The Impact on Estimations of Genetic Correlations of the Use, in Genome Wide Case-Control Studies, of Super-Normal, Unscreened and Family-History Screened Controls

Kenneth S. Kendler MD

Chris Chatzinakos PhD

Silviu-Alin Bacanu PhD

Text of 1979 Words, 3 figures, Appendix

Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA.

Department of Psychiatry, Virginia Commonwealth University, Richmond VA, USA.

Corresponding author: Kenneth S. Kendler, MD, Virginia Institute for Psychiatric and Behavioral Genetics of VCU, Box 980126, Richmond, VA 23298-0126; Phone: (804) 828-8590; FAX: (804) 828-1471; E-mail: Kenneth.Kendler@vcuhealth.org

Funding Information: This work was supported in part by NIH grants P50-AA022537, R01-AA026750 and R01-MH114593.

Author Contributions: KSK and SB conceived of the project. CC wrote the simulation software and ran the subsequent analyses. KSK and SB interpreted the data and wrote the manuscript.

Conflict of Interests: The authors report no competing interests.

ABSTRACT

Traditionally, in *normal* case-control studies of disorder A, individuals with disorder A are screened-out of controls. However, in genome wide association (GWA) studies, controls are sometimes *unscreened* or screened for disorder A and disorder B, producing *super-normal* controls. Using simulations, we examine how the observed genetic correlations between two disorders (A and B) are influenced by the use of unscreened, normal, and super-normal controls. Normal controls produce unbiased estimates of the genetic correlation. However, unscreened and super-normal controls both bias upward the genetic correlations. The strength of the bias increases with increasing population prevalences for the two disorders. With super-normal controls, the magnitude of bias is stronger when the true genetic correlation is low. The opposite is seen with the use of unscreened controls. Adding screening of first-degree relatives of controls substantially increases the bias in genetic correlations with super-normal controls but has minimal impact when normal controls are used.

Genetic correlations calculated from genome wide association (GWA) studies are widely used to quantify the genetic contribution to disorder comorbidity (Anttila et al., 2018; Bulik-Sullivan et al., 2015; Cross-Disorder Group of the Psychiatric Genomics Consortium(PGC-CDG), 2013). In this report, we examine one area of concern in such studies – that estimates of genetic correlations (r_g) between disorders assessed using case-control GWA studies can be biased by the method of control selection.

In epidemiology, the proper selection of controls in case-control studies is a subject of considerable concern and debate (Hodge, Subaran, Weissman, & Fyer, 2012; Lubin & Gail, 1984; Wacholder, McLaughlin, Silverman, & Mandel, 1992; Wacholder, Silverman, McLaughlin, & Mandel, 1992a; Wacholder, Silverman, McLaughlin, & Mandel, 1992b). The principle defended by most methodologists is that controls should resemble cases in all characteristics *except* not having the disorder for which cases are selected – here disorder A. We term controls selected on this principle *normal* controls. Nonetheless, the collection of controls in many biomedical studies are neither properly done nor adequately reported (Lopez, Scheutz, Errboe, & Baelum, 2007; Malay & Chung, 2012).

In genetic epidemiology, some studies utilize *super-normal* controls screened not only for the disorder being studied but also for other often-related disorders (Chen et al., 2005; Schwartz & Susser, 2011). That is, if cases are selected for having disorder A, controls are selected for not having disorder A and for not having an additional screened-out disorder B (or multiple screened-out disorders B, C and D). In family studies, the use of super-normal controls produces spurious co-aggregation between disorders A and B the magnitude of which increases with increasing population prevalence of disorder B (Kendler, 1990). Case-control GWA studies have also utilized super-normal controls (Ikeda et al., 2011; Otowa et al., 2016; Schwartz & Susser, 2010; Sullivan et al., 2008).

Controls in case-control studies can be selected based not only on their own phenotype but also on the phenotype of close relatives (Wray et al., 2012). When normal controls or super-normal controls

are screened in this fashion, we call these *normal-FH* and *super-normal-FH* controls, respectively, where FH stands for "family history."

Finally, because screening of potential controls can be effortful and expensive, *unscreened* controls can be used (Kirov et al., 2009; O'Donovan et al., 2008). So, if cases are selected for having disorder A, investigators might take a population-based sample and not screen it so that it contains cases of A at approximately population prevalence.

Recently, van Rheenen et al. (van Rheenen, Peyrot, Schork, Lee, & Wray, 2019) provided a thorough review of the properties of GWA derived genetic correlations, finding the measure to be robust to many misspecifications of the causal model, case control proportions, and stratification. However, they provide two examples of what they call "double-screening control cohorts," equivalent to what we have termed super normal controls. Consistent with results found in family studies (Kendler, 1990), they report minimal bias when the two disorders (schizophrenia and bipolar illness in their example) are relatively rare. However, when the two disorders are common (asthma and hay fever in their example), r_g between the two disorders is substantially inflated.

In this report, using simulation, we systematically explore the impact of different approaches to control selection in GWA studies on the estimation of r_g, addressing four specific questions. First, as predicted by theory, does the use of normal controls in GWA studies produce unbiased estimates of r_g? Second, attempting to replicate and expand upon the results presented by van Rheenen et al. (van Rheenen et al., 2019), do we find that the use of super-normal controls in GWA studies upwardly bias estimates of r_g between disorders A and B (r_{gAB})? If so, what is the effect on this bias of the true value of r_g and the population prevalences of A and B? Third, if we use normal-FH or super-normal-FH controls, does that contribute further to biases in r_{gAB}? Finally, what bias occurs in r_{gAB} when the two disorders are studied using unscreened controls and how is that bias affected by the population prevalence of disorders A and B and the true r_g between them?

METHODS

Due to the difficulty of performing theoretical calculations, we resorted to computationally intensive simulations using the Julia language version 0.6.4 (https://julialang.org/). For simplicity and computational tractability, we assumed that i) the studies of the two disorders of interest (A and B) are non-overlapping with total sample sizes $N_1 = N_2 = 5{,}000$, for ii) GWAS consisted of $M = 5{,}000$ independent Single Nucleotide Polymorphisms (SNPs) and iii) super-normals are defined as screened out if they have either disorder A or B. The experiment assumes that the genetic correlation is computed, as it is commonly performed, from the GWAS summary statistics using equation (1) from Bulik-Sullivan et al (Bulik-Sullivan et al., 2015). Under the above settings for sample sizes and number of markers, the genetic covariance is simply $E(z_{1j}z_{2j})$, where z_{1j} and z_{2j} are the j-th SNP's Z-score for the first and second disorder, respectively. Similarly, the heritabilities of the traits are ${h_i}^2 = E\left({z_{ij}}^2\right) - 1$, for $i = 1, 2$.

