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, i.e. $E\left(\hat{\beta}_x^s\right) = E\left(\hat{\beta}_x^{ur}\right) = \beta_x$.

In pop-GWAS of sample size N:

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

The variance term in the denominator is

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

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

In sib-GWAS of size N_{pair}:

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

The variance term in the denominator is

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

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

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

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

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

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

For siblings we have:

$$\mathrm{cov}(X_{1M}, X_{2M}) = \mathrm{var}(X_M)/2$$

$$\mathrm{cov}(X_{1P}, X_{2P}) = \mathrm{var}(X_P)/2$$

Assuming symmetry between siblings:

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

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

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

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

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

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

$$= \mathrm{var}(X_M) + \mathrm{var}(X_P)$$

which equals to $\mathrm{var}(\Delta X)$ under random mating. Put differently, assortative mating increases variance of the genotypes in the population and the covariance of sibling genotypes by the same amount, and so $\mathrm{var}(\Delta X)$ does not change with assortative mating.

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 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,

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

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:

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

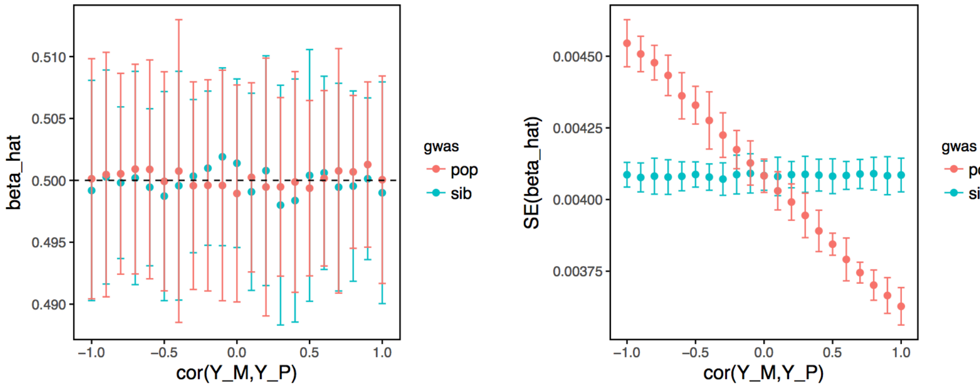
Both $\mathrm{var}(\Delta X)$ and $\mathrm{var}(\Delta Z)$ do not change with assortative mating. In the polygenic limit, the numerator includes variance and covariance of the sib differences in genotypes across all other loci. As shown above, the covariance terms are zero, and variance terms do not depend on mating pattern. Therefore, $\mathrm{var}\left(\hat{\beta}_x^s\right)$ does not change with assortative mating.

Simulations

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

1. One locus case

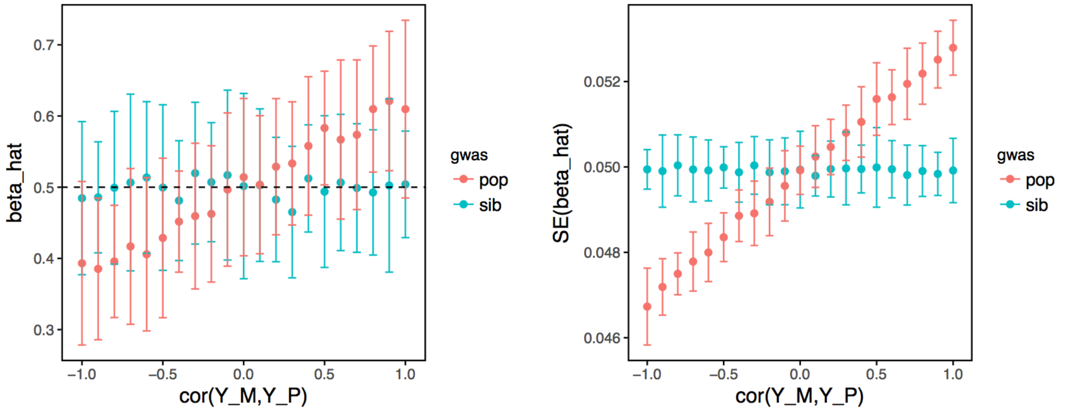
Simulation parameters:
minor allele frequency $p = 0.2$; $\beta = 0.5$; $\mathrm{var}(e) = 2\beta^2 p(1 - p)$; pop-GWAS size $N = 15{,}000$; sib-GWAS size $N_{pair} = 30{,}000$



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 $\beta = 0.5$; $\mathrm{var}(e) = 2M\beta^2 p(1-p)$, 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.

