European Journal of Human Genetics (2006) 14, 941–952

www.nature.com/ejhg

ARTICLE

Variance components model with disequilibria

Ao Yuan*[,1], Guanjie Chen[1], Qi Yang[2], Charles Rotimi[1] and George Bonney[1]

[1]*National Human Genome Center, Howard University, Washington DC, USA;* [2]*Department of Computer Science and Software Engineering, University of Wisconsin-Platteville, WI, USA*

The variance components (VC) model has been popular for genetic analysis. It has received wide applications in a variety of genetic practices, and been extended to various forms for different settings. However, most of the existing VC models are, explicitly or implicitly, under the assumption of the Hardy–Weinberg and/or linkage equilibria, which is impractical in some realistic settings since more or less deviations from this assumption are common. We propose a new VC model that incorporates both these disequilibria, and includes the existing models as special cases. The corresponding variance components are computed for some commonly used relative pairs conditional on the observed marker identity-by-descent data. Parameters can be estimated by the traditional methods such as the maximum likelihood estimate. Simulation studies suggest that this extended model improves inference significantly over the existing models when deviations of these disequilibria are present.
European Journal of Human Genetics (2006) **14,** 941–952. doi:10.1038/sj.ejhg.5201645; published online 24 May 2006

Keywords: Hardy–Weinberg disequilibrium; kinship coefficient; linkage disequilibrium; IBD; variance components

Introduction

The variance components(VC) models[1–8] has received much attention and wide applications in quantitative genetic trait studies, as this method requires few model assumptions. It has been extended to various forms for different data structures under different algorithms and model assumptions. Lange and Boehnke[9] extended it to multivariate traits, Duggirala *et al*[10] applied it to dichotomous traits, Amos *et al*[11] studied the least squares algorithm of it. Andrade *et al*[12] extended it to longitudinal pedigree data. This model and its variants have been used extensively in genetic linkage analysis. However, most of the existing VC models are, explicitly or implicitly, under the assumption of the Hardy–Weinberg and/ or linkage equilibria. These fundamental assumptions are sometimes not easy to justify, and in practice they are often more or less deviated. In linkage analysis the latter assumption may be inappropriate, since putative disease locus are usually in linkage disequilibrium(LD) with the flanking marker loci.[13] Almasy *et al*[14] proposed a combined linkage/disequilibrium analysis in which the LD are incorporated into the VC model. There are some VC models for combined linkage and association studies,[15] a VC model incorporated with the two disequilibria is of practical meaning, and has not been in the literature. Here we consider such model in the settings of Hardy–Weinberg and/or LD, as an extension of the existing VC models. In our model the LD is parameterized via the trait-marker composite genotype, differently from that in Almasy *et al*[14] in which the LD is parameterized via the trait-marker alleles. The correspondindg variance components are computed for some commonly used relative pairs conditional on the observed marker identity-by-descent (IBD) data. Parameters can be estimated by the traditional methods such as the maximum likelihood estimate (MLE) under the normal model assumption. This extended VC model is expected to have more accurate estimation of parameters, can be used for linkage and combined linkage and LD mapping (association study), using pedigree data, and have more power for such analysis.

*Correspondence: Dr Ao Yuan, National Human Genome Center, Howard University, 2216 Sixth Street, N. W., Washington, DC, 20059, USA.
Tel: +1 202 806 4361; Fax: +1 202 265 0871;
E-mail: ayuan@howard.edu
Received 22 September 2005; revised 5 April 2006; accepted 6 April 2006; published online 24 May 2006

The common VC model

We first describe the likelihood of the commonly used variance components model, for example as in Amos.[5] Since the total likelihood is a product of likelihood over all the families under study, we only present the model for a given family for the sake of simplicity.

Let Y_i be the trait value of the ith individual in the family.The VC model describing the trait value is

$$Y_i = \mu + g_i + G_i + \sum_{j=1}^{J} \eta_j x_{ij} + e_i, \tag{1}$$

where μ is the overall mean, g_i is the unobserved random major gene effect at the trait locus with alleles denoted by A and B, G_i is the unobserved polygenic effects,

$$g_i = \begin{cases} a, & \text{if individual } i \text{ has genetype AA} \\ d, & \text{if individual } i \text{ has genetype AB} \\ -a, & \text{if individual } i \text{ has genetype BB} \end{cases},$$

where the η_j's are effects associated with the covariates x_{ij}'s, and e_i is the residual random error. The usual assumption is that g_i, G_i and e_i are uncorrelated and $E(g_i) = E(G_i) = E(e_i) = 0$. Let p be the population proportion of allele A. Under the Hardy–Weinberg assumption one has $E(g_i) = a(2p-1) + 2p(1-p)d = 0$. The covariance between individuals i and j is

$$\mathrm{Cov}(Y_i, Y_j) = \begin{cases} \sigma_a^2 + \sigma_d^2 + \sigma_G^2 + \sigma_e^2, & \text{if } i = j \\ 2\Phi_{ij}\sigma_a^2 + \Delta_{7ij}\sigma_d^2 + 2\Phi_{ij}\sigma_G^2, & \text{if } i \neq j \end{cases}, \tag{2}$$

where $\sigma_a^2 = 2p(1-p)[a-d(2p-1)]^2$ is the additive genetic variance due to the locus, $\sigma_d^2 = 4p^2(1-p)^2d^2$ is the dominant genetic variance, $\Phi_{ij} = \Delta_{7ij}/2 + \Delta_{8ij}/4$ is the kinship coefficient[16] between individuals i and j, and Δ_{7ij}, Δ_{8ij}, Δ_{9ij} are the condensed kinship coefficient,[17] between individuals i and j. The Δ_{kij}s$(k = 1, \ldots, 9)$ are the probabilities for the nine possible condensed IBD status as divided by Jacquard,[17] in which Δ_{7ij}, Δ_{8ij} and Δ_{9ij} are commonly used in practice. They are the population probabilities of sharing 2, 1 and 0 genes IBD for individuals (i, j), without regard to their particular genotypes, but only (i, j)'s kinship relationships, and under the Mendelian inheritance. Also, $2\Phi_{ij}$ is the expected proportion of gene IBD for individuals (i, j), at this locus.

For linkage analysis, usually IBD sharing data $\{\pi_{ij}\}$ $\{\pi_{ij} = 0, 1, 2\}$ between a relative pair individual i and j, at marker locus is available, Amos[5] proposed the following model for the conditional covariance

$$\mathrm{Cov}(Y_i, Y_j|\pi_{ij}) = \begin{cases} \sigma_a^2 + \sigma_d^2 + \sigma_G^2 + \sigma_e^2, & \text{if } i = j \\ f(\theta, \pi_{ij})\sigma_a^2 + g(\theta, \pi_{ij})\sigma_a^2 + 2\Phi_{ij}\sigma_G^2, & \text{if } i \neq j \end{cases}, \tag{3}$$

where θ is the recombination fraction between the trait and the marker loci. The values of $f(\theta, \pi_{ij})$ and $g(\theta, \pi_{ij})$ can be found.[5] It is noted that $g(\theta, \pi_{ij}) = 0$ for most human relative pairs except full sibs and it's related to the possibility of sharing two allales IBD.

VC model with disequilibria

In this section we derive VC models with disequilibria in different settings, by incorporating these parameters into the covariances (2).

Hardy–Weinberg disequilibrium at trait locus

We first consider incorporating the Hardy–Weinberg disequilibrium at the trait locus into the VC model, without marker information Let A_k denote allele k at the trait locus $(k = 1, \ldots, K)$, p_k its proportion in the population, P_{kl} the corresponding proportion of the genotype A_kA_l. One way to deal with the deviation from the Hardy–Weinberg assumption is the use of the within population inbreeding cofficient[18,19] f at the trait locus, which is the odds that at any gene, both alleles of the gene pair were inherited from the same ancestor. Let $I(\cdot)$ be the indicator function. Given f we have

$$p_{kl} := p(A_kA_l) = (1-f)p_kp_l + fp_kI(l = k). \tag{4}$$

Here $0 \leq f \leq 1$, and $f = 0$ corresponds to Hardy–Weinberg equilibrium. Let $p_{(kl)(km)}$ be the conditional probability that two individuals have genotype (A_kA_l, A_kA_m) or (A_lA_k, A_mA_k) at the trait locus given that they share A_k IBD (Assuming random mating and phase known, these are the only cases they share A_k IBD. The possibilities for the cases A_kA_l, A_mA_k or (A_lA_k, A_kA_m) are negligible). Let Y be the trait value of a general individual and g be his/her genotype, and $\mu_{kl} = E(Y|g = A_kA_l)$. Following Fisher[1] and Lange,[16] let α_k's be the optimal additive genetic effects in the sense that they minimize the sum of squared residuals $\Sigma_k\Sigma_l\delta_{kl}^2p_{kl}$, where $\delta_{kl} = \mu_{kl} - \alpha_k - \alpha_l$. We show in Appendix A that

$$\begin{aligned} p_{(kl)(km)} &= \frac{p_{kl}p_{km}}{p_k} \\ &= p_k[(1-f)p_l + fI(l = k)] \\ &\quad [(1-f)p_m + fI(m = k)], \end{aligned} \tag{5}$$

and

$$\mathrm{Cov}(Y_i, Y_j|f) = \begin{cases} (1 + (f/2))\sigma_a^2 + (1-f)\sigma_d^2 + f\sigma_0^2 + \sigma_G^2 + \sigma_e^2, & \text{if } i = j \\ \Delta_{7ij\gamma 7}(f) + \Delta_{8ij\gamma 8}(f) + 2\Phi_{ij}\sigma_G^2, & \text{if } i \neq j \end{cases}, \tag{6}$$

where $\gamma_7(f) = (1 + (f/2))\sigma_a^2 + (1-f)\sigma_d^2 + f\sigma_0^2$, $\gamma_8(f) = ((1+f)^2/2)\sigma_a^2$, $\sigma_a^2 = 2\Sigma_k\alpha_k^2p_k$, $\sigma_d^2 = \Sigma_k\Sigma_l\delta_{kl}^2p_kp_l$, $\sigma_0^2 = \Sigma_k\delta_{kk}^2p_k$ is the part of variance explained by the optimal additive genetic effects, and $\alpha_k = \Sigma_l\mu_{kl}p_{kl}/[(1+f)p_k]$ for all k.

