Using GREML to estimate SNP- heritability among ‘unrelated’ individuals

Matthew Keller

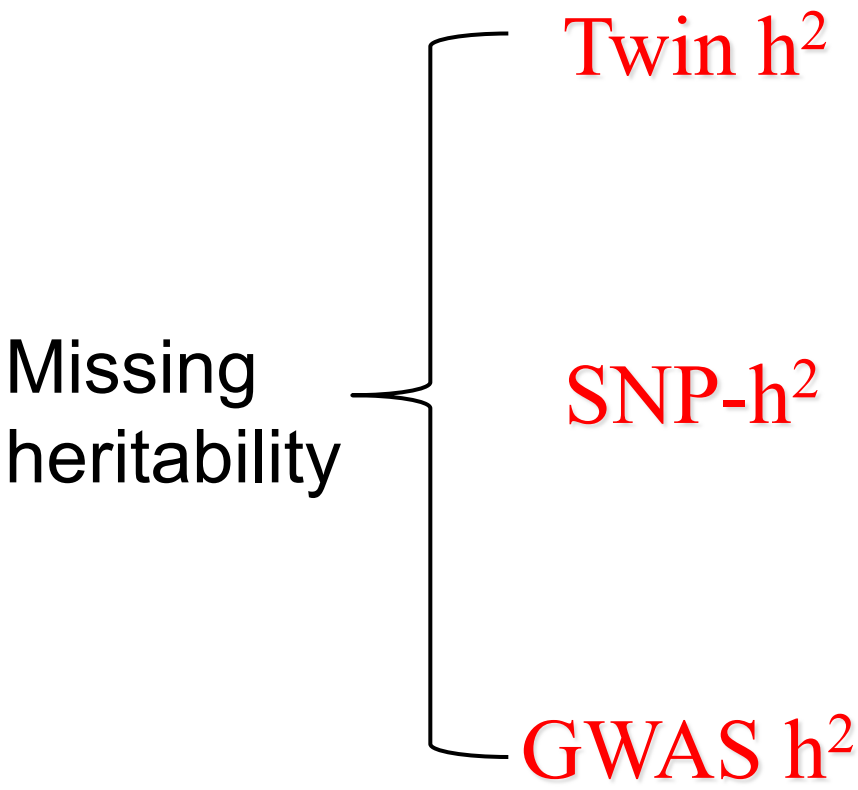
University of Colorado at Boulder

Three flavors of h^2

- Twin h^2
  - estimate of the narrow-sense (but in practice, broad-sense) h^2
  - Can be biased if strong assumptions are unmet
- SNP-h^2
  - Estimate of the narrow-sense h^2 captured by causal variants tagged by SNPs in your analysis
- GWAS h^2
  - Sum of r^2 from all independent genome-wide significant SNPs from a GWAS

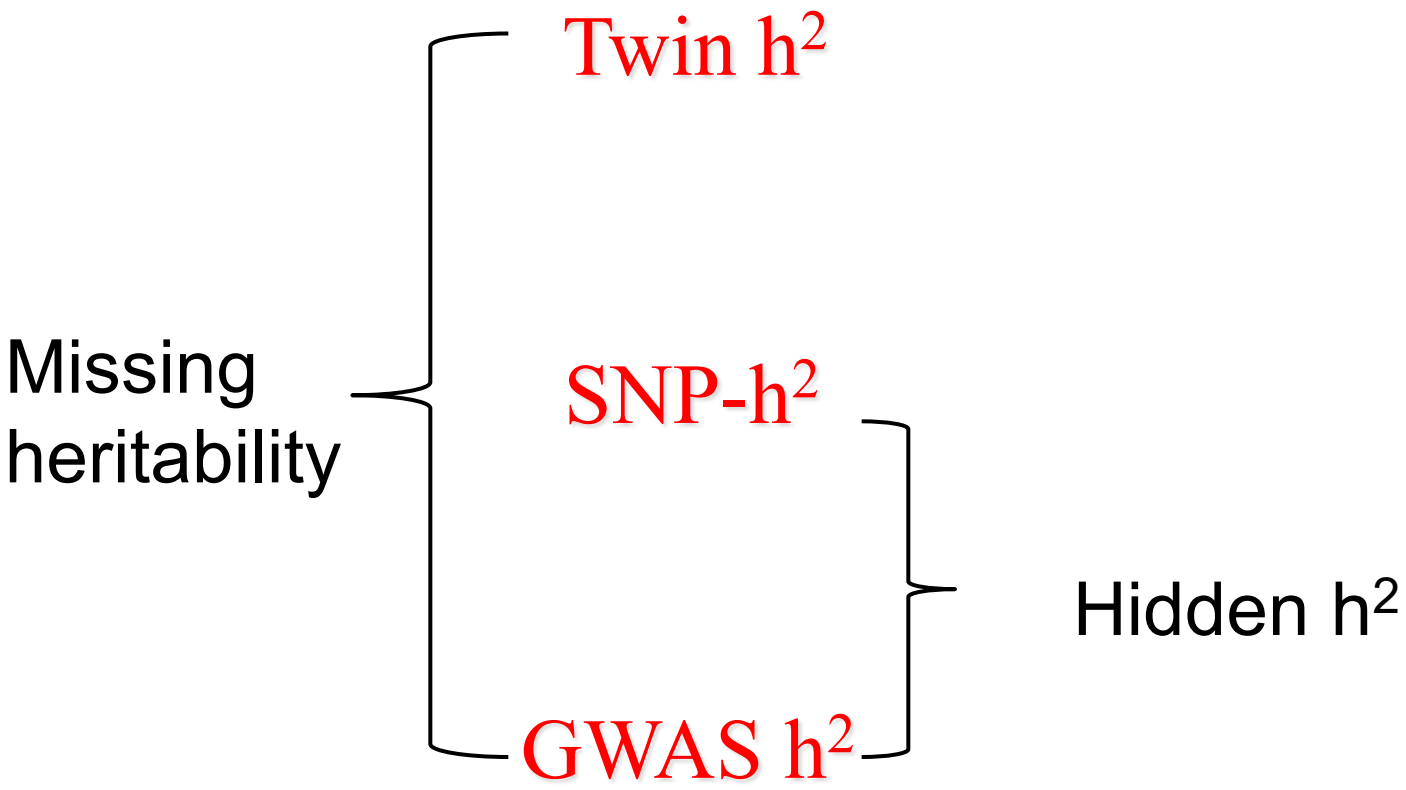
Three types of missing h^2

- Almost always Twin h^2 > SNP-h^2 > GWAS h^2



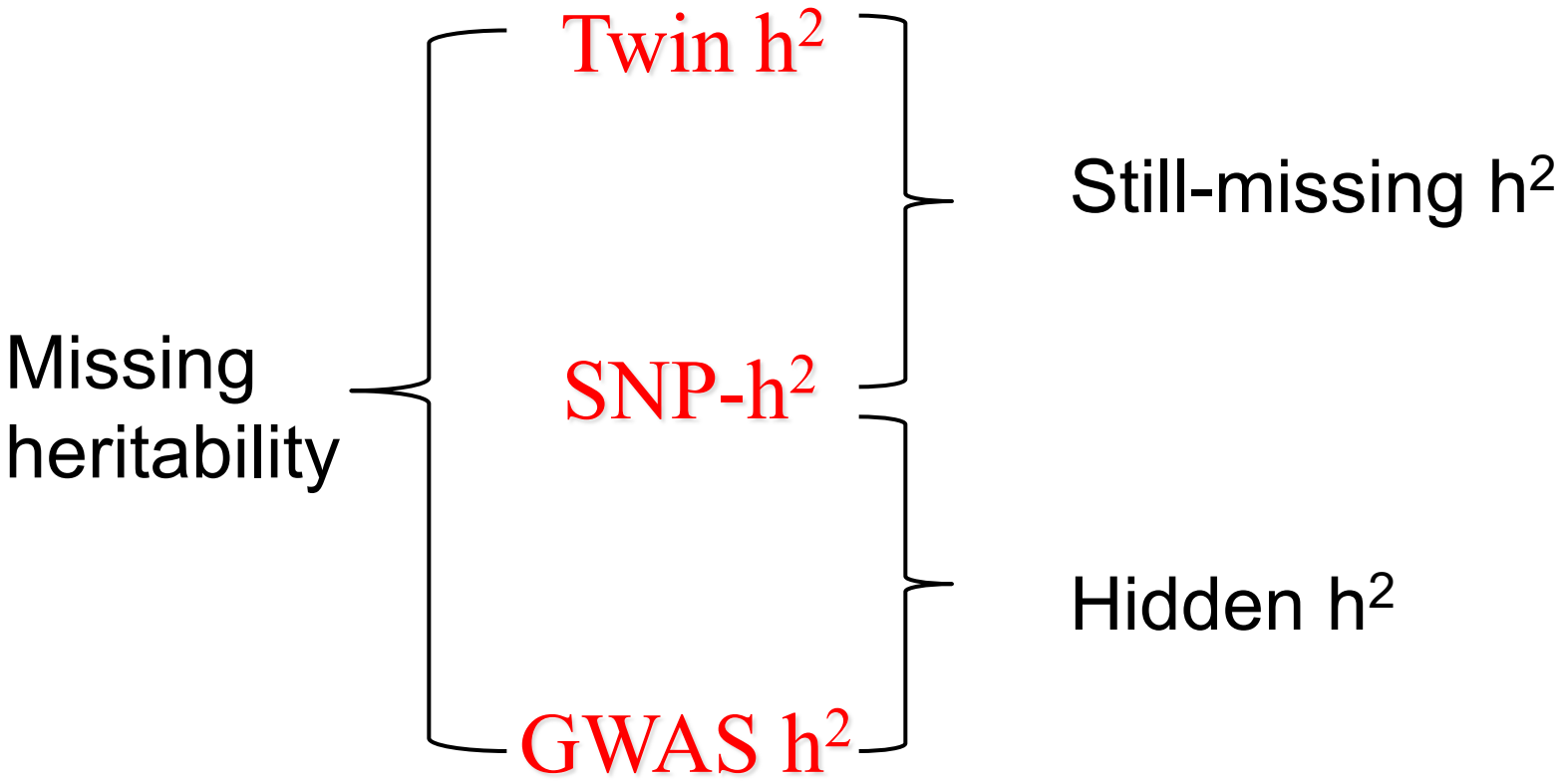
Three types of missing h^2

- Almost always Twin h^2 > SNP-h^2 > GWAS h^2



Three types of missing h^2

- Almost always Twin h^2 > SNP-h^2 > GWAS h^2



Why care about SNP-h^2?

- h^2_{snp} should be unbiased by environmental factors that increase close relative similarity. In particular, doesn't rely on rMZ > rDZ due only to genetic differences (although still assumes no relationship between genetic & env. similarity among distant relatives)
- Estimating h^2_{snp} from binned SNPs allows for estimates of relative importance of different SNP annotations. E.g., allows for estimates of allelic spectra (distribution of CV MAF)
- Can estimate r_g between low prevalence disorders that are impractical or impossible to estimate using twins/family designs
- As we continue to capture lower MAF SNPs through imputation or sequencing, GREML (but not LDSC) estimates of h^2_{snp} approach full narrow-sense h^2.

Multiple approaches to estimating h^2_{snp}

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

Multiple approaches to estimating h^2_{snp}

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

  TODAY
- Bayesian approaches

pihat

$\hat{\pi}$ = E(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*

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

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

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

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

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

H-E REGRESSION

Regression estimates of h^2

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

product of centered scores (here, z-scores)

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

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

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

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

Regression estimates of h^2

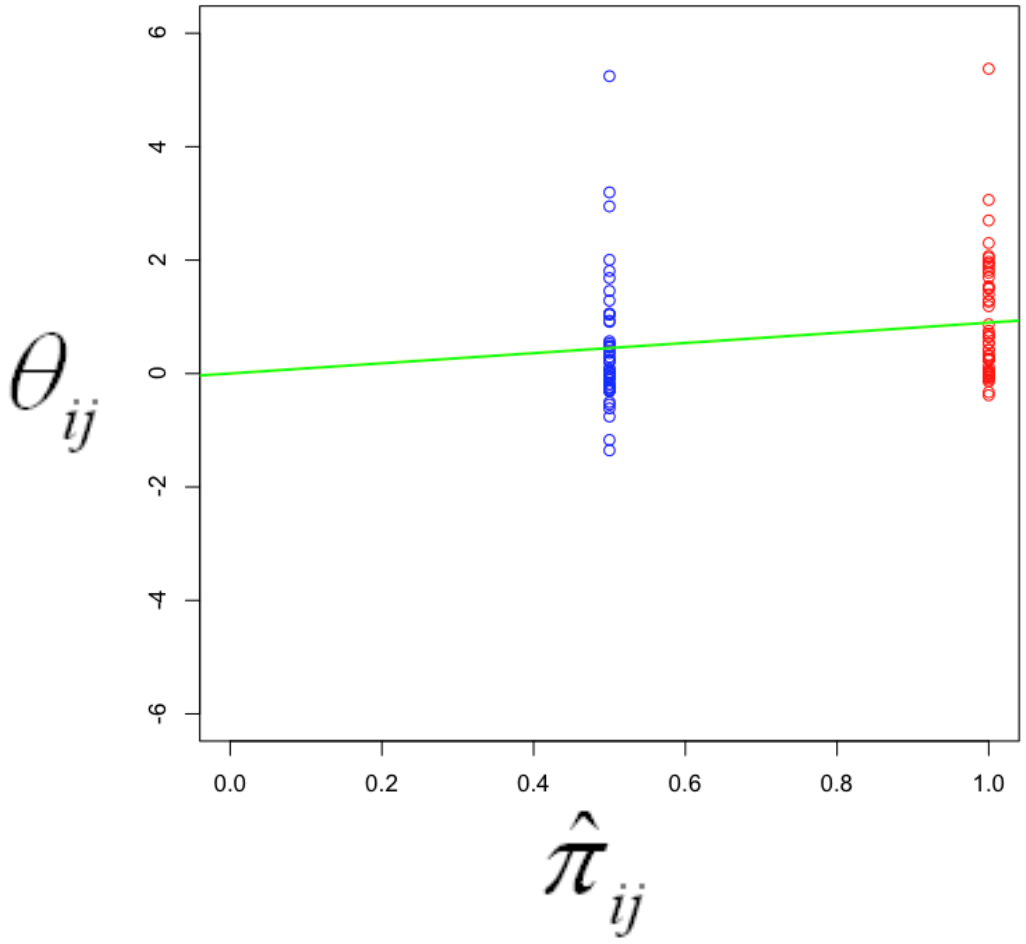
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



Regression estimates of h^2

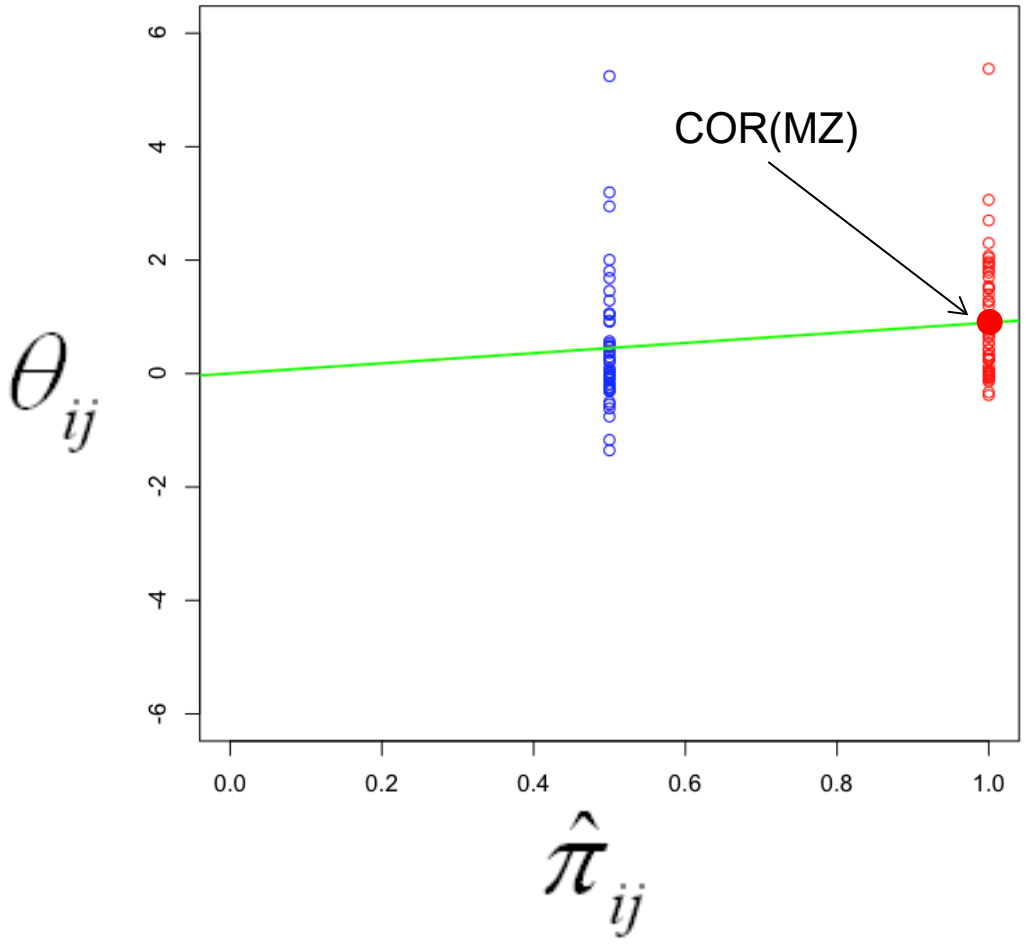
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



Regression estimates of h^2

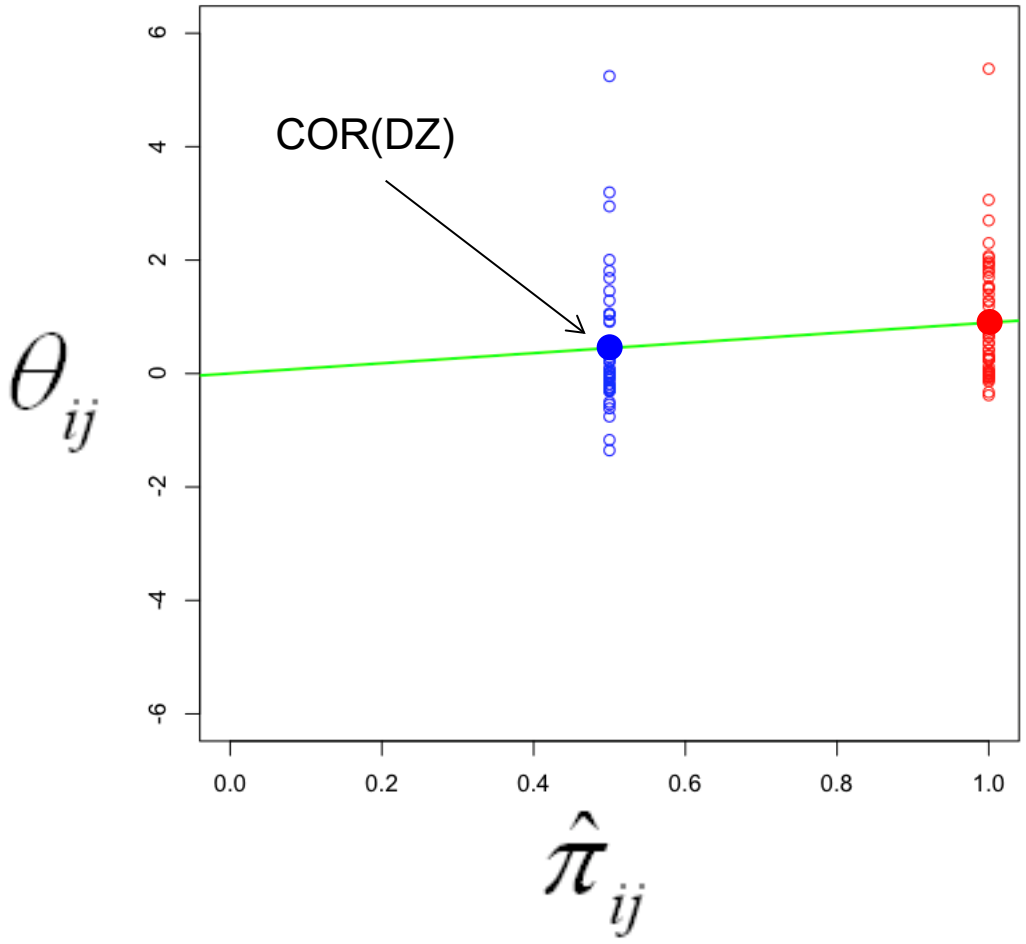
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



