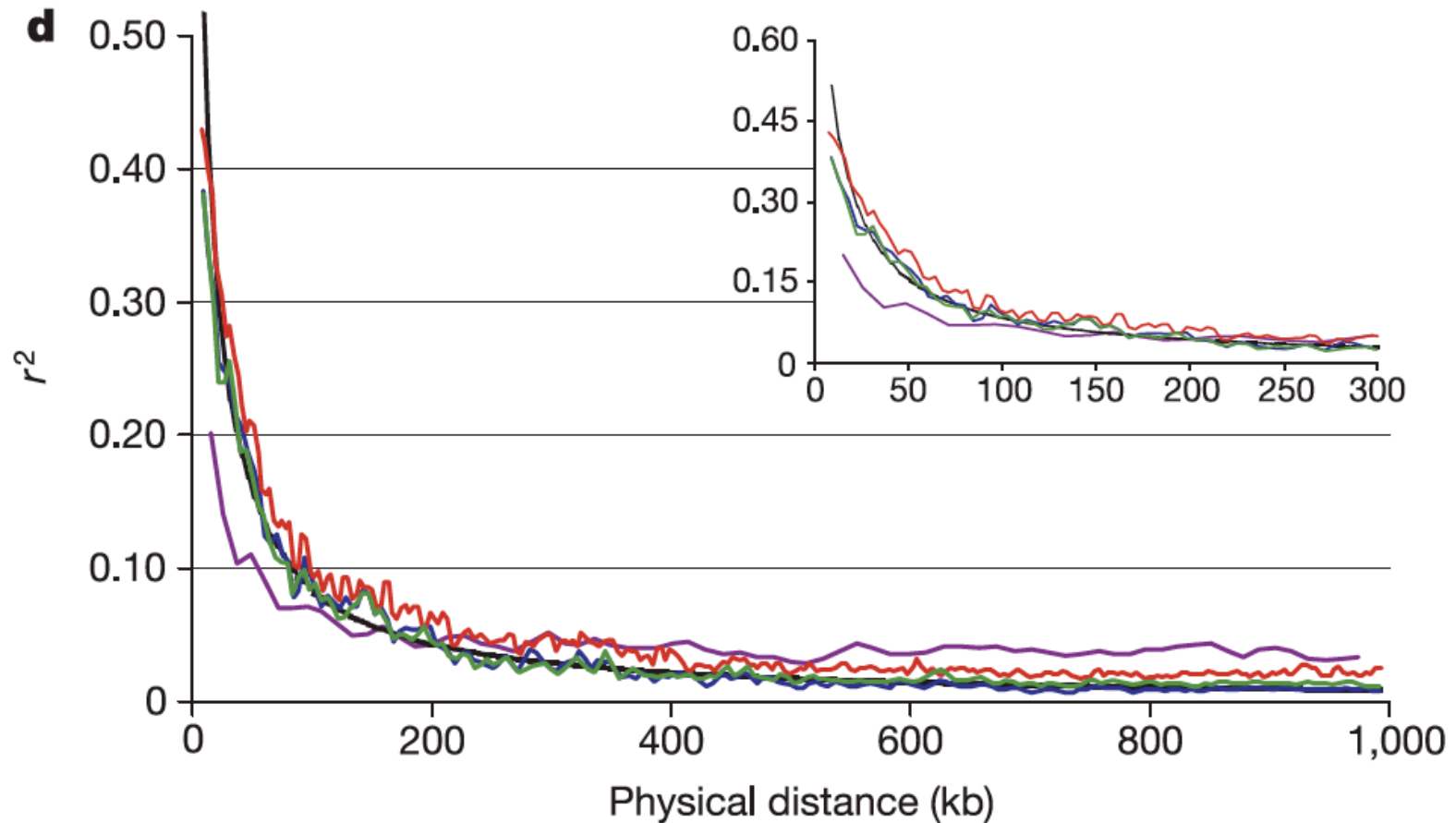
Decay of Linkage over successive generations



SNPs can tag other nearby SNPs...

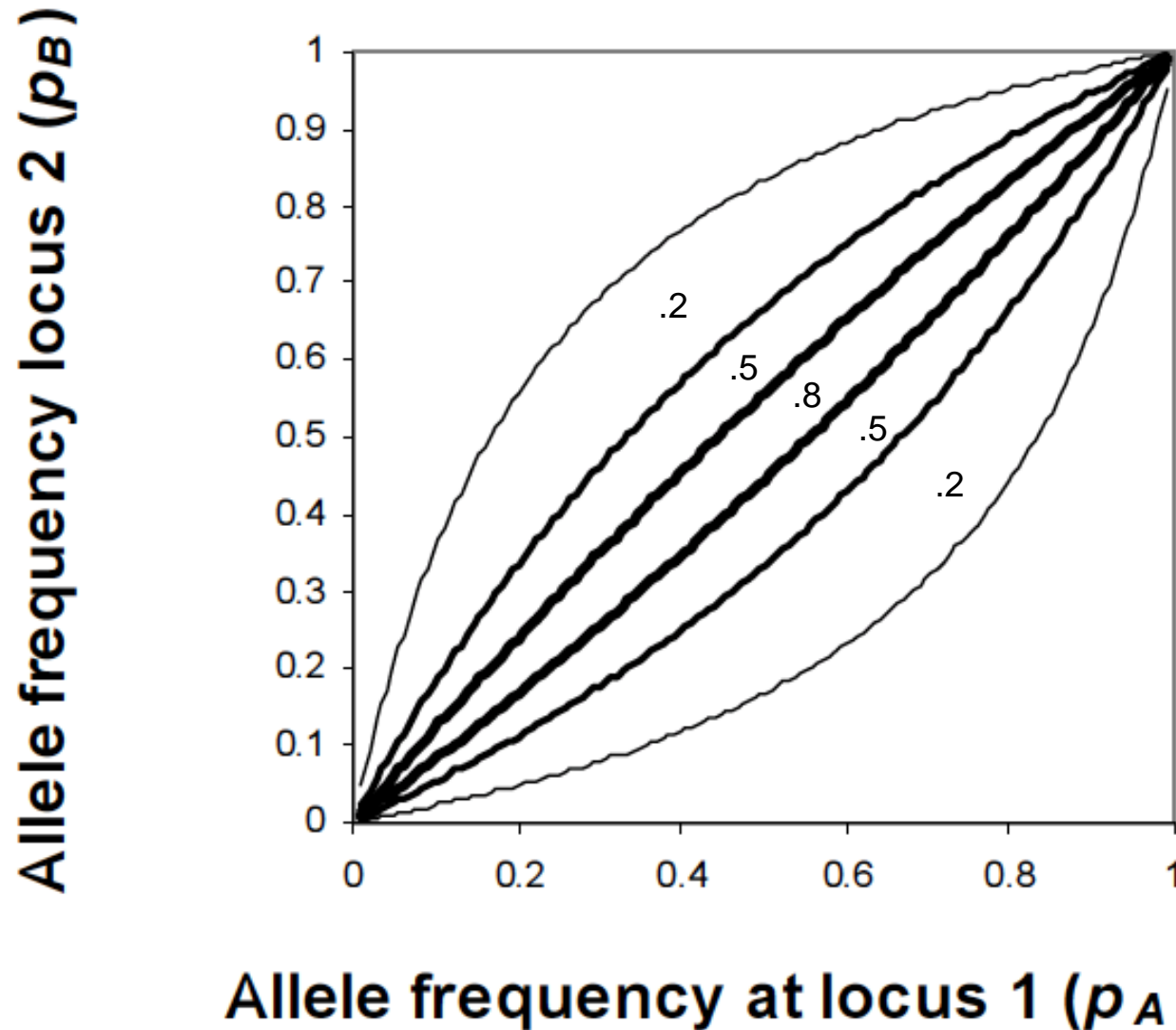


LD drops as a function of distance



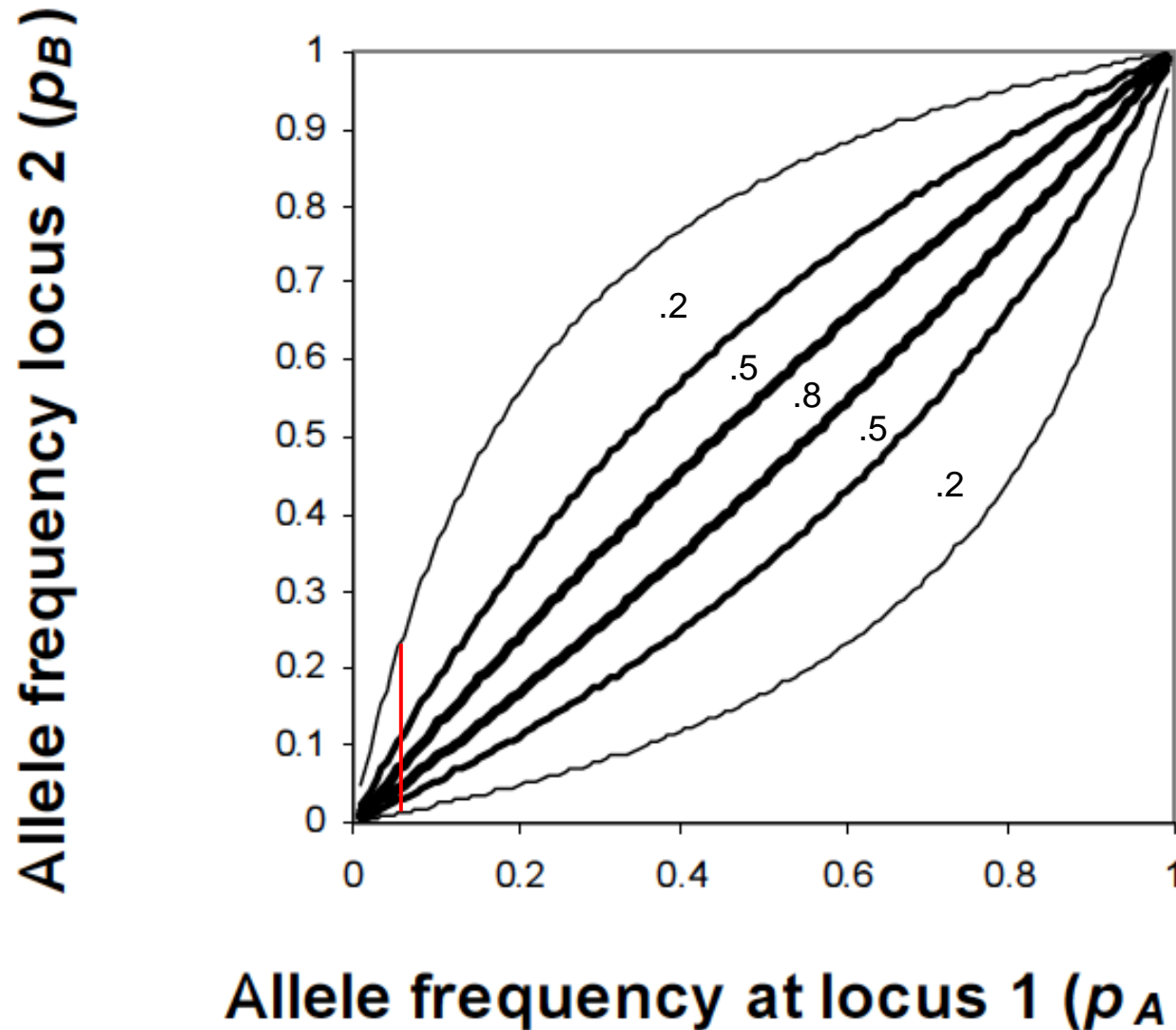
...and high LD possible only if the two alleles are of similar frequencies.

Possible range of allele frequencies given LD (r^2) between 2 SNPs



...and high LD possible only if the two alleles are of similar frequencies.

Possible range of allele frequencies given LD (r^2) between 2 SNPs



Imputation

Observed Genotypes

Observed Genotypes

. . . . A A A . . .
. . . . G C A . . .

Study
Sample

Reference Haplotypes

C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C

HapMap

Identify matches among Reference

Observed Genotypes

. **A** **A** **A**
. **G** **C** **A**

Reference Haplotypes

C	G	A	G	A	T	C	T	C	C	T	T	C	T	T	C	T	G	T	G	C
C	G	A	G	A	T	C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
C	C	A	A	G	C	T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
C	G	A	A	G	C	T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
C	G	A	G	A	C	T	C	T	C	C	G	A	C	C	T	T	A	T	G	C
T	G	G	G	A	T	C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
C	G	A	G	A	T	C	T	C	C	C	G	A	C	C	T	T	G	T	G	C
C	G	A	G	A	C	T	C	T	T	T	T	C	T	T	T	T	G	T	A	C
C	G	A	G	A	C	T	C	T	C	C	G	A	C	C	T	C	G	T	G	C
C	G	A	A	G	C	T	C	T	T	T	T	C	T	T	C	T	G	T	G	C

Impute missing genotypes

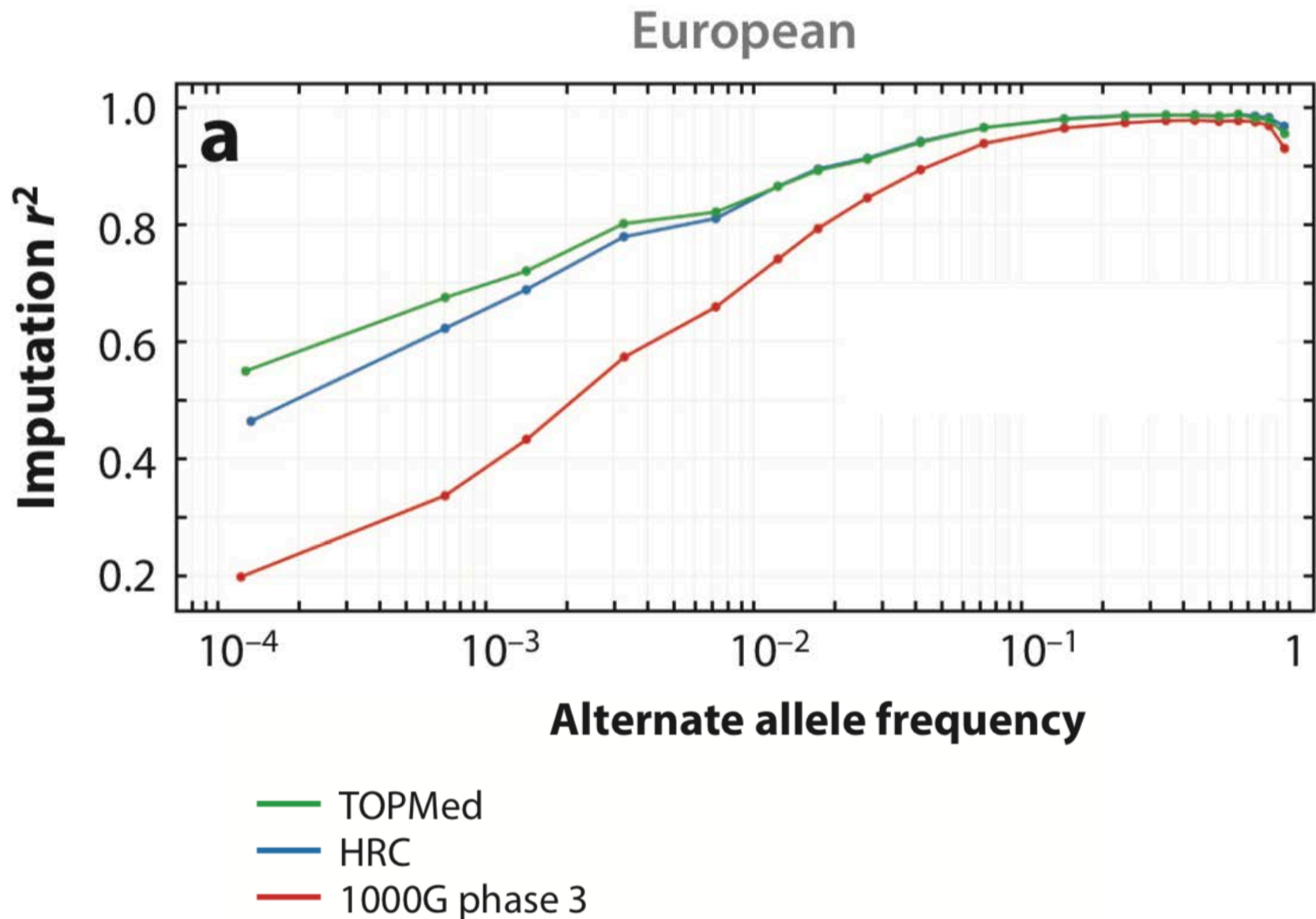
Observed Genotypes

c	g	a	g	A	t	c	t	c	c	c	g	A	c	c	t	c	A	t	g	g
c	g	a	a	G	c	t	c	t	t	t	t	C	t	t	t	c	A	t	g	g

Reference Haplotypes

C	G	A	G	A	T	C	T	C	C	T	T	C	T	T	C	T	G	T	G	C
C	G	A	G	A	T	C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
C	C	A	A	G	C	T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
C	G	A	A	G	C	T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
C	G	A	G	A	C	T	C	T	C	C	G	A	C	C	T	T	A	T	G	C
T	G	G	G	A	T	C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
C	G	A	G	A	T	C	T	C	C	C	G	A	C	C	T	T	G	T	G	C
C	G	A	G	A	C	T	C	T	T	T	T	C	T	T	T	T	G	T	A	C
C	G	A	G	A	C	T	C	T	C	C	G	A	C	C	T	C	G	T	G	C
C	G	A	A	G	C	T	C	T	T	T	T	C	T	T	C	T	G	T	G	C

Imputation quality as function of AF



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, ~ 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

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

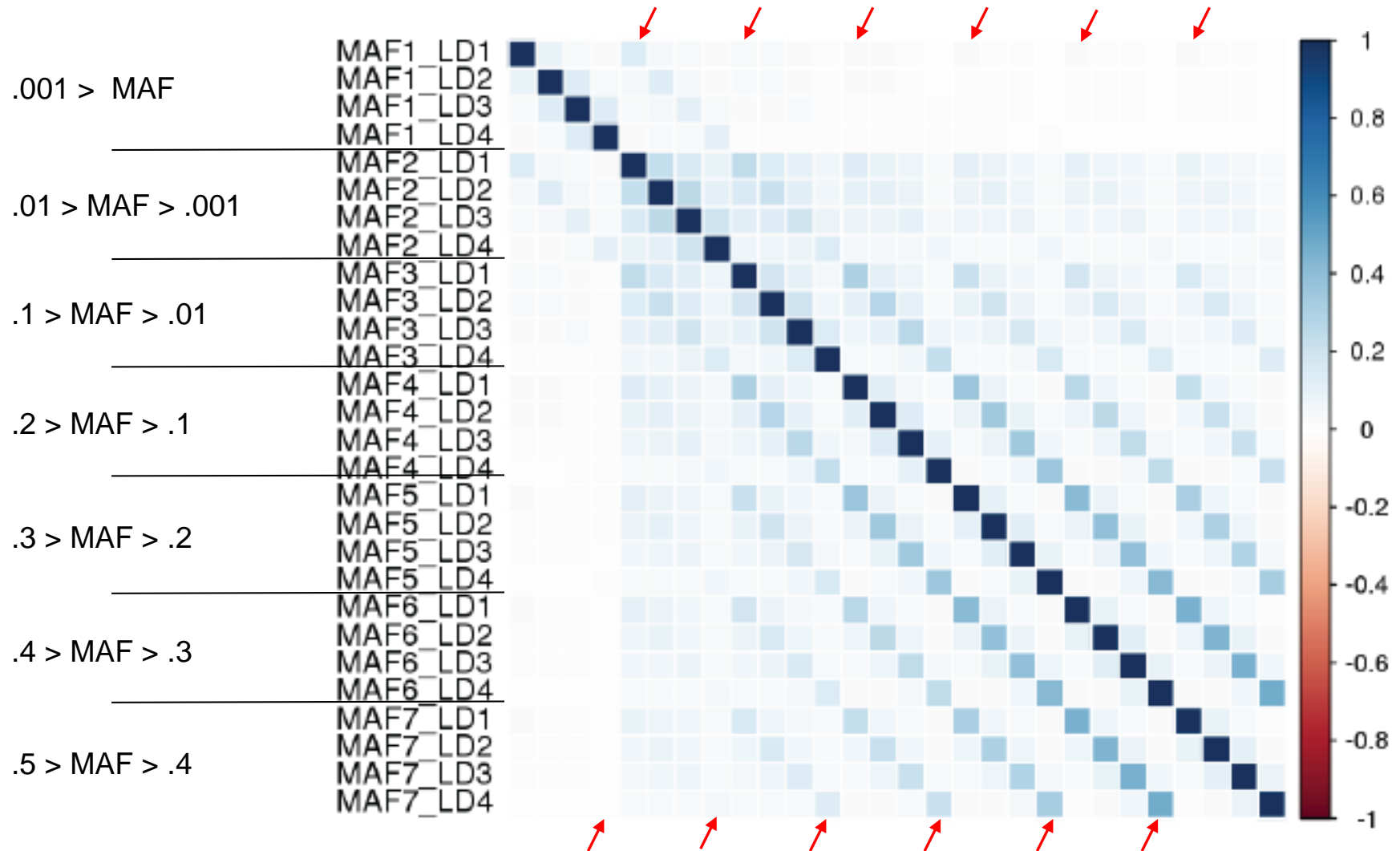
$$\hat{\pi}_{jk} = \frac{1}{m} \sum_i \left(\frac{x_{ij} - 2p_i}{\sqrt{2p_i(1 - p_i)}} \right) \left(\frac{x_{ik} - 2p_i}{\sqrt{2p_i(1 - p_i)}} \right)$$

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

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

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

Cor. of $\hat{\pi}$ across MAF/LD binned GRMs



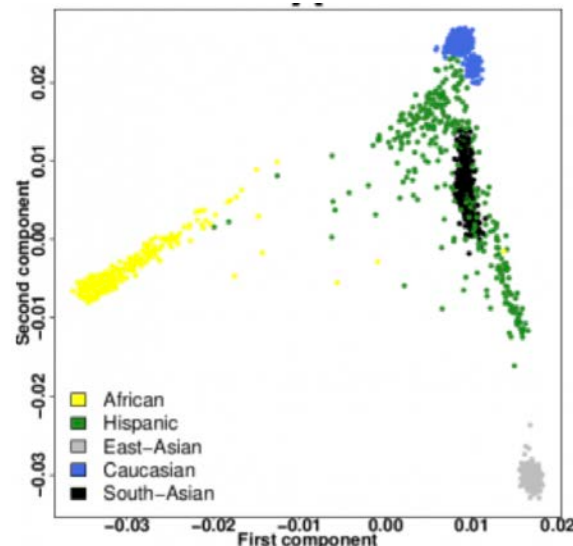
STRATIFICATION

Stratification

- Population subdivision (non-random mating) creates allele frequency differences between populations due to random genetic drift

Stratification

- Population subdivision (non-random mating) creates allele frequency differences between populations due to random genetic drift
- A principal components analysis (PCA) of n -by- n matrix of p ihats (called the GRM - genomic relationship matrix) can pull out axes that differentiate such subdivision (ancestry).