Note that if $f = 0$, (6) reduces to (2). The α_k's and is δ_{kl}'s are the optimal additive major gene effects and the residual effects.[16]

Linkage to marker

Now we consider the case with marker information available in addition to the trait locus data. Let $\pi_{ij} (= 0, 1, 2)$ to be the number of IBD allele sharing between individuals i and j at the marker locus, π'_{ij} be the corresponding

unobserved number at the trait locus, and θ be the recombination fraction between the two loci. Expressions for $\mathrm{Cov}(Y_i, Y_j|\pi_{ij}=k)$ can be found by the formula

$$\mathrm{Cov}(Y_i, Y_j/\pi_{ij}=k)=\sum_{l=0}^{2}\mathrm{Cov}\left(Y_i, Y_j|\pi'_{ij}=l\right)P\left(\pi'_{ij}=l|\pi_{ij}=k\right).$$

Usually, for each individual the IBD data π_{ij} is not directly available. However, their probabilities $P(\pi_{ij}=k)(k=0,1,2)$ can be computed from the corresponding observed marker genotypes. So the covariances between individual pair (i,j) in a given family is

$$\mathrm{Cov}(Y_i, Y_j)=\sum_{k=0}^{2}\mathrm{Cov}(Y_i, Y_j|\pi_{ij}=k)P(\pi_{ij}=k). \qquad (7)$$

Covariance with Hardy–Weinberg disequilibrium at trait given marker IBD

In the previous section, we derived the variance components under Hardy–Weinberg equilibrium at the trait locus. Here we give these componenets with the linked marker information, that is, conditional on the trait-marker IBD data. In this case the variance components are

$$\begin{aligned}&\mathrm{Cov}(Y_i, Y_j|f,\pi_{ij})\\&=\Delta_{7ij}(\pi_{ij})\mathrm{Cov}\left(Y_i, Y_j|f,\pi'_{ij}=2\right)\\&+\Delta_{8ij}(\pi_{ij})\mathrm{Cov}\left(Y_i, Y_j|f,\pi'_{ij}=1\right) \qquad (8)\\&=\begin{cases}\left(1+\frac{f}{2}\right)\sigma_a^2+(1-f)\sigma_d^2+f\sigma_0^2+\sigma_G^2+\sigma_e^2, & \text{if } i=j\\ \Delta_{7ij}(\pi_{ij})\gamma_7(f)+\Delta_{8ij}(\pi_{ij})\gamma_8(f)+2\Phi_{ij}\sigma_G^2, & \text{if } i\neq j\end{cases},\end{aligned}$$

where $\Delta_{7ij}(\pi_{ij})=P(\pi'_{ij}=2|\pi_{ij})$, $\Delta_{8ij}(\pi_{ij})=P(\pi'_{ij}=1|\pi_{ij})$ and $\Delta_{9ij}(\pi_{ij})=P(\pi'_{ij}=0|\pi_{ij})$ are the conditional IBD sharing at the trait locus given the IBD sharing at the marker locus, for individuals (i,j). The derivation is the same as that for $\mathrm{Cov}(Y_i, Y_j|f)$ with Δ_{7ij} and Δ_{8ij} replaced by $\Delta_{7ij}(\pi_{ij})$ and $\Delta_{8ij}(\pi_{ij})$, whose values are obtained from the relationships

$$\begin{aligned}\Delta_{7ij}(\pi_{ij})&=P(\pi'_{ij}=2,\pi_{ij})/p(\pi_{ij}),\\ \Delta_{8ij}(\pi_{ij})&=P(\pi'_{ij}=1,\pi_{ij})/p(\pi_{ij}),\\ \Delta_{9ij}(\pi_{ij})&=P(\pi'_{ij}=0,\pi_{ij})/p(\pi_{ij})\end{aligned}$$

and the known values of $P(\pi'_{ij}=0,\pi_{ij})$ as listed in the literatures cited before. Note here given $\pi'_{ij}=0$, Y_i and Y_j are independent, and $\mathrm{Cov}(Y_i, Y_j|f,\pi'_{ij}=0)=0$, thus we don't have the term for $\Delta_{9ij}(\pi_{ij})$.

Since in real data the set $\{\pi_{ij}\}$ is unobservable, we only have the computed the set of probabilities $\{P(\pi_{ij}=k)\}$, thus the covariance is

$$\mathrm{Cov}(Y_i, Y_j|f)=\sum_{k=0}^{2}\mathrm{Cov}(Y_i, Y_j|f,\pi_{ij}=k)P(\pi_{ij}=k). \qquad (9)$$

Covariance with LD between trait and marker

In linkage analysis, LD between the trait locus and the genotype marker locus should be taken into consideration. In this section we compute the covariances between relative pairs when in addition to the case of LD is also present between the trait and marker loci. Let a_k and a_ka_l denote the alleles and genotypes at the marker locus, q_k and q_{kl} be the corresponding population frequencies. Since the within-population inbreeding coefficient f is common for any locus in the genome of the given population, f describes the relationship between the marker genotype frequencies q_ks allele frequencies $q_{kl}s$, in the same way as it did between the p_ks and $p_{kl}s$ at the trait locus. That is, we have

$$q_{kl}=(1-f)q_kq_l+fq_kI(l=k). \qquad (10)$$

Let $G=\left(\frac{A_kA_l}{a_ra_s}\right)$ be a general notation for the trait-marker composite genotype. We assume

$$p_{(kl,rs)}=p_{kl}(q_{rs}-\zeta p_{rs})+\zeta p_{kl}I((r,s)=(k,l)). \qquad (11)$$

It is easy to check that under (11), $\Sigma_k\Sigma_l p_{(kl,rs)}=p_{rs}$ and $\Sigma_r\Sigma_s p_{(kl,rs)}=p_{kl}$, the probabilities of composite genotypes satisfy such *consistent* condition with its marginal probabilities. Here $0\leqslant\zeta\leqslant1$ is the LD parameter, and it should not be confused with the definition of LD that is used in some texts, such as in Weir[20] or Almasy *et al*.[14] Note that $\zeta=0$ corresponds to linkage equilibrium. Also, ζ manifests the vertical connection between the trait and marker loci, while the recombination fraction describes the horizontal link between the alleles.

For a relative pair, let $p_{(kl)(mn)|}\pi'_{ij}=P(A_kA_l, A_mA_n|\pi'_{ij})$ be the conditional probability that individual i has trait genotype A_kA_l and individual j has trait genotype A_mA_n given their IBD value π'_{ij} at this locus, $p_{klm|}\pi'_{ij}=$ ½$P(A_kA_l, A_kA_m|\pi'_{ij})+$ ½$P(A_kA_l, A_mA_k|\pi'_{ij})$ be the probability when they also share one allele identical by state (IBS) at the trait; $p_{kl|}\pi'_{ij}=P(A_kA_l, A_kA_l|\pi'_{ij})$ be the probability when they share both alleles IBS at the trait locus. We have (Appendix B)

$$\begin{aligned}&\mathrm{Cov}(Y_i, Y_j|f,\zeta,\pi_{ij})\\&=\begin{cases}(1+f/2)\sigma_a^2+(1-f)\sigma_d^2+f\sigma_0^2+\sigma_G^2+\sigma_e^2, & \text{if } i=j\\ \sum_{k=7}^{9}\Delta_{kij}(\pi_{ij})\gamma_k(f,\zeta,\pi_{ij})+2\Phi_{ij}\sigma_G^2, & \text{if } i\neq j\end{cases},\end{aligned} \qquad (12)$$

where $\gamma_7(f,\zeta,\pi_{ij})$, $\gamma_8(f,\zeta,\pi_{ij})$ and $\gamma_9(f,\zeta,\pi_{ij})$ denote respectively $\mathrm{Cov}(g_i, g_j|f,\zeta,\pi_{ij},\pi'_{ij}=2)$, $\mathrm{Cov}(g_i, g_j|f,\zeta,\pi_{ij},\pi'_{ij}=1)$ and $\mathrm{Cov}(g_i, g_j|f,\zeta,\pi_{ij},\pi'_{ij}=0)$. Note that by conditioning on the IDB values at both the trait and marker loci, we cannot assert $\mathrm{Cov}(g_i, g_j|f,\zeta,\pi_{ij},\pi'_{ij}=0)=0$ as we did for the previous section. We have $\gamma_7(f,\zeta,\pi_{i,j})\equiv\gamma_7(f)$,

$$\begin{aligned}&\gamma_8(f,\zeta,\pi_{ij})\\&=\begin{cases}\gamma_8(f), & \pi_{ij}=0\\ \gamma_8(f), & \pi_{ij}=1\\ (1-\zeta)\gamma_8(f)-\zeta(1+f)^2\sigma_{1,1}+2\zeta\sigma_{1,2}+\zeta^2\sigma_{1,3}, & \pi_{ij}=2\end{cases},\end{aligned}$$

and

$$\gamma_9(f,\zeta,\pi_{ij}) = \begin{cases} 0, & \pi_{ij}=0 \\ \zeta^2(1+f)^2\sigma_2^2, & \pi_{ij}=1 \\ \zeta^2\sigma_3^2, & \pi_{ij}=2 \end{cases},$$

where

$$\sigma_{1,1} = \sum_k \alpha_k^2 p_k \sum_s \frac{(q_{ks}-\zeta_{pks})[(1-f)p_s + fI(s=k)]}{q_{ks}-\zeta_{pks}+\zeta(1-f)p_s+\zeta fI(s=k)},$$

$$\sigma_{1,2} = \sum_k \sum_l \sum_m (\alpha_k+\alpha_l+\delta_{kl})(\alpha_k+\alpha_m+\delta_{km}) \times \frac{p_{(kl)(km)}(q_{km}-\zeta_{pkm})}{q_{km}-\zeta_{pkm}+\zeta(1-f)p_m+f\zeta I(k=m)},$$

$$\sigma_{1,3} = \sum_k \sum_l (\alpha_k+\alpha_l+\delta_{kl})^2 \times \frac{p_{kl}^2}{pk[q_{kl}-\zeta_{pkl}+\zeta(1-f)p_l+f\zeta]},$$

$$\sigma_2^2 = \sum_k \alpha_k^2 p_k^2/q_k,\ \sigma_3^2 = \sum_k \sum_l (\alpha_k+\alpha_l+\delta_{kl})^2 p_{kl}^2$$

which is also written as

$$\begin{aligned} &2(1-f)^2\sum_k\sum_l \alpha_k^2 p_k^2 p_l^2 + 8f(1-f)\sum_k \alpha_k^2 p_k^3 \\ &+ 4f^2\sum_k \alpha_k^2 p_k^2 + (1-f)^2\sum_k\sum_l \alpha_{kl}^2 p_k^2 p_l^2 \\ &2f(1-f)\sum_k \delta_{kk}^2 p_k^3 + f^2\sum_k \delta_{kk}^2 p_k^2 + 4(1-f)^2\sum_k\sum_l \alpha_k\delta_{kl} p_k^2 p_l^2 \\ &+ 8f(1-f)\sum_k \alpha_k\delta_{kk} p_k^3 + 4f^2\sum_k \alpha_k\delta_{kk} p_k^2 \\ &+ 2(1-f)^2(\sum_k \alpha_k p_k^2)^2 := 2(1-f)^2\sigma_{3,1} + 8f(1-f)\sigma_{3,2} \\ &+ 4f^2\sigma_{3,3} + (1-f)^2\sigma_{3,4} + 2f(1-f)\sigma_{3,5} + f^2\sigma_{3,6} \\ &+ 4(1-f)^2\sigma_{3,7} + 8f(1-f)\sigma_{3,8} + 4f^2\sigma_{3,9} + 2(1-f)^2\sigma_{3,10}. \end{aligned}$$