Regression estimates of h^2

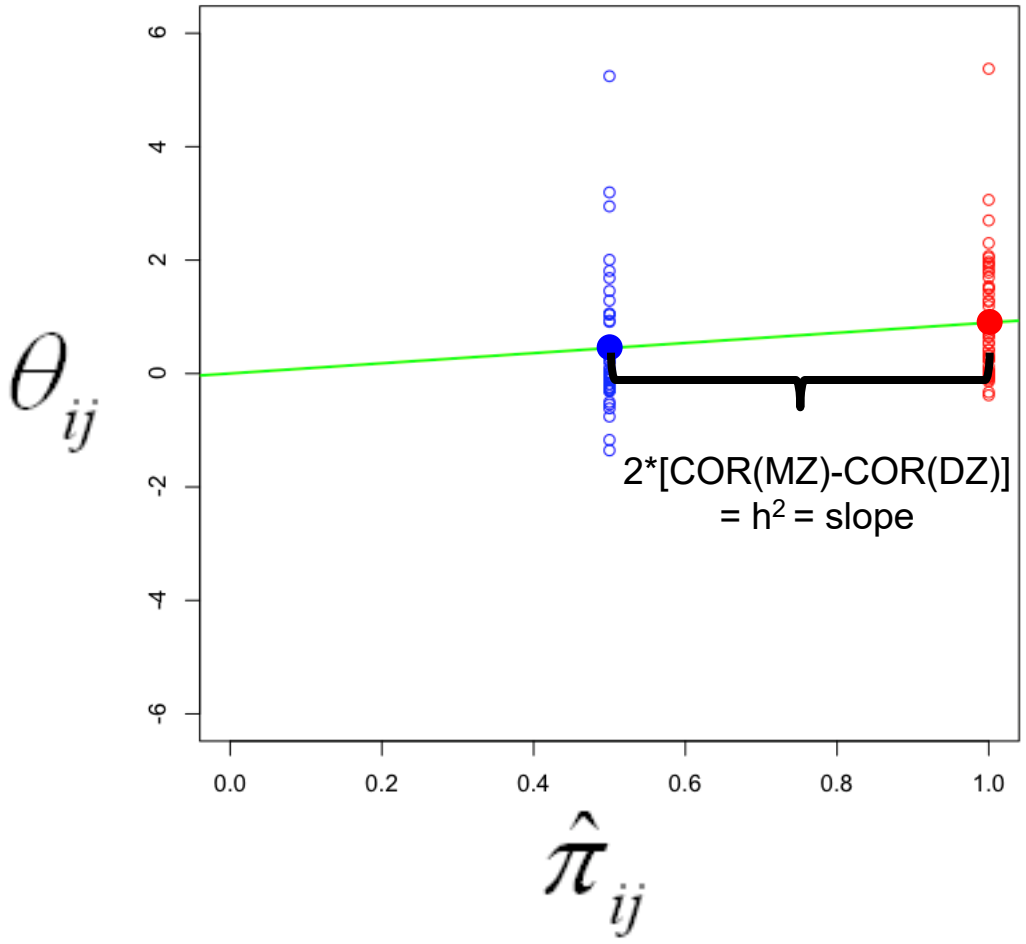
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



Regression estimates of h^2

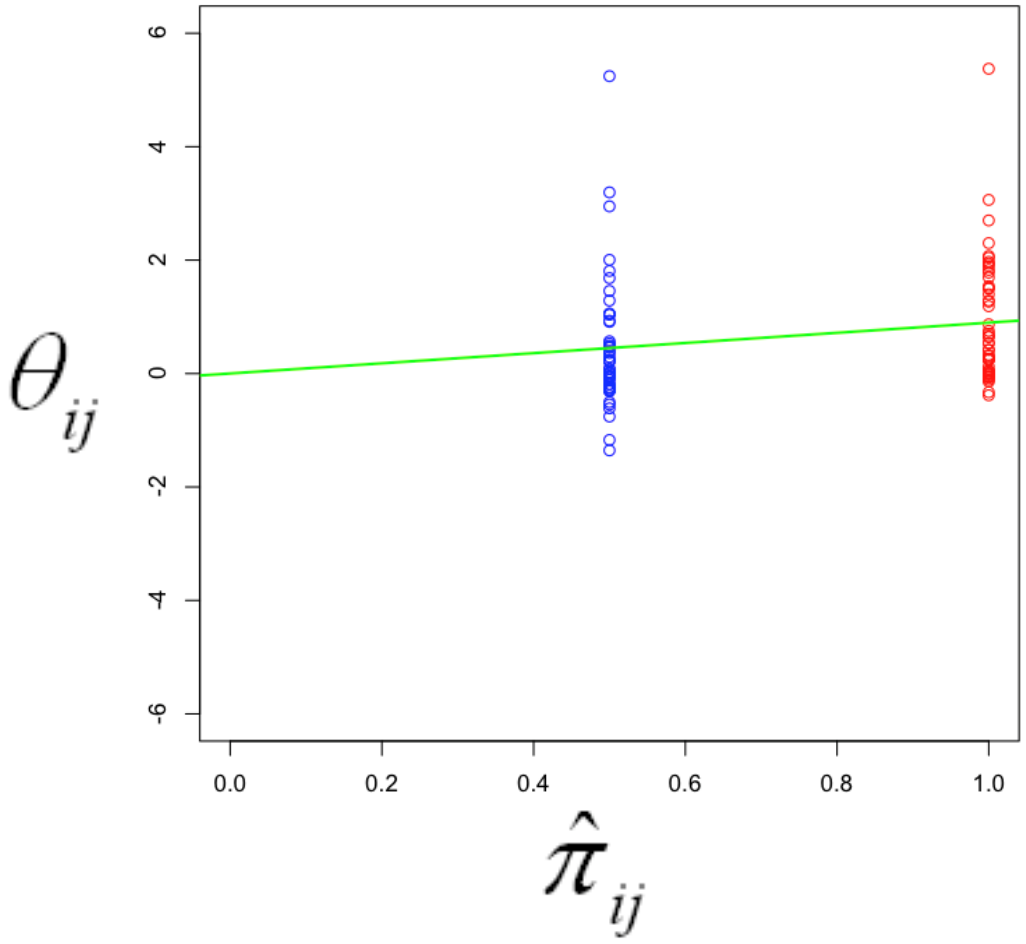
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



Regression estimates of h^2

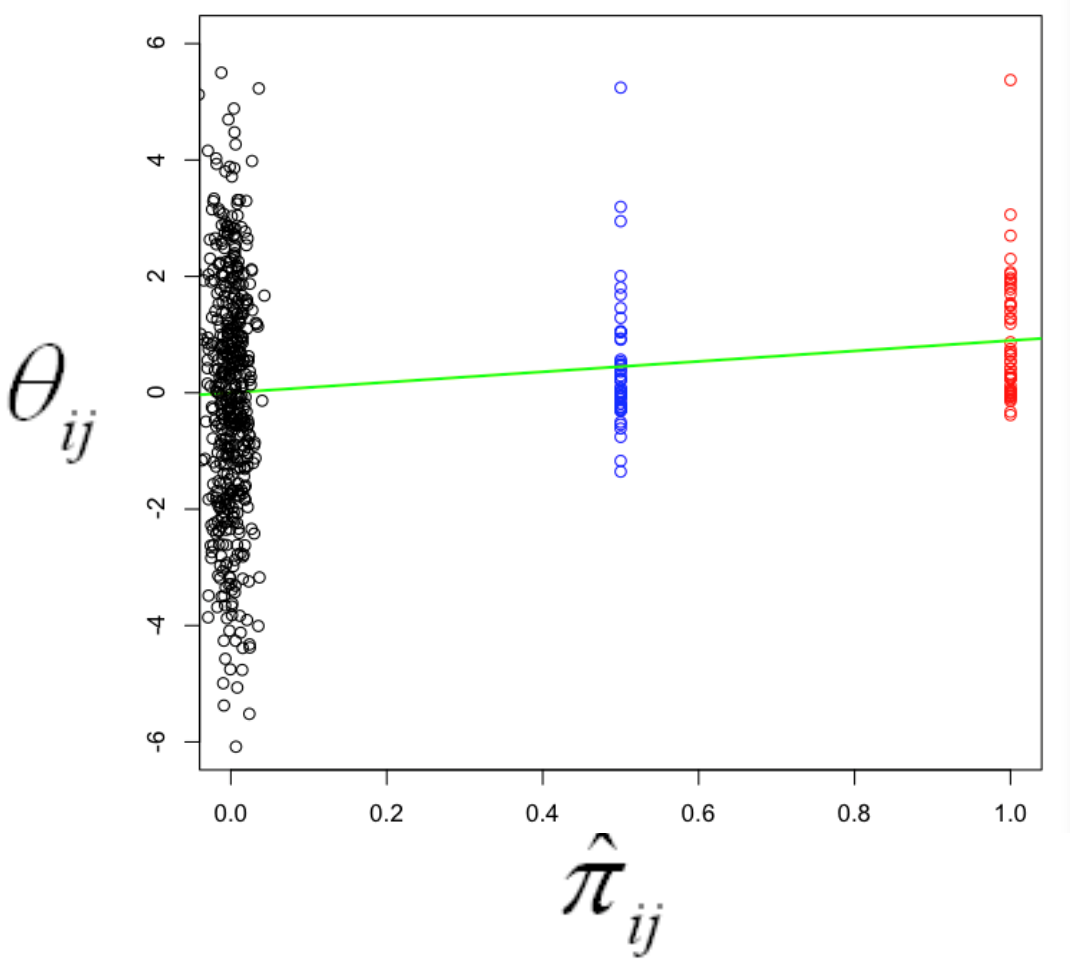
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



Regression estimates of h^2_{snp}

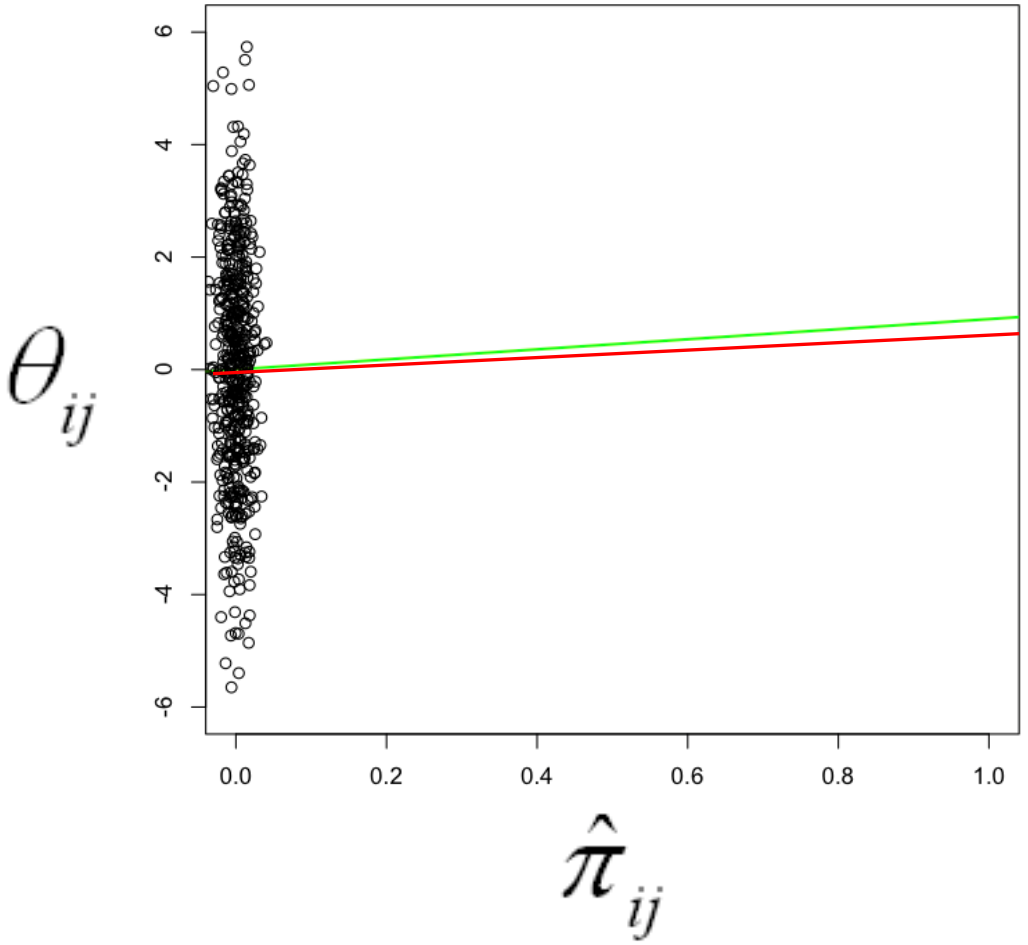
$$\theta_{ij} = Z_i Z_j$$

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

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

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

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



GREML

nature genetics

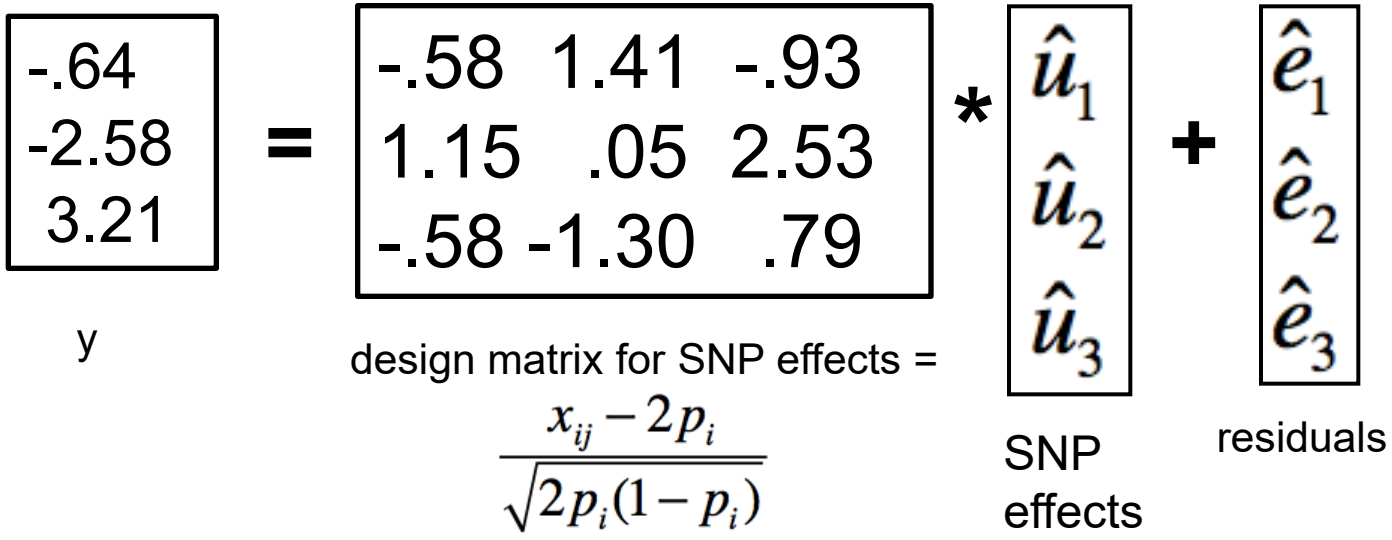
2010

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

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

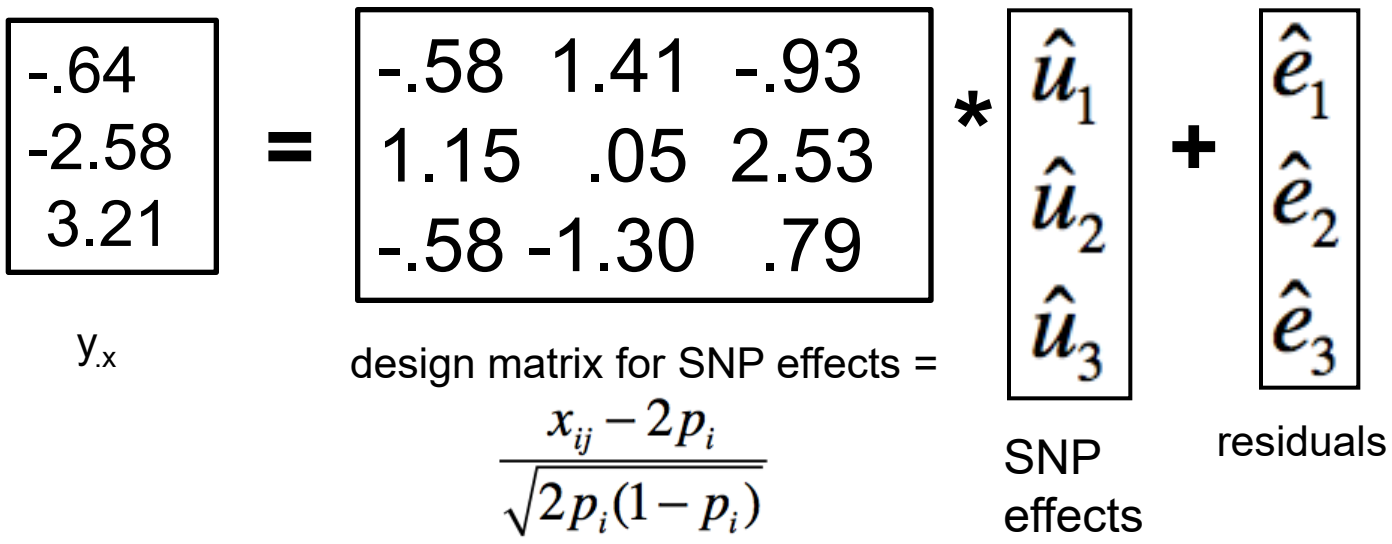
GREML Model

$$y = Z\hat{u} + \hat{e}$$



GREML Model

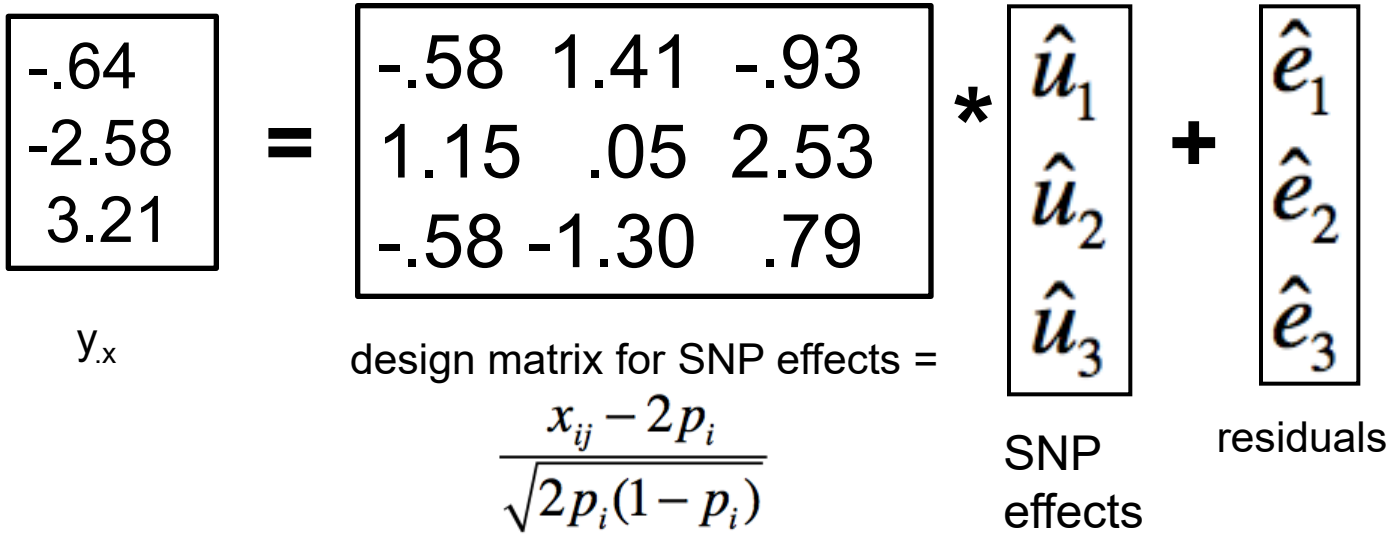
$$y = Z\hat{u} + \hat{e}$$



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

$$y = Z\hat{u} + \hat{e}$$



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^2_u and thus σ^2_A)

$$\begin{aligned}\text{var}(y) &= ZZ'\sigma^2_u + I\sigma^2_e \\ &= ZZ'(\sigma^2_A / m) + I\sigma^2_e \\ &= G\sigma^2_A + I\sigma^2_e\end{aligned}$$

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

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

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

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

Identity matrix

GREML

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

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

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})

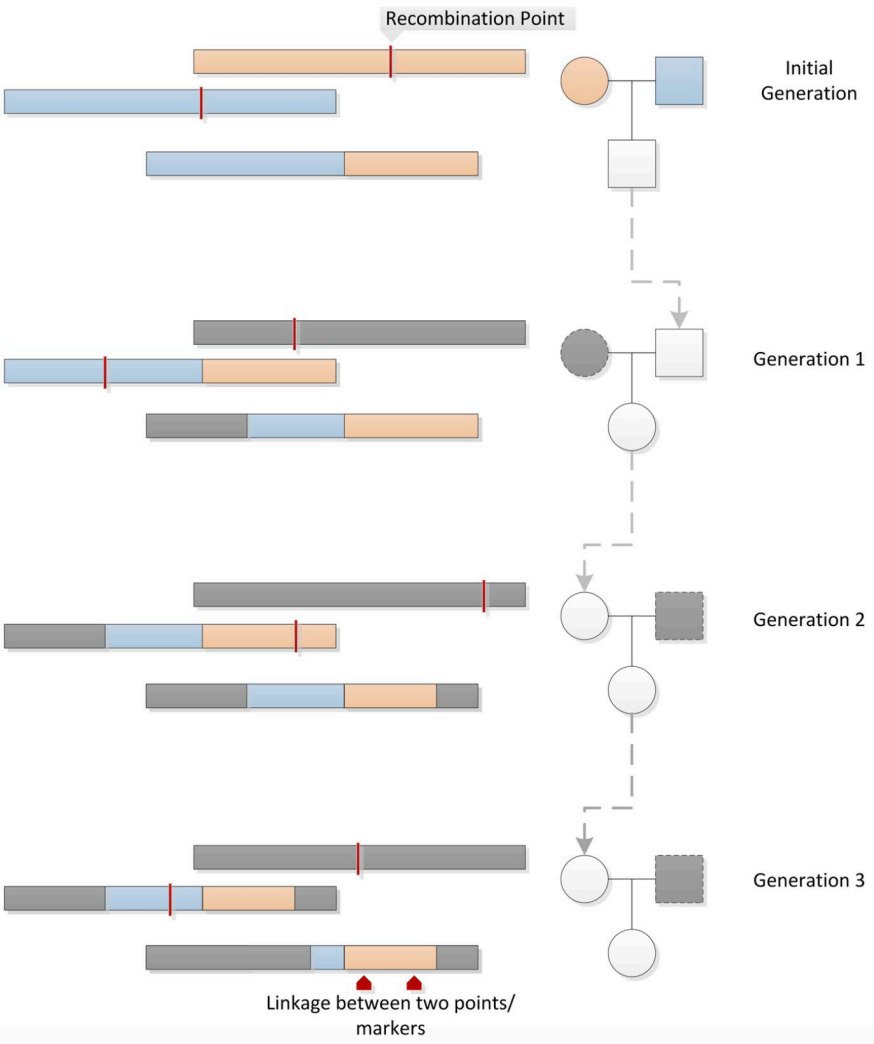
- If close relatives included (e.g., sibs), $h^2_{snp} \cong h^2$ estimated from a family-based method, because great influence of extreme pihats. Interpret h^2_{snp} as from these designs.
- If use 'unrelateds' (e.g., pihat < .05):
  - h^2 estimate 'uncontaminated' by shared environment and non-additive 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_{snp} = proportion of V_P due to CVs tagged by (in LD with) SNPs used in the GRM.
  - Typically, $h^2_{snp} < h^2$. It is the max r^2 possible from a PRS using those SNPs.

