

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

Wouter J. Peyrot,<sup>1,2,\*</sup> Dorret I. Boomsma,<sup>3</sup> Brenda W.J.H. Penninx,<sup>1</sup> and Naomi R. Wray<sup>2,\*</sup>

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.<sup>1</sup> 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,<sup>2</sup> 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,<sup>3</sup> whereas only two loci were found in one study on major depressive disorder (MDD [MIM: 608516])<sup>4</sup> but none in another,<sup>5</sup> no loci

for attention-deficit/hyperactivity disorder (ADHD [MIM: 143465]),<sup>6</sup> and only single-study genome-wide significant loci for autism spectrum disorder (ASD [MIM: 209850]).<sup>7–9</sup> 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<sup>10,11</sup> (although these are expected to explain only a small proportion of variation),<sup>12</sup> lower family-based heritability of MDD (~0.4 versus ~0.8 for SCZ, ASD, and ADHD),<sup>13</sup> and higher prevalence and greater heterogeneity of MDD.<sup>14</sup> 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.<sup>15</sup>

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.<sup>16,17</sup> 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,<sup>18–21</sup> 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

<sup>1</sup>Department of Psychiatry, VU University Medical Center & GGZ inGeest, Amsterdam 1081 HL, the Netherlands; <sup>2</sup>Queensland Brain Institute, University of Queensland, Brisbane 4072, Australia; <sup>3</sup>Department of Biological Psychology, VU University Amsterdam, Amsterdam 1081 BT, the Netherlands

\*Correspondence: [w.peyrot@ggzingeest.nl](mailto:w.peyrot@ggzingeest.nl) (W.J.P.), [naomi.wray@uq.edu.au](mailto:naomi.wray@uq.edu.au) (N.R.W.)

<http://dx.doi.org/10.1016/j.ajhg.2015.12.017>. ©2016 by The American Society of Human Genetics. All rights reserved.

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.<sup>22–25</sup> 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<sup>23</sup> and of 0.26 in parents of ASD probands.<sup>22</sup> For ADHD a spouse correlation of 0.11 on the ADHD index in population samples has been reported.<sup>25</sup> 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.<sup>2,16</sup> 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<sup>26</sup>

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

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<sup>2,27</sup>

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

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,  $F = N_{\text{false controls}} / (N_{\text{false controls}} + N_{\text{true controls}}) = N_{\text{false controls}} / N_{\text{controls}}$ . We closely followed the derivations of Golan et al. (paragraphs 1.2 and 1.3 of their Supplemental Materials)<sup>27</sup> to derive an updated equation (Table S1) validated by simulation (Table S2),