Since the genetic covariance between the relative pair can be written as

$$\mathrm{Cov}(g_i,g_j|f,\zeta,\pi_{ij}) = \sum_{\pi'_{ij}} P(\pi'_{ij}|\pi_{ij})\,\mathrm{Cov}(g_i,g_j|f,\zeta,\pi_{ij},\pi'_{ij}),$$

by (12), when $\pi'_{ij}=2$ or $\pi_{ij}=0$, the expression for genetic variance between a relative pair is the same regardless LD is present or not. In fact, from the derivation in Appendix B, this conclusion is true for any consistent composite genotype specification: under random mating and any consistent specification $P(G)$ of the composite genotype, the IBD status (π'_{ij},π_{ij}) of a relative pair (i,j) contributes information of LD to their genetic variance at the trait locus only if $\pi'_{ij}\le 1$ and $\pi_{ij}\ge 1$.

Again in practice, given the estimated IBD probabilities, the covariance is computed as

$$\mathrm{Cov}(Y_i,Y_j|f,\zeta) = \sum_{k=0}^{2} \mathrm{Cov}(Y_i,Y_j|f,\zeta,\pi_{ij}=k)P(\pi_{ij}=k). \quad (13)$$

Parameter estimation

Let $\beta=(\mu,\eta_1,\ldots,\eta_j)^{\mathrm{T}}$ be the parameters in the mean, $\alpha=(\theta,f,\zeta,\sigma_a^2,\sigma_d^2,\sigma_G^2,\sigma_e^2,\sigma_0^2,\sigma_1^2,\sigma_2^2,\sigma_3^2,\sigma_{11},\sigma_{12},\sigma_{13})^{\mathrm{T}}$ be the parameters in the covariance matrices, $\mathbf{y}_k$ be the observations of all the members in the kth family, and $\mu_k=\mu_k(\beta)=E(\mathbf{Y}_k)=X_k\beta$, where X_k is the covariate matrix for the kth family, and n_k is the total number of individuals in this family. The commonly used model based estimation method is MLE, while the common model for quantitative trait is the normal distribution. Under these assumptions, the likelihood of the kth family is $L_k(\alpha,\beta|\mathbf{Y}_k)=\phi(\mathbf{Y}_k-\mu_k|\Omega_k)$, where $\phi(\mathbf{Y}-\mu|\Omega)$ is the density of the n_k dimensional normal $N(\mu,\Omega)$ distribution, $\Omega_k=(\omega_{ij})_{n_k\times n_k}$ is the covariance matrix of the kth family, with

$$\begin{aligned} \omega_{ij} &= \mathrm{Cov}(Y_i,Y_j|f,\zeta,g_{ij}) \\ &= \sum_{r=0}^{2}\mathrm{Cov}(Y_i,Y_j|f,\zeta,\pi_{ij}=r)P(\pi_{ij}=r|g_{ij}) \end{aligned}$$

as specified in (12) in the most general case. The $P(\pi_{ij}=r|g_{ij})$'s can be obtained by some common IBD computation methods. The covariances can also take any of the more specific form (8), (6), (3) and (2) in the equilibrium case. Here we used (Y_i,Y_j) for (Y_{ki},Y_{kj}), the (i,j)th relative pair in the kth family. The total likelihood is thus $L(\alpha,\beta|\mathbf{Y})=\Pi_{k=1}^{k}L_k(\alpha,\beta|\mathbf{Y}_k)$, and the log-likelihood, omitting the normalizing constant, is

$$\log L(\alpha,\beta|\mathbf{Y}) = -\frac{1}{2} \times \sum_{k=1}^{K}\log|\Omega_k| + (\mathbf{Y}_k-\mu_k)'\Omega_k^{-1}(\mathbf{Y}_k-\mu_k). \quad (14)$$

The MLE is the parametric value $(\hat{\alpha},\hat{\beta})$ that maximizes (14), and it has many desired optimality properties.

Power

The power of the method can be easily estimated and will shown is dependent only on the parameters α in the covariance matrix. Let H_0:$\alpha=\alpha_0$ and H_1:$\alpha=\alpha_1$ ($H_0\subset H_1$ or α_0 be part of α_1) be the null and alternative hypothesis considered in the previous sections, $\dim(H_1)-\dim(H_0)=k$ and $f(\cdot|\alpha,\beta)$ be the density of the model considered. Let $\hat{\alpha}_0$ and $\hat{\alpha}_1$ be the MLE of α under H_0 and H_1, respectively. Note our hypothesis only involves α, not the parameters β in the mean specification. Let

$$T_n = -2\log\frac{L(\hat{\alpha}_0,\hat{\beta}|\mathbf{Y})}{L(\hat{\alpha}_1,\hat{\beta}|\mathbf{Y})} \quad \text{and}$$

$$D(\alpha_1||\alpha_0) = \int f(x|\alpha_1,\beta)\log\frac{f(x|\alpha_1,\beta)}{f(x|\alpha_0,\beta)}\mathrm{d}x$$

be the relative entropy (Kullback–Leibler divergence) between the two densities $f(\cdot|\alpha_1,\beta_1)$ and $f(\cdot|\alpha_0,\beta_0)$. It is known that $D(\alpha_1||\alpha_0)\ge 0$ with equality hold if and only if $\alpha_1=\alpha_0$. Assuming homogenous familial structures for all

the families, for give level $\gamma>0$, the asymptotic power q_n for the likelihood ratio test of H_0 vs H_1, with a dataset of size (number of families) n, is (Appendix C)

$$q_n = P\big(T_n > \chi_k^2(1-\gamma)\big) \approx P\big(V_k > \chi_k^2(1-\gamma) - 2nD(\alpha_1||\alpha_0)\big), \quad (15)$$

where V_k is the χ^2 random variable with k degrees of freedom and $\chi_k^2(1-\gamma)$ is its $1-\gamma$ upper quantile.

Given $f(\cdot|\cdot,\cdot)$, α_1 and α_0, $D(\alpha_1||\alpha_0)$ can be easily computed. In fact, since our model $f(\cdot|\cdot,\cdot)$ is multivariate normal, it is easy to see that

$$D(\alpha_1||\alpha_0) = \frac{1}{2}\log\frac{|\Omega(\alpha_0)|}{|\Omega(\alpha_1)|} + \frac{1}{2}[tr((\Omega^{1/2}(\alpha_1))'\Omega^{-1}(\alpha_0)\Omega^{1/2}(\alpha_1)) - d],$$

where $d=\dim(\mu)$, $\Omega(\cdot)$ is the Ω_k's with the elements given in (12), in which the τ_{ij}'s take the theoretical mean values. To plot the power surface, we fix the parameter values at their MLE, except those for f and ζ. Then for a given $\gamma>0$ and a set of selected (f,ζ) values, we can compute $q_n=q_n(\gamma,f,\zeta)$ for different γ, f, ζ and n.

Application

Simulation study

Data of 10 000 sibpairs are simulated in our study. We give some detailed description of how the two levels of disequilibria are incorporated in the simulation process. It can be described in the following three steps.

Step 1 For each sibpair we simulate the their trait genotypes g_is and the marker IBD probabilities π_{ij}s. Let $G_i=(a_r a_s)/(A_k A_l)$ be the composite genotype of the trait and marker for the ith individual, with lower case letters $a_r a_s$ for marker genotype. we simulate (G_i, G_j) for each sibpair, and π_{ij} is generated along. We first generate the composite genotypes G_f of the father and G_m of the mother by the probability given in (11) with $\zeta=0.1$, and p_{kl} and q_{rs} are given (4) and (10) with $f=0.12$, $p_1=0.55$, $p_2=0.45$, $q_1=0.65$ and $q_2=0.35$. Although (G_f, G_m) are not part of the data to be used in the computation, they are needed to generate the sibs composite genotypes. Now given (G_f, G_m) we generate G_i, G_j and π_{ij} as below. Let $G_f=(a_{f1}\ a_{f2})/(A_{f1}\ A_{f2})$, $G_m=(a_{m1}\ a_{m2})/(A_{m1}\ A_{m2})$. During meiosis, if there is no recombination (with probability $1-\theta$, $\theta=0.25$), G_f splits into two gametes (a_{f1}/A_{f1}) and (a_{f2}/A_{f2}). Then one of the gametes is selected with probability 0.5 to pass to the next generation. Here we only consider the recombination at the marker, since we want the IBD π_{ij} at the marker. The recombination at the trait is similar, and we omit it for simplicity, since this will not affect the probabilities of the G_is. Similarly, G_m will split into (a_{m1}/A_{m1}) and (a_{m2}/A_{m1}), or (a_{m2}/A_{m1}) and (a_{m1}/A_{m2}), and one of the gamets is selected with probability 0.5 to pass to the next generation. For example, if for the father, there is recombination during meiosis and (a_{f1}/A_{f1}) is selected, and for the mother there is no recombination during meiosis and (a_{m1}/A_{m1}) is selected, then $G_i=(a_{f1}\ a_{m1})/(A_{f2}\ A_{m1})$ and $g_i=(A_{f2}A_{m1})$. Repeat the above process to get, say, $G_j=(a_{f2}\ a_{m1})/(A_{f1}\ A_{m1})$ and $g_j=(A_{f1}A_{m1})$. Since at the marker locus, sibpair (i, j) has a composite genotype $(a_{f1}a_{m1}, a_{f2}a_{m2})$, we have $\pi_{ij}=1$, which comes from the common maternal allele a_{m1}.