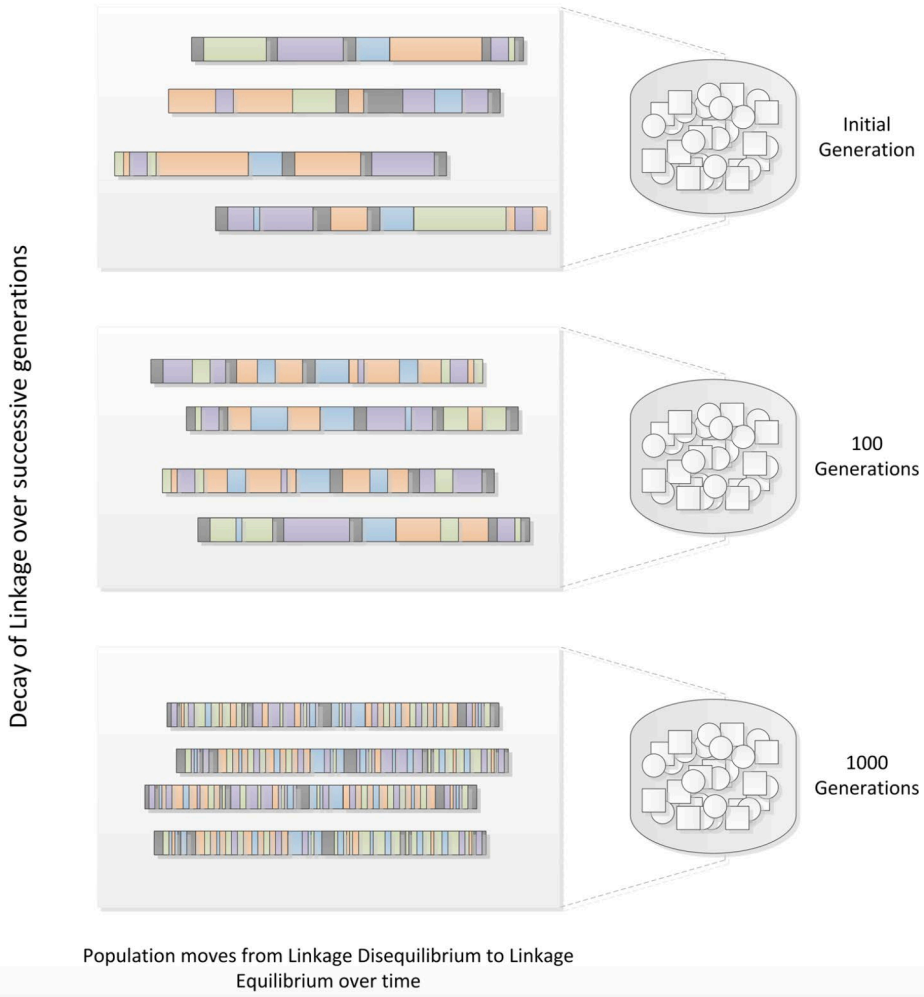
LD

Linkage disequilibrium (LD)

- Statistical association (e.g., r^2) between 2 SNPs
- Typically arises from a mutation that occurs on a haplotype. This mutation will co-segregate with nearby SNPs. As it rises in frequency, so too will nearby SNPs.
- 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.

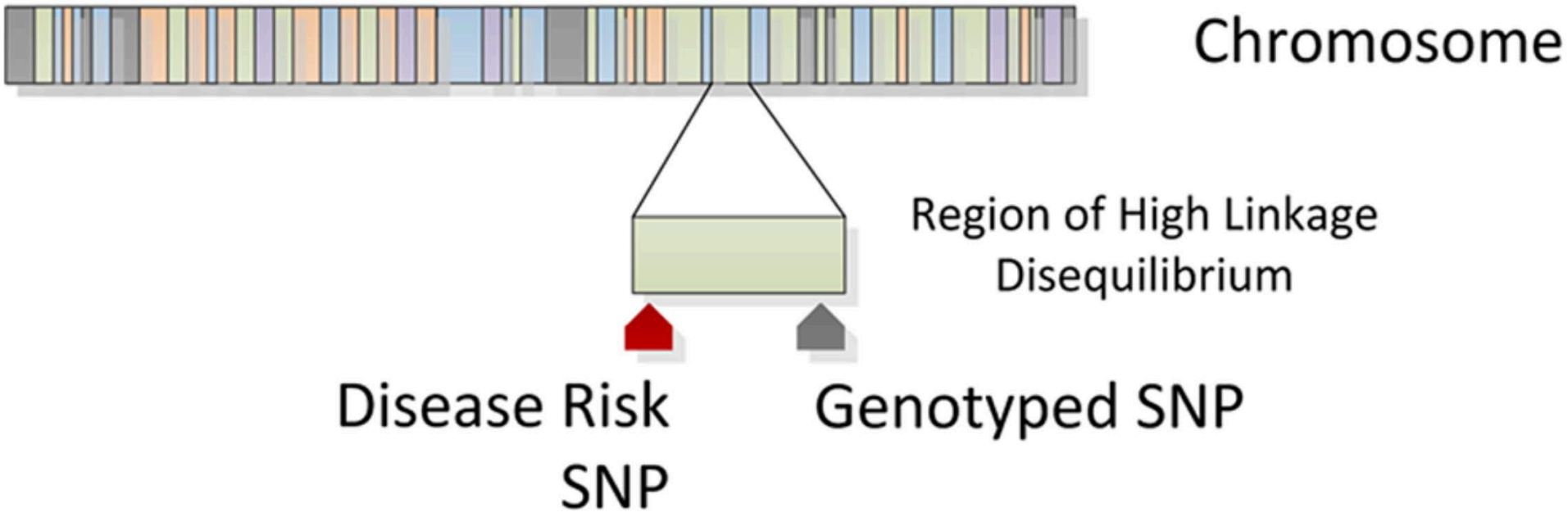
How LD arises & decays





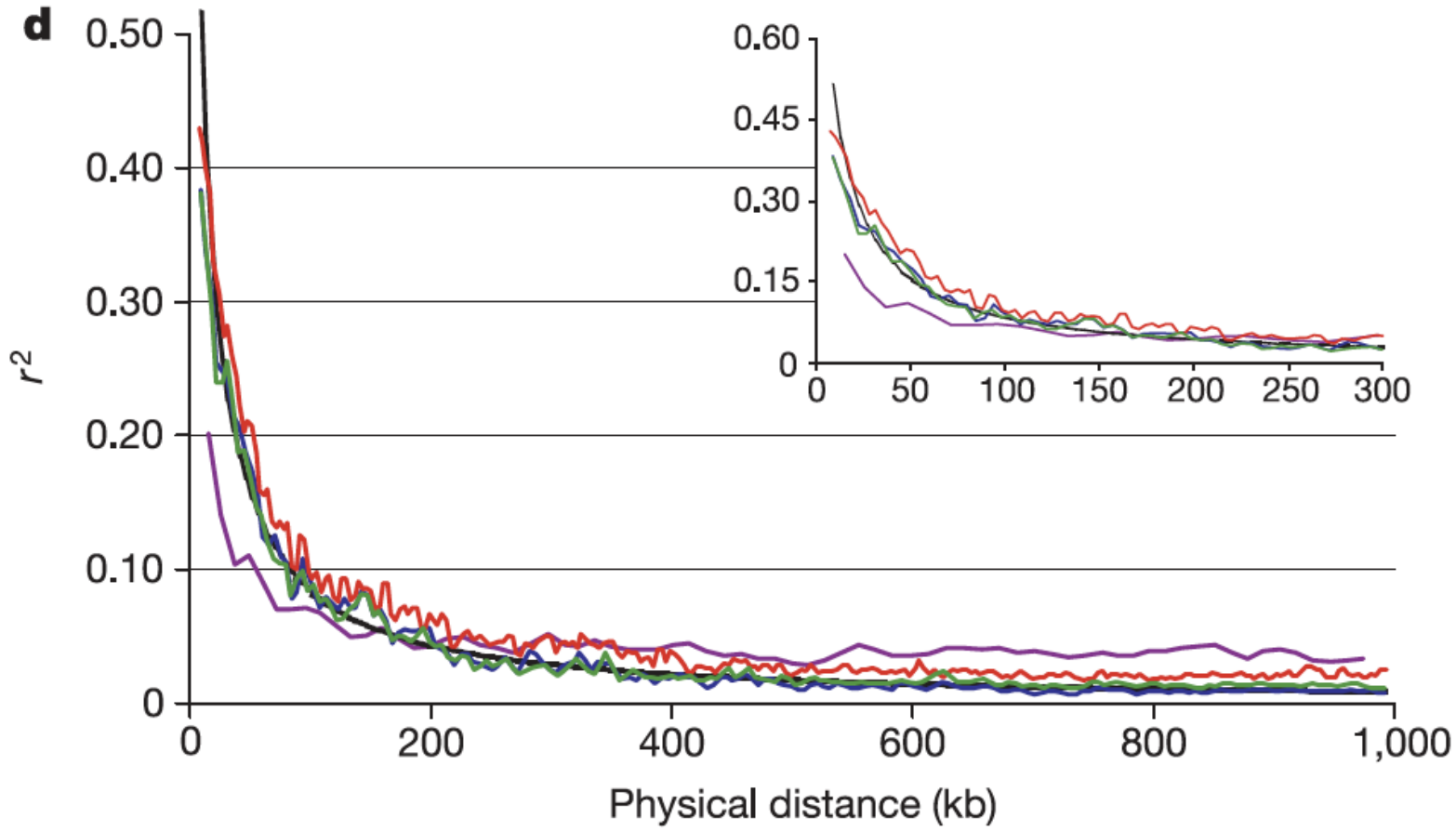
Bush & Moore, *PLoS Comp Bio*, 2012

SNPs can tag other nearby SNPs…



Bush & Moore, *PLoS Comp Bio*, 2012

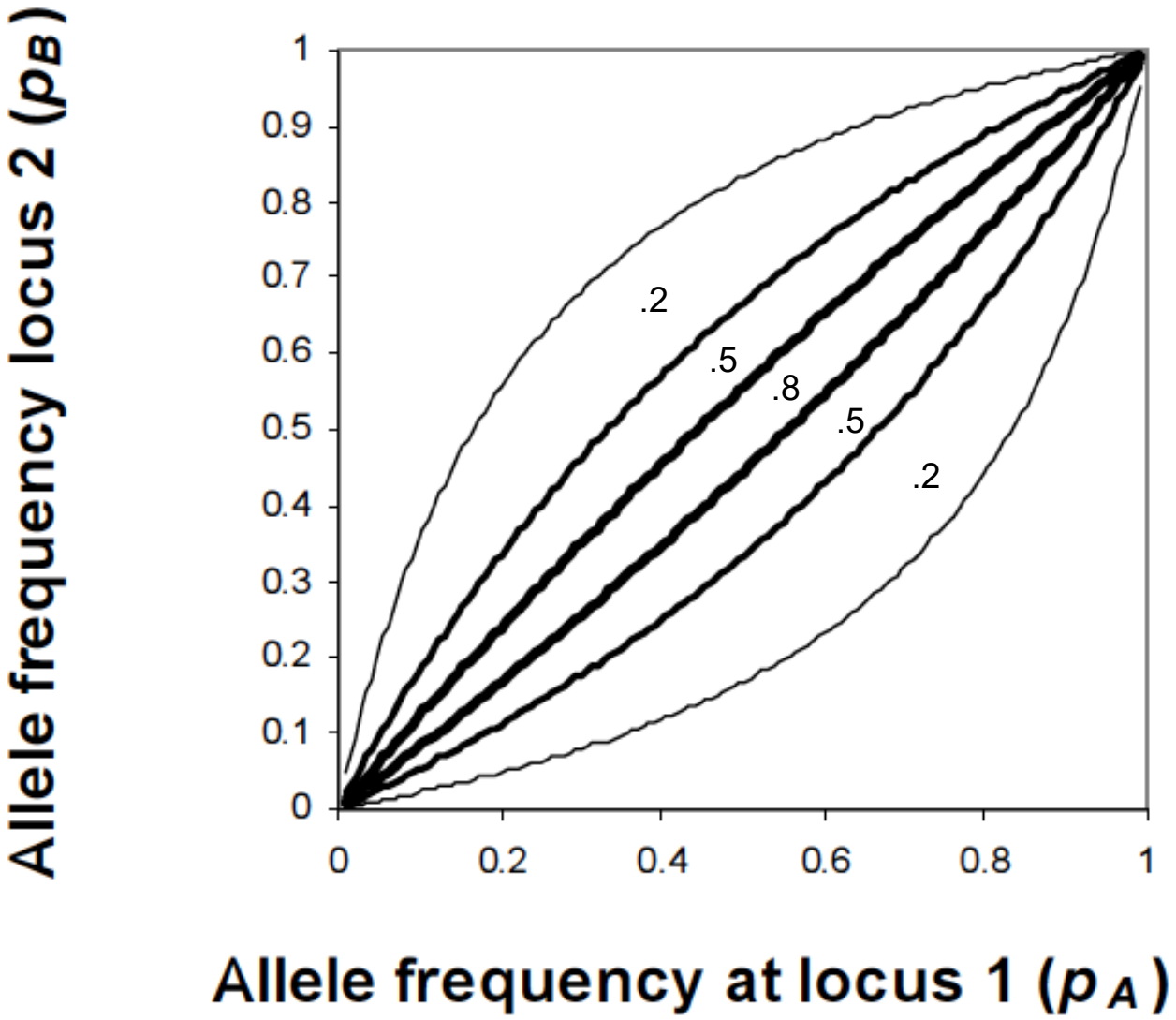
LD drops as a function of distance



Dawson et al, *Nature*, 2002

…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



Wray, *TRHG*, 2006

High LD possible only if the two alleles are of similar frequencies.
Here A/B are major and a/b are minor alleles on haplotypes.

Max LD $r^2=1.0$

a & b; A & B "in phase"

A	B
A	B
A	B
A	B
A	B
A	B
a	b
a	b
a	b

Min LD $r^2=0.0$

Random phase

A	b
A	B
A	B
A	b
A	B
A	B
a	b
a	B
a	B

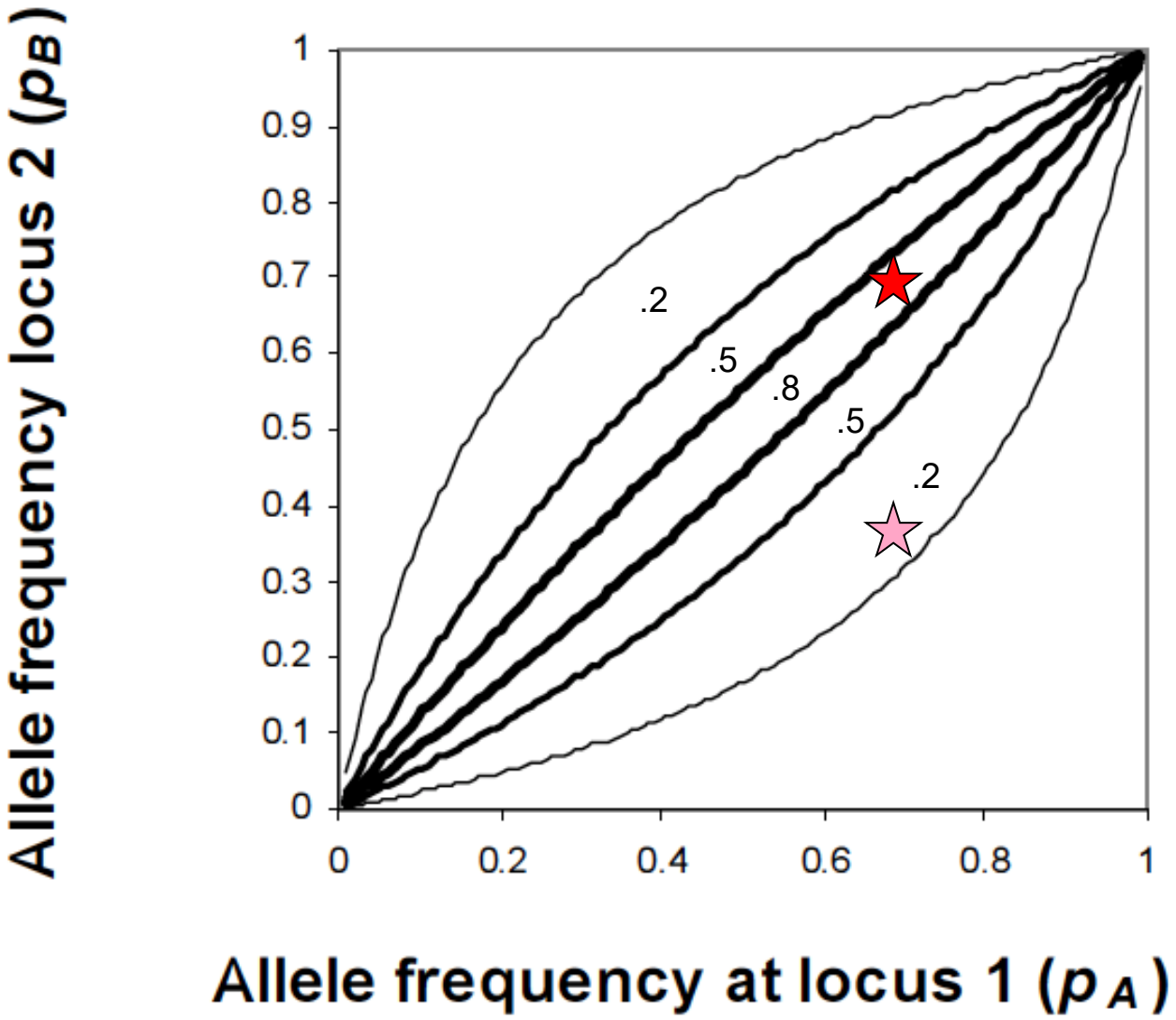
Max LD $r^2=.25$

a & B; A & b "in phase"

A	B
A	B
A	B
A	b
A	b
A	b
a	B
a	B
a	B

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

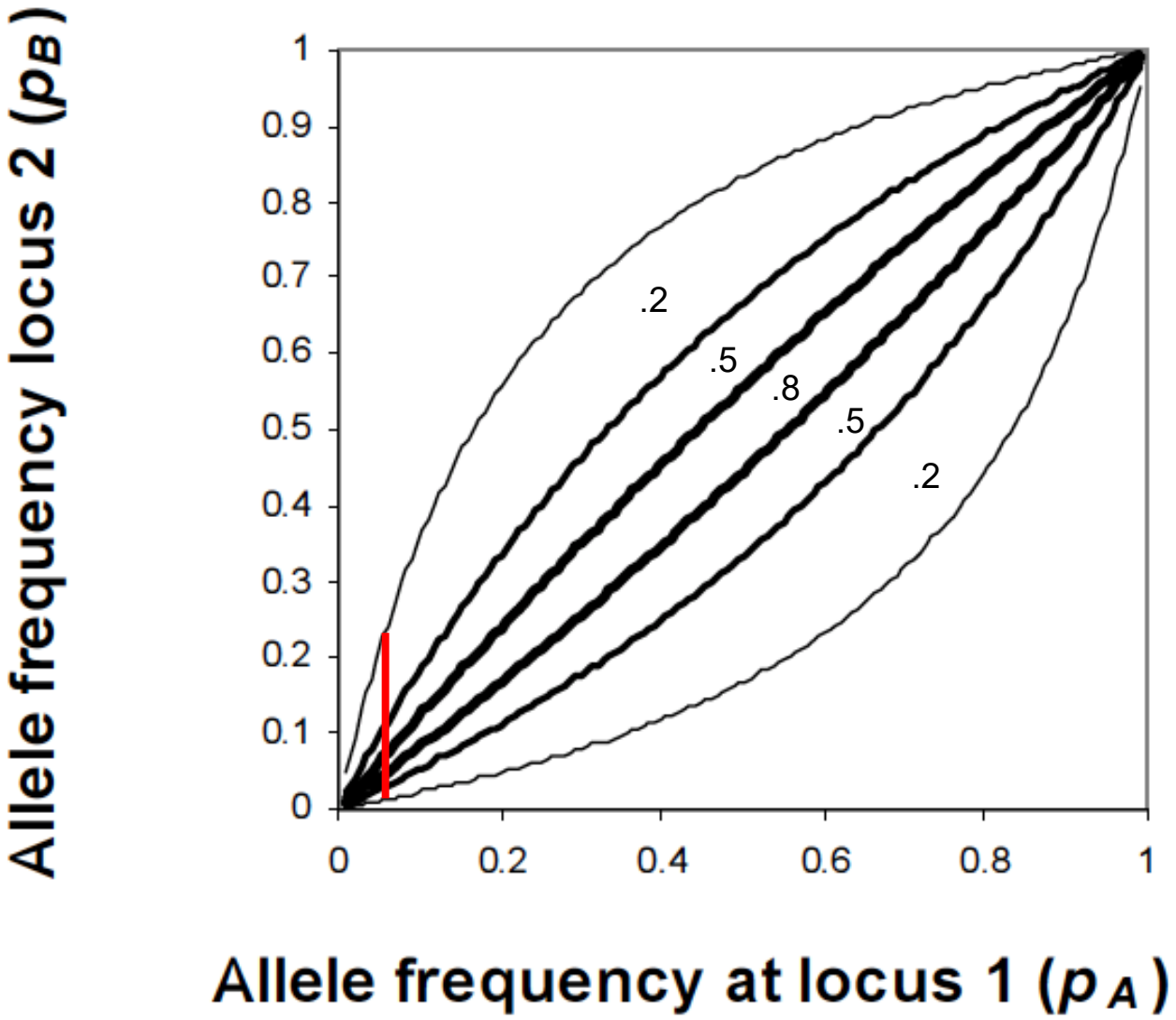
Possible range of allele frequencies given LD (r^2) between 2 SNPs



