REPORT

Disease and Polygenic Architecture: Avoid Trio Design and Appropriately Account for Unscreened Control Subjects for Common Disease

Wouter J. Peyrot,[1,2,*] Dorret I. Boomsma,[3] Brenda W.J.H. Penninx,[1] and Naomi R. Wray[2,*]

Genome-wide association studies (GWASs) are an optimal design for discovery of disease risk loci for diseases whose underlying genetic architecture includes many common causal loci of small effect (a polygenic architecture). We consider two designs that deserve careful consideration if the true underlying genetic architecture of the trait is polygenic: parent-offspring trios and unscreened control subjects. We assess these designs in terms of quantification of the total contribution of genome-wide genetic markers to disease risk (SNP heritability) and power to detect an associated risk allele. First, we show that trio designs should be avoided when: (1) the disease has a lifetime risk $> 1\%$; (2) trio probands are ascertained from families with more than one affected sibling under which scenario the SNP heritability can drop by more than 50% and power can drop as much as from 0.9 to 0.15 for a sample of 20,000 subjects; or (3) assortative mating occurs (spouse correlation of the underlying liability to the disorder), which decreases the SNP heritability but not the power to detect a single locus in the trio design. Some studies use unscreened rather than screened control subjects because these can be easier to collect; we show that the estimated SNP heritability should then be scaled by dividing by $(1 - K \times u)^2$ for disorders with population prevalence K and proportion of unscreened control subjects u. When omitting to scale appropriately, the SNP heritability of, for example, major depressive disorder ($K = 0.15$) would be underestimated by 28% when none of the control subjects are screened.

Optimal experimental design of genetic studies of disease for discovery of associated loci depends on the underlying genetic architecture of the trait. Although the true genetic architecture of the trait is usually not known, different experimental designs aim at exposing causal loci of differing population frequencies. For example, the optimal experimental design to detect de novo mutations is a trio design in which affected probands and their parents are genotyped.[1] In contrast, genome-wide association studies (GWASs) are an optimal design for a genetic architecture that includes many common causal loci of small effect (a polygenic architecture). Here, we consider two designs of GWASs, which we show deserve careful consideration: designs based on parent-offspring trios and designs based on unscreened control subjects. We assess these designs in terms of quantification of the total contribution to disease risk of genome-wide genetic markers, via estimation of so-called SNP heritability,[2] and the power to detect an associated risk allele.

Our study is motivated by experiences with GWAS designs for psychiatric disorders, but our results are parameterized based on baseline disease risk and heritability, and are, therefore, applicable to the full range of diseases and disorders with a polygenic genetic architecture of underlying risk. For psychiatric disorders, GWASs have had variable success in detecting genome-wide significant common SNPs. On the one hand, 108 significant loci were recently found for schizophrenia (SCZ [MIM: 181500]) in a sample comprising 36,989 case subjects,[3] whereas only two loci were found in one study on major depressive disorder (MDD [MIM: 608516])[4] but none in another,[5] no loci for attention-deficit/hyperactivity disorder (ADHD [MIM: 143465]),[6] and only single-study genome-wide significant loci for autism spectrum disorder (ASD [MIM: 209850]).[7–9] Sample size is pivotal in explaining this discrepancy, because much smaller numbers of cases were included for MDD (5,303 and 9,240, respectively), ADHD (2,960), and ASD (2,705, 1,984, and 1,553, respectively) than for SCZ. Other contributing factors have, nevertheless, been proposed, such as the impact of de novo mutations in ASD[10,11] (although these are expected to explain only a small proportion of variation),[12] lower family-based heritability of MDD (~0.4 versus ~0.8 for SCZ, ASD, and ADHD, assuming a similar genetic architecture between disorders),[13] and higher prevalence and greater heterogeneity of MDD.[14] Here, we show that the trio design, which is regularly applied in ASD and ADHD, and use of unscreened control subjects deserves careful consideration in the context of an underlying polygenic architecture, which is an important consideration for design of future studies that strive to increase sample size.[15]

The impact of trio design and the use of unscreened control subjects on the SNP heritability have, to the best of our knowledge, not yet been described, probably because the methods for estimation of SNP heritability were developed only in recent years.[16,17] The impact on the power to detect a single locus has, on the other hand, been studied in the pre-GWAS era of candidate genes,[18–21] but we could find no clear-cut comparison of the power to detect an associated risk allele with trio studies versus screened control studies, and we will therefore also give an overview of these differences. We investigate the trio design and the

[1]Department of Psychiatry, VU University Medical Center & GGZ inGeest, Amsterdam 1081 HL, the Netherlands; [2]Queensland Brain Institute, University of Queensland, Brisbane 4072, Australia; [3]Department of Biological Psychology, VU University Amsterdam, Amsterdam 1081 BT, the Netherlands
*Correspondence: w.peyrot@ggzingeest.nl (W.J.P.), naomi.wray@uq.edu.au (N.R.W.)
http://dx.doi.org/10.1016/j.ajhg.2015.12.017.

use of unscreened control subjects by analytical derivation followed by simulation studies to validate theory. Assortative mating (correlation in liability between spouses) is included in our trio design analyses, because this has been reported for a range of psychiatric disorders.[22–25] For example, a spouse correlation on the social responsiveness scale (a quantitative measure of autistic traits) of 0.29 has been reported in a population sample[23] and of 0.26 in parents of ASD probands.[22] For ADHD a spouse correlation of 0.11 on the ADHD index in population samples has been reported.[25] In trio designs, genotypes of proband cases are compared to genotypes of pseudocontrol subjects (the non-transmitted parental alleles).

SNP Heritability Calculations

The SNP heritability estimates the total proportion of variance tagged by common SNPs from a genome-wide association study.[2,16] If samples with GWAS data are population samples, then the variance estimated on the observed scale ($\hat{h}_o^2$) is expressed with the Robertson's transformation on the liability scale ($\hat{h}_l^2$) as[26]

$$\hat{h}_l^2 = \hat{h}_o^2 \frac{K(1-K)}{z^2}, \qquad \text{(Equation 1)}$$

where z denotes the height of the standard normal density function at the threshold corresponding to a baseline disease risk K. Quantification on the liability scale is most interpretable because it allows direct comparisons of estimates of heritability from family data that are reported on this scale to estimates of variance explained by individual genome-wide significant loci. However, usually GWAS samples are oversampled for case subjects compared to population samples and the transformation of proportion of variance attributable to SNPs estimated from case-control data ($\hat{h}_{occ}^2$) must also account for the proportion of cases in the sample P by[2,27]

$$\hat{h}_l^2 = \hat{h}_{occ}^2 \frac{K^2(1-K)^2}{P(1-P)z^2}, \qquad \text{(Equation 2)}$$

which reduces to Equation 1 when the sample is a population sample and $P = K$. However, these transformations assume that control subjects are screened. To account for control subjects being unscreened, we define F as the proportion of falsely classified control subjects, $F = N_{false\ controls}/(N_{false\ controls} + N_{true\ controls}) = N_{false\ controls}/N_{controls}$. We closely followed the derivations of Golan et al. (paragraphs 1.2 and 1.3 of their Supplemental Materials)[27] to derive an updated equation (Table S1) validated by simulation (Table S2),

$$\hat{h}_l^2 = \hat{h}_{occ}^2 \frac{K^2(1-K)^2}{P(1-P)(1-F)^2 z^2}, \qquad \text{(Equation 3)}$$

which reduces to Equation 2 when $F = 0$ and control subjects are screened. If a proportion u of the control subjects are a random sample from the population, then one can assume that $F \approx Ku$. Therefore, if it is unknown whether control subjects are screened or not, the potential underestimation when all control subjects are unscreened ($u = 1$) of the SNP heritability $\hat{h}_l^2$ estimated from the standard Equation 2 can be assessed as $\hat{h}_l^2(1-K)^2$ and thus depends on baseline risk K. In trio designs where probands are ascertained randomly, the pseudocontrol subjects are equivalent to unscreened control subjects under a polygenic model (Figure S1).

For the trio design, the SNP heritability was derived for a disease parameterized with normally distributed phenotypic (l) and genetic (G) liabilities with means $E(l) = E(G) = 0$ and variances $V_l = 1$ and $V_G = h_l^2$, the true heritability on the liability scale in the parental generation.[28] Under the liability-threshold model, individuals are deemed affected when their liability (l) is larger than threshold (T) such that $P(l > T \mid l \sim N(0,1)) = K$. Parental assortative mating was taken into account by parameterizing a spouse liability correlation of ρ_l and genetic correlation of $\rho_G = h_l^2 \rho_l$.[28] The $E(G)$ of proband case subjects and pseudocontrol subjects were derived by considering the variance-covariance matrix of l and G of individuals that could contribute to a trio design (proband, sibling, mother, father, pseudocontrol). To account for the affected proband, the variance-covariance matrix of random families was conditioned on the proband being affected by accounting for the reduction in variance as a result of the Bulmer effect[29] in related individuals described by Tallis.[30] To account for a second affected sibling, the variance-covariance matrix was further conditioned on the sibling also being affected. Details of these derivations are provided in the Supplemental Methods and were validated with a simulation study in R (Tables S3 and S4).

Figure 1A displays the SNP heritability assessed from unscreened control subjects (Figure 1A, dashed line), which is equivalent to estimates from pseudocontrol subjects from random families with at least one affected proband (Figure 1A, dotted line) and screened control subjects (Figure 1, solid lines). Although the standard transformation (Equation 2) applied to derive estimates of SNP heritability on the liability scale ($\hat{h}_l^2$) is expected to give unbiased estimates of the true SNP heritability when case subjects are randomly ascertained and control subjects are screened (Figure 1A, solid line), the transformation underestimates h_l^2 by a factor $(1-K)^2$ when diseases are common (high K) and control subjects are unscreened or are pseudocontrol subjects (Figure 1A, dashed and dotted line). The estimated heritability from the Equation 2 transformation $\hat{h}_l^2$ severely underestimates h_l^2 when data result from a trio design with probands ascertained from multiplex families (Figure 1B, dotted line), for example, $\hat{h}_l^2 = 0.31$ for $K = 0.05$ and $h_l^2 = 0.5$, because the mean liability of pseudocontrol subjects is greater than the average in the population and so the contrast in genetic values between case subjects and pseudocontrol subjects is less than between case subjects and screened control subjects (Table 1, additional

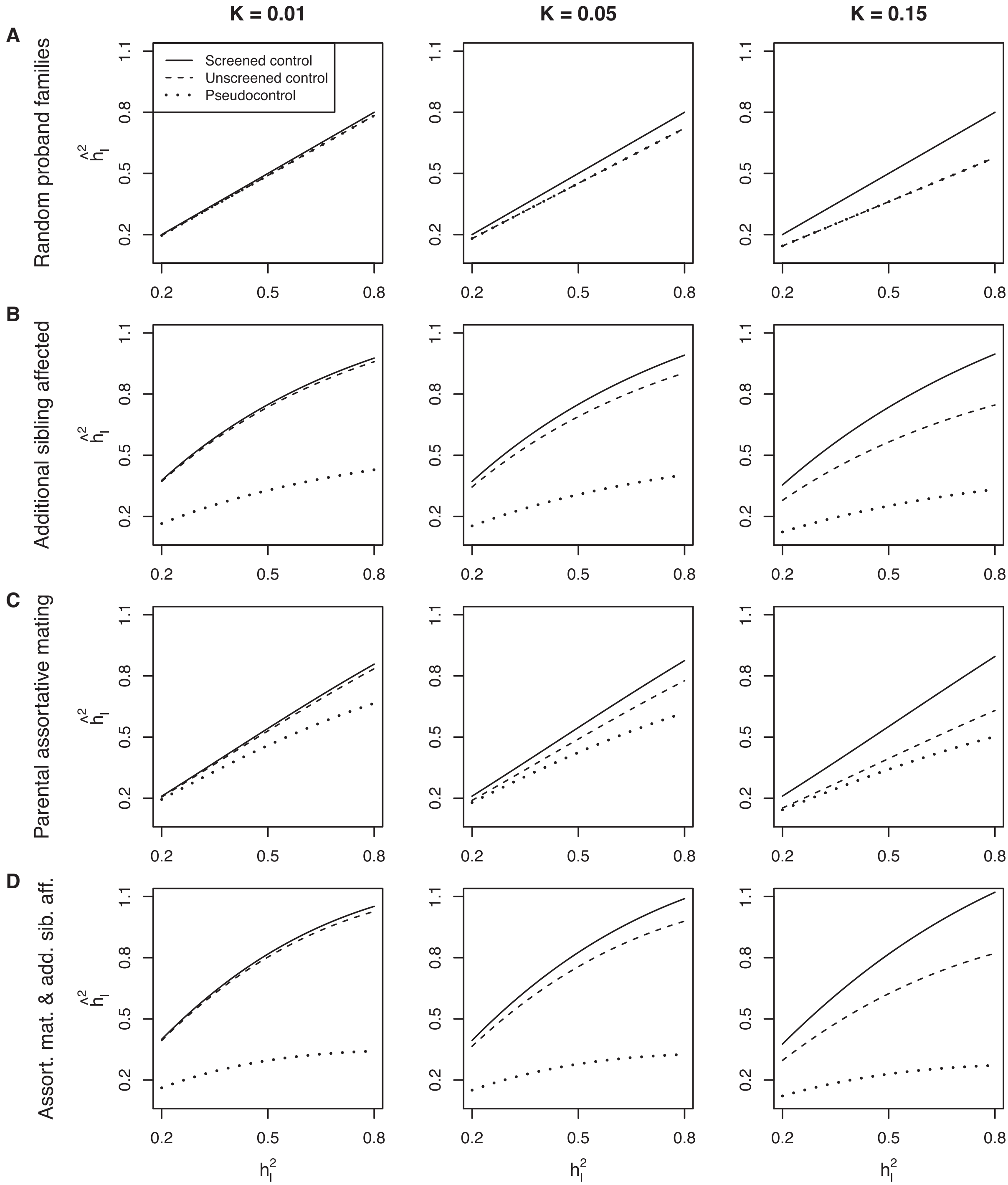


Figure 1. Relationship between the True SNP Heritability and Its Estimates Based on the Standard Transformation with Equation 2 from Trio Data, Screened Controls, and Unscreened Controls

The SNP heritability $\widehat{h_l^2}$ that would be estimated based on the standard liability transformation equation (Equation 2) for GWASs using pseudocontrol subjects (dotted lines), unscreened control subjects (dashed lines), and screened control subjects (solid lines) compared to the true parental SNP heritability h_l^2 for designs based on randomly ascertained proband families (A), families with an additional affected sibling (B), in the context of parental assortative mating with a correlation on the liability scale of $\rho_l = 0.3$ (C), and families with an additional affected sibling in the context of parental assortative mating (D) for disorders with lifetime risk $K = 0.01$, 0.05, and 0.15. The pseudocontrol subjects of random proband families are equivalent to unscreened control subjects (dashed and dotted lines overlap in A), and the slope of these lines are defined by $(1 - K)^2$, i.e., the underestimation of $\widehat{h_l^2}$ when mistakenly applying Equation 2 rather than Equation 3 to transform the heritability on the observed scale to the liability scale when none of the control subjects are screened.

Table 1. **Mean Genetic Liabilities and SNP Heritability Estimated from the Standard Transformation with Equation 2 from GWAS using Trio Design, Screened Control Subjects, or Unscreened Control Subjects for Actual Parental Heritability 0.5**

		Mean Genetic Liability ($E(G)$)				$\widehat{h}_l^2$ Assessed from Proband		
			Control					
K	h_l^2 Parents	Case	Screened	Unscreened	Pseudo	Screened	Unscreened	Pseudo
Random Proband Families								
0.01	0.5	1.333	−0.013	0.000	0.000	0.500	0.490	0.490
0.05	0.5	1.031	−0.054	0.000	0.000	0.500	0.451	0.451
0.15	0.5	0.777	−0.137	0.000	0.000	0.500	0.361	0.361
Additional Sibling Affected								
0.01	0.5	1.634	−0.013	0.000	0.543	0.749	0.736	0.328
0.05	0.5	1.275	−0.054	0.000	0.424	0.750	0.690	0.307
0.15	0.5	0.972	−0.137	0.000	0.323	0.735	0.565	0.251
Parental Assortative Mating								
0.01	0.5	1.386	−0.016	0.000	0.097	0.542	0.530	0.459
0.05	0.5	1.075	−0.060	0.000	0.075	0.547	0.490	0.424
0.15	0.5	0.812	−0.148	0.000	0.057	0.552	0.395	0.341
Additional Sibling Affected and Parental Assortative Mating								
0.01	0.5	1.706	−0.016	0.000	0.670	0.818	0.803	0.296
0.05	0.5	1.335	−0.060	0.000	0.525	0.826	0.756	0.278
0.15	0.5	1.021	−0.148	0.000	0.402	0.818	0.624	0.230

The mean genetic liabilities $E(G)$ are displayed for probands, unrelated screened control subjects, unrelated unscreened control subjects, and their pseudocontrol subjects as well as the SNP heritability $\widehat{h}_l^2$ estimated from Equation 2 from comparing case subjects to these three sets of control subjects, for different parameterization of baseline disease risk K and a fixed underlying heritability of $h_l^2 = 0.5$. The probands are parameterized in line with Figure 1 to be selected from random proband families (Figure 1A), families with an additional affected sibling (Figure 1B), families in the context of parental assortative mating (Figure 1C), and families with an additional affected sibling in the context of assortative mating (Figure 1D), respectively.

sibling affected), which is not fully compensated by the fact that case subjects from multiplex families have higher mean liability than randomly selected case subjects (Table 1, random proband families). In contrast, when case subjects are selected from multiplex families and control subjects are screened, the estimated SNP heritability based on the standard transformation is an overestimate of h_l^2 (for example, $\widehat{h}_l^2 = 0.75$ for $K = 0.05$ and $h_l^2 = 0.5$). When control subjects are unscreened, the SNP heritability is found between the SNP heritabilities from screened and pseudocontrol subjects (Figure 1, dashed lines), when SNP heritabilities are estimated by Equation 2. In the context of assortative mating, a trio design comparison of probands to pseudocontrol subjects yields decreased $\widehat{h}_l^2$ (Figure 1C; Table 1, parental assortative mating; spouse correlation $\rho_l = 0.3$). Again, comparing the probands to screened control subjects (from the offspring generation) does in fact overestimate the heritability in the parent generation h_l^2; this is, however, a well-known consequence of assortative mating and is not restricted to the trio design ($V_{G,offspring} = V_{G,parents} + (1/2)\rho_{G,parents} V_{G,parents}$).[29] The most pronounced difference between screened and pseudocontrol control subjects is found for probands with an additional affected sibling in the context of parental assortative mating (Figure 1D; Table 1, additional sibling affected and parental assortative mating).

