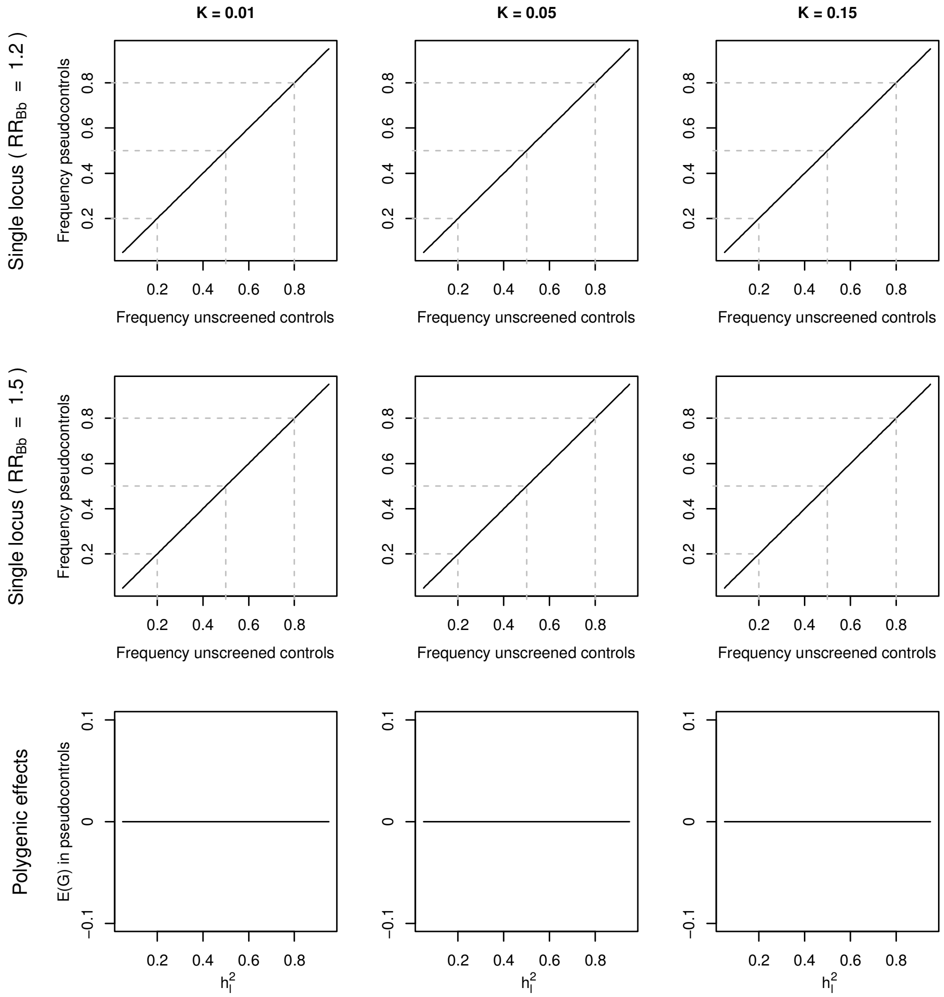
The American Journal of Human Genetics

Supplemental Data

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

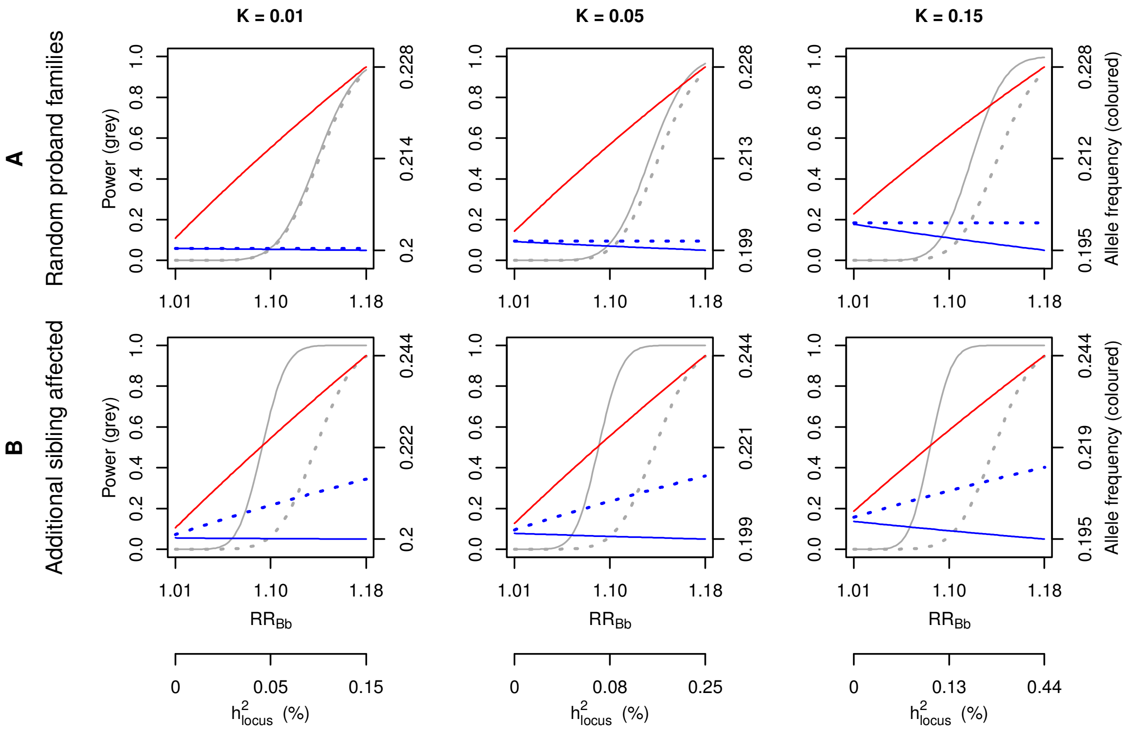
Wouter J. Peyrot, Dorret I. Boomsma, Brenda W.J.H. Penninx, and Naomi R. Wray

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



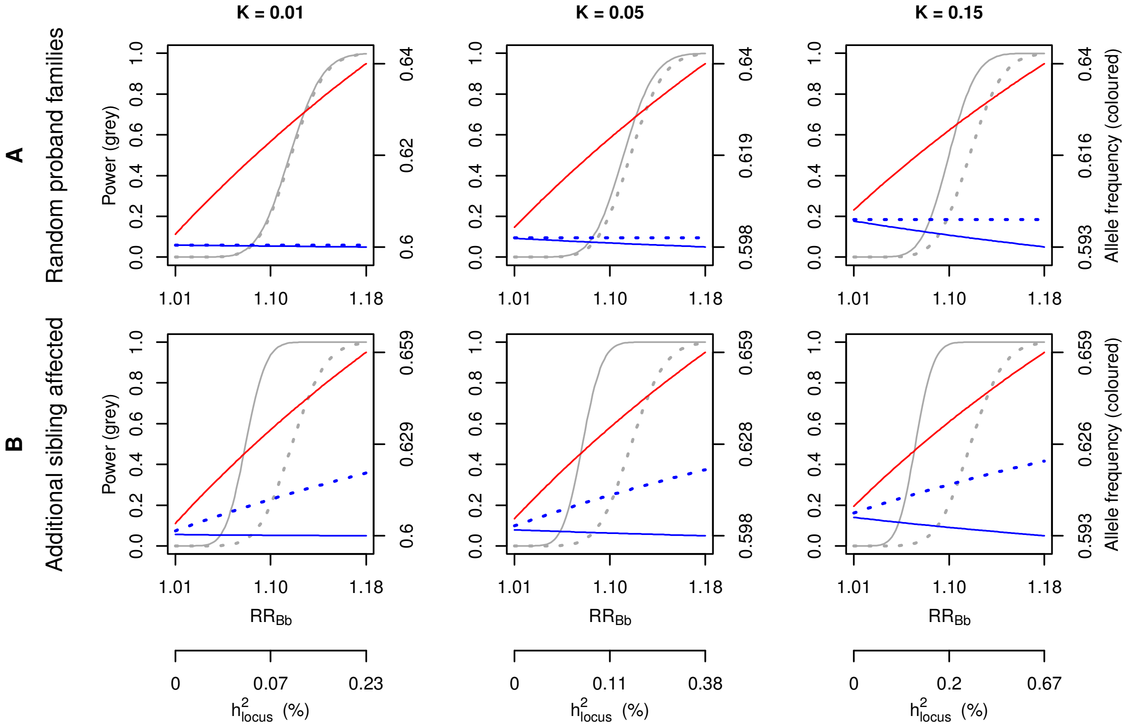
Pseudocontrols of random families with at least one affected proband case are equal to unscreened controls (i.e. population mean) as displayed for the allele frequency of single loci of different effect-size (first two rows) and the mean genetic liability $E(G)$ (population mean equals 0) for variable 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



