Heredity (2005) 94, 338–346
© 2005 Nature Publishing Group All rights reserved 0018-067X/05 $30.00
www.nature.com/hdy

Estimating the correlation of pairwise relatedness along chromosomes

X-S Hu

Department of Forest Sciences, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4

The 'spatial' pattern of the correlation of pairwise relatedness among loci within a chromosome is an important aspect for an insight into genomic evolution in natural populations. In this article, a statistical genetic method is presented for estimating the correlation of pairwise relatedness among linked loci. The probabilities of identity-in-state (IIS) are related to the probabilities of identity-by-descent (IBS) for the two- and three-loci cases. By decomposing the joint probabilities of two- or three-loci IBD, the probability of pairwise relatedness at a single locus and its correlation among linked loci can be simultaneously estimated. To provide effective statistical methods for estimation, weighted least square (LS) and maximum likelihood (ML) methods are evaluated through extensive Monte Carlo simulations. Results show that the ML method gives a better performance than the weighted LS method with haploid genotypic data. However, there are no significant differences between the two methods when two- or three-loci diploid genotypic data are employed. Compared with the optimal size for haploid genotypic data, a smaller optimal sample size is predicted with diploid genotypic data.

Heredity (2005) **94,** 338–346. doi:10.1038/sj.hdy.6800586
Published online 8 September 2004

Keywords: pairwise relatedness; identity-by-descent; identity-in-state; correlation

Introduction

Pairwise relatedness at a single locus and its correlation among linked loci along chromosomes give insights into patterns of genomic evolution in natural populations. Pairwise relatedness reveals the genetic similarity between individuals due to recently shared ancestors, and is affected by several evolutionary forces (Wright, 1969). The correlation of pairwise relatedness and its 'spatial' pattern along chromosomes may reflect differential processes of co-evolution that have occurred among different regions of chromosomes. Hitchhiking effects and background selection (Maynard Smith and Haigh, 1974; Charlesworth *et al*, 1993) can enhance a positive regional correlation of pairwise relatedness. Recombination separates linked alleles that have a common ancestor, and hence alters the coalescence process for linked loci (eg Hudson, 1991), resulting in a negative correlation in pairwise relatedness for the two loci. Random drift or founder effects can also reduce the correlation in pairwise relatedness. In general, the joint effects of different forces can bring about a nonuniform distribution of the correlation of relatedness along chromosomes. Thus, a combination of linkage maps and the correlation of pairwise relatedness among loci can help us understand the 'patchy' pattern of genomic evolution.

The significance of pairwise relatedness has long been appreciated in understanding population genetic structure and evolution (Wright, 1922, 1969; Cotterman, 1940; Jacquard, 1974). Estimation of the pairwise relatedness is now simplified by the advent of abundant polymorphic markers, such as microsatellite markers and single-nucleotide polymorphisms (SNPs) (eg Brookes, 1999). The populations studied can either be populations with known pedigree or populations with unknown pedigree. There have been numerous studies exploring the marker-based statistical methods for the purpose of estimating pairwise relatedness (eg Thompson, 1975; Pamilo and Crozier, 1982; Lynch, 1988; Queller and Goodnight, 1989; Ritland, 1996; Lynch and Ritland, 1999; Wang, 2002; Milligan, 2003). However, these previous analyses mainly focus on the average pairwise relatedness per locus and have not been coupled with assessment of the genome-wide diversity in relatedness. The heterogeneity of pairwise relatedness along chromosomes cannot be assessed using these methods.

Previous theories on kinship mapping, the relationship between the joint probability of identity-by-descent (IBD) of linked markers and the recombination fraction (Morton *et al*, 1971; Morton and Simpson, 1983) did not examine the interaction among linked loci in terms of the correlation of relatedness. A recent theoretical study evaluated the role of drift, gene flow, selfing, and mutation in affecting the association of gene identity, and shows that a small identity disequilibrium (ID) within subpopulations is present (Vitalis and Couvet, 2001a). ID is termed as the difference between the joint two-loci probability of identity-in-state (IIS) and the expected product of each component's IIS probability. Such an ID approach is essentially distinct from the correlation of relatedness, which assesses the difference in terms of IBD. Unlike Vitalis and Couvet (2001b), who apply ID to jointly estimate the local effective population size and migration rate, I use the joint probabilities of two- and three-loci IBD to address the correlation of pairwise relatedness among loci.

Correspondence: X-S Hu, Department of Renewable Resources, 751 General Service Building, University of Alberta, Edmonton, Canada AB T6G 2H1. E-mail: xin-sheng.hu@ualberta.ca
Published online 8 September 2004

The purpose of this article is to develop a new method for jointly estimating the pairwise relatedness at a single locus and its correlation among loci. Unlike previous methods where pairwise relatedness is estimated at individual loci and then weighted to gain the average (eg Ritland, 1996; Lynch and Ritland, 1999; Wang, 2002), the present method relies on the estimation of the joint probabilities of IBD at two or three loci.