Power Calculations

The power to detect an associated risk allele in a case-control association test follows from the non-centrality parameter NCP of the χ^2 test statistic. This NCP is expressed in terms of sample size N, proportion of case subjects in the study v, the allele frequency in case subjects p_{case}, the allele frequency in control subjects $p_{control}$, and the mean allele frequency in the sample $\overline{p} = vp_{case} + (1-v)p_{control}$ as

$$NCP = \frac{(p_{case} - p_{control})^2}{\overline{p}(1-\overline{p})\left(\frac{1}{2N*v} + \frac{1}{2N*(1-v)}\right)} \quad \text{(Equation 4)}$$

and the power as $P(\, x > \sqrt{NCP} + x_T \mid z \sim N(0,1))$, where x_T is the z -value quantile-function of the standard normal distribution for the desired significance threshold, here set at $\alpha = 5 \times 10^{-8}$ ($x_T = -5.45$). The power of different experimental designs is reflected in the appropriate expressions of p_{case} and $p_{control}$. We parameterize a disease with a baseline lifetime disease risk K, a di-allelic locus with risk allele frequency $P(B) = p$, non-risk allele frequency $P(b) = q = 1 - p$,

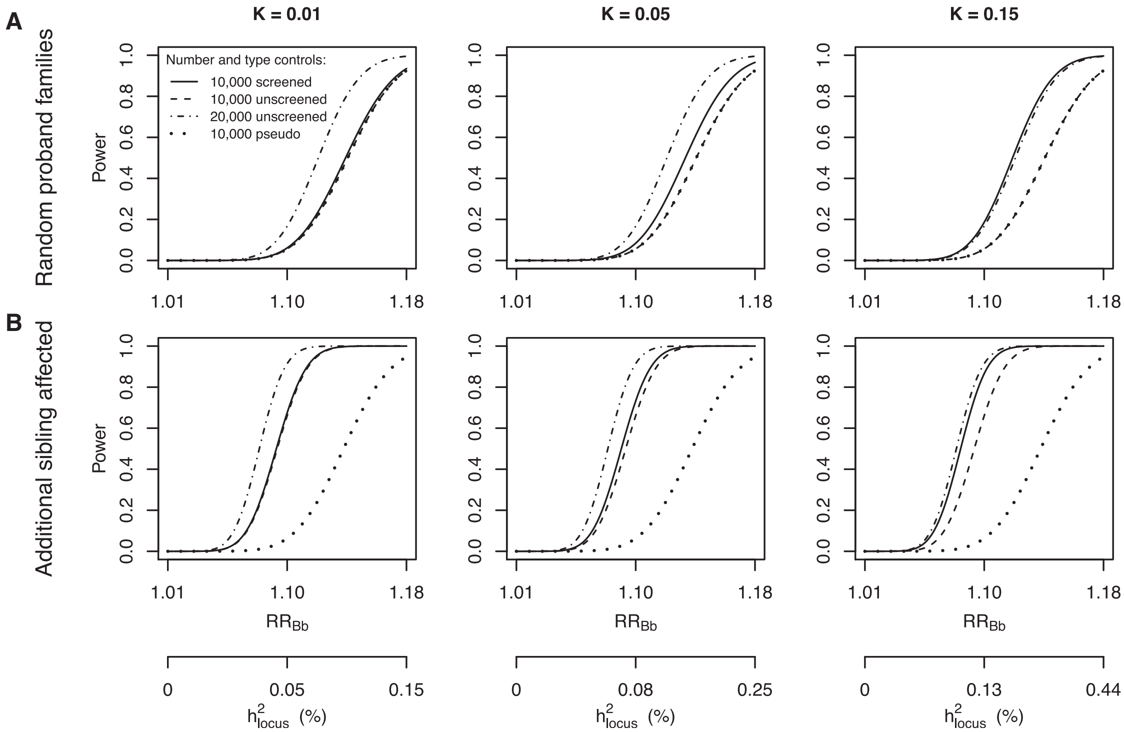


Figure 2. Power to Detect a Single Risk Variant in Association Studies of 10,000 Case Subjects that Use a Trio Design, Screened Control Subjects, or Unscreened Control Subjects

Power of association analysis comparing 10,000 probands to 10,000 screened control subjects (solid line), 10,000 unscreened control subjects (dashed), 20,000 unscreened control subjects (dot-dashed), and 10,000 pseudocontrol subjects (dotted) to detect a single associated risk variant for a risk allele with frequency $p = 0.2$, for a baseline disease risk $K = 0.01$, 0.05, and 0.15. Power was estimated for risk variants with underlying additive effect ($RR_{BB} = RR_{Bb}^2$) for random ascertainment of probands (A) and probands from families with an additional affected sibling (B). Note that pseudocontrol subjects from random families are equivalent to unscreened control subjects and that the dotted and dashed lines in (A) overlap. The variation explained on the liability scale was approximated by $h^2_{locus} \approx 2p(1-p)(RR_{Bb}-1)^2/i^2$, where i equals z/K the mean liability of probands, and z the height of the standard normal density function at the threshold corresponding with disease of lifetime risk K.

relative risk of heterozygotes $RR_{Bb} = P(Disease \mid Bb)/P(Disease \mid bb)$, and relative risk of the homozygotes $RR_{BB} = P(Disease \mid BB)/P(Disease \mid bb)$.[31,32] When control subjects are screened, power follows from $p_{case} = k_{bb}RR_{Bb}p(1 + p(RR_{Bb} - 1))/K$, where $k_{bb} = P(Disease \mid \mathrm{bb}) = K/(q^2 + 2pqRR_{Bb} + p^2RR_{BB})$ and $p_{control} = ((1 - k_{bb}RR_{Bb})p(1 - p) + (1 - k_{bb}RR_{BB})p^2)/(1 - K)$,[32] which agrees with the genetic power calculator of Purcell et al.[33] When control subjects are unscreened, the power of an association study is expressed by Equation 4 with $p_{control} = p$. For the trio design, power was assessed by substituting in Equation 4 the allele frequency in probands and pseudocontrol subjects (the non-transmitted alleles of the parents). When trios are ascertained from families with an additional affected sibling or when there is assortative mating, the risk allele frequency in control subjects can be derived from combined and conditional genotype frequencies of an individual, the affected sibling, and the parents. Under assortative mating, expressions are dependent on spouse liability correlation $\rho_{liability}$, which results in the correlation between the parental genotypes as $\rho_{locus} = \rho_{liability} h^2_{locus}$.[28] It follows that assortative mating (for example, $\rho_{liability} = 0.3$) has no impact on the power to detect a single locus for loci typical of polygenic architecture that explain less than 1% of variation ($\rho_{locus} = 0.3 \times 0.01 = 0.003$).[28] When assuming a small RR_{Bb} typical of complex genetic disease and a multiplicative model on the disease scale ($RR_{BB} = RR_{Bb}^2$, implying additively on the underlying risk scale), the variance attributable to the risk locus can be approximated by $h^2_{locus} \approx 2pq(RR_{Bb} - 1)^2/i^2$ with $i = z/K$ the mean liability of case subjects and z the height of the standard normal density function at the threshold corresponding to a baseline disease risk K.[32] The expressions to derive allele frequencies in trios are closed but complex (Supplemental Methods) and were validated by simulation (Table S5).

Figure 2 displays the power to detect an associated risk allele for probands from random trios with an affected proband (Figure 2A) and multiplex trios with an additional affected sibling (Figure 2B), when the risk allele has a frequency of $P(B) = p = 0.2$ for disorders with baseline risk $K = 0.01$, 0.05, and 0.15 in a sample of n = 10,000 trios (probands versus pseudcontrol subjects) against RR_{Bb} given an underlying additive effect ($RR_{BB} = RR_{Bb}^2$) (dotted line). Note that pseudocontrol subjects from random families are equivalent to unscreened control subjects (Figure S1), which are displayed in Figure 2 for 10,000 unscreened control subjects (dashed line) and 20,000 unscreened control subjects (dot-dashed line) compared to 10,000 probands.

Table 2. **Maximum Power Difference between Trio Design and Screened Control Subject Studies with 20,000 Subjects**

		Allele Frequencies			Power (n = 20,000)		n (Power = 0.8)	
K	RR_{Bb}	Proband	Pseudo	Screened	Pseudo	Screened	Pseudo	Screened
Proband from Random Proband Families								
0.01	1.147	0.223	0.200	0.200	0.56	0.58	25,226	24,714
0.05	1.144	0.222	0.200	0.199	0.51	0.63	26,327	23,712
0.15	1.135	0.221	0.200	0.196	0.39	0.74	29,670	21,297
Proband from Families with an Additional Affected Sibling								
0.01	1.115	0.228	0.209	0.200	0.17	0.91	39,201	17,307
0.05	1.113	0.227	0.209	0.199	0.15	0.92	40,533	16,923
0.15	1.108	0.226	0.208	0.197	0.11	0.94	44,574	15,945

The loci with allele frequency $p = 0.2$ from Figure 2 that result in most pronounced decrease in power for pseudocontrol compared to screened control studies for a sample of 10,000 case subjects and 10,000 control subjects are displayed in detail. The power difference depends on the baseline disease risk K, its effect size RR_{Bb}, and whether the probands are from random proband families or families with an additional affected sibling (compare to solid and dotted lines, respectively, in Figure 2). For these loci, the allele frequencies in probands, pseudocontrol subjects, and screened control subjects is displayed, as well as the power given a sample size of n = 20,000 (50% case subjects) and the required sample size to obtain a power of 0.8. Note that pseudocontrol subjects from random families are equivalent to unscreened population control subjects.

The solid line on each graph is the power for 10,000 probands compared to 10,000 unrelated screened control subjects. Figure 2A shows that there is little to be gained in screening control subjects for diseases of lifetime morbid risk < 1%, but for more common disorders (such as ADHD and MDD), there is an important gain in power, which can also be gained by increasing the number of unscreened control subjects. When trios come from families with an additional affected sibling, the case subjects have an increased probability of carrying the risk allele and so when matched with screened control subjects, there is a gain in power compared to random ascertainment of case subjects (solid line in Figure 2B versus solid line in Figure 2A). For example, when $p = 0.2$, $RR_{Bb} = 1.2$, then $p_{proband\ B} = 0.248$ and $p_{proband\ A} = 0.231$, respectively (these frequencies do not depend on K). However, when the association study is of case subjects from multiplex families compared to pseudocontrol subjects, there is little gain in power compared to trios based on randomly selected case subjects (dotted line in Figure 2B versus dotted line in Figure 2A), because the pseudocontrol subjects also have increased probability of carrying the risk allele ($p_{pseudocontrol\ B} = 0.215$ and $p_{pseudocontrol\ A} = 0.2$). The maximum power difference between using screened and pseudocontrol subjects depends on RR_{Bb}, K, sample size, and whether probands are ascertained randomly or from families with an additional affected sibling (Table 2), but is found for a sample comprising 20,000 subjects at $RR_{Bb} = 1.11$ and $K = 0.15$ for probands with additional affected siblings, under which scenario a total sample size of n = 15,945 is needed when control subjects are screened versus n = 44,574 for the pseudocontrol trio design, respectively, to obtain a power of 0.8. For unscreened control subjects (equivalent to pseudocontrol subjects from random families), the most pronounced decrease in power in a sample of 20,000 subjects is found for a locus with $RR_{Bb} = 1.14$ in disease with $K = 0.15$ where unscreened control subjects yield a power of 0.39 and screened control subjects a power of 0.74. As expected, the impact of using screened control subjects is higher for more common disorders. Allele frequencies in probands, pseudocontrol subjects, and screened control subjects for all Figure 2 scenarios are presented in Figure S2. Furthermore, the power differences between pseudocontrol and screened control studies are consistent for other risk allele frequencies, e.g., $p = 0.6$ (Figure S3) underlying actual recessive ($RR_{Bb} = 1$; Figure S4) and dominant ($RR_{Bb} = RR_{BB}$; Figure S5) effects. In addition, to select only trios with unaffected parents has no impact on power of pseudocontrol studies, because although the risk allele frequency in pseudocontrol subjects decreases, the frequency in case subjects decreases proportionally (Figure S6). When unscreened control subjects are much easier to obtain then screened control subjects, the loss of power due to not screening can be balanced by increasing the number of unscreened control subjects, which is illustrated for different numbers of unscreened control subjects in Figure S7. Note that Equation 4 defines a limit to the power-gain from increasing the number of unscreened control subjects, but that when increasing number of unscreened control subjects from 10,000 to 20,000, the loss of power due to not screening is balanced for all scenarios under consideration here. In Figure 2, the additional x axis is variance explained by the locus, and therefore the results generalize to many combinations of p and RR_{Bb} that together explain the same locus variance.[31] Although association studies have similar power to detect a locus based on RR_{Bb} regardless of baseline disease risk K, the variance explained by a locus is much larger for high K. Therefore, to detect a risk allele that explains the same proportion of genetic variance, a much larger sample size is needed for larger K (Figure 3).

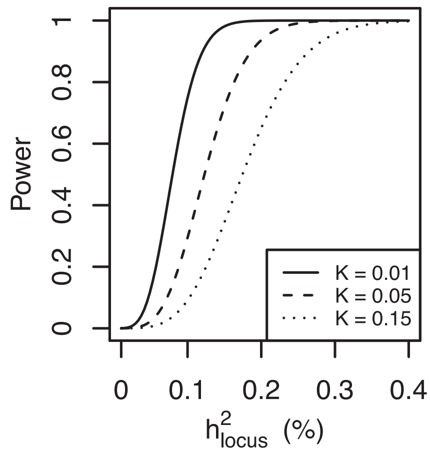


Figure 3. Power to Detect an Associated Locus by the Proportion of Variation It Explains
The power to detect an associated locus depends on the proportion of variation it explains on the liability scale h^2_{locus}, the baseline disease risk K, and is displayed for random case versus screened control. For a locus with the same h^2_{locus}, larger sample sizes are required for larger K. h^2_{locus} can be approximated by $2p(1-p)(RR_{Bb}-1)^2/i^2$, where i equals z/K the mean liability of probands, and z the height of the standard normal density function at the threshold corresponding with disease of lifetime risk K. The (complex) relation between allele frequency p, RR_{Bb}, and the non-centrality parameter NCP given h^2_{locus} results in an identical relation between power and h^2_{locus} for varying p.

To summarize our findings, our results generate two important conclusions that trio-based samples and unscreened control subjects for common diseases deserve careful consideration when the underlying genetic architecture is highly polygenic. We have quantified this in two ways, first by the underestimation of SNP heritability through application of the inappropriate transformation equation, and second by power calculations of association analysis. We derived a transformation equation for the SNP heritability that is appropriate for unscreened control samples (Equation 3).

The use of trio designs most commonly occurs for pediatric diseases and disorders in which it is relatively easy to obtain blood samples from parents. Trio designs are needed to detect de novo causal mutations,[34] to determine accurately phased haplotypes,[34] and to undertake parent-of-origin analyses implied by a hypothesis of parental imprinting.[35] Trio designs have also been considered for detection of gene-environment interaction.[36,37] In the pre-GWAS era, trio designs were recommended to protect against potential bias from population stratification,[1] and although this quality is also sometimes promoted for trio GWAS, with genome-wide SNP data, other strategies, such as genomic principal components[38] or mixed model association analysis,[39] appropriately account for population stratification without the need to incur 50% higher costs by genotyping three samples to generate two genomes. While acknowledging the benefits of parent-offspring trios under some experimental paradigms, trio-design GWASs have been undertaken without full regard of the implications to power under the genetic architecture implicated by the GWAS paradigm. We draw the following conclusions.

(1) If the case probands of trios are ascertained randomly, then the resulting case-pseudocontrol study is equivalent to a case-unscreened control design under a polygenic genetic architecture and has little impact on the SNP heritability and power for disorders that are less common, but for more common disorders there is important decrease in SNP heritability (Figure 1A) and loss of power (Figure 2A), inadvertently contributing to the missing heritability problem. For example, in a study on MDD (lifetime risk $K \sim 0.15$)[13,40] where all control subjects are unscreened, the SNP heritability (say 0.3) would reduce by a factor of 0.72 ($0.72 \times 0.3 = 0.22$) (hence underestimated by 28%) when not accounting for the unscreened control subjects (i.e., applying Equation 2 rather than Equation 3). For disorders such as MDD, even when control subjects have been screened, it is likely that some control subjects remain misclassified, because onset can occur throughout the lifetime. Naturally, it should also be noted that when super-control subjects are used (control subjects screened to be at the lower end of the liability distribution, for example based on low scores for the personality trait neuroticism in the context of MDD), SNP heritability estimates based on the standard transformation equation would be biased upward. The loss of power due to including unscreened control subjects can be compensated by increasing the number of control subjects (Figures 2 and S7), in particular in the context of the continuously decreasing costs for genotyping, but this requires caution when estimating the SNP heritability, because Equation 3 should then be applied rather than the standard Equation 2.

