Web Based Supplementary Material for Combining Disease Models to test for Gene-Environment Interaction in Nuclear Families by Thomas J. Hoffmann, Stijn Vansteelandt, Christoph Lange, Edwin K. Silverman, Dawn L. DeMeo, and Nan M. Laird

Web Appendix A: Conditional independence of the gene and environment given the sufficient statistic for parental mating type

Here we drop the subscript i from our notation, and partition P into P^{obs} and P^{mis}, the genotypes of the observed and missing parents, where P^{obs} can be null. The independence of $\mathbf{X}$ and $\mathbf{Z}$ given S follows from two properties of the sufficient statistic: it depends only on observed genotype data and the distribution of $\mathbf{X}$ given S and P^{mis} does not depend on P^{mis}. We assume

(1) $\mathbf{X}$ and $\mathbf{Z}$ are independent given P.

(2) S is solely determined by $\mathbf{X}$ and P^{obs}.

A formal argument can be given as follows. Because S is a function only of P^{obs} and $\mathbf{X}$, we can write

$$Pr(\mathbf{X}, \mathbf{Z}, P) \propto Pr(\mathbf{X}, \mathbf{Z}, S, P^{\text{mis}})$$

but by assumption 1,

$$Pr(\mathbf{X}, \mathbf{Z}, S, P^{\text{mis}}) = Pr(\mathbf{X}|S, P^{\text{mis}})Pr(\mathbf{Z}|S, P^{\text{mis}})Pr(P^{\text{mis}}|S)Pr(S)$$

because S and P^{mis} together give P. Additionally, by definition of the sufficient statistic, $Pr(\mathbf{X}|S, P^{\text{mis}}) = Pr(\mathbf{X}|S)$, hence summing of the above equation over the P^{mis} and dividing by $Pr(S)$ gives $Pr(\mathbf{X}, \mathbf{Z}|S) \propto Pr(\mathbf{X}|S)Pr(\mathbf{Z}|S)$, and hence $Pr(\mathbf{X}, \mathbf{Z}|S)$ factors.

Web Appendix B: Further details of the TX method

Here we fill in the missing details for the derivation in Section 2.1. A detailed derivation of the likelihood for the log link proceeds as follows. Continuing from equation 3 in the paper, the retrospective likelihood is given by

$$\begin{aligned}\mathcal{L}_i^{\text{TX}} &= Pr(\mathbf{X}_i|\mathbf{Y}_i = \mathbf{1}, \mathbf{Z}_i, S_i) \\ &= \frac{Pr(\mathbf{Y}_i = 1|\mathbf{X}_i, \mathbf{Z}_i, S_i)Pr(\mathbf{X}_i, \mathbf{Z}_i|S_i)}{\sum_{\mathbf{X}^\star \in S_i} Pr(\mathbf{Y}_i = 1|\mathbf{X}^\star, \mathbf{Z}_i, S_i)Pr(\mathbf{X}^\star, \mathbf{Z}_i|S_i)}.\end{aligned}$$

Hoffmann et al. (2009) and Web Appendix A show that conditional independence of the genotype and exposure given the parents (P) implies conditional independence of the genotype and exposure given the sufficient statistic for parental mating type (i.e. $\mathbf{X} \perp \mathbf{Z}|P \Rightarrow \mathbf{X} \perp \mathbf{Z}|S$). Using this result, and the assumption in equation 2, we have

$$\begin{aligned}\mathcal{L}_i^{\text{TX}} &= \frac{\left\{\prod_j Pr(Y_{ij} = 1|\mathbf{X}_{ij}, Z_{ij}, S_i)\right\} Pr(\mathbf{X}_i|S_i)}{\sum_{\mathbf{X}^\star \in S_i} \left\{\prod_j Pr(Y_{ij} = 1|\mathbf{X}_j^\star, Z_{ij}, S_i)\right\} Pr(\mathbf{X}^\star|S_i)} \\ &\frac{e^{\boldsymbol{\beta}_{\text{ge}}^T \sum_j \mathbf{X}_{\text{ge},ij} Z_{ij} + \boldsymbol{\beta}_{\text{g}}^T \sum_j \mathbf{X}_{ij}} Pr(\mathbf{X}_i|S_i)}{\sum_{\mathbf{X}^\star \in S_i} e^{\boldsymbol{\beta}_{\text{ge}}^T \sum_j \mathbf{X}_{\text{ge},j}^\star Z_{ij} + \boldsymbol{\beta}_{\text{g}}^T \sum_j \mathbf{X}_j^\star} Pr(\mathbf{X}^\star|S_i)}.\end{aligned}$$