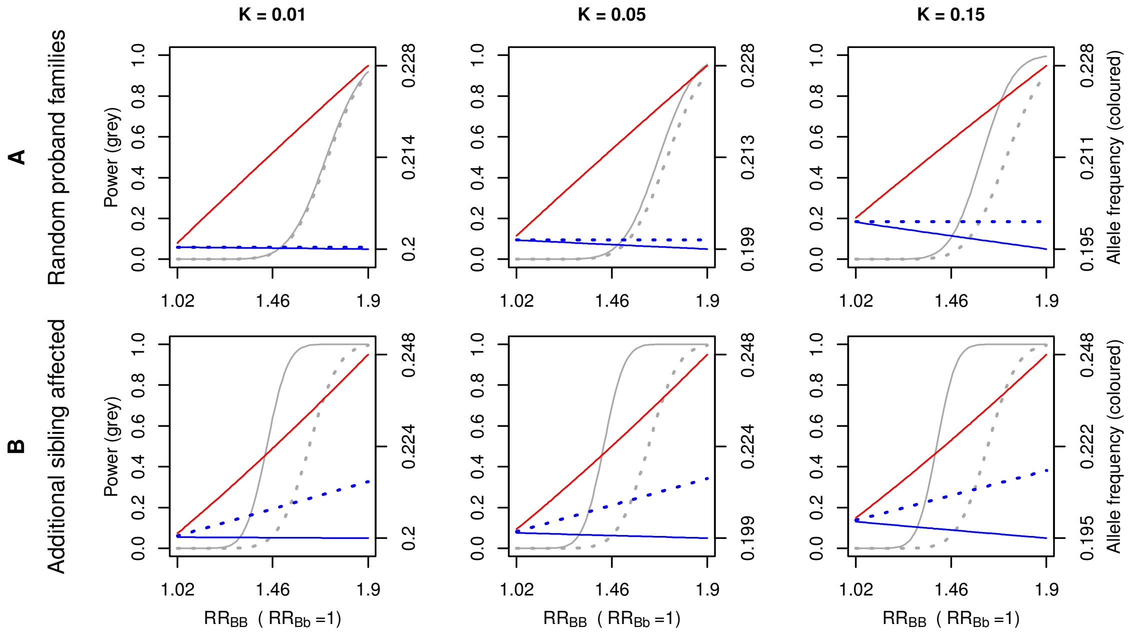
Power to detect a single SNP with risk allele frequency $p = 0.2$ for case vs screened controls (solid grey line) and case vs pseudocontrol (dotted grey line). The allele frequencies of proband cases are displayed as the red solid line, the allele frequency of screened controls as the solid blue line, 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