Stratification

- Population subdivision (non-random mating) creates allele frequency differences between populations due to random genetic drift
- A principal components analysis (PCA) of n -by- n matrix of p ihats (called the GRM - genomic relationship matrix) can pull out axes that differentiate such subdivision (ancestry).
- If there are mean phenotypic differences between ancestry groups (or differential sampling of ancestries in cases vs. controls), stratification is a potential confounder.
- Typically, control for this using first 5-20 PCs

H-E REGRESSION

Regression estimates of h^2

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

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

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

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

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

Regression estimates of h^2

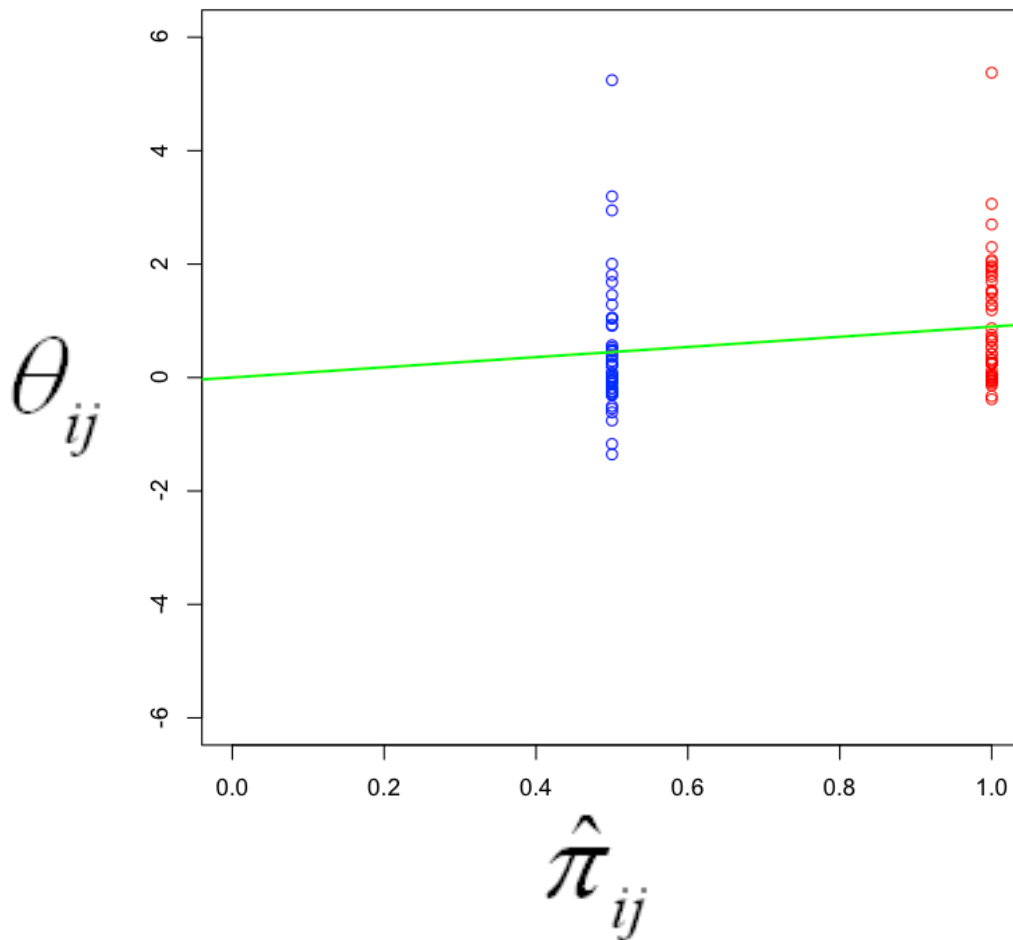
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



Regression estimates of h^2

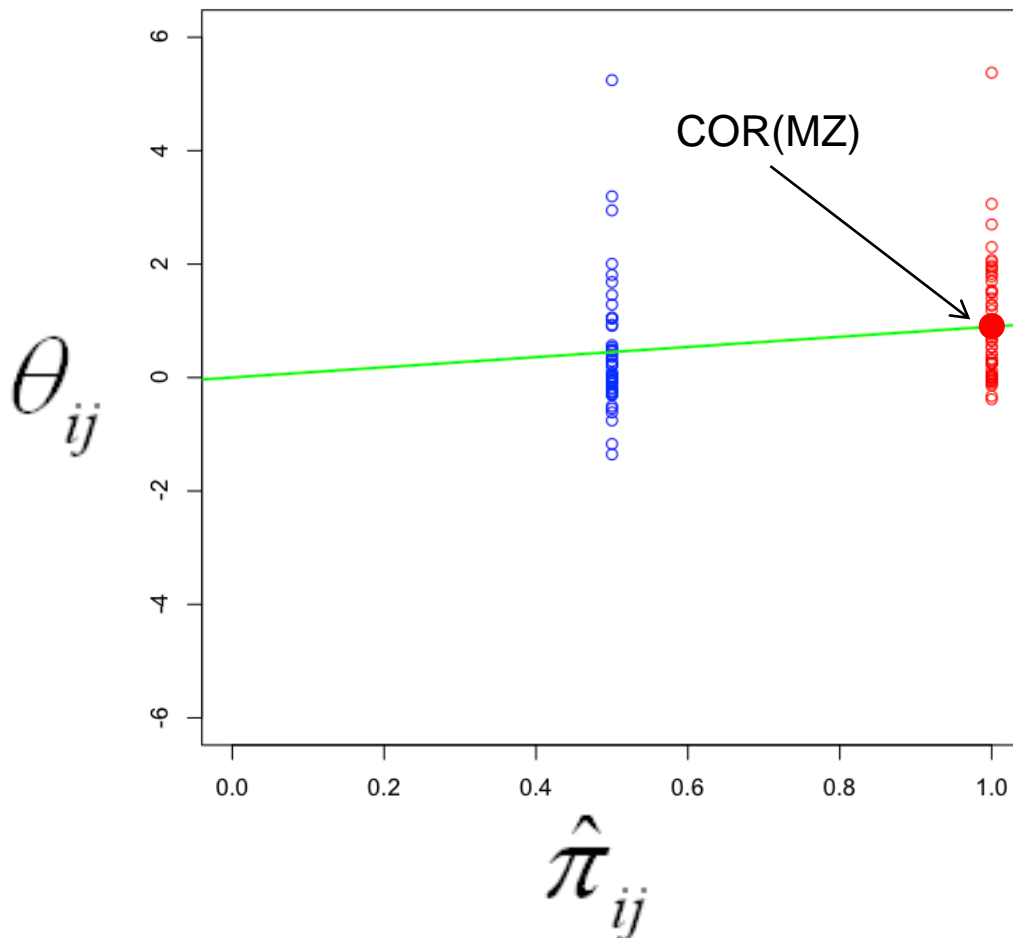
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



Regression estimates of h^2

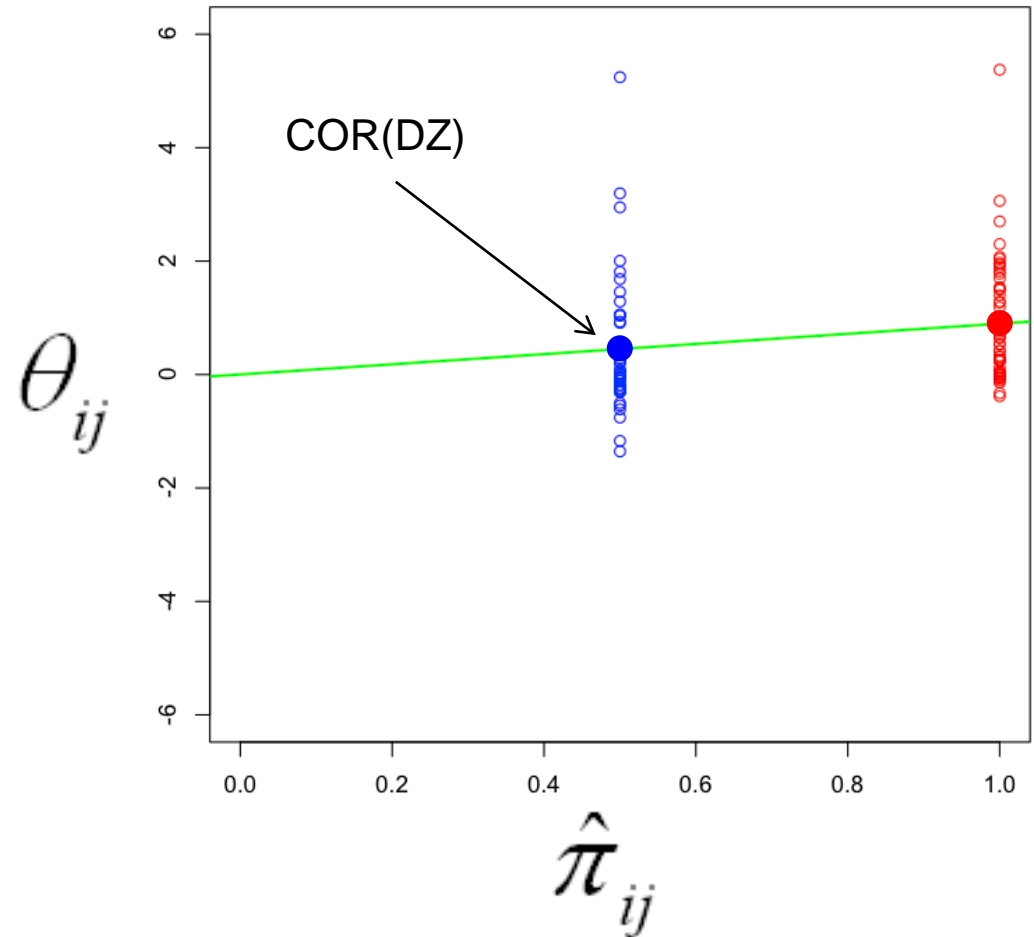
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



Regression estimates of h^2

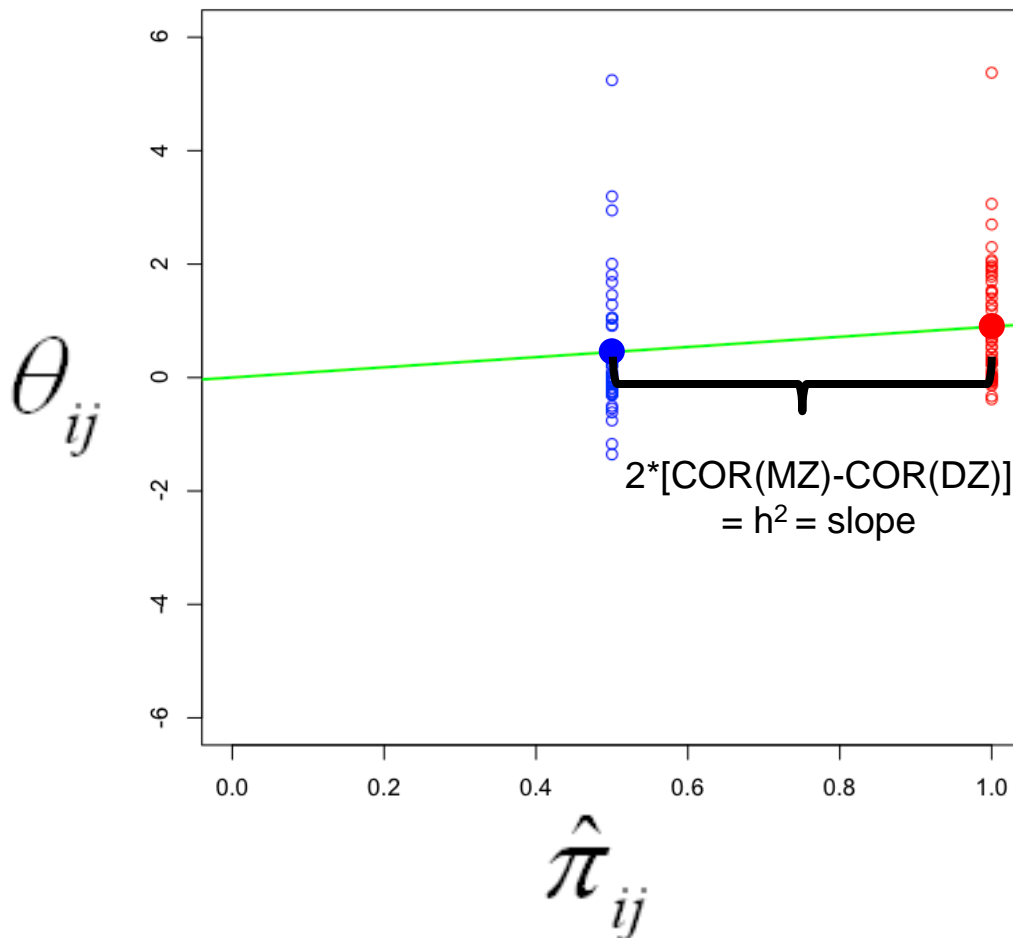
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



Regression estimates of h^2

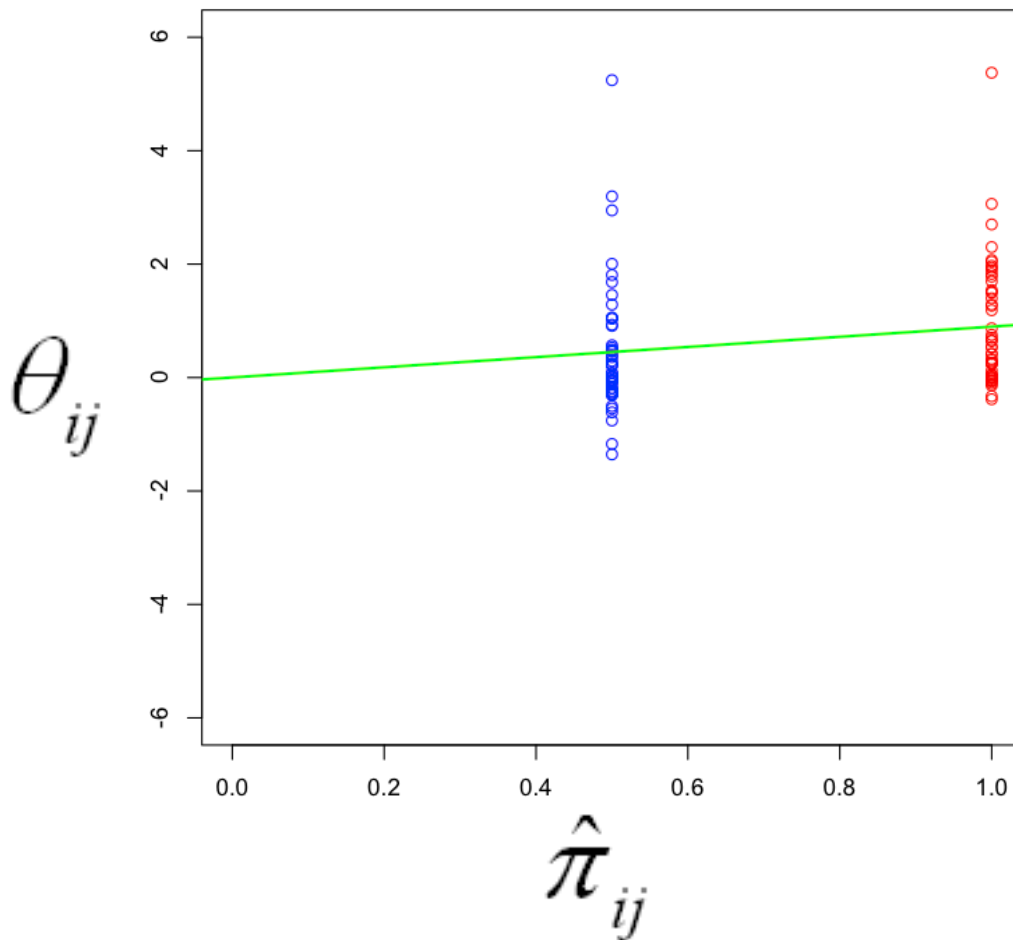
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



Regression estimates of h^2

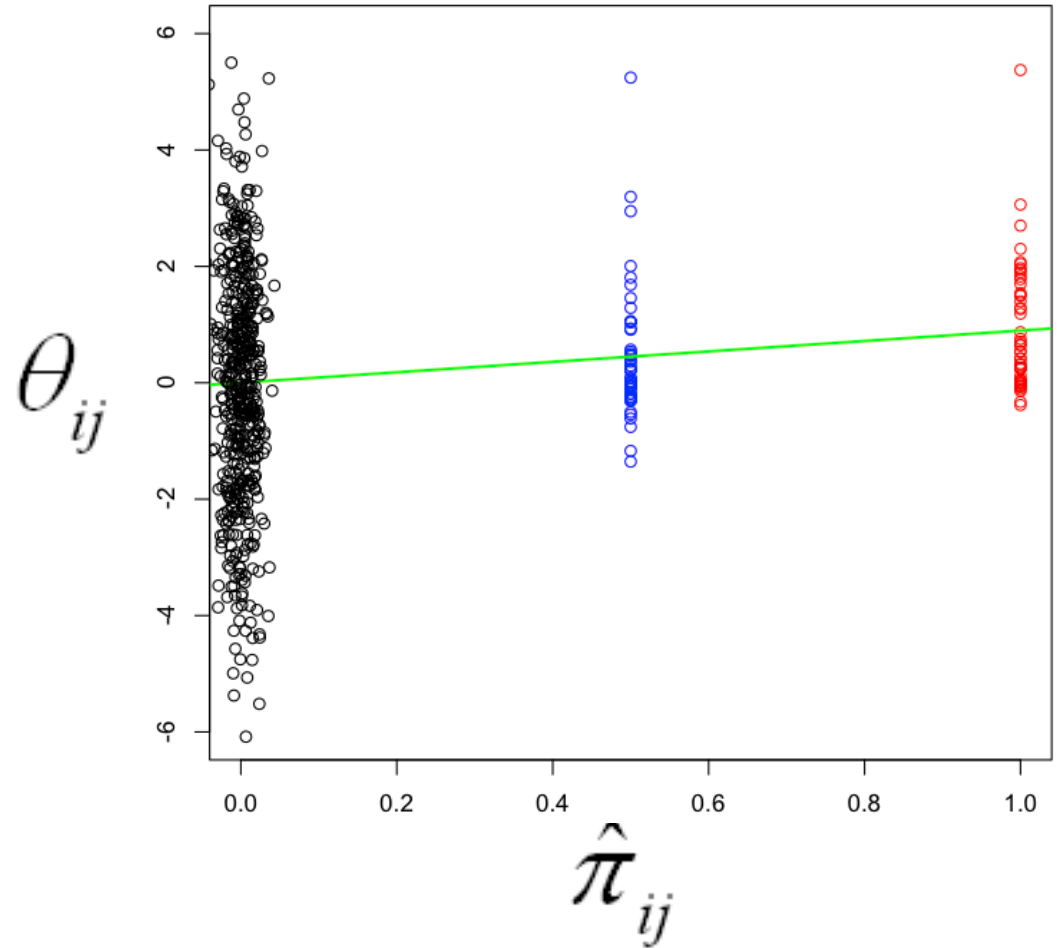
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



Regression estimates of h^2_{snp}

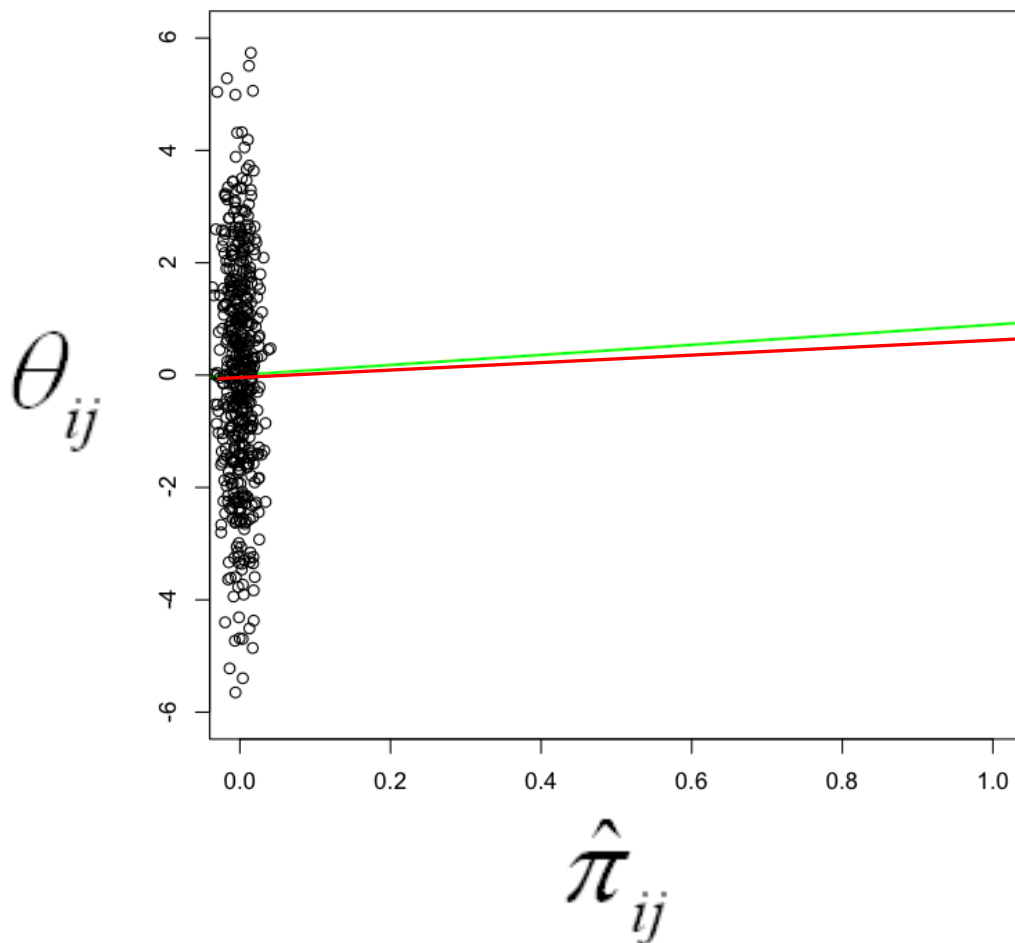
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



GREML

nature
genetics

2010

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

Jian Yang¹, Beben Benyamin¹, Brian P McEvoy¹, Scott Gordon¹, Anjali K Henders¹, Dale R Nyholt¹, Pamela A Madden², Andrew C Heath², Nicholas G Martin¹, Grant W Montgomery¹, Michael E Goddard³ & Peter M Visscher¹

GREML Model

$$y_{.x} = Z\hat{u} + \hat{e}$$

The diagram illustrates the GREML model equation $y_{.x} = Z\hat{u} + \hat{e}$ with numerical values for each term:

- $y_{.x}$ (phenotypic values): $\begin{bmatrix} -.64 \\ -2.58 \\ 3.21 \end{bmatrix}$
- Z (design matrix for SNP effects): $\begin{bmatrix} 1.15 & -.58 & -1.15 \\ -.58 & 1.15 & .58 \\ -.58 & -.58 & .58 \end{bmatrix}$
- \hat{u} (SNP effects): $\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \end{bmatrix}$
- \hat{e} (residuals): $\begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \hat{e}_3 \end{bmatrix}$

The equation is represented as:

$$\begin{bmatrix} -.64 \\ -2.58 \\ 3.21 \end{bmatrix} = \begin{bmatrix} 1.15 & -.58 & -1.15 \\ -.58 & 1.15 & .58 \\ -.58 & -.58 & .58 \end{bmatrix} * \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \end{bmatrix} + \begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \hat{e}_3 \end{bmatrix}$$

Below the design matrix Z , the formula for its elements is given:

$$\text{design matrix for SNP effects} = \frac{x_{ij} - 2p_i}{\sqrt{2p_i(1-p_i)}}$$

Labels below the vectors:

- $y_{.x}$ is labeled below the first vector.
- \hat{u} is labeled "SNP effects" below the third vector.
- \hat{e} is labeled "residuals" below the fourth vector.

GREML Model

$$y_{.x} = Z\hat{u} + \hat{e}$$

-.64
-2.58
3.21

 $=$

1.15	-.58	-1.15
-.58	1.15	.58
-.58	-.58	.58

 $*$

\hat{u}_1
\hat{u}_2
\hat{u}_3

 $+$

\hat{e}_1
\hat{e}_2
\hat{e}_3

$y_{.x}$
design matrix for SNP effects =
SNP effects
residuals

$$\frac{x_{ij} - 2p_i}{\sqrt{2p_i(1-p_i)}}$$

We aren't interested in estimates of each u_i because such individual estimates are unreliable when $m > n$. Instead, estimate the variance of u_i .

GREML Model

$$y_{.X} = Z\hat{u} + \hat{e}$$

The diagram illustrates the GREML model equation $y_{.X} = Z\hat{u} + \hat{e}$. Arrows point from each term in the equation to its corresponding numerical representation:

- $y_{.X}$ is represented by a column vector: $\begin{bmatrix} -.64 \\ -2.58 \\ 3.21 \end{bmatrix}$
- Z is represented by a matrix: $\begin{bmatrix} 1.15 & -.58 & -1.15 \\ -.58 & 1.15 & .58 \\ -.58 & -.58 & .58 \end{bmatrix}$
- \hat{u} is represented by a column vector: $\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \end{bmatrix}$
- \hat{e} is represented by a column vector: $\begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \hat{e}_3 \end{bmatrix}$

The full equation is shown as: $\begin{bmatrix} -.64 \\ -2.58 \\ 3.21 \end{bmatrix} = \begin{bmatrix} 1.15 & -.58 & -1.15 \\ -.58 & 1.15 & .58 \\ -.58 & -.58 & .58 \end{bmatrix} * \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \end{bmatrix} + \begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \hat{e}_3 \end{bmatrix}$

Below the matrix, the text reads: "design matrix for SNP effects = $\frac{x_{ij} - 2p_i}{\sqrt{2p_i(1-p_i)}}$ ".

Below the \hat{u} vector, the text reads: "SNP effects".

Below the \hat{e} vector, the text reads: "residuals".

