QUANTITATIVE TRAITS

Bruce Weir
and
Jérôme Goudet

July 19, 2023

Genetic Model for Trait

Suppose gene **T** affects a trait: its genotype may affect the probability an individual has a disease or its genotype may affect the value of some measurable quantity. There may be other genes also affecting the trait, and there may be non-genetic effects.

Now suppose G is the genotypic effect of **T** on the trait and E is the environmental effect (or all other effects). An individual is observed to have phenotypic (trait) value (disease status or measured value) Y and a simple linear model is

$$Y = G + E$$

The mean environmental effect is taken to be zero, so the mean phenotypic effect is equal to the mean genotypic value.

Genetic Model for Trait

"If we could replicate a particular genotype in a number of individuals and measure them under environmental conditions normal for the population, their mean environmental deviations would be zero, and their mean phenotypic value would consequently be equal to the genotypic value of that particular genotype. This is the meaning of the genotypic value of an individual."

Falconer DS. 1960. Introduction to Quantitative Genetics. Ronald Press, New York. p. 113

Genetic Model for Trait

Extensions to this model can include $G \times E$ interaction, but at present G and E will be considered independent and their variances sum to the variance of Y:

$$\text{Var}(Y) = \text{Var}(G) + \text{Var}(E) \quad \text{or} \quad \sigma_Y^2 = \sigma_G^2 + \sigma_E^2$$

In general, the number of alleles for gene **T** is not known, but a convenient start is to suppose there are two: an ancestral form and a more recent form that may increase the chance of being affected or lead to detrimental values of a measured trait. Write the two alleles as T, t and the three genotypes as TT, Tt, tt. The three genotypic values are G_{TT}, G_{Tt}, G_{tt}.

Additive and Dominance Variance

In a population there is a mean genotypic effect, μ_G, and a variance of genotypic effects, σ^2_G:

$$\begin{aligned}
\mu_G &= \pi_T^2 G_{TT} + 2\pi_T\pi_t G_{Tt} + \pi_t^2 G_{tt} \\
\sigma^2_G &= \pi_T^2 (G_{TT} - \mu_G)^2 + 2\pi_T\pi_t (G_{Tt} - \mu_G)^2 + \pi_t^2 (G_{tt} - \mu_G)^2
\end{aligned}$$

and the variance can be partitioned into additive and dominance components:

$$\begin{aligned}
\sigma^2_G &= \sigma^2_{A_T} + \sigma^2_{D_T} \\
\sigma^2_{A_T} &= 2\pi_T\pi_t[\pi_T(G_{TT} - G_{Tt}) + \pi_t(G_{Tt} - G_{tt})]^2 \\
\sigma^2_{D_T} &= \pi_T^2\pi_t^2(G_{TT} - 2G_{Tt} + G_{tt})^2
\end{aligned}$$

Additive Traits

If the genetic value of trait heterozygotes is the average of the values of the two trait homozygotes, $G_{Tt} = (G_{TT} + G_{tt})/2$ then

$$\begin{aligned} \sigma^2_{A_T} &= 2\pi_T\pi_t(G_{Tt} - G_{tt})^2 \\ \sigma^2_{D_T} &= 0 \end{aligned}$$

and the genetic variance is entirely additive.

If the population has only one of the two trait alleles, $\pi_T\pi_t = 0$ and there is no genetic variance. Otherwise, additive genetic variance is maximized when the two trait alleles are equally frequent.

Heritability

A convenient single parameter to describe the trait genetic variance *in a particular population* is the heritability h^2 defined as

$$h^2 = \frac{\sigma^2_{A_T}}{\sigma^2_Y}$$

or the proportion of phenotypic (trait) variance due to additive allelic effects.

The phenotypic variance, the genetic variance and the additive and dominance variance components all depend on trait genotypic (or allele) frequencies and so are different in different populations.

The genotypic effects G are not known but the variance components and heritability can be estimated.

Association Mapping

Association methods use random samples from a population and are alternatives to methods based on pedigrees or crosses between inbred lines. The associations depend on linkage disequilibrium between marker and trait loci instead of depending on linkage between those loci as in pedigree or line cross methods.

The chances of detecting an association between a trait and a genetic marker are high if:

The trait has a high genetic component.
The marker has high linkage disequilibrium with the trait genes (depends on allele frequencies at trait and marker loci).
The sample sizes are large.

