Assortative mating: sib- vs pop-GWAS

This is a note on how sib- and pop-GWAS perform under assortative mating. Simple scenarios are considered analytically to build intuition. More complicated scenarios are investigated with simulations.

1. Effect size estimates and their standard error

Case 1. One locus model (i.e. monogenic trait).

In this scenario the model for phenotype, Y, and for sib differences, ΔY, are:

$$Y = \beta_x X + e$$

$$\Delta Y = \beta_x \Delta X + \Delta e$$

where X is the genotype, β_x is the effect size, and e the noise.

It is straightforward to show that effect sizes from sib- and pop-GWAS are unbiased independent of assortative mating:

$$E\left(\hat{\beta}_x^{ur}\right) = \frac{\text{cov}(X,Y)}{\text{var}(X)} = \frac{\beta_x \text{var}(X) + \text{cov}(X,e)}{\text{var}(X)} = \beta_x$$

$$E\left(\hat{\beta}_x^{s}\right) = \frac{\text{cov}(\Delta X,\Delta Y)}{\text{var}(\Delta X)} = \frac{\beta_x \text{var}(\Delta X) + \text{cov}(\Delta X,\Delta e)}{\text{var}(\Delta X)} = \beta_x$$

assuming $\text{cov}(X,e) = 0$ and $\text{cov}(\Delta X,\Delta e) = 0$.

In pop-GWAS of sample size N:

$$\text{var}\left(\hat{\beta}_x^{ur}\right) \approx \frac{\text{var}(e)}{N\ \text{var}(X)}$$

The variance term in the denominator is

$$\text{var}(X) = \text{var}(X_M + X_P) = \text{var}(X_M) + \text{var}(X_P) + 2\ \text{cov}(X_M, X_P)$$

where X_M and X_P are maternally and paternally transmitted alleles. The covariance term, $\text{cov}(X_M, X_P)$, is nonzero under assortative mating, and increases with the parental phenotypic covariance. Thus, $\text{var}\left(\hat{\beta}_x^{ur}\right)$ decreases with assortative mating.

Note 1: All quantities above (and in the remaining) are written with reference to the generation following the last generation of assortative mating. For example, consider a population assortative mating at generation t. The term $\text{cov}(X_M, X_P)$ in equation above refers to the

covariance between maternally and paternally transmitted alleles in generation $t+1$. Specifically, $\text{cov}(X_M, X_P)|_{t+1}$ is a function of the $\text{cov}(X_M, X_P)|_t$ induced by assortative mating in all previous generations, and the degree of assortative mating in generation t. For convenience, a subscript denoting the generation is only used when not referring to generation $t+1$.

In sib-GWAS of size N_{pair}:

$$\text{var}\left(\hat{\beta}_x^s\right) \approx \frac{\text{var}(\Delta e)}{N_{pair}\ \text{var}(\Delta X)}$$

The variance term in the denominator is

$$\text{var}(\Delta X) = \text{var}(X_1 - X_2) = \text{var}(X_1) + \text{var}(X_2) - 2\ \text{cov}(X_1, X_2) \qquad [1]$$

where 1 and 2 denote the two siblings. Assuming symmetry between siblings:

$$\text{var}(X_1) = \text{var}(X_2) = \text{var}(X)$$

The covariance between sibling genotypes can be written in terms of transmitted alleles:

$$\text{cov}(X_1, X_2) = \text{cov}(X_{1M} + X_{1P}, X_{2M} + X_{2P})$$

$$= \text{cov}(X_{1M}, X_{2M}) + \text{cov}(X_{1M}, X_{2P}) + \text{cov}(X_{1P}, X_{2M}) + \text{cov}(X_{1P}, X_{2P}) \qquad [2]$$

For siblings we have:

$$\text{cov}(X_{1M}, X_{2M}) = [\text{var}(X_M)|_t\ +\ \text{cov}(X_M, X_P)|_t]/2$$

$$\text{cov}(X_{1P}, X_{2P}) = [\text{var}(X_P)|_t\ +\ \text{cov}(X_M, X_P)|_t]/2$$

These relationships are derived as follows: for each sibling we are sampling from maternal and paternal alleles with replacement. Half of the time the same allele is chosen, in which case their covariance equals the variance of the allele in the parental generation (the first terms on the right hand sides). And half of the time the transmitted alleles are different, in which case their covariance equals their covariance in the parental generation (the second terms on the right hand sides).