(2) If case probands are ascertained from multiplex families, then the SNP heritability and power of GWASs are substantially reduced when using pseudocontrol subjects even for less-common disorders (see Figures 1B and 2B, respectively; modeled on families with two affected siblings). Even in the absence of deliberate ascertainment of multiplex families, studies are likely to be biased by self-ascertainment because parents from multiplex families might be more concerned with the genetic origins of the disorder. In fact, 43.6% of the 1,369 families included in the Autism Genome Project (AGP) had two or more children affected with ASD while counting up to third-degree relatives.[7] However, the proportion of multiplex families is often not reported, as is the case for the family-based studies,[41–43] contributing to the most recent ADHD meta-analysis,[6] which leaves the loss in power due to included multiplex families unknown,

but likely. In addition, in a number of families with a first affected child, parents will stop having children, so that a second affected child is never observed. Our results are consistent with the simplex versus multiplex and simulation results of Klei et al. in analyses of ASD samples.[44]

(3) Assortative mating considerably decreases the SNP heritability assessed from trio design compared to screened control subjects also for small K (Figure 1C), but it does not impact the power to detect a single locus under a polygenic model, because of the small proportions of variation explained by single loci (<1%). Assortative mating is possibly common for psychiatric disorders[22–25] and needs to be considered when interpreting SNP heritability in general and for trio design in particular. These results and point (2) could explain why lower SNP-based heritabilities were found in the ADHD pseudocontrol samples from the Psychiatric Genomics Consortium compared to case-control samples (see Table S5 of Lee et al.).[14]

We also take the opportunity to re-emphasize that parameterization of power in terms of genotype relative risk can be misleading because the same RR_{Bb} operating in common disease implies a much higher proportion of variance explained by the locus compared to a locus operating in a less common disease. For example, when the risk allele has frequency $p = 0.2$ and effect size $RR_{Bb} = 1.1$, the locus explains 0.05%, 0.08%, and 0.13% of the variance in disease liability for a disorder of frequency $K = 0.01$, 0.05, and 0.15, respectively. Hence, to detect a locus that explains the same proportion of variance in liability, much larger samples are needed for common disorders (Figure 3). For example, samples of n = 4,059 (50% case subjects, 50% screened control subjects) are needed to detect a locus that explains 0.5% of the variance in liability for a disorder lifetime risk $K = 0.01$ ($RR_{Bb} = 1.39$), compared to samples of n = 9,181 when the disorder risk is $K = 0.15$ ($RR_{Bb} = 1.21$). Similar arguments have been used to explain that much larger GWAS samples are needed for MDD compared to schizophrenia.[45]

To the best of our knowledge, the impact of the trio design and use of unscreened control subjects on the SNP heritability has not yet been addressed, but our power analyses build upon a rich literature exploring the characteristics of family-based association studies in the pre-GWAS era. Ferreira et al. showed that the trio-based transmission disequilibrium test (TDT) has less power when an additional (non-genotyped) sibling is affected compared to random families with one affected sibling.[18] Li et al.,[19] Risch and Teng,[46] and Risch[47] showed that case-control studies are generally more powerful when case subjects are from families with an additional affected sibling, which is in line with our results (Figure 2B compared to Figure 2A). Teng and Risch found that family-based approaches have less power than case-unrelated control strategies for families with multiple affected siblings.[20] Of note, our paper focuses on the pseudocontrol trio design, because this is how the trio design is typically applied in GWASs; however, the TDT has often been applied for candidate genes and could yield more power for rare disorders as has been indicated by Laird et al.[21] The power to detect a locus with the use of unscreened control subjects can readily be calculated with the online power calculator of Purcell et al.[33] or the Quanto software from Gauderman.[48] Nevertheless, our study adds also to the current literature on the power to detect a single locus, because we directly compare pseudocontrol studies to screened control studies for multiplex families and assortative mating. As expected, there is overall similarity between consequences of design for the power to detect a single risk variant and expected SNP heritability, but in this study we have formalized these expectations and also shown that such similarity does not hold when considering assortative mating that impacts the estimated SNP heritability but not in power to detect a single risk variant.

To conclude, we advise against the use of trio designs for disorders with a polygenic genetic architecture, such as psychiatric disorders, and we advise careful consideration when using unscreened control subjects for prevalent disorders, because these designs can result in an underestimated SNP heritability and decreased power to detect an associated risk allele.

Supplemental Data

Supplemental Data include seven figures, five tables, and Supplemental Methods and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.12.017.

Acknowledgments

We thank P.M. Visscher and G.B. Chen for the useful discussions about the derivation of Equation 3. This work was supported by an EMGO+ Travel Grant (number 2001228) to W.J.P. and by the National Health and Medical Research Council grants and fellowship (1050218, 613602, 1078901) to N.R.W. In addition, B.W.J.H.P. received research funding from Jansen Research.

Received: September 17, 2015
Accepted: December 17, 2015
Published: February 4, 2016

Web Resources

The URLs for data presented herein are as follows:

OMIM, http://www.omim.org/
R statistical software, http://www.r-project.org/

References

1. Spielman, R.S., and Ewens, W.J. (1996). The TDT and other family-based tests for linkage disequilibrium and association. Am. J. Hum. Genet. *59*, 983–989.

2. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.
3. Ripke, S., Neale, B.M., Corvin, A., Walters, J.T.R., Farh, K.-H., Holmans, P.A., Lee, P., Bulik-Sullivan, B., Collier, D.A., Huang, H., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
4. Cai, N., Bigdeli, T.B., Kretzschmar, W., Li, Y., Liang, J., Song, L., Hu, J., Li, Q., Jin, W., Hu, Z., et al.; CONVERGE consortium (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature *523*, 588–591.
5. Ripke, S., Wray, N.R., Lewis, C.M., Hamilton, S.P., Weissman, M.M., Breen, G., Byrne, E.M., Blackwood, D.H.R., Boomsma, D.I., Cichon, S., et al.; Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium (2013). A mega-analysis of genome-wide association studies for major depressive disorder. Mol. Psychiatry *18*, 497–511.
6. Neale, B.M., Medland, S.E., Ripke, S., Asherson, P., Franke, B., Lesch, K.-P., Faraone, S.V., Nguyen, T.T., Schäfer, H., Holmans, P., et al.; Psychiatric GWAS Consortium: ADHD Subgroup (2010). Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. J. Am. Acad. Child Adolesc. Psychiatry *49*, 884–897.
7. Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., Sykes, N., Pagnamenta, A.T., et al. (2010). A genome-wide scan for common alleles affecting risk for autism. Hum. Mol. Genet. *19*, 4072–4082.
8. Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J.T., Abrahams, B.S., Salyakina, D., Imielinski, M., Bradfield, J.P., Sleiman, P.M.A., et al. (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature *459*, 528–533.
9. Weiss, L.A., Arking, D.E., Daly, M.J., and Chakravarti, A.; Gene Discovery Project of Johns Hopkins & the Autism Consortium (2009). A genome-wide linkage and association scan reveals novel loci for autism. Nature *461*, 802–808.
10. Ronemus, M., Iossifov, I., Levy, D., and Wigler, M. (2014). The role of de novo mutations in the genetics of autism spectrum disorders. Nat. Rev. Genet. *15*, 133–141.
11. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.
12. Gratten, J., Visscher, P.M., Mowry, B.J., and Wray, N.R. (2013). Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. Nat. Genet. *45*, 234–238.
13. Sullivan, P.F., Daly, M.J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat. Rev. Genet. *13*, 537–551.
14. Lee, S.H., Ripke, S., Neale, B.M., Faraone, S.V., Purcell, S.M., Perlis, R.H., Mowry, B.J., Thapar, A., Goddard, M.E., Witte, J.S., et al.; Cross-Disorder Group of the Psychiatric Genomics Consortium; International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat. Genet. *45*, 984–994.
15. Sullivan, P.F. (2010). The psychiatric GWAS consortium: big science comes to psychiatry. Neuron *68*, 182–186.
16. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
17. Chen, G.-B. (2014). Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. Front. Genet. *5*, 107.
18. Ferreira, M.A.R., Sham, P., Daly, M.J., and Purcell, S. (2007). Ascertainment through family history of disease often decreases the power of family-based association studies. Behav. Genet. *37*, 631–636.
19. Li, M., Boehnke, M., and Abecasis, G.R. (2006). Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. Am. J. Hum. Genet. *78*, 778–792.
20. Teng, J., and Risch, N. (1999). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. Genome Res. *9*, 234–241.
21. Laird, N.M., and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. Nat. Rev. Genet. *7*, 385–394.
22. Virkud, Y.V., Todd, R.D., Abbacchi, A.M., Zhang, Y., and Constantino, J.N. (2009). Familial aggregation of quantitative autistic traits in multiplex versus simplex autism. Am. J. Med. Genet. B. Neuropsychiatr. Genet. *150B*, 328–334.
23. Constantino, J.N., and Todd, R.D. (2005). Intergenerational transmission of subthreshold autistic traits in the general population. Biol. Psychiatry *57*, 655–660.
24. Lichtenstein, P., Björk, C., Hultman, C.M., Scolnick, E., Sklar, P., and Sullivan, P.F. (2006). Recurrence risks for schizophrenia in a Swedish national cohort. Psychol. Med. *36*, 1417–1425.
25. Boomsma, D.I., Saviouk, V., Hottenga, J.-J., Distel, M.A., de Moor, M.H.M., Vink, J.M., Geels, L.M., van Beek, J.H.D.A., Bartels, M., de Geus, E.J.C., and Willemsen, G. (2010). Genetic epidemiology of attention deficit hyperactivity disorder (ADHD index) in adults. PLoS ONE *5*, e10621.
26. Dempster, E.R., and Lerner, I.M. (1950). Heritability of threshold characters. Genetics *35*, 212–236.
27. Golan, D., Lander, E.S., and Rosset, S. (2014). Measuring missing heritability: inferring the contribution of common variants. Proc. Natl. Acad. Sci. USA *111*, E5272–E5281.
28. Falconer, D., and Mackay, T. (1996). Introduction to Quantitative Genetics (Essex: Longman).
29. Bulmer, M. (1985). The Mathematical Theory of Quantitative Genetics (Oxford: Clarendon press).
30. Tallis, G.M. (1987). Ancestral covariance and the Bulmer effect. Theor. Appl. Genet. *73*, 815–820.
31. Yang, J., Wray, N.R., and Visscher, P.M. (2010). Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. Genet. Epidemiol. *34*, 254–257.
32. Witte, J.S., Visscher, P.M., and Wray, N.R. (2014). The contribution of genetic variants to disease depends on the ruler. Nat. Rev. Genet. *15*, 765–776.
33. Purcell, S., Cherny, S.S., and Sham, P.C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics *19*, 149–150.
34. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A.; 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

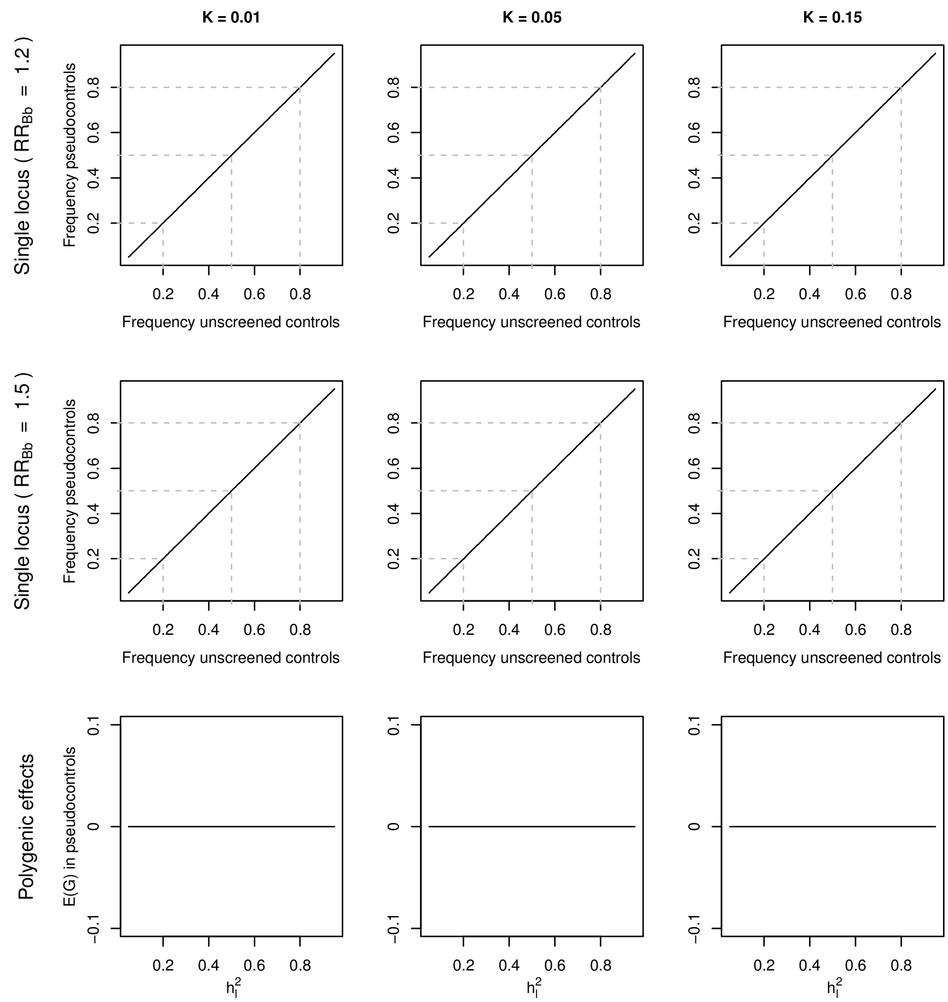
35. Solter, D. (1988). Differential imprinting and expression of maternal and paternal genomes. Annu. Rev. Genet. *22*, 127–146.
36. Cordell, H.J. (2009). Estimation and testing of gene-environment interactions in family-based association studies. Genomics *93*, 5–9.
37. Gauderman, W.J., Thomas, D.C., Murcray, C.E., Conti, D., Li, D., and Lewinger, J.P. (2010). Efficient genome-wide association testing of gene-environment interaction in case-parent trios. Am. J. Epidemiol. *172*, 116–122.
38. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
39. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. *46*, 100–106.
40. de Graaf, R., ten Have, M., van Gool, C., and van Dorsselaer, S. (2012). Prevalence of mental disorders and trends from 1996 to 2009. Results from the Netherlands Mental Health Survey and Incidence Study-2. Soc. Psychiatry Psychiatr. Epidemiol. *47*, 203–213.
41. Neale, B.M., Lasky-Su, J., Anney, R., Franke, B., Zhou, K., Maller, J.B., Vasquez, A.A., Asherson, P., Chen, W., Banaschewski, T., et al. (2008). Genome-wide association scan of attention deficit hyperactivity disorder. Am. J. Med. Genet. B. Neuropsychiatr. Genet. *147B*, 1337–1344.
42. Mick, E., Todorov, A., Smalley, S., Hu, X., Loo, S., Todd, R.D., Biederman, J., Byrne, D., Dechairo, B., Guiney, A., et al. (2010). Family-based genome-wide association scan of attention-deficit/hyperactivity disorder. J. Am. Acad. Child Adolesc. Psychiatry *49*, 898–905. e3.
43. Elia, J., Gai, X., Xie, H.M., Perin, J.C., Geiger, E., Glessner, J.T., D'arcy, M., deBerardinis, R., Frackelton, E., Kim, C., et al. (2010). Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. Mol. Psychiatry *15*, 637–646.
44. Klei, L., Sanders, S.J., Murtha, M.T., Hus, V., Lowe, J.K., Willsey, A.J., Moreno-De-Luca, D., Yu, T.W., Fombonne, E., Geschwind, D., et al. (2012). Common genetic variants, acting additively, are a major source of risk for autism. Mol. Autism *3*, 9.
45. Wray, N.R., Pergadia, M.L., Blackwood, D.H.R., Penninx, B.W.J.H., Gordon, S.D., Nyholt, D.R., Ripke, S., MacIntyre, D.J., McGhee, K.A., Maclean, A.W., et al. (2012). Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. Mol. Psychiatry *17*, 36–48.
46. Risch, N., and Teng, J. (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res. *8*, 1273–1288.
47. Risch, N. (2001). Implications of multilocus inheritance for gene-disease association studies. Theor. Popul. Biol. *60*, 215–220.
48. Gauderman, W.J. (2003). Candidate gene association analysis for a quantitative trait, using parent-offspring trios. Genet. Epidemiol. *25*, 327–338.

The American Journal of Human Genetics
Supplemental Data

Disease and Polygenic Architecture: Avoid Trio Design and Appropriately Account for Unscreened Control Subjects for Common Disease

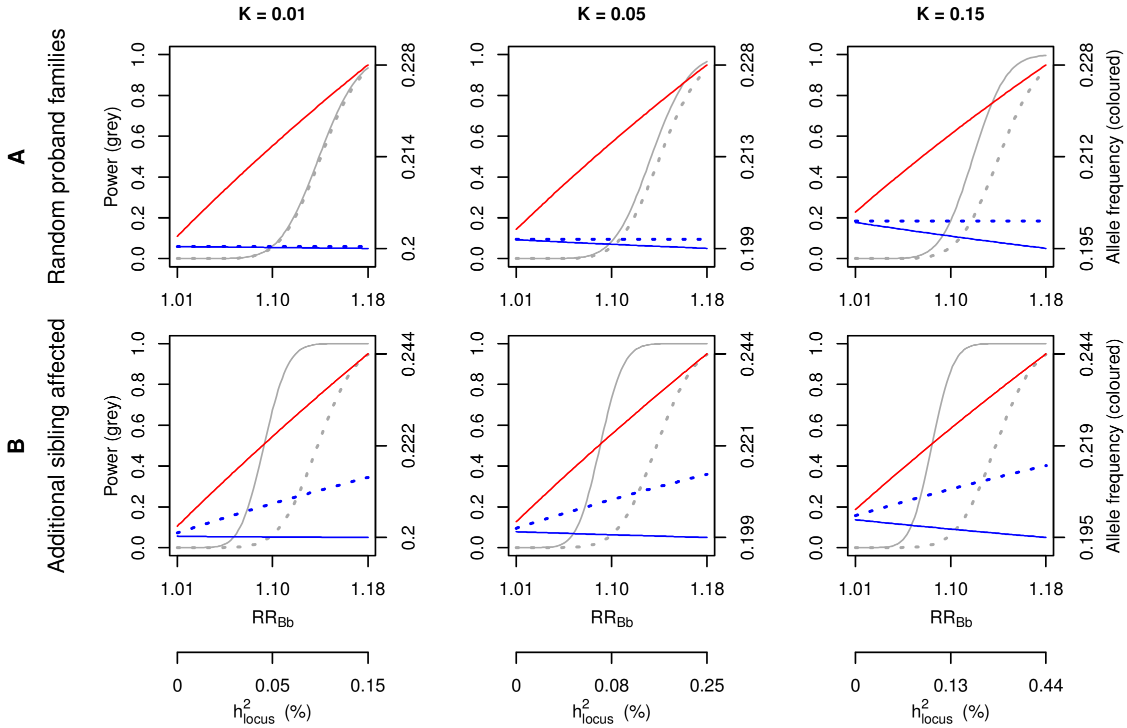
Wouter J. Peyrot, Dorret I. Boomsma, Brenda W.J.H. Penninx, and Naomi R. Wray

Figure S1. Pseudocontrols of random families with at least one affected proband case are equal to unscreened controls.