Then the log-likelihood is given by equation 4. Then the test statistic is computed as described in Section 2.1 following after equation 4. For the test statistic, we need a few derivatives that we present here now. The derivative of $\ell_{i,TX}$ is given in equation 5, where

$$\begin{aligned}&E\left\{\sum_j \begin{pmatrix} \mathbf{X}_{\text{ge},ij} Z_{ij} \\ \mathbf{X}_{ij} \end{pmatrix} \middle| \mathbf{Y}_i = 1, \mathbf{Z}_i, S_i; \boldsymbol{\beta}\right\} \\ &= \frac{\sum\limits_{\mathbf{X}^\star \in S_i} \left\{\sum\limits_j \begin{pmatrix} \mathbf{X}_{\text{ge},j}^\star Z_{ij} \\ \mathbf{X}_j^\star \end{pmatrix}\right\} e^{\boldsymbol{\beta}_{\text{ge}}^T \sum_j \mathbf{X}_{\text{ge},j}^\star Z_j + \boldsymbol{\beta}_{\text{g}}^T \sum_j \mathbf{X}_j^\star} Pr(\mathbf{X}^\star|S_i)}{\sum\limits_{\mathbf{X}^\star \in S_i} e^{\boldsymbol{\beta}_{\text{ge}}^T \sum_j \mathbf{X}_{\text{ge},j}^\star Z_j + \boldsymbol{\beta}_{\text{g}}^T \sum_j \mathbf{X}_j^\star} Pr(\mathbf{X}^\star|S_i)}.\end{aligned}$$

Lastly, to compute W_i, we need the following second derivatives. Let

$$A_{i,a,b} = \sum_{\mathbf{g}^\star \in S_i} \left(\sum_j \mathbf{X}_{\text{ge},j}^\star Z_{ij}\right)^{\otimes a} \left\{\left(\mathbf{X}_{\text{ge},j}^\star \mathbf{X}_j^\star\right)^{\otimes b}\right\}^T e^{\boldsymbol{\beta}_{\text{ge}}^T \sum_j \mathbf{X}_{\text{ge},j}^\star Z_{ij} + \boldsymbol{\beta}_{\text{g}}^T \sum_j \mathbf{X}_j^\star} Pr(\mathbf{X}^\star|S_i),$$

where we define $M^{\otimes 0} = 1$, $M^{\otimes 1} = M$, and $M^{\otimes 2} = MM^T$. Then we have

$$\frac{\partial}{\partial \beta_{\text{nuis}}} U_i^{\beta_{\text{ge}}}(\boldsymbol{\beta}) = \frac{A_{i,1,1}A_{i,0,0} - A_{i,1,0}A_{i,0,1}}{A_{i,0,0}^2} \tag{1}$$

$$\frac{\partial}{\partial \beta_{\text{nuis}}} U_i^{\beta_{\text{e}}}(\boldsymbol{\beta}) = \frac{A_{i,0,2}^T A_{i,0,0} - \left(A_{i,0,1}^T\right)^{\otimes 2}}{A_{i,0,0}^2}. \tag{2}$$