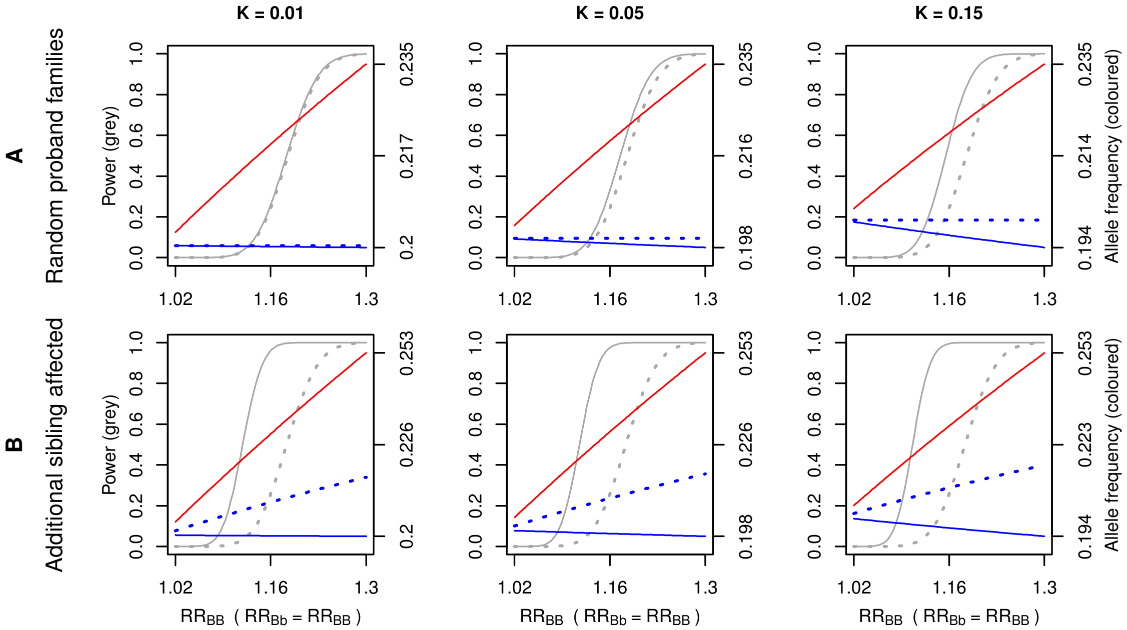
The power is displayed for a risk allele with frequency p=0.6, and results indicate that the conclusions do not depend on the allele frequency (noting that in Figure S2 a locus with p=0.2 was displayed). See the legend of Figure S2 for details.

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



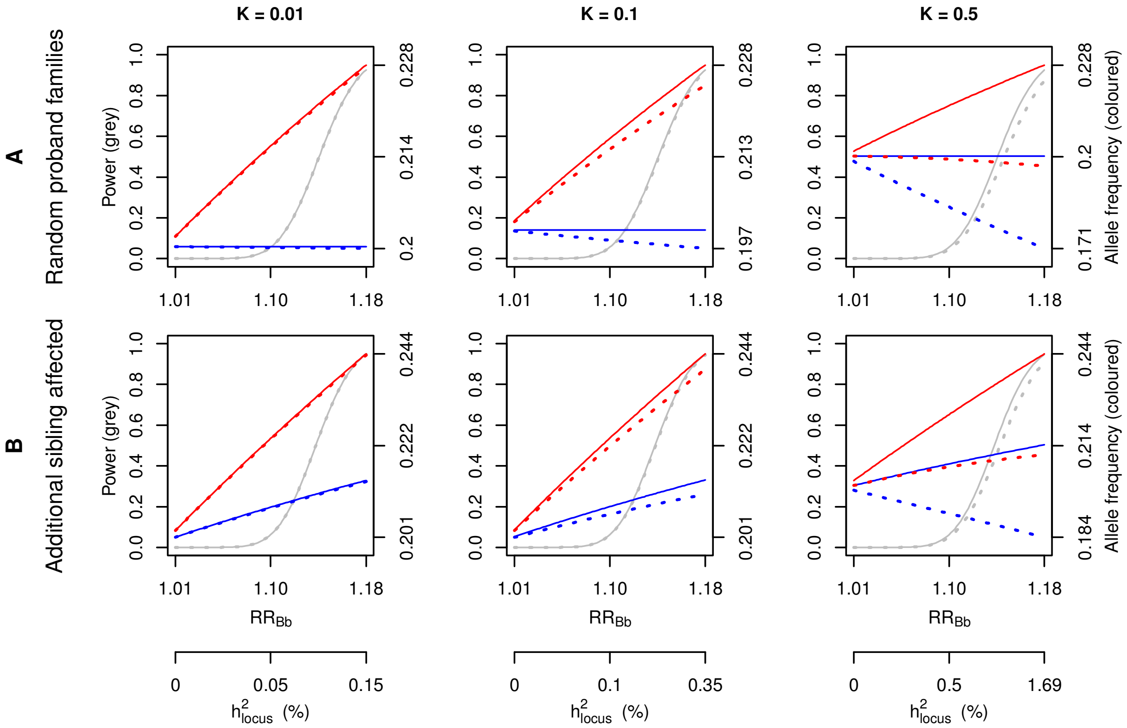
Power to detect the additive effect a single SNP with risk allele frequency $p = 0.2$ with an underlying recessive effect for case vs screened controls (solid grey line) and case vs pseudocontrol (dotted grey line). The allele frequency of cases is displayed as the red solid line, the allele frequency 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