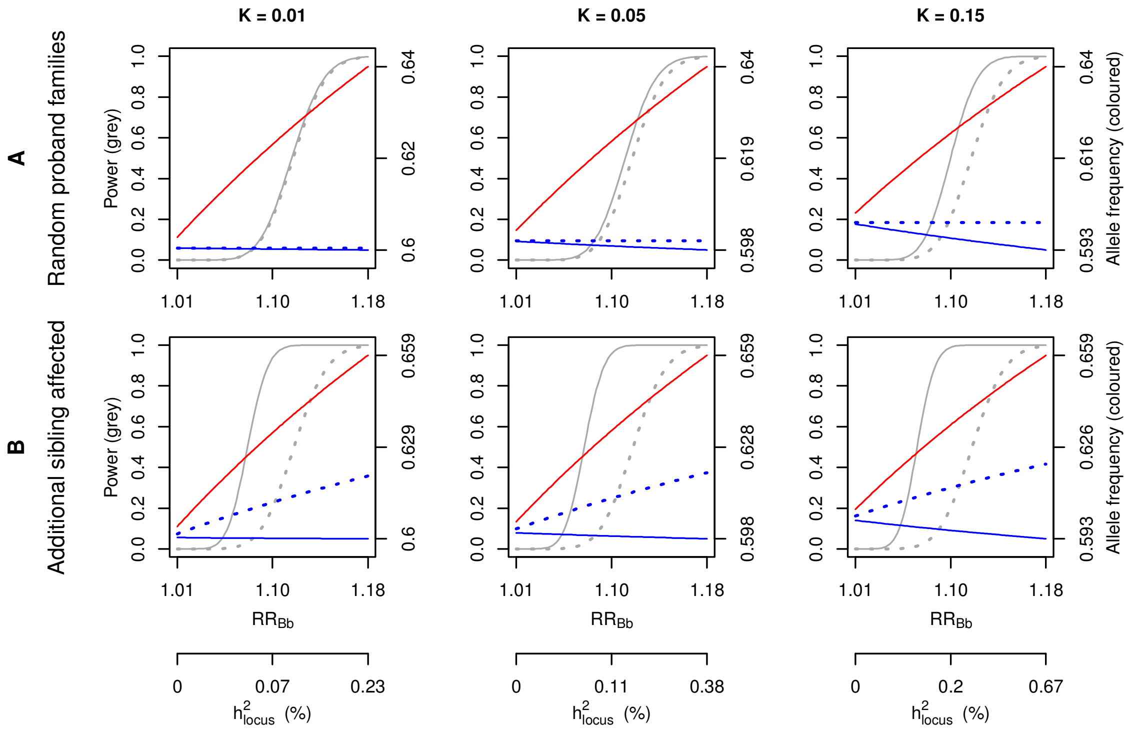
Pseudocontrols of random families with at least one affected proband case are equal to unscreened controls (i.e. population mean) as displayed for the allele frequency of single loci of different effect-size (first two rows) and the mean genetic liability $E(G)$ (population mean equals 0) for variable heritability h_l^2(bottom row) and different baseline population risk K. The equivalence is exact and follows from the closed formulas provided in the R scripts, but is non-trivial to display in equations, because multiple sequential probabilities were needed to derive at the allele frequency and mean genetic liability in pseudocontrols. The equivalence can be understood intuitively by realizing that the non-transmitted alleles of random proband family are, in fact, part of the population background.

Figure S2. Power to detect a single SNP in trio-design and unscreened control studies, p=0.2



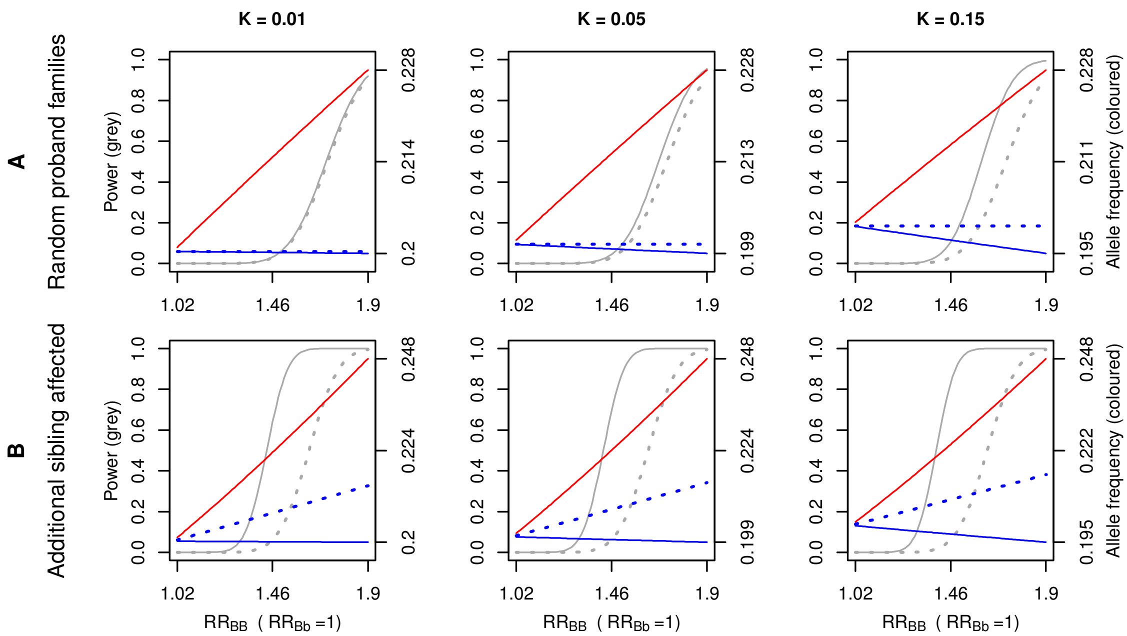
Power to detect a single SNP with risk allele frequency $p = 0.2$ for case vs screened controls (solid grey line) and case vs pseudocontrol (dotted grey line). The allele frequencies of proband cases are displayed as the red solid line, the allele frequency of screened controls as the solid blue line, and the allele frequency of pseudocontrols in the dotted blue line. The allele frequencies of pseudocontrols from proband random families equal unscreened population controls, which is reflected by the horizontal blue dotted lines at 0.2 in Panel A. Note that the grey lines equal the solid and dotted lines in Main Figure 2; the unscreened controls are not displayed in the Supplemental Figures, because they will always have an allele frequency equal to the population frequency.

Figure S3. Power to detect a single SNP in trio-design and unscreened control studies, p=0.6



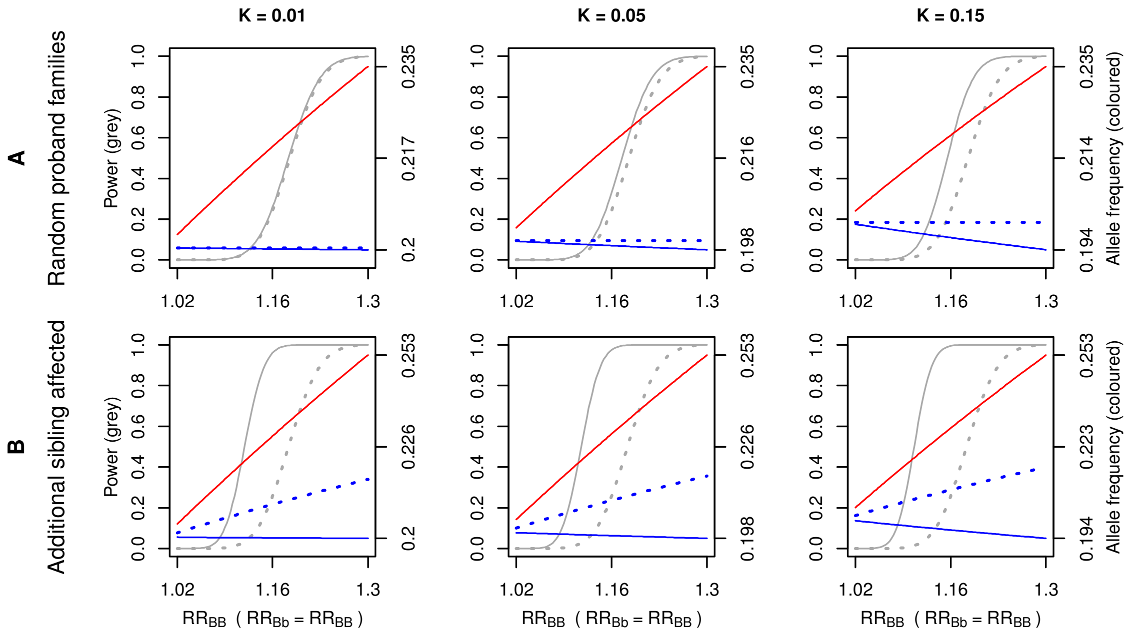
The power is displayed for a risk allele with frequency p=0.6, and results indicate that the conclusions do not depend on the allele frequency (noting that in Figure S2 a locus with p=0.2 was displayed). See the legend of Figure S2 for details.

Figure S4. Power in trio design to detect SNP with underlying recessive effect



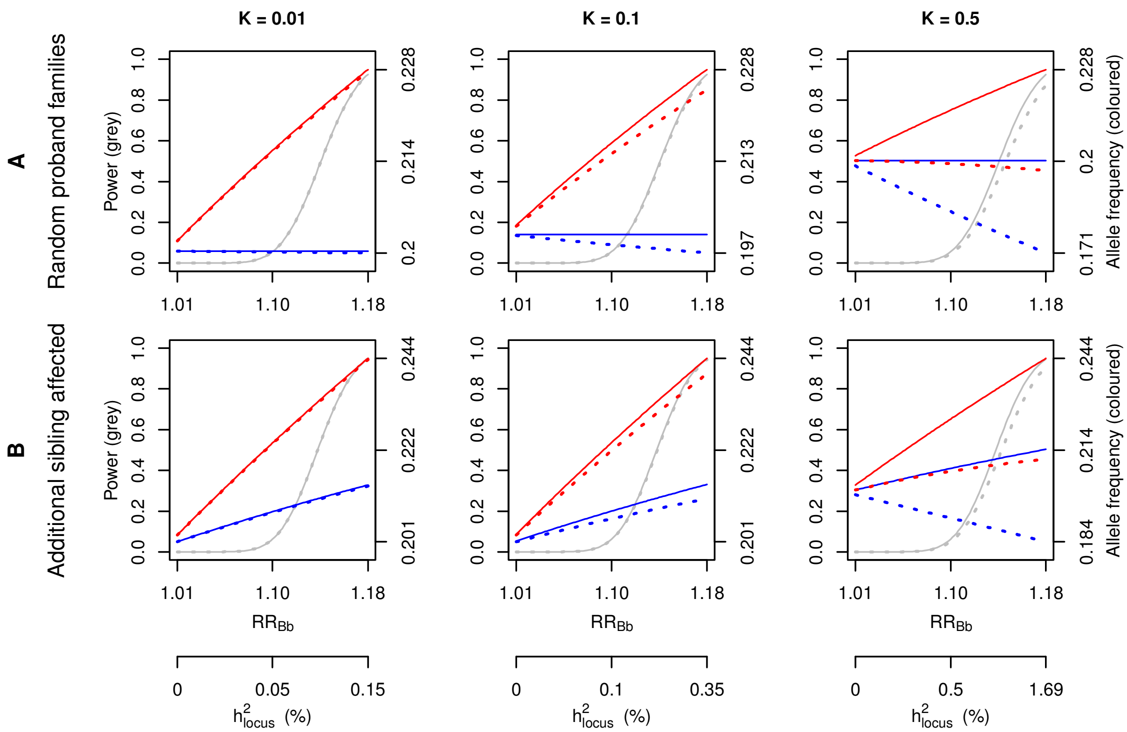
Power to detect the additive effect a single SNP with risk allele frequency $p = 0.2$ with an underlying recessive effect for case vs screened controls (solid grey line) and case vs pseudocontrol (dotted grey line). The allele frequency of cases is displayed as the red solid line, the allele frequency of screened controls as the solid blue line, and the allele frequency of pseudocontrols in the dotted blue line. Note that the RR_{BB} are being displayed for a larger range than in Figure S2 ($1.9 > 1.18^2 = 1.39$), i.e. an actual recessive allele results in less power given RR_{BB}.

Figure S5. Power in trio design to detect SNP with underlying dominant effect



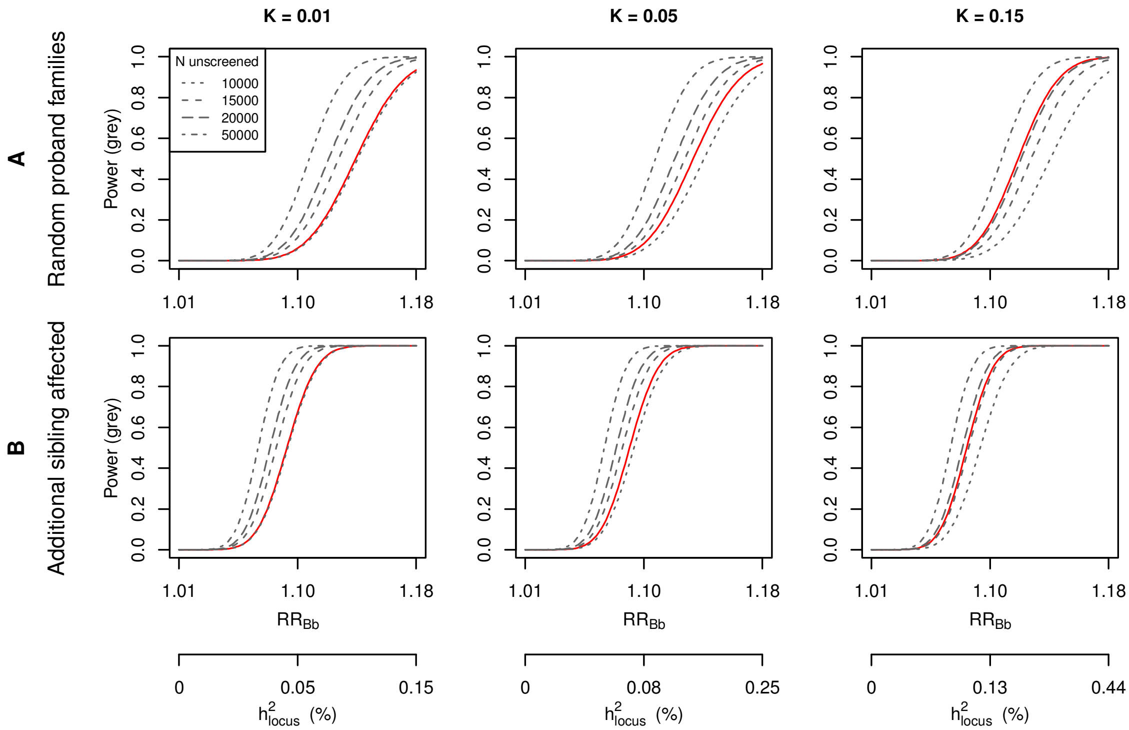
Power to detect the additive effect a single SNP with risk allele frequency $p = 0.2$ with an actual dominant effect for case vs screened controls (solid grey line) and case vs pseudocontrol (dotted grey line). The allele frequency of cases is displayed as the red solid line, the allele frequency of screened controls as the solid blue line, and the allele frequency of pseudocontrols in the dotted blue line. Note that the RR_{BB} are being displayed for a smaller range than in Figure S2 ($1.3 < 1.18^2 = 1.39$), i.e. a dominant allele results in more power given RR_{BB}.

Figure S6. Power to detect SNP in trios with unaffected parents



Power to detect a single SNP with risk allele frequency $p = 0.2$ for cases vs pseudocontrols without conditioning on parents (solid grey line) and case vs pseudocontrol restricted to trios with unaffected parents (dotted grey line). The allele frequency of cases from trios without conditioning on parents is displayed as the red solid line, and the allele frequency of their pseudocontrols as the solid blue line. The allele frequency in cases from trios with unaffected parents is displayed as the red dotted line, and the allele frequency in their pseudocontrols as the dotted blue line. To summarize: solid=no selection on parents; dotted=unaffected parents; grey=power; red=allele frequency case; blue=allele frequency pseudocontrol. Note that the grey lines overlap, i.e. selecting trios with unaffected parents does not increase power in pseudocontrol studies. Furthermore, note that for $K = 0.1$ and $K = 0.5$ the allele frequencies are lower in trios from unaffected parents, but this difference is proportional for cases and pseudocontrol resulting in no power-difference.

Figure S7. Power to detect a risk variant from screened *vs.* unscreened controls studies



Power to detect a risk variant with risk allele frequency $p = 0.2$ for 10,000 proband cases vs 10,000 screened controls (solid red line) and 10,000 proband cases vs respectively 10,000 unscreened controls (dotted line), 15,000 unscreened controls (short dashed), 20,000 unscreened controls (long dashed), and 50,000 unscreened controls (dot-dashed).