Let the two disorders under investigation have prevalences K_1 / K_2 , (equal) heritabilities ${h_1}^2$= ${h_2}^2$ and genetic correlation ρ. The genetic correlation matrix of the two disorders can be written as $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. The varying parameters in the simulation design consisted of true genetic correlation, heritability and prevalence of the two disorders A and B (Table 1). 10 simulations were performed at each setting, such that the average standard error of the estimates was < 0.02.

The simulation procedure is as follows:

1) Simulate (liability) random effects for the i-th SNP ($i = 1, \dots, M$) $c_i \sim \text{Bivariate Normal}(0, \Sigma)$.
2) Draw at random the minor allele frequency (MAF) of the i-th SNP: $p_i \sim \text{Uniform}(0.05, 0.5)$.
3) Independently simulate the genotype for the father and the mother of the proband at each SNP. E.g., if the genotype at the i-th SNP is $G_i =$ the number of minor alleles, then $G_i \sim \text{Binomial}(2, p_i)$.

4) Simulate proband and sib by randomly sampling one allele from each parent

5) For each individual, assign the standard normal (Gaussian) liability for the j-th disorder as $L = \sqrt{{h_j}^2}\frac{\sum_1^M c_i\,(X_i - 2\,p_i)}{\sqrt{\sum_1^M 2\,{c_i}^2\,p_i\,(1-p_i\,)}} + \varepsilon$, where $\varepsilon \sim \text{Normal}\,(0, 1 - {h_j}^2)$ $(1 - {h_j}^2$ being the variance of the normal deviate).

6) For the j-th trait assign case/control status for each member of the family when the upper percentile of the liability under a Gaussian distribution is less/bigger than K_j. (Assign SN status when a subject is a control for both disorders.)

7) number of cases and controls/SN controls equals $\frac{N_1}{2} = \frac{N_2}{2} = 2{,}500$.

8) Compute Z-score for the i-th ($i = 1{,}2$) trait and j-th SNP as $Z_{ij} = \sqrt{N_i - 3}\,Cor(G_j, Y)$, where G_j is the genotype vector for cases and controls and Y is the phenotype vector (0 for controls and 1 for cases)

RESULTS

We began by validating our simulation procedure by showing that for normal controls, the estimates of r_{gAB} are unbiased (Fig. 1, left panel). Prior to starting the simulation experiment, we also validated that the simulated liabilities for the two traits yield the desired genetic correlations (data not shown). We next examined the bias in r_{gAB} resulting from the use of super-normal controls with results depicted in figure 2, left panel. When both disorders A and B are rare (K1=K2=1.0%), the upward bias of r_{gAB} is very modest across the range of values of the true r_{gAB}. When disorder K1 is rare (1.0%), and disorder K2 has a prevalence of 7.5% or 15%, the upward bias becomes more pronounced with greater bias when the true value of r_{gAB} is low. For example, when K2=7.5 and 15% and the true r_{gAB} equals 0.10, the estimates values (SE) of r_{gAB} are +0.18 (0.01) and +0.20 (0.01), respectively. The upward bias on r_{gAB} is slightly greater when both K1=K2=7.5% and becomes pronounced when the prevalences of K1 and K2 are 7.5 and 15%, and both are 15%. For example, if K1=K2=15%, when the true value of r_{gAB} equals 0, 0.10,

0.30 and 0.50, the estimated values of r_{gAB} equal, respectively, +0.31 (0.02), +0.40 (0.01), +0.54 (0.01), and +0.68 (0.01).

We then examined the effects of using normal-FH and super-normal-FH controls. As seen in figure 1 right panel, using normal-FH controls produced no appreciable bias in the estimated r_{gAB}. However, compared to standard super-normal control, super-normal-FH controls produced a greater upward bias on r_{gAB} (figure 2, right panel) which was particularly strong when the true value of r_g between the two disorders equaled zero. For example, when K1=K2=1.0%, K1=K2=7.5% and K1=K2=15.0%, the estimated values of r_{gAB} equaled, respectively, +0.04 (0.01), +0.27 (0.01), and +0.50 (0.01).

Finally, we examined the impact of unscreened controls on estimates of r_{gAB} (figure 3). At low levels of true genetic correlation, the upward biases were modest. However, these biases became progressively stronger with higher true genetic correlations and higher prevalence disorders producing, in several situations, implausible values of genetic correlations exceeding unity.

DISCUSSION

We draw four major conclusions from these simulations. First, to obtain accurate estimates of r_g using GWA results, studies should use normal controls. Second, replicating and extending the results of van Rheenen et al. (van Rheenen et al., 2019), we find that using super-normal controls will result in upward biases the magnitude of which are strongly related to the prevalence of the pair of disorders and their true r_g. At low prevalences (≤ 1%), the bias introduced is modest while it can become quite substantial at high prevalences (≥ 15%). The upward bias on estimate of r_g is also substantially stronger at low true correlations and, with high prevalence disorders, can produce evidence for substantial r_gs where none exists. The replication of the results for super-normal controls is reassuring given that van Rheenen et al (van Rheenen et al., 2019) derived their results theoretically, while we used a simulation design

coupled with empirical estimation of correlation (Bulik-Sullivan et al., 2015) in order to explore other types of controls (e.g. unscreened or screened together with close relatives).

Third, while screening the close relatives of controls produced no appreciable bias in r_g when used with normal controls, screening relatives using a super-normal control paradigm substantially exacerbated the upward bias on r_gs. Finally, while the use of unscreened controls in GWA studies can result in savings of research resources, they pose a risk of substantially upwardly biasing evidence for genetic correlations especially when used with relatively common disorders.