Marker-Trait Genotypes

Until the trait locus is identified, the trait genotype cannot be observed, but maybe it can be inferred, and the location of the locus estimated, from observations on the trait and the genotype at a genetic marker **M**.

Although there are several types of genetic markers, attention here will be restricted to those with only two alleles M, m: e.g. SNPs. Individuals with the same marker genotype can have different trait genotypes and a way to describe joint marker-trait genotypes is needed.

Marker-Trait Genotype Frequencies

With random mating, (two-locus) genotype frequencies are products of gamete frequencies. For example

$$\mathrm{Pr}(MM, TT) \ = \ \mathrm{Pr}(MT)^2$$

and gamete frequencies involve allele frequencies and linkage disequilibria:

$$\mathrm{Pr}(MT) \ = \ p_M p_T + D_{MT}$$

Two-allele Genotypes

	TT	Tt	tt
MM	P_{MT}^2	$2P_{MT}P_{Mt}$	P_{Mt}^2
Mm	$2P_{MT}P_{mT}$	$2P_{MT}P_{mt} + 2P_{Mt}P_{mT}$	$2P_{Mt}P_{mt}$
mm	P_{mT}^2	$2P_{mT}P_{mt}$	P_{mt}^2

Two-allele Gametes

	T	t
M	$P_{MT} = p_M p_T + D_{MT}$	$P_{Mt} = p_M p_t - D_{MT}$
m	$P_{mT} = p_m p_T - D_{MT}$	$P_{mt} = p_m p_t + D_{MT}$

Linkage disequilibrium can be regarded as the covariance of marker and trait allele frequencies, and this can be transformed to the correlation of marker and trait allele frequencies:

$$\begin{aligned} \rho_{MT} &= \frac{D_{MT}}{\sqrt{p_M p_m p_T p_t}} \\ \rho^2_{MT} &= \frac{D^2_{MT}}{p_M p_m p_T p_t} \end{aligned}$$

Marker and Trait Variables

The trait genotypic values are not known, but they can be summarized by the additive and dominance components of variance.

An analogous system can be considered for the marker genotypes. These genotypes are observed and their genotypic values can be assigned. Consider variables X for marker locus **M**. As before, a Hardy-Weinberg assumption provides the following expressions for the mean and variance:

$$\begin{aligned} \mathcal{E}(X) = \mu_X &= p_M^2 X_{MM} + 2p_M p_m X_{Mm} + p_m^2 X_{mm} \\ \text{Var}(X) &= \sigma^2_{A_M} + \sigma^2_{D_M} \end{aligned}$$

Components of Variance and Covariance

The additive and dominance components of variance for the marker are

$$\begin{aligned}\sigma^2_{A_M} &= 2p_Mp_m[p_M(X_{MM}-X_{Mm})+p_m(X_{Mm}-X_{mm})]^2\\ \sigma^2_{D_M} &= p^2_Mp^2_m(X_{MM}-2X_{Mm}+X_{mm})^2\end{aligned}$$

and these lead to the following expression for the covariance of X and G:

$$\text{Cov}(G,X) = \rho_{MT}\sigma_{A_T}\sigma_{A_M}+\rho^2_{MT}\sigma_{D_T}\sigma_{D_M}$$

Correlation of Trait and Marker Variables

If *either* X or G are purely additive (e.g. X is allele dosage), so that σ_{D_T} or σ_{D_M} is zero, then

$$\text{Cov}(G, X) \;=\; \rho_{MT}\sigma_{A_T}\sigma_{A_M}$$

If *both* X of G are purely additive, then $\sigma^2_G = \sigma^2_{A_T}$ and $\sigma^2_X = \sigma^2_{A_M}$ and

$$\rho_{GX} \;=\; \rho_{MT}$$

so the correlation between the trait and marker effects is just the linkage disequilibrium between trait and marker loci.

As the trait Y is $G + E$, in this case $\text{Cov}(X, Y) = \text{Cov}(X, G)$ and $\text{Corr}(X, Y) = (\rho_{MT}\sigma_{A_T}\sigma_{A_M})/(\sigma_X\sigma_Y)$. In this additive case, $\text{Corr}(X, Y) = \rho_{MT}\sqrt{h^2}$ (and $\text{Corr}(X, G) = \rho_{MT}$).