In my analyses of pairwise relatedness and its correlation among loci, data can be randomly sampled from either the haploids or diploids genotyped with codominant markers, each with an arbitrary number of alleles. The use of haploid and of diploid genotypic data is evaluated through extensive Monte Carlo simulations. The preexisting approaches are based on sampling diploids (eg Ritland, 1996; Lynch and Ritland, 1999). However, the approach of using the haploid genotypic data allows the method to be applicable to specific chromosomes, such as the sex chromosomes, and has an advantage of ignoring the effects of mating system.

Two statistical methods are extensively evaluated through simulation study: the weighted least squares (LS) and the maximum likelihood (ML) methods. Application of the simulation results to practical analysis is discussed.

Two-loci relatedness

Throughout the study, the concept of pairwise relatedness at a single locus refers to the probability that an allele randomly sampled from one individual is IBD with an allele at the same locus randomly sampled from another individual (eg Jacquard, 1974; Ritland, 1996; Lynch and Ritland, 1999). Rousset (2002) reviewed some properties of this relatedness definition.

Haploid data:

Consider a pair of two linked codominant loci in a population, denoted by A and B, with numbers of n_A and n_B alleles, respectively. Let p_u and q_v be the frequencies of alleles A_u $(u=1,2,\ldots,n_A;$ $\sum_u p_u = 1)$ and B_v $(v=1,2,\ldots,n_B;$ $\sum_v q_v = 1)$, respectively. Following the classical definition on the probability of IBD at a single locus (Jacquard, 1974), four parameters are defined in the two-loci case: κ_{11} $(0 \le \kappa_{11} \le 1)$ is the probability that the two alleles at each of the two loci are IBD, κ_{10} $(0 \le \kappa_{10} \le 1)$ is the probability that the two alleles at the A locus are IBD but the two alleles at the B locus are not, κ_{01} $(0 \le \kappa_{01} \le 1)$ is the probability that the two alleles at the B locus are IBD but the two alleles at the A locus are not, and κ_{00} $(=1-\kappa_{11-}\kappa_{10}-\kappa_{01})$ is the probability that the two alleles at each of the two loci are not IBD. Note that there are several definitions of the probability of two-loci descent (Whitlock *et al*, 1993; Vitalis and Couvet, 2001a, b; Laurie and Weir, 2003), but the present definition of κ_{11} is actually the same as the definition F_{11} of Whitlock *et al* (1993).

The basic approach for estimating the relatedness is to use the probabilities of IIS to infer the relatedness parameters, similar to previous studies (eg Ritland, 1996). Denote by $P^{uv}_{u'v'}$ the probability that a pair of two-loci gametes have genotypes A_uB_v and $A_{\acute{u}}B_{\acute{v}}$ $(u,\ u'=1,2,\ldots,n_A;\ v,\ v'=1,2,\ldots,n_B)$. When linkage disequilibrium (LD) is absent, $P^{uv}_{u'v'}$ can be decomposed as

$$\begin{aligned} P^{uv}_{u'v'} = {} & \delta_{uu'}\delta_{vv'}p_u q_v \kappa_{11} + \delta_{uu'}(2-\delta_{vv'})p_u q_v q_{v'}\kappa_{10} \\ & + (2-\delta_{uu'})\delta_{vv'}p_u p_{u'} q_v \kappa_{01} \\ & + \frac{(2-\delta_{uu'})(2-\delta_{vv'})}{2^{(1-\delta_{uu'})(1-\delta_{vv'})}} p_u p_{u'} q_v q_{v'} \kappa_{00} \end{aligned} \tag{1}$$

where $\delta_{uu'}$ is Kronecker delta variable, which is equal to unity when $u=u'$ and zero otherwise. The factor $2^{-(1-\delta_{uu'})(1-\delta_{vv'})}$ in the coefficient of κ_{00} on the right side of equation (1) is introduced so that the coupling and repulsion linkage phases can be separated, which is distinct from the previous four-gene case at a single locus (Lynch and Ritland, 1999). Note that the gamete and allele frequencies in equation (1) are assumed known beforehand with sufficient accuracy, as in previous studies (eg Ritland, 1996, 2000; Lynch and Ritland, 1999).