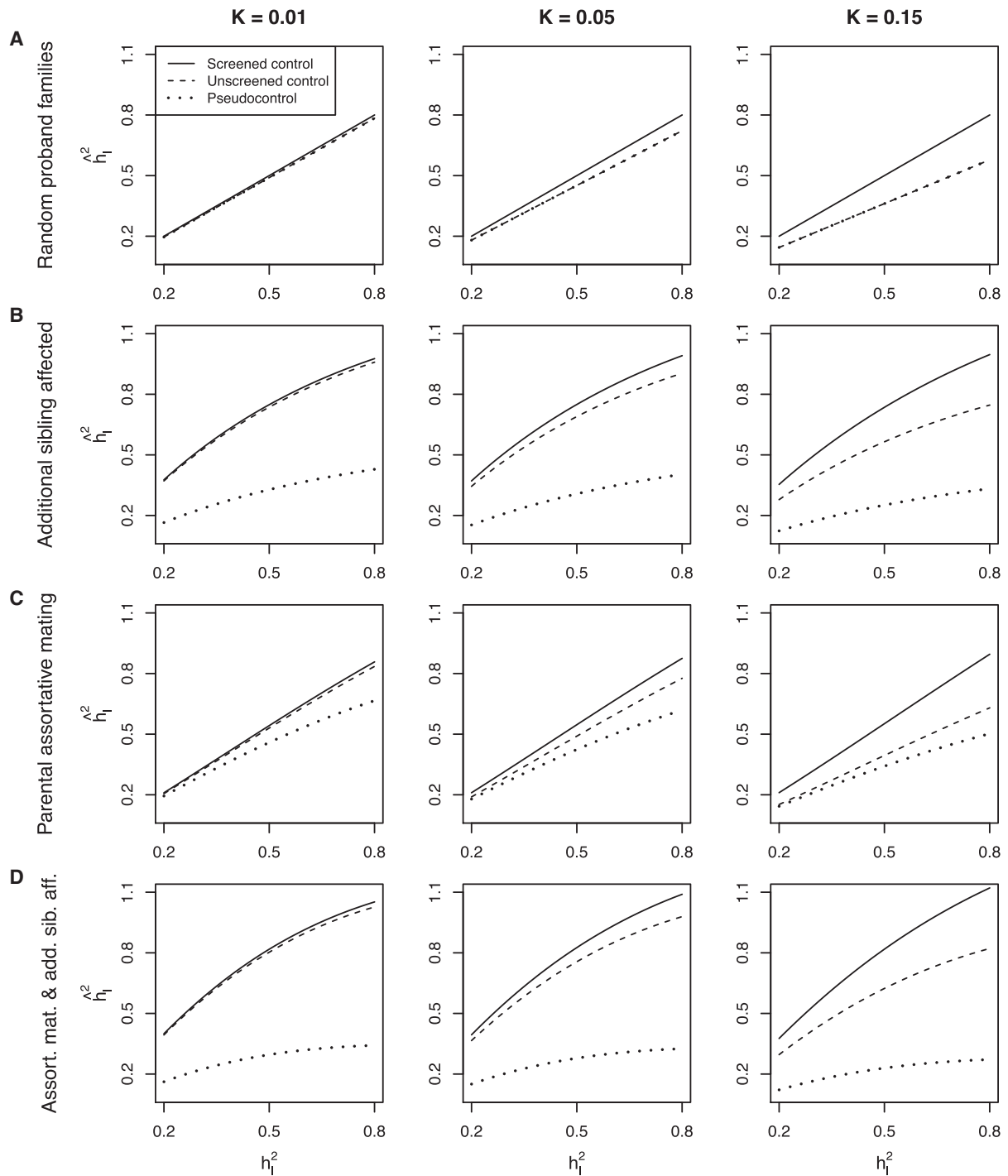
$$\hat{h}_l^2 = \hat{h}_{occ}^2 \frac{K^2(1-K)^2}{P(1-P)(1-F)^2 z^2}, \quad (\text{Equation 3})$$

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.<sup>28</sup> Under the liability-threshold model, individuals are deemed affected when their liability ( $l$ ) is larger than threshold ( $T$ ) such that  $P(l > T | l \sim N(0, 1)) = K$ . Parental assortative mating was taken into account by parameterizing a spouse liability correlation of  $\rho_l$  and genetic correlation of  $\rho_G = h_l^2 \rho_l$ .<sup>28</sup> 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<sup>29</sup> in related individuals described by Tallis.<sup>30</sup> 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  $\hat{h}_1^2$  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  $h_1^2$  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  $\hat{h}_1^2$  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**

K	$h_i^2$ Parents	Mean Genetic Liability ( $E(G)$ )				$\hat{h}_i^2$ Assessed from Proband		
		Case	Control		Pseudo	Screened	Unscreened	Pseudo
			Screened	Unscreened				
<b>Random Proband Families</b>								
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
<b>Additional Sibling Affected</b>								
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
<b>Parental Assortative Mating</b>								
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
<b>Additional Sibling Affected and Parental Assortative Mating</b>								
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  $\hat{h}_i^2$  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  $h_i^2 = 0.5$ . 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  $h_i^2$  (for example,  $\hat{h}_i^2 = 0.75$  for  $K = 0.05$  and  $h_i^2 = 0.5$ ). 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  $\hat{h}_i^2$  (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  $h_i^2$ ; this is, however, a well-known consequence of assortative mating and is not restricted to the trio design ( $V_{G,offspring} = V_{G,parents} + (1/2)\rho_{G,parents}V_{G,parents}$ ).<sup>29</sup> The most pronounced difference between screened and pseudocontrol subjects is found for probands with an additional affected sibling in the context of parental assortative