Step 2 Simulate each pair's covariates. The mean μ_I of the ith individual is given by (1). Specifically, we take $\mu=23$, $g_i=1$ if individual i has genotype $A_1A_1=0$ if A_1A_2, and $=-1$ if A_1A_2. We take $G_i\sim N(0,\sigma_G^2)$ with $\sigma_G^2=0.2$. Two covariates are genetated, x_{i1} and x_{i2}, stand for age (years) and sex index for the ith individual, $x_{i2}=1$ for female and $=0$ for male. The coefficient for age is $\eta_1=0.2$ and that for sex is $\eta_2=1.5$. e_i is the random error from $N(0, 1)$ distribution.we always assume the first dib is younger with $x_{i1}\sim U[10, 60]$, then for the second sib, with $x_{j1}=x_{i1}+z$ with $z\sim U[1, 10]$. For x_{i2}, using the gender ratio from the real data, we sample $z\sim U(0, 1)$, if $z\leq 0.54$ let $x_{i2}=1$ (female) otherwise 0 (male).

Step 3 Simulates the sibpair covariance matrices $\Omega_{ij}=\mathrm{Cov}(Y_i, Y_j)=(\omega_{ij})$ and the final observed trait values. By (3.9), $\omega_{11}=\omega_{22}=(1+f/2)\sigma_a^2+(1-f)\sigma_d^2+f\sigma_0^2+\sigma_G^2+\sigma_e^2$, $\sigma_a^2=2\Sigma_k\alpha_k^2p_k$, $\alpha_d^2=\Sigma_{k,l}\delta_{kl}^2p_kp_l$, $\sigma_0^2=\Sigma_k\delta_{kk}^2p_k$ and p_k is the population proportion of allele A_k, $\delta_{kl}=\mu_{kl}-\alpha_k-\alpha_l$, $\alpha_k=\Sigma_l\mu_{kl}p_{kl}/[(1+f)p_k]$, $p_{kl}=(1-f)p_kp_l+f_{pk}I(l=k)$ is the population proportion of genotype A_kA_l, and $\mu_{kl}=E(Y|g=A_kA_l)=\mu+g_{kl}+\eta_1E(x_{i1})+\eta_2E(x_{i2})=\mu+g_{kl}+40\eta_1+0.54\eta_2$, $\omega_{12}=\omega_{21}=\sum_{k=7}^{9}\Delta_{kij}(\pi_{ij})\gamma_k(f,\zeta,\pi_{ij})+2\Phi_{ij}\sigma_G^2$ as is given in (3.9), and for sibpairs $\Phi_{ij}=1/4$. $\Delta_{kij}(\pi_{ij})$ is defined after (8) and can be found in Wright,[18] where they are implemented in terms of the recombination fraction θ. The marker IBD data π_{ij}s are generated above, the trait IBD π'_{ij} are unknown, but only the conditional probability $P(\pi'_{ij}|\pi_{ij})$s are used, which are easily derived.[20] The $\gamma_k(f,\zeta,\pi_{ij})$s are defined after (12). The definition of $\sigma_{1,2}$ involved $p_{(kl)(km)}$ which is given in the definition of the $\gamma_k(f,,\zeta,\pi_{ij})$s. Now we have implemented Ω_{ij} and are ready to simulated the y_is. We simulate the data pairwise. For a sibpair (y_i, y_j), denote $Y=(y_i, y_j)$ and $\mu=(\mu_i, \mu j)$. We sample $Z\sim N(\mathbf{0}, I_2)$, the two-dimensional standard normal distribution, and let $Y=\Omega_{ij}^{1/2}Z+\mu$, and simulate such Y 10 000 times.

For $\gamma_8(f,\zeta,\pi_{ij})$ in te case $\pi_{ij}=2$, $\sigma_{1,1}$, $\sigma_{1,2}$ and $\sigma_{1,3}$ are not independently estimable, so in this case we write $\gamma_8(f,,\zeta,2)=(1-\zeta)\gamma_8(f)=\sigma_4^2$, where $\sigma_4^2=-\zeta(1+f)\sigma_{1,1}+2\zeta\sigma_{1,2}+\zeta^2\sigma_{1,3}$ viewed as a single parameter to be estimated.

Table 1 displays the values of the real parameters of interest from the simulation, and their MLE estimates (estimated standard deviation in bracket) under H_0: $f=\zeta=0.0$ and H_1: all parameters free, respectively.

The difference 2(log *likelihood*(H_1)−log *likelihood*(H_0))= 20.9934, with a P-value of 0.000106 under a χ^2 distribution

Table 1 Parameter estimates for the simulated data under H_0 and H_1

Parameter	*Real*	*Under H_0*	*Under H_1*
f	0.12	0.00	0.1253 (0.1473)
ζ	0.10	0.00	0.1089 (0.2229)
θ	0.25	0.2470 (0.1052)	0.2468 (0.1163)
μ	23	23.0430 (4.3663)	23.4938 (4.8503)
η_1	1.5	1.3661 (2.7195)	1.5611 (2.9335)
η_2	0.2	0.2057 (0.1053)	0.1882 (0.1164)
Log-likelihood		−89736.12	−89725.63

Table 2 Parameter estimates for the AADM data under H_0 and H_1

Parameter estimates	*Under H_0*	*Under H_1*
μ	22.874 (2.8530)	23.582 (0.000365)
η_1	0.075 (0.0495)	0.053 (0.000063)
θ	0.5	0.00157 (0.000384)
f	0	0.2349 (0.00026)
ζ	0	0.0530 (0.00285)
σ_g^2	4.146 (1.146)	3.5419 (4.947)
σ_e^2	5.6214 (0.7581)	5.6423 (4.039)
Log-likelihood	−1136.60	−1130.34

with two degrees of freedom, that is, the evidence of rejecting H_0 is very strong. This example shows that incorporating the disequilibria mechanism into the variance components model can improve the inference significantly when such disequilibria are present.

Real data application

We used the AADM data (African-American *Diabetes mellitus*) to illustrate the method. The data is from an international collaboration between West Africa and US investigators in mapping type II diabetes susceptibility genes in West African ancestral populations of African-Americans. Affected sib-pairs along with unaffected spouse controls were being enrolled. Eligible participants were invited to study clinics to obtain detailed epidemiological, familial and medical history information. For detailed description of the data, see Rotimi *et al*.[21] For this data we computed the model parameter estimates using VC model (2), or under the hypothesis of equilibria, H_0: $f=\zeta=0$; and under the VC model with Hardy–Weinberg/LD (12), H_1:f and ζ are free parameters, to fit the data. The response variable is BMI, the covariate is age. The results are shown in Table 2, where the estimated standard deviations are listed inside the brackets.

The −2 loglikelihood difference is 12.5076 with a P-value of 0.0058, which is highly significant. So the inference should be based on H_1. We see a large Hardy–Weinberg disequilibrium at the triat locus, suggesting that the genetic background of the sample under study is not as simple as assumed by the existing VC model. The low recombination rate (0.0016) indicates that the trait and marker loci are tightly linked, and the LD between the trait and marker is non-negligible. The overall BMI of this sample is 23.58, and the age effect is 0.053, which are quite common for normal populations.

The power depends on all the parameters in the model, we highlight its dependence on (f, η) to study its relationship with these two parameters. Using (15) and the parameters above, the following Figure 1 shows the powers of the likelihood ratio test for H_0 *vs* H_1, for various combinations of f, ζ, and n.

Since the LD depends on the unobservable trait genotype, its needs larger sample size to detect. For the real data, with the observations and the estimated parameter setting, it is easy to detect the HWE disequilibrium with reasonable sample size, while it is very difficult to detect the LD, or requires very large sample size to achieve high power. For the simulated data-parameter setting, the powers are high for the joint HWE disequilibrium, LD and the joint HWE and LD disequilibria.

The software for this extended VC model is written in SAS; the current version is for sibpair familial structure only, and is available upon request from the second author at gchen@genomecenter.howard.edu. The CPU time to compute the parameter estimates depends on the machine, data size, number of regressors, pedigree structure and starting values for the parameters etc. For the two examples above, with suitably chosen starting values, the CPU times for computing the MLEs are 27.24 and 27.33 s on our machine.

Discussion

We have generalized the VC model to the cases of the Hardy–Weinberg and LD or both, this gives more practical application of this popular model. In some practices, these disequilibria are not justified. In these cases, the existing VC model is clearly inadequate, and our generalized VC model might be beneficial in more estimates, and in enhancing the inference power of parameters of interest. Also this generalized model can be used in testing these disequilibria by forming the corresponding likelihood ratio statistic, along with the parameter estimates. Other inferences on one or both of the two disequilibria are sometimes also of direct interest, which are now available under this generalized VC model.

We computed the variance components for some common relative pairs. The cases of other relative pairs are similar and straightforward. We considered the parameter estimation in several ways and computed the IBD under some common cases.

