APPENDIX / AVAILABLE FROM THE AUTHRORS

It is assumed that the study population is a stratified population with a total of *J* strata ($j = 1,2,...,J$). In each stratum, the occurrence of disease follows a log-linear risk model, that is, for a subject in the *j*th stratum, $\log(\text{risk}) = \alpha_j + \mathbf{x}^{\text{t}}_{(\text{gene, exposure, sex})}\boldsymbol{\beta}$, where α_j represents the log background risk for the *j*th stratum, $\mathbf{x}_{(\text{gene, exposure, sex})}$ is a vector of data or codes regarding the gene, the environmental exposure, the sex, and any possible cross-product terms between them, for the subject under concern ($\mathbf{x}^{\text{t}}$ is the transpose of $\mathbf{x}$), and $\boldsymbol{\beta}$ is a vector of parameters of present interest (to be estimated).

Supposed that a sample of *n* (*i*=1,…,*n*) case-spouse pairs has been recruited. Let T_i (T'_i) represent the stratum to which the *i*th proband(his/her spouse) belongs. Let $\text{G}_i = \text{A}_{1i}\text{A}_{2i}$ ($\text{G}'_i = \text{A}_{3i}\text{A}_{4i}$) represent the genotype for the *i*th proband(his/her spouse)($\text{A}_{1i}, \text{A}_{2i}, \text{A}_{3i}$, and A_{4i} are 'alleles'). Let D_i represent the event that the *i*th proband is a case (i.e., he/she contracts the disease). Let $\text{E}_i(\text{S}_i)$ represent the environmental exposure(sex) of the *i*th proband, and $\text{E}'_i(\text{S}'_i)$, the environmental exposure(sex) of his/her spouse. Further, we let U_i represent the set that contains as its two elements the genotype of the *i*th proband and the genotype of the *i*th spouse, that is, $\text{U}_i = \{\text{A}_{1i}\text{A}_{2i}, \text{A}_{3i}\text{A}_{4i}\}$. And we let V_i be the set containing as its elements the four alleles in the *i*th case-spouse pair, that is, $\text{V}_i = \{\text{A}_{1i}, \text{A}_{2i}, \text{A}_{3i}, \text{A}_{4i}\}$.

Conditioned on $\text{U}_i, \text{D}_i, \text{E}_i, \text{E}'_i, \text{S}_i$, and S'_i, the probability that the *i*th proband has genotype of $\text{A}_{1i}\text{A}_{2i}$ and the *i*th spouse has genotype of $\text{A}_{3i}\text{A}_{4i}$ is denoted as Q_i. With elementary algebra, this conditional probability is (the index i was suppressed for simplicity):

$$\text{Q} = \Pr(\text{G} = \text{A}_1\text{A}_2, \text{G}' = \text{A}_3\text{A}_4 \,|\, \text{U}, \text{D}, \text{E}, \text{E}', \text{S}, \text{S}')$$

$$= \sum_{j=1}^{J} \Pr(\text{T} = j \,|\, \text{D}, \text{E}, \text{E}', \text{S}, \text{S}') \cdot \frac{\Pr(\text{G} = \text{A}_1\text{A}_2, \text{G}' = \text{A}_3\text{A}_4, \text{D}, \text{E}, \text{E}', \text{S}, \text{S}', \text{T} = j)}{\left\{ \begin{array}{l} \Pr(\text{G} = \text{A}_1\text{A}_2, \text{G}' = \text{A}_3\text{A}_4, \text{D}, \text{E}, \text{E}', \text{S}, \text{S}', \text{T} = j) \\ + \Pr(\text{G} = \text{A}_3\text{A}_4, \text{G}' = \text{A}_1\text{A}_2, \text{D}, \text{E}, \text{E}', \text{S}, \text{S}', \text{T} = j) \end{array} \right\}}.$$

