Marker heritability

Biases, confounding factors, current methods, and best practices

Luke Evans, Matthew Keller

Background – What Matt Keller presented

- GREML-SC: single genetic relatedness matrix (GRM) to estimate heritability (h^2_{SNP})
- Relate allele sharing at genome-wide SNPs to phenotypic similarity
- *Genetic Relatedness Matrix (GRM) is a proxy for allele sharing at causal variants (CVs)*

Background – GCTA-style approach

- Unrelated individuals (e.g., $A_{ij} < 0.05$)
- Common markers from SNP arrays
  - (e.g., MAF > 0.05, m = 500,000 – 2.5M SNPs)
- Low-moderate stratification in samples
  - E.g., UK Biobank, GoT2D, AMD
  - Homogeneous populations, e.g., North Finland Birth Cohort, Sardinia

Background – GCTA-style approach

GRM:

- $A_{ij} = \frac{1}{m}\sum_{k}^{m}\frac{(x_{ik}-2p_k)(x_{jk}-2p_k)}{2p_k(1-p_k)}$

Phenotype:

- $y_i = g_i + e_i$
- $var(\boldsymbol{y}) = \boldsymbol{A}\sigma_v^2 + \boldsymbol{I}\sigma_e^2$

$h^2_{SNP} = \sigma^2_v / (\sigma^2_v + \sigma^2_e)$

Examine biases using real genotypes and simulated phenotypes

Genotypes: Haplotype Reference Consortium whole genome sequences

- Relatively homogeneous subset
- Build GRM from
  - Axiom array positions only
  - Whole genome sequence variants
  - Vary the MAF of markers for GRM

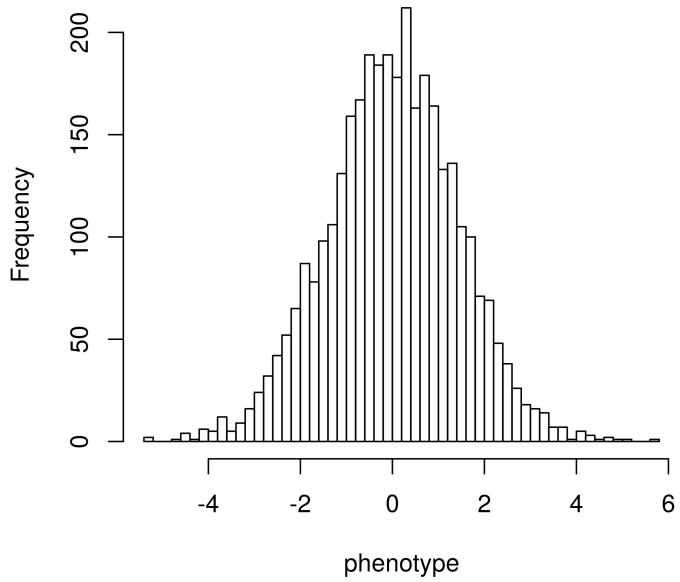
GRM ELEMENTS:

$$A_{ij} = \frac{1}{m}\sum_{k}^{m}\frac{(x_{ik}-2p_k)(x_{jk}-2p_k)}{2p_k(1-p_k)}$$

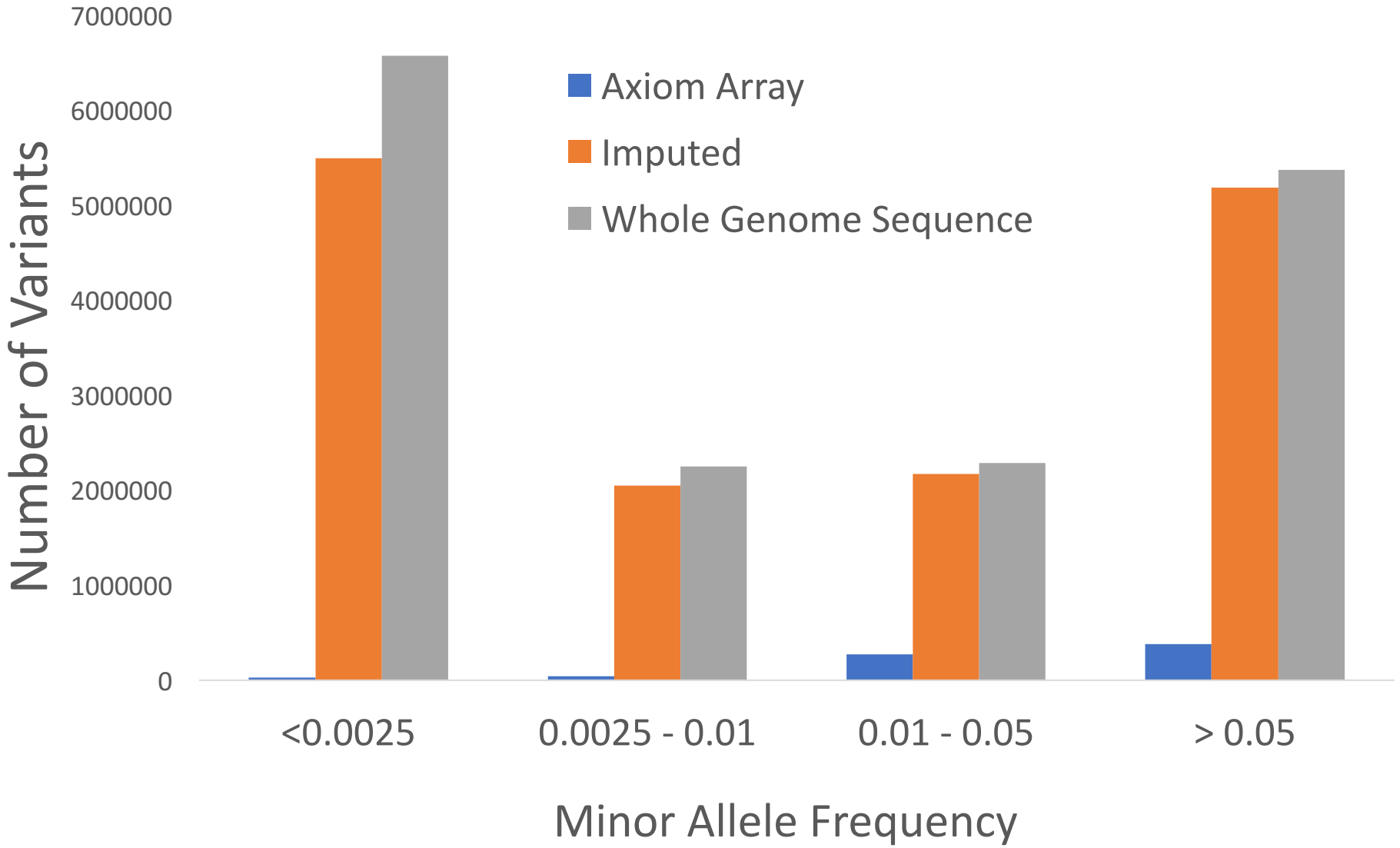
Examine biases using real genotypes and simulated phenotypes

Phenotypes: Simulated from whole genome sequence

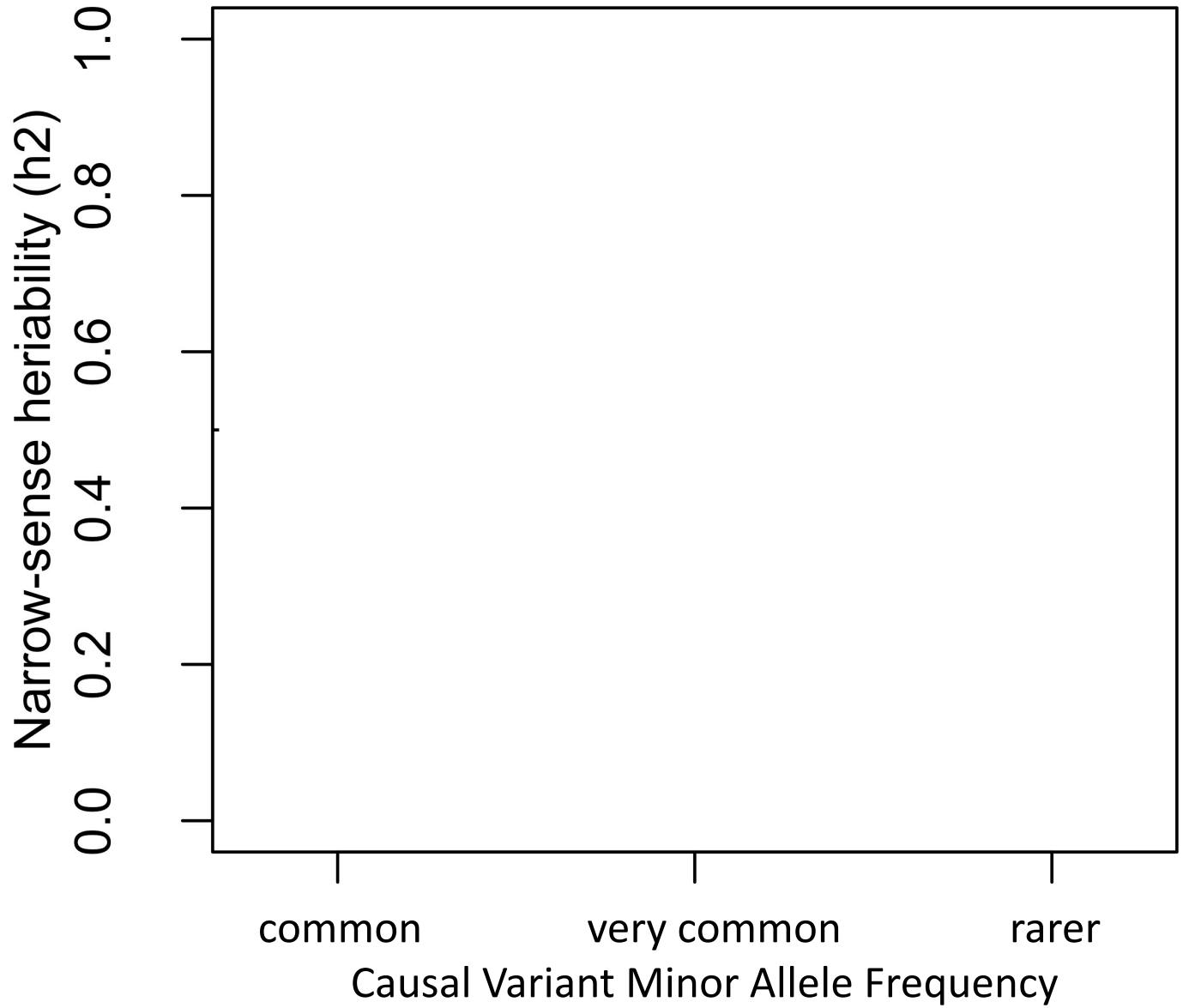
- 1,000 CVs drawn randomly from sequence data
- Vary the MAF of CVs
- $y_i = g_i + e_i$
- $g_i = \sum w_{ik} \beta_k$