Note 2: Throughout the text it is assumed that allele frequencies are not changing over time, i.e. $\text{var}(X_M)|_t = \text{var}(X_M)|_{t+1}$ and $\text{var}(X_P)|_t = \text{var}(X_P)|_{t+1}$. Therefore, as stated in "Note 1", the generation subscripts are omitted for these quantities, even if they represent the quantitates in the previous generations.

Assuming symmetry between siblings:

$$\text{cov}(X_{1M}, X_{2P}) = \text{cov}(X_{1P}, X_{2M}) = \text{cov}(X_M, X_P)$$

Plugging these relationships into the sibling covariance equation above, eq. 2, we have:

$$\mathrm{cov}(X_1, X_2) = [\mathrm{var}(X_M) + \mathrm{var}(X_P) + 2\,\mathrm{cov}(X_M, X_P)|_t]/2 + 2\,\mathrm{cov}(X_M, X_P)$$

And the $\mathrm{var}(\Delta X)$ in eq.1 becomes:

$$\mathrm{var}(\Delta X) = \mathrm{var}(X_1) + \mathrm{var}(X_2) - 2\,\mathrm{cov}(X_1, X_2)$$

$$= 2\left[\mathrm{var}(X_M) + \mathrm{var}(X_P) + 2\,\mathrm{cov}(X_M, X_P)\right] - \mathrm{var}(X_M) - \mathrm{var}(X_P) - 2\,\mathrm{cov}(X_M, X_P)|_t - 4\,\mathrm{cov}(X_M, X_P)$$

$$= \mathrm{var}(X_M) + \mathrm{var}(X_P) - 2\,\mathrm{cov}(X_M, X_P)|_t$$

Now consider two scenarios: (i) $\mathrm{cov}(X_M, X_P)|_t = 0$. This refers to the case of assortative mating starting at generation t (and random mating up to that). Under this scenario $\mathrm{var}(\Delta X)$ does not change with assortative mating (assortative mating for only one generation).
(ii) $\mathrm{cov}(X_M, X_P)|_t \neq 0$. This refers to the case where assortative mating has been going on for more than one generation (or any other process that generates such covariance between alleles, e.g. population structure). $\mathrm{cov}(X_M, X_P)$ increases with time (number of generations of assortative mating) until it reaches its maximum value at steady state (i.e. continuous stable assortative mating over time), $\mathrm{cov}(X_M, X_P)|_{ss}$. My intuition is that $\mathrm{cov}(X_M, X_P)|_{ss}$ increases with the strength of assortative mating within one generation. Therefore, considering the population at equilibrium, $\mathrm{var}(\Delta X)$ decreases with assortative mating, and $\mathrm{var}\left(\hat{\beta}_x^s\right)$ increases.

Note 3: The noise terms, e and Δe (here and also in the polygenic case below) represent the non-genetic contributions, and not the residuals when regressing on a particular genetic effect. Assortative mating on the phenotype, in addition to genetic components, induces correlations in non-genetic components in the parents. Here I assume no transmission of the non-genetic parental components to the children. Consequently, e and Δe will be independent of assortative mating in the parental generation.

Case 2. Two loci trait.

Now, the model is

$$Y = \beta_x X + \beta_z Z + e$$

where X and Z are the genotypes at the two loci (assumed to be independent under random mating), and β_x and β_z are the corresponding effect sizes.