The numerator can be expressed as the product of three terms:

$$\Pr(G = A_1A_2, G' = A_3A_4, D, E, E', S, S', T = j) = \Pr(D|G = A_1A_2, G' = A_3A_4, E, E', S, S', T = j) \cdot \Pr(G = A_1A_2, G' = A_3A_4 \mid E, E', S, S', T = j) \cdot \Pr(E, E', S, S', T = j).$$

With the assumed disease model, the first term of the product is

$$\begin{aligned}\Pr(D|G = A_1A_2, G' = A_3A_4, E, E', S, S', T = j) &= \Pr(D|G = A_1A_2, E, S, T = j) \\ &= \exp[\alpha_j + \mathbf{x}^{t}_{(A_1A_2,E,S)}\boldsymbol{\beta}].\end{aligned}$$

The second term of the product can be shown to be (the parenthesis shows the assumption used in each step of derivation; A1: mating is restricted to subjects in the same stratum; A2: mating is independent to genotypes, for males and females in each and every stratum; A3: environmental exposures are independent to genotypes, for males and females in each and every stratum; A4: the genotype frequencies for males are equal to the corresponding frequencies for females, in each and every stratum):

$$\begin{aligned}&\Pr(G = A_1A_2, G' = A_3A_4 \mid E, E', S, S', T = j) \\ &= \Pr(G = A_1A_2, G' = A_3A_4 \mid E, E', S, S', T = T' = j) && \text{(A1)} \\ &= \Pr(G = A_1A_2 \mid E, S, T = j) \cdot \Pr(G' = A_3A_4 \mid E', S', T' = j) && \text{(A2)} \\ &= \Pr(G = A_1A_2 \mid S, T = j) \cdot \Pr(G' = A_3A_4 \mid S', T' = j) && \text{(A3)} \\ &= \Pr(G = A_1A_2 \mid T = j) \cdot \Pr(G = A_3A_4 \mid T = j). && \text{(A4)}\end{aligned}$$

Therefore, we see that the numerator in Q is

$$\Pr(G = A_1A_2, G' = A_3A_4, D, E, E', S, S', T = j) = \exp[\alpha_j + \mathbf{x}^{t}_{(A_1A_2,E,S)}\boldsymbol{\beta}] \cdot \Pr(G = A_1A_2 \mid T = j) \cdot \Pr(G = A_3A_4 \mid T = j) \cdot \Pr(E, E', S, S', T = j).$$

Similarly, we can show that the second term of the denominator in Q is

$$\Pr(G = A_3A_4, G' = A_1A_2, D, E, E', S, S', T = j) = \exp[\alpha_j + \mathbf{x}^{t}_{(A_3A_4,E,S)}\boldsymbol{\beta}] \cdot \Pr(G = A_3A_4 \mid T = j) \cdot \Pr(G = A_1A_2 \mid T = j) \cdot \Pr(E, E', S, S', T = j).$$

Thus, we have (with the index i denoting the ith case-spouse pair)

$$\mathrm{Q}_i = \sum_{j=1}^{J} \Pr(\mathrm{T}_i = j \,|\, \mathrm{D}_i, \mathrm{E}_i, \mathrm{E}_i', \mathrm{S}_i, \mathrm{S}_i') \cdot \frac{\exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_{1i}\mathrm{A}_{2i},\mathrm{E}_i,\mathrm{S}_i)}\boldsymbol{\beta}]}{\exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_{1i}\mathrm{A}_{2i},\mathrm{E}_i,\mathrm{S}_i)}\boldsymbol{\beta}] + \exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_{3i}\mathrm{A}_{4i},\mathrm{E}_i,\mathrm{S}_i)}\boldsymbol{\beta}]}$$
$$= \frac{\exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_{1i}\mathrm{A}_{2i},\mathrm{E}_i,\mathrm{S}_i)}\boldsymbol{\beta}]}{\exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_{1i}\mathrm{A}_{2i},\mathrm{E}_i,\mathrm{S}_i)}\boldsymbol{\beta}] + \exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_{3i}\mathrm{A}_{4i},\mathrm{E}_i,\mathrm{S}_i)}\boldsymbol{\beta}]}.$$