CVs from whole genome sequence include many rare variants

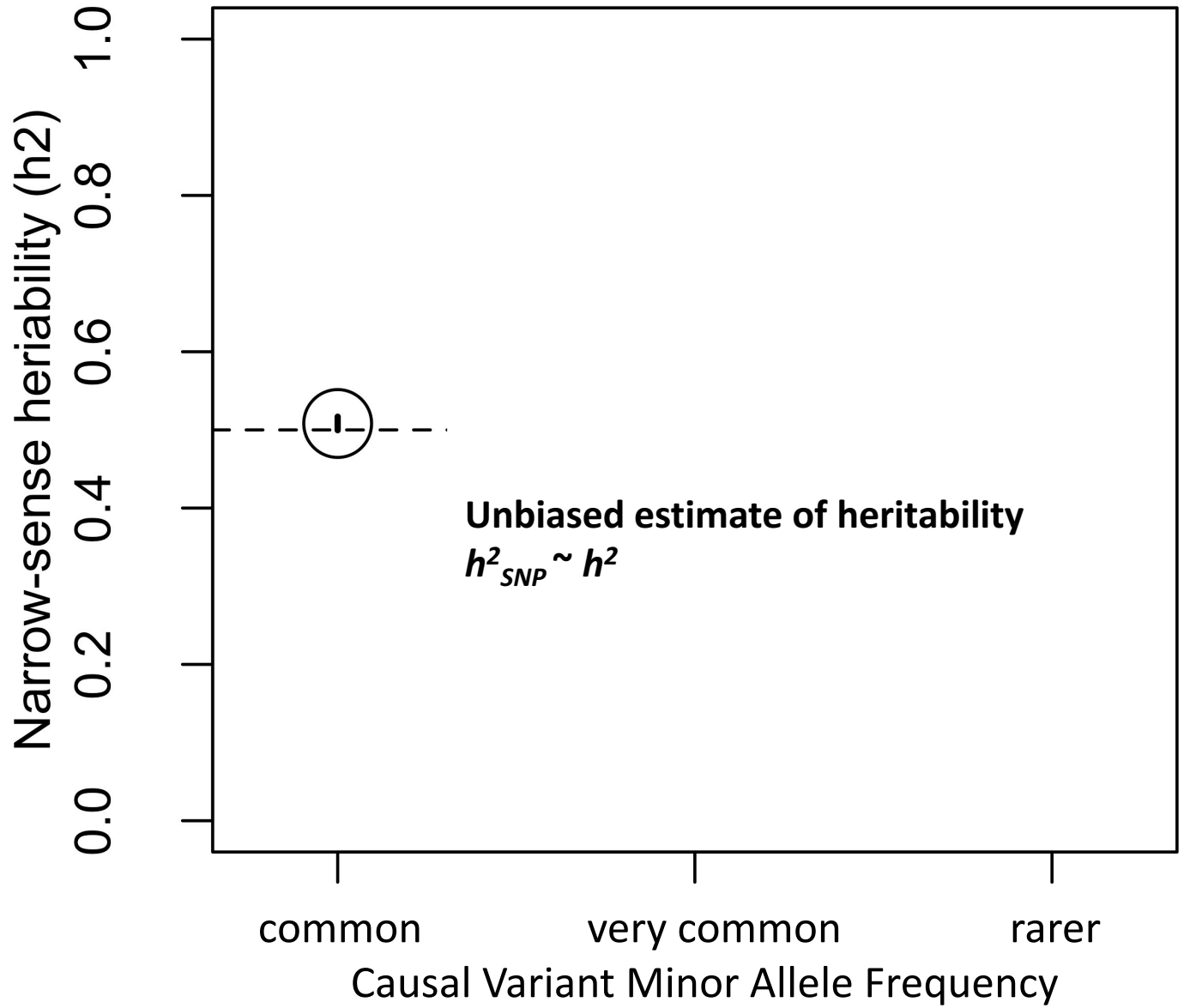


Simulated phenotypes, GRM from common Axiom array



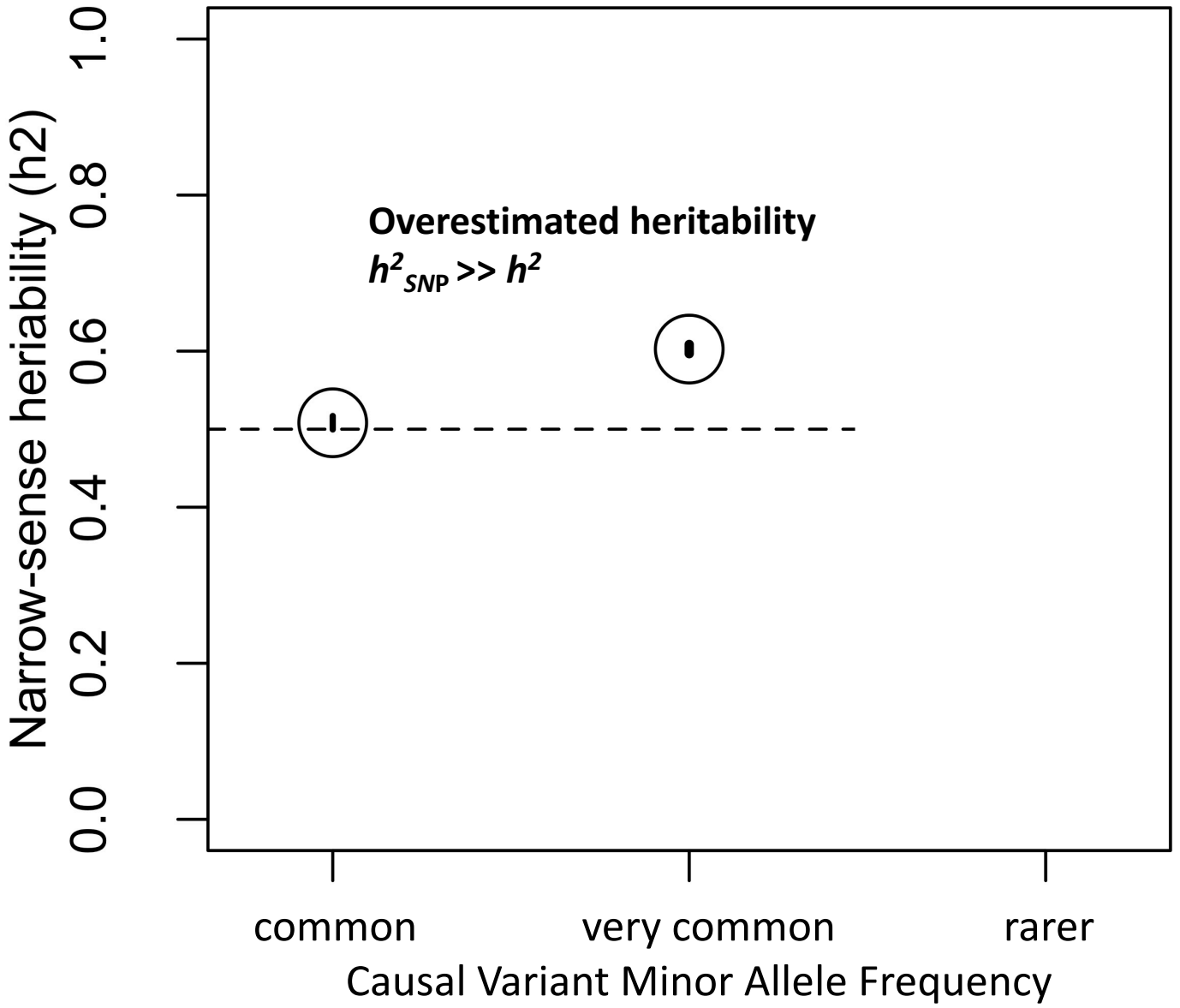
Mean +/- 95% CI
100 replicates

Simulated phenotypes, GRM from common Axiom array



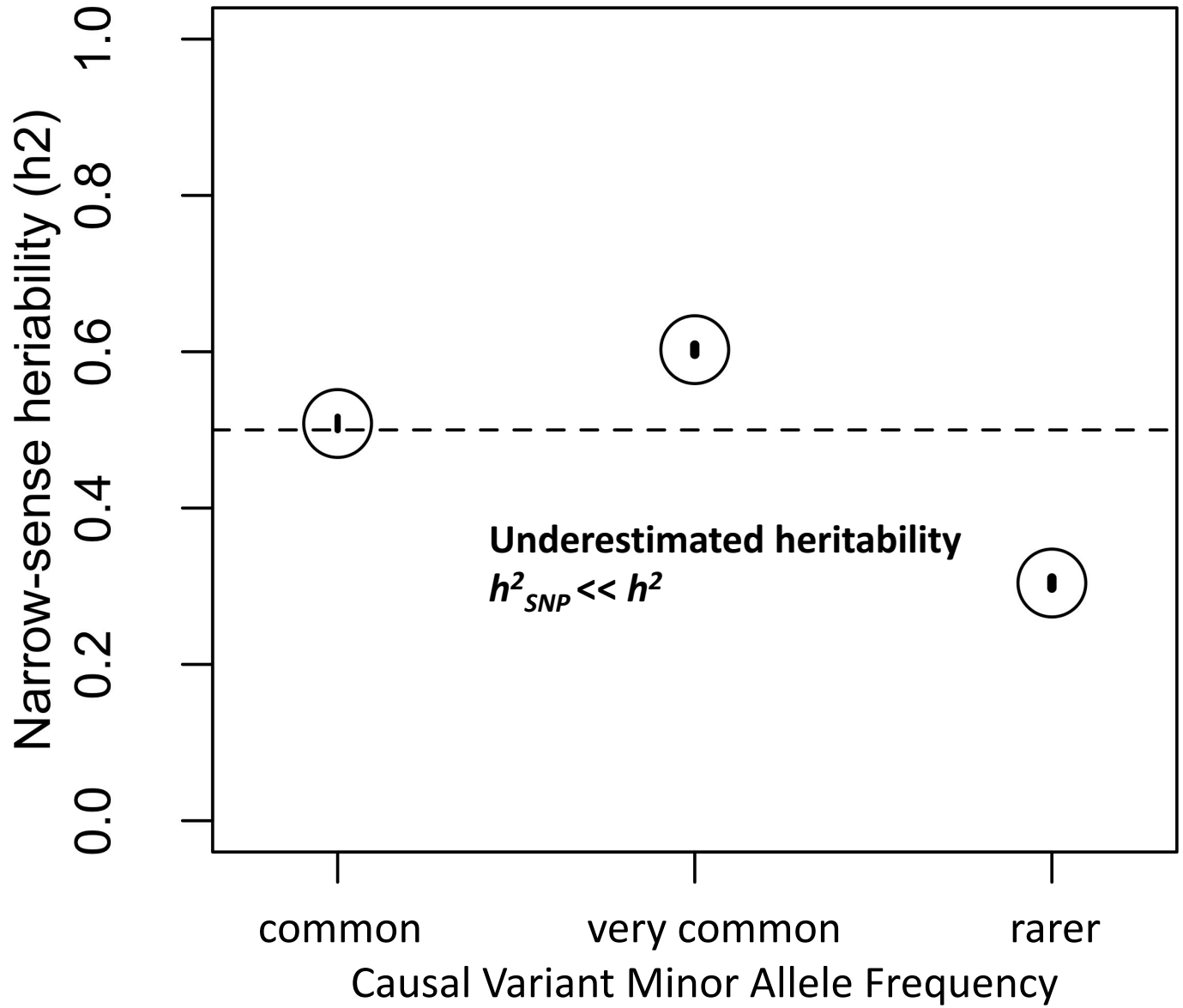
Mean +/- 95% CI
100 replicates

Simulated phenotypes, GRM from common Axiom array



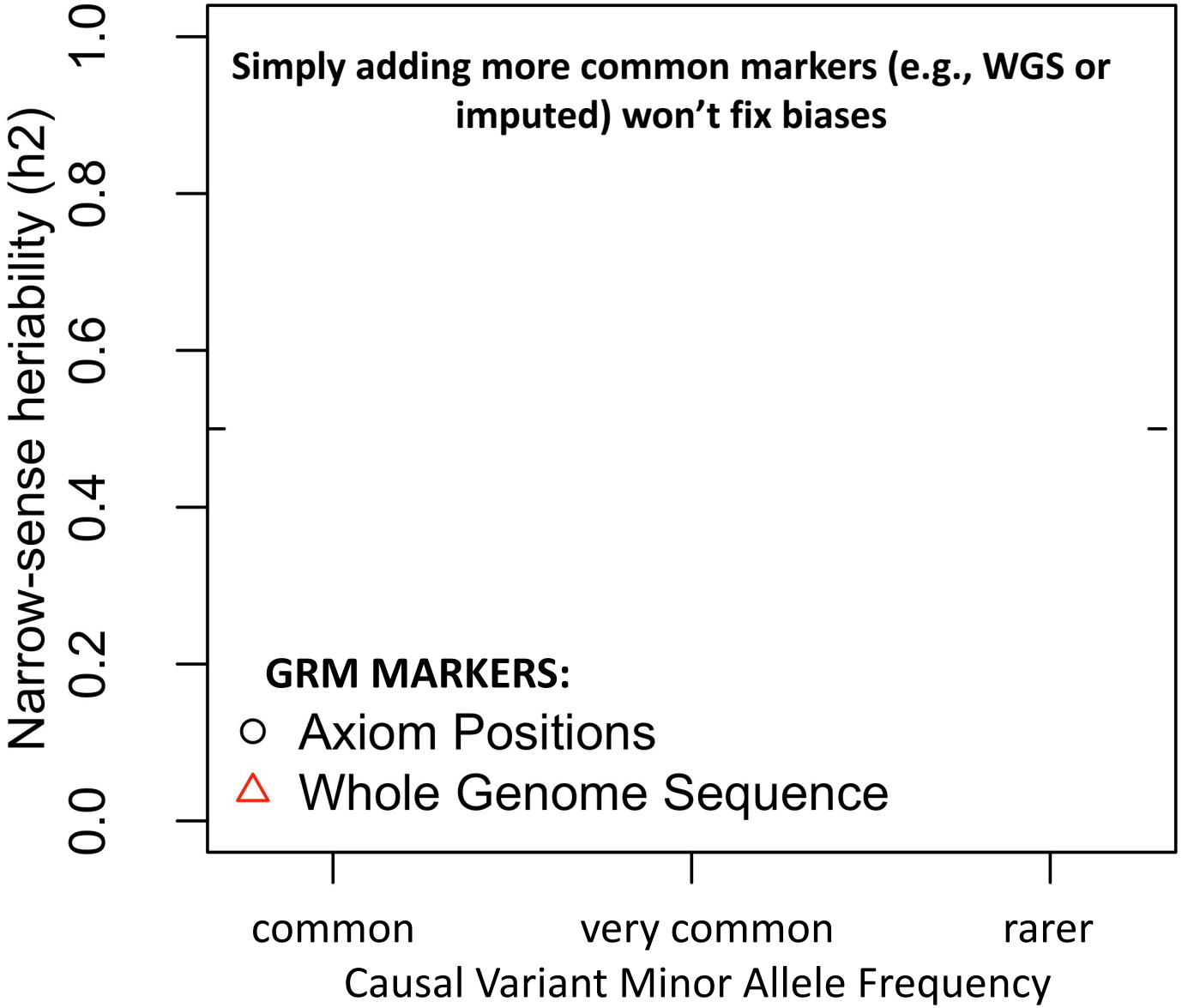
Mean +/- 95% CI
100 replicates

Simulated phenotypes, GRM from common Axiom array



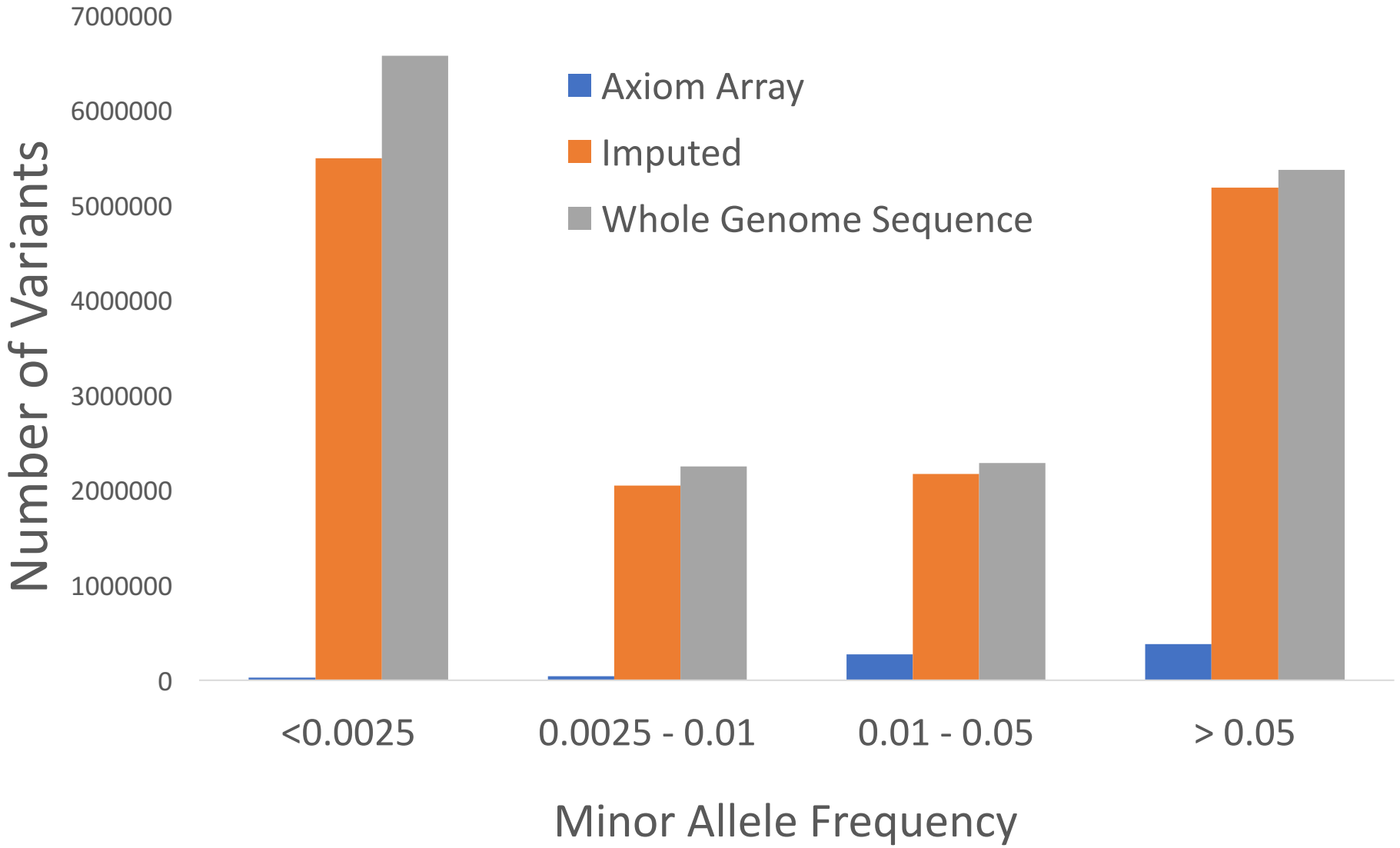
Mean +/- 95% CI
100 replicates

Simulated phenotypes
GRM from *common* Whole Genome Sequence variants

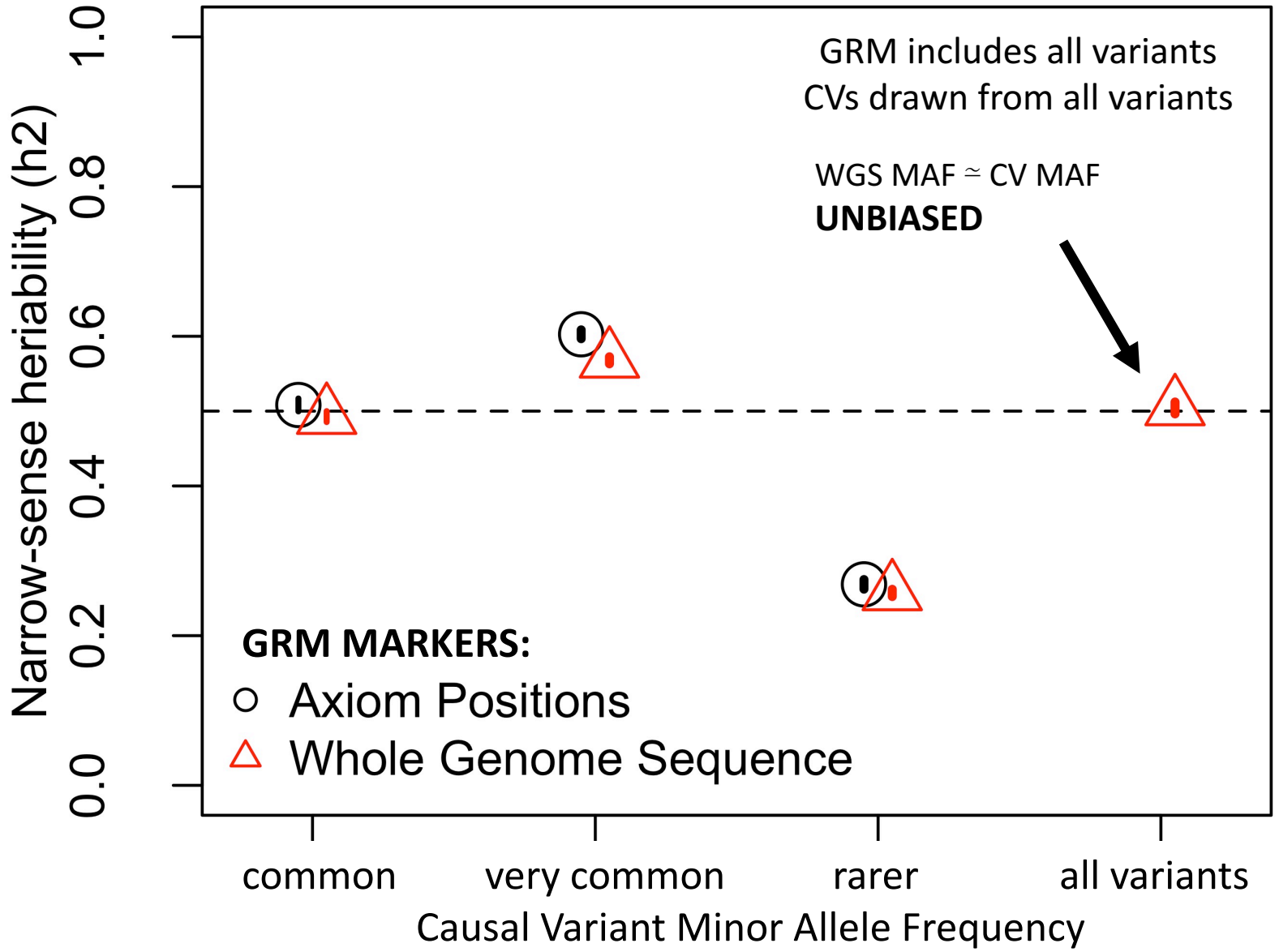


Mean +/- 95% CI