Table S1. Values of the Haseman Elston cross-product accounting for falsely classified controls

$y_{true,i}$	$y_{true,j}$	$y_{assumed,i}$	$y_{assumed,j}$	$\mathbb{P}_{ij}$	Z_{ij}
1	1	0	0	$((1-P_{assumed})F)^2$	$\frac{P_{assumed}}{1-P_{assumed}}$
1	1	0	1	$(1-P_{assumed})FP_{assumed}$	-1
1	1	1	0	$P_{assumed}(1-P_{assumed})F$	-1
1	1	1	1	$P^2_{assumed}$	$\frac{1-P_{assumed}}{P_{assumed}}$
1	0	1	0	$P_{assumed}(1-P_{assumed})(1-F)$	-1
1	0	0	0	$(1-P_{assumed})F(1-P_{assumed})(1-F)$	$\frac{P_{assumed}}{1-P_{assumed}}$
0	1	0	1	$(1-P_{assumed})(1-F)P_{assumed}$	-1
0	1	0	0	$(1-P_{assumed})(1-F)(1-P_{assumed})F$	$\frac{P_{assumed}}{1-P_{assumed}}$
0	0	0	0	$((1-P_{assumed})(1-F))^2$	$\frac{P_{assumed}}{1-P_{assumed}}$

To adjust the transformation from the heritability on the observed scale $\hat{h}^2_o$ to the liability scale $\hat{h}^2_l$ for a proportion $F = \frac{N_{false\ controls}}{N_{all\ controls}}$ of falsely classified controls, we closely followed the derivations of Golan et al, which we recommend for further reading (paragraphs 1.2 and 1.3 of their Supplemental Materials).[1] The adjusted expected values of the cross-product Z_{ij} used for Haseman Elston-regression follow from considering the true disease status y_{true} and assumed disease status $y_{assumed}$ with probabilities

$$\mathbb{P}(y_{true} = 1 \ \& \ y_{assumed} = 1) = P_{assumed}$$
$$\mathbb{P}(y_{true} = 1 \ \& \ y_{assumed} = 0) = (1-P_{assumed})F$$
$$\mathbb{P}(y_{true} = 0 \ \& \ y_{assumed} = 0) = (1-P_{assumed})(1-F)$$

The 9 possible pairs, their probabilities $\mathbb{P}_{ij}$ and values of cross-product Z_{ij} are displayed in the Table. The expected values of $\mathbb{E}[Z_{ij}|y_{true,i}, y_{true,j}]$ follow as:

$$\mathbb{E}[Z_{ij}|y_{true,i} = y_{true,j} = 1] = \frac{\sum \mathbb{P}_{ij|y_{true,i}=y_{true,j}=1} Z_{ij|y_{true,i}=y_{true,j}=1}}{\sum \mathbb{P}_{ij|y_{true,i}=y_{true,j}=1}} = \frac{P_{assumed}(1-P_{assumed})(1-F)^2}{(P_{assumed}+(1-P_{assumed})F)^2}$$
$$\mathbb{E}[Z_{ij}|y_{true,i} \neq y_{true,j}] = \frac{P_{assumed}(F-1)}{(P_{assumed}+(1-P_{assumed})F)}$$
$$\mathbb{E}[Z_{ij}|y_{true,i} = y_{true,j} = 0] = \frac{P_{assumed}}{1-P_{assumed}}$$

Given these $\mathbb{E}[Z_{ij}|y_{true,i}, y_{true,j}]$ the derivation of Golan et al can be followed with $P_{Golan} = P_{true} = P_{assumed} + (1-P_{assumed})F$ to derive at the transformation of the observed to the liability scale as: $\hat{h}^2_l = \frac{K^2(1-K)^2}{P(1-P)(1-F)^2 z^2} \hat{h}^2_{occ}$, where $P = P_{assumed}$.

Table S2. Simulation of falsely classified controls

Simulation parameters				Haseman-Elston regression					
				$\hat{h}^2_{occ}$		$\hat{h}^2_l$ (assuming F=0)		$\hat{h}^2_l$ (corrected for F)	
K	h^2_l	P	F	Mean	SE	Mean	SE	Mean	SE
Parameters of Major Depressive Disorder									
0.2	0.4	0.5	0	0.3048	0.0131	0.3983	0.0171	0.3983	0.0171
0.2	0.4	0.5	0.1	0.2467	0.0112	0.3224	0.0146	0.3980	0.0180
0.2	0.4	0.5	0.2	0.1834	0.0095	0.2396	0.0124	0.3744	0.0194
0.2	0.4	0.25	0	0.2288	0.0062	0.3985	0.0107	0.3985	0.0107
0.2	0.4	0.25	0.1	0.1795	0.0088	0.3127	0.0153	0.3861	0.0189
0.2	0.4	0.25	0.2	0.1545	0.0055	0.2691	0.0096	0.4204	0.0150
Parameters of Schizophrenia									
0.01	0.8	0.5	0	1.4699	0.0130	0.8113	0.0072	0.8113	0.0072
0.01	0.8	0.5	0.005	1.4358	0.0116	0.7924	0.0064	0.8004	0.0065
0.01	0.8	0.5	0.01	1.4096	0.0157	0.7780	0.0087	0.7938	0.0089
0.01	0.8	0.25	0	1.0927	0.0055	0.8041	0.0040	0.8041	0.0040
0.01	0.8	0.25	0.005	1.0829	0.0078	0.7969	0.0057	0.8049	0.0058
0.01	0.8	0.25	0.01	1.0737	0.0049	0.7901	0.0036	0.8061	0.0037
Additional parameter settings to further validate the derived equation									
0.2	0.8	0.5	0	0.6282	0.0182	0.8207	0.0238	0.8207	0.0238
0.2	0.8	0.5	0.1	0.4964	0.0117	0.6485	0.0153	0.8006	0.0189
0.2	0.8	0.5	0.2	0.4062	0.0076	0.5307	0.0100	0.8293	0.0156
0.2	0.8	0.25	0	0.4608	0.0077	0.8028	0.0135	0.8028	0.0135
0.2	0.8	0.25	0.1	0.3722	0.0061	0.6484	0.0107	0.8005	0.0132
0.2	0.8	0.25	0.2	0.2956	0.0062	0.5150	0.0109	0.8047	0.0170
0.01	0.4	0.5	0	0.7287	0.0108	0.4022	0.0059	0.4022	0.0059
0.01	0.4	0.5	0.005	0.6993	0.0148	0.3859	0.0082	0.3898	0.0082
0.01	0.4	0.5	0.01	0.7022	0.0132	0.3876	0.0073	0.3954	0.0074
0.01	0.4	0.25	0	0.5395	0.0047	0.3970	0.0035	0.3970	0.0035
0.01	0.4	0.25	0.005	0.5393	0.0076	0.3969	0.0056	0.4009	0.0057
0.01	0.4	0.25	0.01	0.5375	0.0064	0.3956	0.0047	0.4036	0.0048

To validate the Equation 3, $\hat{h}^2_l = \frac{K^2(1-K)^2}{P(1-P)(1-F)^2 z^2}\hat{h}^2_{occ}$, we performed a simulation study in line with Golan et al (Supplemental Materials paragraph 5.3).[1]

1. MAFs of 10,000 SNPs in full linkage equilibrium were randomly sampled from $U[0.05,0.5]$, and the effect sizes were randomly sampled from $N(0, h^2_l/10{,}000)$.
2. An individual was generated by
   a. Randomly assigning alleles with the probabilities given by the MAFs
   b. Standardizing the allele counts by $(allele\ count - 2 * MAF)/\sqrt{2MAF(1-MAF)}$.
   c. Assessing the genetic liability G as the product of the standardized allele counts with the effects
   d. Assessing the phenotypic liability l as $G + E$ with E randomly drawn from $N(0, 1 - h^2_l)$

e. Defining disease status $y = 1$ for those with $l > T$ with T the liability threshold corresponding to a proportion of K cases

3. Step 2 was repeated until we obtained $2{,}000$ cases, an additional $F * 2{,}000$ cases which we labeled as controls, and $(1 - F) * 2{,}000$ true controls. The cases and controls were saved in a single ped-file.
4. Plink was used to transform the ped-file to a bim-file,[2] and GCTA[3] to estimate the genetic relationship matrix and to perform cross-product Haseman-Elston regression with the "--HEreg" option yielding $\hat{h}^2_{occ}$.
5. Steps 1-4 were repeated 10 times. The mean of these 10 point-estimates of the SNP-heritability are displays, as well as their standard error (SE) estimated as their standard deviation divided by $\sqrt{10}$.
6. The mean $\hat{h}^2_o$ was, first, transformed to the liability scale assuming $F = 0$ (i.e. with Equation 2, $\hat{h}^2_l = \frac{K^2(1-K)^2}{P(1-P)z^2}\hat{h}^2_{occ}$), and second, with Equation 3, $\hat{h}^2_l = \frac{K^2(1-K)^2}{P(1-P)(1-F)^2z^2}\hat{h}^2_{occ}$. Simulation illustrates that Equation 3 appropriately accounts for unscreened controls, because the actual simulated h^2_l fall within the approximate 95% confidence interval of the mean $\hat{h}^2_l$ from simulation (mean ± 1.96*SE).

Table S3. Analytical derivation of genetic liabilities in trios versus simulation

				Screened controls		Case		Pseudo control		Case | sib aff		Ps contr | sib aff	
Method	K	h_l^2	ρ_l	$\sigma^2(G)$	$E(G)$	$\sigma^2(G)$	$E(G)$	$\sigma^2(G)$	$E(G)$	$\sigma^2(G)$	$E(G)$	$\sigma^2(G)$	$E(G)$
Sim	0.001	0.8	0	0.7932	-0.0027	0.2052	2.6945	0.8059	-0.0014	0.2134	2.9642	0.6400	0.9853
Ana	0.001	0.8	0	0.7933	-0.0027	0.2034	2.6937	0.8000	0.0000	0.2133	2.9529	0.6347	0.9788
Sim	0.001	0.8	0.5	0.9450	-0.0058	0.2259	2.8185	0.9360	0.4686	0.2415	3.1014	0.7186	1.4582
Ana	0.001	0.8	0.5	0.9451	-0.0058	0.2250	2.8182	0.9396	0.4697	0.2381	3.0970	0.7162	1.4595
Sim	0.001	0.4	0	0.3982	-0.0013	0.2502	1.3461	0.3991	0.0003	0.2417	1.6929	0.3489	0.5700
Ana	0.001	0.4	0	0.3983	-0.0013	0.2508	1.3468	0.4000	0.0000	0.2384	1.7045	0.3622	0.5674
Sim	0.001	0.4	0.5	0.4377	-0.0017	0.2688	1.4265	0.4392	0.1287	0.2519	1.8069	0.3818	0.7377
Ana	0.001	0.4	0.5	0.4377	-0.0017	0.2668	1.4286	0.4386	0.1299	0.2506	1.8200	0.3896	0.7484
Sim	0.01	0.8	0	0.7596	-0.0216	0.2218	2.1327	0.7996	-0.0004	0.2342	2.3623	0.6462	0.7870
Ana	0.01	0.8	0	0.7595	-0.0215	0.2220	2.1322	0.8000	0.0000	0.2344	2.3578	0.6432	0.7813
Sim	0.01	0.8	0.5	0.8914	-0.0350	0.2488	2.2414	0.9403	0.3723	0.2674	2.4906	0.7281	1.1794
Ana	0.01	0.8	0.5	0.8913	-0.0350	0.2492	2.2423	0.9403	0.3737	0.2642	2.4889	0.7282	1.1733
Sim	0.01	0.4	0	0.3899	-0.0109	0.2552	1.0664	0.4015	-0.0012	0.2451	1.3546	0.3632	0.4459
Ana	0.01	0.4	0	0.3899	-0.0108	0.2555	1.0661	0.4000	0.0000	0.2437	1.3561	0.3637	0.4513
Sim	0.01	0.4	0.5	0.4270	-0.0128	0.2720	1.1315	0.4375	0.1025	0.2571	1.4517	0.3905	0.5990
Ana	0.01	0.4	0.5	0.4271	-0.0129	0.2723	1.1323	0.4386	0.1029	0.2568	1.4509	0.3916	0.5965
Sim	0.1	0.8	0	0.6157	-0.1558	0.2682	1.4039	0.8004	-0.0003	0.2844	1.5857	0.6633	0.5286
Ana	0.1	0.8	0	0.6157	-0.1560	0.2682	1.4040	0.8000	0.0000	0.2818	1.5844	0.6615	0.5261
Sim	0.1	0.8	0.5	0.7104	-0.1982	0.3073	1.4969	0.9420	0.2497	0.3265	1.7023	0.7538	0.8060
Ana	0.1	0.8	0.5	0.7102	-0.1984	0.3071	1.4968	0.9419	0.2495	0.3208	1.6993	0.7530	0.8035
Sim	0.1	0.4	0	0.3539	-0.0780	0.2670	0.7020	0.3998	0.0000	0.2567	0.9043	0.3668	0.3016
Ana	0.1	0.4	0	0.3539	-0.0780	0.2671	0.7020	0.4000	0.0000	0.2562	0.9040	0.3671	0.3009
Sim	0.1	0.4	0.5	0.3851	-0.0873	0.2859	0.7480	0.4392	0.0677	0.2724	0.9727	0.3971	0.4003
Ana	0.1	0.4	0.5	0.3851	-0.0873	0.2858	0.7483	0.4387	0.0680	0.2713	0.9721	0.3961	0.3997

Legend to Table S3.

We validated the analytical estimations (see Supplemental Methods) of the mean genetic liabilities $E(G)$ with a simulation study. The heritability h_l^2, phenotypic correlation between parents ρ_l, the population disease frequency K, and corresponding threshold T were defined as described in the main text. Hereby, the variance-covariance matrix of the genetic liabilities of the parents was defined as

$$\Sigma(G_m, G_f) = \begin{pmatrix} h_l^2 & \rho_l h_l^2 h_l^2 \\ \rho_l h_l^2 h_l^2 & h_l^2 \end{pmatrix}$$

with $V_G = h_l^2 V_l = h_l^2$. Subsequently, the genetic liabilities of the mothers and fathers were randomly drawn from this bivariate normal distribution. The genetic liabilities of the first and second sibling were independently defined as $G_s = \frac{1}{2}G_m + \frac{1}{2}G_f + G_{residual}$, where $G_{residual}$ represent Mendelian variation and was randomly drawn from the normal distribution with mean 0 and variation $\frac{1}{2}V_G$.[4] The phenotypes l of the siblings were than independently defined as $l_s = G_s + E_s$, with E_s randomly drawn from $N(0,1 - h_l^2)$. To conclude, the genetic liability of the complement $c1$ of the first sibling $s1$ was defined as $G_{c1} = G_m + G_f - G_{s1}$. In this manner, $l_{s1}, G_{s1}, l_{s2}, G_{s2}, G_m, G_f$ and G_{c1} were defined for 10^8 families. We note that the value of $\sigma^2(G_s)$ thus simulated was in line with previous theoretical derivations $V_G + \frac{1}{2}\rho_G V_G$.[4,5] The respective variances, covariances and means were estimated from this simulation study and were in line with the theoretically derived values (see Table S3). Simulations were performed in R.[6]

Table S4. Heuristic prediction of assessed heritability in trios versus simulation

Simulation parameters				$\hat{h}_l^2$ screened control			$\hat{h}_l^2$ pseudocontrol		
				Simulation			Simulation		
K	h_l^2	sib aff	ρ_l	Mean	SE	Pred. $\hat{h}_l^2$	Mean	SE	Pred. $\hat{h}_l^2$
0.3	0.8	Y	0	0.9885	0.0225	0.9864	0.2182	0.0196	0.2331
0.3	0.8	N	0.5	0.9741	0.0155	0.9833	0.3303	0.0139	0.3221
0.3	0.8	Y	0.5	1.2126	0.0113	1.2214	0.1452	0.0129	0.1736
0.1	0.8	Y	0	0.9888	0.0122	0.9957	0.3613	0.0158	0.3682
0.1	0.8	N	0.5	0.9418	0.0152	0.9447	0.5001	0.0129	0.5114
0.1	0.8	Y	0.5	1.2115	0.0105	1.1839	0.2822	0.0107	0.2638
0.01	0.8	Y	0	0.9899	0.0069	0.9764	0.4249	0.0073	0.4287
0.01	0.8	N	0.5	0.8810	0.0096	0.8945	0.6054	0.0067	0.6022
0.01	0.8	Y	0.5	1.1072	0.0045	1.0987	0.3135	0.0057	0.2985
0.3	0.4	Y	0	0.6153	0.0127	0.5913	0.1397	0.0213	0.1491
0.3	0.4	N	0.5	0.4643	0.0162	0.4640	0.2154	0.0180	0.1860
0.3	0.4	Y	0.5	0.6995	0.0210	0.6957	0.1438	0.0132	0.1362
0.1	0.4	Y	0	0.6435	0.0140	0.6340	0.2257	0.0118	0.2391
0.1	0.4	N	0.5	0.4539	0.0086	0.4591	0.3002	0.0104	0.3043
0.1	0.4	Y	0.5	0.7240	0.0117	0.7379	0.1998	0.0083	0.2154
0.01	0.4	Y	0	0.6531	0.0056	0.6445	0.2952	0.0059	0.2824
0.01	0.4	N	0.5	0.4507	0.0075	0.4524	0.3573	0.0043	0.3655
0.01	0.4	Y	0.5	0.7451	0.0057	0.7391	0.2604	0.0093	0.2518

To formally get from the $E(G)$ (Table S3) of cases and controls to the SNP-heritability $\hat{h}_l^2$ that would be assessed is non-trivial, because no normal distribution thresholds exist to define the pseudocontrols or the probands with an additional affected sibling (which form a non-random subset of all cases not defined by a specific threshold). $\hat{h}_l^2$ was therefore heuristically derived and validated with a simulation study of individual level SNP-data. In short, for any baseline disease frequency K, a unique set of T, z, and i can be found such that K equals $P(l > T | l \sim N(0,1))$, z the height of the standard normal distribution at T, and $i = z/K$ the mean l of cases, which results in a mean G in cases of ih_l^2. We numerically inverted this equation in R to find an unique equivalent-K matching the difference between $E(G_{case}) - E(G_{(pseudo)control})$. The equivalent-$K$, corresponding equivalent-z and Equation 3 yields the heritability that would be assessed with Haseman-Elston regression (Pred. $\hat{h}_l^2$), and was validated with simulation study:

1. Following Golan et al,[1] the MAFs of 10,000 SNPs in full linkage disequilibrium were randomly sampled from $U[0.05,0.5]$, and the effect sizes were randomly sampled from $N(0, h_l^2/10{,}000)$.
2. An individual was generated by
    a. Randomly assigning alleles with the probabilities given by the MAFs
    b. Standardizing the allele counts by $(allele\ count - 2 * MAF)/\sqrt{2MAF(1 - MAF)}$.

c. Assessing the genetic liability G as the product of the standardized allele counts with the effects
d. Assessing the phenotypic liability l as $G + E$ with E randomly drawn from $N(0, 1 - h_l^2)$
e. Defining disease status $y = 1$ for those with $l > T$ with T the liability threshold corresponding to a proportion of K cases

3. Assortative mating ρ_l was simulated following
   a. The genotypes and phenotypes of 600 men l_{men} and 600 women l_{women} were simulated
   b. A vector V was simulated as $V = \rho_l l_{men} + N(0,1 - \rho_l^2)$ so that $cor(l_{men}, V) = cov(l_{men}, V)/(\sigma_{l_{men}}\sigma_V) = cov(l_{men}, \rho_l l_{men})/(1\sigma_V) = \rho_l/\sqrt{\sigma^2_{\rho_l l_{men}} + 1 - \rho_l^2} = \rho_l$
   c. Subsequently, the l_{women} were ordered in line with V thereby ensuring $cor(l_{men}, l_{women}) = \rho_l$
4. For the 600 pair of spouses, families were generated as follows
   a. Kid-1 got one random allele from the father and one from the mother for all of the 10,000 loci. Subsequently, l and disease status y were generates as described above.
   b. The genetic complement of Kid-1 was formed by the non-transmitted alleles of the parents
   c. Kid-2 was generated as Kid-1
5. Affected proband (Kid-1) were selected as cases. Depending on the type of families simulated, we additionally conditioned on $y_{Kid-2} = 1$.
6. Unaffected Kid-1's were selected as screened controls.
7. Step 2-6 were repeated until 2,000 cases and 2,000 screened controls were collected
8. Cross-product Haseman-Elston regression yielded the $\hat{h}^2_{occ}$ for case vs screened controls and case vs pseudocontrols, which were than transformed to the liability scale with $\hat{h}_l^2 = \hat{h}^2_{occ}\frac{K^2(1-K)^2}{P(1-P)z^2}$
9. Steps 1-8 were repeated 10 times for the different setting of K, h_l^2, and ρ_l. The mean of these 10 point-estimates of the SNP-heritability are displays, as well as their standard error (SE) estimated as their standard deviation divided by $\sqrt{10}$.
10. The heuristically predicted $\hat{h}_l^2$ are within or very close to the *ballpark* 95% confidence interval of the mean $\hat{h}_l^2$ from simulation (mean ± 1.96*SE), which justifies the use of this heuristic approach for Main Figure 1.