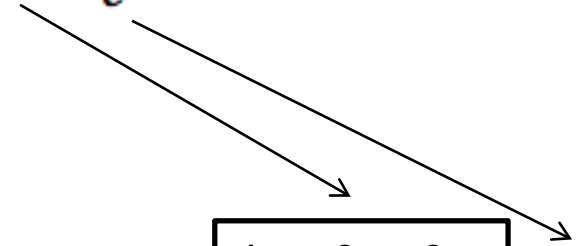
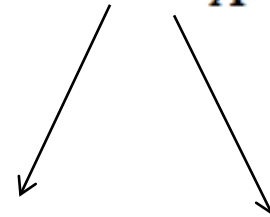
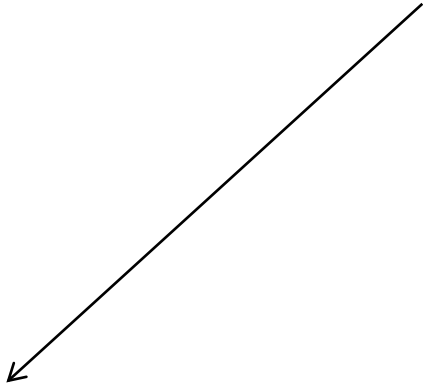
We assume $u \sim N(0, \sigma_u^2)$ and are iid

and therefore $\sigma_A^2 = \sum_{i=1}^m \sigma_u^2 = m\sigma_u^2$

GREML Model

(we treat u as random and estimate σ_u^2 and thus σ_A^2)

$$\begin{aligned}\text{var}(y_{.X}) &= ZZ' \sigma_u^2 + I \sigma_e^2 \\ &= ZZ' (\sigma_A^2 / m) + I \sigma_e^2 \\ &= G \sigma_A^2 + I \sigma_e^2\end{aligned}$$



.41	1.65	-2.05
1.65	6.66	-8.28
-2.05	-8.28	10.3

=

.99	-.02	-.01
-.02	1.0	.01
-.01	.01	1.02

$\hat{\sigma}_A^2$

+

1	0	0
0	1	0
0	0	1

$\hat{\sigma}_e^2$

observed n-by-n
var/covar matrix
of y

Genomic Relationship Matrix (GRM)
at measured SNPs. Each element =

Identity
matrix

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

GREML

$$\begin{aligned}\text{var}(y_{.X}) &= ZZ' \sigma_u^2 + I \sigma_e^2 \\ &= ZZ' (\sigma_A^2 / m) + I \sigma_e^2 \\ &= G \sigma_A^2 + I \sigma_e^2\end{aligned}$$

$$\begin{bmatrix} .41 & 1.65 & -2.05 \\ 1.65 & 6.66 & -8.28 \\ -2.05 & -8.28 & 10.3 \end{bmatrix} = \begin{bmatrix} .99 & -.02 & -.01 \\ -.02 & 1.0 & .01 \\ -.01 & .01 & 1.02 \end{bmatrix} \hat{\sigma}_A^2 + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \hat{\sigma}_e^2$$

observed var/covar = implied var/covar

REML find values of $\hat{\sigma}_A^2$ & $\hat{\sigma}_e^2$ that maximizes the likelihood of the observed data. Intuitively, this makes the observed and implied var-covar matrices be as similar as possible.

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 great influence of extreme phenotypes. 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

RUNNING GCTA

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 > ~.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 ascertainments) 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})

GCTA command & input

COMMAND:

gcta
--grm-bin <path>/SNPs.rel05

gcta --bfile SNPs --make-grm-bin --out SNPs.rel05

--pheno <path>/test.phen

test.phen (no header line; columns are family ID, individual ID and phenotypes)

011	0101	0.98
012	0102	-0.76
013	0103	-0.06
.....		

--covar <path>/test.covar

test.covar (no header line; columns are family ID, individual ID and discrete covariates)

01	0101	F	Adult	0
02	0203	M	Adult	0
03	0305	F	Adolescent	1
.....				

--qcovar <path>/test.qcovar

test.qcovar (no header line; columns are family ID, individual ID and quantitative covariates)

01	0101	-0.024	0.012
02	0203	0.032	0.106
03	0305	0.143	-0.056
.....			

--reml --out SNPgrm.randomCV

GCTA command & output

COMMAND:

```
gcta --grm-bin <path>/SNPs.rel05 --pheno <path>/pheno.txt --covar  
<path>/cov.txt --reml --out SNPgrm.randomCV
```

OUTPUT: cat SNPgrm.randomCV.hsq

Source	Variance	SE
V(G)	0.300098	0.275857
V(e)	1.730548	0.279257
Vp	2.030646	0.049529
V(G)/Vp	0.147785	0.135820
logL	-2876.706	
logL0	-2877.338	
LRT	1.264	
Df	1	
Pval	0.1305	
N	3363	

GCTA command & output

COMMAND:

```
gcta --grm-bin <path>/SNPs.rel05 --pheno <path>/pheno.txt --covar  
<path>/cov.txt --reml --out SNPgrm.randomCV
```

OUTPUT: cat SNPgrm.randomCV.hsq

Source	Variance	SE
V(G)	0.300098	0.275857
V(e)	1.730548	0.279257
Vp	2.030646	0.049529
V(G)/Vp	0.147785	0.135820
logL	-2876.706	
logL0	-2877.338	
LRT	1.264	
Df	1	
Pval	0.1305	
N	3363	

← h^2_{snp} & SE

95% CI:

$0.147 - 1.96 \times 0.134 = -0.12$

$0.147 + 1.96 \times 0.134 = 0.41$

GCTA command & output

COMMAND:

```
gcta --grm-bin <path>/SNPs.rel05 --pheno <path>/pheno.txt --covar  
<path>/cov.txt --reml --out SNPgrm.randomCV
```

OUTPUT: cat SNPgrm.randomCV.hsq

Source	Variance	SE
V(G)	0.300098	0.275857
V(e)	1.730548	0.279257
Vp	2.030646	0.049529
V(G)/Vp	0.147785	0.135820
logL	-2876.706	
logL0	-2877.338	
LRT	1.264	
Df	1	
Pval	0.1305	
N	3363	



Likelihood Ratio Test

Testing if $V(G) > 0$

$-2*(-2877.338 - -2876.706) = 1.26$

χ^2 test, 1 df

LD-score regression

nature
genetics

2015

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan¹⁻³, Po-Ru Loh^{1,4}, Hilary K Finucane^{4,5}, Stephan Ripke^{2,3}, Jian Yang⁶, Schizophrenia Working Group of the Psychiatric Genomics Consortium⁷, Nick Patterson¹, Mark J Daly¹⁻³, Alkes L Price^{1,4,8} & Benjamin M Neale¹⁻³

How does LD shape association?

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Affiliations | Contributions | Corresponding author

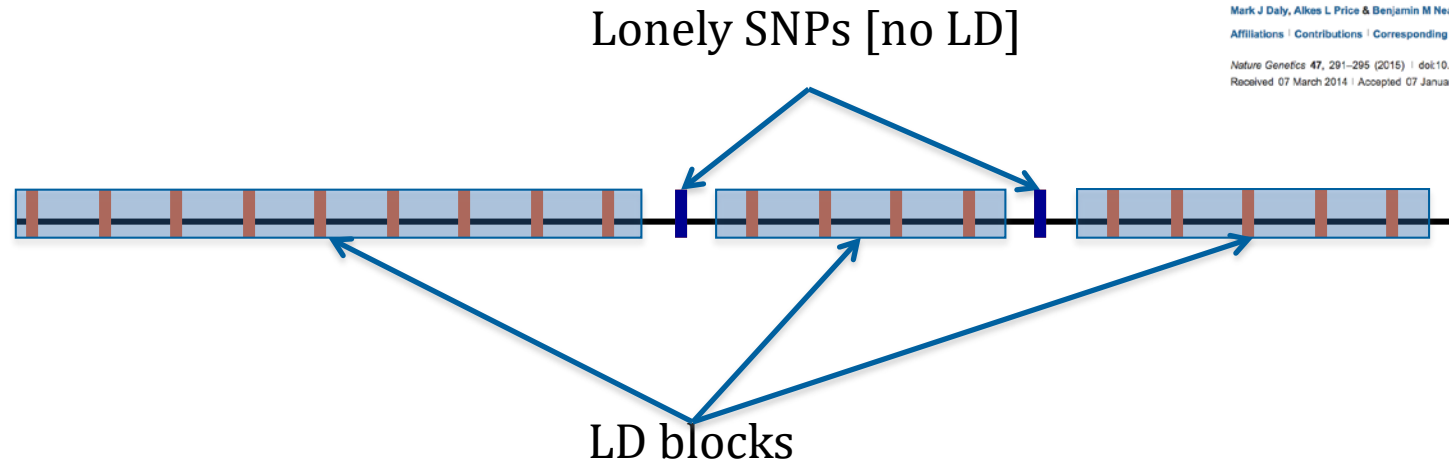
Nature Genetics 47, 291–295 (2015) | doi:10.1038/ng.3211

Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015



Cribbed from Neale, Boulder Workshop, 2017

How does LD shape association?



LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

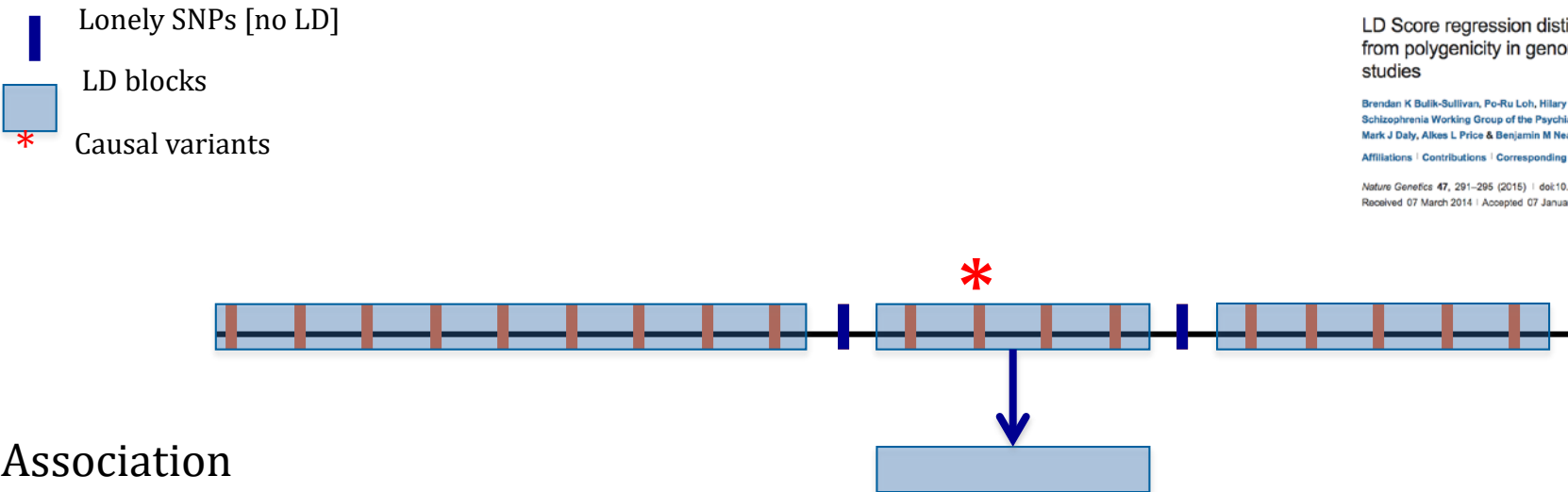
Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics 47, 291–295 (2015) | doi:10.1038/ng.3211

Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015

How does LD shape association?



LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics 47, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015

All markers correlated with a causal variant show association

How does LD shape association?

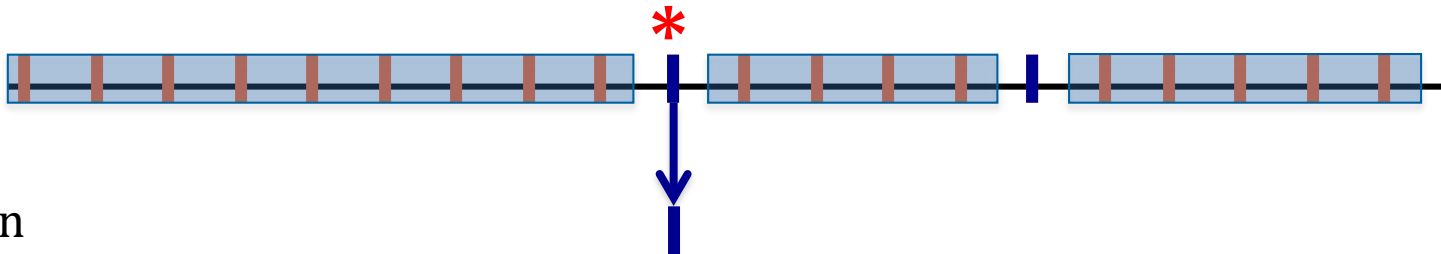
- Lonely SNPs [no LD]
- LD blocks
- Causal variants

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics 47, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015



Lonely SNPs only show association if they are causal

What happens under polygenicity?

- Lonely SNPs [no LD]
- LD blocks
- Causal variants

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics 47, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015

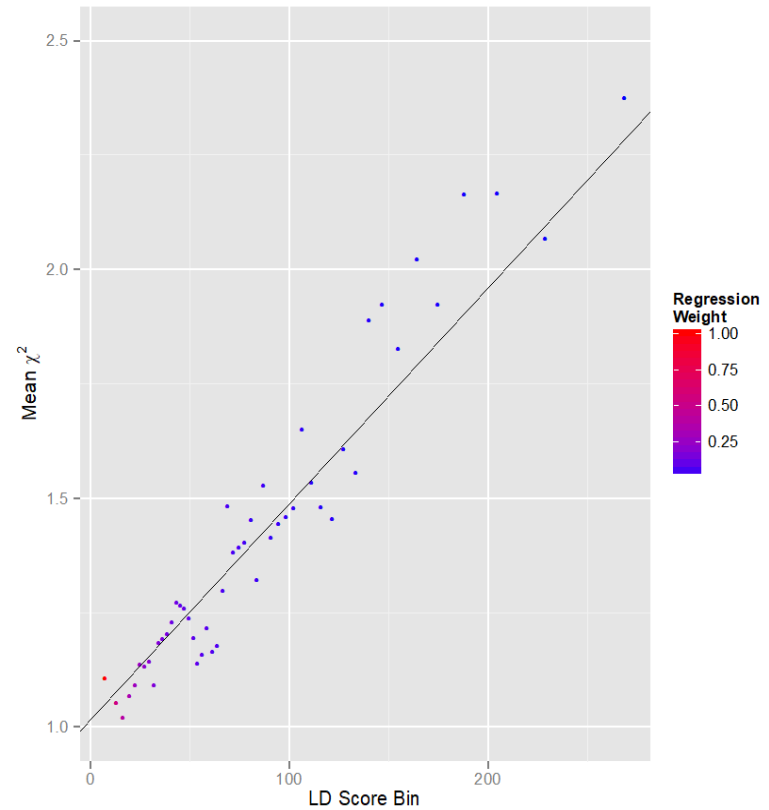
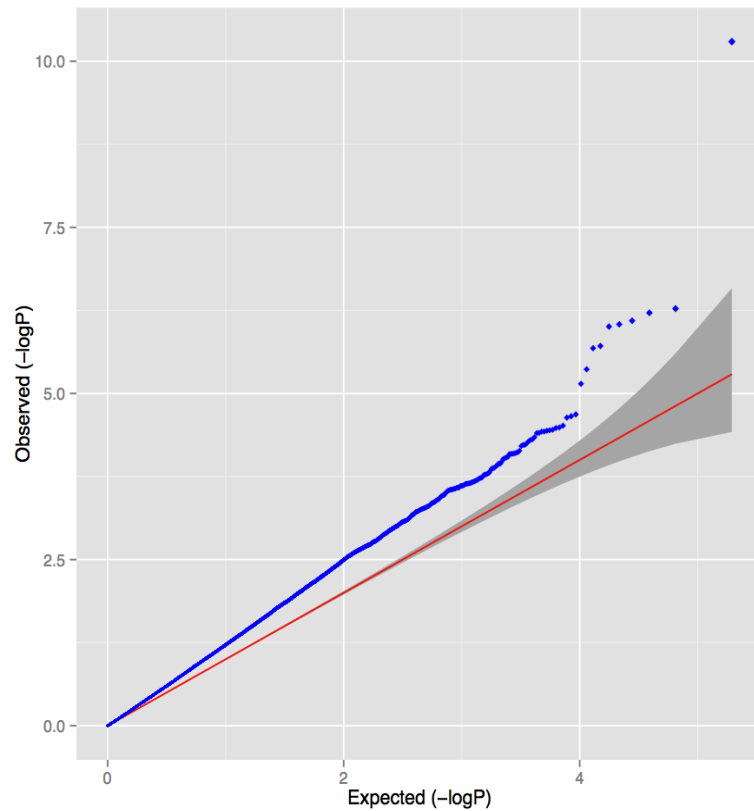


Assuming a random CV placement, we see SNPs with more LD friends showing more association

The more you tag, the more likely you are to tag a causal variant

Simulated polygenic architecture

$\Lambda = 1.30$ LD score intercept = 1.02



What happens under stratification?

- Lonely SNPs [no LD]
- LD blocks
- Causal variants

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Affiliations | Contributions | Corresponding author

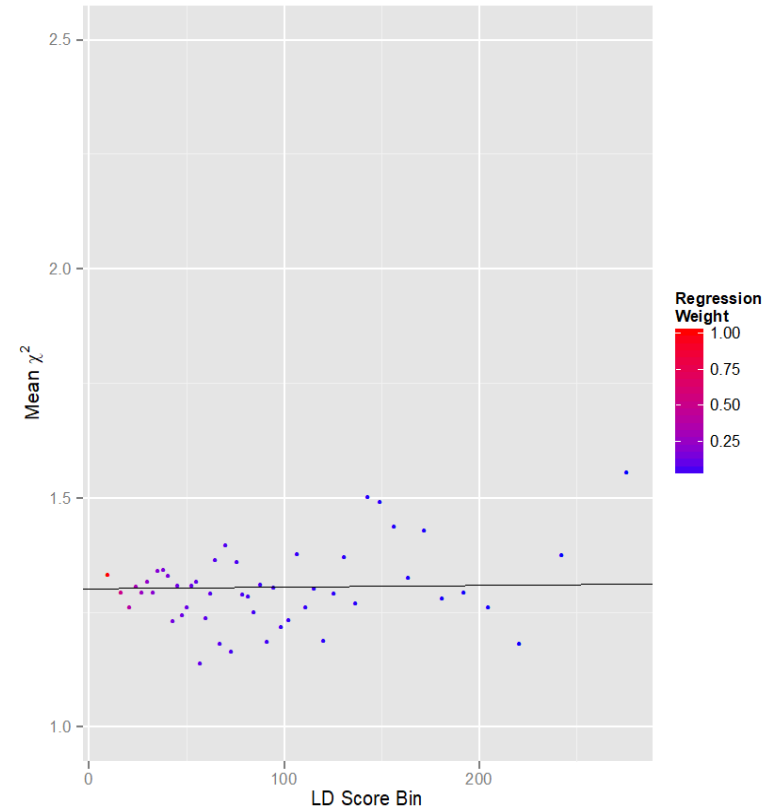
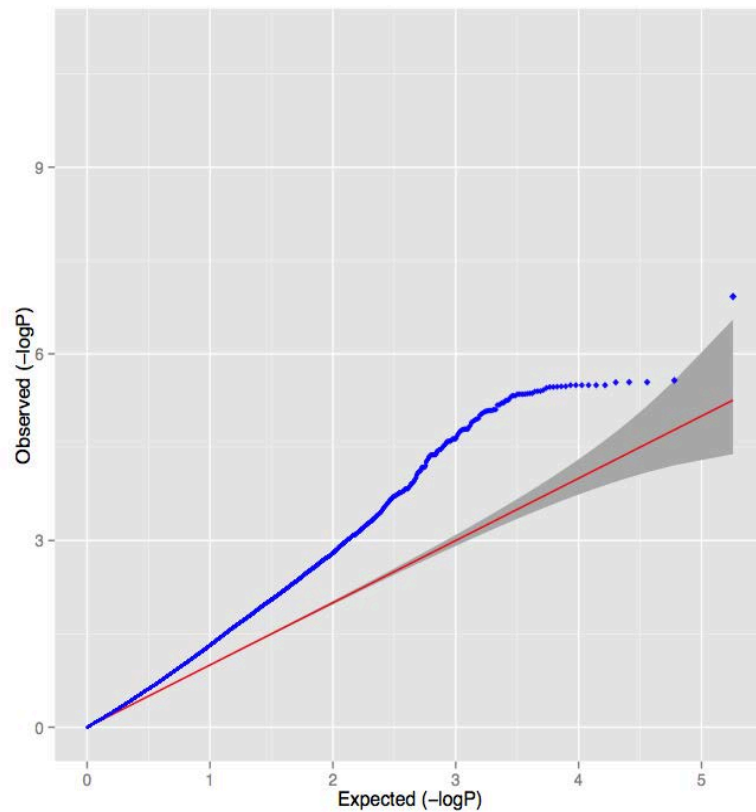
Nature Genetics 47, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015



Under pure drift we expect LD to have no relationship to differences in allele frequencies between populations

UK controls versus Sweden controls

$\Lambda = 1.30$ LD score intercept = 1.32



LD Score regression

What is the $E[\chi^2]$ for variant j ?

$$E[\chi_j^2] = 1 + Na + \frac{h_g^2 N}{M} l_j$$

where N =sample size, M =# of SNPs, a =inflation due to confounding, h_g^2 is heritability (total obs.) and l_j is the *LD Score*: $l_j = \sum_{k \neq j} r_{jk}^2$

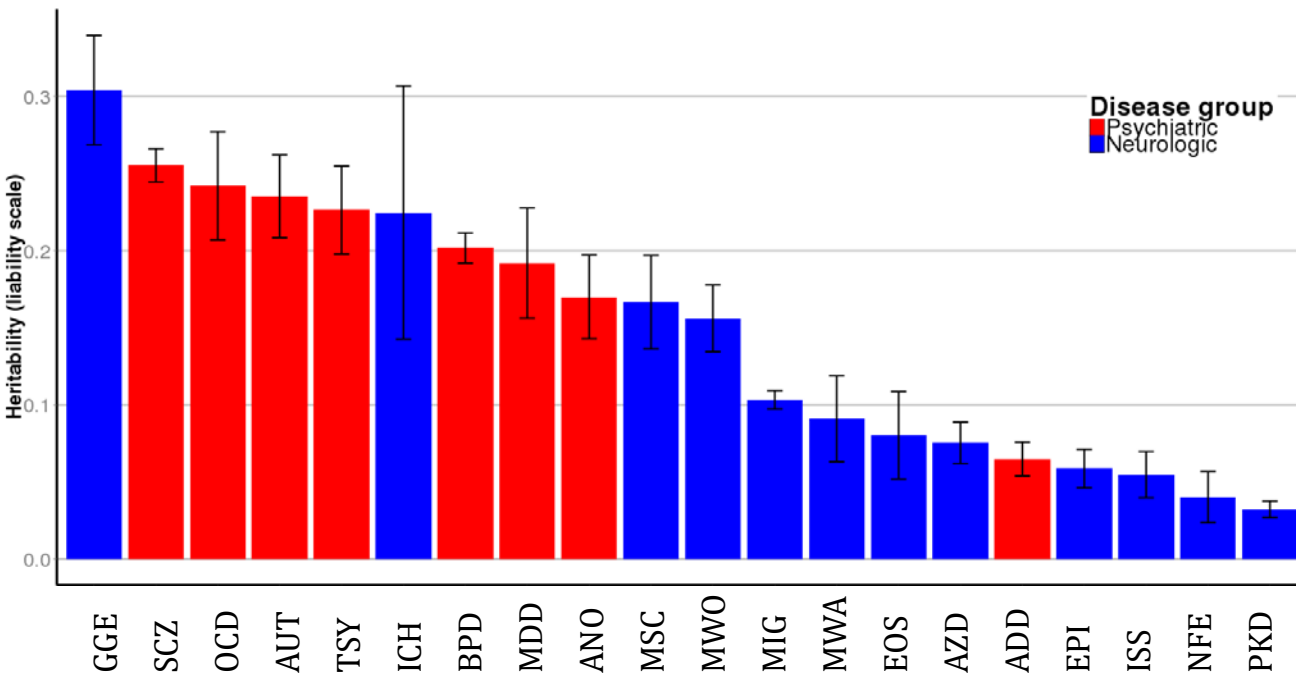
Thus, an estimator of heritability is:

$$\hat{h}^2 = \frac{M}{N} \hat{\beta}_1$$

Advantages of LD-score regression

- It only requires GWAS summary statistics. It is therefore fast, easy to use, and can be employed on very large sample sizes.
- Because it relies on LD-2 relationship, and because rare SNPs have very low LD with untyped SNPs, it breaks down when estimating SNPs with $MAF < .01$ (shown below).
- However, it should provide decent (\sim unbiased) estimates of $r_g \cong r_{snp} = \frac{COV_{snp}(T1,T2)}{\sqrt{VG_{snp}(T1)*VG_{snp}(T2)}}$

Univariate heritability from common variation



GGE = Generalized Epilepsy
 SCZ = Schizophrenia
 OCD = Obsessive Compulsive Disorder
 AUT = Autism
 TSY = Tourette's Syndrome
 ICH = Intracerebral Hemorrhage
 BPD = Bipolar Disorder
 MDD = Major Depressive Disorder
 ANO = Anorexia Nervosa
 MSC = Multiple Sclerosis
 MWO = Migraine without Aura
 MIG = Migraine
 MWA = Migraine with Aura
 EOS = Early Onset Stroke
 AZD = Alzheimer's Disease
 ADD = Attention Deficit/Hyperactivity
 EPI = Epilepsy (all)
 ISS = Ischemic Stroke
 NFE = Non-acquired focal epilepsy
 PKD = Parkinson's Disease

LDAK

ARTICLES

nature
genetics

2017

Reevaluation of SNP heritability in complex human traits

Doug Speed¹, Na Cai^{2,3}, the UCLEB Consortium⁴, Michael R Johnson⁵, Sergey Nejentsev⁶ & David J Balding^{1,7}

ARTICLE

2012

Improved Heritability Estimation from Genome-wide SNPs

Doug Speed,^{1,*} Gibran Hemani,² Michael R. Johnson,³ and David J. Balding¹

Changing GREML assumptions by weighting $\hat{\pi}$

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

- A more general form of this formula is:

$$\hat{\pi}_{jk} = \frac{1}{\sum W_i} \sum_i W_i (x_{ij} - 2p_i)(x_{ik} - 2p_i) [2p_i(1 - p_i)]^\alpha$$

- Which reduces to the typical formulation above when:

$$W_i = 1 \ \forall i = 1 \dots m \quad \& \quad \alpha = -1$$

- The choice of W_i and α are arbitrary, and depend on implicit assumptions about which types of SNPs tag CVs & CV effect sizes. If we heavily weight a certain type of SNP (e.g., those on even chromosomes), we assume such SNPs better tag CVs.