100 replicates

CVs from whole genome sequence include many rare variants



Simulated phenotypes, Axiom array or Whole Genome GRM



Mean +/- 95% CI
100 replicates

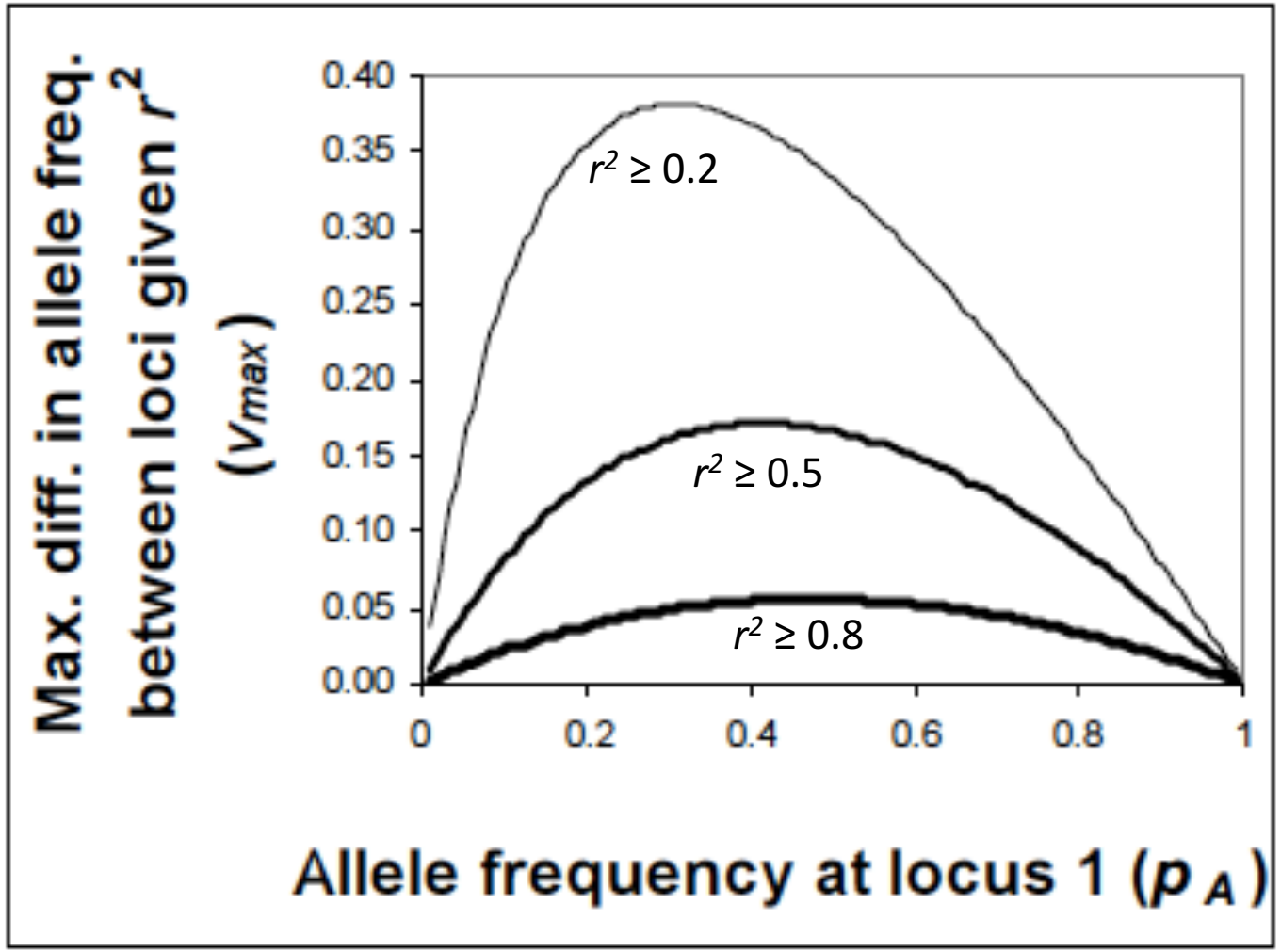
Why is there a relationship between GREML-SC heritability estimates and MAF?

- Unbiased estimates when marker MAF is the same as the CV MAF
- Underestimated when CVs are rarer than the markers used
- Overestimated when CVs are more common than the markers

- *MAF is related to LD, and LD is related to biases in h^2 estimation*
- Details: Wray 2005 Twin Res. Hum. Gen., Speed et al. 2012 AJHG, Yang et al. 2015 NG

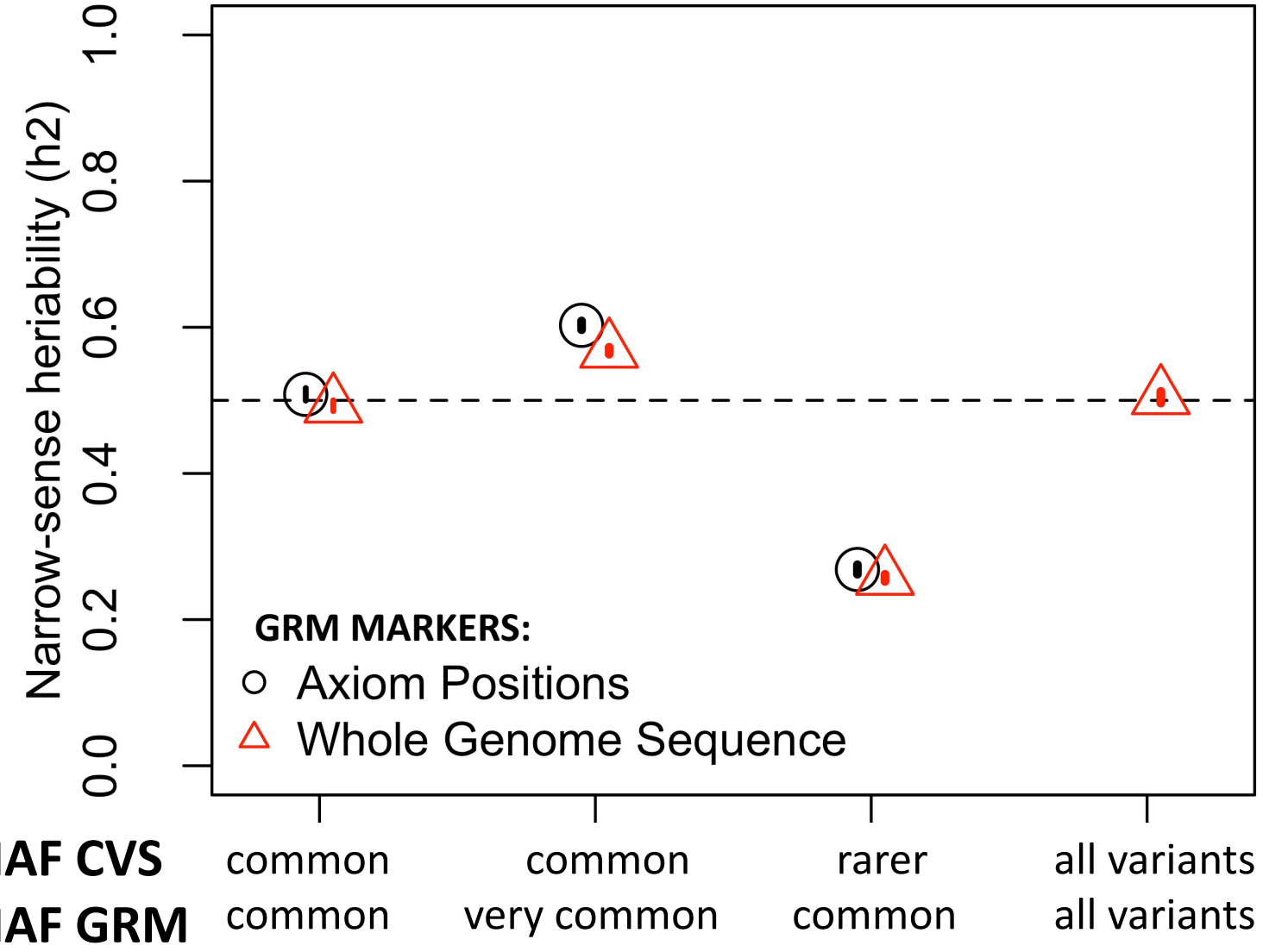
MAF is related to LD – Wray 2005 Twin Res. Hum. Gen. Figure 1