Measured Traits

Suppose $Y = G + E$ where G is the genetic effect of locus **T** and E are all other effects. These other effects are supposed to have mean zero and to be independent of both G and the marker variable X. Then

$$\begin{aligned}
\mathcal{E}(Y) &= \mathcal{E}(G) \\
\text{Cov}(X, Y) &= \text{Cov}(X, G) \\
\sigma^2_Y &= \sigma^2_{A_T} + \sigma^2_{D_T} + \sigma^2_E
\end{aligned}$$

Trait values Y may be regressed on marker variables X. The regression coefficient is

$$\beta_{Y.X} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\rho_{MT}\sigma_{A_T}\sigma_{A_M} + \rho^2_{MT}\sigma_{D_T}\sigma_{D_M}}{\sigma^2_{A_M} + \sigma^2_{D_M}}$$

Additive Marker Variable

Variable X may be chosen to be additive, e.g $X_{MM} = 2, X_{Mm} = 1, X_{mm} = 0$ so that $\sigma^2_{A_M} = 2p_M p_m$, $\sigma^2_{D_M} = 0$, and then the regression of trait on (coded) marker genotype has slope

$$\beta_{Y.X} = \rho_{MT}\frac{\sigma_{A_T}}{\sigma_{A_M}}$$

A zero slope means either that there is no trait additive variance (unlikely) or that there is no linkage disequilibrium between marker and trait alleles. This, in turn, suggests that trait and marker loci are not close.

Trait Mean in Inbred Populations

Finding the mean and variance for quantitative traits in populations where there is inbreeding and/or relatedness and Hardy-Weinberg equilibrium does not hold, requires modification of genotype probabilities. For inbred populations, the mean trait value requires the inbreeding coefficient. For a random member of a population inbred to an extent F (relative to a reference population), the genotype probabilities are

$$\begin{aligned} P_{TT} &= \pi_T^2 + F\pi_T\pi_t \\ P_{Tt} &= 2\pi_T\pi_t(1-F) \\ P_{tt} &= \pi_t^2 + F\pi_T\pi_t \end{aligned}$$

Trait Mean in Inbred Populations

The expected trait value μ_F in an inbred population is

$$\mu_F = \mu_0 + FH$$

where μ_0 is the value in a HWE population ($F = 0$) and $H = \pi_T\pi_t(G_{TT} - 2G_{Tt} + G_{tt})$ is a measure of dominance.

Inbreeding Depression

From data on 1.4 million individuals, "F_{ROH} equivalent to the offspring of first cousins is associated with a 55% decrease in the odds of having children."

Note the need for estimation of individual inbreeding coefficients.

Clark DW, et al. 2019. Nature Communications. Published October 31, 2019:

Clark et al., 2019

Nearly one billion people live in populations where consanguineous marriages are common.

Burden of disease thought to be disproportionately due to increased homozygosity of rare, recessive variants.

The fraction of each autosomal genome in ROH > 1.5 Mb correlates well with pedigree-based estimates of inbreeding.

Genetic Variance and Covariance in Inbred Populations

For individual j, the genetic variance for an additive trait is

$$\sigma^2_{G_j} \;=\; (1 + F_j)\sigma^2_A$$

For individuals j, j', the genetic covariance for an additive trait is

$$\text{Cov}_{G_{jj'}} \;=\; 2\theta_{jj'}\sigma^2_A$$

regardless of inbreeding.

Total Variances and Covariances for Additive Trait

Trait values have both genetic and environmental components. The simplest model of $Y = G + E$ leads to the variance of trait values Y among individuals j in a non-inbred population of unrelated individuals:

$$\text{Var}(Y_j) \;=\; \sigma^2_A + \sigma^2_E$$

This is also referred to as the phenotypic variance σ^2_P.

For an additive trait and for individuals that have no shared environment, the variance-covariance matrix for a sample of related and inbred individuals has elements

$$\begin{aligned} \text{Var}(Y_j) &= (1 + F_j)\sigma^2_A + \sigma^2_E \\ \text{Cov}(Y_j, Y_{j'}) &= 2\theta_{jj'}\sigma^2_A \end{aligned}$$

Genetic Relationship Matrix

Vector $\boldsymbol{Y}$ of trait values for individuals $i = 1, 2, \ldots n$ has GRM

$$\boldsymbol{G} = \begin{bmatrix} (1+F_1) & 2\theta_{12} & \ldots & 2\theta_{1n} \\ 2\theta_{21} & (1+F_2) & \ldots & 2\theta_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 2\theta_{n1} & 2\theta_{2n} & \ldots & (1+F_n) \end{bmatrix}$$

The trace of this matrix is

$$\text{tr}(\boldsymbol{G}) = \sum_{j=1}^{n} (1+F_j) = n(1+F_I)$$

and the sum of the off-diagonal elements is

$$\Sigma_{\boldsymbol{G}} = \sum_{j=1}^{n} \sum_{\substack{j'=1 \\ j \neq j'}}^{n} 2\theta_{jj'} = n(n-1)2\theta_S$$

to define the average inbreeding and kinship values F_I, θ_S for the sample.

