Supplementary information S3 | Estimating heritability using relatedness coefficients

Simulating a phenotype
Quantitative trait phenotypes with a heritability ($h^2$) of 0.5 were simulated for each of the 30 datasets. Random samples of 500 causal variants were selected from two MAF scenarios: I) randomly sample causal variants from polymorphisms with MAF > 0.05. By sampling only from polymorphisms with MAF > 0.05 their allele frequency distribution will similar to that of the SNPs in the genotype panel; II) randomly sample causal variants from the SNPs with MAF < 0.05, representing rare variants in a standard population genetics framework. Whilst this represents a high number of causal variants for a region representing a chromosome, it means the total effect attached to each causal variant will remain small. The phenotypic value for each individual was determined as follows;

1) The effect ($\alpha$) of each causal variant was drawn from a standard normal distribution
2) The overall genetic effect of an individual ($g_i$) was calculated by; $g_i = \sum_{ik} x_{ik} \alpha_{ik}$ where $x_{ik}$ is an indicator for the number of ‘A’ alleles carried at causal variant $k$ by individual $i$.
3) Residual effects ($e_i$) for each individual were drawn from a normal distribution (mean = 0) and variance $\left(\frac{1}{h^2} \sigma_g^2\right) - \sigma_g^2$ where $\sigma_g^2$ is the variance of $g$ in the dataset.
4) The phenotypic value for each individual was calculated by; $y_i = g_i + e_i$.

Estimating $h^2$ from markers
Each of the 30 datasets consists of 1000 ‘unrelated’ individuals simulated under a Wright-Fisher panmictic, non-selfing, diploid population of constant size, as described in S1. Pairwise genetic relationships were estimated from markers within the SNP panel for individuals within each dataset using raw UAR, adjusted UAR and PLINK_IBD methods. Each of these methods produces a 1000 by 1000 relationship matrix ($A$) based on the SNPs in the simulated genotype panel (S1). Each pair of individuals $i$ and $k$ has a relationship coefficient $A_{i,k}$. We fitted a linear model to the simulated phenotype and used restricted maximum likelihood\(^1\) (REML) analyses to estimate the variance explained by the markers within the genotype panel. This process was repeated for each of the methods used to estimate $A_{i,k}$.

Estimates of $h^2$ can be less than the true $h^2$ due to;
1) Sampling error associated with estimating UAR from $m$ SNPs.
2) Incomplete linkage disequilibrium (LD) between causal variants and SNPs. This is exacerbated when MAF distributions of markers and causal variants differ.
Note: Yang et al.\textsuperscript{3} established an empirical linear relationship between $\beta$ and $N$; $\beta = 1 - \frac{(c + \frac{1}{N})}{\text{var}(A_{ij})}$

where $c$ is a constant depending on the causal variant MAF threshold.

References