Conclusions 1

1. In sib regression, neither $E\left(\hat{\beta}_x^s\right)$ nor $\mathrm{var}\left(\hat{\beta}_x^s\right)$ change with assortative mating.
2. Unless the trait is monogenic, effect size estimates from pop-GWAS, $\hat{\beta}_x^{ur}$, are biased.

3. In the case of monogenic trait, $\mathrm{var}\left(\hat{\beta}_x^{ur}\right)$ decreases 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. $\mathrm{var}\left(\hat{\beta}_x^{ur}\right)$ increase with assortative mating.

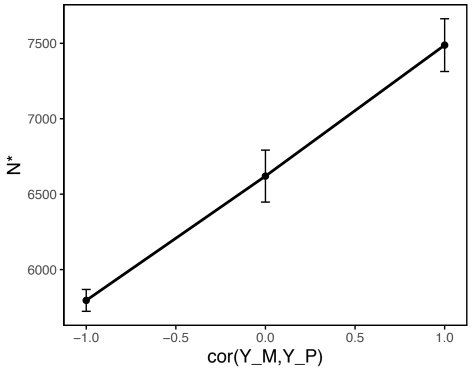
These conclusions (I think) are robust to multi-generational assortative mating.

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.

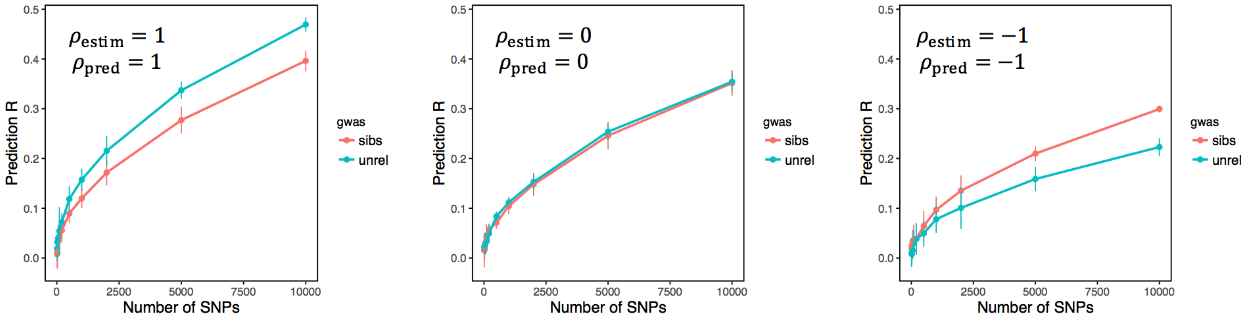
First I consider simulations with following parameters:
Number of loci $M = 10000$ with equal minor allele frequencies and effect sizes for simplicity; across loci $p = 0.2$ and $\beta = 0.5$ (all loci are causal); $\mathrm{var}(e) = 2M\beta^2 p(1-p)$, such that heritability is 0.5; sib-GWAS size $N_{pair} = 10{,}000$, 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: $\mathrm{var}\left(\hat{\beta}_x^{s}\right)$ does not change with assortative mating, but $\mathrm{var}\left(\hat{\beta}_x^{ur}\right)$ increases with assortative mating in the polygenic limit. Therefore, under assortative mating larger pop-GWAS sample sizes are required to match $\mathrm{var}\left(\hat{\beta}_x^{s}\right)$.