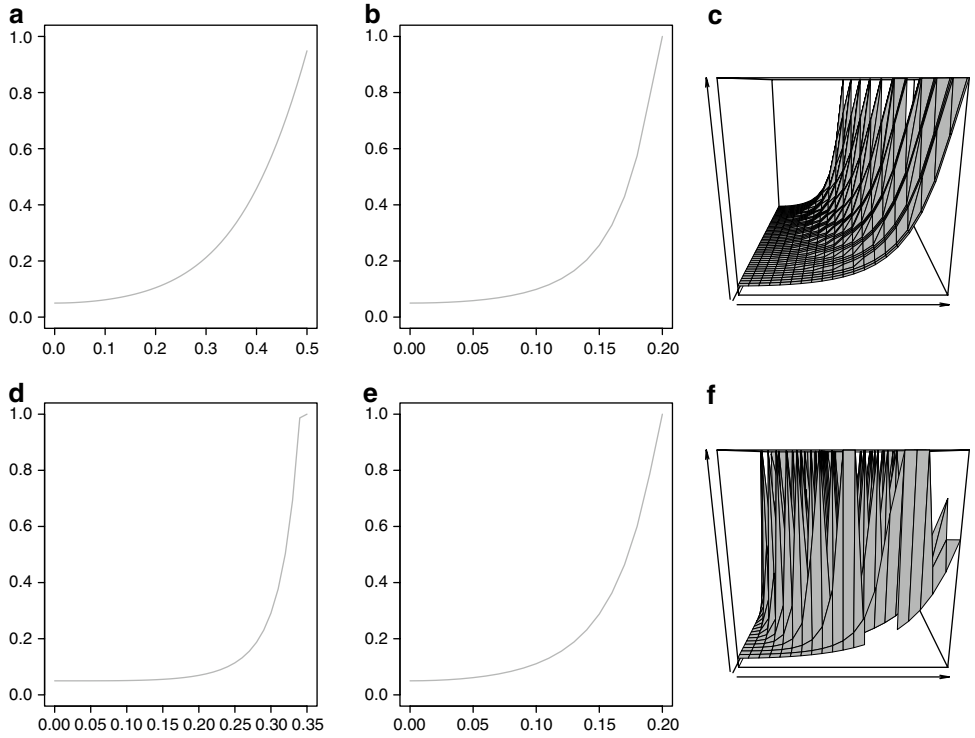

Figure 1 Powers of the likelihood ratio test for H_0 *vs* H_1. (**a**)–(**c**) Power for the real data, with parameters set at the MLE values. (**a**) H_0: $f=0$ *vs* H_1: $f\neq 0$. Horizontal axis for f, vertical axis for power, $n=8\,000\,000$; (**b**) H_0: $\zeta=0$ *vs* H_1: $\zeta\neq 0$. Horizontal axis for ζ, vertical axis for power, $n=675$; (**c**) (the third on the first row): H_0: $(f,\zeta)=(0,0)$ *vs* H_1: $(f,\zeta)\neq(0,0)$. $n=675$. (**d**)–(**f**) Power for the simulated data. The parameters used are $\alpha=(\sigma_4,\sigma_3,\sigma_2,\sigma_g,\sigma_0,\sigma_a,\sigma_d,\theta)=3.08, 16.3, 43.5, 0.45, 21.8, 7.24, 20.8, 0.2468)$, sample size is $n=250$ for the three panels. (**d**): H_0: $f=0$ *vs* H_1: $f\neq 0$; (**e**) H_0: $\zeta=0$ *vs* H_1: $\zeta\neq 0$; (**f**) (the third on the second row): H_0: $(f,\zeta)=(0,0)$ *vs* H_1: $(f,\zeta)\neq(0,0)$.

Further extensions/modifications to implement more features will be similar, such as the multivariate traits,[9] the multipoint VC, dichotomous trait, robust LOD score correction,[7] the conditioning adjustment.[21]

Acknowledgements

We appreciate the suggestions/comments from the Editor and the two anonymous reviewers, which greatly improved the quality of this manuscript. The work was supported by the United States Public Service Grant No AG 16996 and the National Center for Research Resources Grant No 2G12RR003048 from the National Institutes of Health. The AADM study was supported by NIH Grants no. 3T37TW00041 from NCMHD and NHGRI. G Chen and Rotimi were also partly supported by the NIGMS/MBRS program.

References

1 Fisher RA: The correlation between relatives on the supposition of mendel inheritance. *Trans Roy Soc Edinburgh* 1918; **52**: 399–433.
2 Harris DL: Genotypic covariances between inbred relatives. *Genetics* 1964; **50**: 1319–1348.
3 Amos CI, Elston RC: Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genet Epidemiol* 1989; **6**: 306–349.
4 Goldgar DE: Multipoint analysis of human quantitative genetic variation. *Am J Human Genet* 1990; **47**: 957–967.
5 Amos CI: Robust variance-components approach for assessing gene linkage in pedigrees. *Am. J Human Genet* 1994; **54**: 535–543.
6 Almasy L, Blangero J: Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Human Genet* 1998; **62**: 1198–1211.
7 Blangero J, Williams JT, Almasy L: Robust LOD score for variance component-based linkage analysis. *Genetic Epidemiol* 2000; **19** (Suppl 1): S8–S14.
8 Sham PC, Purcell S: Equivalence between Haseman–Elston and variance-components linkage analysis for sib pairs. *Am J Human Genet* 2001; **68**: 1527–1532.
9 Lange K, Boehnke M: Extensions to pedigree analysis. IV covariance components models for multivariate traits. *Am J Med Genet* 1983; **14**: 513–524.
10 Duggriala R, Williams JT, Williams-Blangero S, Blangero J: A variance component approach to dichotomous trait linkage analysis using a threshold model. *Genet Epidemiol* 1997; **14**: 987–992.
11 Amos CI, Gu X, Chen J, Davis BR: Least squares estimation of variance components for linkage. *Genetic Epidemiol* 2000; **19** (Suppl 1): S1–S7.
12 Andrade M, Gueguen R, Visvikis S, Sass C, Siest G, Amos C: Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. *Genetic Epidemiol* 2002; **22**: 221–232.
13 Xiong M, Jin L: Combined linkage and linkage disequilibrium mapping for genome screens. *Genetic Epidemiol* 2000; **19**: 211–234.

14 Almasy L, Williams J, Dyer T, Blangero J: Quantitative trait locus detection using combined linkage/disequilibrium analysis. *Genetic Epidemiol* 1999; **17** (Suppl. 1): S31–S36.
15 Fulker DW, Cherny SS, Sham PC, Hewitt JK: Combined linkage and association sib-pair analysis for quantitative traits. *Am J Human Genet* 1999; **64**: 259–267.
16 Lange K: *Mathematical and statistical methods for genetic analysis*. Berlin: Springer-Verlag, 1997.
17 Jacquard A: *The genetic structure of populations*. New York: Springer-Verlag, 1974.
18 Wright S: The genetic structure of populations. *Ann. Eugenics* 1951; **15**: 323–354.
19 Cockerham CC: Variance of gene frequencies. *Evolution* 1969; **23**: 72–84.
20 Weir B: *Genetic Data Analysis II*. Sinauer Associates, Inc. Publishers: Sunderland, Massachusetts, 1996.
21 Rotimi CN, Dunston GM, Berg K: In search of susceptibility genes for type 2 diabetes in West Africa: The design and results of the first phase of the AADM study. *Ann Epidemiol* 2001; **11**: 51–58.
22 Lange K: Central limit theorems for pedigrees. *J Math Biol* 1978; **6**: 59–66.

Appendix A

We first derive (5). When fixed A_k, the events A_kA_l and A_kA_m are independent, so by (4) we have,

$$\begin{aligned} p_{(kl)(km)} &= P(A_kA_l, A_kA_m) = P(A_k)P(A_kA_l, A_kA_m|A_k) \\ &= P(A_k)P(A_kA_m|A_k) = \frac{P(A_kA_l)P(A_kA_m)}{P(A_k)} \\ &= \frac{p_{kl}p_{km}}{p_k} = p_k[(1-f)p_l + fI(l=k)] \\ &\quad [(1-f)p_m + fI(m=k)]. \end{aligned}$$

For (6), we use the method as in Lange[16] (pp. 87–89). Since $0 = E(g) = \sum_k \sum_l \mu_{kl}$ and the α_k's minimize the squared error $\sum_k \sum_l (\mu_{kl} - \alpha_k - \alpha_l)^2 p_{kl}$, take derivative with respect to α_k, we get

$$\sum_l \delta_{kl} p_{kl} = 0, \quad \text{all } k. \tag{A.1}$$

Sum over k in (A.1) we have

$$\begin{aligned} 0 &= \sum_k \sum_l (\mu_{kl} - \alpha_k - \alpha_l)p_{kl} = -2\sum_k \sum_l \alpha_k p_{kl} \\ &= -2\sum_k \alpha_k. \end{aligned} \tag{A.2}$$

Now (A.1) and (A.2) gives

$$\begin{aligned} 0 &= \sum_l (\mu_{kl} - \alpha_k - \alpha_l)p_{kl} = \sum_l \mu_{kl}p_{kl} - \alpha_k p_k - \sum_l \alpha_l p_{kl} \\ &= \sum_l \mu_{kl}p_{kl} - \alpha_k p_k - \sum_l \alpha_l[(1-f)p_k p_l + fp_k I(l=k)] \\ &= \sum_l \mu_{kl}p_{kl} - \alpha_k p_k - f\alpha_k p_k, \end{aligned}$$

that is,

$$\alpha_k = \frac{1}{(1+f)p_k}\sum_l \mu_{kl}p_{kl}.$$

Then we have

$$\mathrm{Cov}(Y_i, Y_j|f) = \mathrm{Cov}(g_i, g_j|f) + \mathrm{Cov}(G_i, G_j) + \mathrm{Cov}(e_i, e_j).$$

When $i=j$, $\mathrm{Cov}(G_i, G_j) = \sigma_G^2$, $\mathrm{Cov}(e_i, e_j) = \sigma_e^2$ and

$$\mathrm{Cov}(g_i, g_i|f) = E(g_i^2) = \sum_k \sum_l (\alpha_k + \alpha_l + \delta_{kl})^2 p_{kl}.$$

By (A.1), the above is

$$\begin{aligned} &\sum_k \sum_l \alpha_k^2 p_{kl} + \sum_k \sum_l \alpha_l^2 p_{kl} + \sum_k \sum_l \alpha_k \alpha_l p_{kl} \\ &\quad + \sum_k \sum_l \delta_k^2 p_{kl} \end{aligned} \tag{A.3}$$

Since $\sum_l p_{kl} = \sum_l p_{lk} = p_k$, we have

$$\sum_k \sum_l \alpha_k^2 p_{kl} + \sum_k \sum_l \alpha_l^2 p_{kl} + 2\sum_k \alpha_k^2 \sum_l p_{kl} = 2\sum_k \alpha_k^2 p_k = \sigma_a^2;$$

$$\begin{aligned} \sum_k \sum_l \alpha_k \alpha_k p_{kl} &= \sum_k \sum_l \alpha_k \alpha_l [(1-f)p_k p_l + fp_k I(l=k)] \\ &= f\sum_k \alpha_k^2 p_k = \frac{f}{2}\sigma_a^2; \end{aligned}$$

and

$$\begin{aligned} \sum_k \sum_l \delta_{kl}^2 p_{kl} &= \sum_k \sum_l \alpha_{kl}[(1-f)p_k p_l + fp_k I(l=k)] \\ &= (1-f)\sum_k \sum_l \delta_{kl}^2 p_k p_l + f\sum_k \delta_{kk}^2 p_k \\ &= (1-f)\sigma_d^2 + f\sigma_0^2; \end{aligned}$$

so by (A.3) and the above three equations we have

$$\mathrm{Cov}(g_i, g_i|f)(1 + (f/2))\sigma_a^2 + (1-f)\sigma_d^2 + f\sigma_0^2.$$