These results are intuitively plausible. Using normal controls, the signal in a case-control GWA study arises entirely from the fact that all the cases and none of the controls have disorder A. When disorder B is eliminated from only controls, the case-control difference becomes "contaminated" and reflects not only case-control differences in the prevalence of A but also of B. The same will occur when we study disorder B but screen our controls for disorder A. So, if we took the results of such a GWA study and calculated $r_{gAB,}$ our estimates would be biased upward because our genetic risk for disorder A is contaminated with risk variants for disorder B and vice-versa. Within this thought experiment, it is obvious that the strength of the "contamination" of our GWA study will relate directly to the frequency of the screened-out disorder. If rare, it will impact only modestly on the resulting "signal." If we screen relatives of controls for disorder B when studying A, this will further reduce the genetic risk in controls for disorder B, increasing the difference with cases and contaminating the signal and hence the resulting r_{gAB}. On the other hand, unscreened controls induce an increase in genetic correlation due to mainly underestimating the heritability of the trait(s). That is, the heritability is underestimated due to the average noncentrality of statistics being lower when compared to their screened control counterparts (see genetic covariance and heritability estimates in Appendix).

These results should be interpreted in the context of three methodological limitations. First, we only presented results at a single heritability for our two disorders: 0.50. We repeated results for a higher heritability (0.80 – see appendix). The pattern of results was similar. Second, we only presented results where the same control selection method had been used with both disorders. In real life, r_g can be calculated between one disorder assessed using normal controls and another using super-normal controls. Simulations suggested that, in this case, the results were generally in between those found when normal and super-normal controls were assessed for both diseases. Third, we have not considered an expanded super-normal control method that screens for multiple disorders. Here results would depend on the prevalence and true genetic correlations between the range of disorders excluded from controls and would likely produce stronger and more wide-spread biases than presented here when we examined only two disorders at a time.

Reference List

Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L. et al. (2018). Analysis of shared heritability in common disorders of the brain. *Science, 360*. DOI: 10.1126/science.aap8757

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R. et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics, 47,* 1236-1241. DOI: 10.1038/ng.3406

Chen, T. J., Blum, K., Mathews, D., Fisher, L., Schnautz, N., Braverman, E. R. et al. (2005). Are dopaminergic genes involved in a predisposition to pathological aggression? Hypothesizing the importance of "super normal controls" in psychiatricgenetic research of complex behavioral disorders. *Medical Hypotheses, 65,* 703-707. DOI: 10.1016/j.mehy.2005.04.037

Cross-Disorder Group of the Psychiatric Genomics Consortium(PGC-CDG) (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics, 45,* 984-94. DOI: 10.1038/ng.2711

Hodge, S. E., Subaran, R. L., Weissman, M. M., & Fyer, A. J. (2012). Designing case-control studies: decisions about the controls. *American Journal of Psychiatry, 169,* 785-789. DOI: 10.1176/appi.ajp.2012.11111686

Ikeda, M., Aleksic, B., Kinoshita, Y., Okochi, T., Kawashima, K., Kushima, I. et al. (2011). Genome-wide association study of schizophrenia in a Japanese population. *Biological Psychiatry, 69,* 472-478. DOI: 10.1016/j.biopsych.2010.07.010

Kendler, K. S. (1990). The super-normal control group in psychiatric genetics: Possible artifactual evidence for coaggregation. *Psychiatric Genetics, 1,* 45-53.

Kirov, G., Zaharieva, I., Georgieva, L., Moskvina, V., Nikolov, I., Cichon, S. et al. (2009). A genome-wide association study in 574 schizophrenia trios using DNA pooling. *Molecular Psychiatry, 14,* 796-803. DOI: 10.1038/mp.2008.33

Lopez, R., Scheutz, F., Errboe, M., & Baelum, V. (2007). Selection bias in case-control studies on periodontitis: a systematic review. *European Journal of Oral Science, 115,* 339-343. DOI: 10.1111/j.1600-0722.2007.00476.x

Lubin, J. H. & Gail, M. H. (1984). Biased selection of controls for case-control analyses of cohort studies. *Biometrics, 40,* 63-75.

Malay, S. & Chung, K. C. (2012). The choice of controls for providing validity and evidence in clinical research. *Plastic Reconstructive Surgery, 130,* 959-965. DOI: 10.1097/PRS.0b013e318262f4c8

O'Donovan, M. C., Craddock, N., Norton, N., Williams, H., Peirce, T., Moskvina, V. et al. (2008). Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nature Genetics, 40,* 1053-1055. DOI: 10.1038/ng.201

Otowa, T., Hek, K., Lee, M., Byrne, E. M., Mirza, S. S., Nivard, M. G. et al. (2016). Meta-analysis of genome-wide association studies of anxiety disorders. *Molecular Psychiatry, 21,* 1391-1399. DOI: 10.1038/mp.2015.197

Schwartz, S. & Susser, E. (2010). Genome-wide association studies: does only size matter? *American Journal of Psychiatry, 167,* 741-744. DOI: 10.1176/appi.ajp.2010.10030465

Schwartz, S. & Susser, E. (2011). The use of well controls: an unhealthy practice in psychiatric research. *Psychological Medicine, 41,* 1127-1131. DOI: 10.1017/S0033291710001595

Sullivan, P. F., Lin, D., Tzeng, J. Y., van den Oord, E., Perkins, D., Stroup, T. S. et al. (2008). Genomewide association for schizophrenia in the CATIE study: results of stage 1. *Molecular Psychiatry, 13,* 570-584. DOI: 10.1038/mp.2008.25

van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H., & Wray, N. R. (2019). Genetic correlations of polygenic disease traits: from theory to practice. *Nature Reviews Genetics*. DOI: 10.1038/s41576-019-0137-z

Wacholder, S., McLaughlin, J. K., Silverman, D. T., & Mandel, J. S. (1992). Selection of controls in case-control studies. I. Principles. *American Journal of Epidemiology, 135,* 1019-1028. DOI: 10.1093/oxfordjournals.aje.a116396

Wacholder, S., Silverman, D. T., McLaughlin, J. K., & Mandel, J. S. (1992a). Selection of controls in case-control studies. II. Types of controls. *American Journal of Epidemiology, 135,* 1029-1041. DOI: 10.1093/oxfordjournals.aje.a116397

Wacholder, S., Silverman, D. T., McLaughlin, J. K., & Mandel, J. S. (1992b). Selection of controls in case-control studies. III. Design options. *American Journal of Epidemiology, 135,* 1042-1050. DOI: 10.1093/oxfordjournals.aje.a116398

Wray, N. R., Pergadia, M. L., Blackwood, D. H., Penninx, B. W., Gordon, S. D., Nyholt, D. R. et al. (2012). Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Molecular Psychiatry, 17,* 36-48. DOI: 10.1038/mp.2010.109