When LD is present, a more general expression of $P^{uv}_{u'v'}$ can be written as

$$\begin{aligned} P^{uv}_{u'v'} = {} & g_{uv}\delta_{uu'}\delta_{vv'}\kappa_{11} + g_{uv}g_{u'v'}\Bigg(\delta_{uu'}(2-\delta_{vv'})\frac{\kappa_{10}}{p_u} \\ & + (2-\delta_{uu'})\delta_{vv'}\frac{\kappa_{01}}{q_v} + \frac{(2-\delta_{uu'})(2-\delta_{vv'})}{2^{(1-\delta_{uu'})(1-\delta_{vv'})}}\kappa_{00}\Bigg) \end{aligned} \tag{2}$$

where g_{uv} and $g_{u'v'}$ are the frequencies of gametes A_uB_v and $A_{\acute{u}}B_{\acute{v}}$ in the population, respectively.

In the case of two alleles per locus (say A_i, A_j, B_k, and B_l), there are four categories of haplotype pairs according to the number of shared alleles: IIS for both the two alleles of each locus (A_iB_k–A_iB_k, A_iB_l–A_iB_l, A_jB_k–A_jB_k, A_jB_l–A_jB_l), IIS for the two A alleles but not for the two B alleles (A_iB_k–A_iB_l, A_jB_k–A_jB_l) and the reverse case (A_jB_l–A_iB_l, A_iB_k–A_iB_l), and no shared alleles for both loci (A_iB_k–A_jB_l, A_jB_k–A_iB_l). There are 16 types of haplotype pairs, but only 10 of them are distinguishable with codominant markers (Table 1). For an arbitrary number of alleles at each locus, the number of distinguishable haplotype pairs, denoted by n_{AB}, is shown to be equal to $n_An_B(n_An_B+1)/2$.

Diploid data:

When diploid genotypic data are used, the preceding method can be applied as long as the gamete frequencies are available. However, estimation of gamete frequencies requires the assumption of random

Table 1 Probabilities for the two-loci haplotype pairs (two alleles per locus: A_i, A_j; B_k, B_l)

Haplotype pairs	*Coefficient*			
	κ_{11}	κ_{10}	κ_{01}	κ_{00}
A_iB_k–A_iB_k	g_{ik}	$g_{ik}{}^2/p_i$	$g_{ik}{}^2/q_k$	$g_{ik}{}^2$
A_iB_k–A_iB_l	0	$2g_{ik}g_{il}/p_i$	0	$2g_{ik}g_{il}$
A_iB_l–A_iB_l	g_{il}	$g_{il}{}^2/p_i$	$g_{il}{}^2/q_l$	$g_{il}{}^2$
A_jB_k–A_iB_k	0	0	$2g_{jk}g_{ik}/q_k$	$2g_{jk}g_{ik}$
A_jB_k–A_iB_l	0	0	0	$2g_{jk}g_{il}$
A_jB_k–A_jB_k	g_{jk}	$g_{jk}{}^2/p_j$	$g_{jk}{}^2/q_k$	$g_{jk}{}^2$
A_jB_l–A_iB_k	0	0	0	$2g_{jl}g_{ik}$
A_jB_l–A_iB_l	0	0	$2g_{jl}g_{il}/q_l$	$2g_{jl}g_{il}$
A_jB_l–A_jB_k	0	$2g_{jl}g_{jk}/p_j$	0	$2g_{jl}g_{jk}$
A_jB_l–A_jB_l	g_{jl}	$g_{jl}{}^2/p_j$	$g_{jl}{}^2/q_l$	$g_{jl}{}^2$

association of gametes (Hill, 1974; Weir, 1996; Kalinowski and Hedrick, 2001); otherwise, the relative proportion of heterozygotes with repulsion *versus* coupling linkage phases is indistinguishable. When there are nonrandom associations between gametes in forming zygotes (Yang, 2002), expectation–maximization (EM) (Dempster *et al*, 1977) and other methods are not applicable. The following method is only valid under the assumption of random association of gametes.

Denote by H_{AB} the probability that both A and B loci are heterozygous, H_A the probability that the A locus is heterozygous but the B locus is not, H_B the probability that the B locus is heterozygous but the A locus is not, and H_{01} ($H_{01}=1-H_A-H_B-H_{AB}$) the probability that both loci are homozygous. I obtain

$$H_{01}=\sum_u \sum_v \left(g_{uv}\kappa_{11}+g_{uv}^2(\kappa_{00}+\kappa_{10}/p_u+\kappa_{01}/q_v)\right) \tag{3a}$$

$$H_A=2\sum_u \sum_{u'\neq u} \sum_v g_{uv}g_{u'v}(\kappa_{00}+\kappa_{01}/q_v) \tag{3b}$$

$$H_B=2\sum_v \sum_{v'\neq v} \sum_u g_{uv}g_{uv'}(\kappa_{00}+\kappa_{10}/p_u) \tag{3c}$$

$$H_{AB}=2\sum_u \sum_{u'\neq u} \sum_v \sum_{v'\neq v} g_{uv}g_{u'v'}\kappa_{00} \tag{3d}$$

Note that equation (3d) differs from the previous study where only one parameter is considered (Morton and Simpson, 1983). When only one locus is considered, equations (3a)–(3d) reduce to the classical results (Falconer and Mackay, 1996, p 66). In practice, the heterozygote frequencies (H_{01}–H_{AB} variables) can be estimated directly from the genotypic data. Thus, according to equations (3a)–(3d), the three unknown parameters (κ_{11}, κ_{10}, and κ_{01}) can be estimated.