Web Appendix C: Distribution of W_i

From a Taylor series expansion, we have that

$$\begin{aligned}
n^{-1/2}\sum_i U_i^{\text{ge}}(\boldsymbol{\beta}_{\text{ge}}, \hat{\boldsymbol{\beta}}_{\text{nuis}}) &= n^{-1/2}\sum_i U_i^{\text{ge}}(\boldsymbol{\beta}_{\text{ge}}, \boldsymbol{\beta}_{\text{nuis}}) \\
&\quad + n^{-1/2}\sum_i E\left\{\frac{\partial}{\partial \boldsymbol{\beta}_{\text{nuis}}} U_i^{\text{ge}}(\boldsymbol{\beta}_{\text{ge}}, \boldsymbol{\beta}_{\text{nuis}})\right\}\left(\hat{\boldsymbol{\beta}}_{\text{nuis}} - \boldsymbol{\beta}_{\text{nuis}}\right) + o_p(1) \\
n^{-1/2}\sum_i U_i^{\text{nuis}}(\boldsymbol{\beta}_{\text{ge}}, \hat{\boldsymbol{\beta}}_{\text{nuis}}) &= n^{-1/2}\sum_i U_i^{\text{nuis}}(\boldsymbol{\beta}_{\text{ge}}, \boldsymbol{\beta}_{\text{nuis}}) \\
&\quad + n^{-1/2}\sum_i E\left\{\frac{\partial}{\partial \boldsymbol{\beta}_{\text{nuis}}} U_i^{\text{nuis}}(\boldsymbol{\beta}_{\text{ge}}, \boldsymbol{\beta}_{\text{nuis}})\right\}\left(\hat{\boldsymbol{\beta}}_{\text{nuis}} - \boldsymbol{\beta}_{\text{nuis}}\right) + o_p(1) \\
&= n^{-1/2}\sum_i E\left\{\frac{\partial}{\partial \boldsymbol{\beta}_{\text{nuis}}} U_i^{\text{nuis}}(\boldsymbol{\beta}_{\text{ge}}, \boldsymbol{\beta}_{\text{nuis}})\right\}\left(\hat{\boldsymbol{\beta}}_{\text{nuis}} - \boldsymbol{\beta}_{\text{nuis}}\right) + o_p(1).
\end{aligned}$$

Hence $n^{-1/2}\sum_i U_i^{\text{ge}}(\boldsymbol{\beta}_{\text{ge}}, \hat{\boldsymbol{\beta}}_{\text{nuis}}) =$

$$\begin{aligned}
&n^{-1/2}\sum_i U_i^{\text{ge}}(\boldsymbol{\beta}_{\text{ge}}, \boldsymbol{\beta}_{\text{nuis}}) \\
&+ n^{-1/2}\sum_i E\left\{\frac{\partial}{\partial \boldsymbol{\beta}_{\text{nuis}}} U_i^{\text{ge}}(\boldsymbol{\beta}_{\text{ge}}, \boldsymbol{\beta}_{\text{nuis}})\right\} E\left\{\frac{\partial}{\partial \boldsymbol{\beta}_{\text{nuis}}} U_i^{\text{nuis}}(\boldsymbol{\beta}_{\text{ge}}, \boldsymbol{\beta}_{\text{nuis}})\right\}^{-} U_i^{\text{nuis}}(\boldsymbol{\beta}_{\text{ge}}, \hat{\boldsymbol{\beta}}_{\text{nuis}}) \\
&+ o_p(1).
\end{aligned} \tag{3}$$

It now follows from the central limit theorem that W_i follows an asymptotically multivariate normal distribution, with variance that can be consistently estimated by the right hand side of the expression in web equation 3, replacing $\boldsymbol{\beta}_{\text{nuis}}$ with $\hat{\boldsymbol{\beta}}_{\text{nuis}}$. The test has $\text{rank}(\sum_i W_i W_i^T)$ degrees of freedom.

Web Appendix D: Further details of CLR-IJ

Here we fill in the missing details in the derivation in section 2.2.2. We continue from equation 7 in the paper. Calculations follow Chatterjee et al. (2005), but use the more general sibship. We have

$$\begin{aligned}\mathcal{L}_i^{\text{CLR-IJ}} &= Pr\{\mathbf{Y}_{i,A(\mathbf{Y}_i)} = \mathbf{1}, \mathbf{Y}_{i,U(\mathbf{Y}_i)} = \mathbf{0}, \mathbf{X}_i | \mathbf{Y}_{i+}, S_i, \mathbf{Z}_i, \mathbf{C}_i; \boldsymbol{\beta}\} \\ &= Pr\{\mathbf{Y}_{i,A(\mathbf{Y}_i)} = \mathbf{1}, \mathbf{Y}_{i,U(\mathbf{Y}_i)} = \mathbf{0} | \mathbf{X}_i, \mathbf{Y}_{i+}, S_i, \mathbf{Z}_i, \mathbf{C}_i; \boldsymbol{\beta}\} Pr(\mathbf{X}_i | \mathbf{Y}_{i+}, S_i, \mathbf{Z}_i, \mathbf{C}_i; \boldsymbol{\beta}) \\ &:= \mathcal{L}_i^{\text{CLR}} \mathcal{L}_i^\star,\end{aligned}$$