Figures

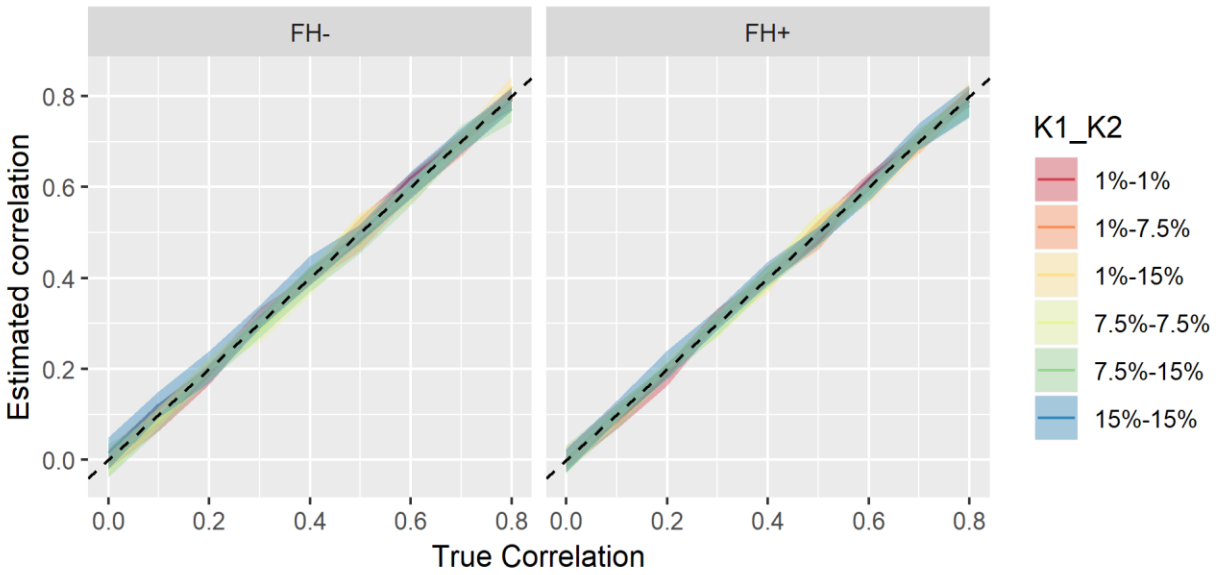

Figure 1. The relationship between the estimated correlation (as-per-Bulik-Sullivan et al. (Bulik-Sullivan et

al., 2015)) and true correlation and true correlation when both disorders use normal controls. The banner settings relate to family history (FH), with (FH+) or without (FH-) family history screening for the condition(s). Each disorder is assumed to have a population prevalence of 1%, 7.5% and 15%, the color of different lines defines the assumed prevalence pair of the disorders traits (K1_K2 -K1 for disorder A and K2 for disorder B). The heritability of each trait was assumed to be 0.5. The black dotted line denotes an unbiased estimator, i.e. estimated and true correlations are equal.

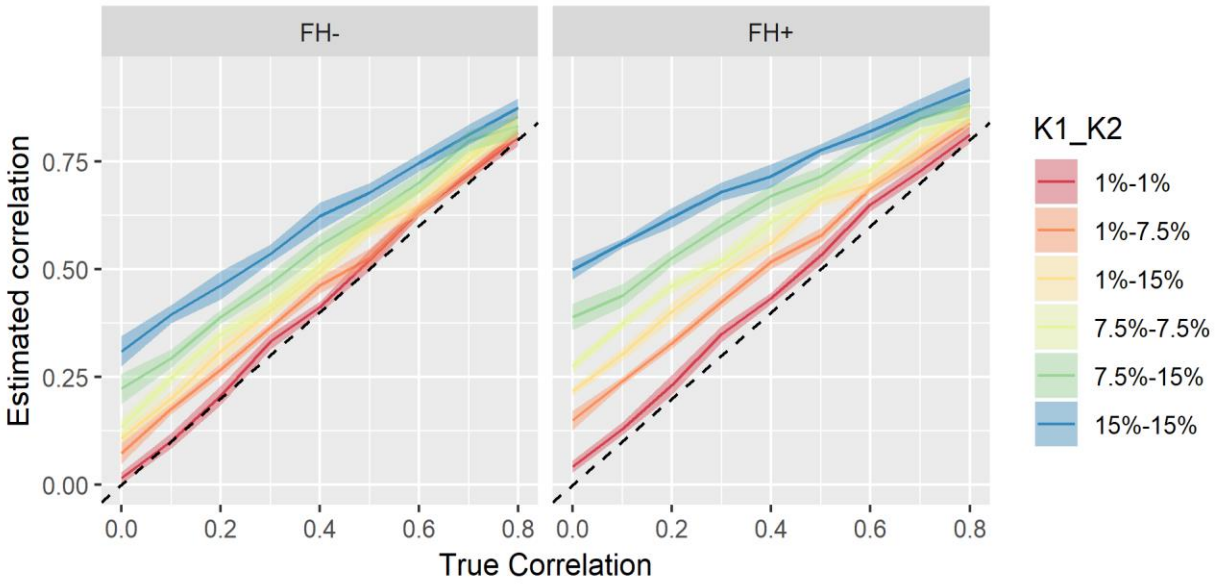

Figure 2. The relationship between the estimated correlation and true correlation when both disorders are studied using-super-normal controls as a function of the true correlation and the prevalence rates of the two disorders (K1 for disorder A and K2 for disorder B). For notation and background see Fig.1.