Correlation of pairwise relatedness: Denote by r_A and r_B the probabilities of pairwise relatedness at the A and B loci, respectively, and c_r the covariance of the probabilities of pairwise relatedness between A and B. If the three unknown parameters ($\hat{\kappa}_{11}$, $\hat{\kappa}_{10}$, and $\hat{\kappa}_{01}$) are estimated, the probability of pairwise relatedness at a single locus (r_A and r_B) and its covariance among loci (c_r) can be calculated from the following equations:

$$\hat{\kappa}_{11}=r_A r_B+c_r \tag{4a}$$

$$\hat{\kappa}_{10}=r_A(1-r_B)-c_r \tag{4b}$$

$$\hat{\kappa}_{01}=(1-r_A)r_B-c_r \tag{4c}$$

Solution to equations (4a)–(4c) is $\hat{r}_A=\hat{\kappa}_{11}+\hat{\kappa}_{10}$, $\hat{r}_B=\hat{\kappa}_{11}+\hat{\kappa}_{01}$, and $\hat{c}_r=\hat{\kappa}_{11}\hat{\kappa}_{00}-\hat{\kappa}_{10}\hat{\kappa}_{01}$.

From equation (4a), c_r can also be viewed as the kinship disequilibrium since it is expressed as the difference between the joint probability of two-loci IBD and the product of single-locus probability of IBD, analogous to the definition of ID (Vitalis and Couvet, 2001a). c_r may be negative when $\kappa_{11}\kappa_{00}<\kappa_{10}\kappa_{01}$, or positive when $\kappa_{11}\kappa_{00}>\kappa_{10}\kappa_{01}$. Theoretically, c_r is associated with the recombination fraction, or inversely proportional to the physical distance between the two linked loci. A smaller distance between two loci implies a stronger correlation of their relatedness. In order to make the correlation be comparable among different pairs of loci, the correlation coefficient of pairwise relatedness is defined as $\hat{c}_r(\hat{r}_A\hat{r}_B(1-\hat{r}_A)(1-\hat{r}_B))^{-1/2}$ so that its value ranges from -1 to 1. This formula can also be proven using the general definition of statistical correlation (see also Hartl and Clark, 1989, pp 53–54).

Three-loci relatedness

Compared with the two-loci analysis, the advantage of a three-loci analysis is that it considers the event of double crossovers during meiosis and hence can give more precise estimates.

Haploid data: The preceding two-loci method can be extended to the three-loci case. Suppose that an additional marker C with n_C (≥ 2) alleles is linked to the A and B markers. The ordering of the three loci is unknown. Denote by κ_{111} ($0\leq\kappa_{111}\leq 1$) the joint probability that the two alleles at each of the three loci are IBD, κ_{110} ($0\leq\kappa_{110}\leq 1$) the joint probability that the two alleles at both A and B loci are IBD but the two alleles at the C locus are not IBD. The definitions of other parameters κ_{101}–κ_{000} can be given in a similar way. The joint probability for a pair of three-loci gametes ($A_uB_vC_w$–$A_{u'}B_{v'}C_{w'}$; w, $\acute{w}=1, \ldots, n_C$), denoted by $P_{u'v'w'}^{uvw}$, can be written in a general formula,

$$\begin{aligned}P_{u'v'w'}^{uvw} = {} & g_{uvw}\delta_{uu'}\delta_{vv'}\delta_{ww'}\kappa_{111}\\ & + g_{uvw}g_{u'v'w'}\left(\delta_{uu'}\delta_{vv'}(2-\delta_{ww'})\frac{\kappa_{110}}{p_uq_v}\right.\\ & + \delta_{uu'}(2-\delta_{vv'})\delta_{ww'}\frac{\kappa_{101}}{p_uo_w}\\ & + (2-\delta_{uu'})\delta_{vv'}\delta_{ww'}\frac{\kappa_{011}}{q_vo_w}\\ & + \frac{\delta_{uu'}(2-\delta_{vv'})(2-\delta_{ww'})}{2^{(1-\delta_{vv'})(1-\delta_{ww'})}}\frac{\kappa_{100}}{p_u}\\ & + \frac{(2-\delta_{uu'})\delta_{vv'}(2-\delta_{ww'})}{2^{(1-\delta_{uu'})(1-\delta_{ww'})}}\frac{\kappa_{010}}{q_v}\\ & + \frac{(2-\delta_{uu'})(2-\delta_{vv'})\delta_{ww'}}{2^{(1-\delta_{uu'})(1-\delta_{vv'})}}\frac{\kappa_{001}}{o_w}\\ & \left. + \frac{(2-\delta_{uu'})(2-\delta_{vv'})(2-\delta_{ww'})}{2^{2-(\delta_{uu'}+\delta_{vv'}+\delta_{ww'})+\delta_{uu'}\delta_{vv'}\delta_{ww'}}}\kappa_{000}\right)\end{aligned} \tag{5}$$