Common SNPs can't be in high LD with very rare SNPs



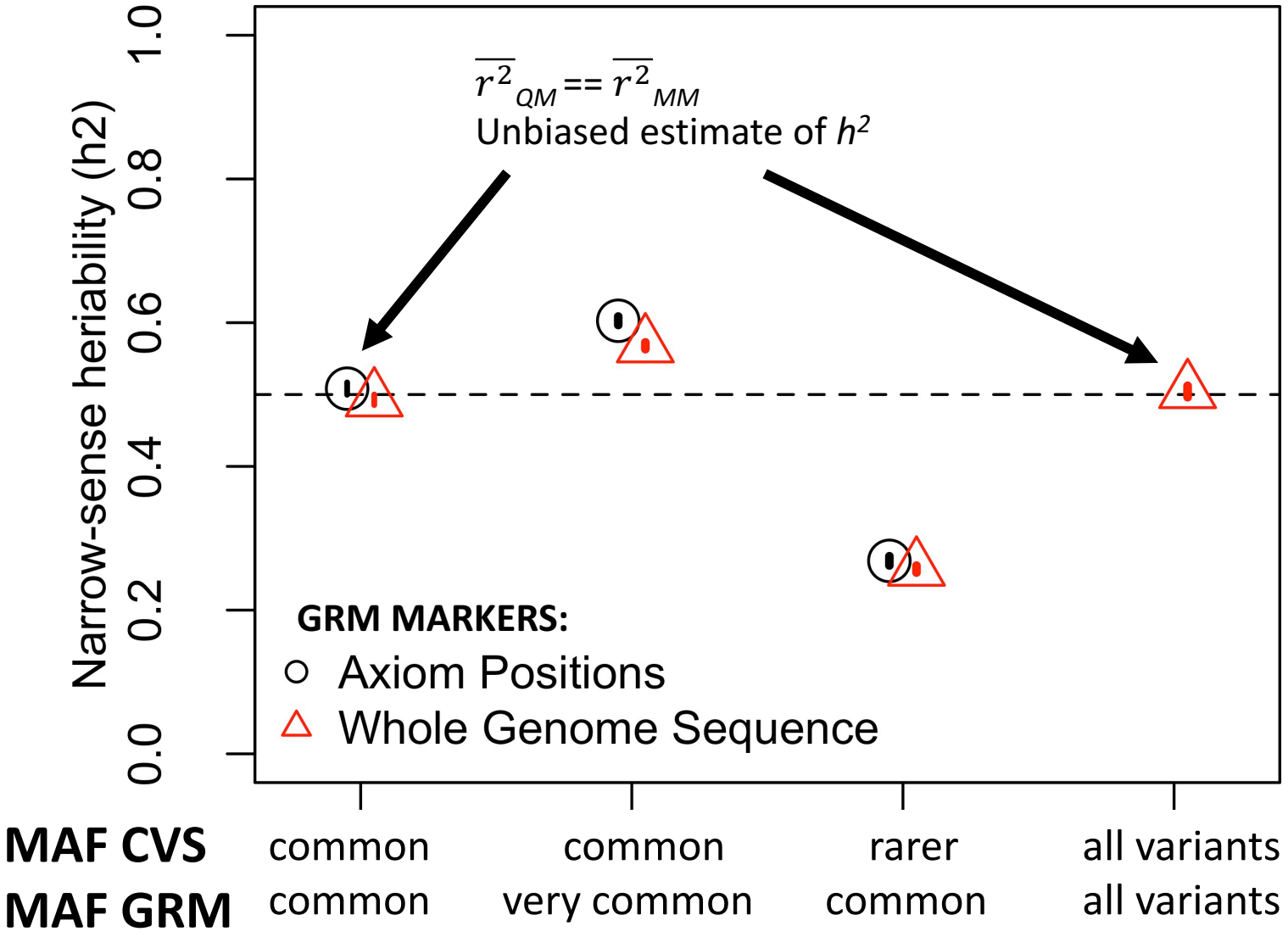
When does GREML-SC correctly estimate h^2?

- LD among markers and between markers and CVs (Yang et al. 2015 NG)
- $h^2_{SNP} = h^2(\overline{r^2}_{QM} / \overline{r^2}_{MM})$
- $\overline{r^2}_{QM}$ = average LD between markers and CV *genome-wide*
- $\overline{r^2}_{MM}$ = average LD among markers *genome-wide*



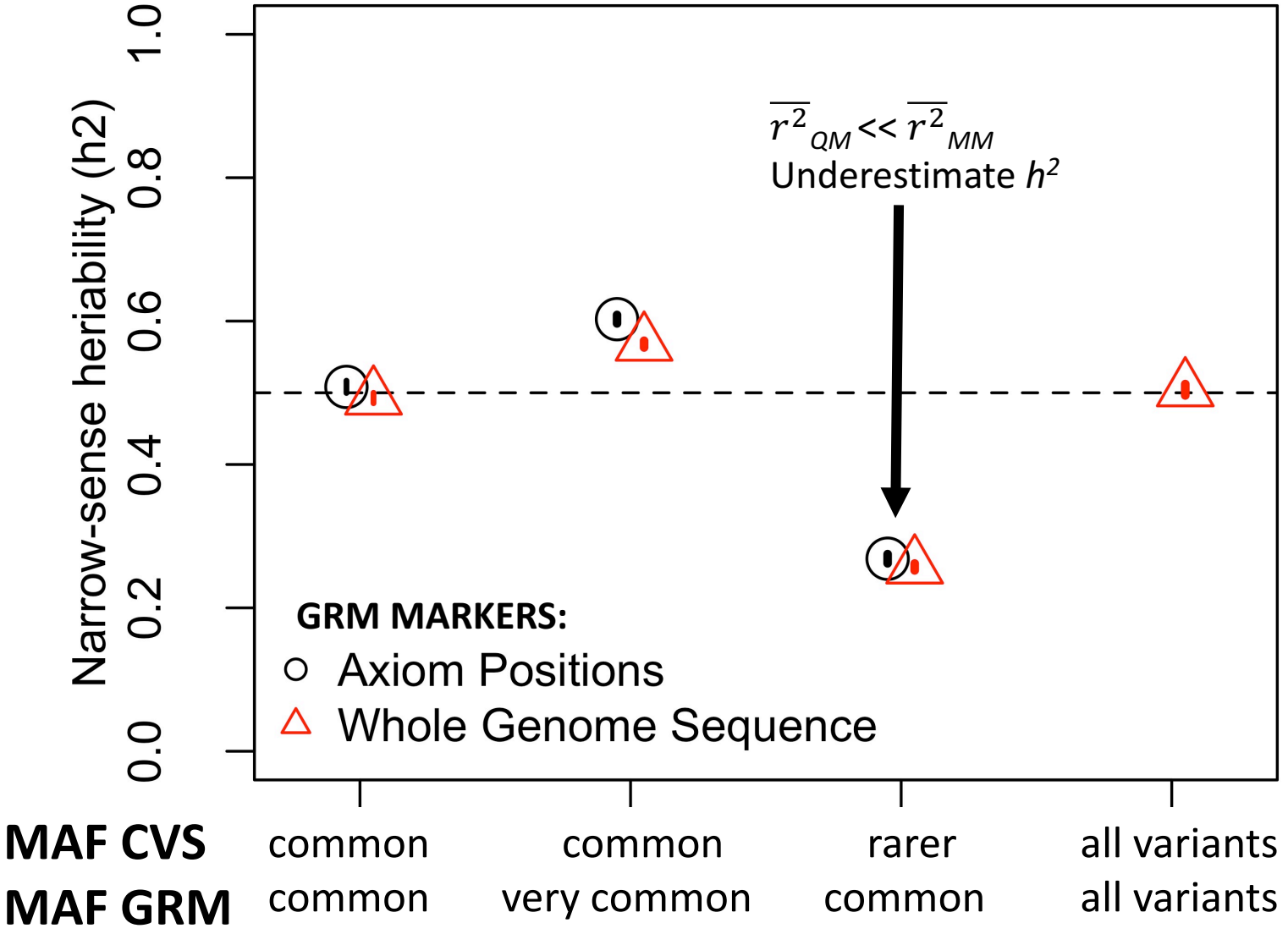
When does GREML-SC correctly estimate h^2?

- LD among markers and between markers and CVs (Yang et al. 2015 NG)
- $h^2_{SNP} = h^2(\overline{r^2}_{QM} / \overline{r^2}_{MM})$
- $\overline{r^2}_{QM}$ = average LD between markers and CV *genome-wide*
- $\overline{r^2}_{MM}$ = average LD among markers *genome-wide*



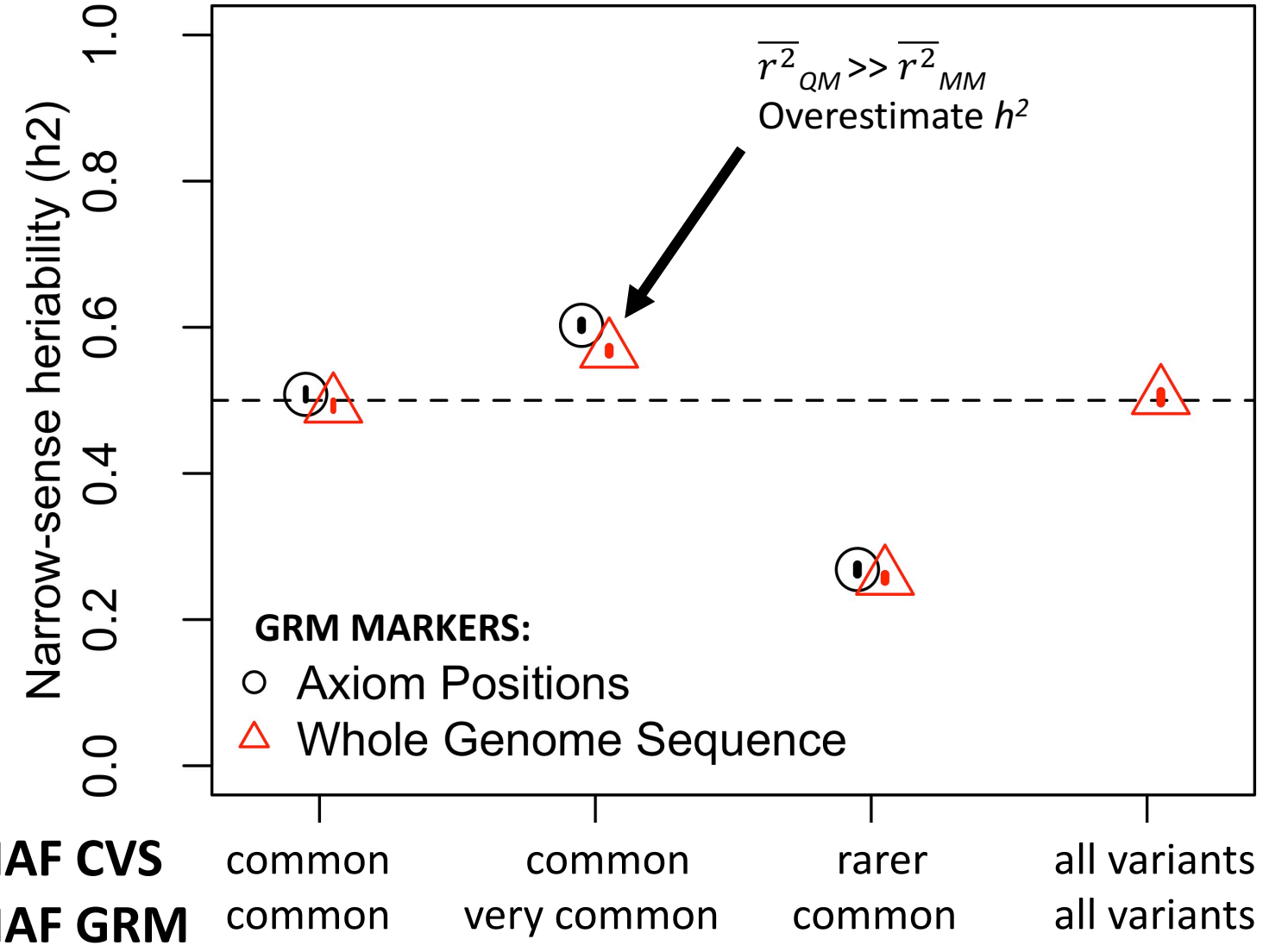
When does GREML-SC correctly estimate h^2?