where $\mathcal{L}_i^{\text{CLR}}$ is the likelihood from conditional logistic regression (Witte et al., 1999; Siegmund et al., 2000; Weinberg, 2000). The term $\mathcal{L}_i^\star$ may resemble the term in $\mathcal{L}_i^{\text{TX}}$, except that the latter only includes affected offspring, whereas the former used here includes all offspring. Now, working with just the second term, we have

$$\begin{aligned}\mathcal{L}_i^\star &= \frac{Pr(\mathbf{Y}_{i+} | \mathbf{X}_i, S_i, \mathbf{Z}_i, \mathbf{C}_i, \boldsymbol{\beta}) Pr(\mathbf{X}_i | S_i, \mathbf{Z}_i, \mathbf{C}_i, \boldsymbol{\beta})}{\sum_{\mathbf{X}^\star \in S_i} Pr(\mathbf{Y}_{i+} | \mathbf{X}_i^\star, S_i, \mathbf{Z}_i, \mathbf{C}_i, \boldsymbol{\beta}) Pr(\mathbf{X}_i^\star | S_i, \mathbf{Z}_i, \mathbf{C}_i, \boldsymbol{\beta})} \\ &= \frac{Pr(\mathbf{Y}_{i+} | \mathbf{X}_i, S_i, \mathbf{Z}_i, \mathbf{C}_i, \boldsymbol{\beta}) Pr(\mathbf{X}_i | S_i)}{\sum_{\mathbf{X}^\star \in S_i} Pr(\mathbf{Y}_{i+} | \mathbf{X}_i^\star, S_i, \mathbf{Z}_i, \mathbf{C}_i, \boldsymbol{\beta}) Pr(\mathbf{X}_i^\star | S_i)}\end{aligned}$$

by the assumption that $\mathbf{X} \perp \mathbf{Z} | P$ (which implies $\mathbf{X} \perp \mathbf{Z} | S$ as shown in Hoffmann et al. (2009) and Web Appendix A). Next, assuming phenotypic independence of the sibs and using equation 1 we have that $Pr(\mathbf{Y}_{i+} | \mathbf{X}_i, S_i, \mathbf{Z}_i, \mathbf{C}_i; \boldsymbol{\beta})$

$$\begin{aligned}&= \sum_{\mathbf{Y}^\star : \mathbf{Y}_+^\star = \mathbf{Y}_{i+}} Pr(\mathbf{Y}_i = \mathbf{Y}^\star | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{C}_i; \boldsymbol{\beta}) \\ &= \sum_{\mathbf{Y}^\star : \mathbf{Y}_+^\star = \mathbf{Y}_{i+}} \left\{ \prod_{j \in A(\mathbf{Y}^\star)} Pr(Y_j = \mathbf{1} | \mathbf{X}_{ij}, Z_{ij}, \mathbf{C}_{ij}; \boldsymbol{\beta}) \right\} \left\{ \prod_{j \in U(\mathbf{Y}^\star)} Pr(\mathbf{Y}_j = \mathbf{0} | \mathbf{X}_{ij}, Z_{ij}, \mathbf{C}_{ij}; \boldsymbol{\beta}) \right\} \\ &= \sum_{\mathbf{Y}^\star : \mathbf{Y}_+^\star = \mathbf{Y}_{i+}} \frac{\prod_{j \in A(\mathbf{Y}^\star)} e^{\alpha_i} e^{h(\mathbf{X_{ij}}, Z_{ij}, \mathbf{C}_{ij}; \boldsymbol{\beta})}}{\prod_{\mathbf{Y}^\star : \mathbf{Y}_+^\star = y_{i+}} \left\{ 1 + \prod_{j \in \mathbf{Y}^\star} e^{\alpha_i} e^{h(\mathbf{X_{ij}}, Z_{ij}, \mathbf{C}_{ij}; \boldsymbol{\beta})} \right\}} \\ &\approx \sum_{\mathbf{Y}^\star : \mathbf{Y}_+^\star = \mathbf{Y}_{i+}} \left\{ \prod_{j \in A(\mathbf{Y}^\star)} e^{\alpha_i} e^{h(\mathbf{X_{ij}}, Z_{ij}, \mathbf{C}_{ij}; \beta)} \right\}\end{aligned}$$