where o_w is the frequency of allele C_w in the population, and g_{uvw} and $g_{u'v'w'}$ are the frequencies of gametes $A_uB_vC_w$ and $A_{u'}B_{v'}C_{w'}$, respectively. For an arbitrary number of alleles at each locus, the number of distinguishable three-loci gamete pairs, denoted by n_{ABC}, is shown to be equal to $n_An_Bn_C(n_An_Bn_C+1)/2$.

Diploid data: As in the two-loci case, denote by H_{AC} the probability that both A and C loci are heterozygous, H_{BC} the probability that both B and C loci are heterozygous, H_{ABC} the probability that the three loci are heterozygous, and H_{O2} the probability that three loci are homozygous. With the assumption of random association of gametes, I obtain

$$H_{O2} = \sum_u \sum_v \sum_w \left(g_{uvw}\kappa_{111} + g_{uvw}^2(\kappa_{000} + \kappa_{110}/p_u q_v \right. \\ + \kappa_{101}/p_u o_w + \kappa_{011}/q_v o_w \\ \left. + \kappa_{100}/p_u + \kappa_{010}/q_v + \kappa_{001}/o_w)\right) \tag{6a}$$

$$H_A = 2\sum_u \sum_{u' \neq u} \sum_v \sum_w g_{uvw} g_{u'vw}(\kappa_{000} + \kappa_{011}/q_v o_w \\ + \kappa_{010}/q_v + \kappa_{001}/o_w) \tag{6b}$$

$$H_B = 2\sum_v \sum_{v' \neq v} \sum_u \sum_w g_{uvw} g_{uv'w}(\kappa_{000} + \kappa_{101}/p_u o_w \\ + \kappa_{100}/p_u + \kappa_{001}/o_w) \tag{6c}$$

$$H_C = 2\sum_w \sum_{w' \neq w} \sum_u \sum_v g_{uvw} g_{uvw'}(\kappa_{000} + \kappa_{110}/p_u q_v \\ + \kappa_{100}/p_u + \kappa_{010}/q_v) \tag{6d}$$

$$H_{AB} = 2\sum_u \sum_{u' \neq u} \sum_v \sum_{v' \neq v} \sum_w g_{uvw} g_{u'v'w}(\kappa_{000} \\ + \kappa_{001}/o_w) \tag{6e}$$

$$H_{AC} = 2\sum_u \sum_{u' \neq u} \sum_w \sum_{w' \neq w} \sum_v g_{uvw} g_{u'vw'}(\kappa_{000} \\ + \kappa_{010}/q_v) \tag{6f}$$

$$H_{BC} = 2\sum_v \sum_{v' \neq v} \sum_w \sum_{w' \neq w} \sum_u g_{uvw} g_{uv'w'}(\kappa_{000} \\ + \kappa_{100}/p_u) \tag{6g}$$

$$H_{ABC} = 2\sum_u \sum_{u' \neq u} \sum_v \sum_{v' \neq v} \sum_w \sum_{w' \neq w} g_{uvw} g_{u'v'w'} \kappa_{000} \tag{6h}$$

The seven unknown parameters (κ_{111}–κ_{001}) can be solved using equations (6a)–(6h), provided that the frequencies of alleles and three-loci gametes are available with sufficient accuracy.