- LD among markers and between markers and CVs (Yang et al. 2015 NG)
- $h^2_{SNP} = h^2(\overline{r^2}_{QM} / \overline{r^2}_{MM})$
- $\overline{r^2}_{QM}$ = average LD between markers and CV *genome-wide*
- $\overline{r^2}_{MM}$ = average LD among markers *genome-wide*



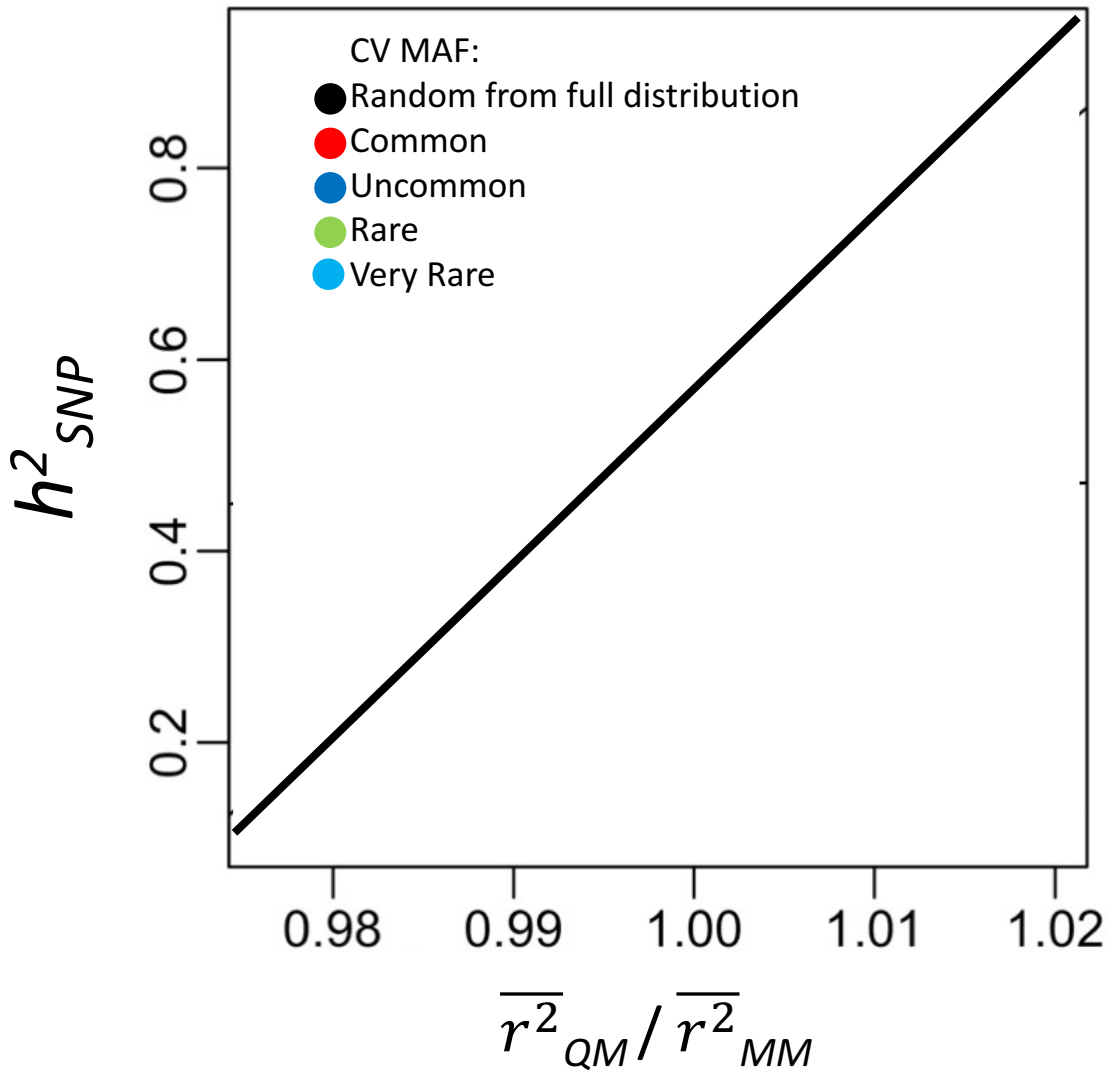
When does GREML-SC correctly estimate h^2?

- LD among markers and between markers and CVs (Yang et al. 2015 NG)
- $h^2_{SNP} = h^2(\overline{r^2}_{QM} / \overline{r^2}_{MM})$
- $\overline{r^2}_{QM}$ = average LD between markers and CV *genome-wide*
- $\overline{r^2}_{MM}$ = average LD among markers *genome-wide*



Heritability estimate related to LD patterns of markers and CVs

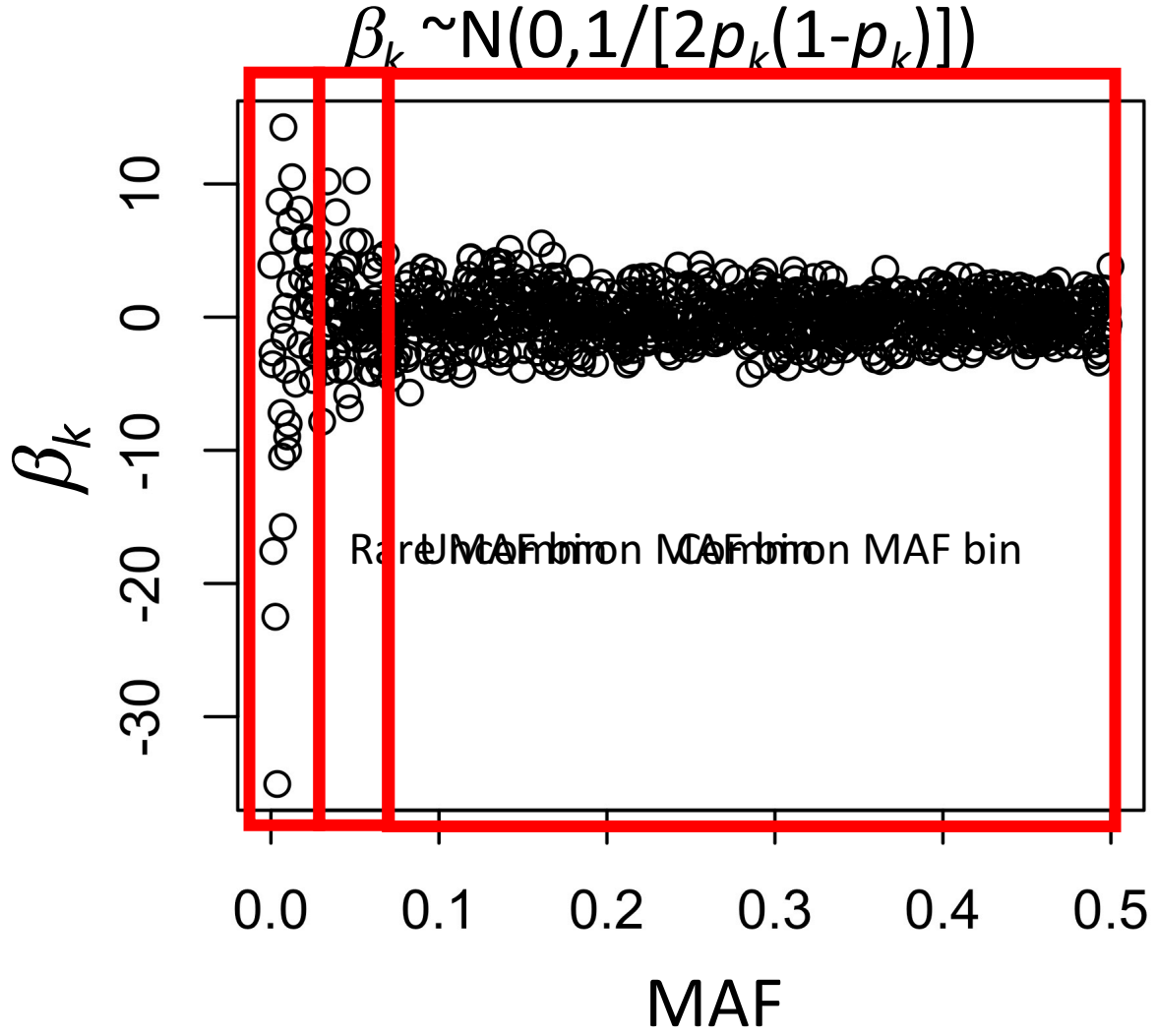
- LD among markers and between markers and CVs (Yang et al. 2015 NG)
- $h^2_{SNP} = h^2(\overline{r^2}_{QM} / \overline{r^2}_{MM})$
- $\overline{r^2}_{QM}$ = average LD between markers and CV *genome-wide*
- $\overline{r^2}_{MM}$ = average LD among markers *genome-wide*



Multiple Component GREML Yang 2011, Yang 2015

- Can correct for many of these biases
- GRMs from various MAF or LD bins
  - Bin variants into MAF and/or LD categories, create a GRM for each
  - GCTA will partition phenotypic variance among all GRMs (plus error)
    - Sum of all genetic variances is the total h^2_{SNP}
    - Partitioned estimates can explore aspects of genetic architecture (e.g., rare vs. common variants)

MAF-stratified approach: Allows the variance to change among MAF bins

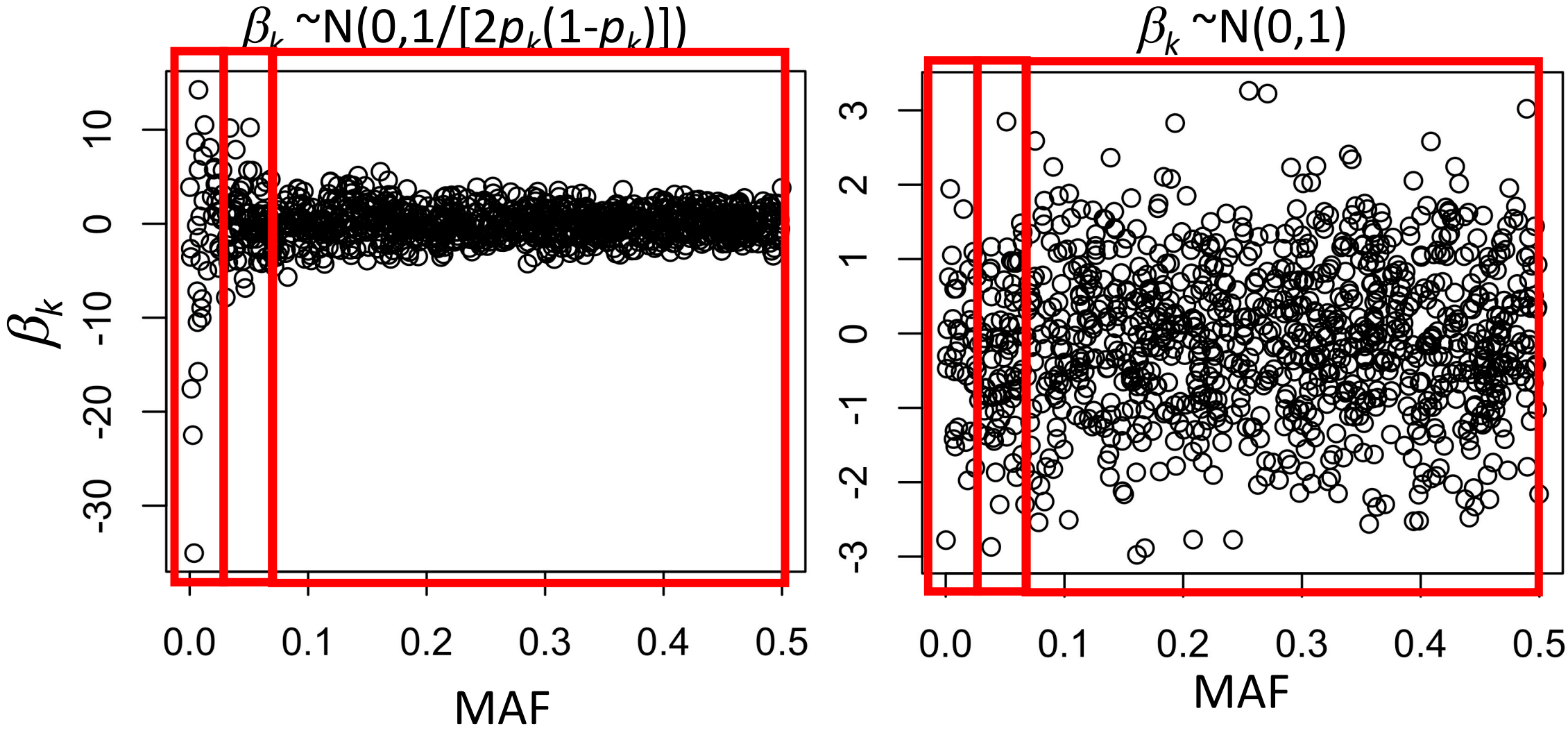


$$\boldsymbol{V} = \sum_i \mathbf{A}_i \sigma^2_i + \boldsymbol{I}\sigma^2_e$$

Relationship between markers and causal variants within bins:

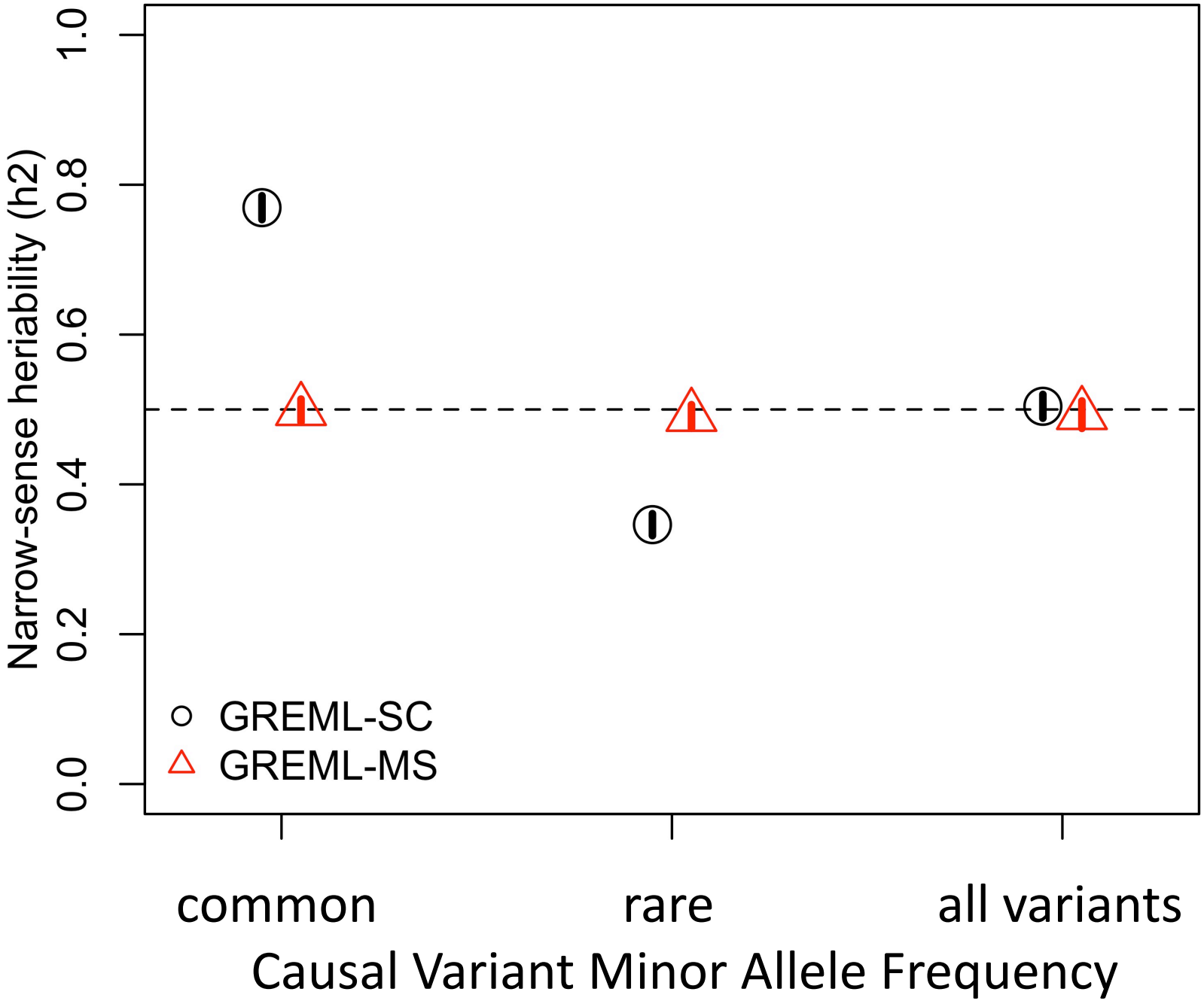
$h^2_{SNP} = h^2(\overline{r^2}_{QM} / \overline{r^2}_{MM})$

MAF-stratified approach: Allows the variance to change among MAF bins

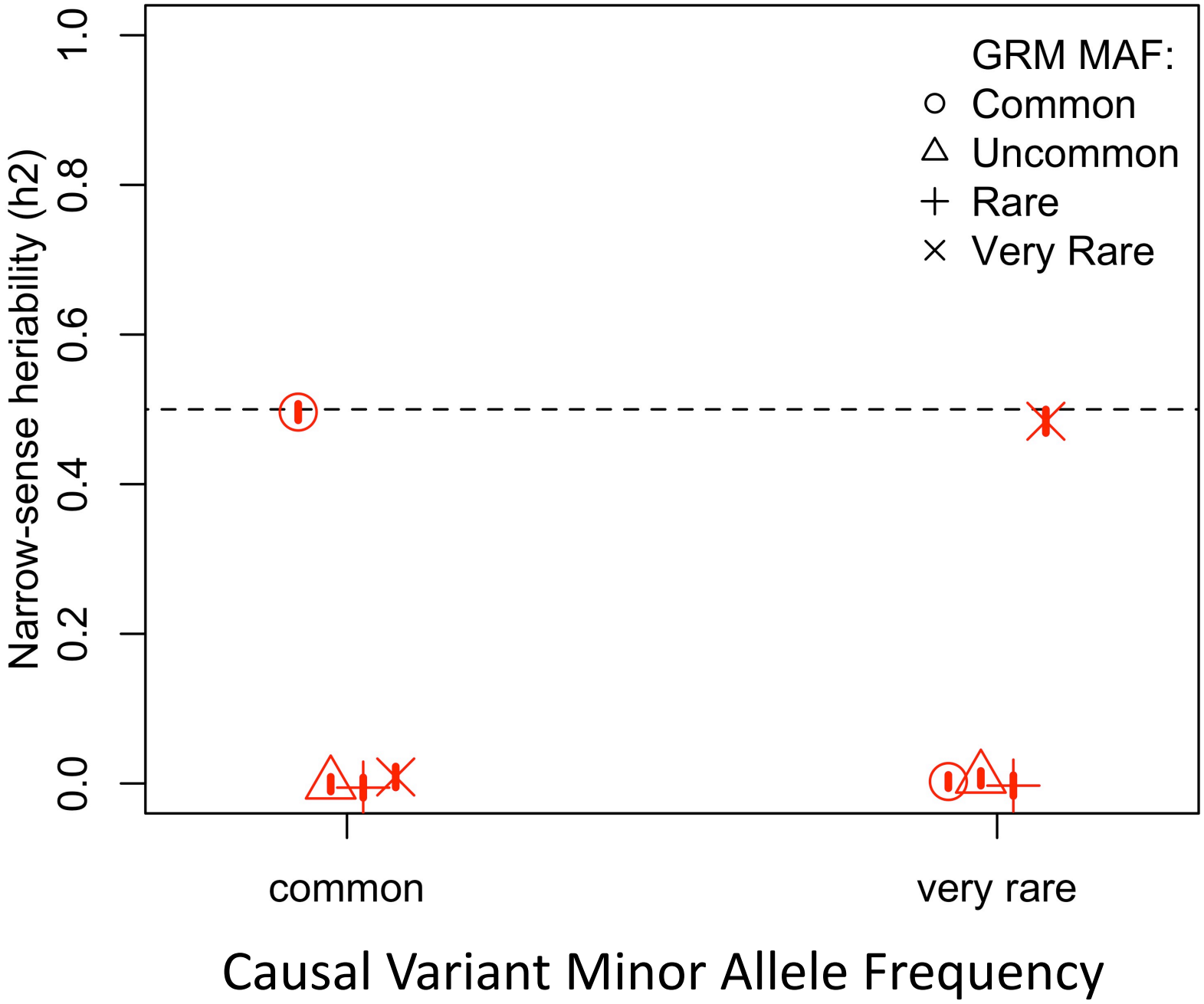


MAF-Stratified GREML is unbiased

Whole genome sequence
4 MAF bins



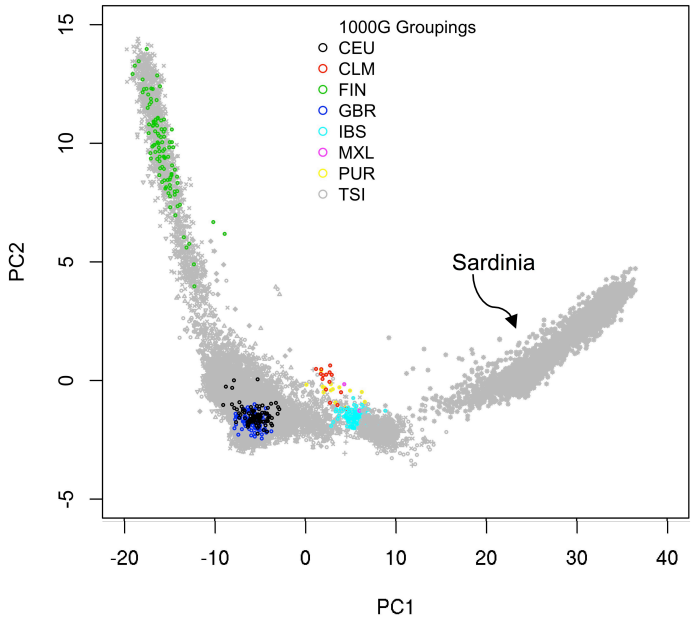
MAF-stratified GREML
Correctly partitions variance to the correct MAF range



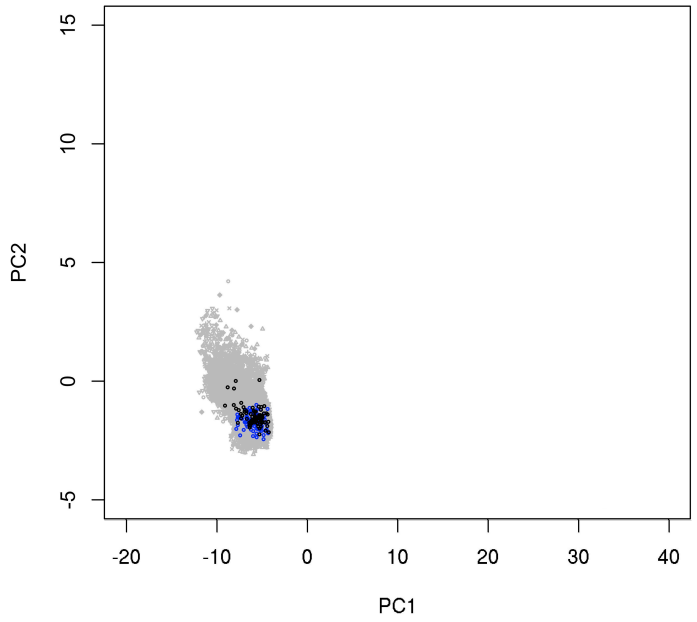
PRACTICAL 2

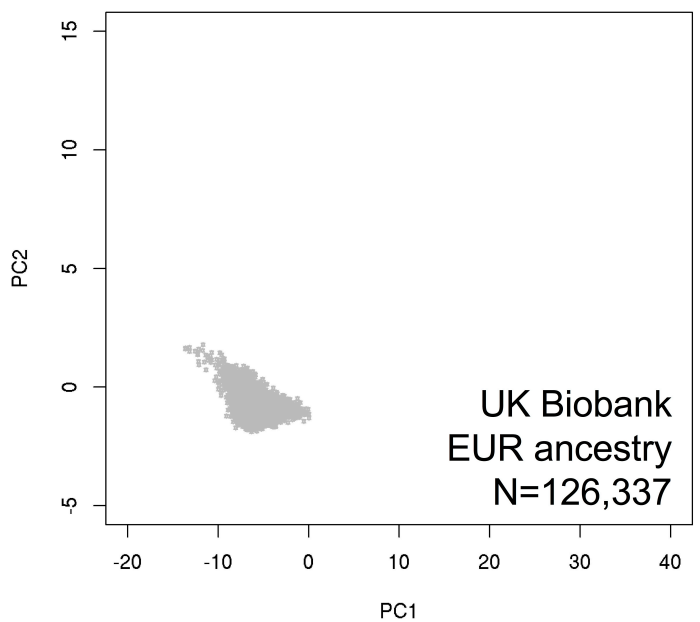
Stratification: Population structure influences LD (and therefore h^2_{SNP})

Europe-wide (HRC data)