under the rare disease assumption, where

$$h(\mathbf{X_{ij}}, Z_{ij}, \mathbf{C}_{ij}; \beta) = \boldsymbol{\beta}_{\text{ge}}^T \mathbf{X}_{\text{ge},j}^\star Z_{ij} + \boldsymbol{\beta}_{\text{nuis}}^T \mathbf{m}(\mathbf{X}_j^\star, Z_{ij}, \mathbf{C}_{ij}).$$

Then we have that

$$\mathcal{L}_i^\star = \frac{\left\{\sum_{\mathbf{Y}^\star:\mathbf{Y}_+^\star=\mathbf{Y}_{i+}} \prod_{j\in A(\mathbf{Y}^\star)} e^{h(\mathbf{X}_{ij},Z_{ij},\mathbf{C}_{ij};\boldsymbol{\beta})}\right\} Pr(\mathbf{X}_i|S_i)}{\sum_{\mathbf{X}^\star\in S_i}\left\{\sum_{\mathbf{Y}^\star:\mathbf{Y}_+^\star=\mathbf{Y}_{i+}} \prod_{j\in A(\mathbf{Y}^\star)} e^{h(\mathbf{X}_{ij},Z_{ij},\mathbf{C}_{ij};\boldsymbol{\beta})}\right\} Pr(\mathbf{X}_i^\star|S_i)}.$$

Putting these two back together leaves us with

$$\begin{aligned}\mathcal{L}_i^{\text{CLR-IJ}} &= \frac{\left\{\prod_{j\in A(\mathbf{Y}_i)} e^{h(\mathbf{X}_{ij},Z_{ij},\mathbf{C}_{ij};\boldsymbol{\beta})}\right\} Pr(\mathbf{X}_i|S_i)}{\sum_{\mathbf{X}^\star\in S_i}\left\{\sum_{\mathbf{Y}^\star:\mathbf{Y}_+^\star=\mathbf{Y}_{i+}} \prod_{j\in A(\mathbf{Y}^\star)} e^{h(\mathbf{X}_{ij},Z_{ij},\mathbf{C}_{ij};\boldsymbol{\beta})}\right\} Pr(\mathbf{X}_i^\star|S_i)}\\ &= \frac{e^{\sum_{j\in A(\mathbf{Y}_i)} h(\mathbf{X}_{ij},Z_{ij},\mathbf{C}_{ij};\boldsymbol{\beta})} Pr(\mathbf{X}_i|S_i)}{\sum_{\mathbf{X}^\star\in S_i,\mathbf{Y}^\star:\mathbf{Y}_+^\star=\mathbf{Y}_{i+}} e^{\sum_{j\in A(\mathbf{Y}^\star)} h(\mathbf{X}_{ij},Z_{ij},\mathbf{C}_{ij};\boldsymbol{\beta})} Pr(\mathbf{X}_i^\star|S_i)}.\end{aligned}$$

Then the log-likelihood is given by equation 7 in the paper. Then the test statistic is computed as described in section 2.2.2, with the derivatives given here. The derivative of the log-likelihood is as given in equation 11, where