Typical (GCTA) assumptions implicit in $\hat{\pi}$

$$\hat{\pi}_{jk} = \frac{1}{\sum W_i} \sum_i W_i (x_{ij} - 2p_i)(x_{ik} - 2p_i)[2p_i(1 - p_i)]^\alpha$$

Assumptions

$$W_i = 1 \quad \forall i = 1 \dots m$$

$$\alpha = -1$$

Consequences

SNPs have equal weight, even if they are poorly imputed and redundantly tag the same CV

Rarer SNPs (which tag rarer CVs) receive more weight, ostensibly due to NS. This means the variance explained per SNP is invariant across MAF:

$$G_i = (X_i - 2p_i)[2p_i(1 - p_i)]^{\alpha/2}$$

$$V[G_i] = [2p_i(1 - p_i)]^\alpha V[(X_i - 2p_i)]$$

$$V[G_i] = [2p_i(1 - p_i)]^\alpha 2p_i(1 - p_i)$$

$$V[G_i] = [2p_i(1 - p_i)]^{\alpha+1}$$

LDAK assumptions implicit in $\hat{\pi}$

$$\hat{\pi}_{jk} = \frac{1}{\sum W_i} \sum_i W_i (x_{ij} - 2p_i)(x_{ik} - 2p_i) [2p_i(1 - p_i)]^\alpha$$

Assumptions

$$W_i = r_i w_i$$
$$w_i \cong \frac{1}{(1 + \sum r_{i,i'}^2)}$$

$$\alpha = -.25$$

Consequences

Where r_i is the imputation INFO score and w_i is the LD score. High LD SNPs receive less weight, and poorly imputed SNPs receive less weight.

Rarer SNPs (which tag rarer CVs) receive less (vs. GCTA) weight. This means the variance explained per SNP increases with MAF:

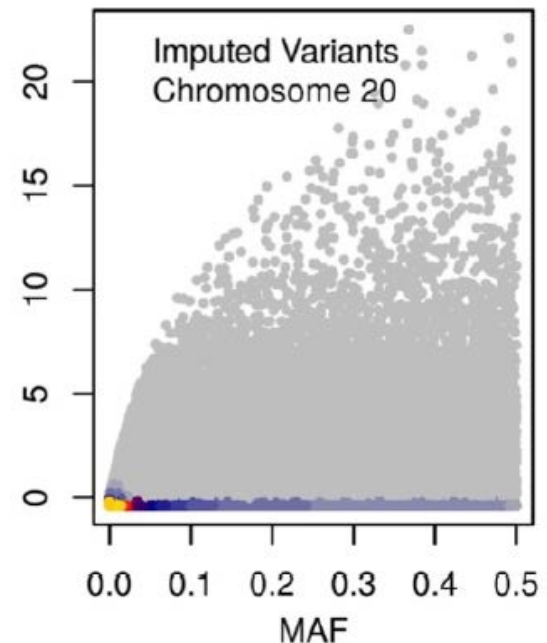
$$V[G_i] = [2p_i(1 - p_i)]^{\alpha+1}$$

Speed & Balding argued that LDAK weights are superior

- Common sense: Redundantly tagged CVs should not have higher effect sizes. Poorly imputed SNPs must tag CVs worse.
- Model Fit: log-likelihood from LDAK models was typically higher than log-likelihood from “GCTA” models
 - Moreover, h^2_{snp} 25-43% higher than GCTA models

Problems with LDAK approach

- Single GRM models depend heavily on assumptions and CV MAF matching the SNP MAF distribution
- Nothing about maximizing likelihoods ensures unbiasedness
- LD and imputation r^2 are highly positively related, but LDAK weights them oppositely. This gives extreme weight to a small number of unusual (well imputed, low LD, high MAF) SNPs



GREML-LDMS-I & -R



2018

Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits

LDMS-I

Luke M. Evans^{1*}, Rasool Tahmasbi¹, Scott I. Vrieze², Gonçalo R. Abecasis³, Sayantan Das³, Steven Gazal^{4,5}, Douglas W. Bjelland¹, Teresa R. de Candia¹, Haplotype Reference Consortium⁶, Michael E. Goddard^{7,8}, Benjamin M. Neale⁵, Jian Yang⁹, Peter M. Visscher⁹ and Matthew C. Keller^{1,10*}

nature
genetics

2015

Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index

LDMS-R

Jian Yang^{1,2,24}, Andrew Bakshi¹, Zhihong Zhu¹, Gibran Hemani^{1,3}, Anna A E Vinkhuyzen¹, Sang Hong Lee^{1,4}, Matthew R Robinson¹, John R B Perry⁵, Ilja M Nolte⁶, Jana V van Vliet-Ostaptchouk^{6,7}, Harold Snieder⁶, The LifeLines Cohort Study⁸, Tonu Esko⁹⁻¹², Lili Milani⁹, Reedik Mägi⁹, Andres Metspalu^{9,13}, Anders Hamsten¹⁴, Patrik K E Magnusson¹⁵, Nancy L Pedersen¹⁵, Erik Ingelsson^{16,17}, Nicole Soranzo^{18,19}, Matthew C Keller^{20,21}, Naomi R Wray¹, Michael E Goddard^{22,23} & Peter M Visscher^{1,2,24}

The LDMS Approach

- Single-GRM models are highly sensitive to assumptions about CV-SNP LD (e.g., that SNPs have same distribution as CVs) and CV effect size-MAF relationships. We don't want our estimates of genetic architecture to depend on our assumptions of genetic architecture.
- Moreover, even if we were to guess at these relationships perfectly for a trait, they are unlikely to hold across all traits.
- Akin to multiple regression, an alternative (LDMS) is to let the data tell us by fitting multiple GRMs, each with SNPs binned according to different MAF levels and LD levels
- Estimates associated with each GRM are free to soak up whatever variance is explained by those MAF/LD SNPs

LDMS justification

- Recall that $\hat{h}_{snp}^2 \cong h^2 \frac{\overline{r}_{MQ}^2}{\overline{r}_{MM}^2}$
- The range of MAF and range of LD will be smaller within a particular MAF/LD bin. Thus, for a given MAF/LD bin k of SNPs (M_k) and CVs (Q_k),

$$\frac{\overline{r}_{M_k Q_k}^2}{\overline{r}_{M_k M_k}^2} \rightarrow 1$$

and thus

$$\hat{h}_{snp,k}^2 \rightarrow h_k^2$$

LDMS-R vs. LDMS-I

- LDMS-R: Create 20 GRMs across 5 MAF bins ($< .001$, $.001-.01$, $.05-.1$, $.1-.25$, $.25-.5$) and 4 quartiles of LD scores, where SNPs take the average LD of SNPs in the surrounding $\sim 200\text{kb}$ region.
- However, SNPs with individually low LD that exist in regions of high LD explain more variation (Gazal et al, *Nature Genetics*, 2017)
- LDMS-I: exactly like LDMS-R except use each individual SNP's LD score for binning
- Because SEs tend to be $\sim 2.5\text{x}$ larger than single-GRM estimates, both require large sample sizes (e.g., $N > 30\text{k}$) and therefore large amounts of RAM (e.g., $> 100\text{ Gb}$)

RUNNING LDMS-I

Create LD quartiles

GCTA LD command:

gcta --bfile <path>/test
--ld-score-region 200
--out LD.txt

test.bed, test.bim, test.fam

```
SNP chr bp freq mean_rsq snp_num max_rsq ldscore_SNP ldscore_region  
rs4475691 1 836671 0.197698 0.000588093 999 0.216874 1.5875 2.75538  
rs28705211 1 890368 0.278112 0.000573876 999 0.216874 1.5733 2.75538  
rs9777703 1 918699 0.0301614 0.00131291 999 0.854464 2.31159 2.75538  
....
```

Create LD quartiles in R:

```
LD <- read.table("LD.txt",header=T)  
quants <- quantile(LD$ldscore_SNP)  
LD1 <- LD$SNP[LD$ldscore_SNP <= quants[2]]  
write.table(LD1,"snp_group1.txt",row.names=F,col.names=F,quote=F)  
<etc...>
```

Create GRMs in GCTA:

```
gcta --bfile <path>/test  
--extract snps_group1.txt  
--make-grm-bin  
--out GRM.1
```

Run LDMS-I using GCTA

COMMAND:

gcta

--mgrm-bin <path>/multi_GRMs.txt

--pheno <path>/test.phen

--covar <path>/test.covar

--qcovar <path>/test.qcovar

--reml --out Multi.SNPgrm

--thread-num 20

Text file

<path>/GRM.1
<path>/GRM.2
...
<path>/GRM.last

As before

You can use multiple cores;
make this as many cores as you
can spare

LDMS-I Output (3 GRM example)

TYPE: cat mgrm.randomCV.hsq

Source	Variance	SE
V(G1)	0.303900	0.184182
V(G2)	0.127654	0.309142
V(G3)	0.653199	0.328909
V(e)	0.926493	0.435653
Vp	2.011246	0.049641
V(G1)/Vp	0.151100	0.091277
V(G2)/Vp	0.063470	0.153765
V(G3)/Vp	0.324773	0.164408
logL	-2872.894	
N	3363	

$$h^2_{SNP} = 0.15 + 0.06 + 0.32 = 0.5391$$

GREML vs. LDAK vs. LDMS-I

Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits

Luke M. Evans^{1*}, Rasool Tahmasbi¹, Scott I. Vrieze², Gonçalo R. Abecasis³, Sayantan Das³, Steven Gazal^{4,5}, Douglas W. Bjelland¹, Teresa R. de Candia¹, Haplotype Reference Consortium⁶, Michael E. Goddard^{7,8}, Benjamin M. Neale³, Jian Yang⁹, Peter M. Visscher⁹ and Matthew C. Keller^{1,10*}



Luke Evans

- We hope it's useful as a guide for best practices and proper interpretation of \hat{h}^2_{SNP} .
- We simulated 16 genetic architectures, 3 levels of stratification, and 3 SNP types (array, imputed, WGS) in order to compare \hat{h}^2_{SNP} across 12 estimation methods (1728 different combos)
- Here I highlight just a few of what I think are the most important points

Overview of Simulation Approach

- Genotypes from real WGS data (n=8k). Choose 1K rare ($MAF < .0025$) or common ($MAF > .05$) CVs.
- Pull out SNPs on UKB array & impute
- Vary 2 CV effect size dimensions ($\lambda_i = u_i [2pq]^{\alpha/2}$):
 - λ -LD (via u_i)
 - λ -MAF (via α)
- Compare \hat{h}_{SNP}^2 to true h^2 ($=.50$) across 3 methods on imputed data
- Repeat this 100 times for different sets of CVs; look at mean (to get bias) and SD (to get SE) \hat{h}_{SNP}^2

Simulation of phenotypes

- CV effect size = $\lambda_i = u_i [2pq]^{\alpha/2}$

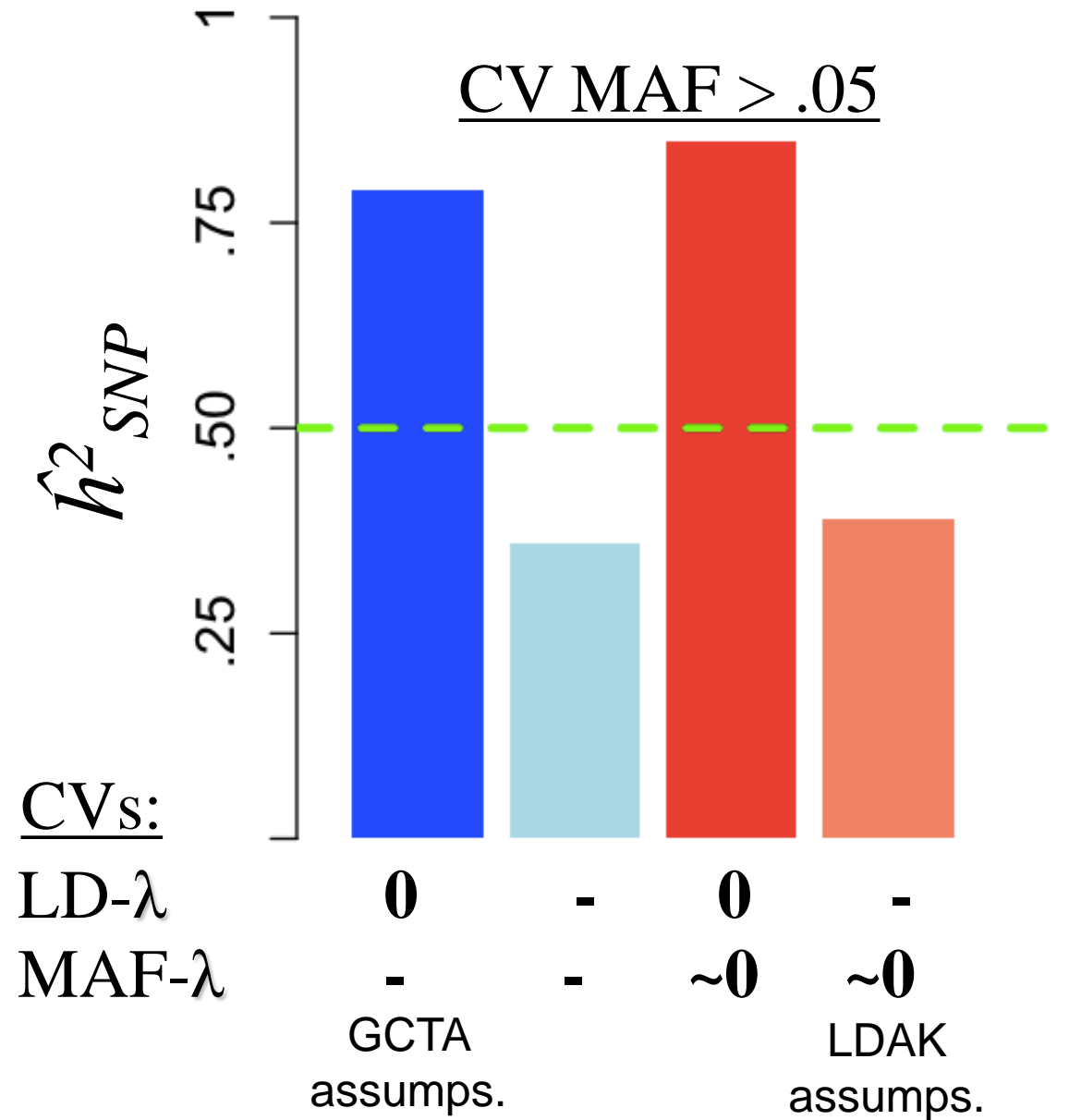
		<u>λ-LD relationship</u>	
		$u_i \sim N(0,1)$	or $u_i \sim N(0, w_i)$
<u>λ-MAF</u> $\alpha = -.25$ or $\alpha = -1$	negative (-)	none (0) Typical GCTA assumpt. V(Gi) invariant	negative (-) V(Gi) decreases w/ LD
	weak (~0)	V(Gi) increases w/ MAF	Typical LDAK assumpt. V(Gi) increases w/ MAF V(Gi) decreases w/ LD

- Breeding values = $A_j = \sum_i \lambda_i x_{ij}$
- Phenotype values = $P_j = A_j + E_j$

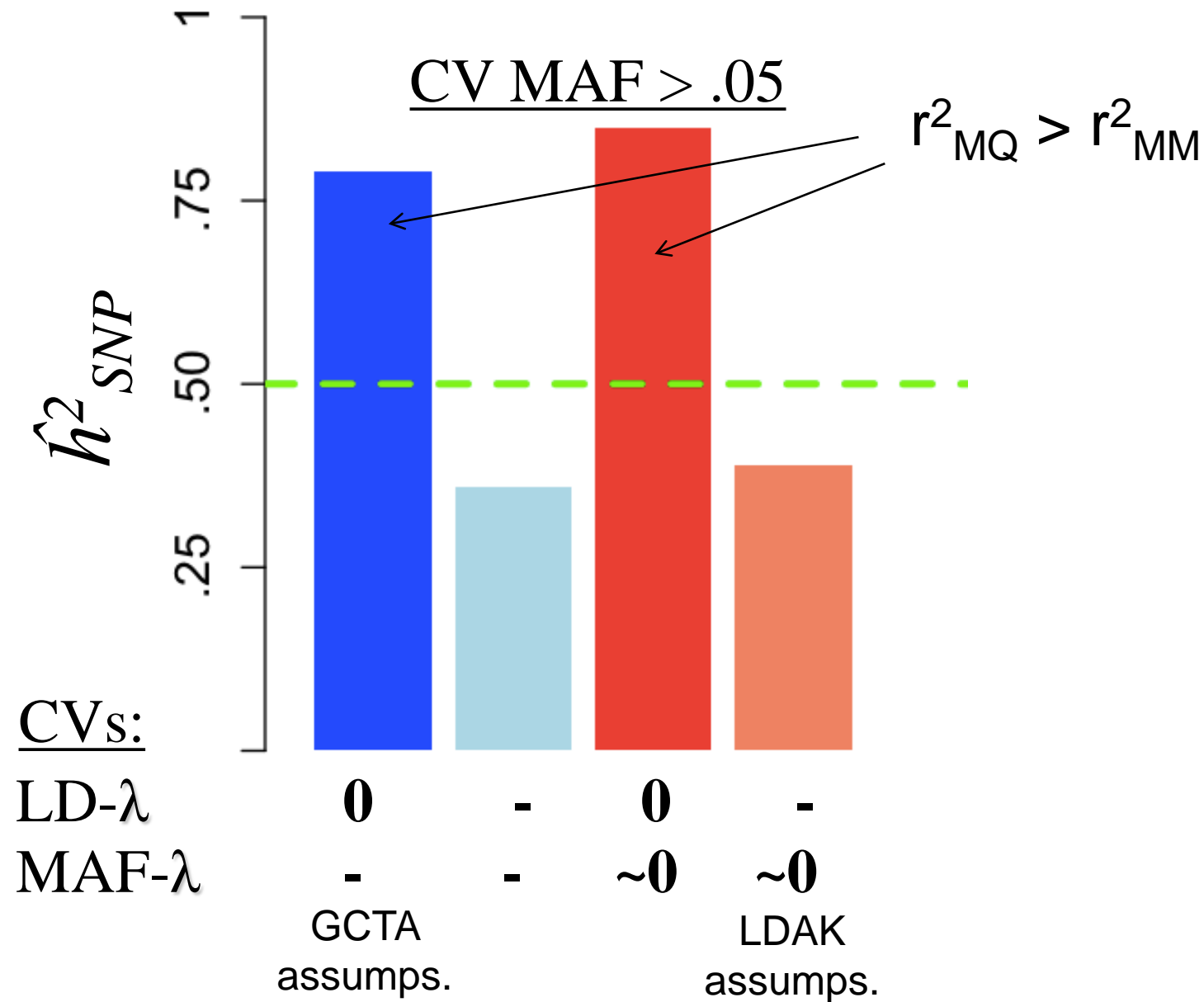
3 Estimation Methods Compared

- GREML-SC: predictor is a single GRM (aka, “GCTA approach”). GRM built as usual from all imputed SNPs with $MAC > 5$ & imputation $r^2 > .3$
- LDAK: predictor is a single GRM from imputed SNPs and weighted by LD and imputation r^2 .
- GREML-LDMS-I: predictors are $k = 8$ GRMs created by binning imputed SNPs into 2 individual LD by 4 MAF categories. Within each bin, GRMs built as usual. $\hat{h}^2_{SNP} = \Sigma(\hat{h}^2_{SNP_k})$