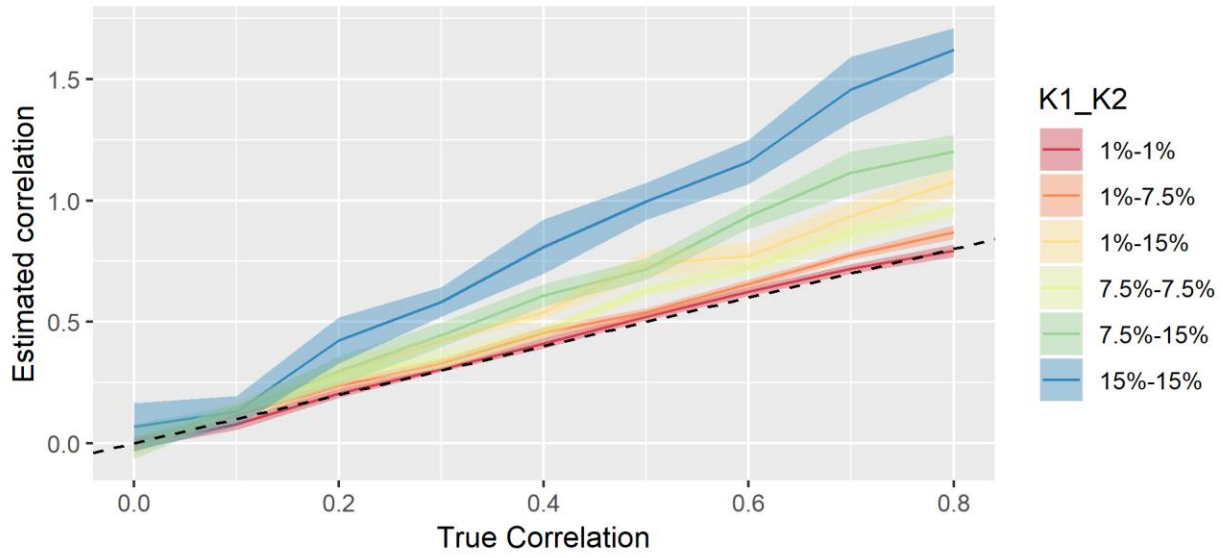

Figure 3. The relationship between the estimated correlation (as-per-Bulik-Sullivan et al (Bulik-Sullivan et al., 2015)) and true correlation when both traits use unscreened controls. For notation and background see Fig.1.

Supplementary Material

Estimates for genetic correlation results for trait heritability of 0.8

When compared to a heritability of 0.5 for both traits, the results are rather similar to their homologues (Fig. S1-3). The one slight exception is that the bias is now diminished when studies use population (Fig. S1). These biases are still sizeable, especially when i) not using normal controls and ii) one of the traits is very common.

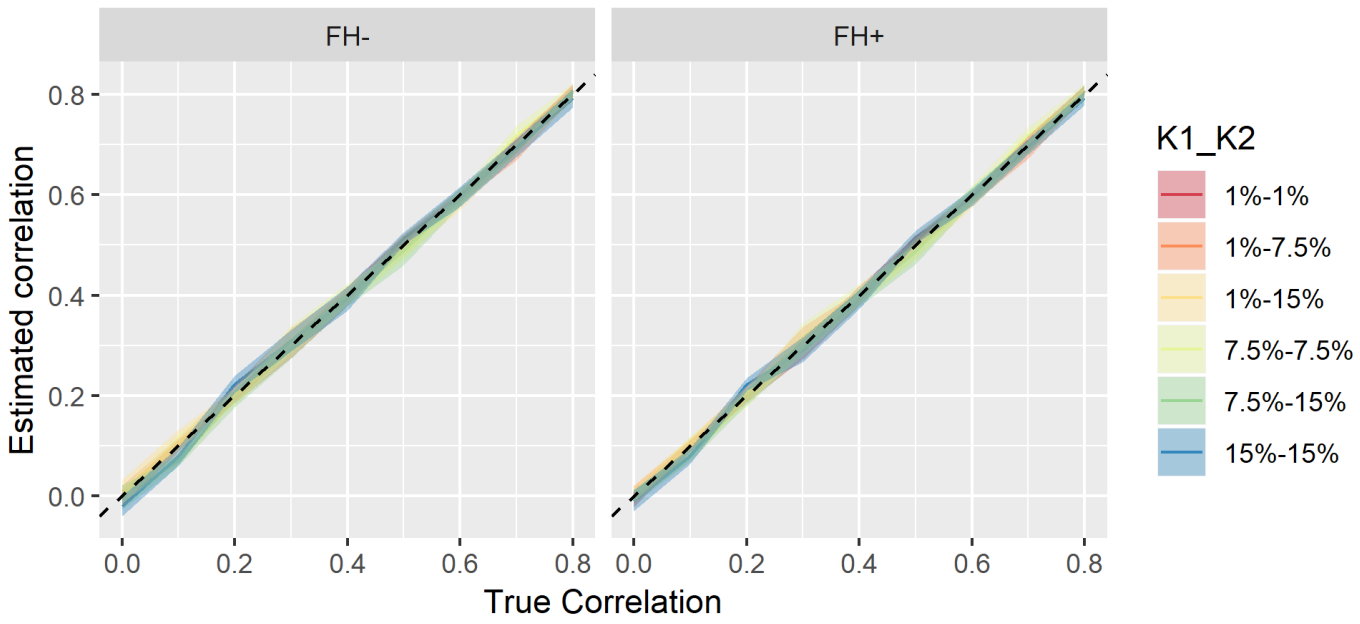

Figure S1. The relationship between the estimated correlation and true correlation and true correlation when both traits use normal controls. The banner settings relate to family history (FH), with (FH+) or without (FH-) family history screening for the condition(s). Each trait is assumed to have prevalence of 1%, 7.5% and 15%, the color of different lines defines the assumed prevalence pair of the traits (K1_K2- where K1 for disorder A and K2 for disorder B). The heritability of each trait was assumed to be 0.8. The black dotted line denotes an unbiased estimator, i.e. estimated and true correlations are equal.