If $i \neq j$, $\mathrm{Cov}(e_i e_j) = 0$, by the central limit theorem of Lange[22] and assume no dominance, we have approximately $\mathrm{Cov}(G_i, G_j) = 2\Phi_{ij}\sigma_G^2$ and

$$\begin{aligned} \mathrm{Cov}(g_i, g_j|f) &= E(g_i g_j|f) = \Delta_{7ij} \sum_k \sum_l (\alpha_k + \alpha_l + \delta_{kl})^2 p_{kl} \\ &+ \Delta_{8ij} \sum_k \sum_l \sum_m (\alpha_k + \alpha_l + \delta_{kl}) \times (\alpha_k + \alpha_m + \delta_{km}) p_{(kl)(km)} \\ &+ \Delta_{9ij} \sum_k \sum_l \sum_m \sum_n (\alpha_k + \alpha_l + \delta_{kl})(\alpha_m + \alpha_n + \delta_{mn}) p_{kl} p_{mn}. \end{aligned}$$

By (A.1) and (A.2), the coefficient Δ_{9ij} is zero. By the calculation for $E(g_i^2)$, the first term above is

$$\Delta_{7ij}\left[\left(1 + \frac{f}{2}\right)\sigma_a^2 + (1-f)\alpha_d^2 + f\sigma_0^2\right], \tag{A.4}$$

the second term is

$$\begin{aligned} \Delta_{8ij}\Bigg[&\sum_k \sum_l \sum_m \alpha_k^2 p_{(kl)(km)} + \sum_k \sum_l \sum_m \alpha_k \alpha_m p_{(kl)(km)} \\ &+ \sum_k \sum_l \sum_m \alpha_k \delta_{km} p_{(kl)(km)} + \sum_k \sum_l \sum_m \alpha_l \alpha_k p_{(kl)(km)} \\ &+ \sum_k \sum_l \sum_m \alpha_l \alpha_m p_{(kl)(km)} + \sum_k \sum_l \sum_m \alpha_l \delta_{km} p_{(kl)(km)} \\ &+ \sum_k \sum_l \sum_m \alpha_k \delta_{kl} p_{(kl)(km)} + \sum_k \sum_l \sum_m \alpha_m \alpha_{kl} p_{(kl)(km)} \\ &+ \sum_k \sum_l \sum_m \delta_{kl} \delta_{km} p_{(kl)(km)} \Bigg]. \end{aligned}$$

From (5) it is easy to check that

$$\begin{aligned}&\sum_m p_{(kl)(km)} = p_{kl}, \quad \sum_l p_{(kl)(km)} = p_{km},\\ &\sum_k p_{(kl)(km)} = (1-f^2)p_l p_m + f^2 p_l I(m=l),\end{aligned} \tag{A.5}$$

so by (A.2) the coefficient of Δ_{8ij} is

$$\begin{aligned}&\sum_k \alpha_k^2 p_k + 2\sum_k\sum_l \alpha_k\alpha_l p_{kl} + \sum_l\sum_m \alpha_l\alpha_m((1-f^2)p_l p_m\\ &+ f^2 p_l I(l=m)) + 2\sum_k\sum_l\sum_m \alpha_l\delta_{km}p_{(kl)(km)}\\ &+ \sum_k\sum_l\sum_m \delta_{kl}\delta_{km}p_{(kl)(km)}.\end{aligned}$$

Since $\sum_k\sum_l \alpha_k\alpha_l p_{kl} = f\sigma_a^2/2$, the above is

$$\begin{aligned}&\left(\frac{1}{2} + f + \frac{f^2}{2}\right) + 2\sum_k\sum_l\sum_m \alpha_l\delta_{km}p_{(kl)(km)}\\ &+ \sum_k\sum_l\sum_m \delta_{kl}\delta_{km}p_{(kl)(km)}.\end{aligned} \tag{A.6}$$

By (5) and (A.1), the middle term in the above is

$$2\sum_k\sum_l \alpha_l(p_{kl}/p_k)\sum_m \delta_{km}p_{km} = 0.$$

By the same way, the last term in (A.6) is

$$\sum_k\sum_l \delta_{kl}(p_{kl}/p_k)\sum_m \delta_{km}p_{km} = 0,$$

so the coefficient of Δ_{8ij} is

$$\frac{(1+f)^2}{2}\sigma_a^2. \tag{A.7}$$

Now collecting terms we have

$$\begin{aligned}\mathrm{Cov}(g_i,g_j|f) =&\Delta_{7ij}\left[\left(1+\frac{f}{2}\right)\sigma_a^2 + (1-f)\sigma_d^2 + f\sigma_0^2\right]\\ &+ \Delta_{8ij}\frac{(1+f)^2}{2}\sigma_a^2.\end{aligned}$$

Appendix B

When $i=j$, $\pi_{ii}'=2$ which is noninformative about trait-marker relationship, so $\mathrm{Cov}(g_i,g_i|f,\zeta,\pi_{ii}) = \mathrm{Cov}(g_i,g_i|f)$, which has the same expression as in (8). When $i\neq j$,

$$\begin{aligned}&\mathrm{Cov}(g_i,g_j|f,\zeta,\pi_{ij}) = \Delta_{7ij}(\pi_{ij})\sum_k\sum_l(\alpha_k+\alpha_l+\delta_{kl})^2 p_{(kl)(kl)|\pi'_{ij}=2}\\ &+ \Delta_{8ij}(\pi_{ij})\sum_k\sum_l\sum_m(\alpha_k+\alpha_l+\delta_{kl})\times(\alpha_k+\alpha_m+\delta_{km})p_{(kl)(km)|\pi'_{ij}=1}\\ &+ \Delta_{9ij}(\pi_{ij})\sum_k\sum_l\sum_m\sum_n(\alpha_k+\alpha_l+\delta_{kl})(\alpha_m+\alpha_n+\delta_{mn})p_{(kl)(mn)|\pi'_{ij}} = 0.\end{aligned} \tag{B.1}$$

We first derive the conditional probabilities $p'_{(kl)(mn)|\pi_{ij}}$s in (B.1). Since conditioning on the IBD status, those quantities are independent of relatedness of the pair, only depend on the relationships among the trait and marker alleles through f and ζ, in other words, given IBD status, different alleles in one configuration are independent with those in the other one. We have

$$p_{(kl)(kl)|\pi'_{ij}=0} = \sum_r\sum_s\sum_u\sum_v P\left(\frac{A_kA_l}{a_ra_s}, \frac{A_kA_l}{a_ua_v}\right).$$

Now the two configurations share A_kA_l in common, if we fix it, the two configurations are independent each other, so we rewrite the above as

$$\begin{aligned}&\sum_r\sum_s\sum_u\sum_v P(A_kA_l)P\left(\frac{A_kA_l}{a_ra_s}, \frac{A_kA_l}{a_ua_v}\middle|A_kA_l\right)\\ &= P(A_kA_l)\sum_r\sum_s\sum_u\sum_v P\left(\frac{A_kA_l}{a_ra_s}\middle|A_kA_l\right)P\left(\frac{A_kA_l}{a_ua_v}\middle|A_kA_l\right)\\ &= P(A_kA_l)\sum_r\sum_s\sum_u\sum_v \frac{P((A_kA_l)/(a_ra_s))}{P(A_kA_l)}\frac{P((A_kA_l)/(a_ua_v))}{P(A_kA_l)}\\ &= \frac{1}{p_{kl}}\sum_r\sum_s\sum_u\sum_v p_{(kl,rs)}p_{(kl,uv)} = p_{kl}.\end{aligned} \tag{B.2}$$

Similarly,

$$\begin{aligned}&p_{(kl)(km)|\pi'_{ij}=0}\\ &= \sum_r\sum_s\sum_u\sum_v P\left(\frac{A_kA_l}{a_ra_s}, \frac{A_kA_m}{a_ua_v}\right)\\ &= P(A_k)\sum_r\sum_s\sum_u\sum_v P\left(\frac{A_kA_l}{a_ra_s}, \frac{A_kA_m}{a_ua_v}\middle|A_k\right)\\ &= P(A_k)\sum_r\sum_s\sum_u\sum_v P\left(\frac{A_kA_l}{a_ra_s}\middle|A_kA_l\right)P\left(\frac{A_kA_m}{a_ua_v}\middle|A_kA_l\right)\\ &= \frac{1}{p_k}\sum_r\sum_s\sum_u\sum_v p_{(kl,rs)}p_{(km,uv)} = \frac{p_{kl}p_{km}}{p_k},\end{aligned} \tag{B.3}$$

$$\begin{aligned}p_{(kl)(mn)|\pi'_{ij}=0} &= \sum_r\sum_s\sum_u\sum_v P\left(\frac{A_kA_l}{a_ra_s}, \frac{A_mA_n}{a_ua_v}\right)\\ &= \sum_r\sum_s P\left(\frac{A_kA_l}{a_ra_s}\right)\sum_u\sum_v P\left(\frac{A_mA_n}{a_ua_v}\right)\\ &= \sum_r\sum_s p_{(kl,rs)}\sum_u\sum_v p_{(mn,uv)} = p_{kl}p_{mn},\end{aligned} \tag{B.4}$$

$$\begin{aligned}p_{(kl)(kl)|\pi'_{ij}=0} = \sum_r\sum_s\sum_u&\left[\frac{1}{2}P\left(\frac{A_kA_l}{a_ra_s}\frac{A_kA_l}{a_ua_v}\right)\right.\\ &\left.+\frac{1}{2}P\left(\frac{A_kA_l}{a_sa_r}, \frac{A_kA_l}{a_ua_r}\right)\right],\end{aligned}$$

and

$$\begin{aligned}&\sum_r\sum_s\sum_u P\left(\frac{A_kA_l}{a_ra_s}, \frac{A_kA_l}{a_ra_u}\right)\\ &= \sum_r\sum_s\sum_u P\left(\frac{A_kA_l}{a_r}\right)P\left(\frac{A_kA_l}{a_ra_s}, \frac{A_kA_l}{a_ra_u}\middle|\frac{A_kA_l}{a_r}\right)\\ &= \sum_r\sum_s\sum_u p_{(kl,r)}\frac{p_{(kl,rs)}}{p_{(kl,r)}}\frac{p_{(kl,ru)}}{p_{(kl,r)}} = \sum_r\sum_s\sum_u\frac{p_{(kl,rs)}p_{(kl,ru)}}{p_{(kl,r)}}\\ &= \sum_r\frac{p_{(kl,r)}p_{(kl,r)}}{p_{(kl,r)}} = p_{kl,}\end{aligned}$$

where $p_{(kl,\,r)}=\sum_s p_{(kl,\,rs)}$. The same reason gives

$$\sum_r\sum_s\sum_u P\left(\frac{A_kA_l}{a_sa_r},\frac{A_kA_l}{a_ua_r}\right)=p_{kl},$$

so we have

$$p_{(kl)(kl)|\pi'_{ij}}=1=p_{kl}. \tag{B.5}$$

Also

$$p_{(kl)(kl)|\pi'_{ij}=1}=\sum_r\sum_s\sum_u\frac{p_{(kl,rs)}p_{(km,ru)}}{p_{(k,r)}}=\sum_r\frac{p_{(kl,r)}p_{(km,r)}}{p_{(k,r)}},$$

where

$$p_{(kl,r)}=\sum_s p(kl,rs)=p_{kl}(q_r-\zeta p_r)+\zeta p_{kl}I(r=k)$$
$$\text{and}p_{(k,r)}=\sum_l p(kl,r)=p_k(q_r-\zeta p_r)+\zeta p_kI(r=k).$$