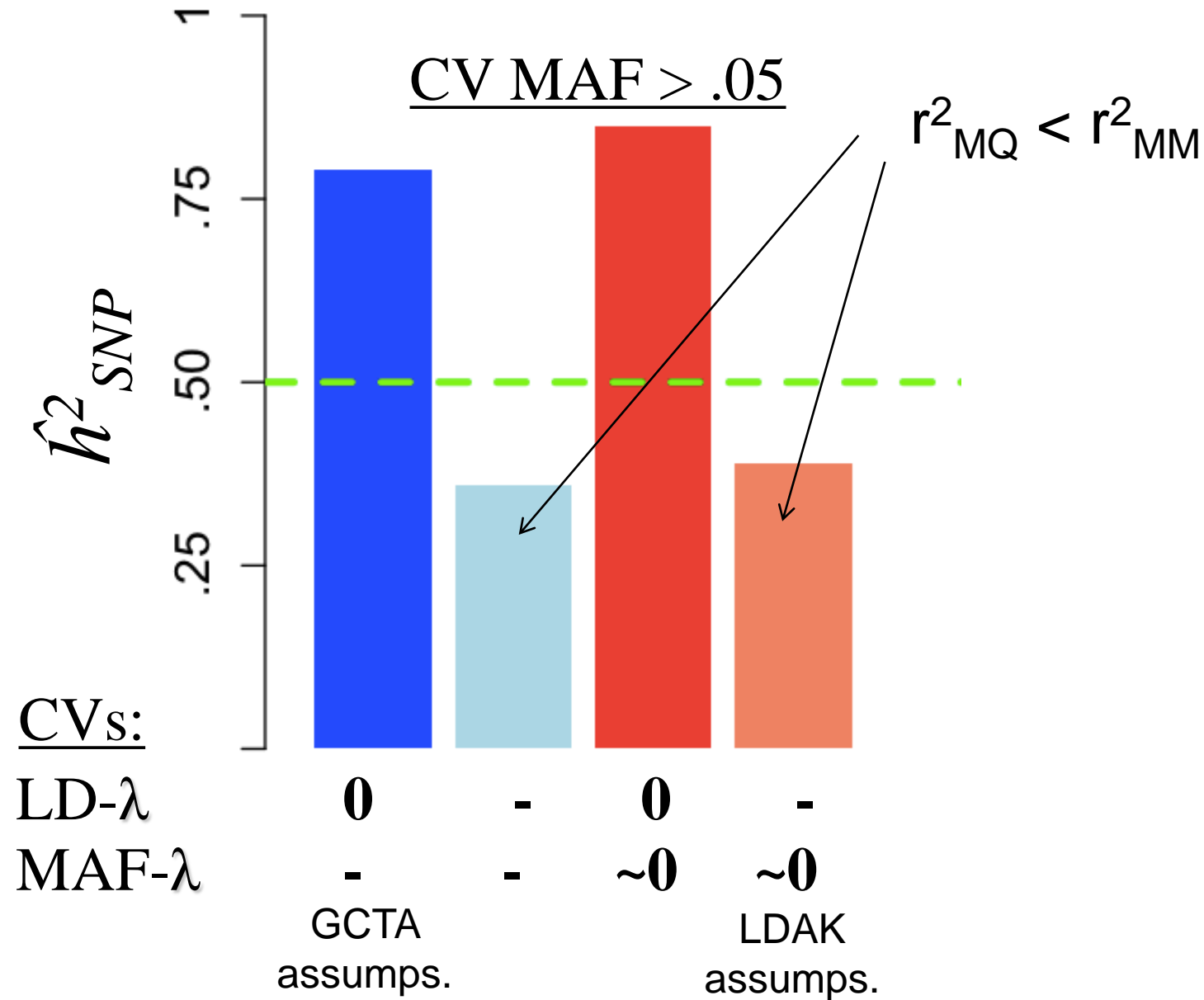
GREML-SC results



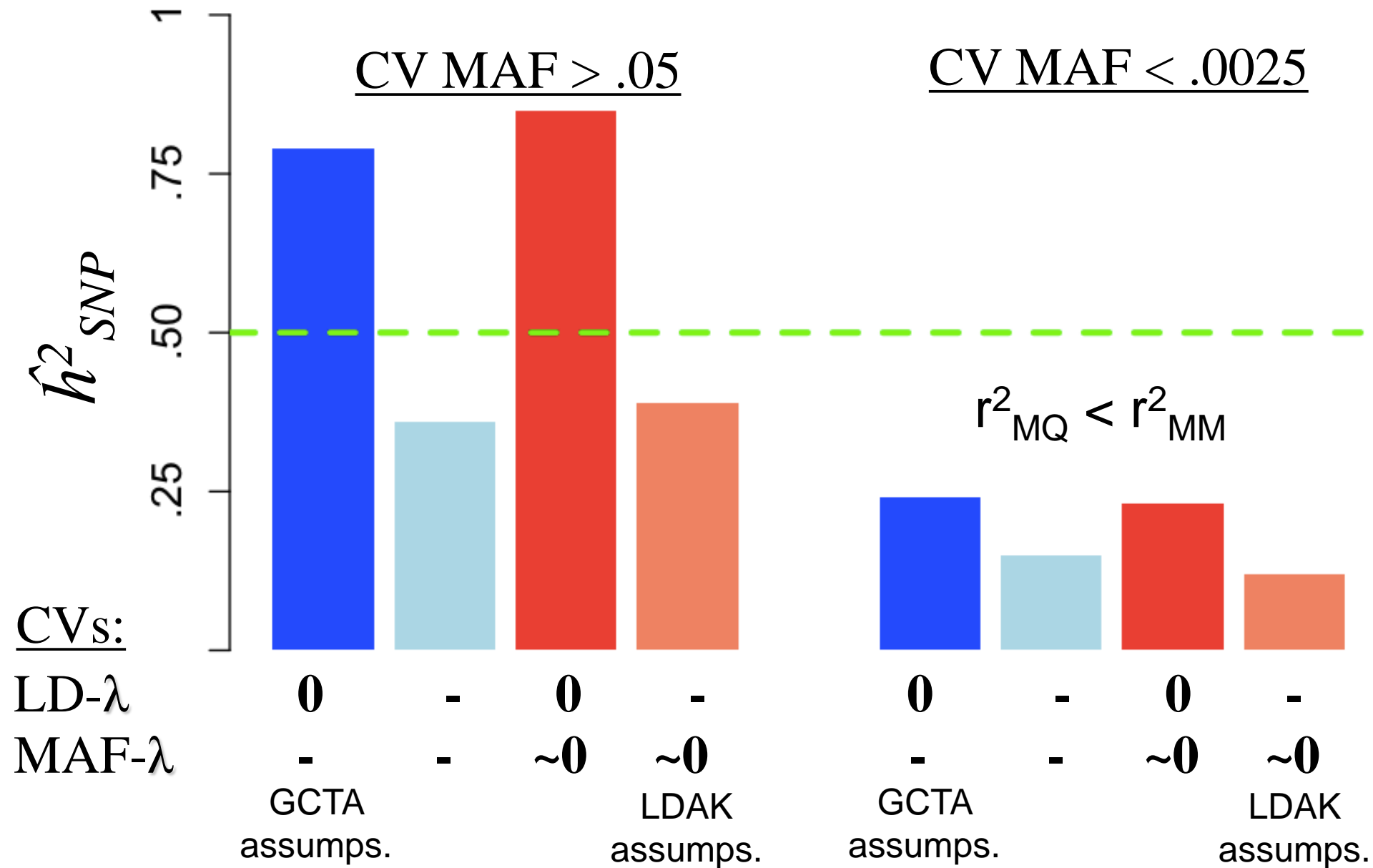
GREML-SC results



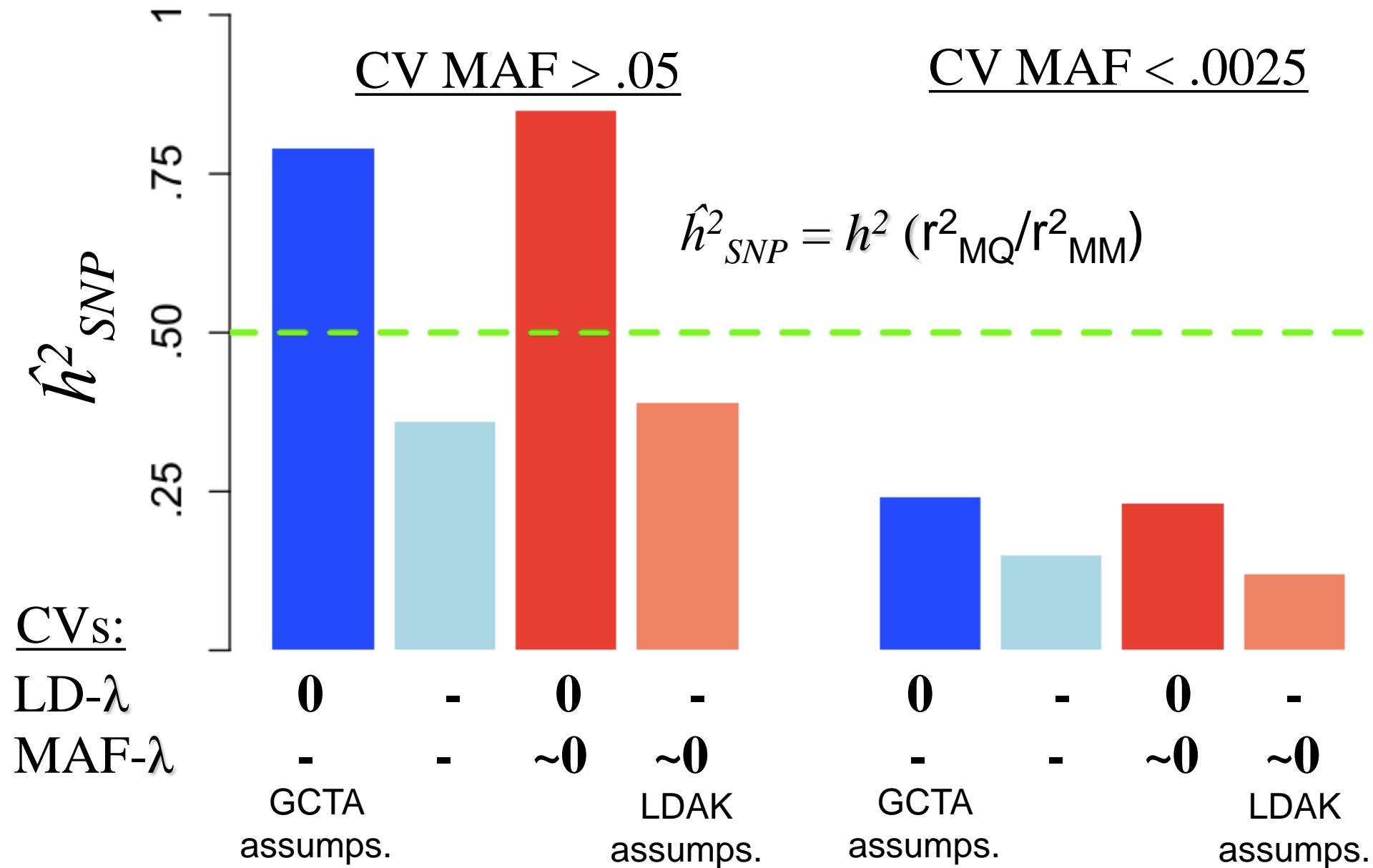
GREML-SC results



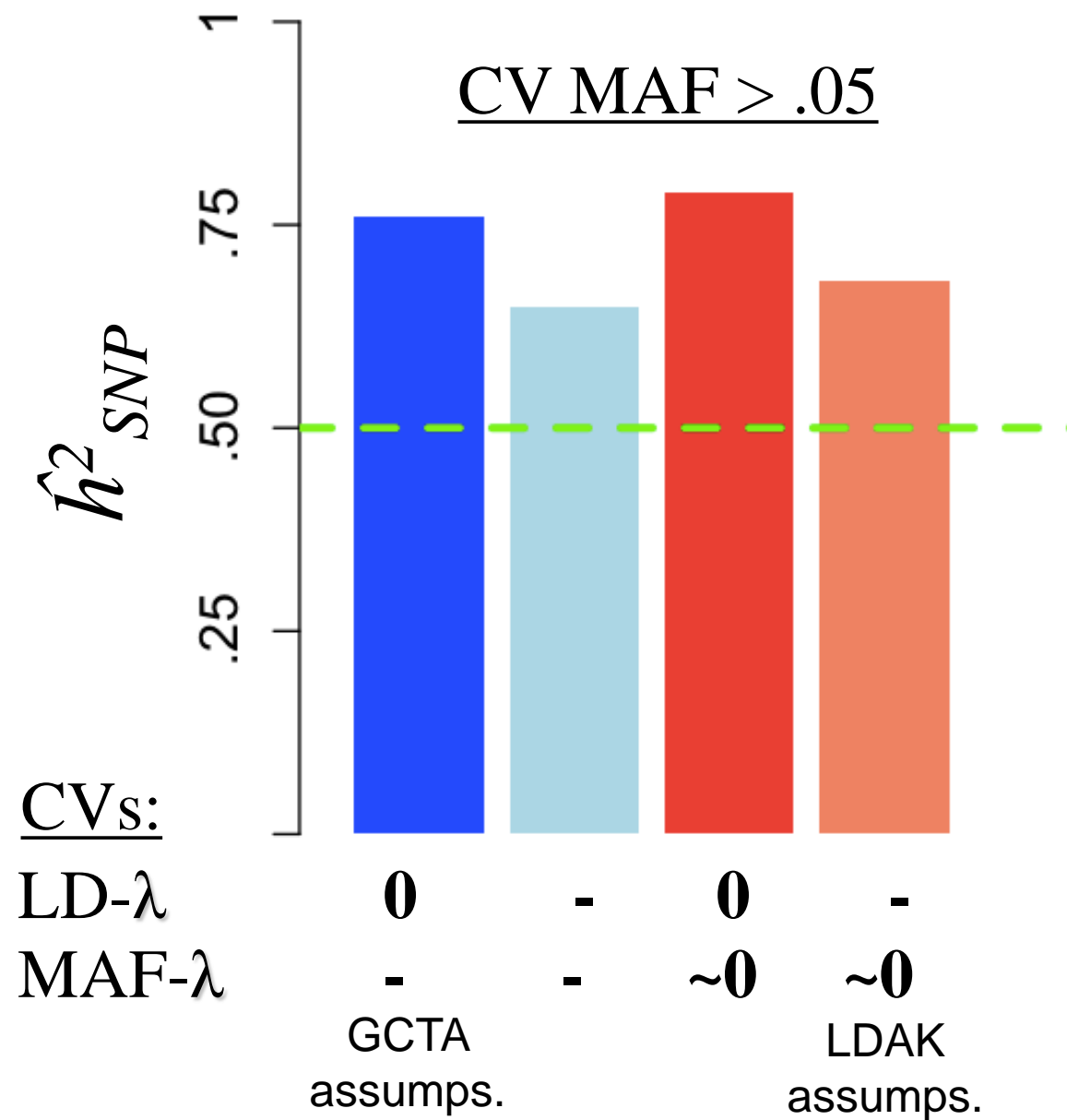
GREML-SC results



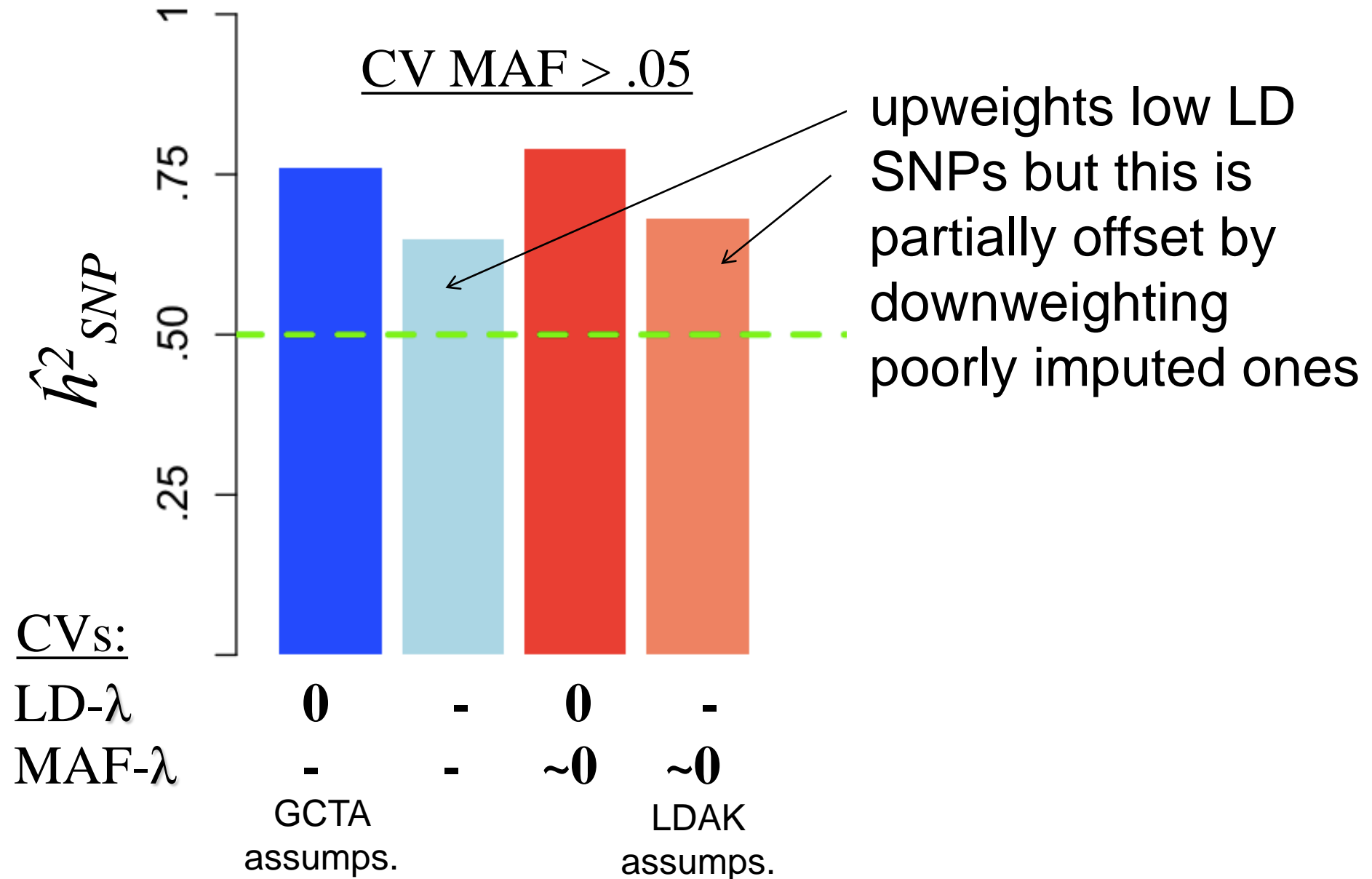
GREML-SC results



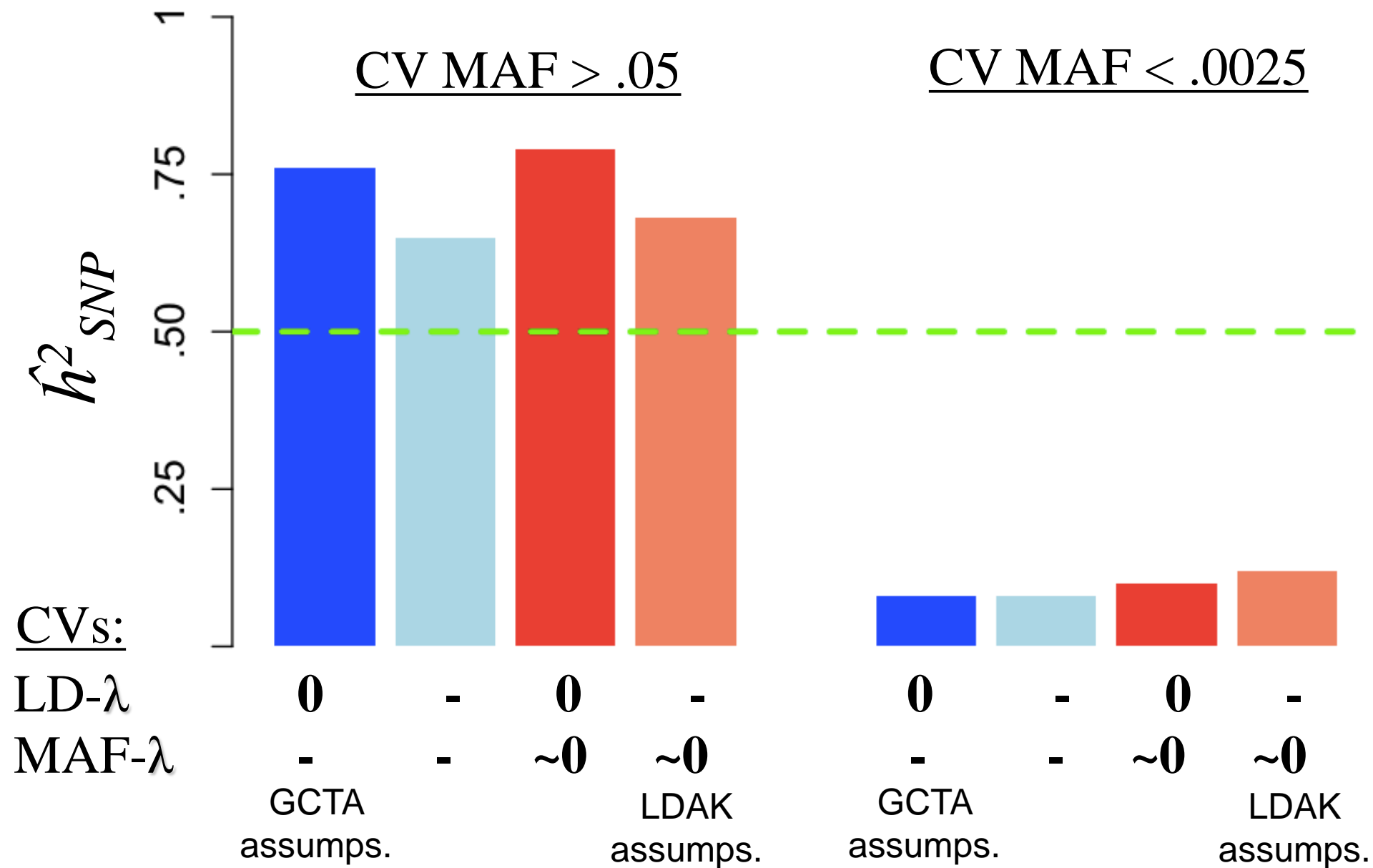
LDAK results



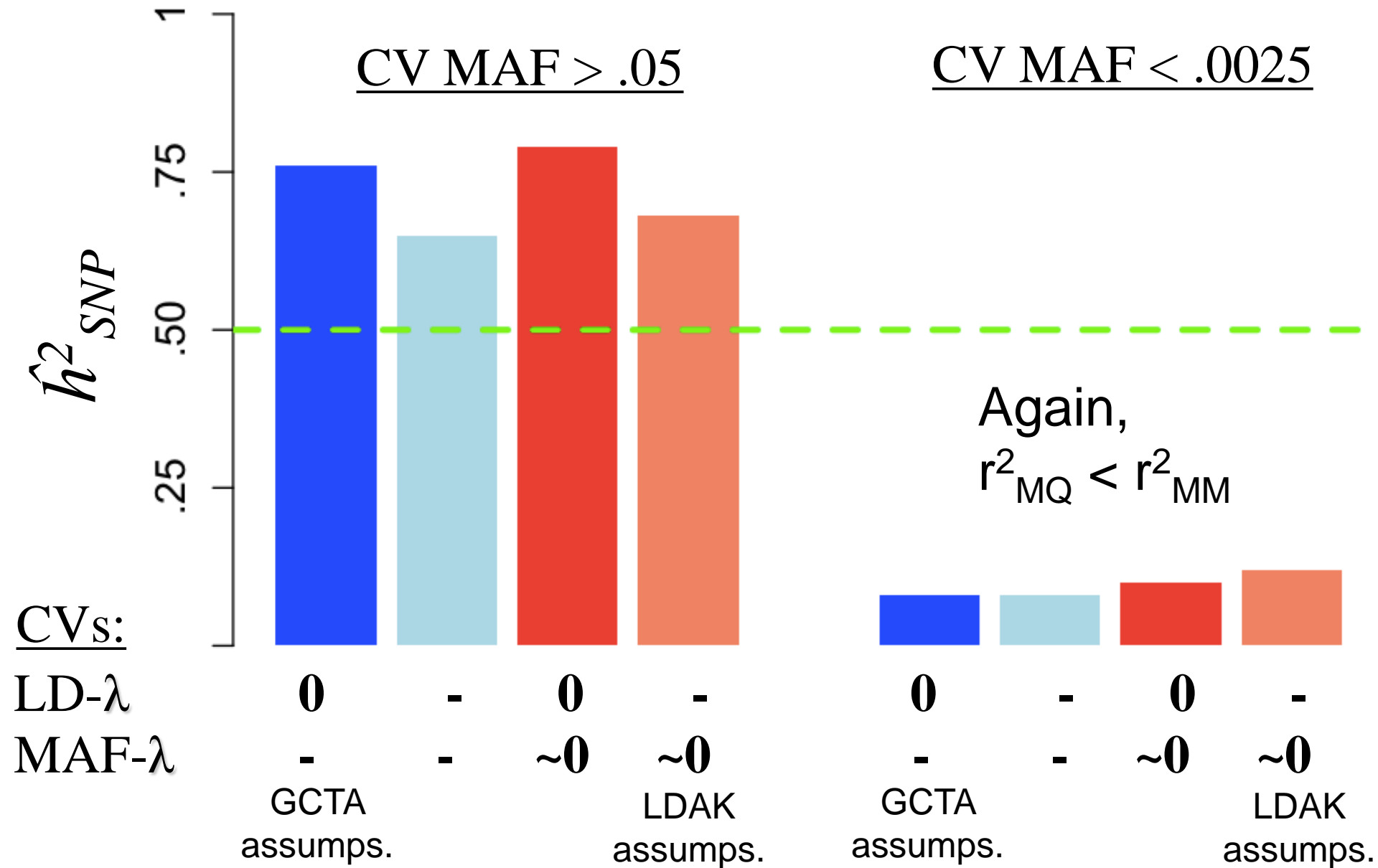
LDAK results



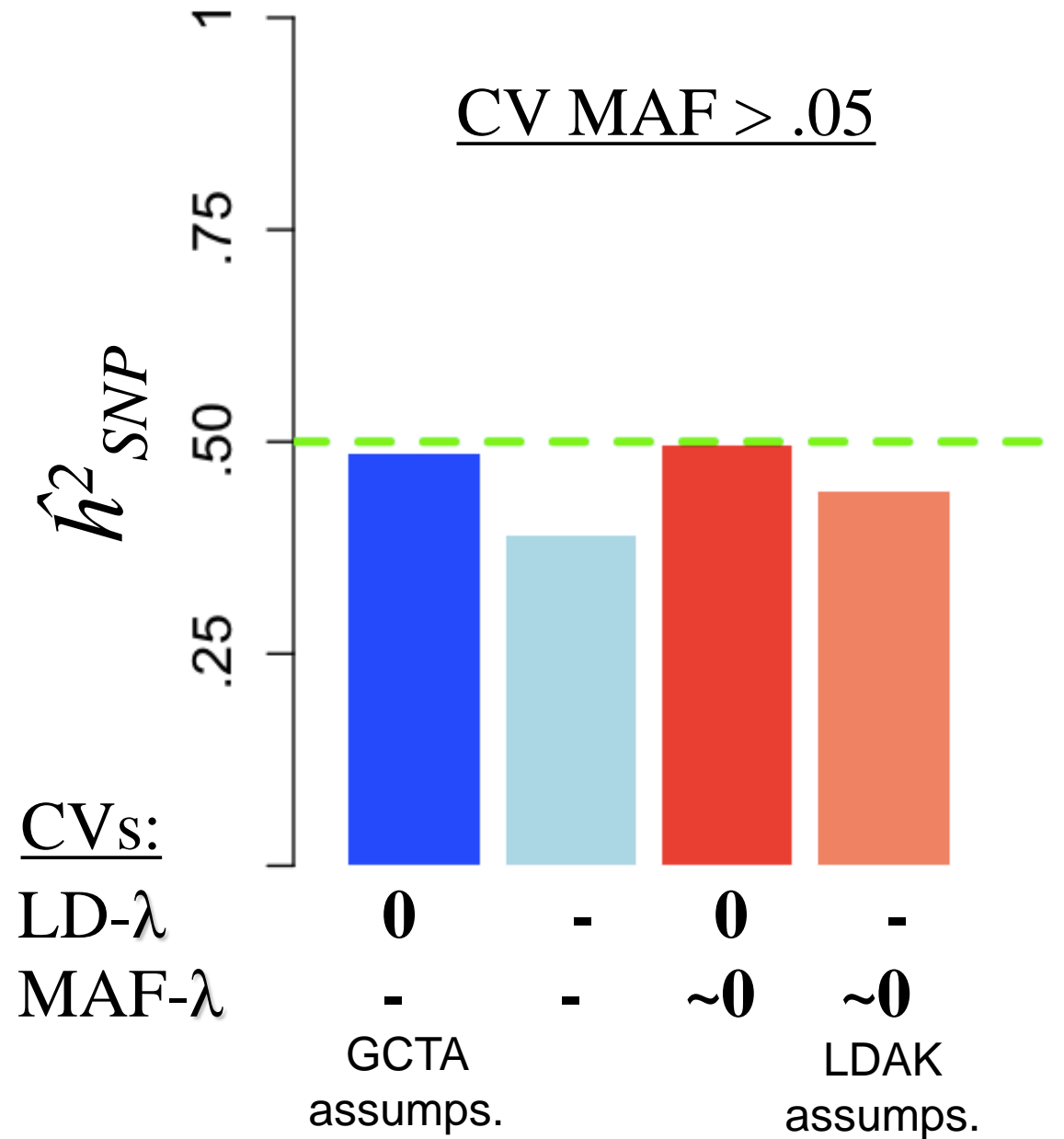
LDAK results



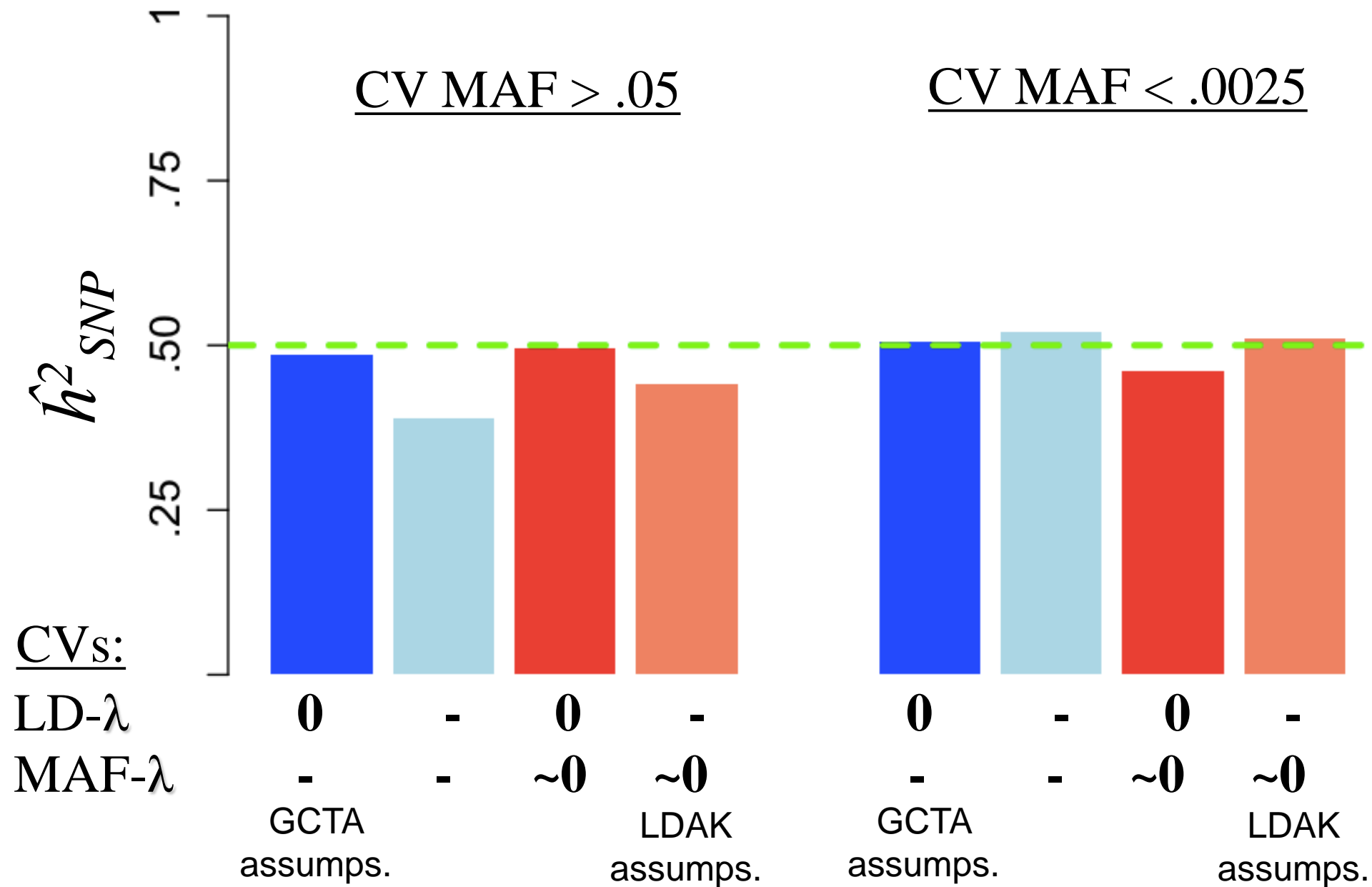
LDAK results



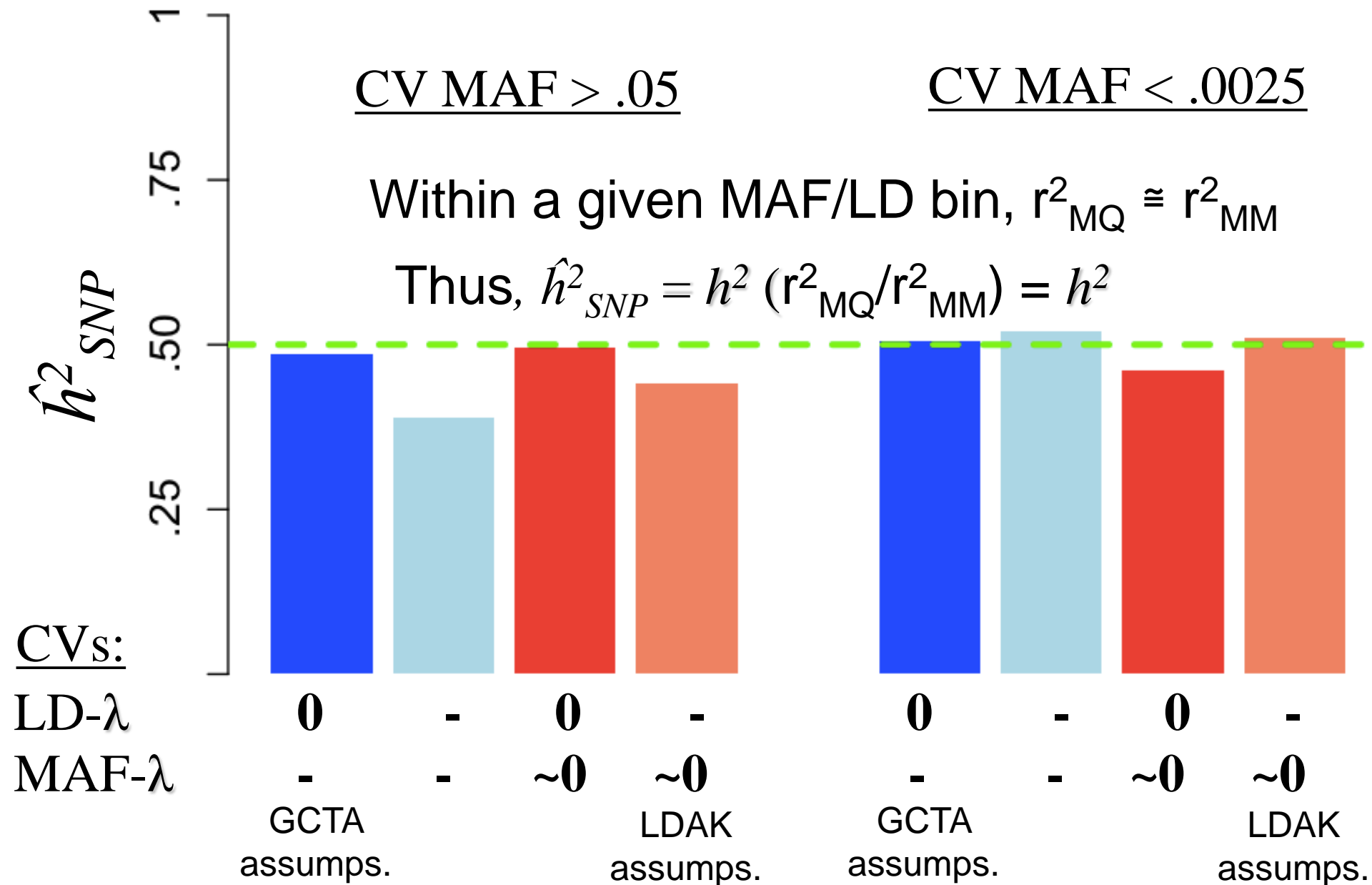
GREML-LDMS-I results



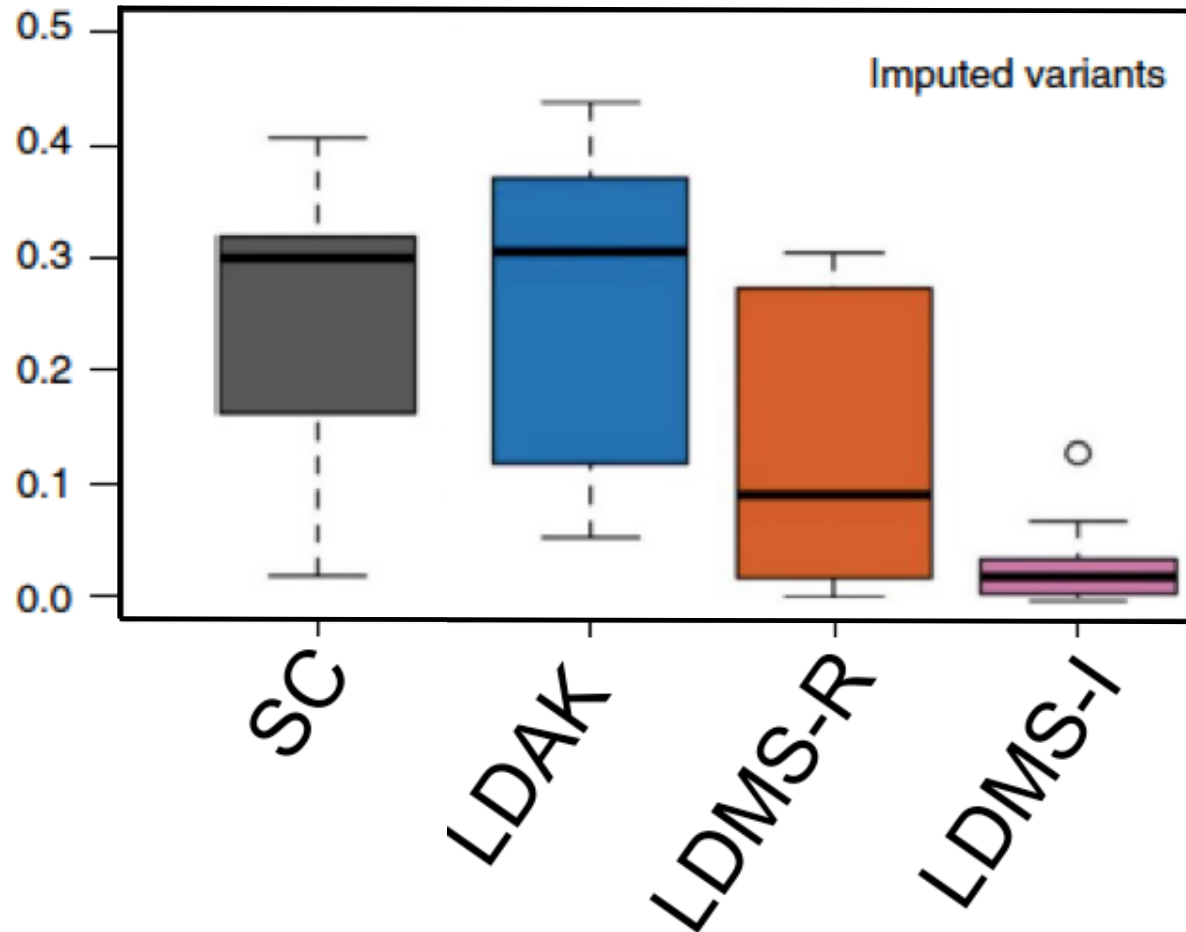
GREML-LDMS-I results



GREML-LDMS-I results



Absolute Bias Across 4 Methods and hundreds of genetic architectures



Regarding LD-Score regression

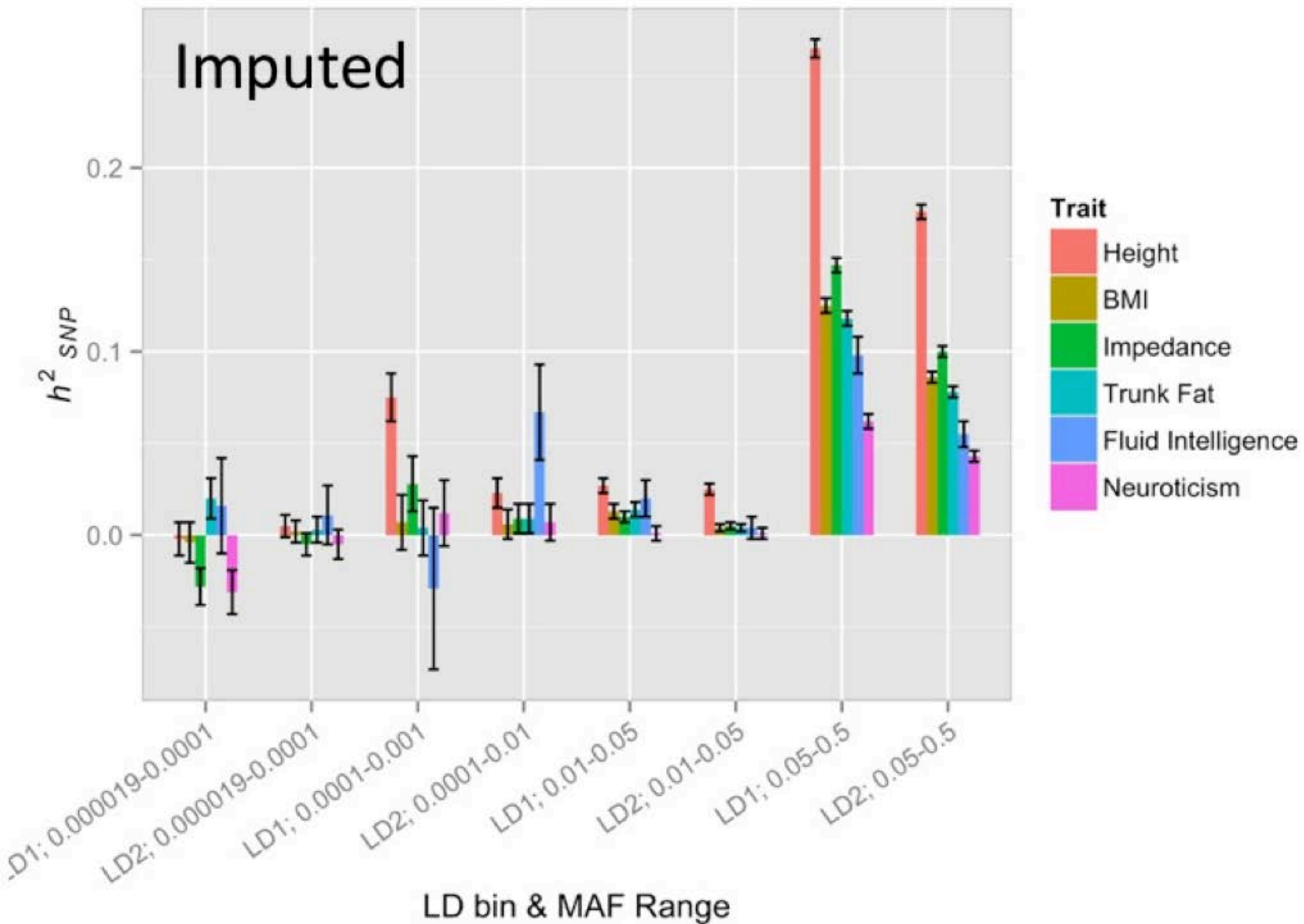
- LD-score regression is robust to stratification and sample overlap. However:
 - it cannot estimate h^2 due to rare CVs, even when using imputed/WGS data
 - it is sensitive to assumptions about LD- λ
 - should provide a lower-bound of \hat{h}^2_{SNP} from other methods
- So long as genetic covariance is affected in the same way as genetic variances, estimates of genetic correlations should be OK.

Summary

- With datasets imputed to large WGS reference panels, \hat{h}^2_{SNP} can estimate full h^2 . It's important that we have unbiased estimators to know the true h^2 and for comparison to twin/family estimates (o/w things will get really confusing).
- Single-GRM approaches (incl. GREML-SC (“GCTA”) and LDAK) are extremely sensitive to CV LD being similar to SNP LD across genome.
 - This is mostly influence by CV vs. SNP MAF, and also by assumptions of LD- λ relationship. MAF- λ less so.
- Binning SNPs by LD & MAF provides ~ unbiased estimates for the CVs tagged by SNPs used in analysis.
 - Even on well-imputed data, you'll still get an underestimate due to extremely rare variants

REAL TRAITS

LDMS-I on UKB phenotypes



STRATIFICATION & LONG-RANGE LD

Chance allele frequency differences b/w populations can induce long-range LD in stratified samples

Population 1

	A	a	
B	.54	.36	.9
b	.06	.04	.1
	.6	.4	

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} = 0$$

Population 2

	A	a	
B	.03	.27	.3
b	.07	.63	.7
	.1	.9	

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} = 0$$

Chance allele frequency differences b/w populations can induce long-range LD in stratified samples