Correlation of pairwise relatedness: Denote by c_{r1}, c_{r2}, and c_{r3} the covariances of pairwise relatedness between the A and B loci, the B and C loci, and the A and C loci, respectively. Let $e_{\theta_A\theta_B\theta_C}$ (θ_A, θ_B, $\theta_C = 0$, 1) be the residual part of $\kappa_{\theta_A\theta_B\theta_C}(\sum_{\theta_A}\sum_{\theta_B}\sum_{\theta_C}\kappa_{\theta_A\theta_B\theta_C} = 1)$ after the deduction of individual components of the covariances of two-loci relatedness. The residual part comes from the effects of double crossover among the three loci. Thus, $\kappa_{\theta_A\theta_B\theta_C}$ can be written in a general form,

$$\kappa_{\theta_A\theta_B\theta_C} = r_A^{\theta_A} r_B^{\theta_B} r_C^{\theta_C}(1-r_A)^{1-\theta_A}(1-r_B)^{1-\theta_B}(1-r_C)^{1-\theta_C} \\ + (-1)^{\theta_A+\theta_B}c_{r1} + (-1)^{\theta_B+\theta_C}c_{r2} + (-1)^{\theta_A+\theta_C}c_{r3} \\ + e_{\theta_A\theta_B\theta_C} \tag{7}$$

There are eight configurations of the sequence of $\theta_A\theta_B\theta_C$, that is, (111), (110), (101), (011), (100), (010), (001), and (000). The probabilities of relatedness at individual loci can be estimated by $\hat{r}_A = \sum_{\theta_B}\sum_{\theta_C}\hat{\kappa}_{1\theta_B\theta_C}$, $\hat{r}_B = \sum_{\theta_A}\sum_{\theta_C}\hat{\kappa}_{\theta_A 1\theta_C}$, and $\hat{r}_C = \sum_{\theta_A}\sum_{\theta_B}\hat{\kappa}_{\theta_A\theta_B 1}$. According to equation (7), the coefficients for the three covariances ($(-1)^{\theta_A+\theta_B}$, $(-1)^{\theta_B+\theta_C}$, and $(-1)^{\theta_A+\theta_C}$) are the same between the partitions of κ_{111} and κ_{000}, κ_{110} and κ_{001}, κ_{101} and κ_{010}, and κ_{011} and κ_{100}. Thus, I can only use the partitions of four three-loci relatedness values (κ_{111}, κ_{110}, $\kappa_{1\ 01}$, and κ_{011}) to estimate c_{r1}, c_{r2}, and c_{r3}. The analytic solution from the least-square method is given by

$$\hat{c}_{r1} = \tfrac{1}{4}(\hat{\kappa}_{111} + \hat{\kappa}_{110} - \hat{\kappa}_{101} - \hat{\kappa}_{011} - \hat{r}_A\hat{r}_B \\ + \hat{r}_C(\hat{r}_A + \hat{r}_B - \hat{r}_A\hat{r}_B)) \tag{8a}$$

$$\hat{c}_{r2} = \tfrac{1}{4}(\hat{\kappa}_{111} - \hat{\kappa}_{110} - \hat{\kappa}_{101} + \hat{\kappa}_{011} - \hat{r}_B\hat{r}_C \\ + \hat{r}_A(\hat{r}_B + \hat{r}_C - \hat{r}_B\hat{r}_C)) \tag{8b}$$

$$\hat{c}_{r3} = \tfrac{1}{4}(\hat{\kappa}_{111} - \hat{\kappa}_{110} + \hat{\kappa}_{101} - \hat{\kappa}_{011} - \hat{r}_A\hat{r}_C \\ + \hat{r}_B(\hat{r}_A + \hat{r}_C - \hat{r}_A\hat{r}_C)) \tag{8c}$$

Denote by R_1, R_2, and R_3 the correlation coefficients of relatedness between the A and B loci, the B and C loci, and the A and C loci, respectively. Estimates of these three correlation coefficients are respectively given by

$$\hat{R}_1 = \hat{c}_{r1}(\hat{r}_A\hat{r}_B(1-\hat{r}_A)(1-\hat{r}_B))^{-1/2} \tag{9a}$$

$$\hat{R}_2 = \hat{c}_{r2}(\hat{r}_B\hat{r}_C(1-\hat{r}_B)(1-\hat{r}_C))^{-1/2} \tag{9b}$$

$$\hat{R}_3 = \hat{c}_{r3}(\hat{r}_A\hat{r}_C(1-\hat{r}_A)(1-\hat{r}_C))^{-1/2} \tag{9c}$$

Monte Carlo simulation

Like previous studies of the pairwise relatedness at a single locus (eg Ritland, 1996; Lynch and Ritland, 1999; Wang, 2002), the aims of the simulations are to examine the effects of (i) sample size, (ii) allele frequency distribution, (iii) the type of data sets (haploid or diploid), and (iv) LD.