Table S5. Analytical derivation of allele frequencies in trios versus simulation

	Genotype relative risk		Random families with at least one affected sibling			Second sibling affected		Second sibling aff. Parents unaffected		Assortative mating parents		
Method	Bb	BB	Case	Scr control	Ps control	Case	Ps control	Case	Ps control	Case	Scr control	Ps control
K=0.01; p=0.2												
Sim	1.00	2.25	0.2381	0.1996	0.1995	0.2723	0.2163	0.2718	0.2155	0.2596	0.1995	0.2052
Ana	1.00	2.25	0.2381	0.1996	0.2000	0.2695	0.2205	0.2688	0.2199	0.2593	0.1994	0.2056
Sim	1.50	2.25	0.2727	0.1993	0.2000	0.3159	0.2316	0.3141	0.2303	0.2865	0.1980	0.2110
Ana	1.50	2.25	0.2727	0.1993	0.2000	0.3171	0.2358	0.3161	0.2349	0.2862	0.1991	0.2109
Sim	2.25	2.25	0.3106	0.1989	0.2002	0.3671	0.2512	0.3660	0.2502	0.3167	0.2012	0.2165
Ana	2.25	2.25	0.3103	0.1989	0.2000	0.3663	0.2475	0.3652	0.2466	0.3169	0.1988	0.2169
K=0.01; p=0.8												
Sim	1.00	2.25	0.8890	0.7991	0.8001	0.9174	0.8424	0.9167	0.8413	0.8909	0.7982	0.8128
Ana	1.00	2.25	0.8889	0.7991	0.8000	0.9179	0.8446	0.9174	0.8437	0.8907	0.7991	0.8131
Sim	1.50	2.25	0.8571	0.7995	0.8004	0.8767	0.8267	0.8763	0.8261	0.8634	0.7992	0.8085
Ana	1.50	2.25	0.8571	0.7994	0.8000	0.8788	0.8283	0.8784	0.8278	0.8637	0.7994	0.8085
Sim	2.25	2.25	0.8181	0.7998	0.7998	0.8233	0.8107	0.8233	0.8104	0.8294	0.8001	0.8029
Ana	2.25	2.25	0.8182	0.7998	0.8000	0.8241	0.8086	0.8239	0.8085	0.8295	0.7997	0.8028
K=0.3; p=0.2												
Sim	1.00	2.25	0.2381	0.1836	0.2000	0.2696	0.2206	0.2415	0.1956	0.2593	0.1730	0.2055
Ana	1.00	2.25	0.2381	0.1837	0.2000	0.2695	0.2205	0.2403	0.1943	0.2593	0.1736	0.2056
Sim	1.50	2.25	0.2727	0.1688	0.2000	0.3171	0.2358	0.2733	0.1980	0.2861	0.1644	0.2109
Ana	1.50	2.25	0.2727	0.1688	0.2000	0.3171	0.2358	0.2732	0.1980	0.2862	0.1628	0.2109
Sim	2.25	2.25	0.3104	0.1527	0.2000	0.3663	0.2475	0.3152	0.2068	0.3169	0.1539	0.2169
Ana	2.25	2.25	0.3103	0.1527	0.2000	0.3663	0.2475	0.3148	0.2060	0.3169	0.1514	0.2169
K=0.3; p=0.8												
Sim	1.00	2.25	0.8889	0.7619	0.8000	0.9178	0.8445	0.8953	0.8062	0.8908	0.7609	0.8131
Ana	1.00	2.25	0.8889	0.7619	0.8000	0.9179	0.8446	0.8958	0.8066	0.8907	0.7602	0.8131
Sim	1.50	2.25	0.8571	0.7755	0.8000	0.8787	0.8283	0.8622	0.8055	0.8637	0.7719	0.8085
Ana	1.50	2.25	0.8571	0.7755	0.8000	0.8788	0.8283	0.8621	0.8056	0.8637	0.7726	0.8085
Sim	2.25	2.25	0.8183	0.7922	0.8000	0.8242	0.8086	0.8184	0.8021	0.8294	0.7893	0.8028
Ana	2.25	2.25	0.8182	0.7922	0.8000	0.8241	0.8086	0.8184	0.8026	0.8295	0.7876	0.8028

Legend to Table S5.

We checked the analytical estimations (described in Supplemental Methods) of allele frequencies with a simulation study. Genotypes were simulated by first randomly assigning each parent two alleles with frequency $p = P(B)$ of the risk allele B. Then, genotypes of the first and second siblings were defined by assigning them a single random allele from both of their parents. The genotypes of the pseudocontrols were defined as the two alleles of the parents not transmitted to the first sibling. Disease status was randomly assigned to parents, siblings, with a probability of disease per genotype of $P(\text{Disease}|\text{Genotype})$ (see Witte et al for details)[7]. Families with the first sibling affected were selected as proband families with the first sibling serving as the proband case. Assortative mating was simulated as the non-random mating fraction $\alpha = 0.3$ (see Supplemental Methods section 2.4 for details), which correspond to a spouse-correlation at the locus of 0.3 (note that this unrealistic large value is merely to validate theory, because assortative mating will have no impact on allele frequency as for a phenotypic spouse-correlation of 0.3 a locus explaining 1% of variance would have a spouse-correlation of only $0.3 * 0.01 = 0.003$). We simulated 10^8 families and compared allele frequencies in different types of cases, controls, and pseudocontrols to the algebraic estimates. Results displayed in this Table validate the analytical estimations described in the Supplemental Methods that were used to make the relevant Figures and Tables.

Supplemental Methods

1. Derivation of genetic liabilities in trio design

The mean genetic liabilities (breeding values) $E(G)$ and their variances were subsequently derived for random families (Section 1.1), families with one affected sibling (Section 1.2), and families with two affected siblings (Section 1.3). Therefore, variance-covariance matrices were derived for these family's phenotypic liabilities and genetic liabilities. The mean genetic liability of screened controls in the offspring generation was derived in Section 1.4. The analytical estimates of the mean genetic liabilities and their variances were validated with a simulation study (Table S3). In Table S4, the derived mean genetic liabilities are used to heuristically predict the SNP-based heritability that would be assessed with Haseman Elston-regression, which is again validated with a simulation study.

Consider a complex disease with a population frequency K and heritability h_l^2 in the parental population. Define phenotype l to represent the underlying liability for disease with variance $V_l = 1$ (the choice for V_l is arbitrary, but conveniently set to 1). The variance of genetic liabilities G equals $V_G = V_l h_l^2 = h_l^2$, while the environmental variance equals $V_E = V_l - V_G = 1 - h_l^2$. Assuming that the parents have a phenotypic correlation of $\rho_l \geq 0$, the genetic correlation follows as $\rho_G = h_l^2 \rho_l$ (page 175 of Falconer and Mackay)[8] and the genetic covariance as $\rho_G V_G$.

1.1 Variances and covariances of genetic liabilities in random families

Consider families with a mother (m), father (f), first sibling $(s1)$, second sibling $(s2)$ and the pseudocontrol of the first sibling (interchangeably referred to as the complement of the first sibling, $c1$). Their genetic liability values are denoted with G_m, G_f, G_{s1}, G_{s2}, respectively. The variance of genetic liabilities in the siblings equals $\sigma^2(G_{s1}) = \sigma^2(G_{s2}) = \sigma^2(G_s) = \sigma^2\left(\frac{1}{2}G_m + \frac{1}{2}G_f\right) + V_{residual}$, where $V_{residual}$ represents Mendelian variation. Bulmer (page 175)[4] proved that $V_{residual} = \frac{1}{2}V_G$, which gives $\sigma^2(G_s) = \sigma^2\left(\frac{1}{2}G_m\right) + \sigma^2\left(\frac{1}{2}G_f\right) + 2\sigma\left(\frac{1}{2}G_m, \frac{1}{2}G_f\right) + \frac{1}{2}V_G = V_G + \frac{1}{2}\rho_G V_G$. In addition, Bulmer showed that the variation of non-genetic effects (E) is not effected by assortative mating, which gives the phenotypic variation of the siblings as $\sigma^2(l_{s1}) = \sigma^2(l_{s2}) = \sigma^2(l_s) = \sigma^2(G_s + E_s) = \sigma^2(G_s) + \sigma^2(E_s) = \sigma^2(G_s) + V_E$. Keeping in mind that $\sigma(G, E) = 0$ per definition, gives $\sigma(l_s, G_s) = \sigma^2(G_s)$, as well as $\sigma(l_{s1}, G_{s2}) = \sigma(l_{s2}, G_{s1}) = \sigma(G_{s1}, G_{s2}) = \sigma\left(\frac{1}{2}G_f + \frac{1}{2}G_m, \frac{1}{2}G_f + \frac{1}{2}G_m\right) = \sigma\left(\frac{1}{2}G_f, \frac{1}{2}G_f\right) + \sigma\left(\frac{1}{2}G_f, \frac{1}{2}G_m\right) + \sigma\left(\frac{1}{2}G_m, \frac{1}{2}A_f\right) + \sigma\left(\frac{1}{2}G_m, \frac{1}{2}G_m\right) = \frac{1}{2}V_G + \frac{1}{2}\rho_G V_G$. The variance of the genetic liabilities in the parents equals $\sigma^2(G_m) = \sigma^2(G_f) = V_G$, and the covariance between fathers and mother equals $\sigma(G_m, G_f) = \rho_G V_G$. The covariance between the siblings and their parents subsequently follows as $\sigma(G_m, l_s) = \sigma(G_f, l_s) = \sigma(G_m, G_s) = \sigma(G_f, G_s) = \sigma\left(G_f, \frac{1}{2}G_m + \frac{1}{2}G_f\right) = \sigma\left(G_f, \frac{1}{2}G_m\right) + \sigma\left(G_f, \frac{1}{2}G_f\right) = \frac{1}{2}V_G + \frac{1}{2}\rho_G V_G$. For the complement of the first sibling, the following covariances are found:

- $\sigma(G_{c1}, l_{s1}) = \sigma(G_{c1}, G_{s1}) = \sigma(G_m + G_f - G_{s1}, G_{s1}) = \sigma(G_m, G_{s1}) + \sigma(G_f, G_{s1}) - \sigma^2(G_{s1}) = V_G + \rho_G V_G - V_G - \frac{1}{2}\rho_G V_G = \frac{1}{2}\rho_G V_G$, and
- $\sigma(G_{c1}, l_{s2}) = \sigma(G_{c1}, G_{s2}) = \sigma(G_m + G_f - G_{s1}, G_{s2}) = (G_m, G_{s2}) + \sigma(G_f, G_{s2}) - \sigma(G_{s1}, G_{s2}) = V_G + \rho_G V_G - \frac{1}{2}V_G - \frac{1}{2}\rho_G V_G = \frac{1}{2}V_G + \frac{1}{2}\rho_G V_G$, and
- $\sigma(G_{c1}, G_m) = \sigma(G_{c1}, G_f) = \sigma(G_m + G_f - G_{s1}, G_f) = \sigma(G_m, G_f) + \sigma^2(G_f) - \sigma(G_{s1}, G_f) = \rho_G V_G + V_G - \frac{1}{2}V_G - \frac{1}{2}\rho_G V_G = \frac{1}{2}V_G + \frac{1}{2}\rho_G$, and finally
- $\sigma^2(G_{c1}) = \sigma^2(G_m + G_f - G_{s1}) = \sigma^2\left(G_m + G_f - \frac{1}{2}G_m - \frac{1}{2}G_f - G_{residual}\right) = \sigma^2\left(\frac{1}{2}G_m, \frac{1}{2}G_f\right) + (-1)^2\sigma^2(G_{residual}) = V_G + \frac{1}{2}\rho_G V_G$

By this, all element were derived of $\sum(l_{s1}, G_{s1}, l_{s2}, G_{s2}, G_m, G_f, G_{c1})$, the 7x7 variance-covariance matrix of random families. The means of $l_{s1}, G_{s1}, l_{s2}, G_{s2}, G_m, G_f$ and G_{c1} all equal zero, noting that assortative mating does not change the mean genetic liability, because $E\left(\frac{1}{2}G_m + \frac{1}{2}G_f + G_{residual}\right) = E\left(\frac{1}{2}G_m\right) + E\left(\frac{1}{2}G_f\right) + E(G_{residual})$, also when $\sigma\left(\frac{1}{2}G_m, \frac{1}{2}G_f\right) > 0$.

1.2 Variances and covariances of genetic liabilities in families with at least one affected sibling

Assortative mating increases the variances of the phenotype l from the parental to the offspring generation with $\frac{1}{2}\rho_G V_G$. The increase in V_l results in a higher disease frequency in the offspring generation, because the liability threshold T remains the same. In order to estimate the reduction in variance in the affected siblings (assume $s1$ to be affected), the offspring population was first described in terms of the standard normal distribution, and than transformed back to the parental scale. The new disease frequency $K_{offspring}$ follows from $P(x > T \mid x \sim N(0, \sqrt{\sigma^2(l_s)}))$, and gives the mean phenotypic value of the affected siblings $s1$ on the standardized liability scale as $i_{offspring} = z_{offspring}/K_{offspring}$, where $z_{offspring}$ is the height of the standard normal distribution $N(0,1)$ at threshold $T_{offspring}$ with $K_{offspring} = P(x > T_{offspring} \mid x \sim N(0,1))$. Bulmer showed (page 153)[4] that the reduction of variation in affected siblings on the standardized liability scale equals $k_{offspring} = i_{offspring}(i_{offspring} - T_{offspring})$, and the variance reduction on the parental liability scale thus equals $k = k_{offspring}/\sigma^2(l_s)$. Tallis showed that given normality of G and l in the family members, the new variances and covariances are given by $\sigma(X, Y|s1\ affected) = \sigma(X, Y) - k\sigma(X, l_{s1})\sigma(Y, l_{s1})$, where X and Y represent all pairwise combinations of $l_{s1}, G_{s1}, l_{s2}, G_{s2}, G_m, G_f$ and G_{c1}.[9] By this, all element are defined of $\sum(l_{s1}, G_{s1}, l_{s2}, G_{s2}, G_m, G_f, G_{c1} \mid s1\ affected)$, the 7x7 variance-covariance matrix of families with one affected sibling. Given these variances and covariances, the means were derived as follows.