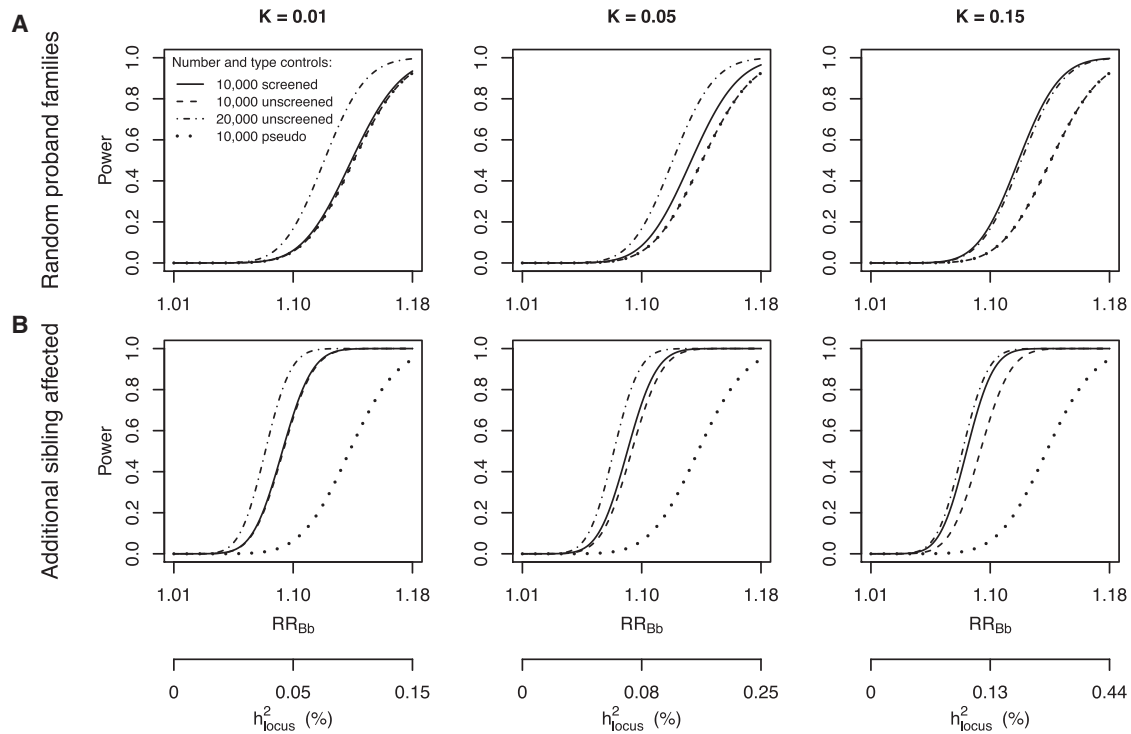
mating (Figure 1D; Table 1, additional sibling affected and parental assortative mating).

### 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  $\chi^2$  test statistic. This  $NCP$  is expressed in terms of sample size  $N$ , proportion of case subjects in the study  $\nu$ , the allele frequency in case subjects  $p_{case}$ , the allele frequency in control subjects  $p_{control}$ , and the mean allele frequency in the sample  $\bar{p} = \nu p_{case} + (1 - \nu)p_{control}$  as