Population 1

	A	a	
B	.54	.36	.9
b	.06	.04	.1
	.6	.4	

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} = 0$$

Population 2

	A	a	
B	.03	.27	.3
b	.07	.63	.7
	.1	.9	

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} = 0$$

Stratified Population

	A	a	
B	.285	.315	.6
b	.065	.335	.4
	.35	.65	

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} = .10$$

However, such “stratification-LD” is typically very small for pairs of common SNPs

Population 1

	A	a	
B	.54	.36	.9
b	.06	.04	.1
	.6	.4	

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} = 0$$

Population 2

	A	a	
B	.44	.36	.8
b	.11	.09	.2
	.55	.45	

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} = 0$$

Stratified Population

	A	a	
B	.490	.360	.85
b	.085	.065	.15
	.575	.425	

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} = .00005$$

But higher b/w rare (often ~ private) SNPs
and common ancestry-informative SNPs

Population 1

	A	a	
B	.003	.897	.9
b	.00	.099	.1
	.003	.997	

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} = 0$$

Population 2

	A	a	
B	.04	.76	.8
b	.01	.19	.2
	.05	.95	

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} = 0$$

Stratified Population

	A	a	
B	.021	.829	.85
b	.005	.145	.15
	.027	.973	

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} = .0004$$

Effects of stratification on r^2_{QM}/r^2_{MM}

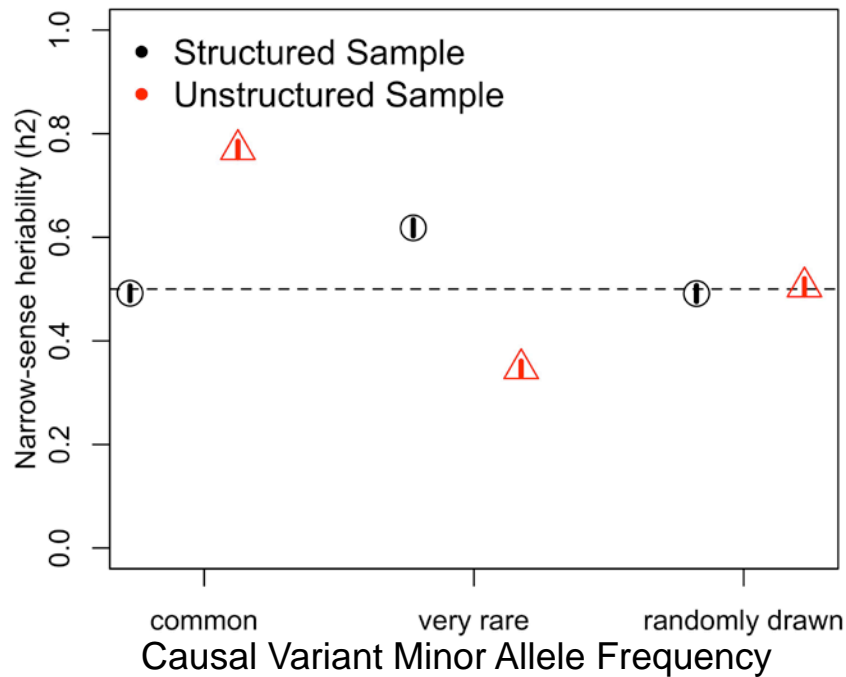
- In general, stratification inflates long-range r^2 between SNPs. However, within a given MAF bin, the ratio of r^2_{QM}/r^2_{MM} is ~ 1 because SNP-SNP & SNP-CV LDs are inflated similarly.

Effects of stratification on r^2_{QM}/r^2_{MM}

- In general, stratification inflates long-range r^2 between SNPs. However, within a given MAF bin, the ratio of r^2_{QM}/r^2_{MM} is ~ 1 because SNP-SNP & SNP-CV LDs are inflated similarly.
- However, across CVs and SNPs of different MAF, stratification induces differences in r^2_{QM} & r^2_{MM} . We observed:
 - For rare CVs, $r^2_{QM}/r^2_{MM} > 1$. Rare (ancestry specific) CVs are tagged by every common SNP that differs in allele frequency across ancestry (note $r^2_{QM}/r^2_{MM} < 1$ in unstratified samples).
 - For very common CVs, $r^2_{QM}/r^2_{MM} \sim 1$. Very common CVs tend to have smaller MAF differences, and therefore less LD with common SNPs than typical between SNPs (note $r^2_{QM}/r^2_{MM} > 1$ in unstratified samples).