Statistical methods: The weighted LS and ML methods are used to estimate the correlation coefficient of pairwise relatedness. With the two-loci haploid genotypic data, estimates of pairwise relatedness with the weighted LS method can be written as

$$\begin{pmatrix} \hat{\bar{y}} \\ \hat{\boldsymbol{\kappa}} \end{pmatrix} = \left(\begin{pmatrix} \mathbf{1}' \\ \mathbf{X}' \end{pmatrix}\mathbf{W}(\mathbf{1X})\right)^{-1}\begin{pmatrix} \mathbf{1}' \\ \mathbf{X}' \end{pmatrix}\mathbf{WY} \tag{10}$$

where $\hat{\boldsymbol{\kappa}}$ is the vector of $(\hat{\kappa}_{11}, \hat{\kappa}_{10}, \hat{\kappa}_{01})'$, $\mathbf{1}$ is the vector of $(1, 1, ..., 1)'_{n_{AB}\times 1}$, $\mathbf{X}$ is the known coefficient matrix with $n_{AB} \times 3$ elements calculated from equation (2), $\mathbf{Y}$ is the known vector $(y_{u'v'}^{uv})_{n_{AB}\times 1}$ in which

$$y_{u'v'}^{uv} = P_{u'v'}^{uv} - \frac{(2-\delta_{uu'})(2-\delta_{vv'})}{2^{(1-\delta_{uu'})(1-\delta_{vv'})}} g_{uv} g_{u'v'}$$

$\mathbf{W}$ is the known diagonal matrix with the diagonal element being $w_{u'v'}^{uv} = 1/P_{u'v'}^{uv}(1-P_{u'v'}^{uv})$, and $\hat{\bar{y}}$ is the estimate of the mean of $y_{u'v'}^{uv}$.

For the ML method, the likelihood function is set as $L \propto \prod_{uv,u'v'} (P^{uv}_{u'v'})^{M^{uv}_{u'v'}}$, where $M^{uv}_{u'v'}$ is the observed number of gamete pairs (A_uB_v–$A_{u'}B_{v'}$) in the random sample with N haploids ($\sum M^{uv}_{u'v'} = N(N-1)$). ML estimates are obtained through Newton–Raphson iteration. The estimate of $\hat{\boldsymbol{\kappa}}$ at the $(t+1)$ step is iteratively calculated by

$$\hat{\boldsymbol{\kappa}}^{t+1} = \hat{\boldsymbol{\kappa}}^{t} + \mathbf{F}^{-1}(\hat{\boldsymbol{\kappa}}^{t})\mathbf{s}(\hat{\boldsymbol{\kappa}}^{t}) \tag{11}$$

where $\mathbf{s}(\boldsymbol{\kappa})$ is the score vector, equal to $(\partial \ln L/\partial\kappa_{11}, \partial \ln L/\partial\kappa_{01}, \partial \ln L/\partial\kappa_{01})'$, and $\mathbf{F}(\boldsymbol{\kappa}) = -E(\mathbf{s}(\boldsymbol{\kappa})\mathbf{s}(\boldsymbol{\kappa})')$ is the Fisher information matrix. The above iterative calculation is continued until $|\hat{\boldsymbol{\kappa}}^{t+1} - \hat{\boldsymbol{\kappa}}^{t}|$ is sufficiently small (convergence).

When the estimates of pairwise relatedness are obtained, the correlation coefficients of pairwise relatedness are calculated according to $\hat{c}_r(\hat{r}_A\hat{r}_B(1-\hat{r}_A)(1-\hat{r}_B))^{-1/2}$ for the two-loci case and equations (9a)–(9c) for the three-loci case. The correlation coefficients of pairwise relatedness in the other cases (two-loci diploid, three-loci haploid and diploid) can be estimated in a way similar to the above two approaches.

Data generation: The simulated samples with the haploid or diploid data are generated in the following steps. Given a set of parameters, including LD, the number of alleles, allele frequencies and the distribution type, and pairwise relatedness (κ_{11}–κ_{00} in the two-loci case and κ_{111}–κ_{000} in the three-loci case) calculate the probabilities for each two-gamete pair according to equation (2) for the haploid case, and the probabilities of each type of heterozygote according to equations (3a)–(3d) and (6a)–(6h) for the diploid case. Then, use this probability distribution (multinomial distribution) to create random samples. It can be shown that a sample of $N/2$ diploids (or N haploids) can generate a total of $N(N-1)$ gamete pairs for either the two- or the three-loci case. Simulation programs in C are available upon request.

In all, 5000 independent data sets are created, and each is used for estimating the correlation coefficients of pairwise relatedness according to the theories described in the preceding two sections. Means and standard deviations of estimates are calculated from these replicated data sets.

Results

With the two-loci haploid data and the weighted LS method, average estimates of the correlation coefficients of pairwise relatedness gradually become consistent with their actual values as the sample size increases (Figure 1a). The standard deviations decrease with the number of haploids (Figure 1b). However, there are large differences in the sample sizes required for obtaining appropriate estimates: 50 haploids for the two-allele case, 120 haploids for the four-allele case, and more than 180 haploids for the eight-allele case (Figure 1a).