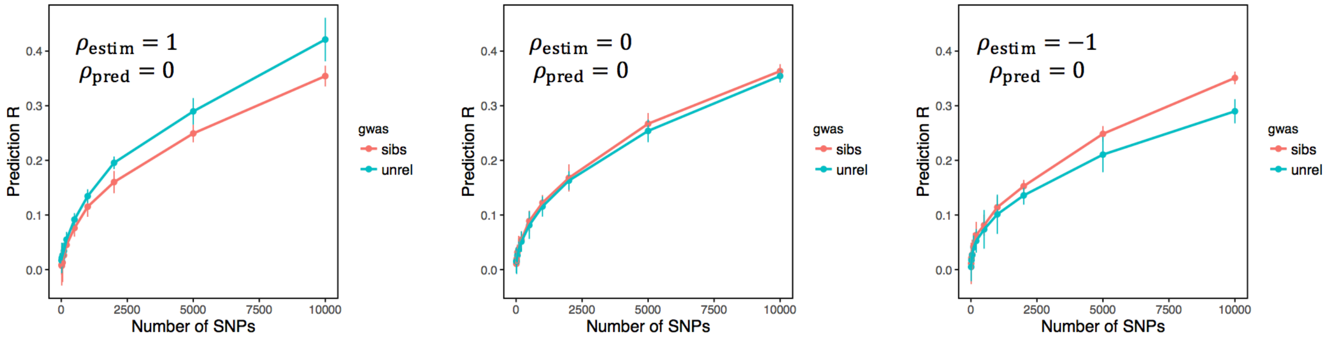
The figure below shows prediction R (correlation between polygenic score and phenotype in the test set) with effect sizes from sib-GWAS, $\mathrm{R}_{\mathrm{sib}}$, and effect sizes from pop-GWAS (with sample size N^*), $\mathrm{R}_{\mathrm{unrel}}$, as a function of SNPs used to build polygenic scores 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.

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:

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

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:

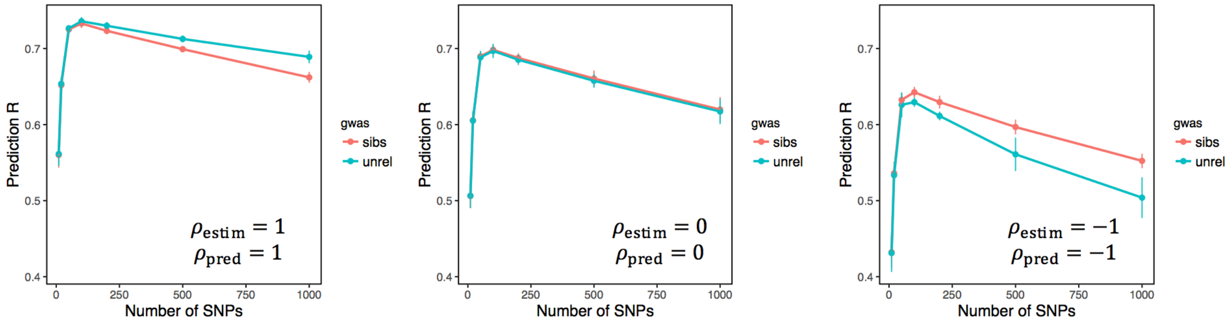
$$R \propto \frac{\sum_i \widehat{\beta_\iota}}{\sqrt{\sum_i \widehat{\beta_\iota}^2}} \qquad [6]$$

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 $\left.\frac{\sum_i \widehat{\beta_\iota}}{\sqrt{\sum_i \widehat{\beta_\iota}^2}}\right|_{ur} - \left.\frac{\sum_i \widehat{\beta_\iota}}{\sqrt{\sum_i \widehat{\beta_\iota}^2}}\right|_{s}$ 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 $p = 0.2$; $\mathrm{var}(e) = 2M\beta^2 p(1-p)$, 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. 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:

$$R \propto \frac{\sum_i \widehat{\beta_\iota}}{\sqrt{\sum_i \widehat{\beta_\iota}^2}} = \frac{\sum_i^{M_c} \widehat{\beta_\iota} + \sum_i^{M_{null}} \widehat{\beta_\iota}}{\sqrt{\sum_i^{M_c} \widehat{\beta_\iota}^2 + \sum_i^{M_{null}} \widehat{\beta_\iota}^2}}$$

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 $\boldsymbol{R}$ 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 $\boldsymbol{R}$ 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, $\boldsymbol{R}$ at the peak, etc.)

Conclusions 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.