In pop-GWAS regressing Y on X:

$$E\left(\hat{\beta}_x^{ur}\right) = \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)} = \frac{\beta_x \mathrm{var}(X) + \beta_z\,\mathrm{cov}(X, Z) + \mathrm{cov}(X, e)}{\mathrm{var}(X)} = \beta_x + \beta_z \frac{\mathrm{cov}(X, Z)}{\mathrm{var}(X)}$$

assuming $\mathrm{cov}(X, e) = 0$. Under assortative mating $\mathrm{cov}(X, Z)$ is nonzero (and increases with strength of assortative mating per generation, as well as the number of assortative mating generations), and so $E\left(\hat{\beta}_x^{ur}\right)$ is biased. In words, X captures the effect of Z, depending on the strength of correlation between the loci induced by assortative mating.

Considering variance of the estimate:

$$\mathrm{var}\left(\hat{\beta}_x^{ur}\right) \approx \frac{{\beta_z}^2\ \mathrm{var}(Z) + \mathrm{var}(e)}{N\ \mathrm{var}(X)}$$

Both $\mathrm{var}(Z)$ and $\mathrm{var}(X)$ increase with assortative mating, and how $\mathrm{var}\left(\hat{\beta}_x^{ur}\right)$ changes depends on the relative change of these quantities. In the polygenic limit, the numerator includes all genotypes other than X, as well as their covariance induced by assortative mating, which would dominate the change in $\mathrm{var}\left(\hat{\beta}_x^{ur}\right)$. Therefore, in the polygenic limit, $\mathrm{var}\left(\hat{\beta}_x^{ur}\right)$ is expected to increase with assortative mating.

In sib-GWAS the model is:

$$\Delta Y = \beta_x \Delta X + \beta_z \Delta Z + \Delta e$$

And regressing ΔY on ΔX,

$$E\left(\hat{\beta}_x^{s}\right) = \frac{\mathrm{cov}(\Delta X, \Delta Y)}{\mathrm{var}(\Delta X)} \qquad [3]$$

The numerator is:

$$\mathrm{cov}(\Delta X, \Delta Y) = \mathrm{cov}(\Delta X, \beta_x \Delta X + \beta_z \Delta Z + \Delta e) = \beta_x \mathrm{var}(\Delta X) + \beta_z \mathrm{cov}(\Delta X, \Delta Z) \quad [4]$$

again assuming that $\mathrm{cov}(\Delta X, \Delta e) = 0$.

Now the covariance term on the right hand side of eq. 4 is

$$\mathrm{cov}(\Delta X, \Delta Z) = \mathrm{cov}(X_1 - X_2, Z_1 - Z_2)$$

$$= \mathrm{cov}(X_1, Z_1) + \mathrm{cov}(X_2, Z_2) - \mathrm{cov}(X_1, Z_2) - \mathrm{cov}(X_2, Z_1) \quad [5]$$

Assuming symmetry between siblings

$$\mathrm{cov}(X_1, Z_1) = \mathrm{cov}(X_2, Z_2)$$

$$\mathrm{cov}(X_1, Z_2) = \mathrm{cov}(X_2, Z_1)$$

Further assuming independent transmission of alleles at these two loci, given the parental alleles (i.e. random transmission):

$$\mathrm{cov}(X_1, Z_1) = \mathrm{cov}(X_1, Z_2)$$

$$\mathrm{cov}(X_2, Z_2) = \mathrm{cov}(X_2, Z_1)$$

Plugging into eq. 5,

$$\mathrm{cov}(\Delta X, \Delta Z) = 0$$

This relationship holds, regardless of the number of assortative mating generations.