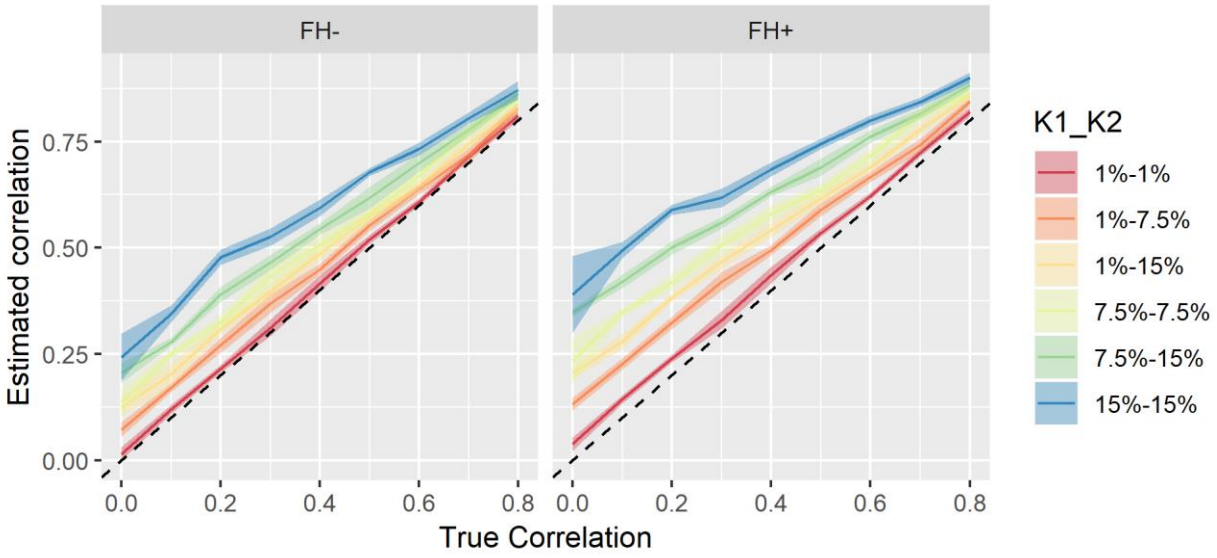

Figure S2. The relationship between the estimated correlation and true correlation and true correlation when both traits use super-normal controls low prevalence traits. For notation and background see Fig.S1.

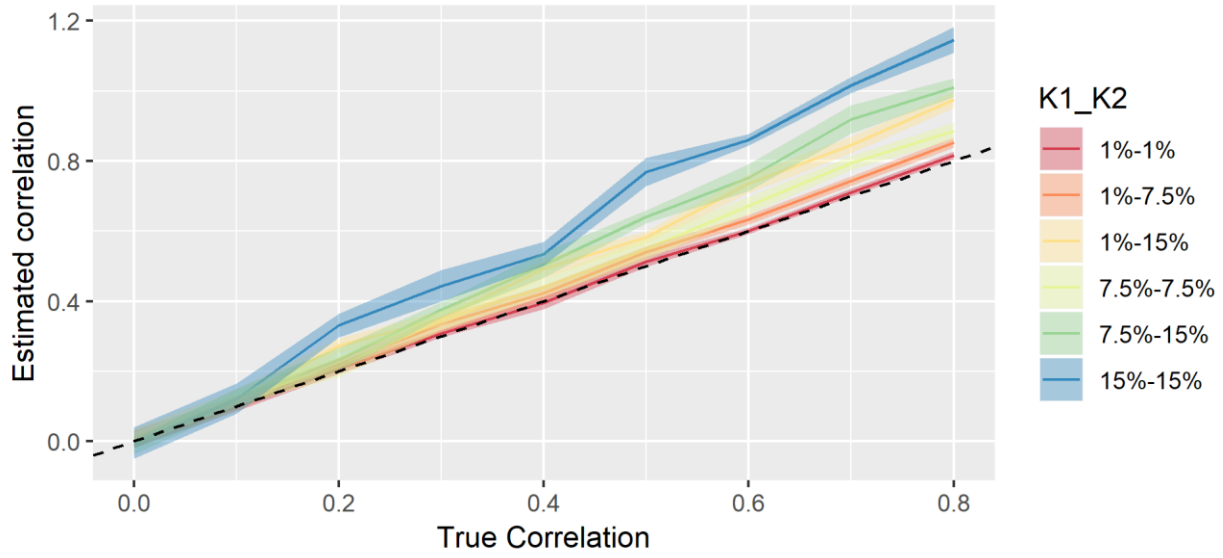

Figure S3. The relationship between the estimated correlation (as-per-Bulik-Sullivan et al.[1]) and true correlation when both traits use unscreened controls.. For notation and background see Fig.S1.

Components of genetic correlation estimation

The genetic correlation is the quotient of genetic covariance and the square root of heritabilities. It is, thus, of interest to assess how the estimates of these two measures vary for unscreened/super-normal controls relative to the normal controls. Consequently, we also estimated heritabilities and genetic covariances under the settings presented in the main manuscript.

The main points relating to genetic covariance estimates are: i) at high prevalences, the use of normal controls induce a slight overestimation only under the FH scenario and population controls a modest underestimation and ii) super-normal controls induce large overestimations, especially for lower true genetic correlations and/or super-normal-FH (Fig. S3-S6). The main feature for estimated heritabilities are: i) at high prevalences, the use of normal controls induces a modest overestimation (only under FH scenario) and population controls a sizeable underestimation ii) super-normal controls induce reasonably large overestimations, especially for lower true genetic correlations and/or super-normal-FH controls (Fig. S7-S9).

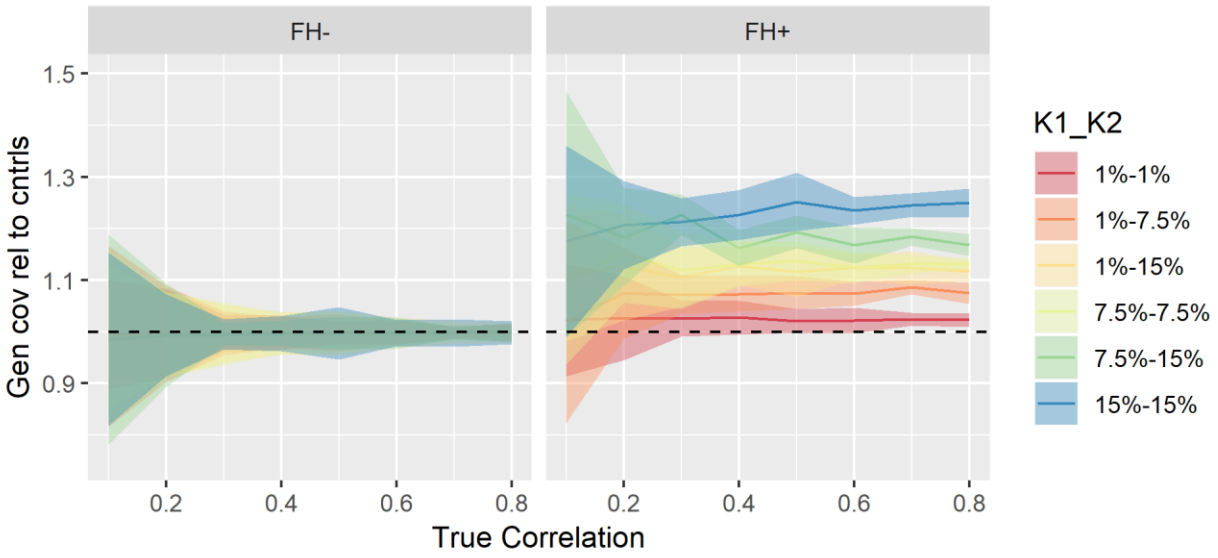

Figure S4. The relationship between the estimated genetic covariance (relative to normal controls) and true correlation when both traits use normal controls. The heritability of each trait was assumed to be 0.8. The true correlation value of 0 was eliminated from the plot due to the division with an estimate of zero (estimated genetic covariance of studies with normal controls), which resulted in extremely large confidence intervals. The black dotted line denotes an unbiased estimation in normal controls (i.e. the horizontal line at y=1). For notation and background see Fig.S1.