- $E(l_{s1}|s1\ aff) = i_{offspring}\sqrt{\sigma^2(l_s)}$
- $E(G_{s1}|s1\ aff) = \{\sigma^2(G_{s1})/\sigma^2(l_{s1})\} * E(l_{s1}|s1\ aff)$
- $E(l_{s2}|s1\ aff) = \{\sigma(l_{s1}, l_{s2})/\sigma^2(l_{s1})\} * E(l_{s1}|s1\ aff)$
- $E(G_{s2}|s1\ aff) = \{\sigma(G_{s1}, G_{s2})/\sigma^2(G_{s1})\} * E(G_{s1}|s1\ aff)$

- $E(G_m|s1\ aff) = E(G_f|s1\ aff) = \left\{(\frac{1}{2}V_G + \frac{1}{2}\rho_G V_G)/\sigma^2(G_s)\right\} * E(G_{s1}|s1\ aff)$, noting that $\frac{1}{2}V_G + \frac{1}{2}\rho_G V_G$ is the part of $\sigma^2(G_s)$ following from the parents contribution $\frac{1}{2}G_f + \frac{1}{2}G_m$.
- $E(G_{c1}|s1\ aff) = E(G_m|s1\ aff) + E(G_f|s1\ aff) - E(G_{s1}|s1\ aff)$

1.3 Variances and covariances of genetic liabilities in families with two affected siblings

To derive variances and covariances within families with two affected siblings, we take the estimates of families with one affected sibling as starting point. However, in order to apply Tallis' method to account of reduction in variance when selecting for an affected sibling, G and l need to be normally distributed in all family members. The distribution of l in the first sibling $s1$ is evidentially non-normal, because he is affected. Nevertheless, the distributions of G and l in the other family members are approximately normally distributed, which was illustrated by simulation (not shown) and can be intuitively understood as follows. The first sibling is affected when l_{s1} exceeds the threshold T. However, because l_{s1} is the sum of G_{s1} and E_{s1} and because G_{s1} and E_{s1} are independent, the violation of normality in $G_{s1|s1\ aff}$ is less than in $l_{s1|s1\ aff}$. In addition, the covariances between $G_{s1|s1\ aff}$ and G and l in the other family members are considerably smaller than 1. Hence, the distribution of G and l in all family members but sibling $s1$ are approximately normally distributed. Furthermore, note that the first and second sibling have equal genetic characteristics when they are both selected to be affected (except for their covariance with the complement, but this characteristic is not needed for this study). The variances and covariances are thus given by

$$\sigma(X, Y \mid s1\ affected\ \&\ s2\ affected) =$$
$$\sigma(X, Y \mid s1\ affected) - k_2\sigma(X, l_{s2}|\ s1\ affected)\sigma(Y, l_{s2}|\ s1\ affected),$$

where X and Y take all pairwise combinations of l_{s2}, G_{s2}, G_m, G_f and G_{c1}. The variance reduction k_2 is derived analoguously as k. The disease frequency in the second siblings $K_{s2\,|\,s1\ affected}$ follows from $P(x > T \mid x \sim N(E(l_{s2}|s1\ aff), \sqrt{\sigma^2(l_{s2}|s1\ affected)}))$, and gives the mean phenotypic value of the affected siblings $s2$ on the standardized liability scale as $i_{s2\,|\,s1\ affected} = z_{s2\,|\,s1\ affected}/K_{s2\,|\,s1\ affected}$, where $z_{s2\,|\,s1\ affected}$ is the height of the standard normal distribution $N(0,1)$ at threshold $T_{s2\,|\,s1\ affected}$ with $K_{s2\,|\,s1\ affected} = P(x > T_{s2\,|\,s1\ affected} \mid x \sim N(0,1))$. The reduction of variation in affected second siblings on the standardized liability scale equals $k_{s2\,|\,s1\ affected} = i_{s2\,|\,s1\ affected}(i_{s2\,|\,s1\ affected} - T_{s2\,|\,s1\ affected})$, and the variance reduction on the parental liability scale thus equals $k_2 = k_{s2\,|\,s1\ affected}/\sigma^2(l_{s2}|s1\ affected)$. This defines $\sum(l_{s2}, G_{s2}, G_m, G_f, G_{c1} \mid s1\ \&\ s2\ affected)$, the 5x5 variance-covariance matrix of families with two affected siblings (leaving out the first sibling $s1$). Given this variance-covariance matrix, the means were derived as:

- $E(l_{s2}|\ s1\ \&\ s2\ aff) = E(l_{s2}|\ s1\ aff) + i_{s2\,|\,s1\ affected}\sqrt{\sigma^2(l_{s2}|\ s1\ affected)}$

- $E(G_{s2}|\ s1\ \&\ s2\ aff) =$
  $E(G_{s2}|\ s1\ aff) + \{i_{s2\ |\ s1\ affected}\sqrt{\sigma^2(l_{s2}|\ s1\ affected)}\} *$
  $\sigma^2(G_{s2}|\ s1\ affected)/\sigma^2(l_{s2}|\ s1\ affected)$
- $E(G_m|s1\ \&\ s2\ aff) = E(G_f|s1\ \&\ s2\ aff) =$
  $E(G_f|s1\ aff) + \delta * \{\frac{1}{2}\sigma^2(G_m|s1\ aff) + \frac{1}{2}\sigma(G_m, G_f|s1\ aff)\}/\{\sigma^2(G_{s2}|s1\ aff)\}$, with $\delta =$
  $E(G_{s2}|\ s1\ \&\ s2\ aff) - E(G_{s2}|\ s1aff)$, while noting that $\frac{1}{2}\sigma^2(G_m|s1\ aff) + \frac{1}{2}\sigma(G_m, G_f|s1\ aff) +$
  $\frac{1}{2}V_{residual} = \sigma^2(G_{s2}|s1\ aff)$.
- $E(G_{c1}|\ s1\ \&\ s2\ aff) = E(G_m|\ s1\ \&\ s2\ aff) + E(G_f|\ s1\ \&\ s2\ aff) - E(G_{s1}|\ s1\ \&\ s2\ aff)$, where
  $E(G_{s1}|\ s1\ \&\ s2\ aff) = E(G_{s2}|\ s1\ \&\ s2\ aff)$.

1.4 Genetic liabilities of screened controls

Screened controls were selected from the offspring generation, i.e. after one generation of assortative mating. In order to apply the useful properties of the standard normal distribution, the liability scale was inverted to regard controls as 'cases', and later transformed back to the original scale of l in the parental generation. The population frequency of screened controls in the offspring generation is $K_{screened\ controls} = 1 - K_{offspring}$, which gives $i_{screened\ controls}$ and $k_{screened\ controls}$ as described previously in Section 1.2. The variation of genetic liabilities follows as $\sigma^2(G_{screened\ controls}) = \sigma^2(G_s) - \{k_{screened\ controls}/\sigma^2(l_s)\} * \sigma(l_s, G_s) * \sigma(l_s, G_s)$, and the mean as $E(G_{screened\ controls}) = -1 * \{\sigma^2(G_{s1})/\sigma^2(l_{s1})\} * i_{screened\ controls}\sqrt{\sigma^2(l_s)}$, where the term is multiplied by -1 to transform the mean back to the original parental liability scale of l.

2. Derivation of a single SNP's risk allele frequency in trio design

First, the risk allele frequencies were analytically derived for screened controls, cases, and cases with unaffected parents ('cases' and 'probands' are used interchangeably) (Section 2.1). Second, risk allele frequencies were derived for cases with affected siblings by applying the first set of derived frequencies and by considering IBD-sharing between cases and their siblings (Section 2.2). Third, all acquired estimates were applied to estimate risk allele frequencies in pseudocontrols (Section 2.3). Next we consider the impact of assortative mating (Section 2.4). To conclude, analytical derivations were validated with a simulation study (Table S5).

2.1 Risk allele frequencies in screened controls, cases, and cases with unaffected parents

This Section closely follows the work of Witte et al.[7] Assume the complex disease of interest has a population frequency $P(\mathrm{D}) = K$, and the locus of interest has risk allele B with frequency $P(\mathrm{B}) = p$, and non-risk allele b with frequency $P(\mathrm{b}) = 1 - p = q$. Given Hardy-Weinberg Equilibrium (HWE), the genotype frequencies are $P(\mathrm{bb}) = q^2$, $P(Bb) = 2pq$, and $P(\mathrm{BB}) = p^2$. Under a multiplicative risk model with relative risk of the heterozygote λ, the risk of disease given genotype $P(\mathrm{D|G})$ can be expressed as $P(\mathrm{D|bb}) = k_{bb}$, $P(D|Bb) = k_{bb}\lambda$, and $P(\mathrm{D|BB}) = k_{bb}\lambda^2$, with k_{bb} the disease risk in subjects with genotype bb. The probabilities of genotypes in cases is given by $P(\mathrm{G|D}) = P(\mathrm{D|G})P(\mathrm{G})/P(\mathrm{D})$, that is $P(\mathrm{bb|D}) = k_{bb}q^2/K$, $P(\mathrm{Bb|D}) = k_{bb}\lambda 2pq/K$, and $P(\mathrm{BB|D}) = k_{bb}\lambda^2 p^2/K$. Affected individuals, thus, have a risk allele frequency of $p_{case} = P(\mathrm{BB|D}) + \frac{1}{2}\,P(\mathrm{Bb|D})$. Analogously, the probabilities of genotypes in unaffected individuals (i.e., screened controls, sc) are given by $p(\mathrm{bb|ND}) = (1 - k_{bb})q^2/(1 - K)$, $P(\mathrm{Bb|ND}) = (1 - k_{bb}\lambda)2pq/(1 - K)$, and $P(\mathrm{BB|ND}) = (1 - k_{bb}\lambda^2)p^2/(1 - K)$, and they have a risk allele frequency of $p_{sc} = P(\mathrm{BB|ND}) + \frac{1}{2}\,P(\mathrm{Bb|ND})$, and non-risk allele frequency $q_{sc} = 1 - p_{sc}$. The offspring of unaffected parents will have genotype frequencies $P(\mathrm{G} \mid \text{parents unaffected})$ of $P(\mathrm{bb|pu}) = q_{sc}^2$, $P(\mathrm{Bb|pu}) = 2p_{sc}q_{sc}$, and $P(\mathrm{BB|pu}) = p_{sc}^2$, noting that HWE is re-established after one generation. Assuming no correlation between genotype and family environment, the $P(\mathrm{D|G})$ in offspring of screened controls are equal to $P(\mathrm{D|G})$ in the baseline population. The probabilities of genotypes in cases (proband) with unaffected parents, therefore, equal $P(\mathrm{bb|D,pu}) = k_{bb}q_{sc}^2/P(\mathrm{D|pu})$, $P(\mathrm{Bb|D,pu}) = k_{bb}\lambda 2p_{sc}q_{sc}/P(\mathrm{D|pu})$, and $P(\mathrm{BB|D,pu}) = k_{bb}\lambda^2 p_{sc}^2/P(\mathrm{D|pu})$, with $P(\mathrm{D|pu}) = k_{bb}q_{sc}^2 + k_{bb}\lambda 2p_{sc}q_{sc} + k_{bb}\lambda^2 p_{sc}^2$. Note that all can be expressed in terms of $p, q = 1 - p, K, \text{and } \lambda$ by realizing that $K = \sum_G P(D|G)P(G) = q^2k_{bb} + 2pqk_{bb}\lambda + p^2k_{bb}\lambda^2$, and thus $k_{bb} = K/(q^2 + 2pq\lambda + p^2\lambda^2)$. To take account of dominance effect, substitute λ with RR_{Bb} and λ^2 with RR_{BB} in the above.

2.2 Risk allele frequencies in proband with an affected sibling

To estimate the risk allele frequency in cases (proband) with affected siblings, the combined probabilities of genotypes in cases and their siblings is required:

$$\boldsymbol{P}(G_{case}, G_{sib}) = \boldsymbol{P}(G_c, G_s) = \begin{pmatrix} P(bb,bb) & P(bb,Bb) & P(bb,BB) \\ P(Bb,bb) & P(Bb,Bb) & P(Bb,BB) \\ P(BB,bb) & P(BB,Bb) & P(BB,BB) \end{pmatrix}$$

The rows of $\boldsymbol{P}(G_c, G_s)$ thus correspond to the three possible genotypes of cases and the columns to the three possible genotypes of their siblings. $\boldsymbol{P}(G_c, G_s)$ is the sum of four matrices: $\boldsymbol{P}(G_c, G_s| \, IBD = 0)$, $\boldsymbol{P}(G_c, G_s| \, IBD = 1(b))$, $\boldsymbol{P}(G_c, G_s| \, IBD = 1(B))$, and $\boldsymbol{P}(G_c, G_s| \, IBD = 2)$, all weighted by $0.25 = \boldsymbol{P}(IBD = 0) = \boldsymbol{P}(IBD = 1)/2 = \boldsymbol{P}(IBD = 2)$. To illustrate, the three row elements of $\boldsymbol{P}(G_s| \, G_c = Bb, IBD = 1(B))$ follow from basic Mendelian reasoning as $P(G_s = bb| \, G_c = Bb, IBD = 1(B)) = 0 * q_{NT|G_c=Bb}$ (the probability that the IDB-allele is b equals 0; the probability that the non-IBD allele is b depends on its frequency in the non-transmitted alleles from the parents given $G_c = Bb$), $P(G_s = Bb| \, G_c = Bb, IBD = 1(B)) = 1 * q_{NT|G_c=Bb}$, and $P(G_s = BB| \, G_c = Bb, IBD = 1(B)) = 1 * p_{NT|G_c=Bb}$ respectively, where $p_{NT|G_c}$ represents the frequency of B in the non-transmitted alleles from parents given G_c, and $q_{NT|G_c} = 1 - p_{p|G_c}$ the frequency of b. Note that $p_{NT|G_c}$ equals $p_{parents}$ when the parental generation is in HWE, however when the parents are unaffected they are not in HWE and derivation of $p_{NT|G_c}$ is slightly more elaborate (described in Appendix A). When IBD=0, the genotypes G_s depend on the distribution of the non-transmitted genotypes, which is also described in Appendix A. In this manner, the four matrices $\boldsymbol{P}(G_s| \, G_c, IBD)$ are defined as:

$$\boldsymbol{P}(G_s| \, G_c, IBD = 0) = \begin{pmatrix} P(NT = bb|G_c = bb) & P(NT = Bb|G_c = bb) & P(NT = BB|G_c = bb) \\ P(NT = bb|G_c = Bb) & P(NT = Bb|G_c = Bb) & P(NT = BB|G_c = Bb) \\ P(NT = bb|G_c = BB) & P(NT = Bb|G_c = BB) & P(NT = BB|G_c = BB) \end{pmatrix}$$

$$\boldsymbol{P}(G_s| \, G_c, IBD = \, 1(b)) = \begin{pmatrix} 2q_{NT|G_c=bb} & 2p_{NT|G_c=bb} & 0 \\ q_{NT|G_c=Bb} & p_{NT|G_c=Bb} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\boldsymbol{P}(G_s| \, G_c, IBD = 1(B)) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & q_{NT|G_c=Bb} & p_{NT|G_c=Bb} \\ 0 & 2q_{NT|G_c=BB} & 2p_{NT|G_c=BB} \end{pmatrix}$$

$$\boldsymbol{P}(G_s| \, G_c, IBD = 2) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

First, the allele frequency in cases with an affected sibling and random parents (in HWE) was derived, where $p_{NT} = p$ irrespective of G_c. Furthermore, define the diagonal matrix with the genotype probabilities in cases, and the diagonal matrix with the probabilities on an affected sibling given the siblings genotype as follows

$\boldsymbol{P}(G_c) = \mathrm{diag}\big(P(G|D)\big) = \mathrm{diag}(P(bb|D), P(Bb|D), P(BB|D))$, and
$\boldsymbol{P}(S = Affected|G_s) = \mathrm{diag}\big(P(D|G)\big) = \mathrm{diag}(P(D|bb), P(D|Bb), P(D|BB))$

Now estimate the combined genotype probabilities of cases and their sibling

$\boldsymbol{P}(\mathrm{G_c}, \mathrm{G_{s=Affected}}|\mathrm{IBD}) = \boldsymbol{P}(\mathrm{G_c}) * \boldsymbol{P}(G_s| \, G_c, IBD) * \boldsymbol{P}(S = Affected|\mathrm{G_s})$, (*Eq 1*) and

$$\boldsymbol{P}\left(G_c, G_{s=Affected}\right) = \sum\nolimits_{IBD} 0.25 * \boldsymbol{P}\left(\mathrm{G_c}, \mathrm{G}_{s=Affected} \middle| \mathrm{IBD}\right)$$

Because of the ascertainment on cases the elements of $\boldsymbol{P}(G_c, G_s)$ do not add up to 1. Hence, $\boldsymbol{P}\left(G_{case}, G_{S=Affected} | case, S = Affected\right) = \boldsymbol{P}(G_c, G_s) / \sum \boldsymbol{P}(G_c, G_s)$. The rows of $\boldsymbol{P}\left(G_{case}, G_{S=Affected} | case, S = Affected\right)$ add up to $P(G_c = \mathrm{bb} | case, S = Affected)$, $P(G_c = \mathrm{Bb} | case, S = Affected)$, and $P(G_c = \mathrm{BB} | case, S = Affected)$ respectively. This defines the risk allele frequency in cases with an affected sibling as $p_{case\,|\,S=Affected} = P(G_c = \mathrm{BB} | case, S = Affected) + \frac{1}{2} P(G_c = \mathrm{Bb} | case, S = Affected)$. Second, the allele frequency in cases with an affected sibling and unaffected parents was derived analoguously but with p_{NT} depending on G_c (see Appendix A in Section 2.5), and with $\boldsymbol{P}(G_c) = \mathrm{diag}\left(p(G|D, parents\ unaffected)\right)$.

2.3 Risk allele frequencies in pseudocontrols

Pseudo-control (pc) genotypes are the genomic complement genotypes from both parents not transmitted to their offspring. Allele frequencies in pseudocontrols depend on the genotypes of the cases selected, on the genotypes and disease statuses of the siblings and their IBD sharing with the cases. The genotype probabilities in pseudocontrols $P\left(\mathrm{G}_{pc} \middle| \mathrm{IBD}, \mathrm{G_c}, \mathrm{G_s}\right)$ were estimated as follows and the sum of these $4 * 3 * 3 = 36$ probabilities for a specific $\mathrm{G_{pc}}$ weighted by the probabilities of the genotypes in cases and controls and their IBD-sharing, gives $P(\mathrm{G_{pc}})$.