Wray, *TRHG*, 2006

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

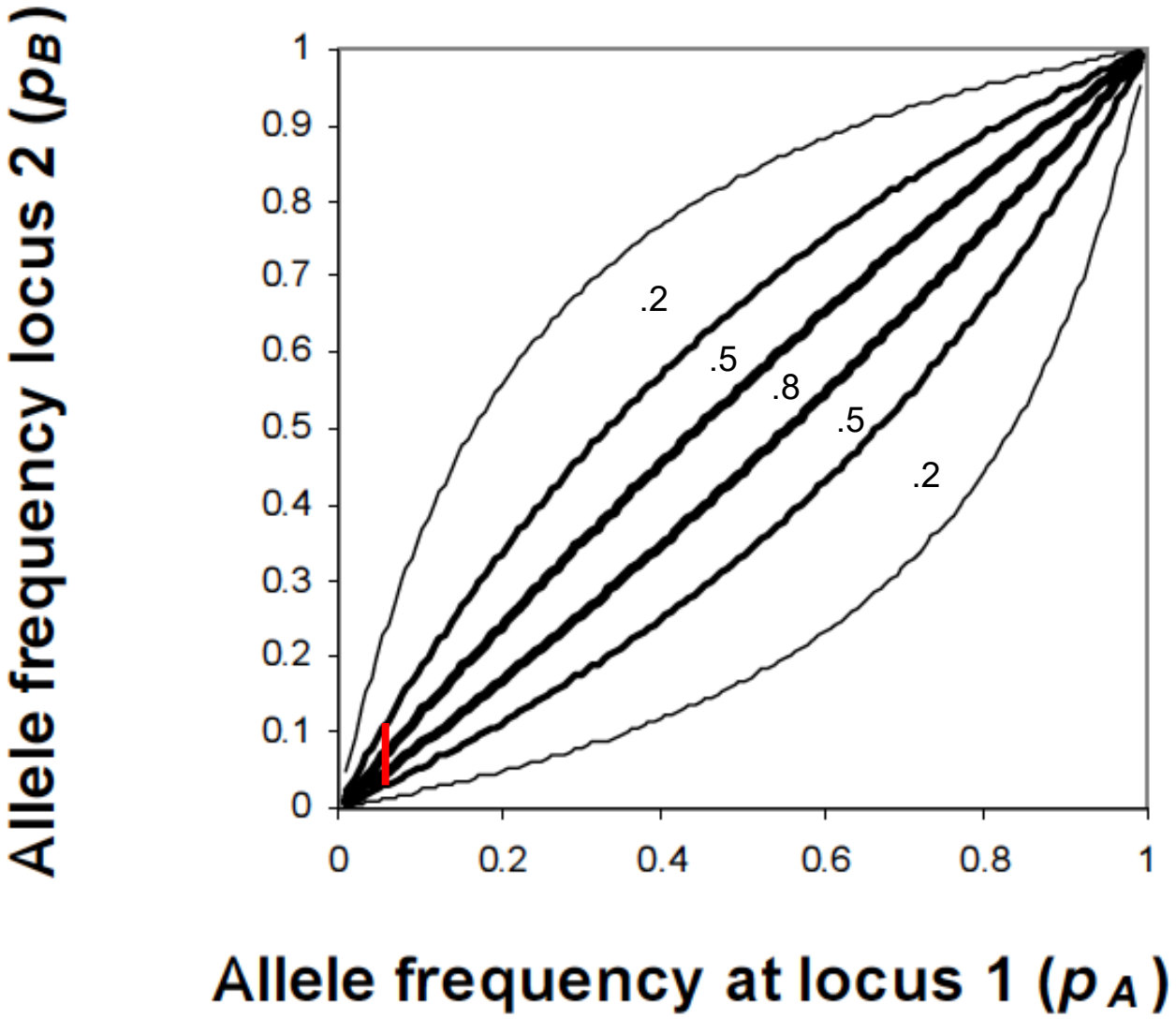
Possible range of allele frequencies given LD (r^2) between 2 SNPs



Wray, *TRHG*, 2006

…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



Wray, *TRHG*, 2006

…AND where the rare allele at SNP 1 is in PHASE with the rare allele at SNP 2

In Phase

	A	a	
B	0.99	0	0.99
b	0	0.01	0.01
	0.99	0.01	

Very high LD

Out of Phase

	A	a	
B	0.98	0.01	0.99
b	0.01	0	0.01
	0.99	0.01	

Very low LD

Why $h^2_{snp} < h^2$ (almost always)

- 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:

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

where

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

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

THE END
(extra slides after)

RUNNING GCTA

SNP QC

- Poor SNP calls can inflate SE and cause downward bias in h^2_{snp}
- Clean data for
  - SNPs missing > ~.05
  - HWE p < 10e-6
  - MAF < ~.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*0.134 = -0.12
0.147-1.96*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

LDAK

ARTICLES

nature genetics

2017

Reevaluation of SNP heritability in complex human traits

Doug Speed[1], Na Cai[2,3], the UCLEB Consortium[4], Michael R Johnson[5], Sergey Nejentsev[6] & 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. Balding[1]

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:

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

- 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 in genes), 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	Consequences
$W_i = 1 \; \forall \; i = 1 \ldots m$	SNPs have equal weight, even if they are poorly imputed and redundantly tag the same CV
$\alpha = -1$	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.

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

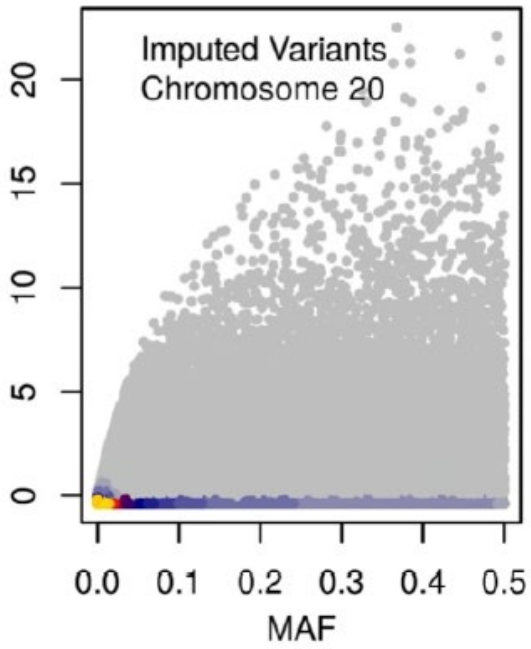
$$V[G_i] = [2p_i(1-p_i)]^{\alpha+1}$$

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

nature genetics

ANALYSIS

https://doi.org/10.1038/s41588-018-0108-x

2018

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

Luke M. Evans[1*], Rasool Tahmasbi[1], Scott I. Vrieze[2], Gonçalo R. Abecasis[3], Sayantan Das[3], Steven Gazal[4,5], Douglas W. Bjelland[1], Teresa R. de Candia[1], Haplotype Reference Consortium[6], Michael E. Goddard[7,8], Benjamin M. Neale[5], Jian Yang[9], Peter M. Visscher[9] and Matthew C. Keller[1,10*]

LDMS-I

nature genetics

2015

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

Jian Yang[1,2,24], Andrew Bakshi[1], Zhihong Zhu[1], Gibran Hemani[1,3], Anna A E Vinkhuyzen[1], Sang Hong Lee[1,4], Matthew R Robinson[1], John R B Perry[5], Ilja M Nolte[6], Jana V van Vliet-Ostaptchouk[6,7], Harold Snieder[6], The LifeLines Cohort Study[8], Tonu Esko[9–12], Lili Milani[9], Reedik Mägi[9], Andres Metspalu[9,13], Anders Hamsten[14], Patrik K E Magnusson[15], Nancy L Pedersen[15], Erik Ingelsson[16,17], Nicole Soranzo[18,19], Matthew C Keller[20,21], Naomi R Wray[1], Michael E Goddard[22,23] & Peter M Visscher[1,2,24]

LDMS-R

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

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

- The range of MAF and range of LD will be smaller within a particular MAF/LD bin. As the MAF & LD range shrink for a given MAF/LD bin k of SNPs (M_k) and CVs (Q_k),

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

  and thus

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

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 within each bin, where SNPs take the average LD of SNPs in the surrounding ~ 200kb region.
- However, SNPs with individually low LD that exist in regions of high LD explain more variation (Gazal et al, *Nature Genetics*, 2017)
- Thus, LDMS-I (unlike LDMS-R) uses each individual SNP's LD score for binning
- Because SEs tend to be ~2.5x larger than single-GRM estimates, both require large sample sizes (e.g., N > 30k) and therefore large amounts of RAM (e.g., >100 Gb)

RUNNING LDMS-I

Create LD quartiles

GCTA LD command:

test.bed, test.bim, test.fam

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

```
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

Text file

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

--pheno <path>/test.phen
--covar <path>/test.covar
--qcovar <path>/test.qcovar
--reml --out Multi.SNPgrm

As before

--thread-num 20

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	

$h2_{SNP}$ = 0.15+0.06+0.32 = 0.5391

GREML vs. LDAK vs. LDMS-I

nature genetics

ANALYSIS

https://doi.org/10.1038/s41588-018-0108-x

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

Luke M. Evans[1*], Rasool Tahmasbi[1], Scott I. Vrieze[2], Gonçalo R. Abecasis[3], Sayantan Das[3], Steven Gazal[4,5], Douglas W. Bjelland[1], Teresa R. de Candia[1], Haplotype Reference Consortium[6], Michael E. Goddard[7,8], Benjamin M. Neale[5], Jian Yang[9], Peter M. Visscher[9] and Matthew C. Keller[1,10*]

- 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 $\hat{h^2}_{SNP}$ 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) $\hat{h^2}_{SNP}$

Simulation of phenotypes

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

λ-LD relationship

$u_i \sim N(0,1)$ or $u_i \sim N(0,w_i)$

λ-MAF (α=-.25 or α=-1)	none (0)	negative (-)
negative (-)	Typical GCTA assump. V(Gi) invariant	V(Gi) decreases w/ LD
weak (~0)	V(Gi) increases w/ MAF	Typical LDAK assump. V(Gi) increases w/ MAF V(Gi) decreases w/ LD

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

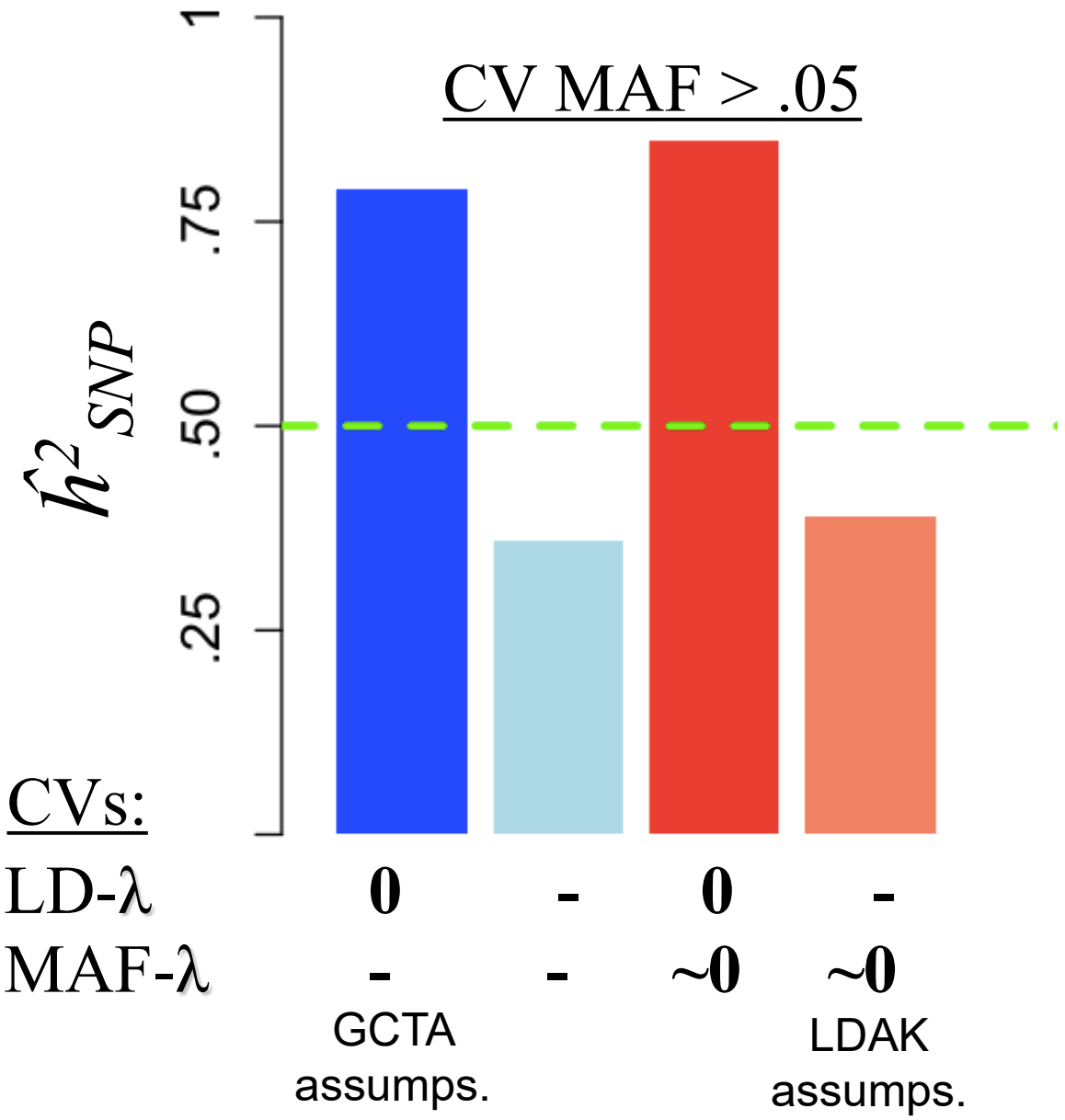
3 Estimation Methods Compared

- GREML-SC: predictor is a single GRM (aka, "GCTA approach"). GRM built as usual from all imputed SNPs with MAC > 5 & imputation $r^2 > .3$

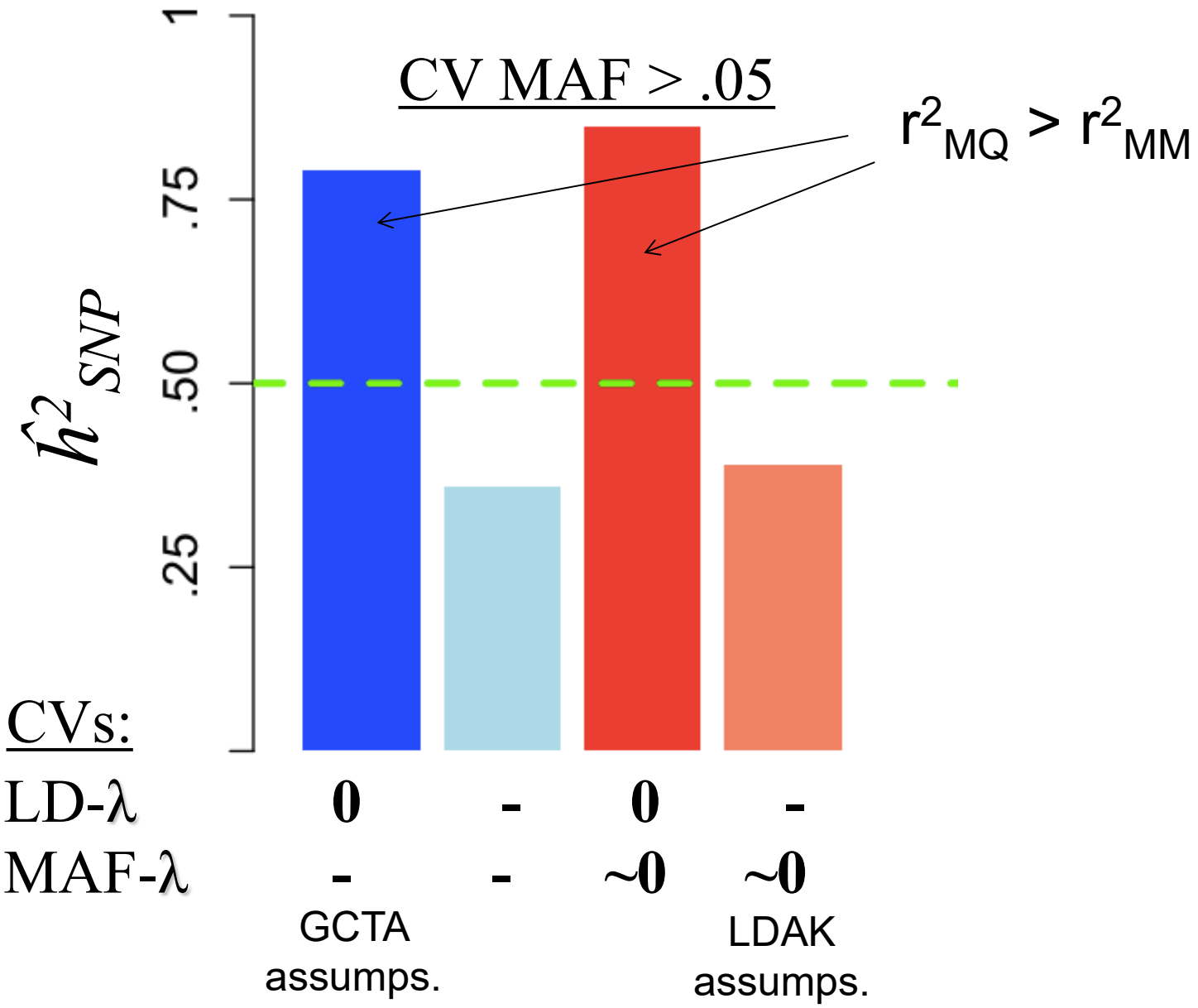
- LDAK: predictor is a single GRM from imputed SNPs and weighted by LD and imputation r^2.