GRM

Historically, the GRM was obtained from known pedigrees, particularly for experimental populations of plants and animals.

Beginning with Yu J, et al. 2006. Nature Genetics 38:203 it has been recognized that pedigree information may not be available, it may not be accurate, and it can be different from the "gold standard" GRM.

Instead, the GRM may be constructed with estimated inbreeding and kinship coefficients.

Heritability Estimation

Heritability

For an additive trait, heritability in a HWE population is

$$h^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2}$$

Estimation of h^2 therefore requires estimation of σ_A^2 and σ_E^2. There are likelihood-based methods for doing that, assuming the trait values are normally distributed.

These notes follow a discussion given by Speed D, et al. 2012. Am J Hum Genet 91:1011

Speed et al. 2012

Instead of having replicates of the history of a single individual, use the trait values for a sample of individuals in some population. Speed et al. used $\widehat{V}_T$ for the sample variance of trait values:

$$\widehat{V}_T = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$$

and $\widehat{V}_R$ for the residual variance once the genotypic effects have been fitted:

$$\widehat{V}_R = \frac{1}{n-1}\sum_{j=1}^{n}(E_j - \bar{E})^2$$

As an estimate of heritability, Speed et al. combined these two sample variances

$$\widehat{h^2} = \frac{\widehat{V}_T - \widehat{V}_R}{\widehat{V}_T}$$

Speed et al. 2012

It can be shown that

$$\begin{aligned}\mathcal{E}(\widehat{V}_T) &= \frac{1}{n}\left[\mathrm{tr}(\boldsymbol{G}) - \frac{1}{n-1}\Sigma_{\boldsymbol{G}}\right]\sigma^2_A + \sigma^2_E \\ \mathcal{E}(\widehat{V}_R) &= \sigma^2_E\end{aligned}$$

so that

$$\mathcal{E}(\widehat{h^2}) = \frac{\frac{1}{n}\left[\mathrm{tr}(\boldsymbol{G}) - \frac{1}{n-1}\Sigma_{\boldsymbol{G}}\right]\sigma^2_A}{\frac{1}{n}\left[\mathrm{tr}(\boldsymbol{G}) - \frac{1}{n-1}\Sigma_{\boldsymbol{G}}\right]\sigma^2_A + \sigma^2_E}$$

and this has a parametric value of

$$\mathcal{E}(\widehat{h^2}) = \frac{(1+F_I-2\theta_S)\sigma^2_A}{(1+F_I-2\theta_S)\sigma^2_A + \sigma^2_E}$$

Expectation of $\widehat{h^2}$

In the case of no identity by descent within or between individuals, $F_W = \theta_S = 0$,

$$\mathcal{E}(\widehat{h^2}) \;=\; \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2} = h^2$$

In general, however, the expectation of $\hat{h}^2$ is

$$\mathcal{E}(\widehat{h^2}) \;=\; \frac{(1 + F_I - 2\theta_S)\sigma_A^2}{(1 + F_I - 2\theta_S)\sigma_A^2 + \sigma_E^2} = \frac{(1 - \theta_S)(1 + f)\sigma_A^2}{(1 - \theta_S)(1 + f)\sigma_A^2 + \sigma_E^2}$$

where $f = (F_I - \theta_S)/(1 - \theta_S)$.

Alternatively,

$$\mathcal{E}(\widehat{h^2}) \;=\; \frac{(1 + F)\sigma_A^2}{(1 + F)\sigma_A^2 + \sigma_E^2}$$

is often given as though the study population has inbreeding but no kinship: $F = f, \theta_S = 0$. It does not seem likely that a natural population could consist of inbred but unrelated individuals.