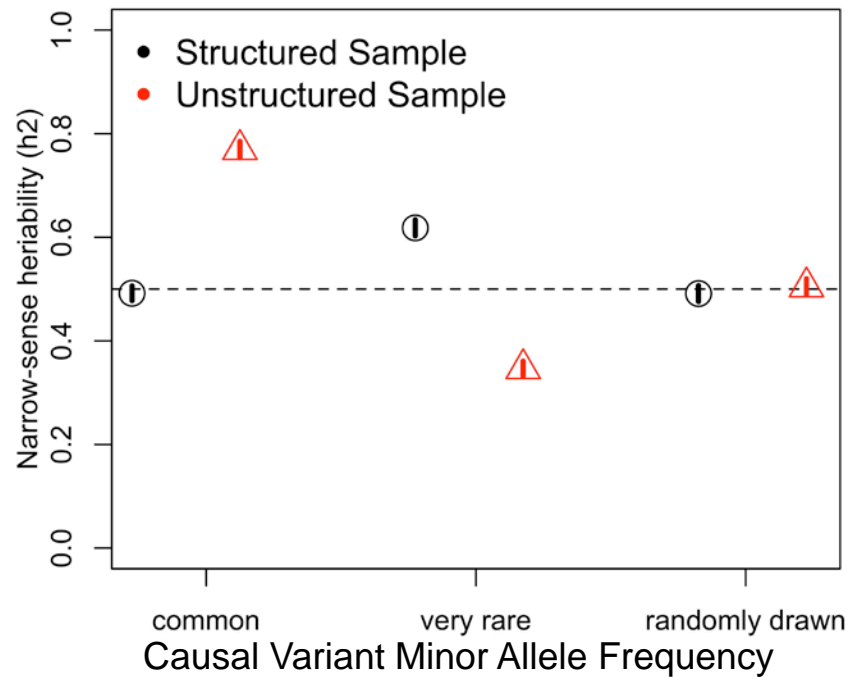
This led to an opposite pattern of bias in stratified ("structured") samples when using single GRM GREML

Single GRM using WGS

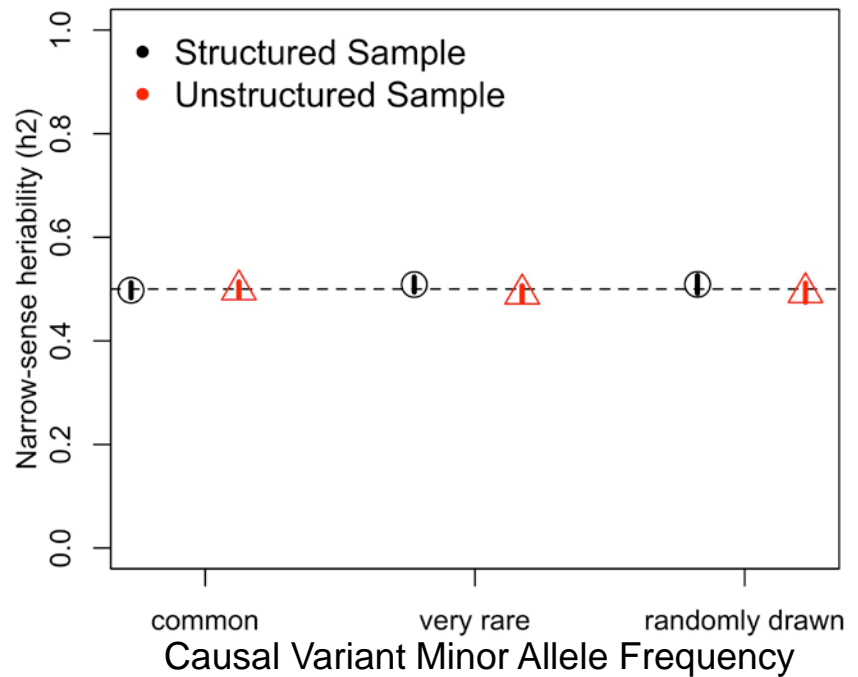


Which once again was corrected by using LDMS GREML

Single GRM using WGS



MAF-stratified GRMs using WGS



ASSORTATIVE MATING

Positive primary phenotypic assortative mating (AM)

- AM: Assortment between mates leading to a correlation between phenotypic (and hence genetic) scores. Often conceptualized as mate choice based on similarity.
- Induces long-range (across chromosome) “directional” LD (δ) b/w CVs
 - δ = covariance among CV effects; under positive AM, $E[\delta] > 0$; allelic effects in the same direction.
- Directional LD increases true V_G & h^2 in the population.
 - This occurs for same reason the variance of a sum of positively correlated $X_i >$ variance of sum of independent X_i
 - For polygenic traits, the vast majority (>99%) of this increase is due to δ between different CVs, not to δ within CVs (homozygosity)

AM effects on $\hat{\pi}_{jk}$

- Assortment has ~ no influence on $\hat{\pi}_{jk}$
- Recall that $E[Z_j Z_k | \hat{\pi}_{jk}] = h^2 \hat{\pi}_{jk}$
- However, this is much different than the reverse conditional*: $E[\hat{\pi}_{jk} | Z_j Z_k] = \frac{r h^2}{m} < \frac{1}{m}$

where r is the mate correlation and m is the # CVs

- This is because δ *between* CVs, the major factor influencing h^2 , plays no role in $\hat{\pi}_{jk}$ (or means in general)

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

However, AM does bias h^2_{snp} estimates

- AM typically leads to upward bias in estimates of equilibrium h^2_{snp}
- Occurs because AM creates positive covariances between CVs and these are correctly reflected in phenotypic covariances between individuals (product of means) but poorly reflected in pihat matrix (mean of products).
- Thus, variance of pihats is too small. Underestimated variance in a predictor leads to overestimates of the coefficients associated with that predictor.
- We derived this bias algebraically in HE regression estimates and confirmed it in simulation.
- REML also upwardly biased, but bias depends on ratio N/m .

Parameter h^2_{snp}

- Define parameter h^2_{snp} : proportion of phenotypic variance tagged by SNPs, accounting for their inter-correlations
- Equilibrium h^2_{snp} : R^2 from linear model $Z \sim X_1 + X_2 + \dots X_m$ for all m SNPs fit simultaneously as $n \rightarrow \infty$
- The parameter depends on how well CVs are tagged by SNPs (e.g., SNP density). Thus, it depends on the SNP chip and the population it is estimated in.

HE regression estimate of h^2_{snp}

$$E[Z_i Z_j] = COV(Z_i, Z_j)$$

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

$$\hat{\beta}_1 = \frac{COV(Z_i Z_j, \hat{\pi}_{ij})}{V(\hat{\pi}_{ij})} = \hat{h}^2_{snp}$$

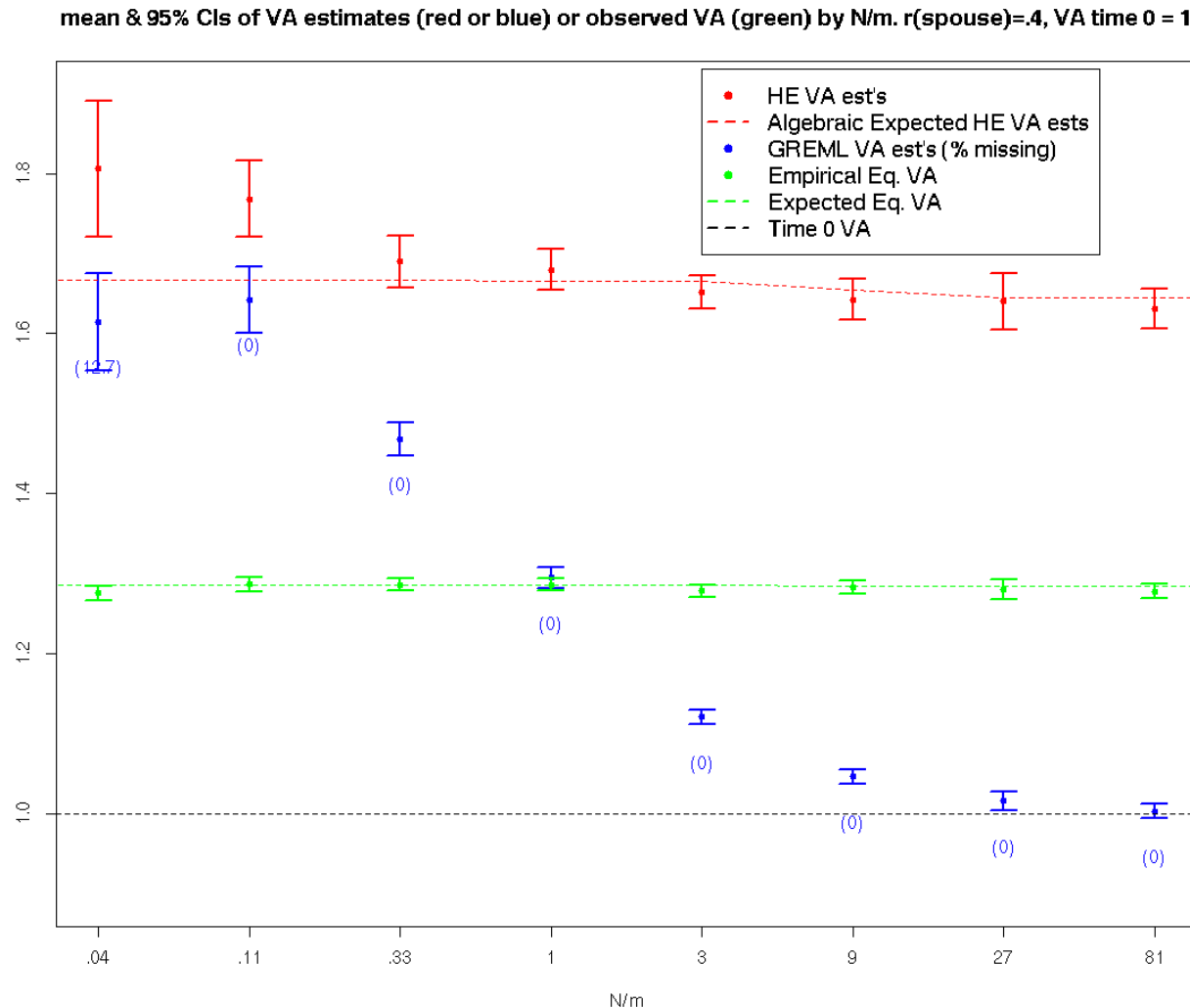
HE regression estimate of h^2_{snp}

$$E[Z_i Z_j] = COV(Z_i, Z_j)$$

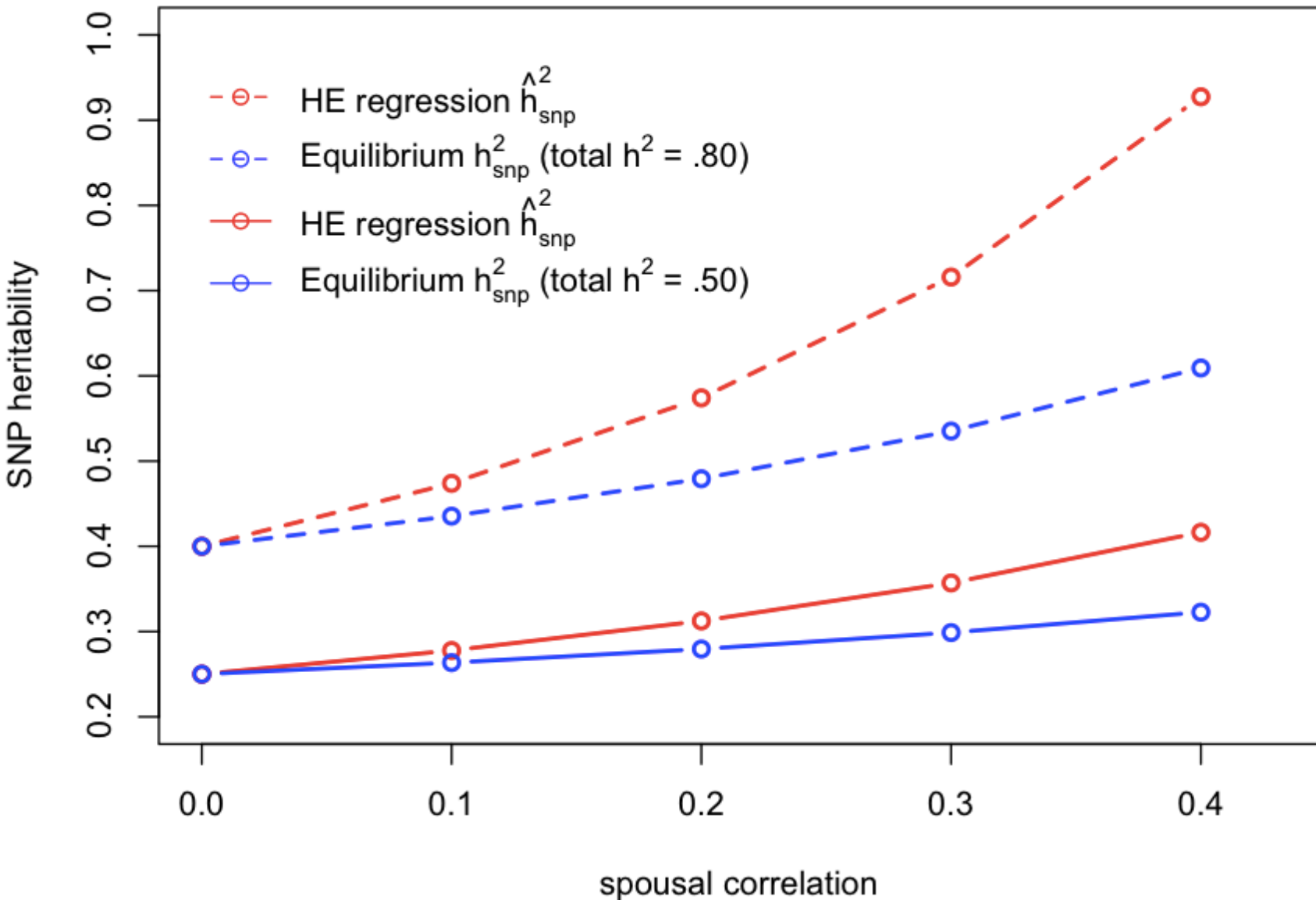
$$E[Z_i Z_j \mid \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij} \cong \frac{h^2_{snp}}{M} + 2M\delta$$

$$\hat{\beta}_1 = \frac{COV(Z_i Z_j, \hat{\pi}_{ij})}{V(\hat{\pi}_{ij})} = \hat{h}^2_{snp} \cong \frac{1}{M} + \left(\frac{2M\delta}{h^2} \right)^2$$

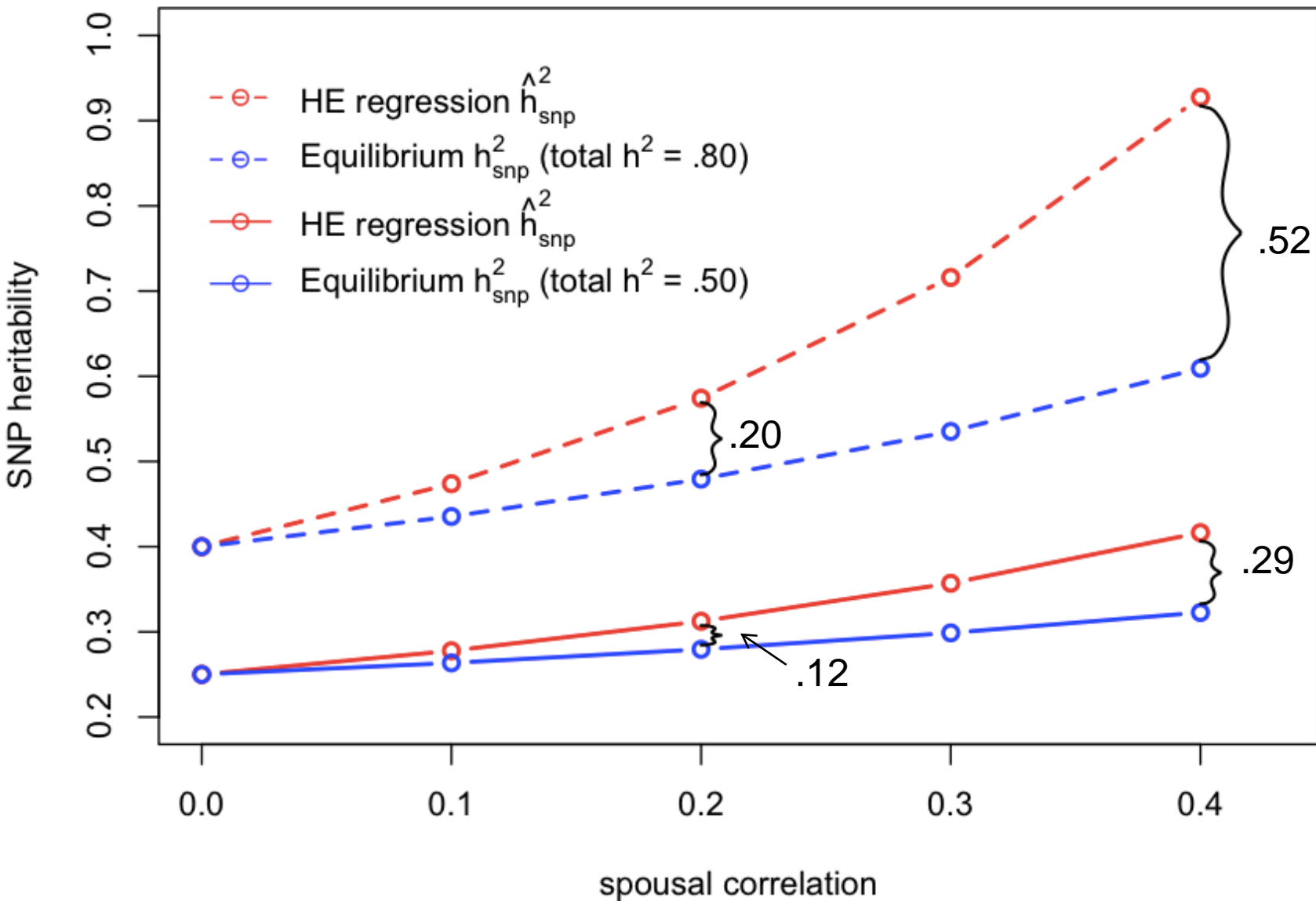
We don't predict HE estimates to change as a function of m or N . GREML estimates are clearly a function of N/m , which occurs because when $N \gg m$, the effects of each SNP are separable.



Predicted h^2_{snp} biases assuming SNPs tag 50% of true VA



Predicted h^2_{snp} biases assuming SNPs tag 50% of true VA



Potential degree of over-estimation for various traits

Trait	r(spouse)	h^2_{ETFD} from literature	h^2_{snp} from literature	corrected h^2_{snp}	% Over- estimated
Extraversion	.01	.23	.15	.15	0
Neuroticism	.08	.24	.16	.155	.03
Height	.20	.70	.45	.39	.15
IQ	.35	.62	.35	.28	.25
Political Pref.	.48	.26	.18	.15	.20

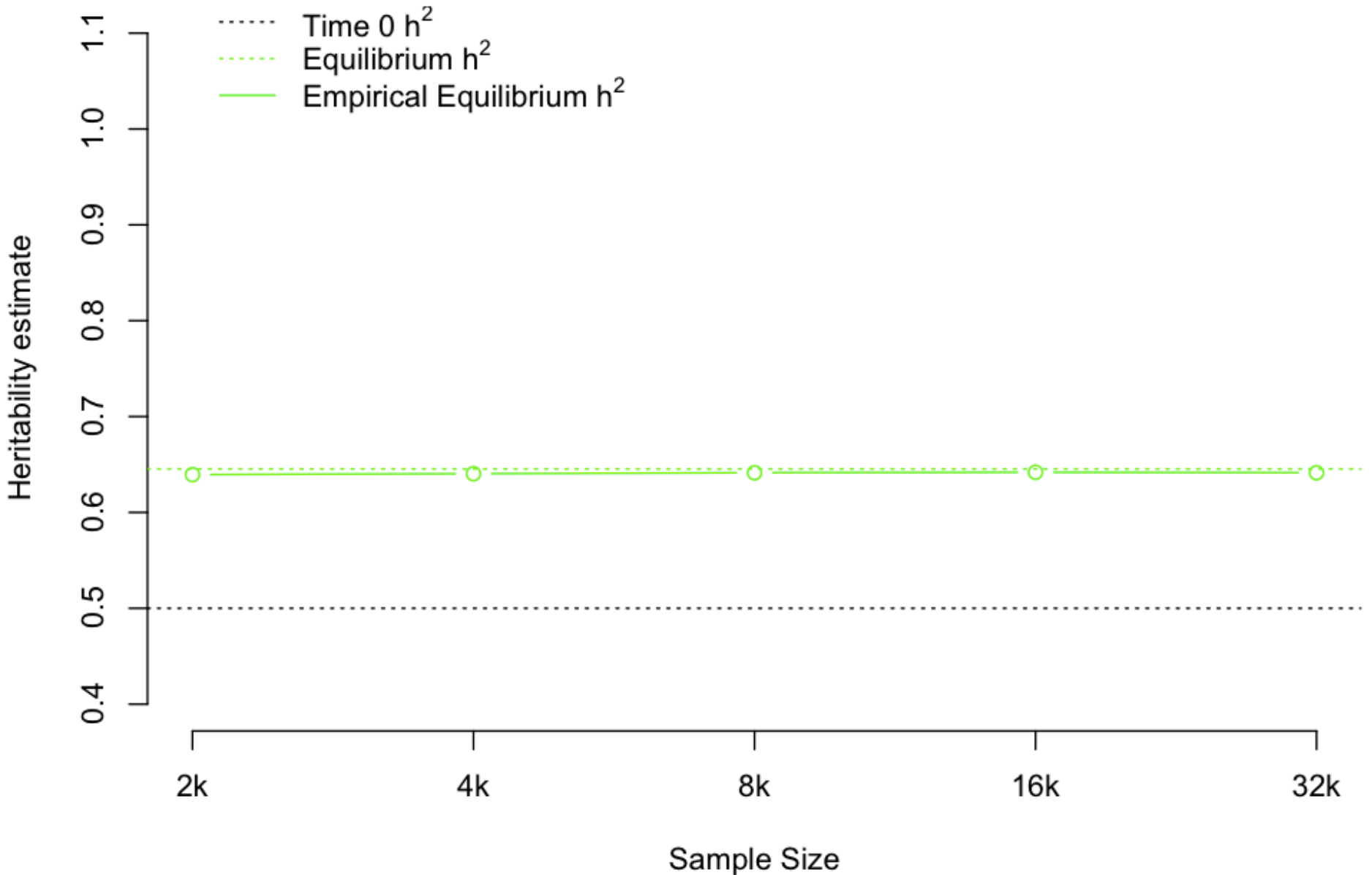
Simulations



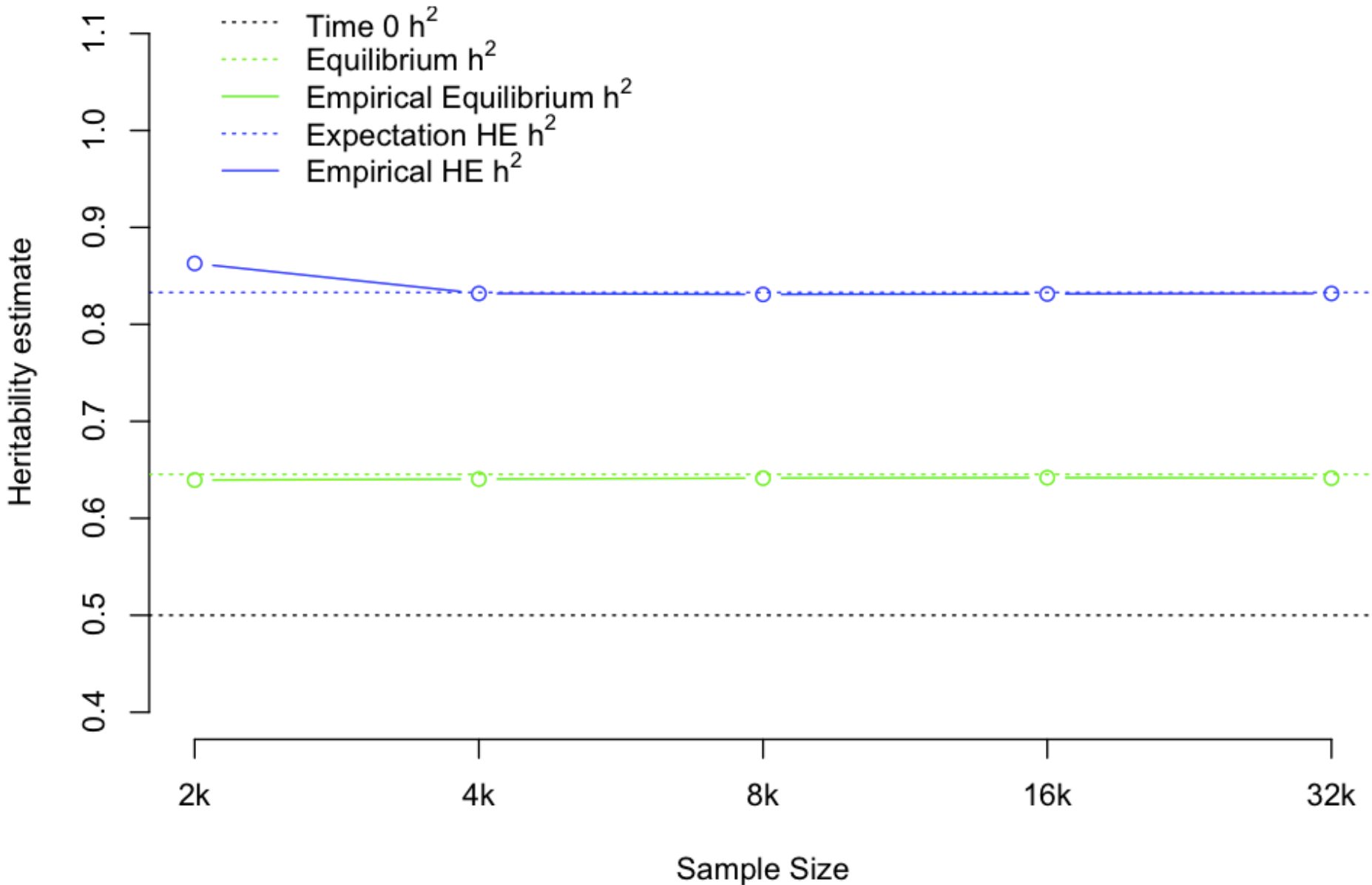
Rasool Tahmasbi

- Simulated populations under AM using GeneEvolve (Tahmasbi & Keller, 2016).
- CVs: 1000
- Heritability: 0.5
- Relative pruning: $>.05$
- Spousal phenotypic correlation: .4
- Took mean of 100 iterations