- 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. $\hat{h}^2_{SNP} = \Sigma(\hat{h}^2_{SNP\_k})$

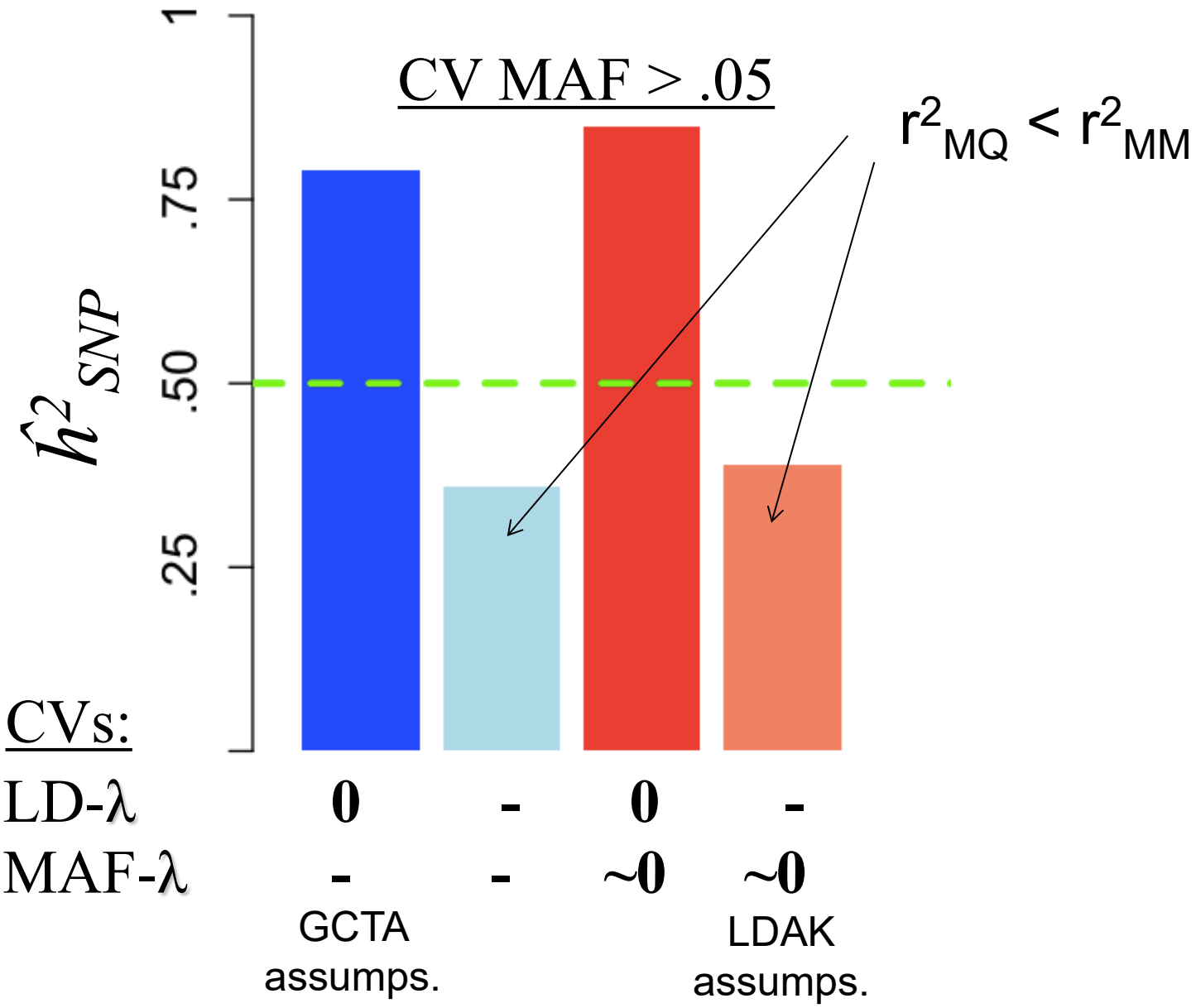
GREML-SC results



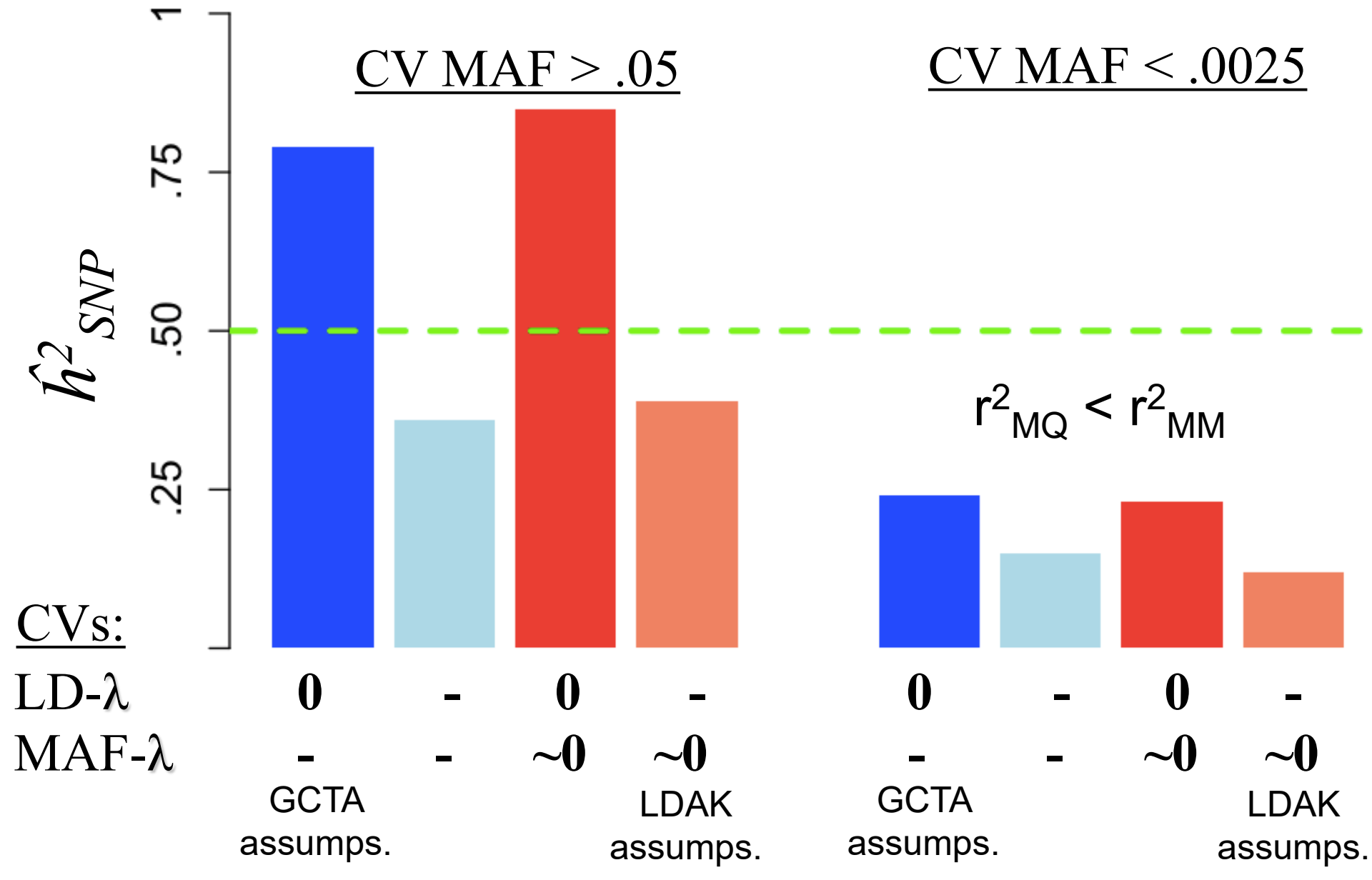
GREML-SC results



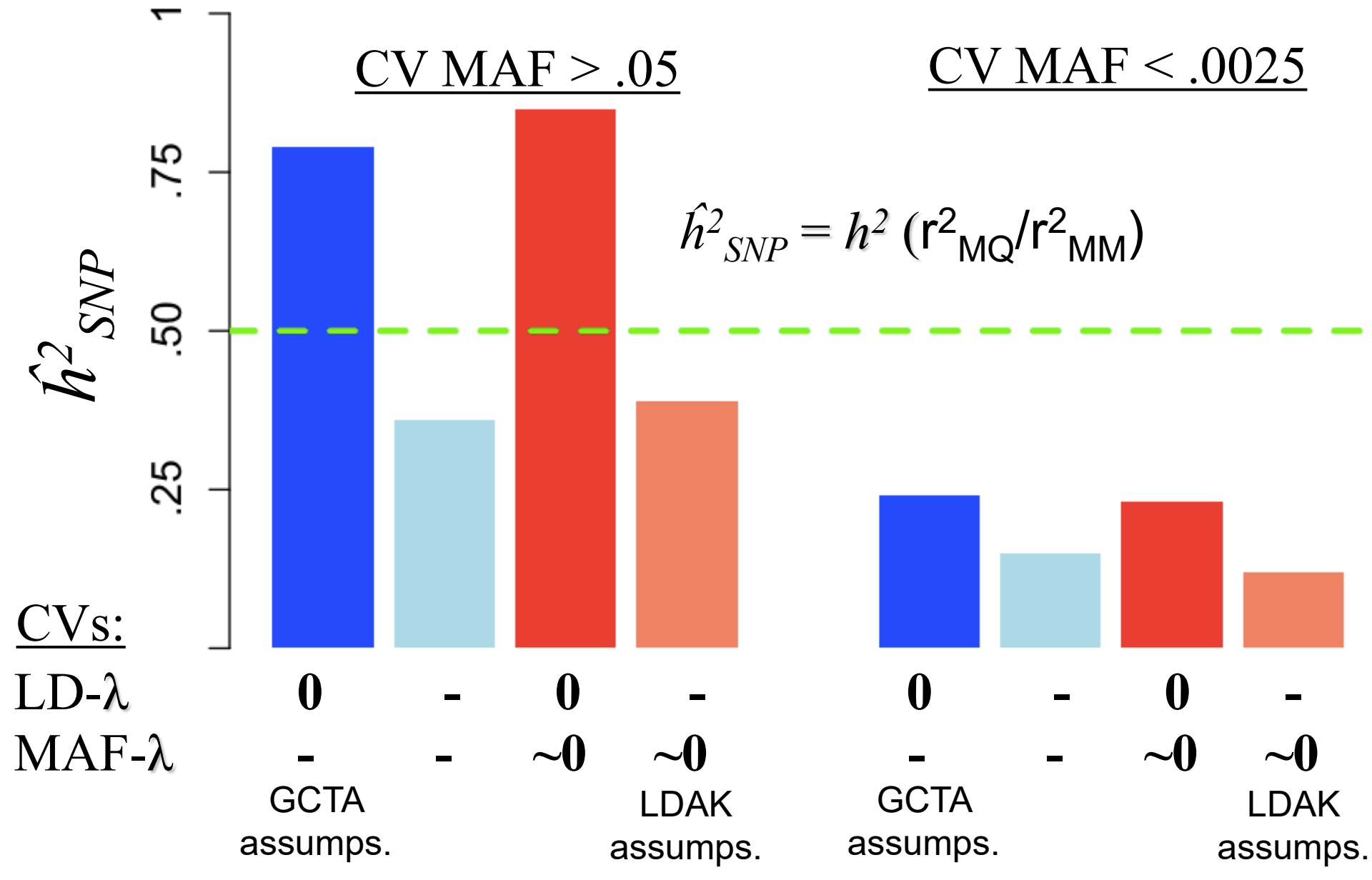
GREML-SC results



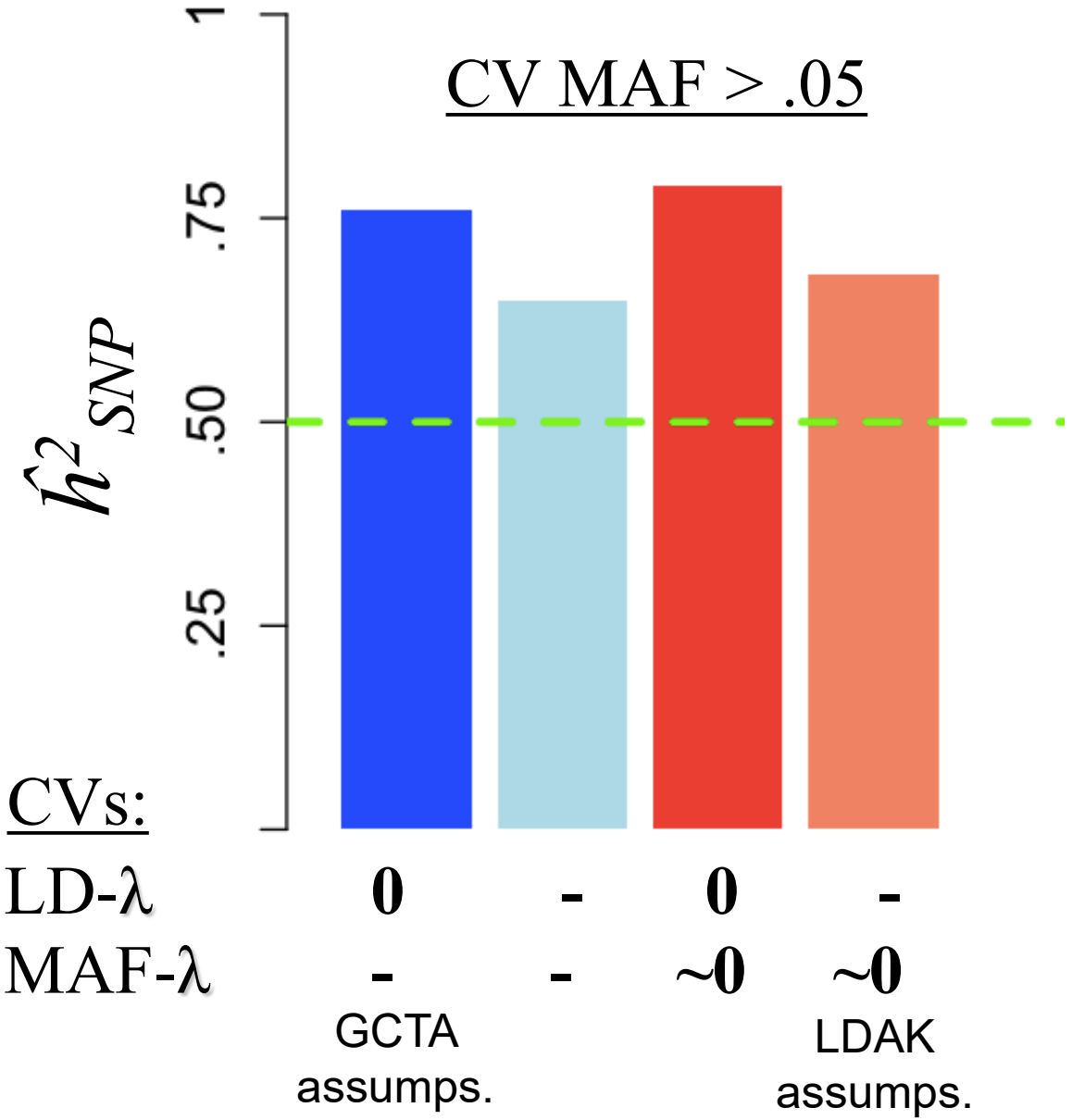
GREML-SC results



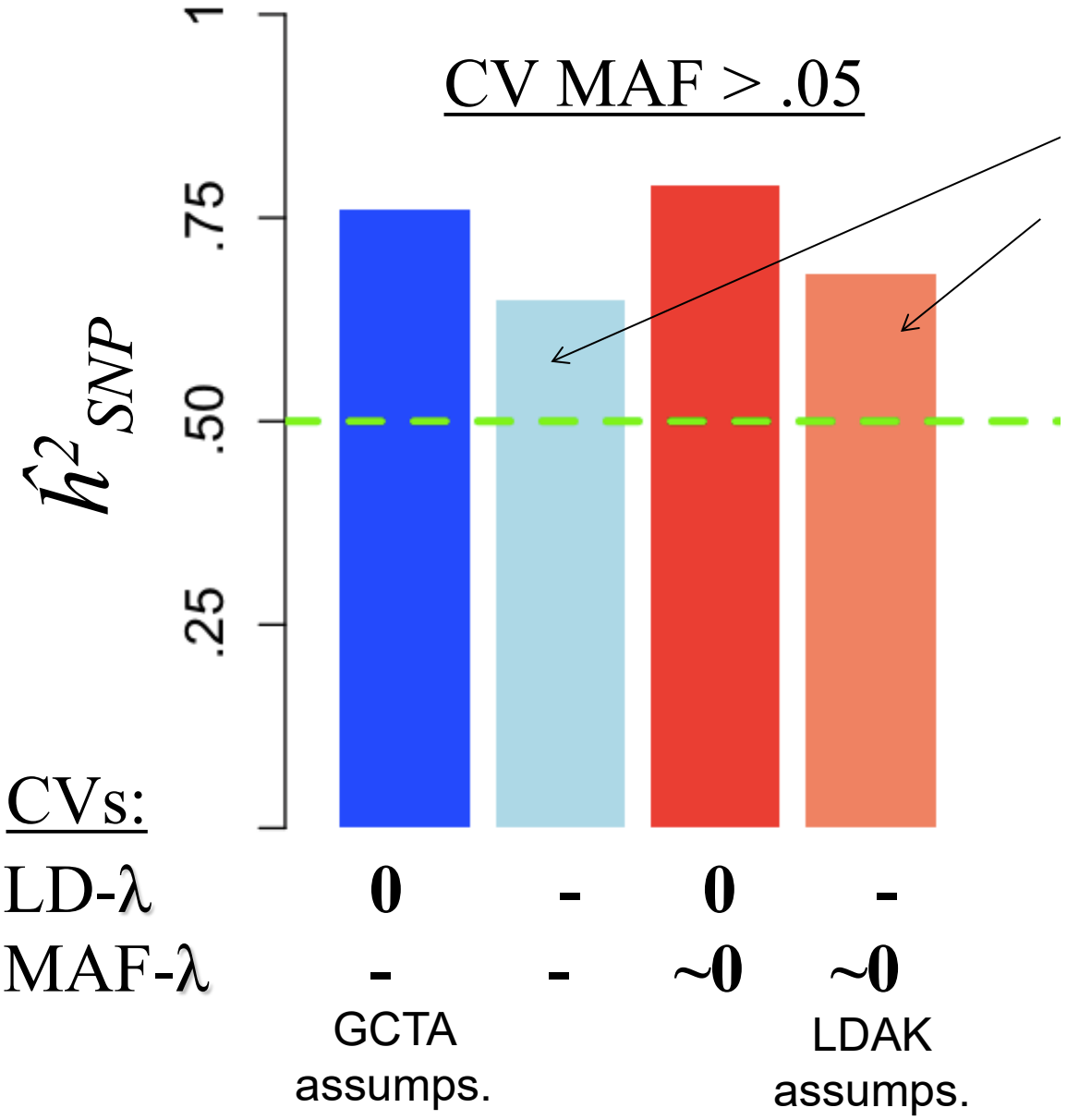
GREML-SC results



LDAK results

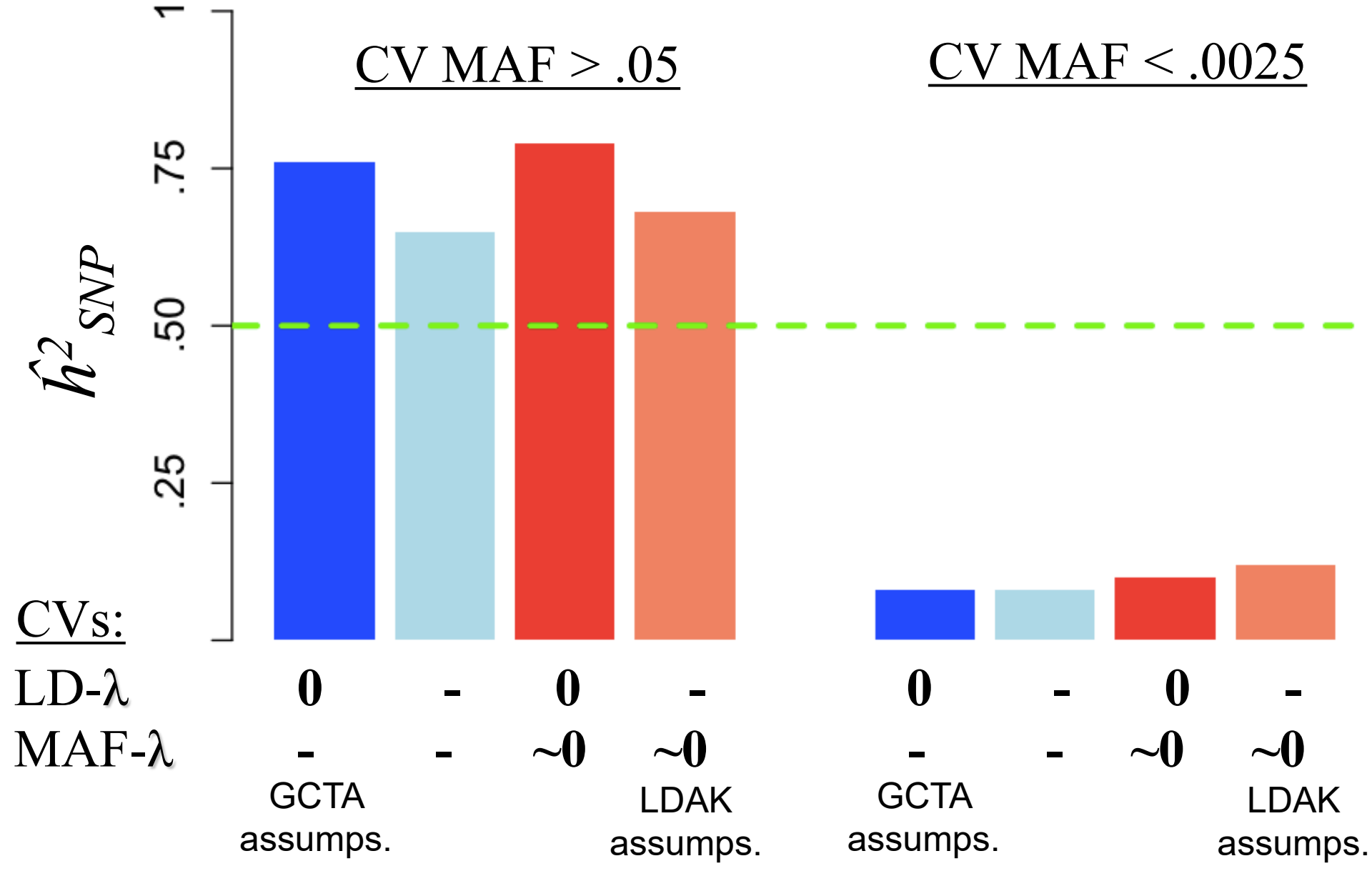


LDAK results

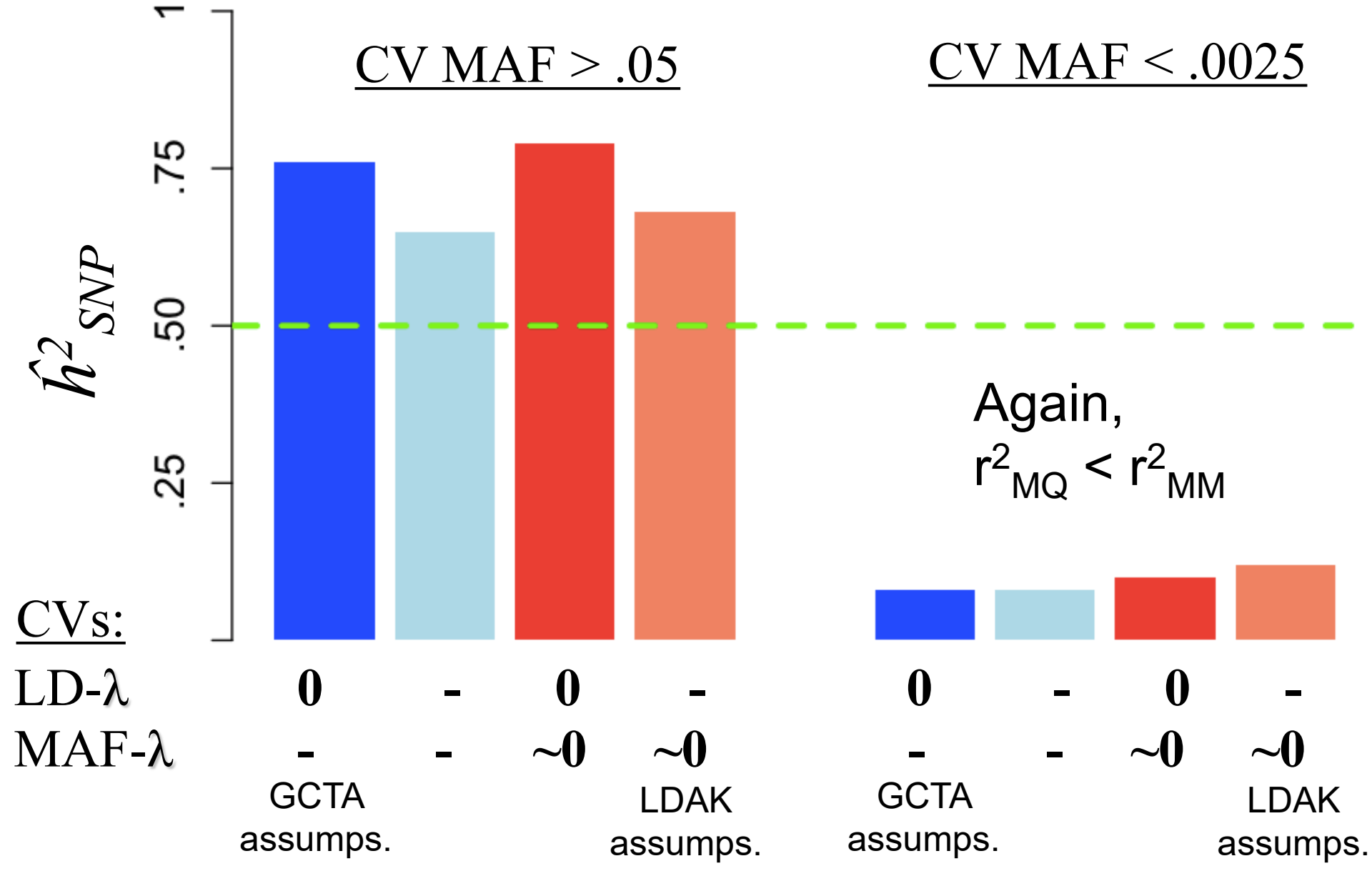


upweights low LD SNPs but this is partially offset by downweighting poorly imputed ones

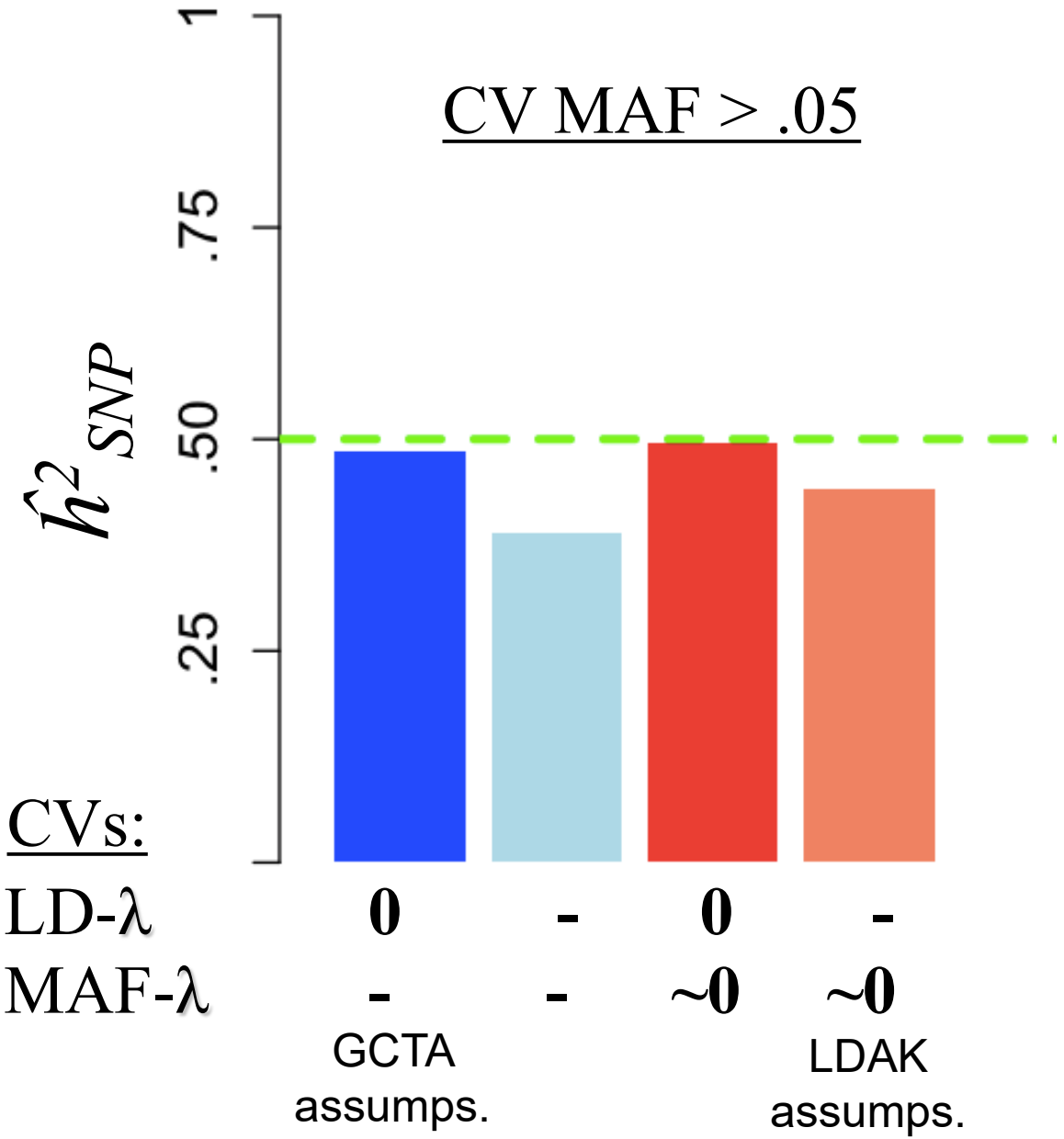
LDAK results



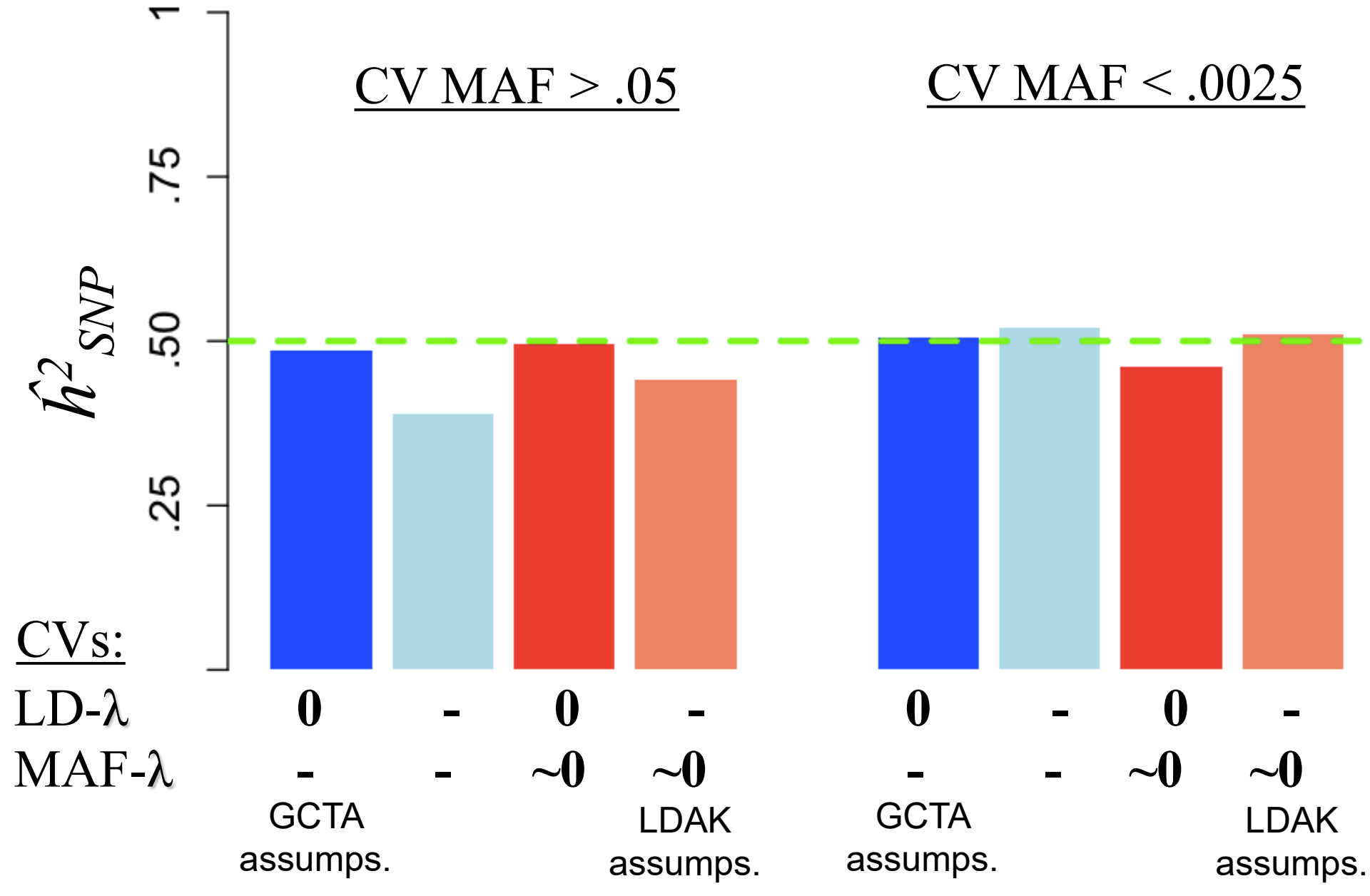
LDAK results



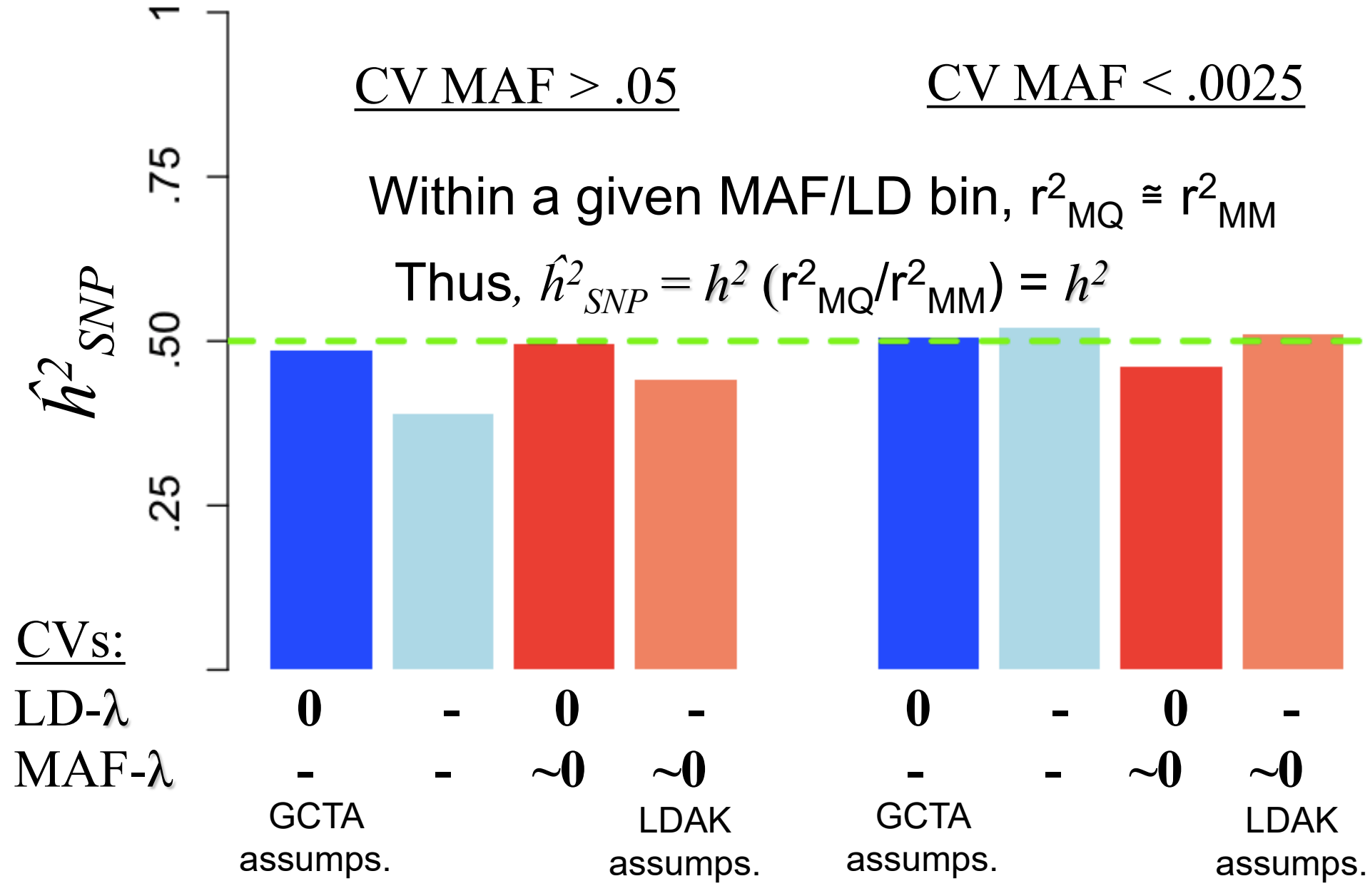
GREML-LDMS-I results



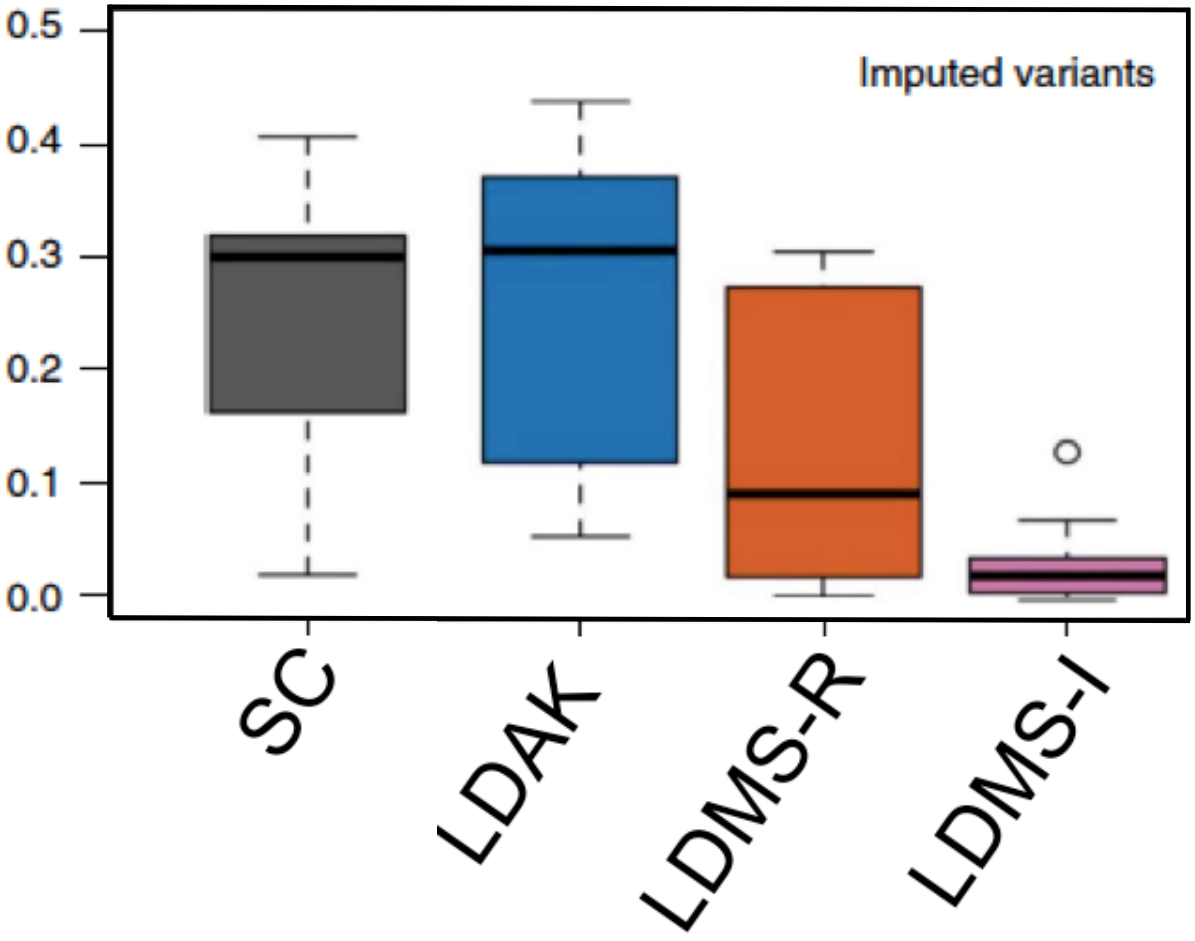
GREML-LDMS-I results



GREML-LDMS-I results



Absolute Bias Across 4 Methods and hundreds of genetic architectures



Regarding LD-Score regression

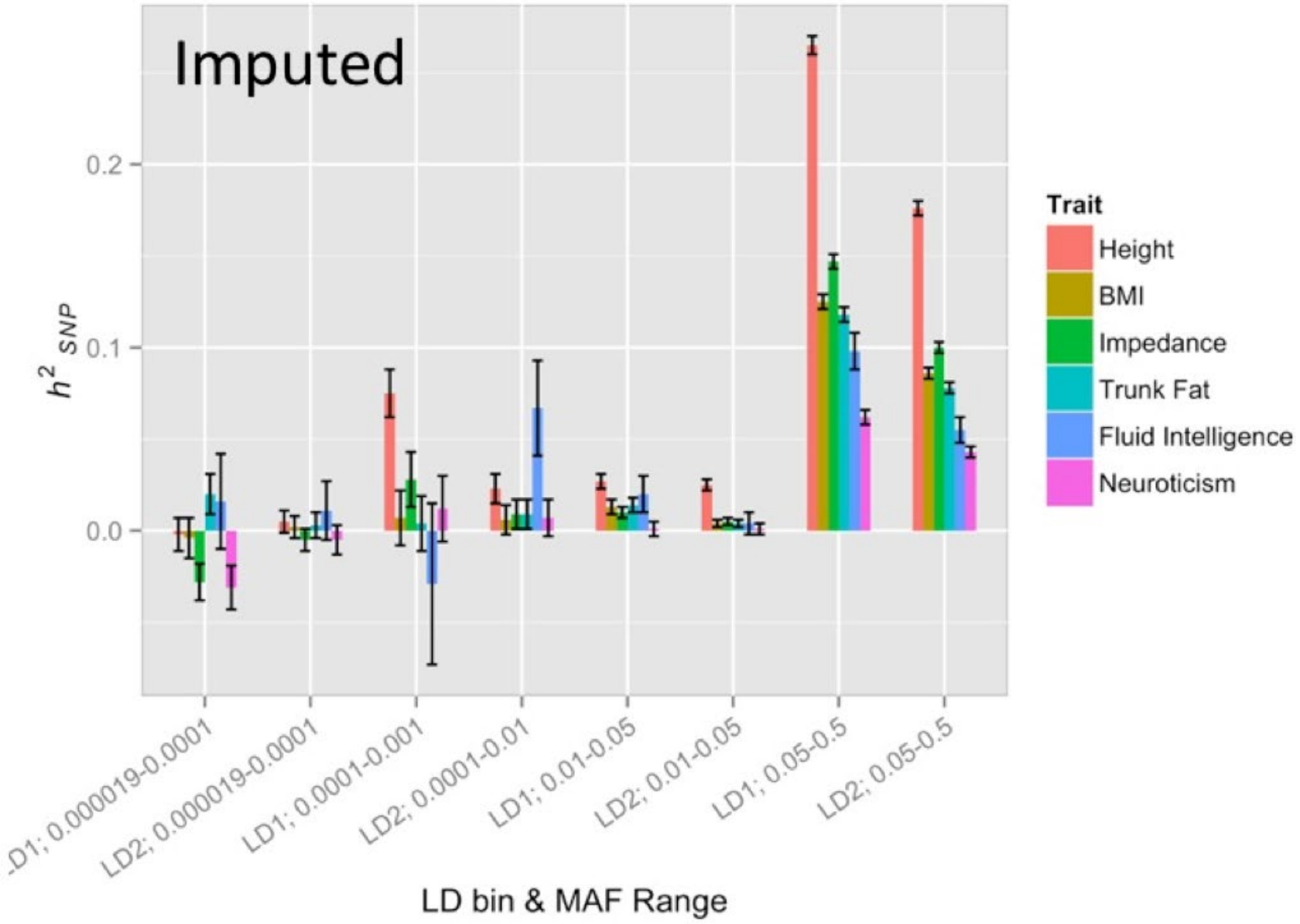
- LD-score regression is robust to stratification and sample overlap. However:
  - it cannot estimate h^2 due to rare CVs, even when using imputed/WGS data
  - it is sensitive to assumptions about LD-λ
  - should provide a lower-bound of $\hat{h}^2_{SNP}$ from other methods
- So long as genetic covariance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

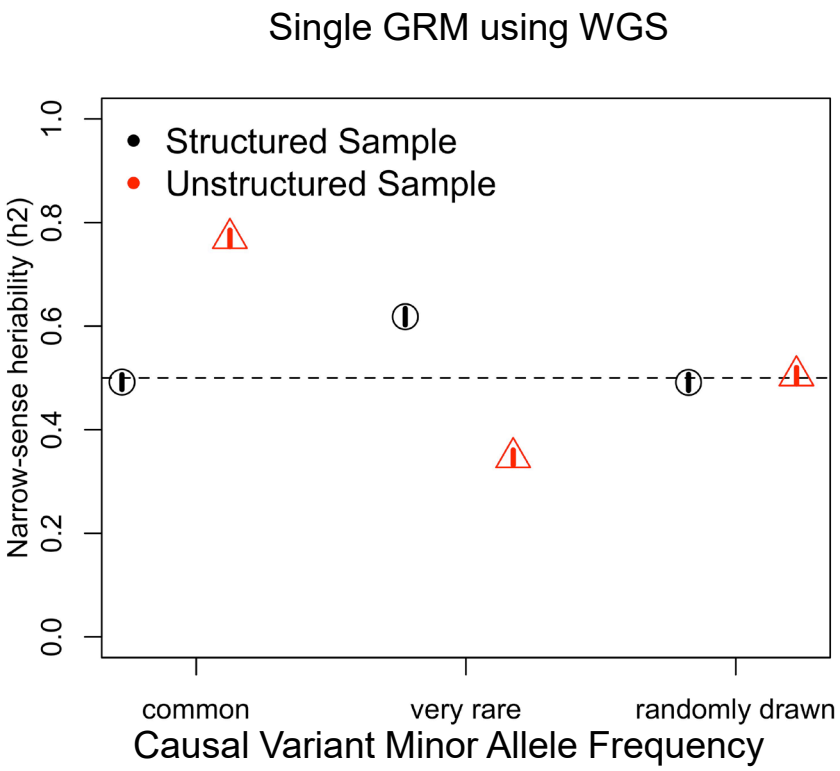
Effects of stratification on r^2_{QM}/r^2_{MM}

- 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.

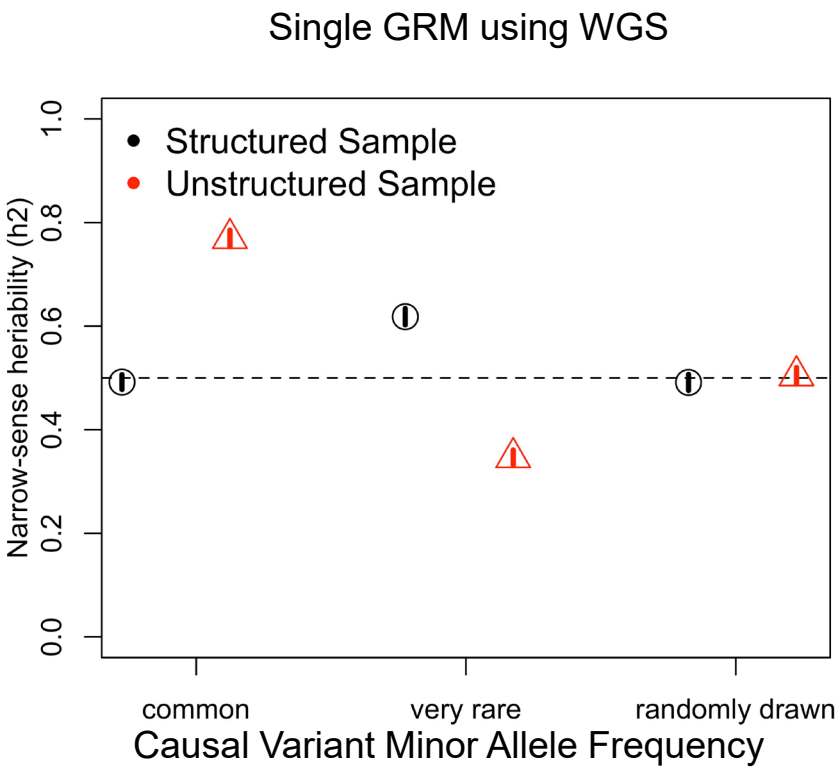
Effects of stratification on r^2_{QM}/r^2_{MM}

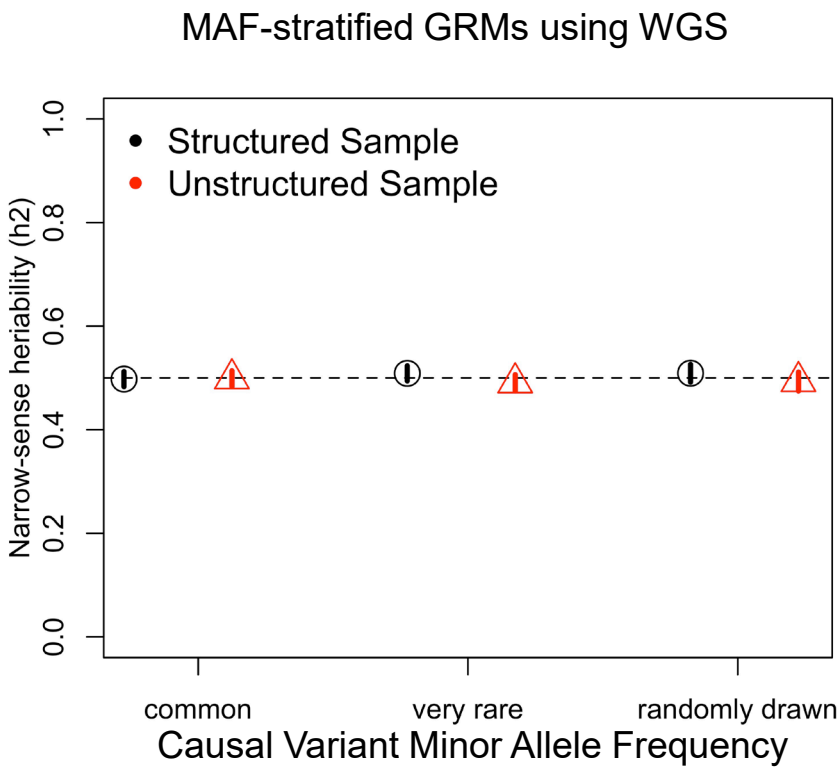
- In general, stratification inflates long-range r^2 between SNPs. However, within a given MAF bin, the ratio of r^2_{QM}/r^2_{MM} is ~ 1 because SNP-SNP & SNP-CV LDs are inflated similarly.