And the conditional likelihood function for the case-spouse data (1:1 case-counterfactual-control analysis) is

$$L_{1:1} = \prod_{i=1}^{n} \mathrm{Q}_i.$$

Let F_i (F_i') represent the allele that the *i*th proband(spouse) inherited from his/her father, and M_i (M_i'), the allele that the *i*th proband(spouse) inherited from his/her mother. Conditioned on V_i, D_i, E_i, $\mathrm{E'}_i$, S_i, and $\mathrm{S'}_i$, the probability that the *i*th proband has genotype of $\mathrm{A}_{1i}\mathrm{A}_{2i}$ and the *i*th spouse has genotype of $\mathrm{A}_{3i}\mathrm{A}_{4i}$ is denoted as R_i. With elementary algebra, this conditional probability is (the index i was suppressed):

$$\mathrm{R} = \Pr(\mathrm{G} = \mathrm{A}_1\mathrm{A}_2, \mathrm{G'} = \mathrm{A}_3\mathrm{A}_4 \,|\, \mathrm{V}, \mathrm{D}, \mathrm{E}, \mathrm{E'}, \mathrm{S}, \mathrm{S'})$$
$$= \sum_{j=1}^{J} \Pr(\mathrm{T} = j \,|\, \mathrm{D}, \mathrm{E}, \mathrm{E}', \mathrm{S}, \mathrm{S}')$$
$$\cdot \frac{\Pr(\mathrm{G} = \mathrm{A}_1\mathrm{A}_2, \mathrm{G}' = \mathrm{A}_3\mathrm{A}_4, \mathrm{D}, \mathrm{E}, \mathrm{E}', \mathrm{S}, \mathrm{S}', \mathrm{T} = j)}{\sum_{h}\sum_{k\neq h}\sum_{\substack{t\neq k\\ t\neq h}}\sum_{\substack{u\neq t\\ u\neq k\\ u\neq h}} \Pr(\mathrm{F} = \mathrm{A}_h, \mathrm{M} = \mathrm{A}_k, \mathrm{F}' = \mathrm{A}_t, \mathrm{M}' = \mathrm{A}_u, \mathrm{D}, \mathrm{E}, \mathrm{E}', \mathrm{S}, \mathrm{S}', \mathrm{T} = j)}$$

$$= \sum_{j=1}^{J} \Pr(\mathrm{T} = j \,|\, \mathrm{D}, \mathrm{E}, \mathrm{E}', \mathrm{S}, \mathrm{S}')$$
$$\cdot \frac{\Pr(\mathrm{G} = \mathrm{A}_1\mathrm{A}_2, \mathrm{G}' = \mathrm{A}_3\mathrm{A}_4, \mathrm{D}, \mathrm{E}, \mathrm{E}', \mathrm{S}, \mathrm{S}', \mathrm{T} = j)}{\left\{\begin{array}{l} \sum_{h}\sum_{k\neq h}\sum_{\substack{t\neq k\\ t\neq h}}\sum_{\substack{u\neq t\\ u\neq k\\ u\neq h}} \Pr(\mathrm{D}|\mathrm{F} = \mathrm{A}_h, \mathrm{M} = \mathrm{A}_k, \mathrm{F}' = \mathrm{A}_t, \mathrm{M}' = \mathrm{A}_u, \mathrm{E}, \mathrm{E}', \mathrm{S}, \mathrm{S}', \mathrm{T} = j) \\ \quad \cdot \Pr(\mathrm{F} = \mathrm{A}_h, \mathrm{M} = \mathrm{A}_k, \mathrm{F}' = \mathrm{A}_t, \mathrm{M}' = \mathrm{A}_u \,|\, \mathrm{E}, \mathrm{E}', \mathrm{S}, \mathrm{S}', \mathrm{T} = j) \\ \quad \cdot \Pr(\mathrm{E}, \mathrm{E}', \mathrm{S}, \mathrm{S}', \mathrm{T} = j) \end{array}\right\}}.$$