$$\begin{aligned}&E\left[\sum_{j\in A(\mathbf{Y}_i)}\left\{\begin{matrix}\mathbf{X}_{\text{ge},ij}Z_{ij}\\ \mathbf{m}(\mathbf{X}_{ij},Z_{ij},\mathbf{C}_{ij})\end{matrix}\right\}\middle|\mathbf{Y}_{i+},\mathbf{Z}_i,\mathbf{C}_i,S_i;\boldsymbol{\beta}\right]\\ &= \frac{\sum_{\substack{\mathbf{X}^\star\in S_i,\\ \mathbf{Y}^\star:\mathbf{Y}_+^\star=\mathbf{Y}_{i+}}}\left[\sum_{j\in A(\mathbf{Y}^\star)}\left\{\begin{matrix}\mathbf{X}^\star_{\text{ge},j}Z_{ij}\\ \mathbf{m}(\mathbf{X}^\star_j,Z_{ij},\mathbf{C}_{ij})\end{matrix}\right\}\right] e^{\sum_{j\in A(\mathbf{Y}^\star)}\boldsymbol{\beta}_{\text{ge}}^T\mathbf{X}^\star_{\text{ge},j}Z_{ij}+\boldsymbol{\beta}_{\text{nuis}}^T\mathbf{m}(\mathbf{X}^\star_j,Z_{ij},\mathbf{C}_{ij})} Pr(\mathbf{X}^\star|S_i)}{\sum_{\substack{\mathbf{X}^\star\in S_i,\\ \mathbf{Y}^\star:\mathbf{Y}_+^\star=\mathbf{Y}_{i+}}} e^{\sum_{j\in A(\mathbf{Y}^\star)}\boldsymbol{\beta}_{\text{ge}}^T\mathbf{X}^\star_{\text{ge},j}Z_{ij}+\boldsymbol{\beta}_{\text{nuis}}^T\mathbf{m}(\mathbf{X}^\star_j,Z_{ij},\mathbf{C}_{ij})} Pr(\mathbf{X}^\star|S_i)}.\end{aligned}$$