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

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  $h_{locus}^2 \approx 2p(1-p)(RR_{Bb} - 1)^2/i^2$ , 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  $RR_{Bb} = P(\text{Disease} | Bb) / P(\text{Disease} | bb)$ , and relative risk of the homozygotes  $RR_{BB} = P(\text{Disease} | BB) / P(\text{Disease} | bb)$ .<sup>31,32</sup> When control subjects are screened, power follows from  $p_{case} = k_{bb}RR_{Bb}p(1 + p(RR_{Bb} - 1))/K$ , where  $k_{bb} = P(\text{Disease} | bb) = K/(q^2 + 2pqRR_{Bb} + p^2RR_{BB})$  and  $p_{control} = ((1 - k_{bb}RR_{Bb})p(1 - p) + (1 - k_{bb}RR_{BB})p^2)/(1 - K)$ ,<sup>32</sup> which agrees with the genetic power calculator of Purcell et al.<sup>33</sup> 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  $\rho_{locus} = \rho_{liability}h_{locus}^2$ .<sup>28</sup> 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 ( $\rho_{locus} = 0.3 \times$

$0.01 = 0.003$ ).<sup>28</sup> 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  $h_{locus}^2 \approx 2pq(RR_{Bb} - 1)^2/i^2$  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  $K$ .<sup>32</sup> 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 pseudocontrol 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.

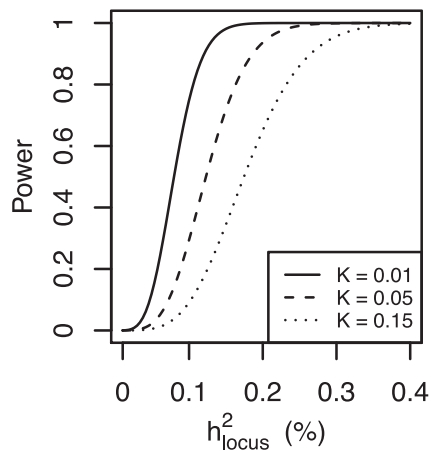
**Table 2. Maximum Power Difference between Trio Design and Screened Control Subject Studies with 20,000 Subjects**

K	$RR_{Bb}$	Allele Frequencies			Power (n = 20,000)		n (Power = 0.8)	
		Proband	Pseudo	Screened	Pseudo	Screened	Pseudo	Screened
<b>Proband from Random Proband Families</b>								
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
<b>Proband from Families with an Additional Affected Sibling</b>								
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  $p_{\text{proband } B} = 0.248$  and  $p_{\text{proband } A} = 0.231$ , 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 ( $p_{\text{pseudocontrol } B} = 0.215$  and  $p_{\text{pseudocontrol } A} = 0.2$ ). 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 than 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.<sup>31</sup> 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,<sup>34</sup> to determine accurately phased haplotypes,<sup>34</sup> and to undertake parent-of-origin analyses implied by a hypothesis of parental imprinting.<sup>35</sup> Trio designs have also been considered for detection of gene-environment interaction.<sup>36,37</sup> In the pre-GWAS era, trio designs were recommended to protect against potential bias from population stratification,<sup>1</sup> and although this quality is also sometimes promoted for trio GWAS, with genome-wide SNP data, other strategies, such as genomic principal components<sup>38</sup> or mixed model association analysis,<sup>39</sup> 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  $K \sim 0.15$ )<sup>13,40</sup> 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.<sup>7</sup> However, the proportion of multiplex families is often not reported, as is the case for the family-based studies,<sup>41–43</sup> contributing to the most recent ADHD meta-analysis,<sup>6</sup> 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 simple versus multiplex and simulation results of Klei et al. in analyses of ASD samples.<sup>44</sup>

- (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<sup>22–25</sup> 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.).<sup>14</sup>

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.<sup>45</sup>

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.<sup>18</sup> Li et al.,<sup>19</sup> Risch and Teng,<sup>46</sup> and Risch<sup>47</sup> 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.<sup>20</sup> 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.<sup>21</sup> 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.<sup>33</sup> or the Quanto software from Gauderman.<sup>48</sup> 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., Bradford, 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., Benyamini, 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., Malter, 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.