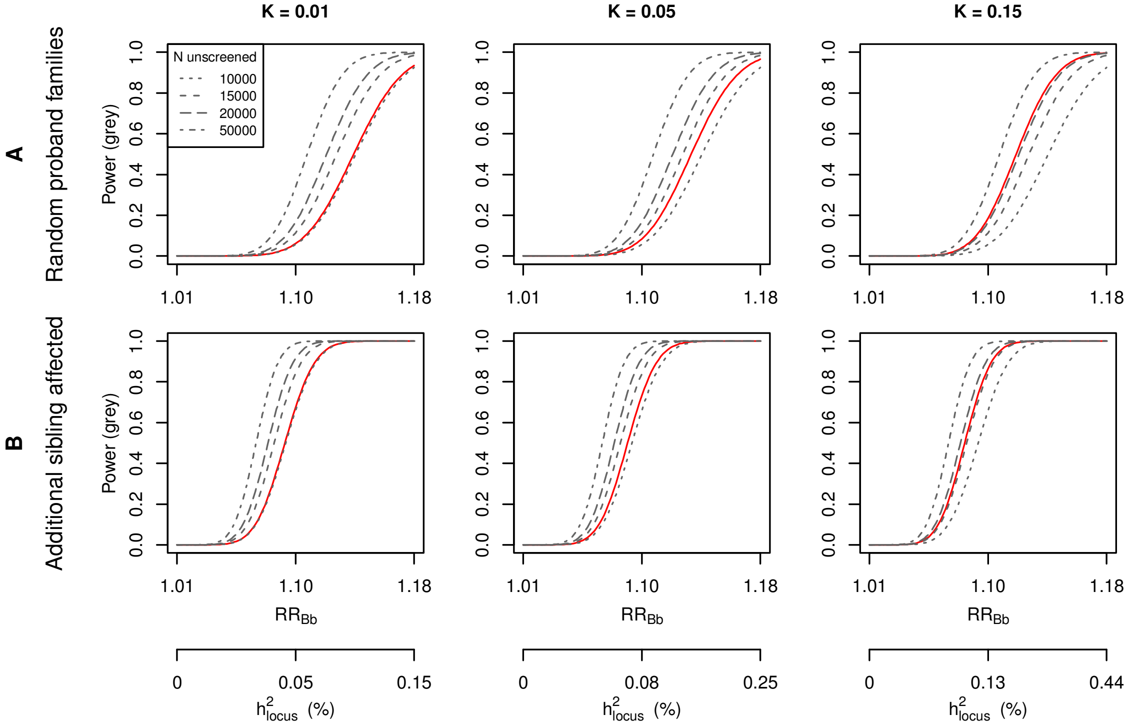
Power to detect the additive effect a single SNP with risk allele frequency $p = 0.2$ with an actual dominant effect for case vs screened controls (solid grey line) and case vs pseudocontrol (dotted grey line). The allele frequency of cases is displayed as the red solid line, the allele frequency of 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

$\boldsymbol{y_{true,i}}$	$\boldsymbol{y_{true,j}}$	$\boldsymbol{y_{assumed,i}}$	$\boldsymbol{y_{assumed,j}}$	$\mathbb{P}_{\boldsymbol{ij}}$	$\boldsymbol{Z_{ij}}$
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	$P_{assumed}^2$	$\frac{1-P_{assumed}}{P_{assumed}}$
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 $\hat{h}_o^2$ to the liability scale $\hat{h}_l^2$ 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 $\mathbb{P}_{\boldsymbol{ij}}$ 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:

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

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

$$\mathbb{E}[Z_{ij}|y_{true,i} = y_{true,j} = 0] = \frac{P_{assumed}}{1-P_{assumed}}$$

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

				$\hat{h}_l^2$ screened control			$\hat{h}_l^2$ pseudocontrol		
Simulation parameters				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 $cor(l_{men}, V) = cov(l_{men}, V)/(\sigma_{l_{men}}\sigma_V) = cov(l_{men}, \rho_l l_{men})/(1\sigma_V) = \rho_l/\sqrt{\sigma_{\rho_l l_{men}}^2 + 1 - \rho_l^2} = \rho_l$
    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 $\hat{h}_{occ}^2$ for case vs screened controls and case vs pseudocontrols, which were than transformed to the liability scale with $\hat{h}_l^2 = \hat{h}_{occ}^2 \frac{K^2(1-K)^2}{P(1-P)z^2}$
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.

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

- $E(G_m|s1\,aff) = E\big(G_f|s1\,aff\big) = \Big\{(\frac{1}{2}V_G + \frac{1}{2}\rho_G V_G)/\sigma^2(G_s)\Big\} * E(G_{s1}|s1\,aff)$, 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 $\frac{1}{2}G_f + \frac{1}{2}G_m$.
- $E(G_{c1}|s1\,aff) = E(G_m|s1\,aff) + E\big(G_f|s1\,aff\big) - E(G_{s1}|s1\,aff)$

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 $K_{s2\,|\,s1\,affected} = P\big(x > T_{s2\,|\,s1\,affected}\,\big|\,x{\sim}N(0,1)\big)$. 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 $\sum\big(l_{s2}, G_{s2}, G_m, G_f, G_{c1}\,\big|\,s1\,\&\,s2\,affected\big)$, 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:

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

- $E(G_{s2}|\ s1\ \&\ s2\ aff) =$ $E(G_{s2}|\ s1\ aff) + \{i_{s2\ |\ s1\ affected}\sqrt{\sigma^2(l_{s2}|\ s1\ affected)}\} *$ $\sigma^2(G_{s2}|\ s1\ affected)/\sigma^2(l_{s2}|\ s1\ affected)$
- $E(G_m|s1\ \&\ s2\ aff) = E\left(G_f\middle|s1\ \&\ s2\ aff\right) =$ $E\left(G_f\middle|s1\ aff\right) + \delta * \{\frac{1}{2}\sigma^2(G_m|s1\ aff) + \frac{1}{2}\sigma(G_m, G_f|s1\ aff)\}/\{\sigma^2(G_{s2}|s1\ aff)\}$, with $\delta =$ $E(G_{s2}|\ s1\ \&\ s2\ aff) - E(G_{s2}|\ s1aff)$, while noting that $\frac{1}{2}\sigma^2(G_m|s1\ aff) + \frac{1}{2}\sigma\left(G_m, G_f\middle|s1\ aff\right) +$ $\frac{1}{2}V_{residual} = \sigma^2(G_{s2}|s1\ aff)$.
- $E(G_{c1}|\ s1\ \&\ s2\ aff) = E(G_m|\ s1\ \&\ s2\ aff) + E\left(G_f\middle|\ s1\ \&\ s2\ aff\right) - E(G_{s1}|\ s1\ \&\ s2\ aff)$, 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 $P(\mathrm{BB|D}) = k_{bb}\lambda^2p^2/K$. 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 $P(\mathrm{BB|D,pu}) = k_{bb}\lambda^2p_{sc}^2/P(\mathrm{D|pu})$, with $P(\mathrm{D|pu}) = k_{bb}q_{sc}^2 + k_{bb}\lambda 2p_{sc}q_{sc} + k_{bb}\lambda^2p_{sc}^2$. 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

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

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

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

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

Where ∘ 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

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

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

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

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.