Expectation of $\widehat{h^2}$

For a population in Hardy-Weinberg equilibrium, $F_I = \theta_S, f = 0$:

$$\mathcal{E}(\widehat{h^2}) \;=\; \frac{(1-\theta_S)\sigma_A^2}{(1-\theta_S)\sigma_A^2 + \sigma_E^2}$$

and then $\widehat{h^2}$ will be close to unbiased if θ_S is low.

Use of Estimated GRM

The Speed et al. estimate uses two sample variances, and does not make explicit use of the GRM. Likelihood-based methods do use the GRM. As the parametric values F_i, θ_{ij} are not generally known, $\boldsymbol{G}$ is replaced by a matrix $\boldsymbol{K}$ of estimates. What is the resulting heritability estimate then estimating?

Use of Allele-sharing GRM

Can estimate half the GRM with $\hat{K}_{\mathrm{AS}}$ having elements $\{\widehat{\psi}_{jj'}\}$ for row j and column j'. If $\tilde{A}_{jj'}$ is the allelic matching proportion, averaged over SNPs, for individuals j and j' including $j = j'$, the ψ estimates are

$$\widehat{\psi}_{jj'} \;=\; \frac{\tilde{A}_{jj'} - \tilde{A}_S}{1 - \tilde{A}_S}$$

where $\tilde{A}_{jj'} = \sum_{l=1}^{L}[1+(X_{jl}-1)(X_{j'l}-1)]/(2L)$ for allelic dosages X_{jl} and $\tilde{A}_S = \sum_{j\neq j'} \tilde{A}_{jj'}/[n(n-1)]$. These estimates have expected values

$$\mathcal{E}(\widehat{\psi}_{jj'}) \;=\; \begin{cases} \frac{\frac{1}{2}(1+F_j)-\theta_S}{1-\theta_S} = 1 + f_j & j = j' \\ \\ \frac{\theta_{jj'}-\theta_S}{1-\theta_S} = \psi_{jj'} & j \neq j' \end{cases}$$

Use of Allele-sharing GRM

As $\Sigma_{\hat{K}_{\text{AS}}} = 0$ by construction, the expectation of the estimated heritability is

$$\mathcal{E}(\widehat{h}^2) \;=\; \frac{\mathcal{E}[\frac{2}{n}\text{tr}(\hat{K}_{\text{AS}})\hat{\sigma}^2_A]}{\mathcal{E}[\frac{2}{n}\text{tr}(\hat{K}_{\text{AS}})\hat{\sigma}^2_A + \hat{\sigma}^2_E]}$$

From the expected values of $\widehat{\psi}_{jj'}$, $\mathcal{E}[\text{tr}(\hat{K}_{\text{AS}})] = n(1+f)/2$ is assumed known and replaces $\widehat{K}_{\text{AS}}$, leading to

$$\mathcal{E}(\widehat{h}^2) \;=\; \frac{(1+f)\sigma^2_A}{(1+f)\sigma^2_A + \sigma^2_E}$$

This replaces F in the classical result with f, reflecting that is f and not F that can be estimated with data from a single population.

Use of Standard GRM

Can also estimate half the GRM with $\hat{K}_{\mathrm{c}0}$ having elements $\{\widehat{k}_{jj'}\}$:

$$\widehat{k}_{jj'} = \frac{\sum_l (X_{jl} - 2\tilde{p}_l)(X_{j'l} - 2\tilde{p}_l)}{\sum_l 4\tilde{p}_l(1 - \tilde{p}_l)}$$

Now all the elements of the GRM sum to zero by construction. In other words $\mathrm{tr}[\hat{K}_{\mathrm{c}0}] + \Sigma_{\hat{K}_{\mathrm{c}0}} = 0$ and the estimated heritability is

$$\widehat{h}^2 = \frac{\frac{2}{n-1}\mathrm{tr}[\hat{K}_{\mathrm{c}0}]\widehat{\sigma}^2_A}{\frac{2}{n-1}\mathrm{tr}[\hat{K}_{\mathrm{c}0}]\widehat{\sigma}^2_A + \widehat{\sigma}^2_E}$$

Use of Standard GRM

Since

$$\mathcal{E}(\hat{k}_{jj}) = \frac{1}{2}(1 + f_j - 4\psi_j)$$

$$\mathcal{E}[\text{tr}(\hat{K}_{\text{C0}})] \;=\; \frac{n}{2}(1 + f)$$

Since $\sum_j \psi_j = 0$, and the expected value of the estimated heritability is

$$\begin{aligned}\mathcal{E}(\hat{h}^2) \;&=\; \frac{\frac{n}{n-1}(1+f)\sigma_A^2}{\frac{n}{n-1}(1+f)\sigma_A^2 + \sigma_E^2} \\ &\approx\; \frac{(1+f)\sigma_A^2}{(1+f)\sigma_A^2 + \sigma_E^2}\end{aligned}$$

as for the allele-sharing estimate.