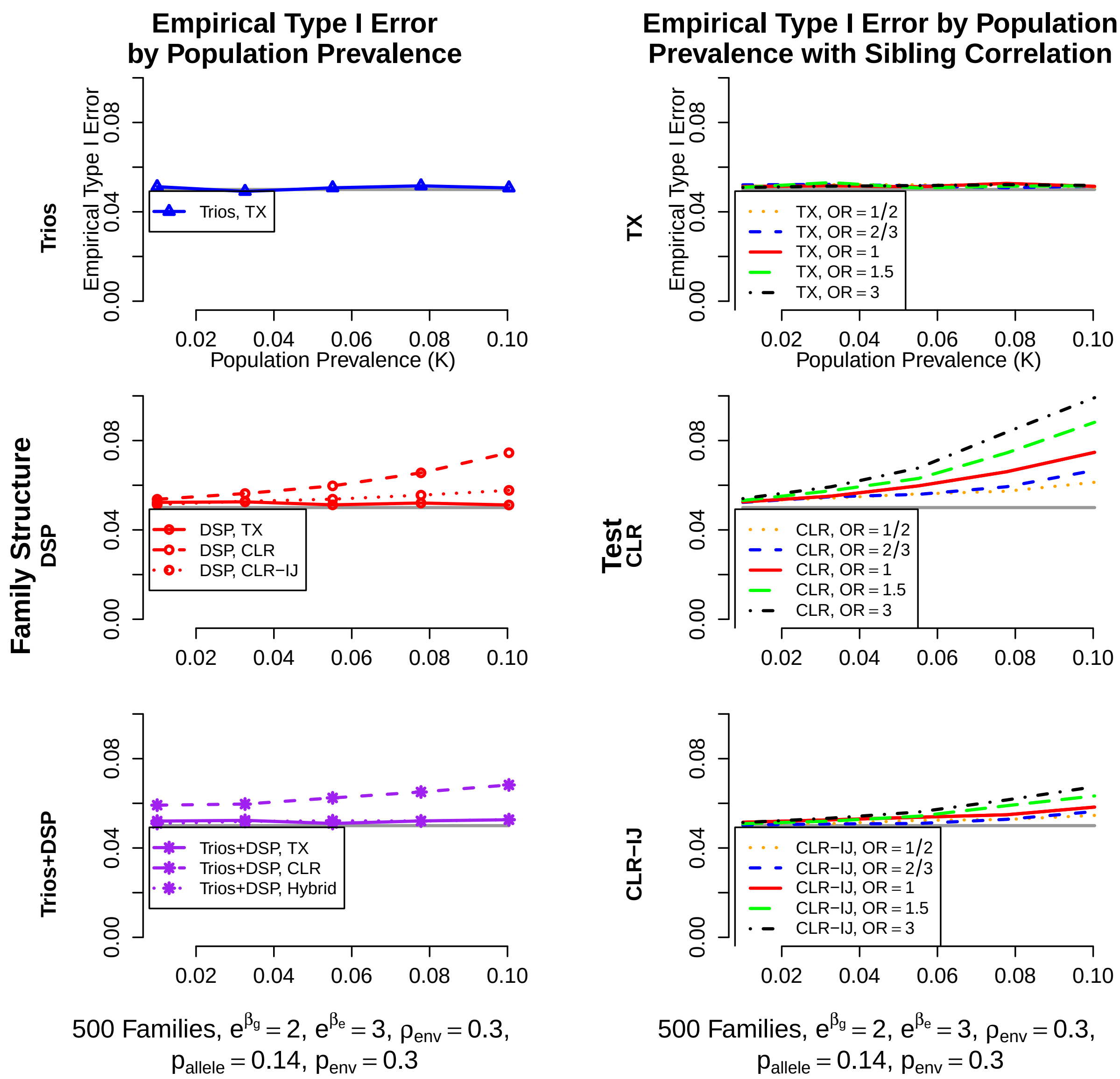
For the second derivatives, redefine

$$\begin{aligned}A_{i,a,b} = \sum_{\mathbf{X}^\star\in S_i,\mathbf{Y}^\star:\mathbf{Y}_+^\star=\mathbf{Y}_{i+}}&\left(\sum_j \mathbf{X}^\star_{\text{ge},j}Z_{ij}\right)^{\otimes a}\left[\left\{\sum_j \mathbf{m}(\mathbf{X}^\star_j,Z_{ij},C_{ij})\right\}^{\otimes b}\right]^T\\ &\times e^{\sum_{j\in A(\mathbf{Y}^\star)}\beta_{ge}\mathbf{X}^\star_{\text{ge},j}Z_{ij}+\beta_g^T\mathbf{X}^\star_j} Pr(\mathbf{X}^\star|S_i).\end{aligned}$$

Then the second derivatives are given by web equations 1 and 2.

Web Appendix E: Additional type I error simulations

Web Figure 1 shows further simulation results looking at type I error under the log link by population prevalence in Web Figure 1(a) and when there is arbitrary phenotype correlation in Web Figure 1(b).



(a) Simulation study to assess type I error for a dichotomous exposure with strong main effects for 500 families under the log link. Based on 100,000 simulations. The gray line is drawn at 0.05.

(b) Simulation study to assess type I error for the effect of arbitrary phenotypic correlation in DSP (failure of assumption 2) under the log link. Based on 100,000 simulations.

Figure 1. Type I error and phenotypic model robustness simulations under the log link.

Web Appendix F: Further power results

In Figure 2 we show the effect of varying the environmental exposure prevalence. The case-control power approaches and surpasses the TX approach for trios as the environmental exposure prevalence increases.