- However, across CVs and SNPs of different MAF, stratification induces differences in r^2_{QM} & $r^2_{MM.}$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



Which once again was corrected by using LDMS GREML





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 pihat

- 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*: $E[\hat{\pi}_{jk} | Z_j Z_k] = \frac{rh^2}{m} < \frac{1}{m}$

  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)

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

*Robinson et al., *Nature Human Behavior*, 2017
Yengo et al., *BioarXiv*, 2018

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 $Z \sim X_1 + X_2 + \ldots X_m$ for all m SNPs fit simultaneously as $n \to \infty$
- 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}

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

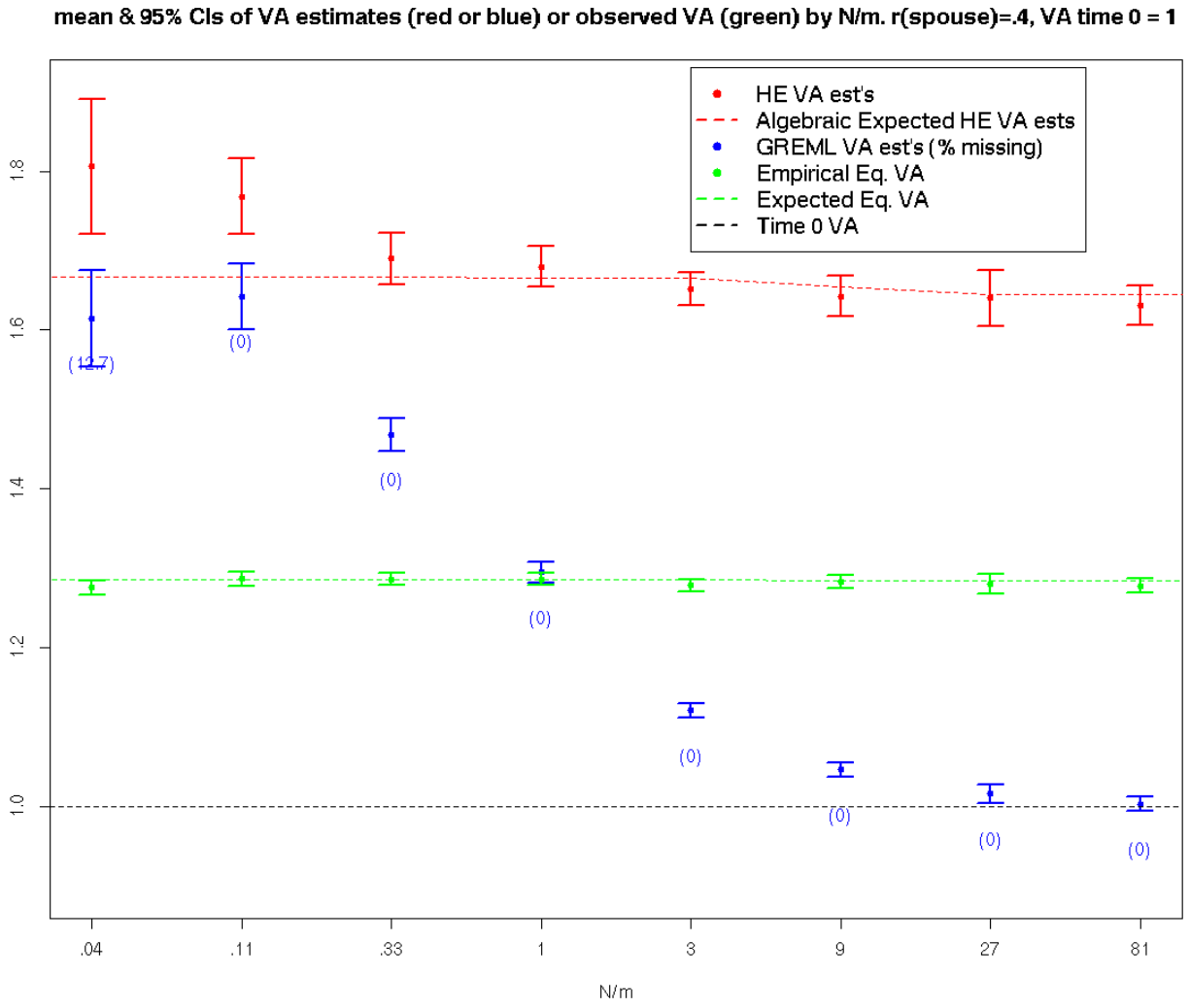
$$E[Z_i Z_j \mid \hat{\pi}_{ij}] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{ij}$$

$$\cong \frac{h^2_{snp}}{M} + 2M\delta$$

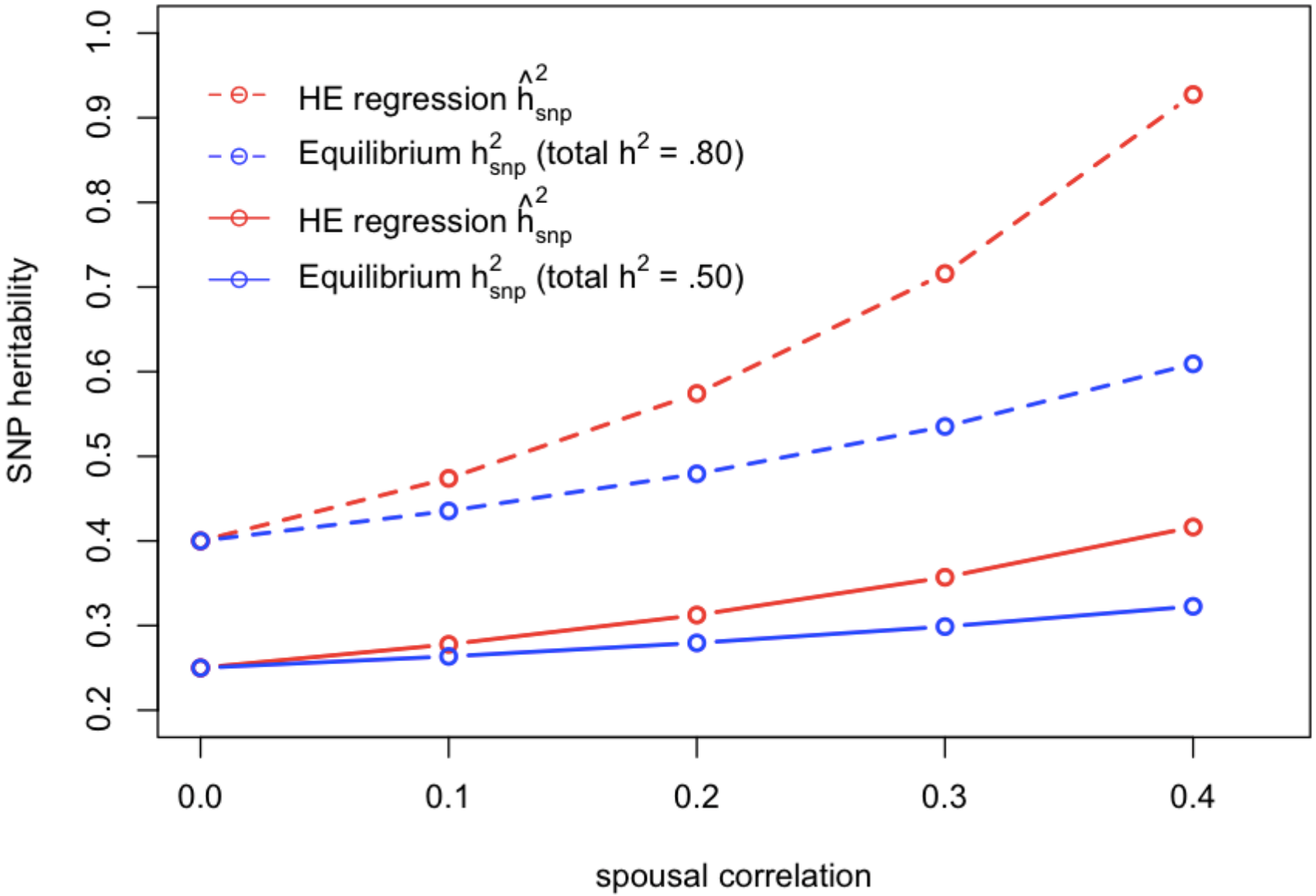
$$\hat{\beta}_1 = \frac{COV(Z_i Z_j, \hat{\pi}_{ij})}{V(\hat{\pi}_{ij})} = \hat{h}^2_{snp}$$

$$\cong \frac{1}{M} + \left(\frac{2M\delta}{h^2}\right)^2$$

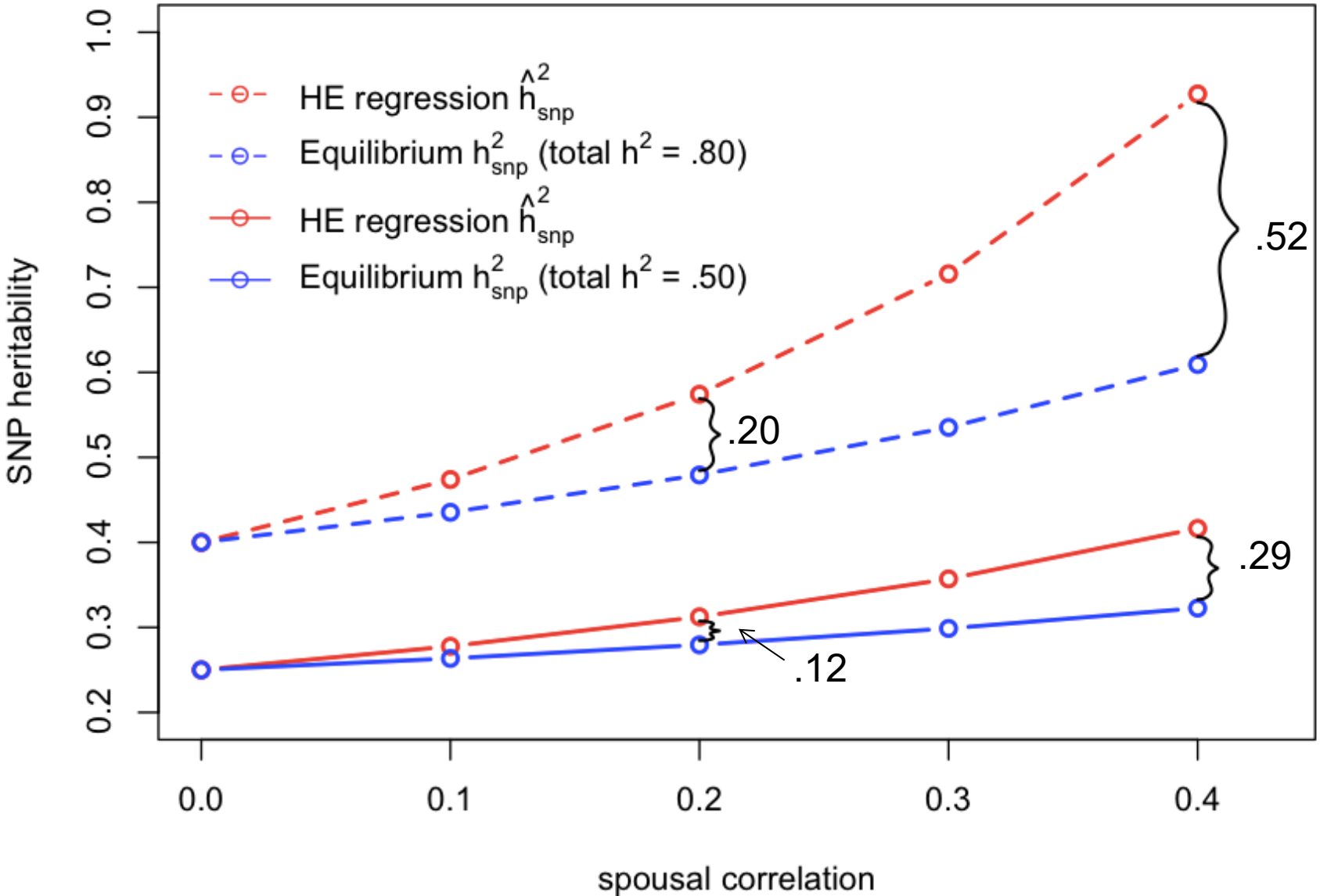
We don't predict HE estimates to change as a function of m or N. GREML estimates are clearly a function of N/m, which occurs because when N>>m, the effects of each SNP are separable.



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

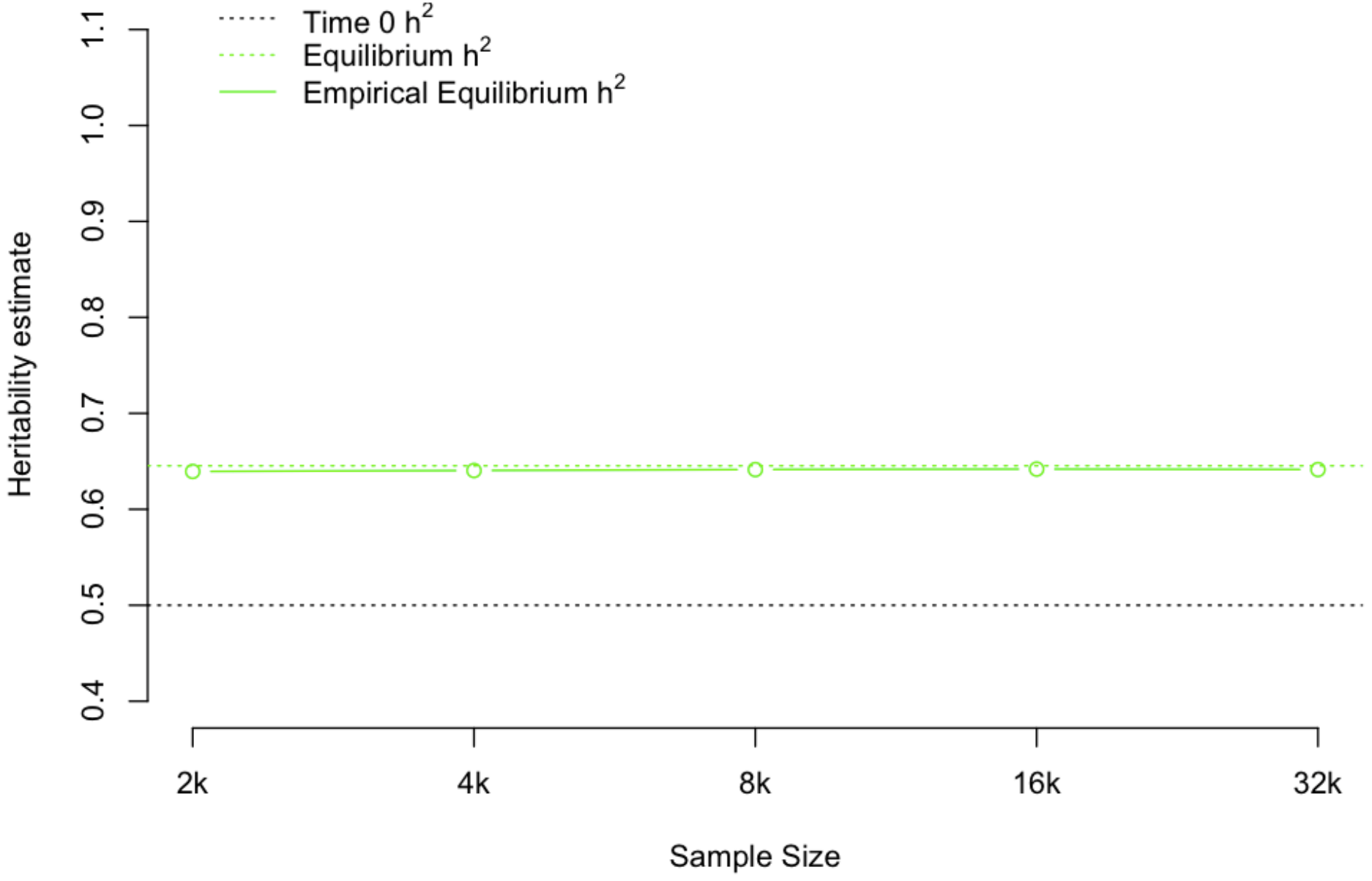
Trait	r(spouse)	h^2_{ETFD} from literature	h^2_{snp} from literature	corrected h^2_{snp}	% Over-estimated
Extraversion	.01	.23	.15	.15	0
Neuroticism	.08	.24	.16	.155	.03
Height	.20	.70	.45	.39	.15
IQ	.35	.62	.35	.28	.25
Political Pref.	.48	.26	.18	.15	.20

Simulations

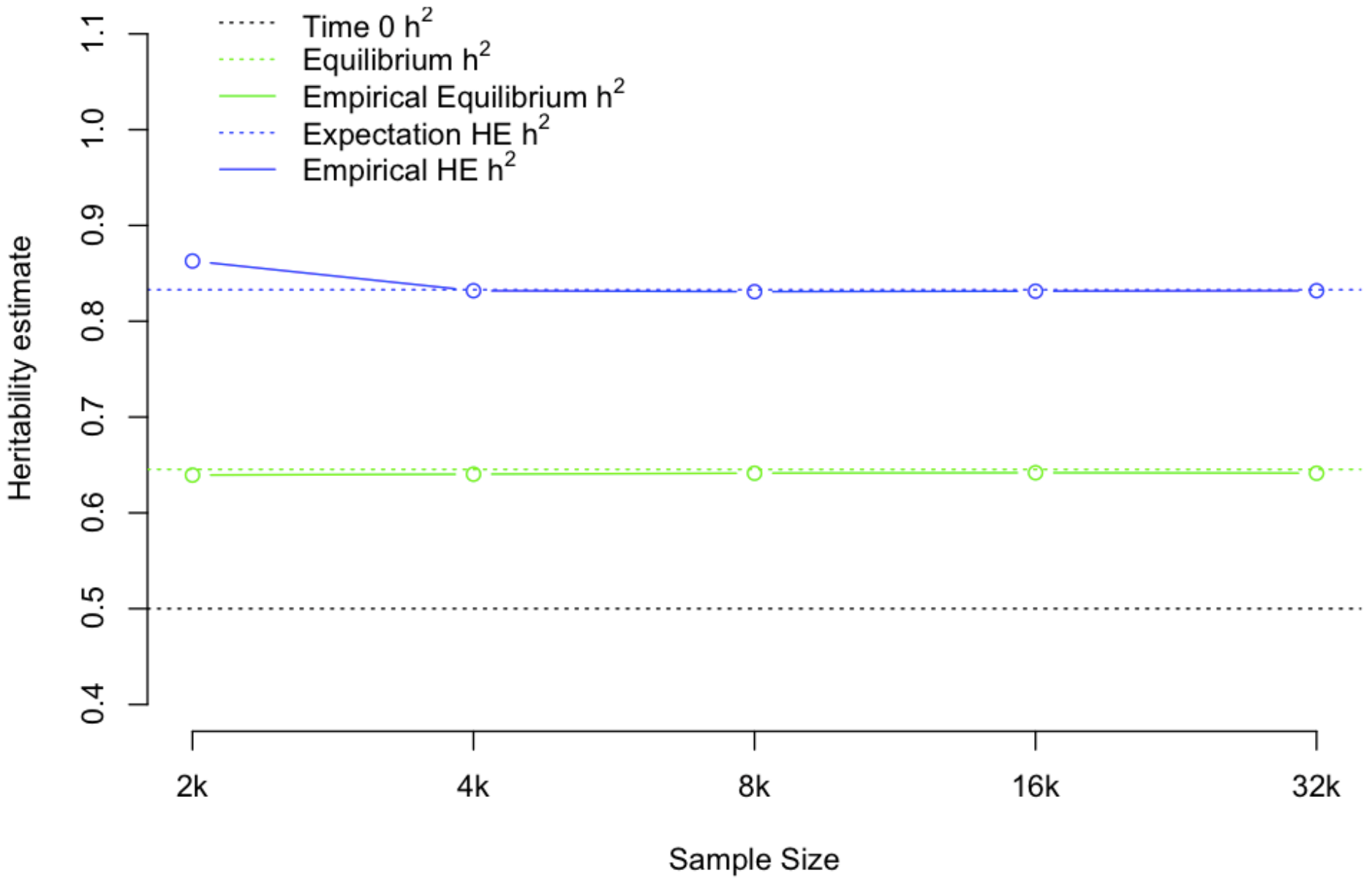
Rasool Tahmasbi