Very different GRMs give the same estimates of heritability.

GWAS

Mixed Linear Model

A simple mixed linear model for a vector of trait values $\boldsymbol{Y}$ is

$$\boldsymbol{Y} \;=\; \boldsymbol{X\beta} + \boldsymbol{g} + \boldsymbol{e} \text{ with } \mathsf{Var}(\boldsymbol{Y}) = \boldsymbol{G}\sigma_A^2 + \mathbf{I}\sigma_e^2$$

The vector $\boldsymbol{Y}$ of trait values for a set of individuals is equated to fixed effects $\boldsymbol{\beta}$, including effects of SNPs of interest and maybe eigenvectors from principal components analysis to account for population structure, plus random effects $\boldsymbol{g}$ for the total polygenic background for each individual. The trait value for an individual is assumed here to depend additively on its constituent alleles.

$\mathbf{G}$ is the Genetic Relatedness Matrix.

Simulation Study

Simulation study with 20,000 SNPs from the 1000 Genomes data on the MXL and ASW combined data set and a trait constructed to have genetic contribution from five SNPs and a population effect added to one of the populations.

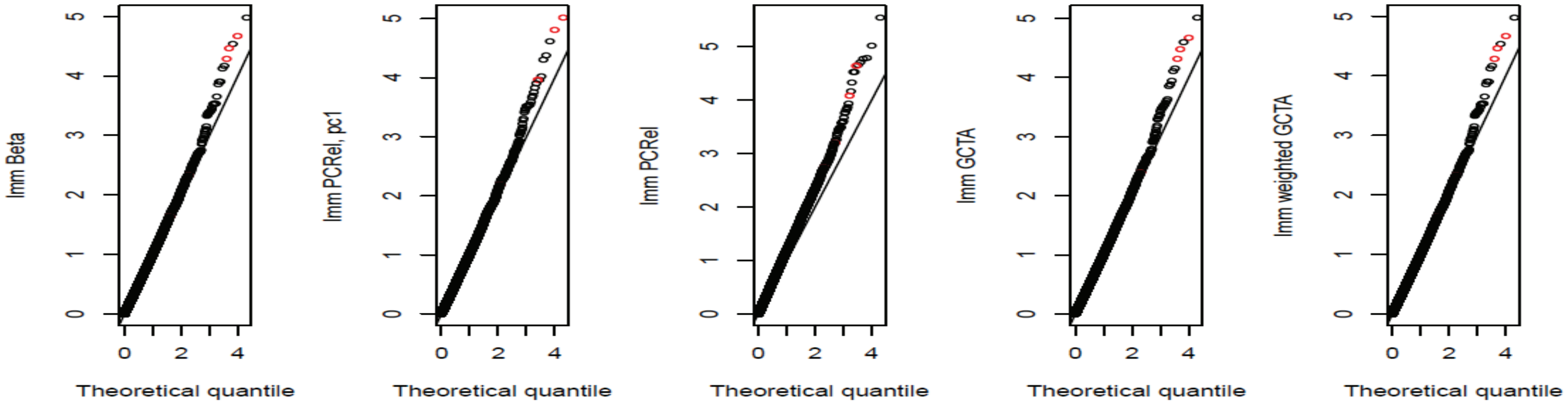
Figure on next slide shows the QQ plots for the strength of association ($-\log_{10} p$ value) for each SNP, with red dots indicating the causal SNPs. The allele-sharing method gives a similar plot to PC-Air/PC-Relate with 1 PC, and does better for PC-Air/PC-Relate with no PCs. Essentially no difference in plots that used the allele-sharing or the Standard GRMs, and both indicated a non-causal SNP as being most associated with the trait.

Conomos MP et al. 2015. Genet Epi 39:276-293.

Conomos MP et aal. 2016. Am J Hum Genet 98:127-148.

Schick UM et al. 2016. Am J Hum Genet 98:229-242.