Thus, eq. 3 becomes:

$$E\left(\hat{\beta}_x^s\right) = \frac{\beta_x \mathrm{var}(\Delta X) + \beta_z \mathrm{cov}(\Delta X, \Delta Z)}{\mathrm{var}(\Delta X)} = \beta_x$$

indicating that sib-GWAS effect size is unbiased under assortative mating. Considering variance of the estimate:

$$\mathrm{var}\left(\hat{\beta}_x^s\right) \approx \frac{{\beta_z}^2\ \mathrm{var}(\Delta Z) + \ \mathrm{var}(\Delta e)}{N_{pair}\ \mathrm{var}(\Delta X)}$$

In the polygenic limit, the numerator includes variance and covariance of the sib differences in genotypes across all other loci. As shown above, the covariance terms are zero, and variance terms do not change (in the case of one generation of assortative mating) or decrease (in the case of more than one generation of assortative mating) with assortative mating. Therefore, $\mathrm{var}\left(\hat{\beta}_x^s\right)$ either does not change or it decreases with assortative mating.

Simulations

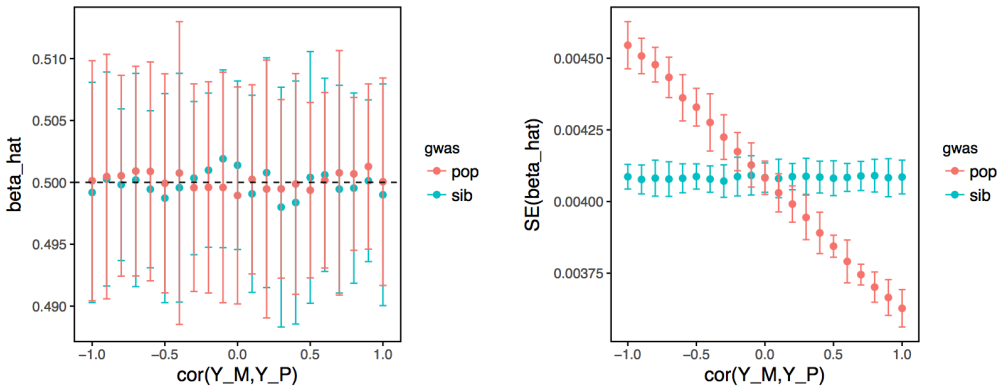
Here I show simple simulation results validating the results above, and vice versa.

Note 4: Details of the simulation procedure to be added. Briefly assortative mating is simulated by sorting and pairing parental phenotypes to generate a given correlation between them. Consider parental phenotypes Y_M and Y_P, two vectors of size N: (i) sort both vectors. (ii) generate random variables, two vectors of size N, using bivariate normal distribution with $\mu = \left(\overline{Y_M}, \overline{Y_P}\right)$ and $\Sigma = \begin{pmatrix} \sigma^2_{Y_M} & \rho\sigma_{Y_M}\sigma_{Y_P} \\ \rho\sigma_{Y_M}\sigma_{Y_P} & \sigma^2_{Y_P} \end{pmatrix}$, where ρ is the target phenotypic correlation between mates. (iii) reorder sorted Y_M and Y_P based on the rank of variables in (ii). *Caveat*: it works better when phenotypes are normally distributed and the sample sizes are large, though it may be good enough for our purpose as long as the realized correlations are used/plotted. Ideally one would want to mimic a realistic assortative mating process.

Note 5: For simplicity only one generation of assortative mating is simulated, but I think results can be qualitatively extended to the case of continued assortative mating (any differences between sib vs pop-GWAS observed for one generation of assortative mating would be amplified with more generations of assortative mating).

1. One locus case

Simulation parameters:
minor allele frequency $p = 0.2$; $\beta = 0.5$; $\text{var}(e) = 2\beta^2 p(1-p)$; pop-GWAS size $N = 15{,}000$; sib-GWAS size $N_{pair} = 30{,}000$