So

$$\begin{aligned}&p_{(kl)(km)|\pi'_{ij}=1}\\&=\sum_r\frac{[p_{kl}(q_r-\zeta p_r)+\zeta p_{kl}I(r=k)][p_{km}(q_r-\zeta p_r)+\zeta p_{km}I(r=k)]}{p_k(q_r-\zeta p_r)+\zeta p_kI(r=k)}\\&=\sum_{r\neq k}\frac{p_{kl}p_{km}}{p_k}(q_r-\zeta p_r)\\&+\frac{[p_{kl}(q_k-\zeta p_k)+\zeta p_{kl}][p_{km}(q_k-\zeta p_k)+\zeta p_{km}]}{p_k(q_k-\zeta p_k)+\zeta p_k}=\frac{p_{kl}p_{km}}{p_k},\end{aligned} \tag{B.6}$$

$$\begin{aligned}&p_{(kl)(mn)|\pi'_{ij}=1}=\sum_r\sum_s\sum_u\left[\frac12\frac{p_{(kl,rs)}p_{(mn,ru)}}{q_r}+\frac12\frac{p_{(kl,sr)}p_{(mn,ur)}}{q_r}\right],\\&\sum_r\sum_s\sum_u\frac{p_{(kl,rs)}p_{(mn,ru)}}{q_r}=\sum_r\frac{p_{(kl,r)}p_{(mn,r)}}{q_r}\\&=p_{kl}p_{mn}\sum_r\frac{1}{q_r}(q_r-\zeta p_r+\zeta I(r=k))\\&\times(q_r-\zeta p_r+\zeta I(r=m))\\&=p_{kl}p_{mn}\left[1+\zeta^2\sum_r\frac{p_r^2}{q_r}-\zeta^2\left(\frac{p_m}{q_m}+\frac{p_k}{q_k}\right)+\zeta^2\frac{I(k=m)}{q_k}\right],\end{aligned}$$

and

$$\begin{aligned}&\sum_r\sum_s\sum_u\frac{p_{(kl,sr)}p_{(mn,ur)}}{q_r}\\&=p_{kl}p_{mn}\left[1+\zeta^2\left(\sum_r\frac{p_r^2}{q_r}-\frac{p_n}{q_n}-\frac{p_l}{q_l}+\frac{I(n=l)}{q_l}\right)\right],\end{aligned}$$

so

$$\begin{aligned}p_{(kl)(mn)|\pi'_{ij}=1}=&p_{kl}p_{mn}\left[1+\zeta^2\left(\sum_r\frac{p_r^2}{q_r}-\frac12\left(\frac{p_m}{q_m}+\frac{p_k}{q_k}+\frac{p_n}{q_n}+\frac{p_l}{q_l}\right)\right.\right.\\&\left.\left.+\frac12\left(\frac{I(k=m)}{q_k}+\frac{I(n=l)}{q_l}\right)\right)\right]\end{aligned} \tag{B.7}$$

Also

$$p_{(kl)(kl)|\pi'_{ij}=2}=\sum_r\sum_s P\left(\frac{A_kA_l}{a_ra_s}\right)=\sum_r\sum_s p_{(kl,rs)=}p_{kl}, \tag{B.8}$$

and

$$\begin{aligned}p_{(kl)(km)|\pi_{ij}=2}&=\sum_r\sum_s P\left(\frac{A_kA_l}{a_ra_s},\frac{A_kA_m}{a_ra_s}\right)\\&=\sum_r\sum_s P\left(\frac{A_k}{a_ra_s}\right)P\left(\frac{A_kA_l}{a_ra_s},\frac{A_kA_m}{a_ra_s}\middle|\frac{A_k}{a_ra_s}\right)\\&=\sum_r\sum_s P\left(\frac{A_k}{a_ra_s}\right)P\left(\frac{A_kA_l}{a_ra_s}\middle|\frac{A_k}{a_ra_s}\right)P\left(\frac{A_kA_m}{a_ra_s}\middle|\frac{A_k}{a_ra_s}\right)\\&=\sum_r\sum_s\frac{p_{(kl,rs)}p_{(km,rs)}}{p_{(k,rs)}},\end{aligned}$$

where $p_{(k,\,rs)}=\sum_l p_{(kl,\,rs)}=p_k(q_{rs}-\zeta p_{rs}+\zeta(1-f)p_sI(r=k)+\zeta I(r=s=k))$, so

$$\begin{aligned}p_{(kl)(km)|\pi'_{ij}=2}=&\frac{p_{kl}p_{km}}{pk}\sum_r\sum_s(q_{rs}-\zeta p_{rs}+\zeta I((r,s)=(k,l)))\\&\times\frac{(q_{rs}-\zeta p_{rs}+\zeta I((r,s)=(k,m)))}{q_{rs}-\zeta p_{rs}+\zeta(1-f)p_sI(r=k)+\zeta I(r=s=k)}\\=&\frac{p_{kl}p_{km}}{p_k}[(1-\zeta)-(q_k-\zeta p_k)]\\&+\sum_s\frac{(q_{ks}-\zeta p_{ks})^2}{q_{rs}-\zeta p_{rs}+\zeta(1-f)p_s+\zeta I(s=k)}\\&+\frac{2\zeta(q_{km}-\zeta p_{km})}{q_{km}-\zeta p_{km}+\zeta(1-f)p_m+\zeta fI(k=m)}\\&+\frac{\zeta^2I(l=m)}{q_{km}-\zeta p_{km}+\zeta(1-f)p_m+\zeta fI(l=m)}\Bigg].\end{aligned} \tag{B.9}$$

Lastly

$$\begin{aligned}p_{(kl)(mn)|\pi'_{ij}=2}&=\sum_r\sum_s P\left(\frac{A_kA_l}{a_ra_s},\frac{A_mA_n}{a_ra_s}\right)\\&=\sum_r\sum_s P(a_ra_s)P\left(\frac{A_kA_l}{a_ra_s},\frac{A_mA_n}{a_ra_s}|a_ra_s\right)\\&=\sum_r\sum_s\frac{p_{(kl,rs)}p_{(mn,rs)}}{q_{rs}}\\&=\sum_r\sum_s[p_{kl}(q_{rs}-\zeta p_{rs})+\zeta p_{kl}I((r,s)=(k,l))]\\&\times\frac{[p_{mn}(q_{rs}-\zeta p_{rs})+\zeta p_{mn}I((r,s)=(m,n))]}{q_{rs}}\\&=p_{kl}p_{mn}\left[1+\zeta^2\sum_r\sum_s\frac{p_{rs}^2}{q_{rs}}-\zeta^2\left(\frac{p_{mn}}{q_{mn}}+\frac{p_{kl}}{q_{kl}}\right)\right.\\&\left.+\zeta^2I((k,l)=(m,n))\right].\end{aligned} \tag{B.10}$$

Now we compute the covariance (B.1) for different values of the π'_{ij}'s. If $\pi'_{ij}=0$, by (B.2)–(B.4) and Appendix A, we have the same expression of (B.1) as in (8).

If $\pi'_{ij}=1$, by (B.5)–(B.7), the coefficient of $\Delta_{7ij}(\pi_{ij})$ and $\Delta_{8ij}(\pi_{ij})$ in (B.1) are the same as that in (A.4) and (A.7); the coefficient of $\Delta_{9ij}(\pi_{ij})$ in (B.1) has four terms corresponding to those in (B.7), the first two terms are zero by the computation in Appendix A, by its symmetry in (k,l,m,n),

the last two terms are

$$\sum_k \sum_l \sum_m \sum_n (\alpha_k + \alpha_l + \delta_{kl})(\alpha_m + \alpha_n + \delta_{mn}) \left(-2\zeta^2 \frac{p_{kl}p_{mn}p_m}{q_m} + \zeta^2 \frac{p_{kl}p_{mn}}{q_k} I(m=k)\right), \tag{B.11}$$

since

$$\sum_k \sum_l (\alpha_k + \alpha_l + \delta_{kl}) p_{kl} = 0,$$

the first term above is zero. By expanding and check each term using (A.1) and (A.2), the second term above, and hence the coefficient of $\Delta_{9ij}(\pi_{ij})$ is

$$\zeta^2 (1+f)^2 \sum_k \alpha_k^2 \frac{p_k^2}{q_k}.$$

If $\pi'_{ij} = 2$, by (B.8), the coefficient of $\Delta_{7ij}(\pi_{ij})$ is the same as before. Now we compute the coefficient of $\Delta_{8ij}(\pi_{ij})$. We expand it in five terms as in (B.9). The first term is $(1+f)^2(1-\zeta)\sigma_a^2/2$ by the computation in Appendix A. Expanding the same way as in Appendix A, the second term is