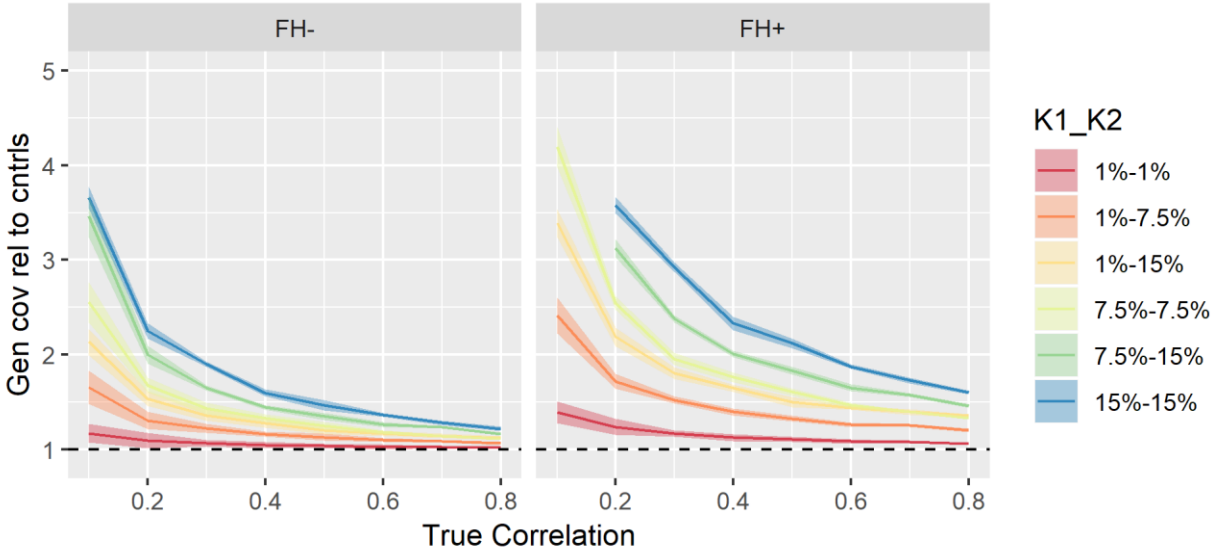

Figure S5. The relationship between the estimated genetic covariance (relative to normal controls) and true genetic correlation when both traits use super-normal controls. The heritability of each trait was assumed to be 0.8. For notation and background

see Fig.S1 and S4.

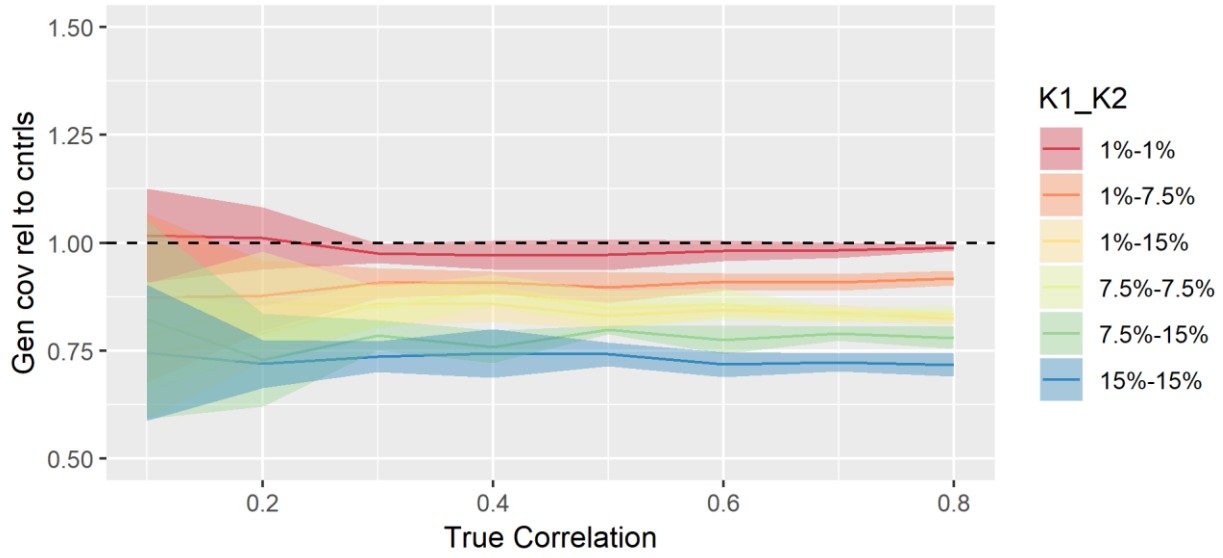

Figure S6. The relationship between the estimated genetic covariance (relative to normal controls) and true genetic correlation when both traits use unscreened controls. The heritability of each trait was assumed to be 0.8. For notation and background see Fig.S1 and S4.