Define the matrices $\boldsymbol{P}\left(\mathrm{G}_{pc} \middle| \mathrm{IBD}, G_c, G_s\right)$ which has rows defined by genotypes of the cases and columns defined by the genotypes of the siblings

$$\begin{pmatrix} P\left(\mathrm{G}_{pc} \middle| \mathrm{IBD}, \mathrm{G_c} = bb, \mathrm{G_s} = bb\right) & P\left(\mathrm{G}_{pc} \middle| \mathrm{IBD}, \mathrm{G_c} = bb, \mathrm{G_s} = Bb\right) & P\left(\mathrm{G}_{pc} \middle| \mathrm{IBD}, \mathrm{G_c} = bb, \mathrm{G_s} = BB\right) \\ P\left(\mathrm{G}_{pc} \middle| \mathrm{IBD}, \mathrm{G_c} = Bb, \mathrm{G_s} = bb\right) & P\left(\mathrm{G}_{pc} \middle| \mathrm{IBD}, \mathrm{G_c} = Bb, \mathrm{G_s} = Bb\right) & P\left(\mathrm{G}_{pc} \middle| \mathrm{IBD}, \mathrm{G_c} = Bb, \mathrm{G_s} = BB\right) \\ P\left(\mathrm{G}_{pc} \middle| \mathrm{IBD}, \mathrm{G_c} = BB, \mathrm{G_s} = bb\right) & P\left(\mathrm{G}_{pc} \middle| \mathrm{IBD}, \mathrm{G_c} = BB, \mathrm{G_s} = Bb\right) & P\left(\mathrm{G}_{pc} \middle| \mathrm{IBD}, \mathrm{G_c} = BB, \mathrm{G_s} = BB\right) \end{pmatrix}$$

Given the parental genotype frequencies $P(G_p = bb)$, $P(G_p = Bb)$ and $P(G_p = BB)$, these $3\ \left(G_{pc}\right) * 4\ (IBD) = 12$ matrices follow from basic Mendelian reasoning and are displayed in Appendix B (Section 2.6). With these matrices the values of $P(\mathrm{G_{pc}} = bb)$, $P(\mathrm{G_{pc}} = Bb)$, and $P(\mathrm{G_{pc}} = BB)$ are separately estimated by

$$\boldsymbol{P}(G_{pc} | \mathrm{G_c}, \mathrm{G_s}, case, S = Affected) = \sum\nolimits_{IBD} 0.25 * \boldsymbol{P}(\mathrm{G_c}, \mathrm{G_{s=Affected}} | \mathrm{IBD}) \circ \boldsymbol{P}\left(\mathrm{G}_{pc} \middle| \mathrm{IBD}, G_c, G_s\right)$$

$$\boldsymbol{P}\left(\mathrm{G_{pc}}\right) = \sum \boldsymbol{P}(G_{pc} | \mathrm{G_c}, \mathrm{G_s}, case, S = Affected)$$

Where $\circ$ represent the Hadamard product of two matrices (i.e., when $A = B \circ C$, than $a_{ij} = b_{ij} * c_{ij}$). The probabilities $P(\mathrm{G_{pc}} = bb)$, $P(\mathrm{G_{pc}} = Bb)$, and $P(\mathrm{G_{pc}} = BB)$ do not add up to 1, because they are defined in terms of the full population. Therefore, $P\left(\mathrm{G_{pc}} \mid case, S = Affected\right)$ equal $P(\mathrm{G_{pc}})/$

$\sum_{\mathrm{G_{pc}}} P(\mathrm{G_{pc}})$. This yields the risk allele frequency in pseudocontrols from cases with affected siblings as $p_{pc\,|\,S=Affected} = P\big(\mathrm{G_{pc}} = BB\big) + \frac{1}{2}P(\mathrm{G_{pc}} = Bb)$.

The following variations yield the estimation for the other sets of pseudocontrols. (i) To estimate p_{pc} (without conditioning on affected siblings), replace $\boldsymbol{P}(\mathrm{G_c}, \mathrm{G_{s=Affected}}|\mathrm{IBD})$ by $\boldsymbol{P}(\mathrm{G_c}, \mathrm{G_s}|\mathrm{IBD})$ by substituting the diagonal matrix $\boldsymbol{P}(S = Affected|G_s)$ in the above for the identity matrix $\mathbb{I}$. (ii) To estimate $p_{pc|P=unaffected}$, adjust the parental genotype probabilities accordingly (no longer in HWE) and set $\boldsymbol{P}(G_c) = \mathrm{diag}\big(p(G|D, parents\ unaffected)\big)$. (iii) To estimated $p_{pc|S=Affected\,\&\,P=unaffected}$, combine the substitutions described in (i) and (ii).

2.4 Assortative mating

The impact of assortative mating on a single locus is expressed as the non-random mating fraction α of parents with similar genotypes. The next generation has the following frequencies[8]

$\boldsymbol{P}(G_c = bb|\ assortative\ mating\ parents) = (1-\alpha)q^2 + \alpha(q^2 + \frac{1}{2}pq)$,
$\boldsymbol{P}(G_c = Bb|\ assortative\ mating\ parents) = (1-\alpha)2pq + \alpha pq$, and
$\boldsymbol{P}(G_c = BB|\ assortative\ mating\ parents) = (1-\alpha)p^2 + \alpha(p^2 + \frac{1}{2}pq)$,

when the parental generation is in HWE, and with p the parental frequency of B and q of b. The genotype probabilities of affected siblings are given by $\boldsymbol{P}(\mathrm{G}|\mathrm{D}, a.m.parents) = P(\mathrm{D}|\mathrm{G})P(\mathrm{G}|a.m.\ parents)/P(\mathrm{D})$ analoguous to Section 2.1. Substituting these as $\boldsymbol{P}(\mathrm{G_c})$ in Eq 1 in Section 2.2

$$\boldsymbol{P}(\mathrm{G_c}, \mathrm{G_s}|\mathrm{IBD}, a.m.parents) = \boldsymbol{P}(\mathrm{G_c}) * \boldsymbol{P}(G_s|\ G_c, IBD) * \mathbb{I},$$

and following the other steps in Sections 2.1 and 2.2 gives the frequencies of cases and pseudocontrol of parents with assortative mating (not selecting of disease-status of parents or siblings). Note that assortative mating changes the probabilities of the combined genotypes of parents, which is described in Appendix A (Section 2.5).

2.5 Appendix A: allele and genotype frequencies of non-transmitted alleles

When the parents are unaffected, they are not in HWE, in which case the non-transmitted allele and genotype frequencies are dependent on the case's (proband's) genotype G_c. These non-transmitted allele and genotype frequencies are needed to derive the combined probabilities of genotypes in cases and their sibling $\boldsymbol{P}(G_c, G_s)$. (Note that these non-transmitted alleles are not the pseudocontrols of interest.) Suppose the genotypes in the parents have frequencies $P(G_p = bb)$, $P(G_p = Bb)$ and $P(G_p = BB)$. The distribution of the genotypes of pairs of parents with a genotype correlation (non-random mating fraction) α is given by

$$\boldsymbol{P}\left(G_{father}G_{mother}\right) = \begin{pmatrix} P(G_f = bb, G_m = bb) \\ P(G_f = bb, G_m = Bb) \\ P(G_f = bb, G_m = BB) \\ P(G_f = Bb, G_m = bb) \\ P(G_f = Bb, G_m = Bb) \\ P(G_f = Bb, G_m = BB) \\ P(G_f = BB, G_m = bb) \\ P(G_f = BB, G_m = Bb) \\ P(G_f = BB, G_m = BB) \end{pmatrix} = \begin{pmatrix} (1-\alpha)P\left(G_p = bb\right)P\left(G_p = bb\right) + \alpha P\left(G_p = bb\right) \\ (1-\alpha)P\left(G_p = bb\right)P\left(G_p = Bb\right) \\ (1-\alpha)P\left(G_p = bb\right)P\left(G_p = BB\right) \\ (1-\alpha)P\left(G_p = Bb\right)P\left(G_p = bb\right) \\ (1-\alpha)P\left(G_p = Bb\right)P\left(G_p = Bb\right) + \alpha P\left(G_p = Bb\right) \\ (1-\alpha)P\left(G_p = Bb\right)P\left(G_p = BB\right) \\ (1-\alpha)P\left(G_p = BB\right)P\left(G_p = bb\right) \\ (1-\alpha)P\left(G_p = BB\right)P\left(G_p = Bb\right) \\ (1-\alpha)P\left(G_p = BB\right)P\left(G_p = BB\right) + \alpha P\left(G_p = BB\right) \end{pmatrix}$$

The distributions of the genotypes of pairs of parents conditional on their offspring G_c are proportional to the pairwise multiplications of the probability of these parental genotypes times the probability of getting offspring with G_c, that is

$$\widetilde{\boldsymbol{P}}\left(G_{father}G_{mother}|G_c = bb\right) = \boldsymbol{P}\left(G_{father}G_{mother}\right)^{*}(1\ \ 0.5\ \ 0\ \ 0.5\ \ 0.25\ \ 0\ \ 0\ \ 0\ \ 0)^T$$
$$\widetilde{\boldsymbol{P}}\left(G_{father}G_{mother}|G_c = Bb\right) = \boldsymbol{P}\left(G_{father}G_{mother}\right) * (0\ \ 0.5\ \ 1\ \ 0.5\ \ 0.5\ \ 0.5\ \ 1\ \ 0.5\ \ 0)^T$$
$$\widetilde{\boldsymbol{P}}\left(G_{father}G_{mother}|G_c = BB\right) = \boldsymbol{P}\left(G_{father}G_{mother}\right) * (0\ \ 0\ \ 0\ \ 0\ \ 0.25\ \ 0.5\ \ 0\ \ 0.5\ \ 1)^T$$

The probabilities of non-transmitted (NT) genotypes are proportional to the sum of the combined parental genotypes resulting in this NT genotype, that is

$$\widetilde{\boldsymbol{P}}(NT = bb|G_c = bb) = (1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0) * \widetilde{\boldsymbol{P}}\left(G_{father}G_{mother}|G_c = bb\right)$$
$$\widetilde{\boldsymbol{P}}(NT = Bb|G_c = bb) = (0\ 1\ 0\ 1\ 0\ 0\ 0\ 0\ 0) * \widetilde{\boldsymbol{P}}\left(G_{father}G_{mother}|G_c = bb\right)$$
$$\widetilde{\boldsymbol{P}}(NT = BB|G_c = bb) = (1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0) * \widetilde{\boldsymbol{P}}\left(G_{father}G_{mother}|G_c = bb\right)$$
$$\widetilde{\boldsymbol{P}}(NT = bb|G_c = Bb) = (0\ 1\ 0\ 1\ 0\ 0\ 0\ 0\ 0) * \widetilde{\boldsymbol{P}}\left(G_{father}G_{mother}|G_c = Bb\right)$$
$$\widetilde{\boldsymbol{P}}(NT = Bb|G_c = Bb) = (0\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 0) * \widetilde{\boldsymbol{P}}\left(G_{father}G_{mother}|G_c = Bb\right)$$
$$\widetilde{\boldsymbol{P}}(NT = BB|G_c = Bb) = (0\ 0\ 0\ 0\ 0\ 1\ 0\ 1\ 0) * \widetilde{\boldsymbol{P}}\left(G_{father}G_{mother}|G_c = Bb\right)$$
$$\widetilde{\boldsymbol{P}}(NT = bb|G_c = BB) = (0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0) * \widetilde{\boldsymbol{P}}\left(G_{father}G_{mother}|G_c = BB\right)$$
$$\widetilde{\boldsymbol{P}}(NT = Bb|G_c = BB) = (0\ 0\ 0\ 0\ 0\ 1\ 0\ 1\ 0) * \widetilde{\boldsymbol{P}}\left(G_{father}G_{mother}|G_c = BB\right)$$
$$\widetilde{\boldsymbol{P}}(NT = BB|G_c = BB) = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1) * \widetilde{\boldsymbol{P}}\left(G_{father}G_{mother}|G_c = BB\right)$$

Scaling gives the exact probabilities of the NT genotypes: $P(NT = bb|G_c = bb) = \tilde{P}(NT = bb|G_c = bb)/\left(\tilde{P}(NT = bb|G_c = bb) + \tilde{P}(NT = Bb|G_c = bb) + \tilde{P}(NT = BB|G_c = bb)\right)$ etc. The allele frequencies $p_{NT|G_c}$ follow directly from the NT genotype frequencies.

2.6 Appendix B: pseudocontrol genotypes conditional on IBD, G_c and G_s

Define the matrices $\boldsymbol{P}\left(\mathrm{G}_{pc}\middle|\mathrm{IBD}, G_c, G_s\right)$ as

$$\begin{pmatrix} P(\mathrm{G}_{pc}|\mathrm{IBD}, \mathrm{G_c} = bb, \mathrm{G_s} = bb) & P(\mathrm{G}_{pc}|\mathrm{IBD}, \mathrm{G_c} = bb, \mathrm{G_s} = Bb) & P(\mathrm{G}_{pc}|\mathrm{IBD}, \mathrm{G_c} = bb, \mathrm{G_s} = BB) \\ P(\mathrm{G}_{pc}|\mathrm{IBD}, \mathrm{G_c} = Bb, \mathrm{G_s} = bb) & P(\mathrm{G}_{pc}|\mathrm{IBD}, \mathrm{G_c} = Bb, \mathrm{G_s} = Bb) & P(\mathrm{G}_{pc}|\mathrm{IBD}, \mathrm{G_c} = Bb, \mathrm{G_s} = BB) \\ P(\mathrm{G}_{pc}|\mathrm{IBD}, \mathrm{G_c} = BB, \mathrm{G_s} = bb) & P(\mathrm{G}_{pc}|\mathrm{IBD}, \mathrm{G_c} = BB, \mathrm{G_s} = Bb) & P(\mathrm{G}_{pc}|\mathrm{IBD}, \mathrm{G_c} = BB, \mathrm{G_s} = BB) \end{pmatrix}$$

Given the parental genotype frequencies $P(G_p = bb)$, $P(G_p = Bb)$ and $P(G_p = BB)$, these $3 * 4 = 12$ matrices follow from basic Mendelian reasoning. Note that IBD=0 (between cases and their siblings) indicates that the pseudocontrol shares both alleles with the sibling; IBD=1 indicates that the pseudocontrol shares the non-IBD allele with the sibling; and IBD=2 indicates that the pseudocontrol and sibling share no alleles. Alleles in the pseudocontrols not shared with the sibling come from the parents with the probabilities derived in Appendix A (Section 2.5). The $\boldsymbol{P}(\mathrm{G}_{pc}|\mathrm{IBD})$ are thus defined as:

$$\boldsymbol{P}(\mathrm{G}_{pc} = bb|\mathrm{IBD} = 0) = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

$$\boldsymbol{P}(\mathrm{G}_{pc} = bb|\mathrm{IBD} = \mathrm{b}) = \begin{pmatrix} q_{NT|G_c=bb} & 0 & 0 \\ q_{NT|G_c=Bb} & 0 & 0 \\ q_{NT|G_c=BB} & 0 & 0 \end{pmatrix}$$

$$\boldsymbol{P}(\mathrm{G}_{pc} = bb|\mathrm{IBD} = \mathrm{B}) = \begin{pmatrix} q_{NT|G_c=bb} & q_{NT|G_c=bb} & 0 \\ q_{NT|G_c=Bb} & q_{NT|G_c=Bb} & 0 \\ q_{NT|G_c=BB} & q_{NT|G_c=BB} & 0 \end{pmatrix}$$

$$\boldsymbol{P}(\mathrm{G}_{pc} = bb|\mathrm{IBD} = 2) = \begin{pmatrix} P(NT = bb|G_c = bb) & P(NT = bb|G_c = bb) & P(NT = bb|G_c = bb) \\ P(NT = bb|G_c = Bb) & P(NT = bb|G_c = Bb) & P(NT = bb|G_c = Bb) \\ P(NT = bb|G_c = BB) & P(NT = bb|G_c = BB) & P(NT = bb|G_c = BB) \end{pmatrix}$$

$$\boldsymbol{P}(\mathrm{G}_{pc} = Bb|\mathrm{IBD} = 0) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

$$\boldsymbol{P}(\mathrm{G}_{pc} = Bb|\mathrm{IBD} = \mathrm{b}) = \begin{pmatrix} p_{NT|G_c=bb} & q_{NT|G_c=bb} & q_{NT|G_c=bb} \\ p_{NT|G_c=Bb} & q_{NT|G_c=Bb} & q_{NT|G_c=Bb} \\ p_{NT|G_c=BB} & q_{NT|G_c=BB} & q_{NT|G_c=BB} \end{pmatrix}$$

$$\boldsymbol{P}(\mathrm{G}_{pc} = Bb|\mathrm{IBD} = \mathrm{B}) = \begin{pmatrix} p_{NT|G_c=bb} & p_{NT|G_c=bb} & q_{NT|G_c=bb} \\ p_{NT|G_c=Bb} & p_{NT|G_c=Bb} & q_{NT|G_c=Bb} \\ p_{NT|G_c=BB} & p_{NT|G_c=BB} & q_{NT|G_c=BB} \end{pmatrix}$$

$$\boldsymbol{P}(\mathrm{G}_{pc} = Bb|\mathrm{IBD} = 2) = \begin{pmatrix} P(NT = Bb|G_c = bb) & P(NT = Bb|G_c = bb) & P(NT = Bb|G_c = bb) \\ P(NT = Bb|G_c = Bb) & P(NT = Bb|G_c = Bb) & P(NT = Bb|G_c = Bb) \\ P(NT = Bb|G_c = BB) & P(NT = Bb|G_c = BB) & P(NT = Bb|G_c = BB) \end{pmatrix}$$

$$\boldsymbol{P}(\mathrm{G}_{pc} = BB|\mathrm{IBD} = 0) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\boldsymbol{P}(\mathrm{G}_{pc} = BB|\mathrm{IBD} = \mathrm{b}) = \begin{pmatrix} 0 & p_{NT|G_c=bb} & p_{NT|G_c=bb} \\ 0 & p_{NT|G_c=Bb} & p_{NT|G_c=Bb} \\ 0 & p_{NT|G_c=BB} & p_{NT|G_c=BB} \end{pmatrix}$$

$$\boldsymbol{P}(\mathrm{G}_{pc} = BB|\mathrm{IBD} = \mathrm{B}) = \begin{pmatrix} 0 & 0 & p_{NT|G_c=bb} \\ 0 & 0 & p_{NT|G_c=Bb} \\ 0 & 0 & p_{NT|G_c=BB} \end{pmatrix}$$

$$\boldsymbol{P}(\mathrm{G}_{pc} = BB|\mathrm{IBD} = 2) = \begin{pmatrix} P(NT = BB|G_c = bb) & P(NT = BB|G_c = bb) & P(NT = BB|G_c = bb) \\ P(NT = BB|G_c = Bb) & P(NT = BB|G_c = Bb) & P(NT = BB|G_c = Bb) \\ P(NT = BB|G_c = BB) & P(NT = BB|G_c = BB) & P(NT = BB|G_c = BB) \end{pmatrix}$$

References

1. Golan, D., Lander, E.S., and Rosset, S. (2014). Measuring missing heritability: Inferring the contribution of common variants. Proc. Natl. Acad. Sci. U. S. A. *111*, E5272–E5281.

2. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

3. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.

4. Bulmer, M. (1985). The mathematical theory of quantitative genetics (Oxford: Clarendon press).

5. Lynch, M., and Walsh, B. (1998). Genetics and analysis of quantitative traits. (Sunderland: Sinauer),.

6. R Core Team (2015). R: A Language and Environment for Statistical Computing.

7. Witte, J.S., Visscher, P.M., and Wray, N.R. (2014). The contribution of genetic variants to disease depends on the ruler. Nat. Rev. Genet. *15*, 765–776.

8. Falconer, D., and Mackay, T. (1996). Introduction to quantitative genetics (Essex: Longman).

9. Tallis, G.M. (1987). Ancestral covariance and the Bulmer effect. Theor. Appl. Genet. *73*, 815–820.