$$\begin{aligned} &- \sum_k \alpha_k^2 p_k (q_k - \zeta p_k) - 2 \sum_k \sum_l \alpha_k \alpha_l p_{kl} (q_k - \zeta p_k) \\ &- \sum_k \sum_l \sum_m \alpha_k \alpha_l \frac{p_{kl}p_{km}}{p_k} (q_k - \zeta p_k) \\ &- 2 \sum_k \sum_l \sum_m \alpha_l \delta_{km} \frac{p_{kl}p_{km}}{p_k} (q_k - \zeta p_k) \\ &- \sum_k \sum_l \sum_m \delta_{kl}\delta_{km} \frac{p_{kl}p_{km}}{p_k} (q_k - \zeta p_k), \end{aligned} \tag{B.12}$$

the last two terms above are zero by (A.1). Since

$$\sum_l \alpha_l p_{kl} = f \alpha_k p_k, \tag{B.13}$$

substitute this into the second and the third term in (B.12), it becomes $-(1+f)^2 \sum_k \alpha_k^2 p_k (q_k - \zeta p_k)$. By expanding the same way, the third term is

$$\begin{aligned} &\sum_k \alpha_k^2 p_k \sum_s \frac{(q_{ks} - \zeta p_{ks})^2}{q_{ks} - \zeta p_{ks} + \zeta(1-f)p_s + \zeta f I(q=k)} \\ &+ 2 \sum_k \sum_l \alpha_k \alpha_l p_{kl} \sum_s \frac{(q_{ks} - \zeta p_{ks})^2}{q_{ks} - \zeta p_{ks} + \zeta(1-f)p_s + \zeta f I(s=k)} \\ &+ \sum_k \sum_l \sum_m \alpha_l \alpha_m \frac{p_{kl}p_{km}}{p_k} \sum_s \frac{(q_{ks} - \zeta p_{ks})^2}{q_{ks} - \zeta p_{ks} + \zeta(1-f)p_s + \zeta f I(s=k)} \\ &+ 2 \sum_k \sum_l \sum_m \alpha_l \delta_{km} \frac{p_{kl}p_{km}}{p_k} \sum_s \frac{(q_{ks} - \zeta p_{ks})^2}{q_{ks} - \zeta p_{ks} + \zeta(1-f)p_s + \zeta f I(s=k)} \\ &+ \sum_k \sum_l \sum_m \delta_{kl}\delta_{km} \frac{p_{kl}p_{km}}{p_k} \sum_s \frac{(q_{ks} - \zeta p_{ks})^2}{q_{ks} - \zeta p_{ks} + \zeta(1-f)p_s + \zeta f I(s=k)}. \end{aligned} \tag{B.14}$$

The last two terms in (B.14) are zero. Substitute (B.13) into the second and third term in (B.14), it becomes

$$(1+f)^2 \sum_k \alpha_k^2 p_k \sum_s \frac{(q_k - \zeta p_{ks})^2}{q_{ks} - \zeta p_{ks} + \zeta(1-f)p_s + \zeta f I(s=k)};$$

now combine the second and the third terms gives

$$-\zeta(1+f)^2 \sum_k \alpha_k^2 p_k \sum_s \frac{(q_{ks} - \zeta p_{ks})[(1-f)p_s + f I(s=k)]}{q_{ks} - \zeta p_{ks} + \zeta(1-f)p_s + \zeta f I(s=k)};$$

the fourth term is

$$2\zeta \sum_k \sum_l \sum_m (\alpha_k + \alpha_l + \delta_{kl})(\alpha_k + \alpha_m + \delta_{km}) \frac{p_{(kl)(km)}(q_{km} - \zeta p_{km})}{q_{km} - \zeta p_{km} + \zeta(1-f)p_m + f\zeta I(k=m)};$$

the fifth term is

$$\zeta^2 \sum_k \sum_l (\alpha_k + \alpha_l + \delta_{kl})^2 \frac{p_{kl}^2}{p_k[q_{kl} - \zeta p_{kl} + \zeta(1-f)p_l + f\zeta]}.$$

For the coefficient of $\Delta_{9ij}(\pi_{ij})$, we expand it in four terms according to (B.10), the first two terms are zero by the computation in Appendix A, so it reduces to

$$\sum_k \sum_l \sum_m \sum_n (\alpha_k + \alpha_l + \delta_{kl})(\alpha_m + \alpha_n + \delta_{mn}) \left(-2\zeta^2 p_{kl} \frac{p_{mn}^2}{q_{mn}} + \zeta^2 p_{kl}p_{mn} I((k,l)=(m,n))\right),$$

the first term above is zero since $\sum\sum(\alpha_k + \alpha_l + \delta_{kl})p_{kl} = 0$, the second term, and hence the coefficient of $\Delta_{9ij}(\pi_{ij})$ is $\zeta^2 \sum_k \sum_l (\alpha_k + \alpha_l + \delta_{kl})^2 p_{kl}^2$, which is

$$\zeta^2 \left[2 \sum_k \sum_l \alpha_k^2 p_{kl}^2 + \sum_k \sum_l \delta_{kl}^2 p_{kl}^2 + 4 \sum_k \sum_l \alpha_k \delta_{kl} p_{kl}^2 + 2 \sum_k \sum_l \alpha_k \alpha_l p_{kl}^2\right].$$

The first term in the bracket above is

$$\begin{aligned} &2 \sum_k \sum_l \alpha_k^2 [(1-f)p_k p_l + f p_k I(l=k)]^2 \\ &= 2(1-f)^2 \sum_k \sum_l \alpha_k^2 p_k^2 p_l^2 + 4f(1-f) \sum_k \alpha_k^2 p_k^3 \\ &+ 2f^2 \sum_k \alpha_k^2 p_k^2; \end{aligned}$$

the second term is

$$\begin{aligned} &\sum_k \sum_l \delta_{kl}^2 [(1-f)p_k p_l + f p_k I(l=k)]^2 \\ &= (1-f)^2 \sum_k \sum_l \delta_{kl}^2 p_k^2 p_l^2 + 2f(1-f) \sum_k \delta_{kk}^2 p_k^3 \\ &+ f^2 \sum_k \delta_{kk}^2 p_k^2; \end{aligned}$$

the third term is

$$\begin{aligned} &4 \sum_k \sum_l \alpha_k \delta_{kl} [(1-f)p_k p_l + f p_k I(l=k)]^2 \\ &= 4(1-f)^2 \sum_k \sum_l \alpha_k \delta_{kl} p_k^2 p_l^2 + 8f(1-f) \sum_k \alpha_k \delta_{kk} p_k^3 \\ &+ 4f^2 \sum_k \alpha_k \delta_{kk} p_k^2; \end{aligned}$$

the fourth term is

$$\begin{aligned}&2\sum_k\sum_l \alpha_k\alpha_l[(1-f)p_kp_l+fp_kI(l=k)]^2\\&=2(1-f)^2\left(\sum_k\alpha_kp_k^2\right)^2+4f(1-f)\sum_k\alpha_k^2p_k^3\\&+2f^2\sum_k\alpha_k^2p_k^2.\end{aligned}$$

Now collect terms, the coefficient of $\Delta_{9ij}(\pi_{ij})$ is

$$\begin{aligned}&\zeta^2\Bigg[2(1-f)^2\sum_k\sum_l\alpha_k^2p_k^2p_l^2+8f(1-f)\sum_k\alpha_k^2p_k^3\\&+4f^2\sum_k\alpha_k^2p_k^2+(1-f)^2\sum_k\sum_l\delta_{kl}^2p_k^2p_l^2\\&+2f(1-f)\sum_k\delta_{kk}^2p_k^3+f^2\sum_k\delta_{kk}^2p_k^2\\&+4(1-f)^2\sum_k\sum_l\alpha_k\delta_{kl}p_k^2p_l^2+8f(1-f)\sum_k\alpha_k\delta_{kk}p_k^3\\&+4f^2\sum_k\alpha_k\delta_{kk}p_k^2+2(1-f)^2\left(\sum_k\alpha_kp_k^2\right)^2\Bigg].\end{aligned}$$

Appendix C

Let $\xi=(\alpha,\beta)$, $\xi_0=(\alpha_0,\beta)$, and define ξ_1 and the hat notations for the corresponding estimates. Let $\hat{I}(\xi)$ be the empirical Fisher information matrix evaluated at (ξ), by Taylor expansion,

$$L\left(\hat{\xi}_0|\mathbf{Y}\right)=L(\xi_0|\mathbf{Y})+\frac{n}{2}\left(\hat{\xi}_0-\xi_0\right)'\hat{I}^{-1}(\xi_0)\left(\hat{\xi}_0-\xi_0\right)+o_P(1),$$
$$L\left(\hat{\xi}_1|\mathbf{Y}\right)=L(\xi_1|\mathbf{Y})+\frac{n}{2}\left(\hat{\xi}_1-\xi_1\right)'\hat{I}^{-1}(\xi_1)\left(\hat{\xi}_1-\xi_1\right)+o_P(1),$$

and its is well known that, under H_1, as $n\to\infty$,

$$\begin{aligned}&\frac{n}{2}\left(\left(\hat{\xi}_1-\xi_1\right)\right)'\hat{I}^{-1}(\xi_1)\left(\hat{\xi}_1-\xi_1\right)\\&-\left(\hat{\xi}_0-\xi_0\right)'\hat{I}^{-1}(\xi_0)\left(\hat{\xi}_0-\xi_0\right)\xrightarrow{D}\chi_k^2.\end{aligned}$$

Also, since the familial structures are homogeneous, so

$$\begin{aligned}\log\frac{L(\alpha_1|\mathbf{Y})}{L(\xi_0|\mathbf{Y})}&=n\frac{1}{n}\sum_{i=1}^n\log\frac{L(\xi_1|\mathbf{Y}_i)}{L(\xi_0|\mathbf{Y}_i)}\\&=nD(\alpha_1\|\alpha_0)+o_P\left(\sqrt{n}\right).\end{aligned}$$

Thus under H_1,

$$\begin{aligned}&-2\log\frac{L\left(\hat{\xi}_0|\mathbf{Y}\right)}{L\left(\hat{\xi}_1|\mathbf{Y}\right)}=2\Bigg(\log\frac{L(\xi_1|\mathbf{Y})}{L(\xi_0|\mathbf{Y})}\\&+\frac{n}{2}\left(\hat{\xi}_1-\xi_1\right)'\hat{I}^{-1}(\xi_1)\left(\hat{\xi}_1-\xi_1\right)-\left(\hat{\xi}_0-\xi_0\right)'\hat{I}^{-1}(\xi_0)\left(\hat{\xi}_0-\xi_0\right)\Bigg)\\&\approx 2nD(\alpha_1||\alpha_0)+\chi_k^2.\end{aligned}$$