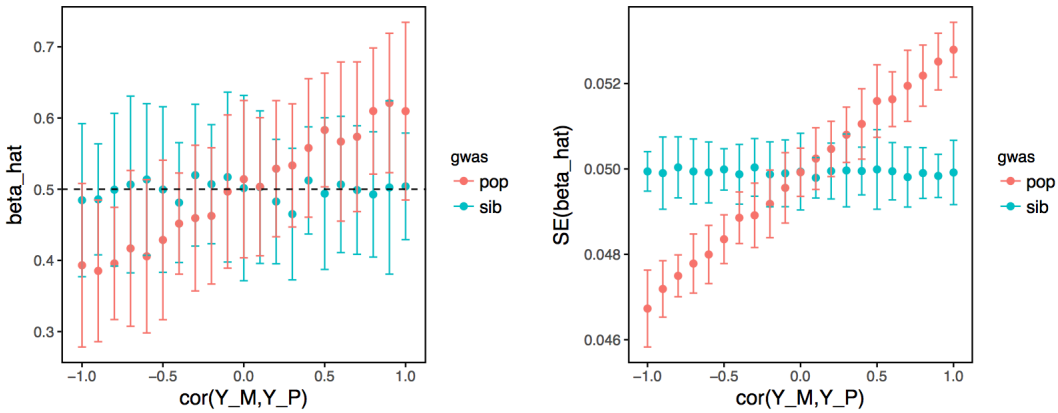


[to be added: replotting results with realized phenotypic correlations, see Note 4]

2. Polygenic case

Simulation parameters:
Number of loci $M = 100$ with equal minor allele frequencies and effect sizes for simplicity; across loci $p = 0.2$ and $\beta = 0.5$; $\text{var}(e) = 2M\beta^2 p(1-p)$, such that heritability is 0.5; pop-GWAS size $N = 20{,}000$; sib-GWAS size $N_{pair} = 30{,}000$

Below are estimated effect size and its standard error at one particular locus:



For both plots ratio of sample sizes were chosen such that pop-GWAS and sib-GWAS yield similar results under random mating. Error bars are standard errors for the parameter of interest, calculated over 20 iterations. X-axis denotes the phenotypic correlation between parents.

Summary 1

1. In sib regression, $E\left(\hat{\beta}_x^s\right)$ does not change with assortative mating.
2. In the case of only one generation of assortative mating, $\text{var}\left(\hat{\beta}_x^s\right)$ dose not change with the strength of parental phenotypic correlation.
3. If assortative mating continues for more than a generation, $\text{var}\left(\hat{\beta}_x^s\right)$ increases with assortative mating. But for a polygenic trait, where the contribution of the focal locus to the trait becomes small relative to the rest of the genome, the trend is reversed, i.e. $\text{var}\left(\hat{\beta}_x^s\right)$ decrease with assortative mating.
4. Unless the trait is monogenic, effect size estimates from pop-GWAS, $E(\hat{\beta}_x^{ur})$, are biased.
5. In the case of a monogenic trait, $\text{var}\left(\hat{\beta}_x^{ur}\right)$ decreases with assortative mating. For a polygenic trait, $\text{var}\left(\hat{\beta}_x^{ur}\right)$ increases with assortative mating.

Based on points 1-5 above, if we consider the plausible case of a polygenic trait and assortative mating for more than one generation:

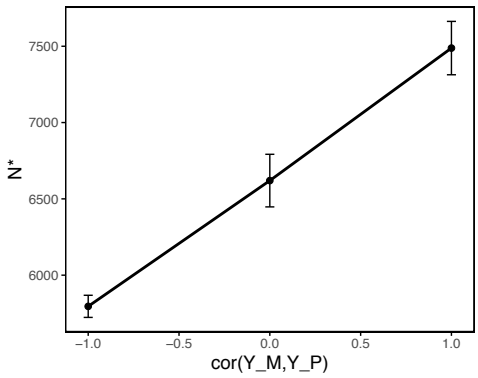
- $E\left(\hat{\beta}_x^s\right)$ does not change with assortative mating, but $E(\hat{\beta}_x^{ur})$ is biased in the direction of assortative mating, and the bias increases with the strength of phenotypic correlation between mates.
- $\text{var}\left(\hat{\beta}_x^s\right)$ decreases, but $\text{var}\left(\hat{\beta}_x^{ur}\right)$ increases with the strength of phenotypic correlation between mates.