Homogeneous Subset

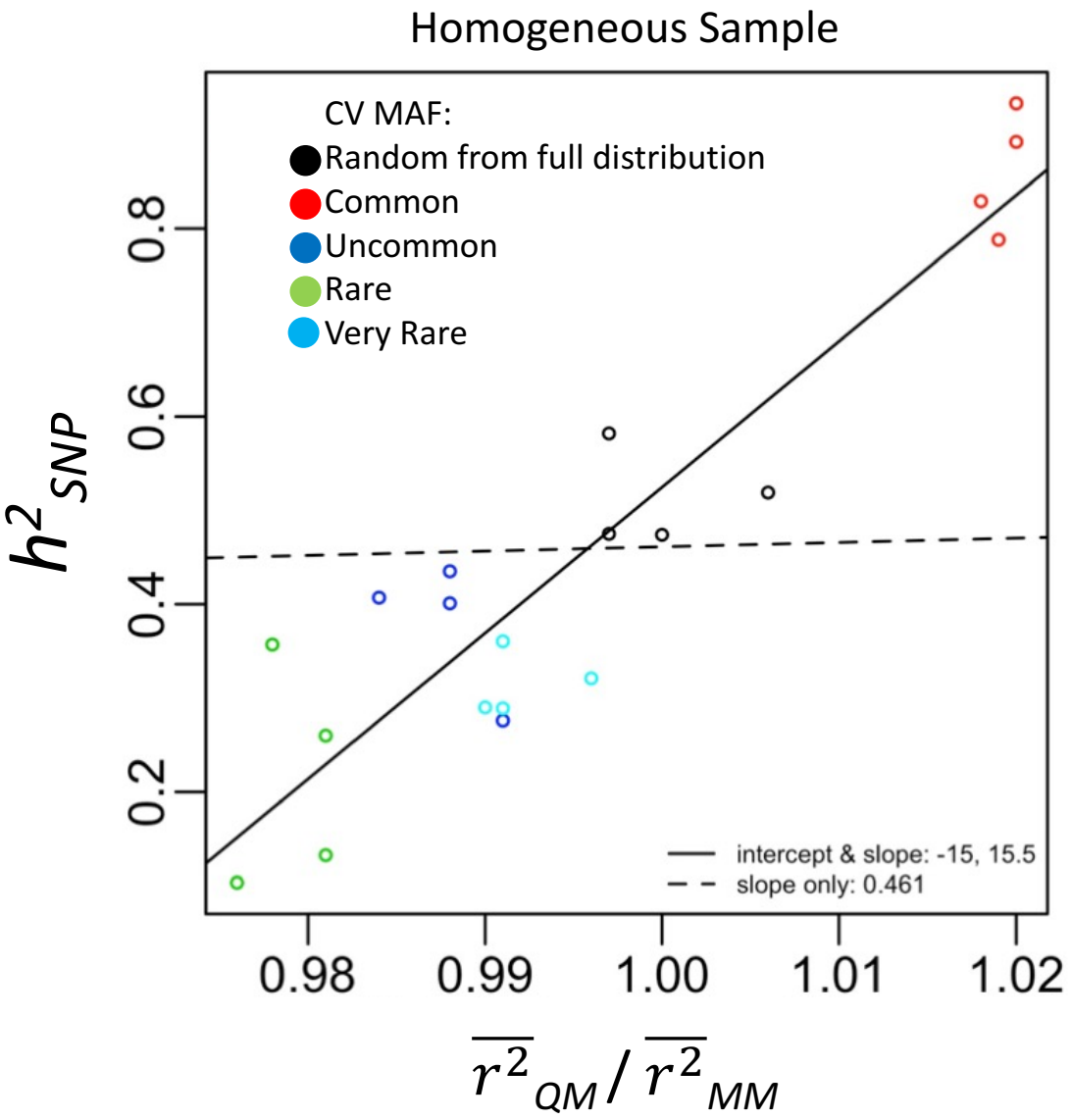


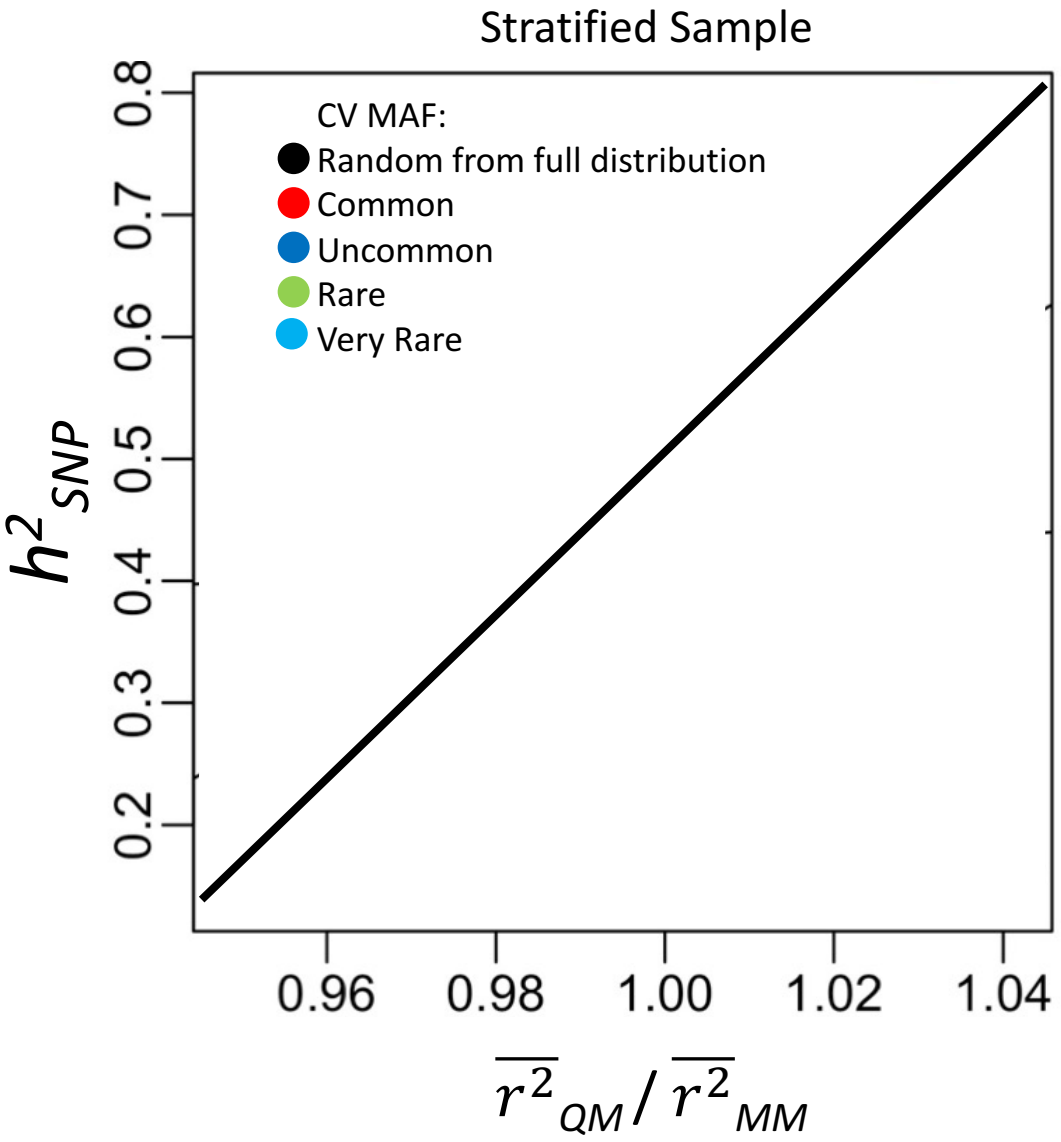


Stratification and confounding

- Remember stratification talks
- Environments can also be confounded with ancestry
- Other covariates – sex, batch, etc.
- Typically, PC scores for some number of axes included as covariates (Price et al. 2010, Yang et al. 2014, etc.)
- Covariates included correct for mean differences, but not the LD effects of stratification

Homogeneous Vs. Stratified Samples: $h^2_{SNP} = h^2(\overline{r^2}_{QM} / \overline{r^2}_{MM})$

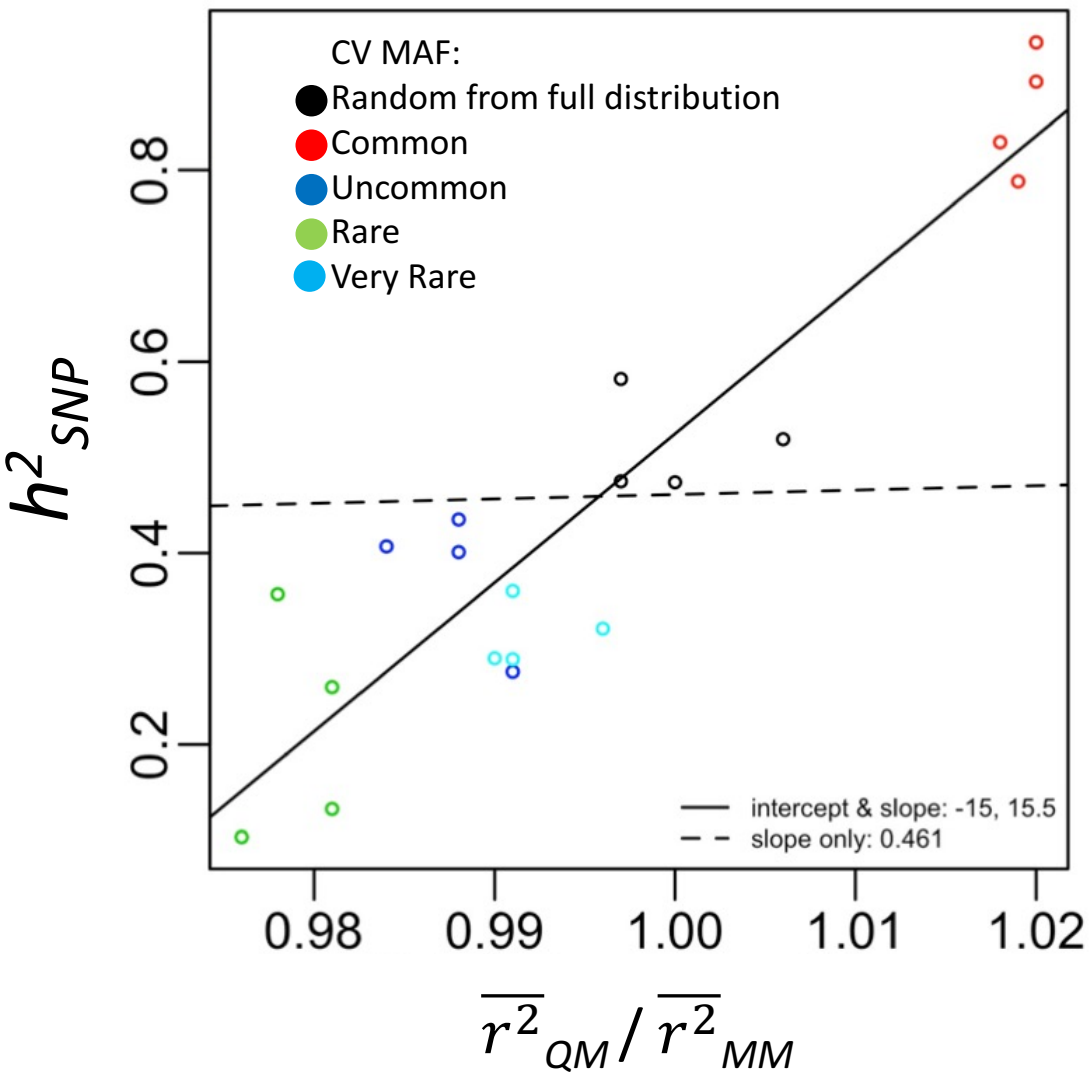




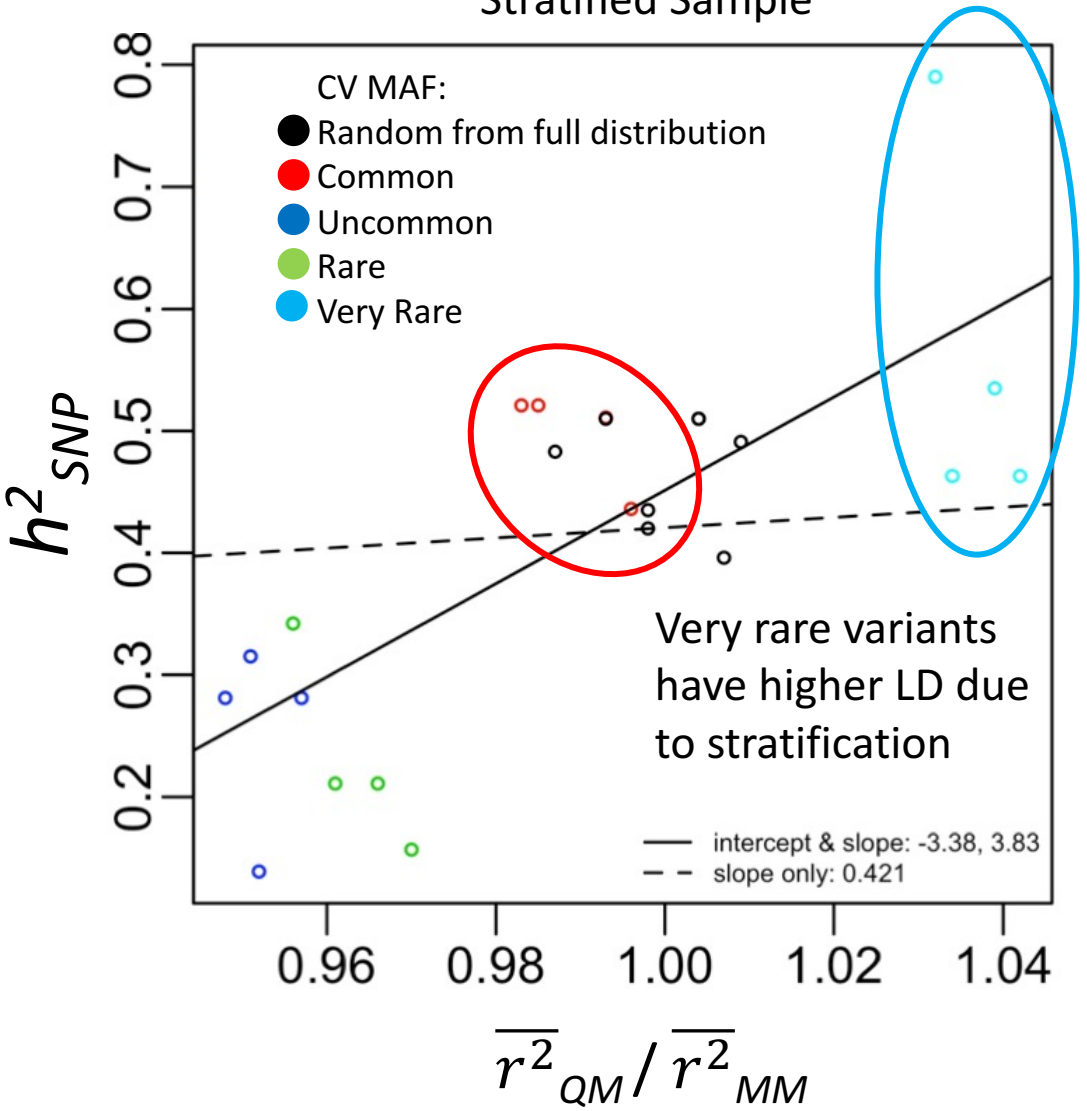
Homogeneous Vs. Stratified Samples: $h^2_{SNP} = h^2(\overline{r^2}_{QM} / \overline{r^2}_{MM})$