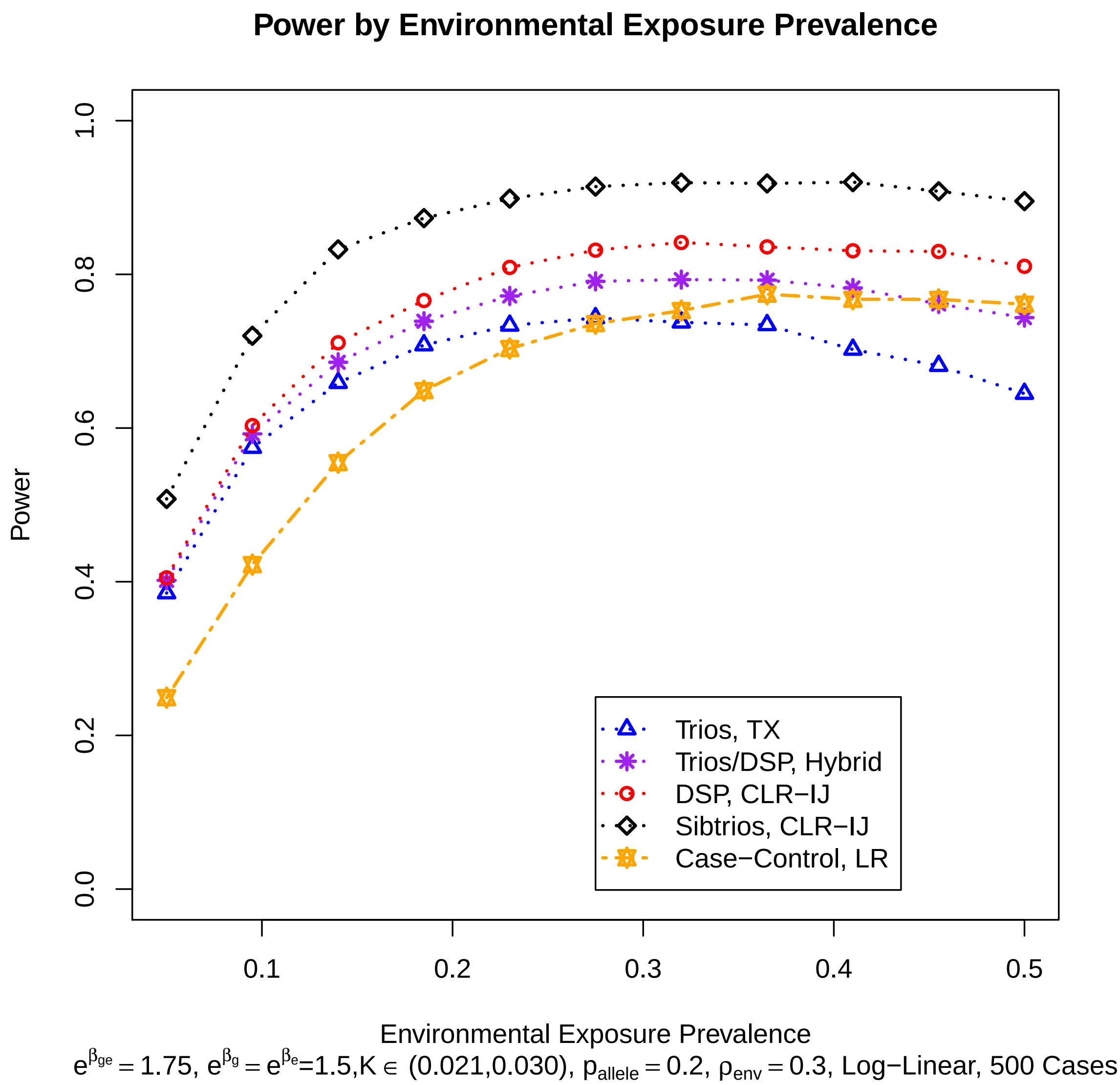


Figure 2. Power of the test for a dichotomous trait by environmental exposure. Based on 10,000 simulations; approximate SE < 0.0025.

Web Appendix G: Additional dataset information

In Web Figure 3 we show the SNPs that had joint test p-value < 0.15, that is the SNPs presented in Table 2.

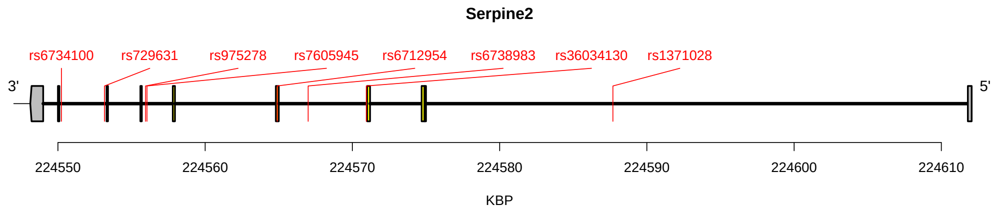


Figure 3. Plot of the SNPs in the Serpine2 gene with joint test p-values < 0.15, as shown in Table 2.

References

Chatterjee, N., Kalaylioglu, Z., and Carroll, R. J. (2005). Exploiting gene-environment independence in family-based case-control studies: increased power for detecting associations, interactions and joint effects. *Genet Epidemiol* **28,** 138–156.

Hoffmann, T. J., Lange, C., Vansteelandt, S., and Laird, N. M. (2009). Gene-environment interaction tests for dichotomous traits in trios and sibships. *Genet Epidemiol* **33,** 691–699.

Siegmund, K. D., Langholz, B., Kraft, P., and Thomas, D. C. (2000). Testing linkage disequilibrium in sibships. *Am J Hum Genet* **67,** 244–248.

Weinberg, C. (2000). Re: "asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs". *Am J Epidemiol* **152,** 689–691.

Witte, J. S., Gauderman, W. J., and Thomas, D. C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am J Epidemiol* **149,** 693–705.