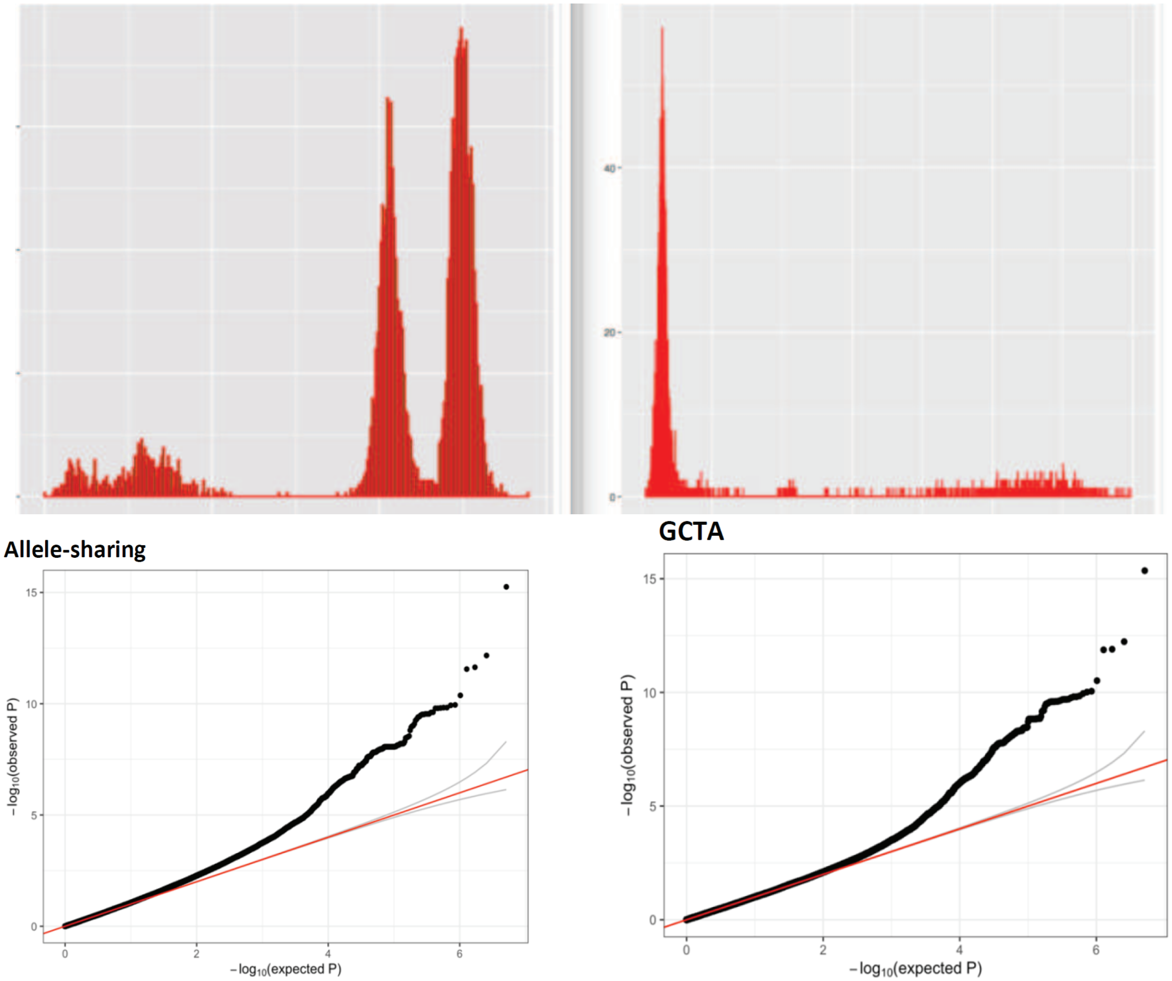
Simulation Study



QQ plots for association mapping of simulated data. Left to right: allele-sharing GRM with no PCs, PC-Air/PC-Relate with 1 PC, PC-Relate with no PCs, Standard GRM, unweighted over SNPs, Standard GRM, weighted over SNPs.

Empirical Study

An empirical study was made of a quantitative trait in a study with European Non-Hispanic, Hispanic and African American participants. The GRM inbreeding elements were quite different for allele-sharing and standard GRMs, but the QQ plots for marker-trait association were very similar.


Allele-sharing
GCTA
−log10(observed P)
−log10(expected P)