The numerator of R is the same as that of Q and has been previously shown to be (with assumptions A1~A4)

$$\begin{aligned}\Pr(\mathrm{G}=\mathrm{A}_1\mathrm{A}_2,\mathrm{G}'=\mathrm{A}_3\mathrm{A}_4,\mathrm{D},\mathrm{E},\mathrm{E}',\mathrm{S},\mathrm{S}',\mathrm{T}=j) &= \exp[\alpha_j+\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_1\mathrm{A}_2,\mathrm{E},\mathrm{S})}\boldsymbol{\beta}]\\ &\quad\cdot\Pr(\mathrm{G}=\mathrm{A}_1\mathrm{A}_2\,|\,\mathrm{T}=j)\cdot\Pr(\mathrm{G}=\mathrm{A}_3\mathrm{A}_4\,|\,\mathrm{T}=j)\\ &\quad\cdot\Pr(\mathrm{E},\mathrm{E}',\mathrm{S},\mathrm{S}',\mathrm{T}=j).\end{aligned}$$

With the A5 assumption that there is no imprinting effect for the gene under study, the first term of the product in the denominator of R is

$$\begin{aligned}&\Pr(\mathrm{D}|\mathrm{F}=\mathrm{A}_h,\mathrm{M}=\mathrm{A}_k,\mathrm{F}'=\mathrm{A}_t,\mathrm{M}'=\mathrm{A}_u,\mathrm{E},\mathrm{E}',\mathrm{S},\mathrm{S}',\mathrm{T}=j)\\ &\quad=\Pr(\mathrm{D}|\mathrm{F}=\mathrm{A}_h,\mathrm{M}=\mathrm{A}_k,\mathrm{E},\mathrm{S},\mathrm{T}=j)\\ &\quad=\exp[\alpha_j+\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_h\mathrm{A}_k,\mathrm{E},\mathrm{S})}\boldsymbol{\beta}].\end{aligned}$$

With the A1~A4 assumptions, the second term of the product in the denominator of R is