2. Prediction in test set

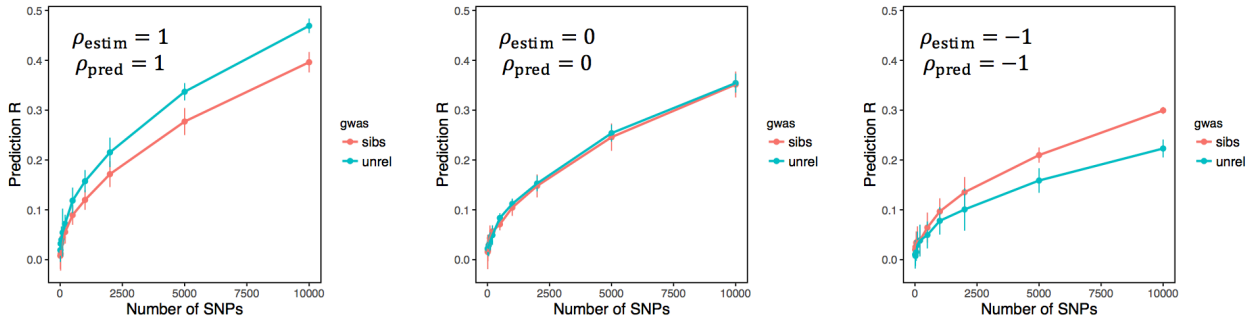
The goal of this section is to investigate how sib- and pop-GWAS perform with respect to prediction, i.e. how well polygenic scores with effect size estimates from the two designs (matched for effective sample size) predict the outcome in an independent test set.

First I consider simulations with following parameters (see Note 5 above):
Number of loci $M = 10000$ with equal minor allele frequencies and effect sizes for simplicity; across loci $p = 0.2$ and $\beta = 0.5$ (all loci are causal); $\text{var}(e) = 2M\beta^2 p(1-p)$, such that heritability is 0.5; sib-GWAS size $N_{pair} = 10{,}000$, test set size $N_{test} = 10{,}000$

The figure below shows how effective sample size changes with assortative mating. This trend is consistent with the conclusions in the previous section: $\text{var}\left(\hat{\beta}_x^s\right)$ does not change with assortative mating (for one generation), but $\text{var}\left(\hat{\beta}_x^{ur}\right)$ increases with assortative mating in the polygenic limit. Therefore, under assortative mating larger pop-GWAS sample sizes are required to match $\text{var}\left(\hat{\beta}_x^s\right)$.



The figure below shows prediction R (correlation between polygenic score and phenotype in the test set) with effect sizes from sib-GWAS, R_{sib}, and effect sizes from pop-GWAS (with sample size N^*), R_{unrel}, as a function of SNPs used to build polygenic scores for three values of phenotypic correlation between parental phenotypes, ρ:

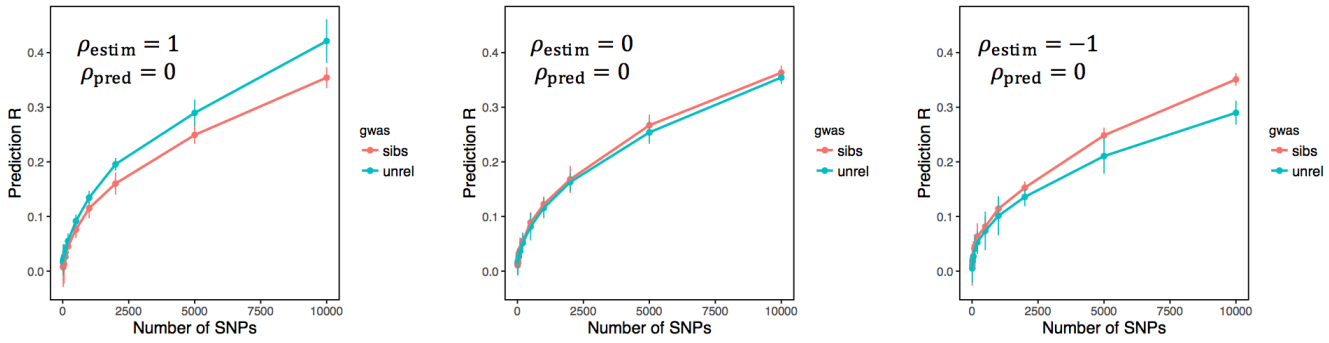


ρ_{estim} is the parental phenotypic correlation used to simulate the pop-GWAS set, and ρ_{pred} is the correlation used to simulate the test set. For all panels above, $\rho_{estim} = \rho_{pred}$, i.e. assortative mating behavior is the same in the pop-GWAS and the test set.

In these simulations no discovery GWAS step is simulated; rather causal SNPs are known and have the same effect size and allele frequencies. Thus the order at which SNPs are added to make the polygenic scores should not make a difference.

As figure shows, under positive assortative mating pop-GWAS outperforms sib-GWAS, while under negative assortative mating pop-GWAS underperforms sib-GWAS. The difference in prediction between sib- and pop-GWAS increases with using more SNPs in the score construction. Furthermore, the prediction accuracy for both designs increases with assortative mating in the test set.

Importantly, the divergence between pop- and sib-GWAS is mostly a consequence of assortative mating in the GWAS set and *not* the test set. To illustrate this, figure below replicates figure above but with no assortative mating in the test sets, i.e. $\rho_{pred} = 0$:



To make sense of these results: Let $S' = \sum_i \hat{\beta}_i X'_i$, denote the polygenic score in the test set, where $\hat{\beta}_i$ is the effect size estimated in GWAS (sib or pop) and X'_i is the genotype at locus i (the prime symbol denotes the quantities in the test set). Then prediction R is proportional to:

$$R \propto \frac{\text{cov}(S', Y')}{\sqrt{\text{var}(S')}} = \frac{\sum_i \hat{\beta}_i \text{ cov}(X'_i, Y')}{\sqrt{\sum_i \hat{\beta}_i^2 \text{ var}\left(X'_i\right) + \sum_i \sum_{j,j\neq i} \hat{\beta}_i \hat{\beta}_j \text{ cov}(X'_i, X'_j)}}$$

Consider a simple scenario, where (1) the test set is infinitely large such that all variability in prediction accuracy is from uncertainty in the beta estimates from GWAS, (2) there is no assortative mating in the test set, and (3) all causal SNPs have the same allele frequencies and effect sizes (as simulated above). Then:

$$R \propto \frac{\sum_i \hat{\beta}_i}{\sqrt{\sum_i \hat{\beta}_i^2}} \quad [6]$$

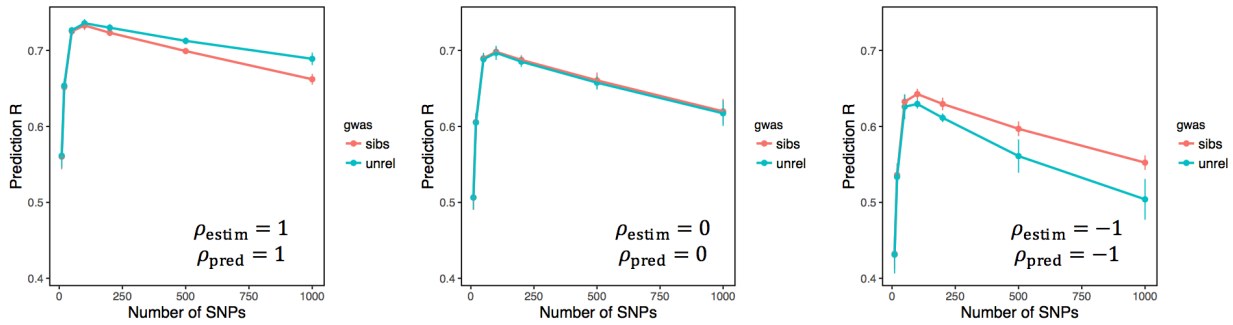
Now comparing sib- and pop-GWAS consider: $\hat{\beta}^s \sim N(\beta, \sigma_s^2)$ and $\hat{\beta}^{ur} \sim N(\beta + \beta_{bias}, \sigma_{ur}^2)$. Matching sample sizes then $\sigma_{ur}^2 = \sigma_s^2$. β_{bias} is the bias captured in pop-GWAS, which

is positive under positive assortative mating, and negative under negative assortative mating.