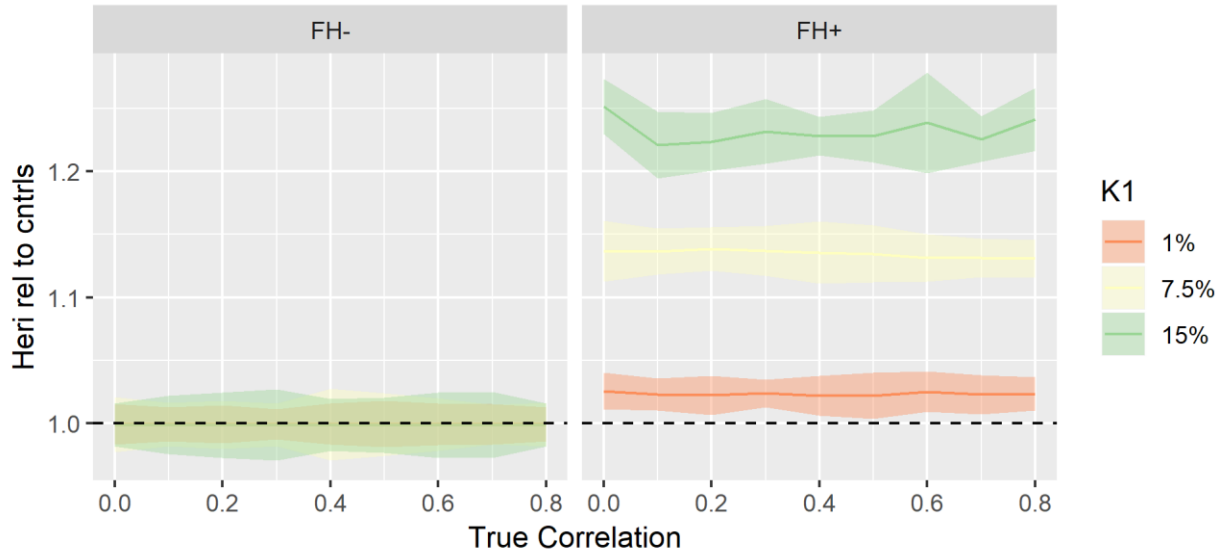

Figure S7. The relationship between the estimate in heritability (relative to normal controls) and true correlation when both traits use normal controls. The black dotted line denotes an unbiased estimation in normal controls (i.e. the horizontal line at y=1). For notation and background see Fig.S1 and S4.

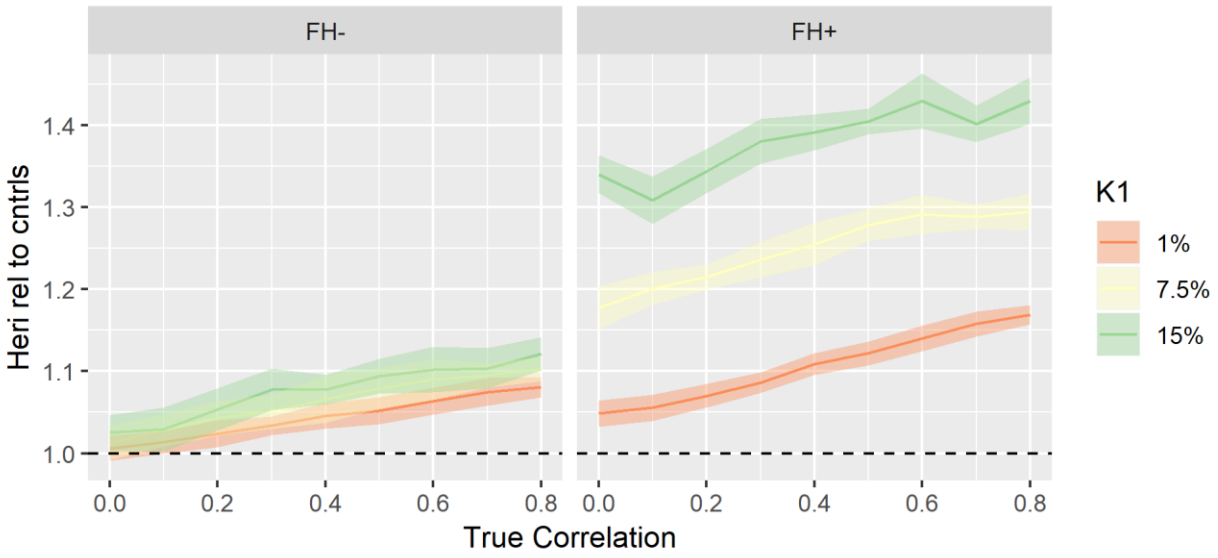

Figure S8. The relationship between the estimated heritability (relative to normal controls) and true genetic correlation when both traits use super-normal controls. The heritability of each trait was assumed to be 0.8. For notation and background see Fig.S1 and S7.

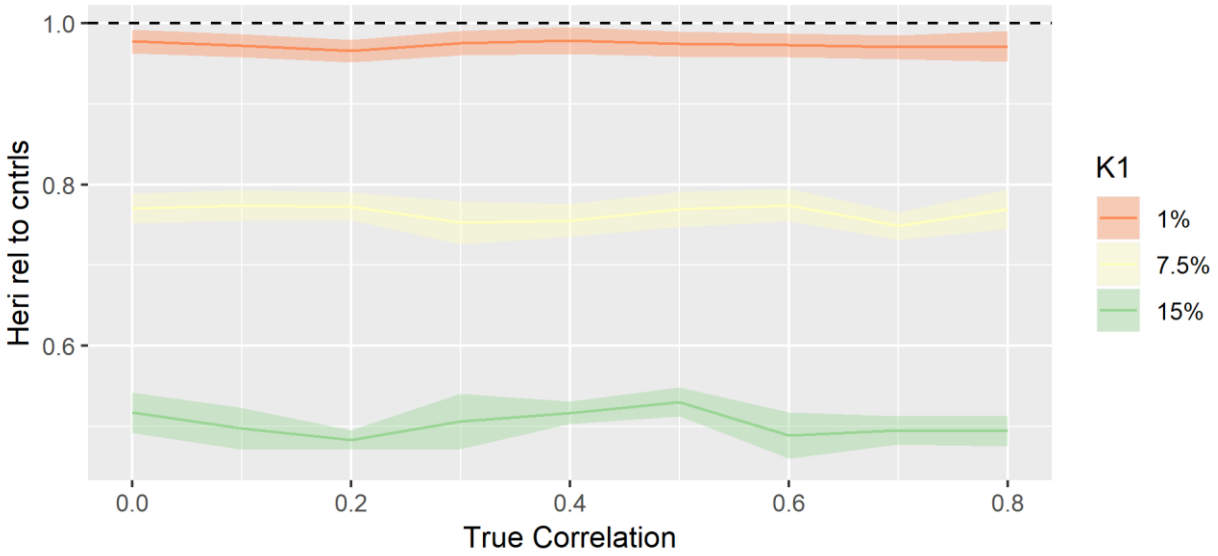

Figure S9. The relationship between the estimated heritability (relative to normal controls) and true genetic correlation when both traits use unscreened controls. The heritability of each trait was assumed to be 0.8. For notation and background see Fig.S1 and S7.