$$\begin{aligned}&\Pr(\mathrm{F}=\mathrm{A}_h,\mathrm{M}=\mathrm{A}_k,\mathrm{F}'=\mathrm{A}_t,\mathrm{M}'=\mathrm{A}_u\,|\,\mathrm{E},\mathrm{E}',\mathrm{S},\mathrm{S}',\mathrm{T}=j)\\ &\quad=\Pr(\mathrm{F}=\mathrm{A}_h,\mathrm{M}=\mathrm{A}_k\,|\,\mathrm{T}=j)\cdot\Pr(\mathrm{F}=\mathrm{A}_t,\mathrm{M}=\mathrm{A}_u\,|\,\mathrm{T}=j).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{R}&=\sum_{j=1}^{J}\Pr(\mathrm{T}=j\,|\,\mathrm{D},\mathrm{E},\mathrm{E}',\mathrm{S},\mathrm{S}')\\ &\quad\cdot\frac{\exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_1\mathrm{A}_2,\mathrm{E},\mathrm{S})}\boldsymbol{\beta}]\cdot\Pr(\mathrm{G}=\mathrm{A}_1\mathrm{A}_2\,|\,\mathrm{T}=j)\cdot\Pr(\mathrm{G}=\mathrm{A}_3\mathrm{A}_4\,|\,\mathrm{T}=j)}{\sum_{h}\sum_{k\neq h}\sum_{\substack{t\neq k\\ t\neq h}}\sum_{\substack{u\neq t\\ u\neq k\\ u\neq h}}\exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_h\mathrm{A}_k,\mathrm{E},\mathrm{S})}\boldsymbol{\beta}]\cdot\Pr(\mathrm{F}=\mathrm{A}_h,\mathrm{M}=\mathrm{A}_k\,|\,\mathrm{T}=j)\cdot\Pr(\mathrm{F}=\mathrm{A}_t,\mathrm{M}=\mathrm{A}_u\,|\,\mathrm{T}=j)}\\ &=\sum_{j=1}^{J}\Pr(\mathrm{T}=j\,|\,\mathrm{D},\mathrm{E},\mathrm{E}',\mathrm{S},\mathrm{S}')\\ &\quad\cdot\frac{\left\{\begin{array}{r}\exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_1\mathrm{A}_2,\mathrm{E},\mathrm{S})}\boldsymbol{\beta}]\cdot\left[\Pr(\mathrm{F}=\mathrm{A}_1,\mathrm{M}=\mathrm{A}_2|\mathrm{T}=j)+\Pr(\mathrm{F}=\mathrm{A}_2,\mathrm{M}=\mathrm{A}_1|\mathrm{T}=j)\right]\\ \cdot\left[\Pr(\mathrm{F}=\mathrm{A}_3,\mathrm{M}=\mathrm{A}_4|\mathrm{T}=j)+\Pr(\mathrm{F}=\mathrm{A}_4,\mathrm{M}=\mathrm{A}_3|\mathrm{T}=j)\right]\end{array}\right\}}{\sum_{h}\sum_{k\neq h}\sum_{\substack{t\neq k\\ t\neq h}}\sum_{\substack{u\neq t\\ u\neq k\\ u\neq h}}\exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_h\mathrm{A}_k,\mathrm{E},\mathrm{S})}\boldsymbol{\beta}]\cdot\Pr(\mathrm{F}=\mathrm{A}_h,\mathrm{M}=\mathrm{A}_k\,|\,\mathrm{T}=j)\cdot\Pr(\mathrm{F}=\mathrm{A}_t,\mathrm{M}=\mathrm{A}_u\,|\,\mathrm{T}=j)}.\end{aligned}$$

With the A6 assumption that each and every stratum is in Hardy-Weinberg equilibrium, we have

$$\Pr(\mathrm{F}=x,\mathrm{M}=y\,|\,\mathrm{T}=j)=\Pr(\mathrm{F}=x\,|\,\mathrm{T}=j)\cdot\Pr(\mathrm{M}=y\,|\,\mathrm{T}=j),$$

for arbitrary alleles x and y and arbitrary stratum j. Thus we see that (with the index i denoting the ith case-spouse pair):

$$\mathrm{R}_i = \sum_{j=1}^{J} \Pr(\mathrm{T}_i = j \,|\, \mathrm{D}_i, \mathrm{E}_i, \mathrm{E}_i', \mathrm{S}_i, \mathrm{S}_i') \cdot \frac{4 \cdot \exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_{1i}\mathrm{A}_{2i}, \mathrm{E}_i, \mathrm{S}_i)} \boldsymbol{\beta}]}{2 \cdot \sum_{h} \sum_{k \neq h} \exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_{hi}\mathrm{A}_{ki}, \mathrm{E}_i, \mathrm{S}_i)} \boldsymbol{\beta}]}$$

$$= \frac{\exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_{1i}\mathrm{A}_{2i}, \mathrm{E}_i, \mathrm{S}_i)} \boldsymbol{\beta}]}{\sum_{h=1}^{3} \sum_{k=h+1}^{4} \exp[\mathbf{x}^{\mathrm{t}}_{(\mathrm{A}_{hi}\mathrm{A}_{ki}, \mathrm{E}_i, \mathrm{S}_i)} \boldsymbol{\beta}]}.$$

And the conditional likelihood function for the case-spouse data (1:5 case-counterfactual-controls analysis) is

$$L_{1:5} = \prod_{i=1}^{n} \mathrm{R}_i.$$