Under this model, it can be shown that the expected value of $\left.\frac{\sum_i \hat{\beta}_i}{\sqrt{\sum_i \hat{\beta}_i^2}}\right|_{ur} - \left.\frac{\sum_i \hat{\beta}_i}{\sqrt{\sum_i \hat{\beta}_i^2}}\right|_{s}$ increases with β_{bias} (i.e. assortative mating) and with number of SNPs, consistent with simulations.

So far, all loci were considered to be causal. The following simulations take 10% of the loci to be causal, and the rest to be null. Specifically:
Number of causal loci $M_c = 100$ with equal effect sizes for simplicity $\beta = 0.5$; Number of null loci $M_{null} = 900$ with $\beta = 0$; across loci $p = 0.2$; $\text{var}(e) = 2M\beta^2 p(1-p)$, such that heritability is 0.5; sib-GWAS size $N_{pair} = 10{,}000$, test set size $N_{test} = 10{,}000$

Under this scenario:



In all cases, with increasing SNP number prediction accuracy increases (all included loci are causal), until it reaches a maximum at the point beyond which added loci are null (same as above simulations no discovery GWAS step is simulated; first causal SNPs are added to the score followed by null SNPs). Beside this trend, the qualitative observations are similar to the case above where all loci were causal. Perhaps counterintuitively, the difference between sib- and pop-GWAS grows with incorporating more null SNPs. We can make sense of this by re-writing eq. 6 in terms of causal and null SNPs:

$$R \propto \frac{\sum_i \hat{\beta}_i}{\sqrt{\sum_i \hat{\beta}_i^2}} = \frac{\sum_i^{M_c} \hat{\beta}_i + \sum_i^{M_{null}} \hat{\beta}_i}{\sqrt{\sum_i^{M_c} \hat{\beta}_i^2 + \sum_i^{M_{null}} \hat{\beta}_i^2}}$$

Now, the expected value of the quantity $\sum_i^{M_{null}} \hat{\beta}_i$ in the numerator is zero, but the quantity $\sum_i^{M_{null}} \hat{\beta}_i^2$ in the denominator is always greater than zero, and grows with number of null SNPs leading to decrease in prediction accuracy. To what degree adding more null SNPs decreases R depends on the magnitude of $\sum_i^{M_{null}} \hat{\beta}_i^2$ relative to $\sum_i^{M_c} \hat{\beta}_i^2$. The quantity $\sum_i^{M_c} \hat{\beta}_i^2$ is larger under assortative mating, but $\sum_i^{M_{null}} \hat{\beta}_i^2$ is independent of assortative mating. Thus, the rate at which R drops with incorporating null SNPs varies by assortative mating (I think whether the difference grows or shrinks would depend on model specifics, i.e. the standard error of the effect size estimates, R at the peak, etc.)

[to be added: results with more realistic effect size and allele frequencies, also adding a GWAS]

Summary 2

Considering prediction:

1. Under positive (negative) assortative mating (in the GWAS set) pop-GWAS outperforms (underperforms) sib-GWAS.
2. The prediction accuracy for both designs increases with assortative mating in the *test* set.
3. Under assortative mating (in the GWAS set), the expected difference between sib- and pop-GWAS grows with including more *causal* SNPs.
4. Under assortative mating (in the GWAS set), the expected difference between sib- and pop-GWAS changes (grows or shrinks depending on parameters) with including more *null* SNPs.