- 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

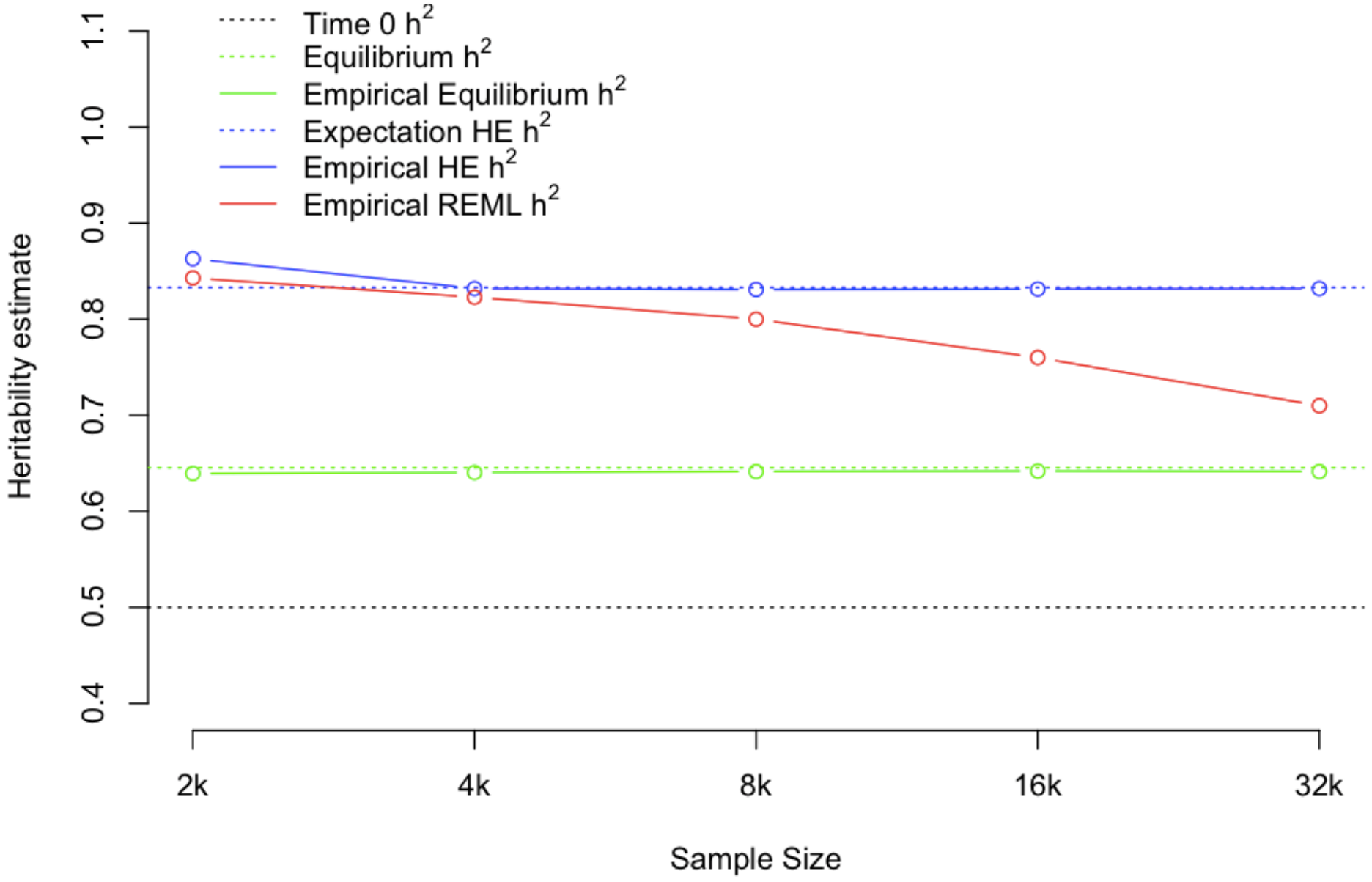
GeneEvolve Simulation Results



GeneEvolve Simulation Results



GeneEvolve Simulation Results

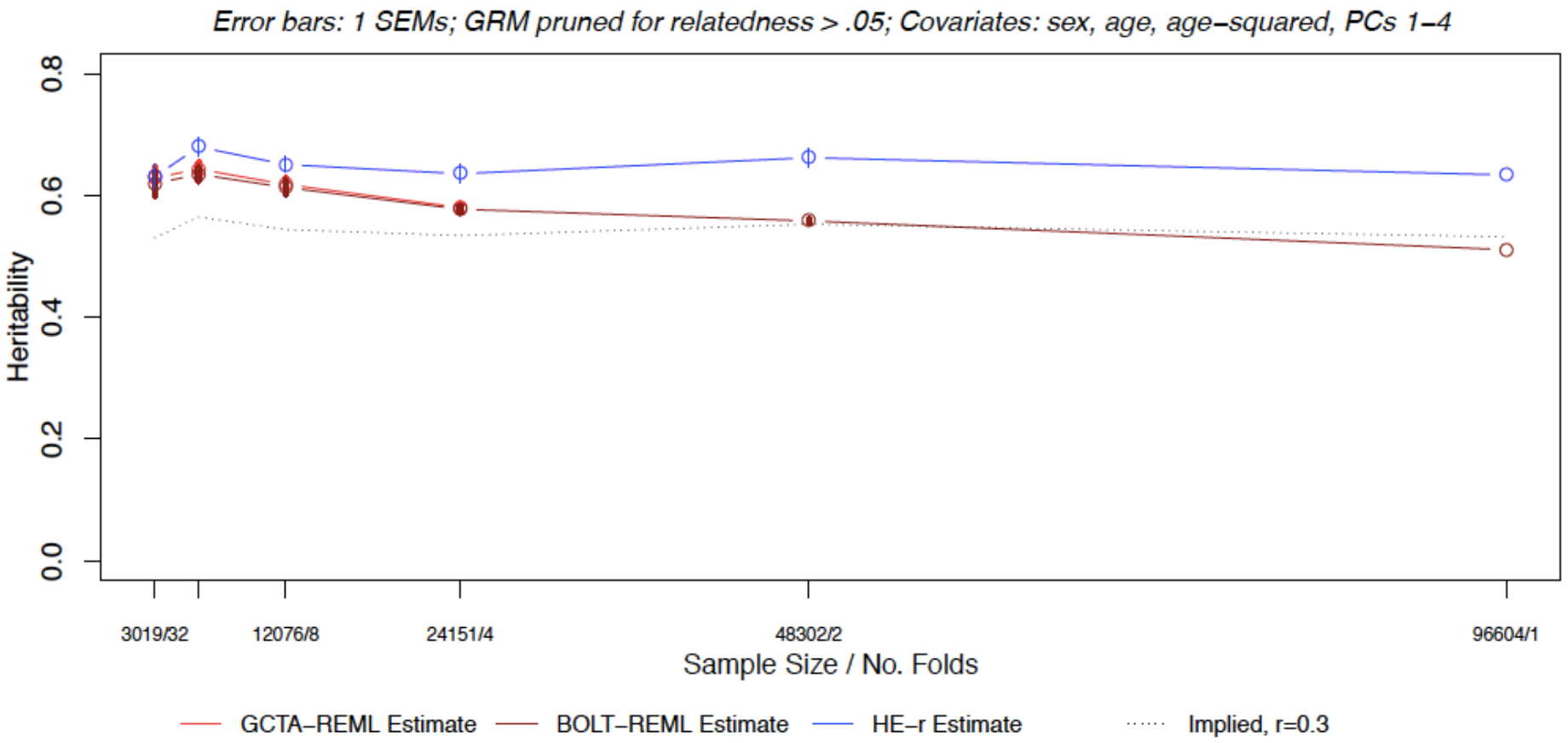


HE vs. GREML estimates

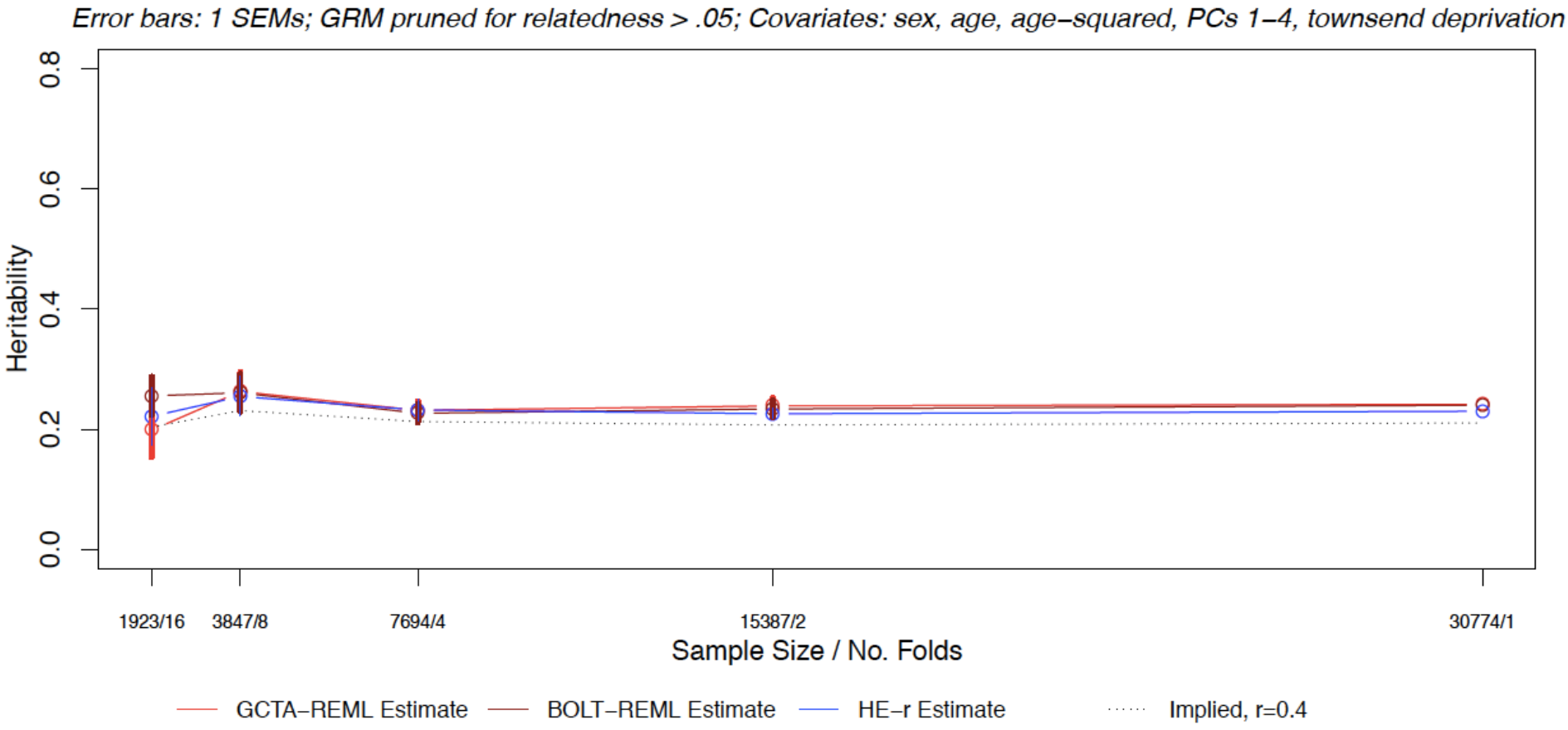
- For most realistic situations, m>>N, and thus GREML and HE estimates are similar: both over-estimate equilibrium h^2_{snp}
- We can vary N (holding m constant) to see if AM is biasing estimates in real data

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

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



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



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 ~ .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)

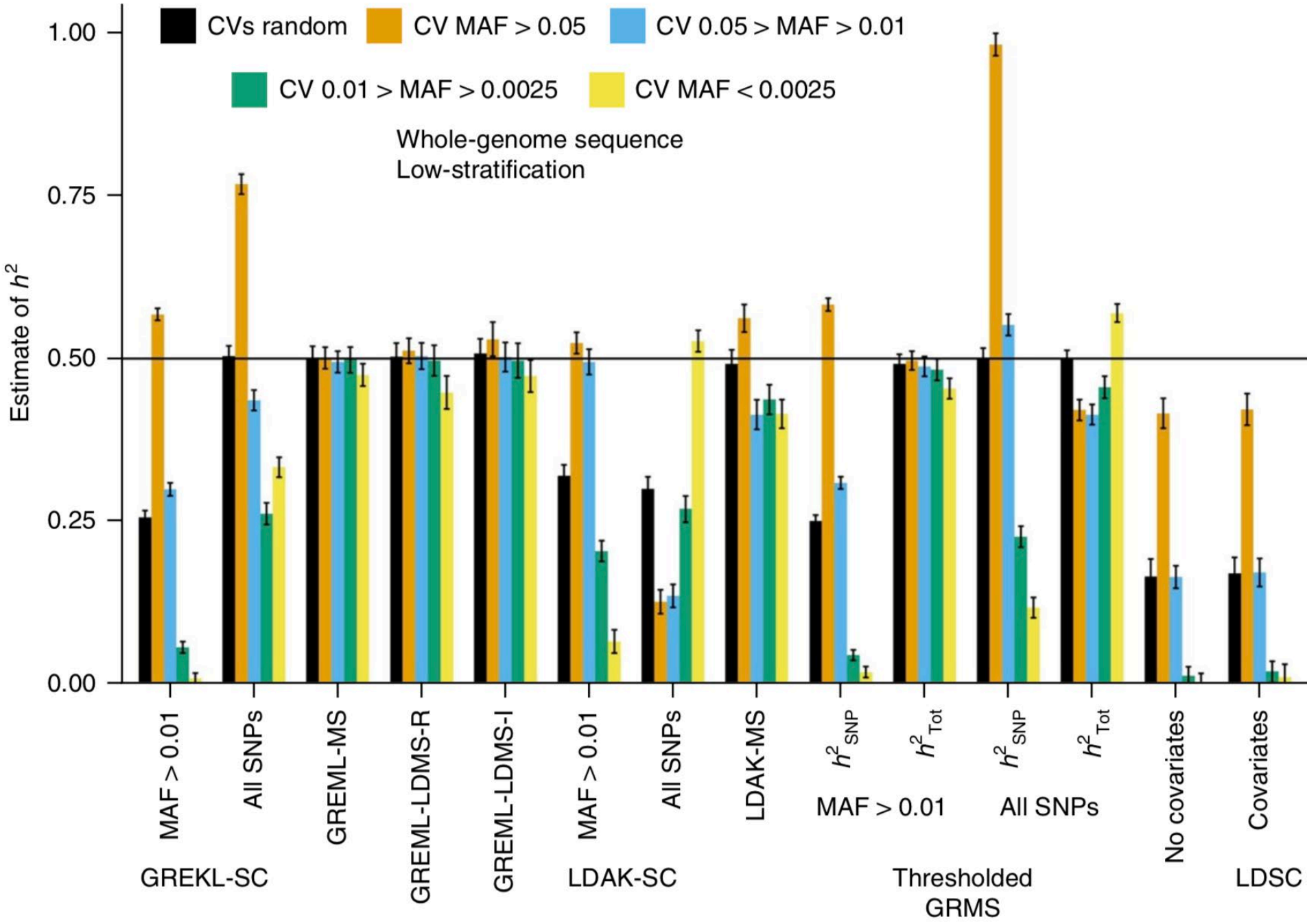

CVs random
CV MAF > 0.05
CV 0.05 > MAF > 0.01
CV 0.01 > MAF > 0.0025
CV MAF < 0.0025
Whole-genome sequence
Low-stratification
Estimate of h^2
1.00
0.75
0.50
0.25
0.00
MAF > 0.01
All SNPs
GREML-MS
GREML-LDMS-R
GREML-LDMS-I
MAF > 0.01
All SNPs
LDAK-MS
h^2_{SNP}
h^2_{Tot}
h^2_{SNP}
h^2_{Tot}
No covariates
Covariates
GREKL-SC
LDAK-SC
MAF > 0.01
All SNPs
Thresholded GRMS
LDSC

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

Method	Description	Major assumptions	Simulation findings regarding $\hat{h}^2_{SNP}$	Computational issues
GREML-SC[5]	Often called the GCTA approach. Originally applied to common array SNPs only. Estimates $\hat{h}^2_{SNP}$, 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.
LDAK-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.
LDAK-MS[15]	A multicomponent extension of LDAK-SC that bins SNPs by MAF.	Requires that the same assumptions of LDAK-SC hold within each GRM.	Less biased on average than LDAK-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.