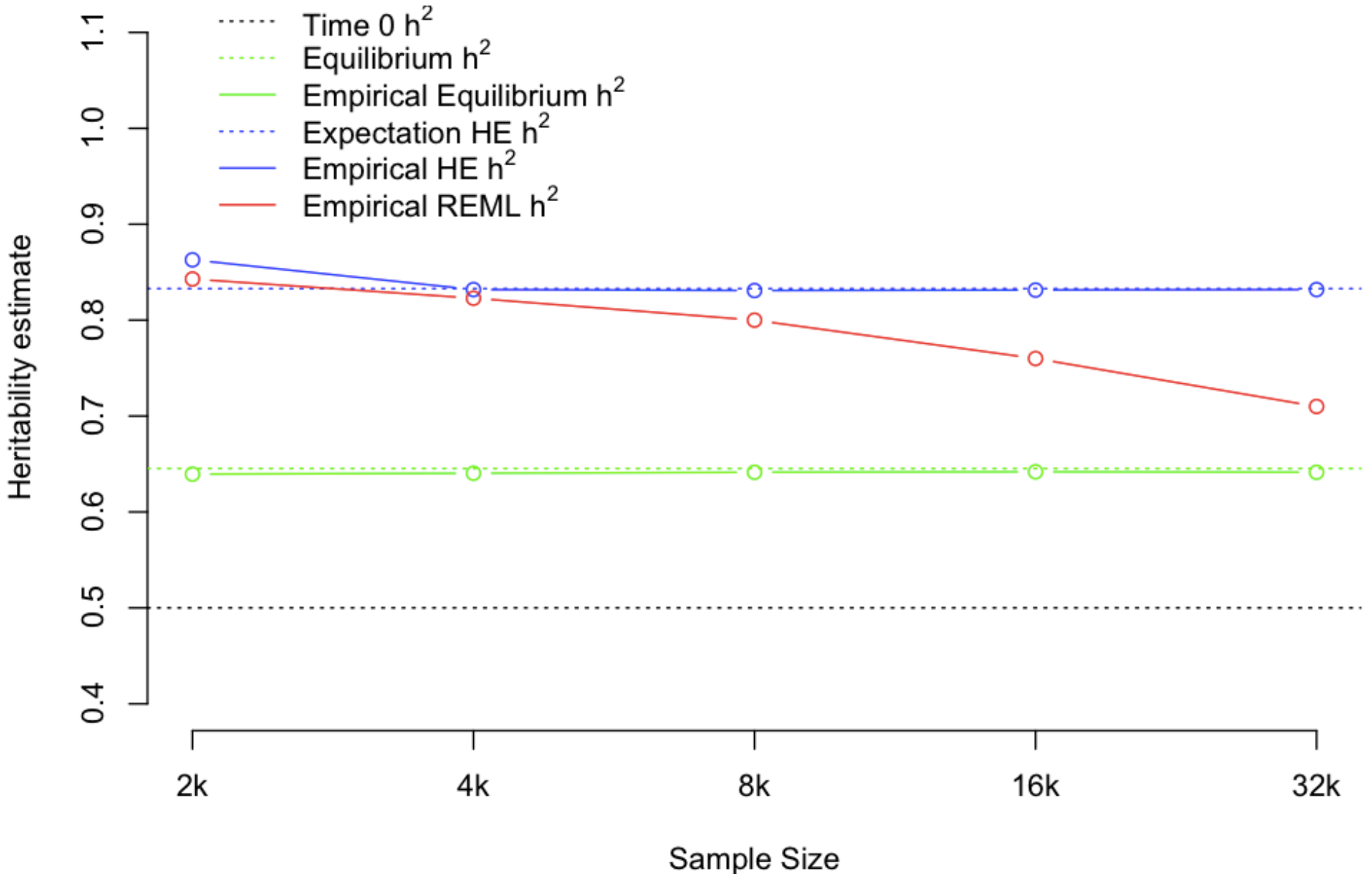
GeneEvolve Simulation Results



GeneEvolve Simulation Results



GeneEvolve Simulation Results

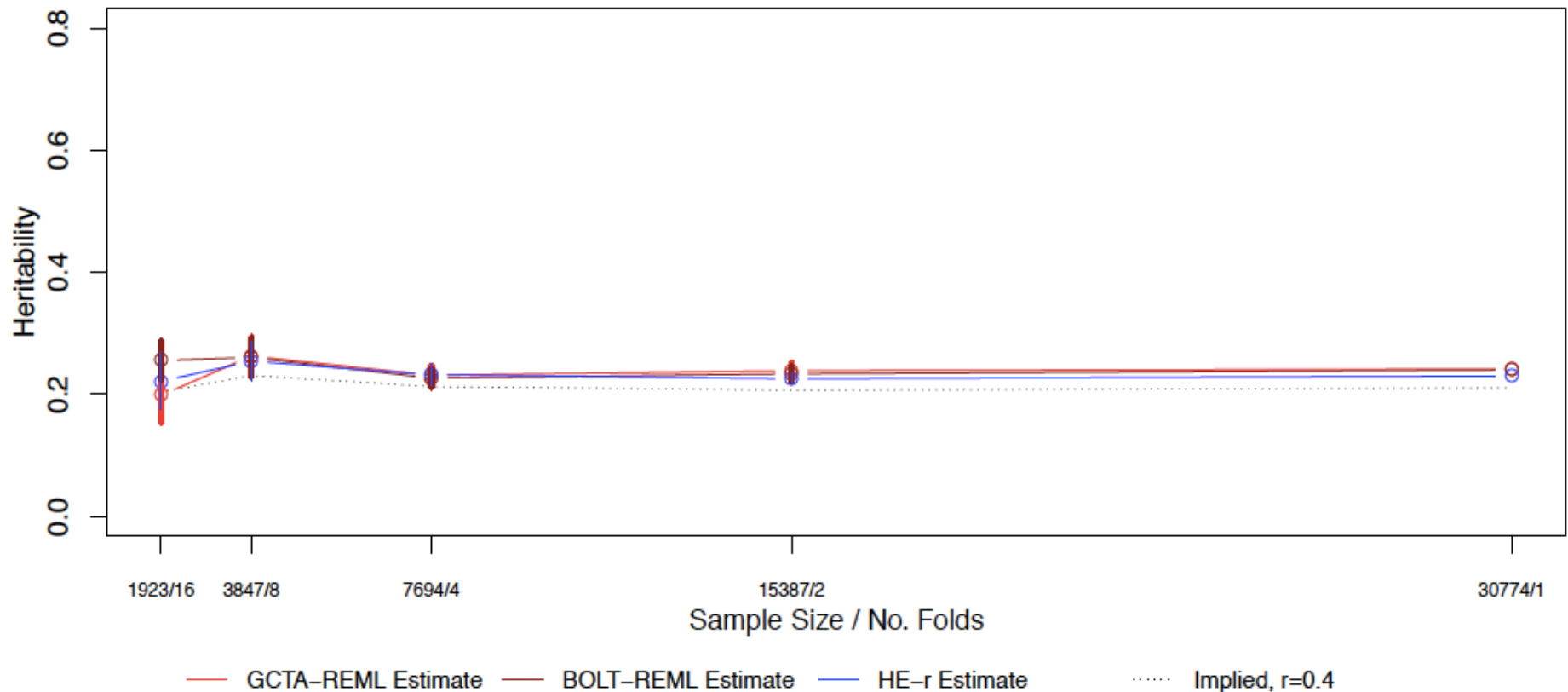


HE vs. GREML estimates

- For most realistic situations, $m \gg N$, and thus GREML and HE estimates are similar: both over-estimate equilibrium h^2_{snp}
- We can vary N (holding m constant) to see if AM is biasing estimates in real data

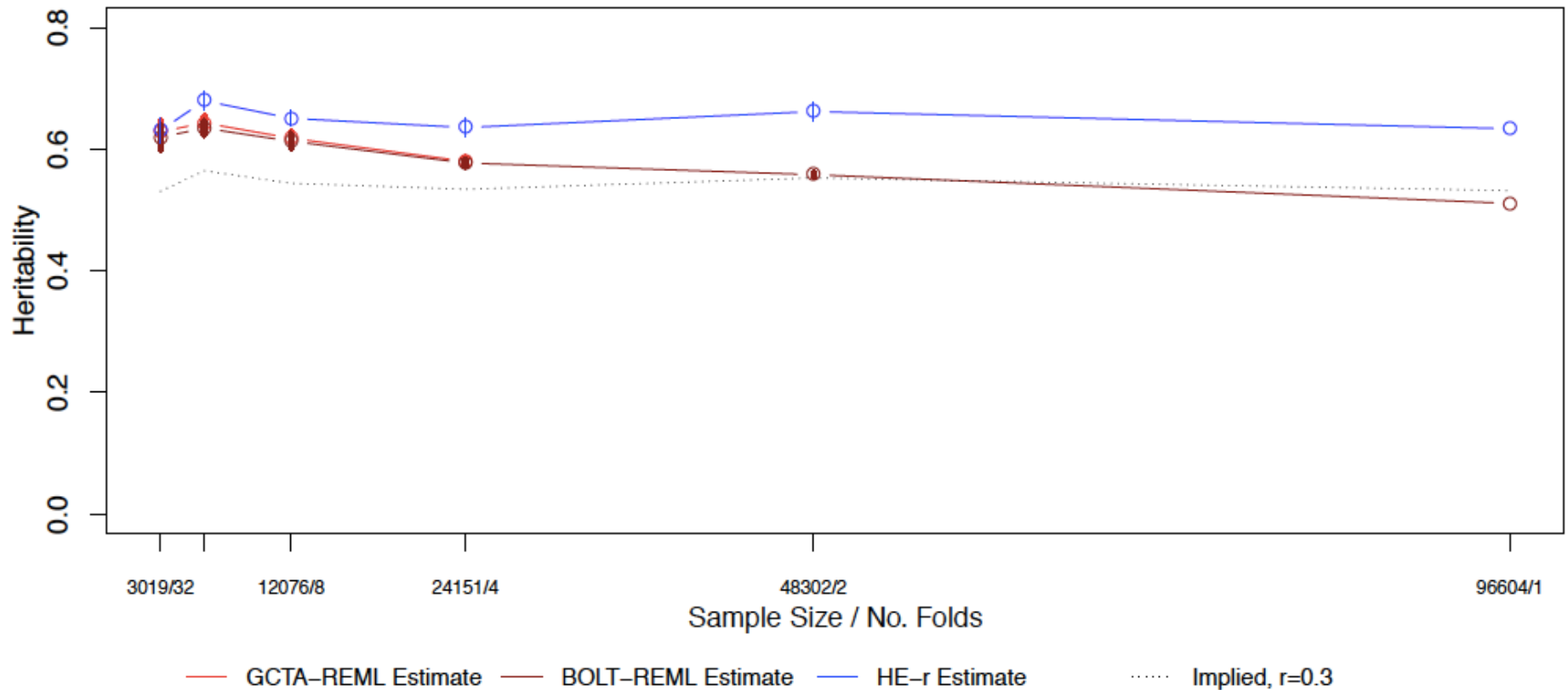
HE (blue) vs. REML (red) h^2_{snp} estimates of systolic BP - UK Biobank

Error bars: 1 SEMs; GRM pruned for relatedness > .05; Covariates: sex, age, age-squared, PCs 1-4, townsend deprivation



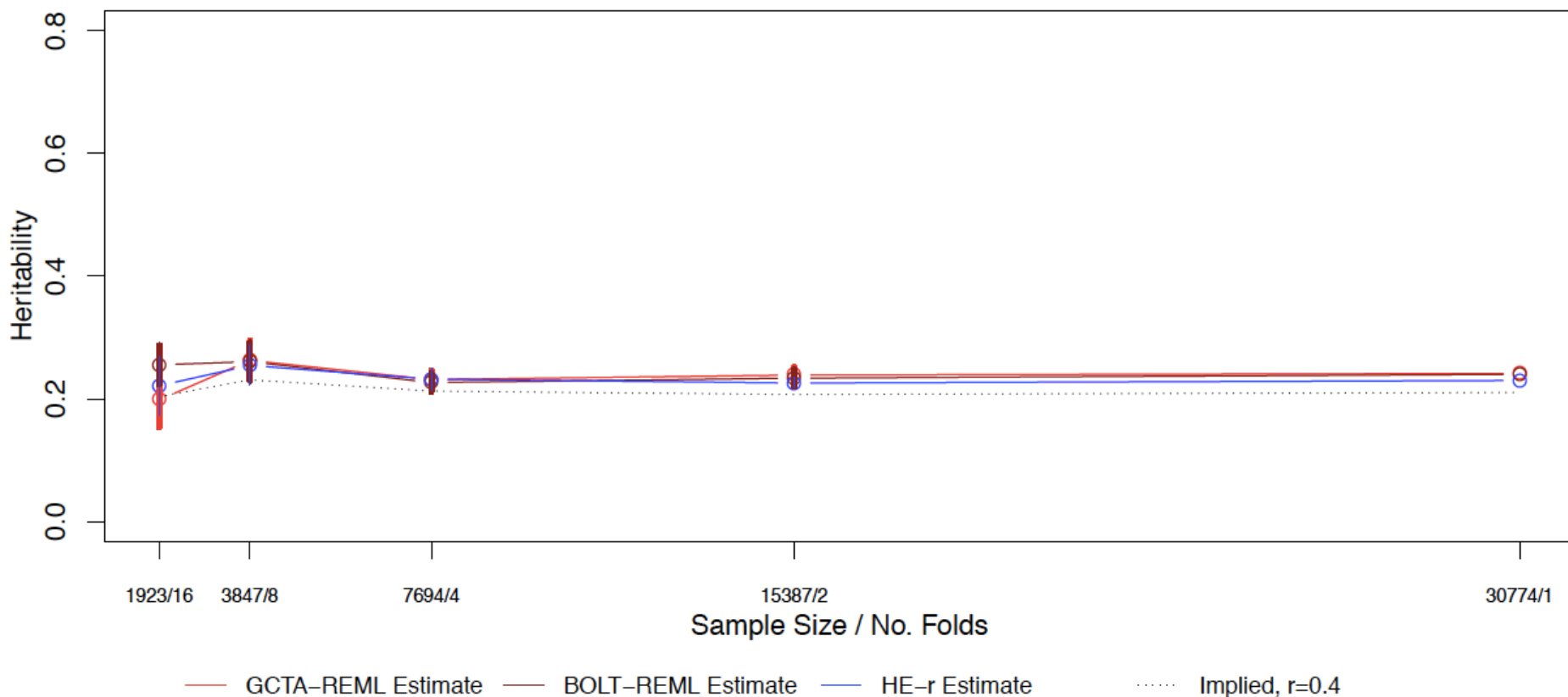
HE (blue) vs. REML (red) h^2_{snp} estimates of height - UK Biobank

Error bars: 1 SEMs; GRM pruned for relatedness > .05; Covariates: sex, age, age-squared, PCs 1-4



HE (blue) vs. REML (red) VA estimates of fluid IQ - UK Biobank

Error bars: 1 SEMs; GRM pruned for relatedness > .05; Covariates: sex, age, age-squared, PCs 1-4, townsend deprivation



Summary – bias due to AM

- AM creates upward biases in HE and REML h^2_{SNP} estimates
 - We see evidence for this in UK Biobank data for height but not for fluid IQ
 - Natural selection creates negative LD among CVs. The combined effect of AM and NS could cancel each other out.
- Remaining issues:
 - Unsure how to account for the bias. LDMS GREML does not help.
 - Need to understand the effects of NS on h^2_{snp}

Big picture: Using SNPs to estimate h^2

- There has been a great deal of excitement about using SNPs to estimate h^2
- Large sequence reference panels (TopMed) allow SNPs to be imputed down to MAF $\sim .0001$.
 - h^2_{snp} will approach h^2
 - Also allows investigation of allelic spectra, and importance of biological/evolutionary annotations
 - By understanding true h^2 , can begin understanding importance of familial environmental factors
- However, it is crucial to understand the factors that can bias these estimates
 - LDMS accounts for biases due to MAF & stratification
 - But not for biases caused by AM (and probably NS)

Acknowledgements

Collaborators

Consortia/Databases

Haplotype Reference Consortium

UK Biobank

University of Queensland

Peter Visscher

Jian Yang

Naomi Wray

Mike Goddard

CU

Matt Jones

Broad Institute

Ben Neale

Postdoctoral

Fellows

Teresa de Candia

Luke Evans

Rasool Tahmasbi

Graduate

Students

Emma Johnson

Richard Border



Funding

NIMH K01 MH085812 (Keller)

NIMH R01 MH100141 (Keller)

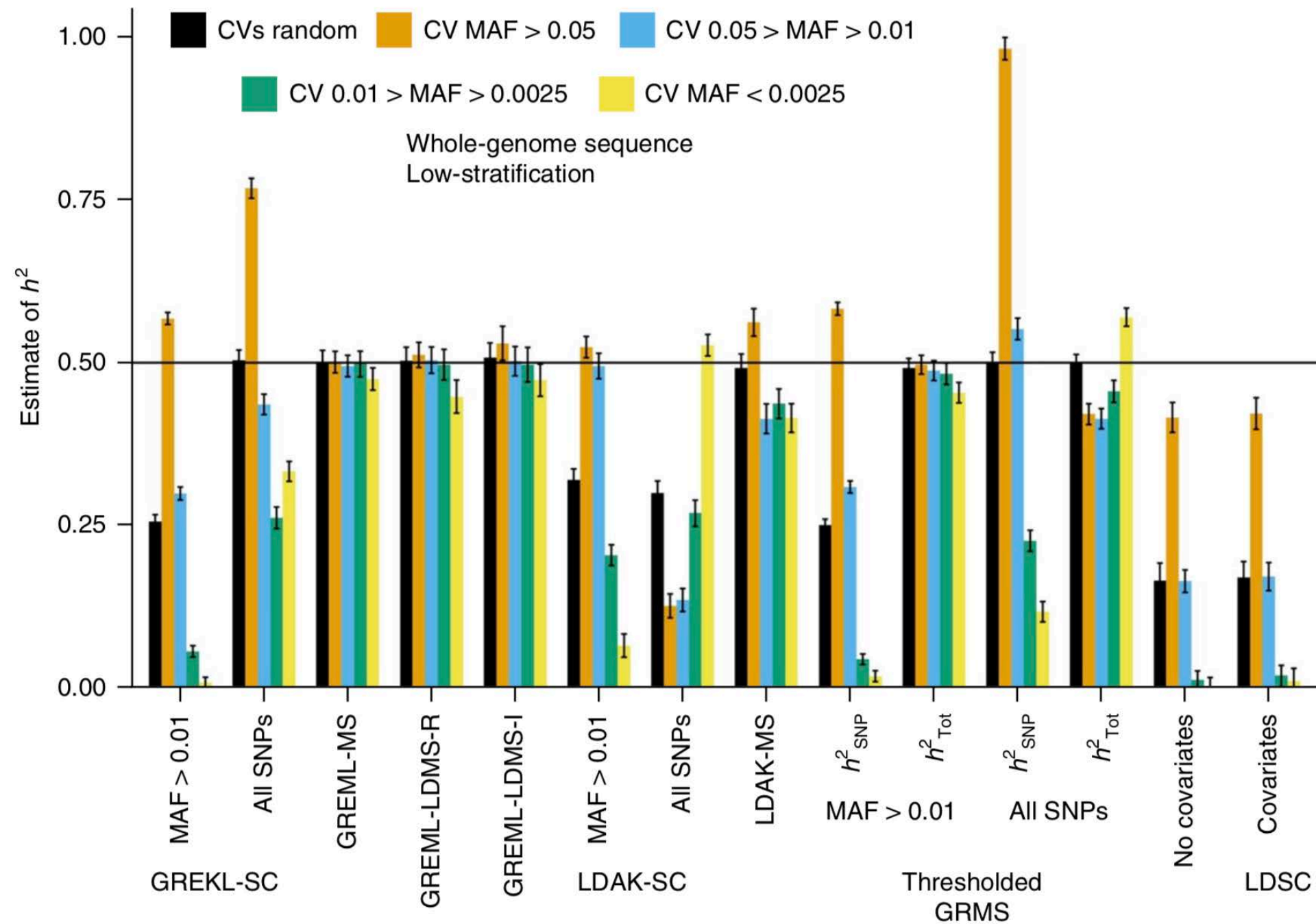


Table 1 | Summary of commonly applied methods and a description of findings from simulations

Method	Description	Major assumptions	Simulation findings regarding \hat{h}_{SNP}^2	Computational issues
GREML-SC ⁵	Often called the GCTA approach. Originally applied to common array SNPs only. Estimates \hat{h}_{SNP}^2 , the amount of h^2 caused by CVs tagged by SNPs used to create the GRM.	(i) Genetic similarity is uncorrelated with environmental similarity; (ii) an infinitesimal model; (iii) SNP effects are normally distributed, independent of LD, and inversely proportionate to MAF ($\alpha = -1$).	Biased to the degree that the average LD among SNPs is different from the average LD between SNPs and CVs. This occurs in stratified samples and when MAF and LD distributions of SNPs do not match those of CVs.	Simple model tractable with large samples (>100,000).
GREML-MS ¹¹	The first multicomponent approach, usually applied by binning SNPs according to their MAF, annotation, or physical regions to explore genetic architecture.	Requires that the same assumptions of GREML-SC hold within each GRM.	Biased when CVs have generally higher or lower levels of LD than the SNPs used to make the GRM. Relatively large standard errors.	Run times and memory requirements higher than GREML-SC and increase as a function of the number of variance components estimated.
GREML-LDMS-R ⁷	A multicomponent approach that bins imputed SNPs by their MAF and regional LD.	Same as GREML-MS.	Use of regional LD scores can lead to biases when CVs have different LD on average compared to surrounding SNPs. Relatively large standard errors.	Same as GREML-MS.
GREML-LDMS-I	A multicomponent approach introduced here that bins imputed SNPs by their MAF and individual LD.	Same as GREML-MS.	Appears to be the least biased approach, even when traits have complex genetic architectures. Relatively large standard errors.	Same as GREML-MS.
LDAC-SC ^{15,20}	Introduced to account for redundant tagging of CVs by common SNPs. Recently modified to incorporate error due to imputation and to alter the MAF effect-size relationship.	Same as GREML-SC, except that allelic effects are a function of LD. Extended to assume that effects are also a function of imputation quality and weakly inversely proportionate to MAF ($\alpha = -0.25$).	Can correct for the overestimation observed in GREML-SC from redundant tagging of CVs, but otherwise about as biased as GREML-SC when assumptions are unmet, although the biases are sometimes in different directions.	Same as GREML-SC.
LDAC-MS ¹⁵	A multicomponent extension of LDAC-SC that bins SNPs by MAF.	Requires that the same assumptions of LDAC-SC hold within each GRM.	Less biased on average than LDAC-SC, but more biased than GREML-LDMS-I or -R). Relatively large standard errors.	Same as GREML-MS.
Threshold GRMs ²⁴	A multicomponent approach with two GRMs: the normal (unthresholded) GRM built from all SNPs and a second GRM with entries set to 0 if below a threshold. Conducted in samples that include close relatives.	Same as GREML-SC for the unthresholded GRM. Assumes no shared environmental influences among close relatives.	Estimates associated with unthresholded GRM similar to those of GREML-SC. When used in samples that include close relatives, the second GRM captures pedigree-associated variation but can be upwardly biased by shared environmental influences.	See GREML-SC.
LD score regression ¹⁹	Uses the slope from χ^2 (from GWAS) regressed on SNPs' LD scores to estimate the h^2 due to CVs in LD with common SNPs.	Infinitesimal model with allelic effects normally distributed.	Largely robust to confounding due to stratification and shared environmental influences. Estimates h^2 due to common CVs only, even when used on imputed or WGS data. Underestimates h^2 if the trait is not highly polygenic.	The most computationally efficient method of those compared and tractable for very large datasets.