Rare, ancestry-informative alleles are in high LD, driving up LD scores

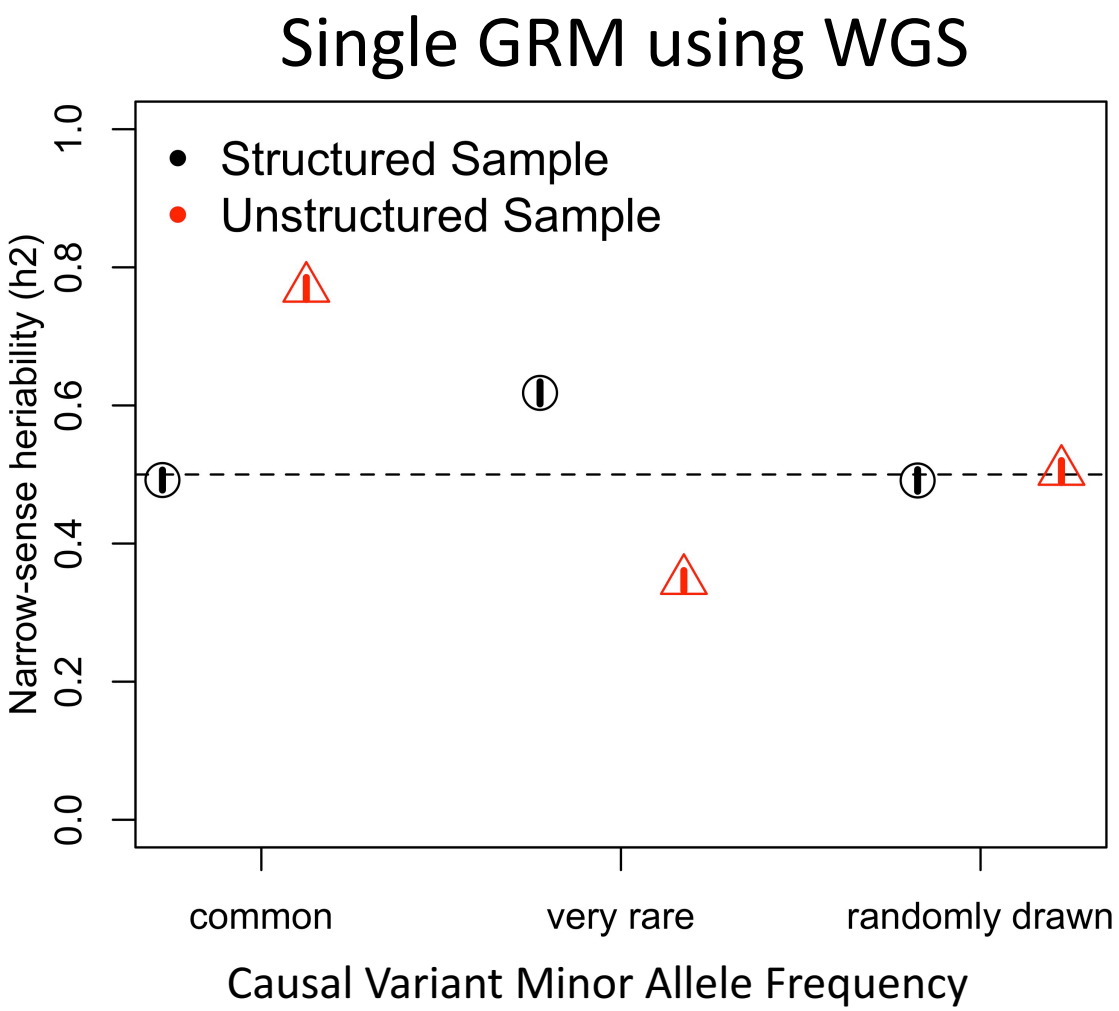
Homogeneous Sample

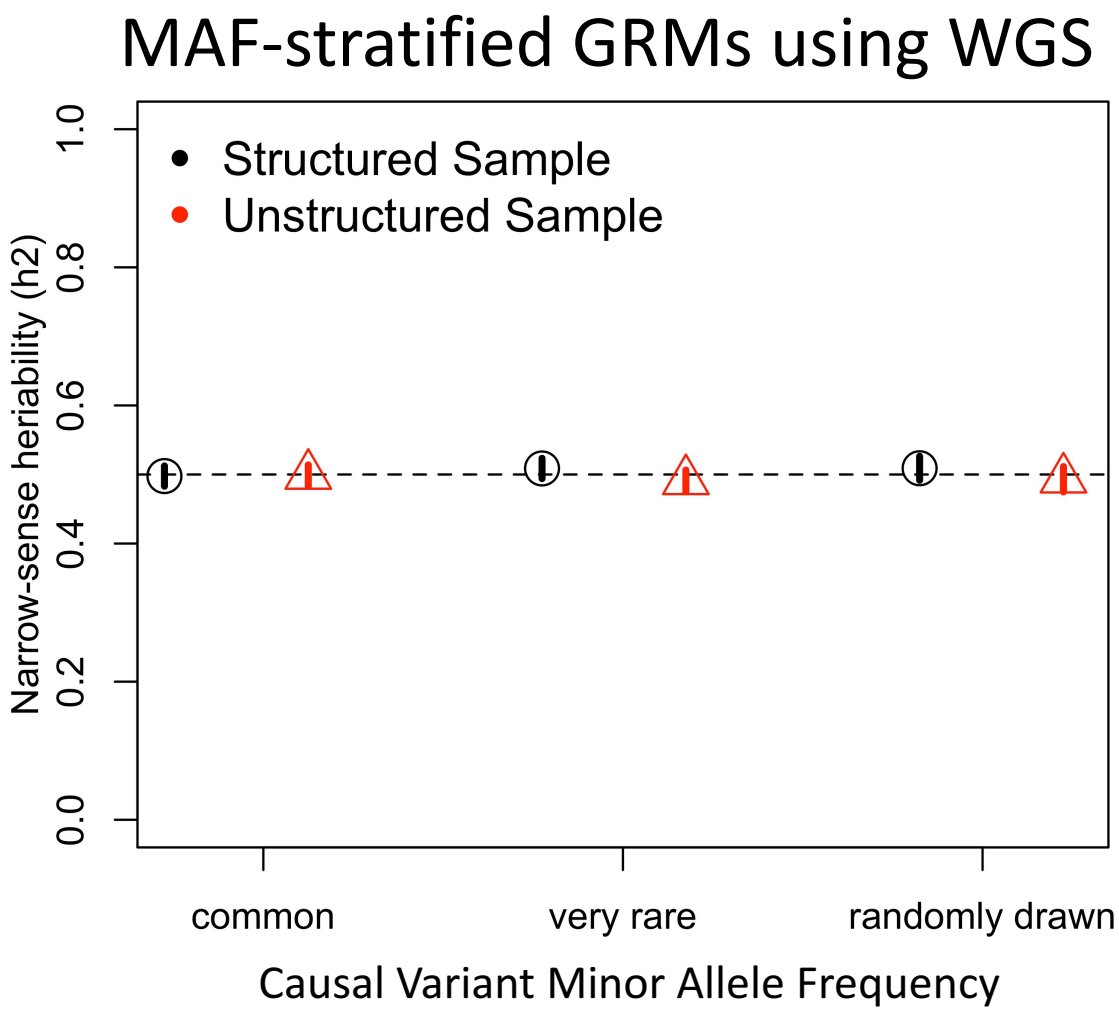


Stratified Sample

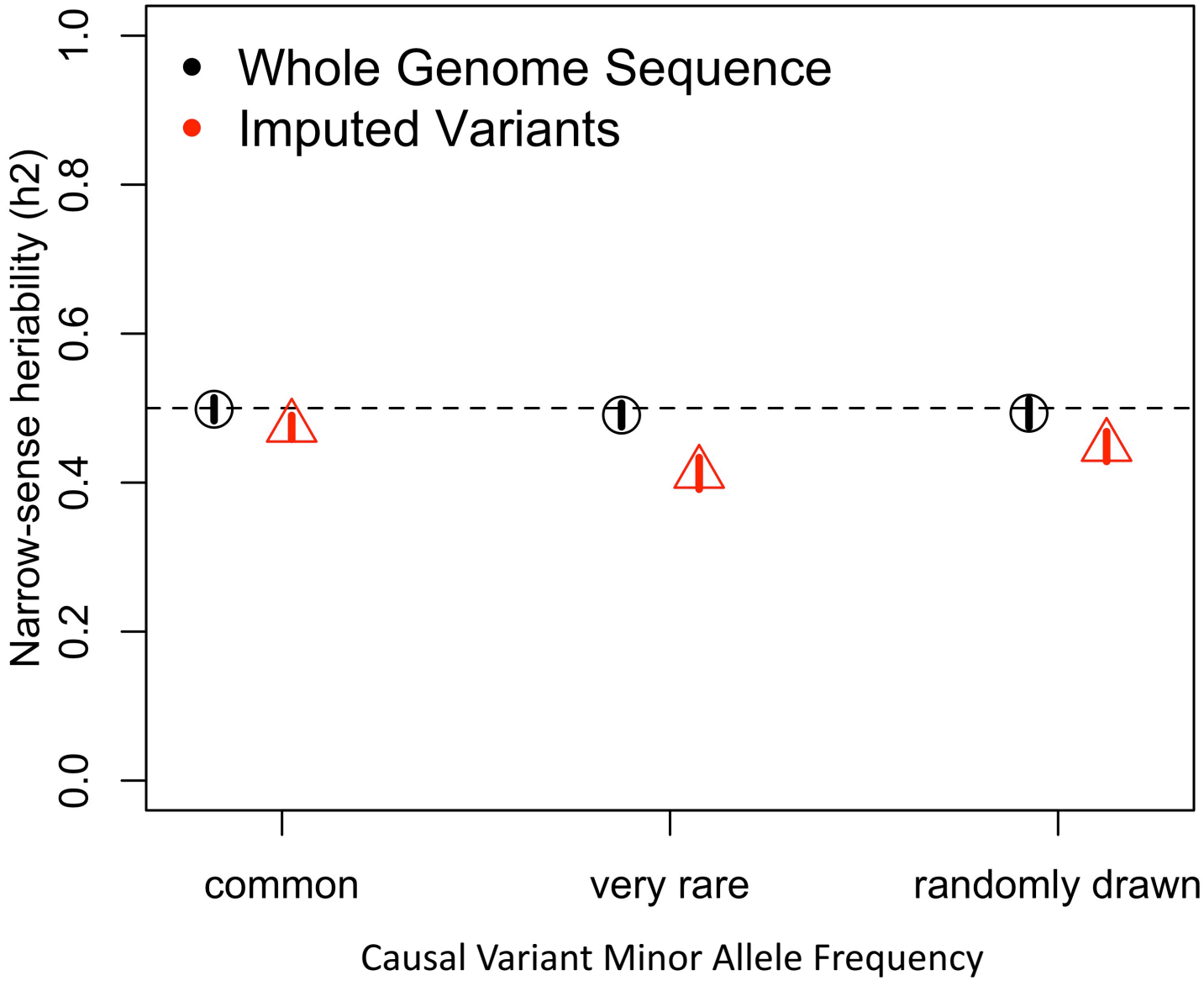


MAF-stratified or single component in homogeneous samples vs. structured samples





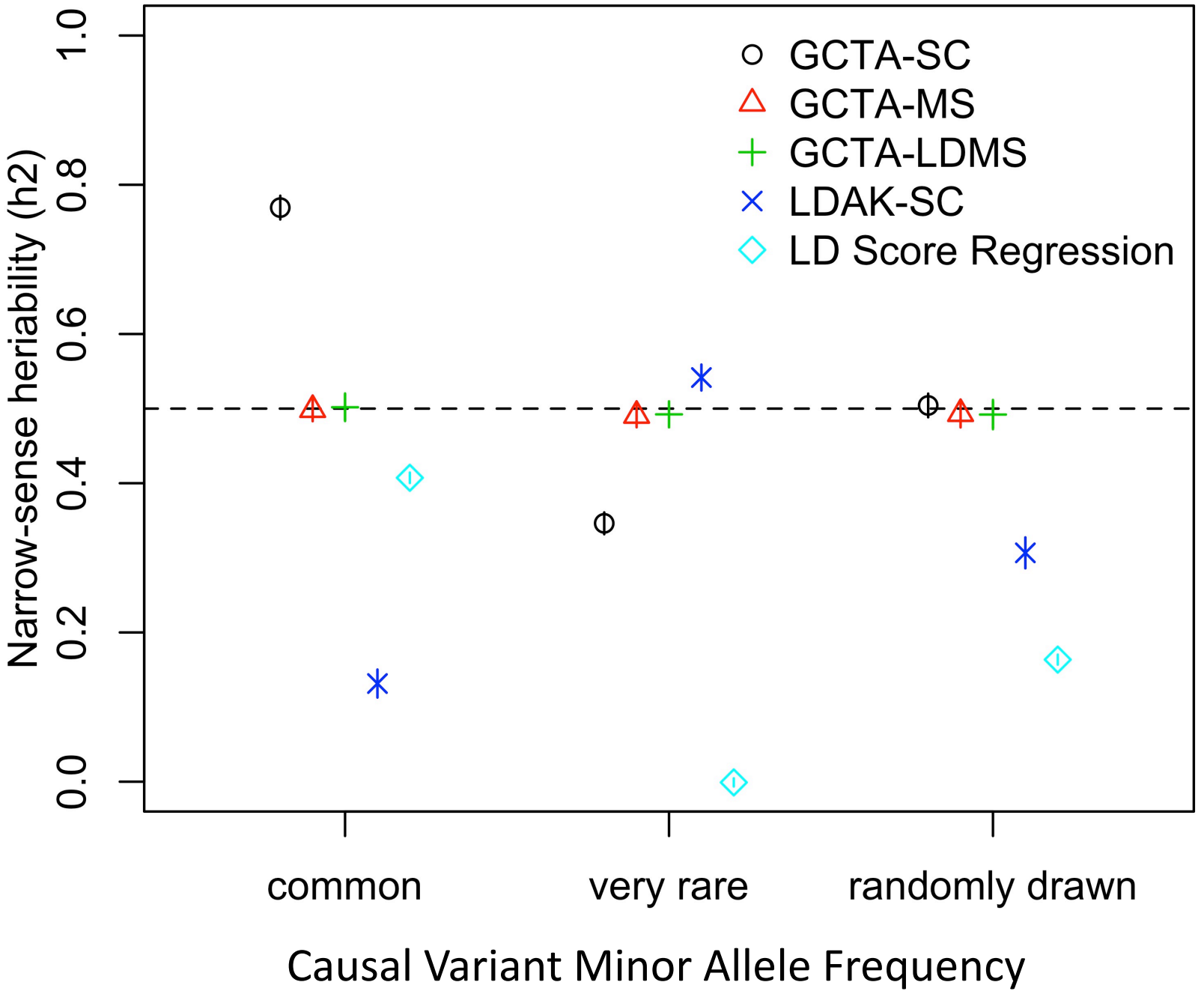
Imputation vs. genome sequence – GREML-MS



Numerous methods developed

Relative performance varies

Dependent on model & assumptions



BEST PRACTICES:

- Careful QC, appropriate covariates
- Whole genome sequence is best
- Impute! Use the Haplotype Reference Consortium.
- Remove related individuals – these share confounding environmental effects, but this is avoided using unrelated samples.
- Carefully interpret results from studies that use a single GRM in GREML. There are clear biases from this approach, yet most have used GREML-SC.
- GREML-MS or GREML-LDMS are much preferred.