A difference between the weighted LS and ML methods is that the ML method can give better estimates when the sample size is small. For example, when the number of haploids is greater than 40 for the four- or eight-allele case, an appropriate estimate of the correlation coefficient of pairwise relatedness can be obtained (Figure 2a, b).

The effects of the distribution of allele frequency (uniform *versus* triangular distribution; see Lynch and Ritland, 1999) on the LS or ML methods are very small. Although the triangular distribution produces a slightly greater standard deviation than does the uniform distribution (Figure 3), there are no differences in obtaining the unbiased average of the correlation coefficients of pairwise relatedness.

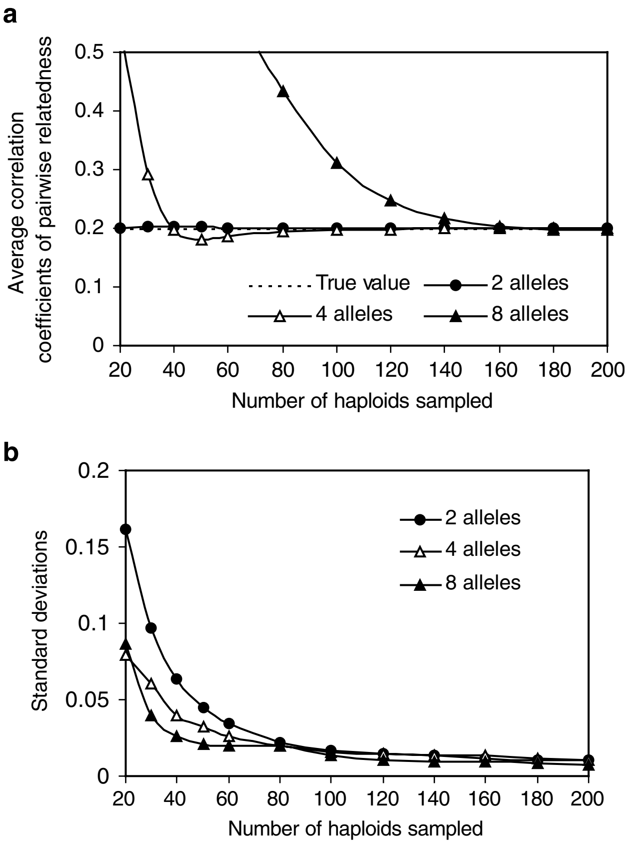

Figure 1 Effects of sample size: (**a**) average correlation coefficients of pairwise relatedness; (**b**) standard deviations. The results are obtained from 5000 independent runs under the uniform distribution of allele frequencies, with the two-loci haploid data, the weighted LS method, and linkage equilibrium. Two-loci relatedness values are set as $\kappa_{11} = 0.3$, $\kappa_{10} = 0.2$, and $\kappa_{01} = 0.2$.

Unbiased estimates of the correlation coefficient of pairwise relatedness can be obtained when LD is present and the sample size is appropriate. For example, there are no significant differences when LD is changed from 0 to 0.2 in the two-allele case (Figure 4a, b). Also, no significant differences are observed between the weighted LS and ML methods.

With the two-loci diploid data, an unbiased estimate of the correlation coefficient of pairwise relatedness can be obtained with each of the two methods when an appropriate sample size is provided. Compared with the haploid case, the optimal sample size is smaller. For example, a good estimate can be obtained with sampling 40 diploids in the four-allele case (Figure 5). Both the weighted LS and ML methods have the same performance (Figure 5).

With the three-loci diploid data, the unbiased average estimates of the three correlation coefficients of pairwise relatedness can be simultaneously obtained when the sample size is appropriate. For example, when the sample size is more than 80 individuals, the three

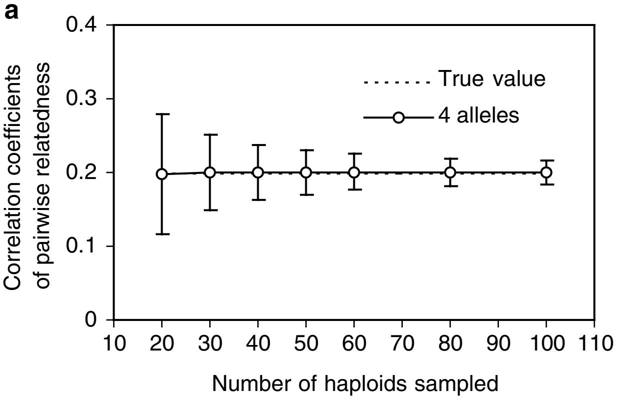

Figure 2 Effects of sample size: (**a**) four-allele case; (**b**) eight-allele case. The results are obtained from 5000 independent runs under the uniform distribution of allele frequencies, with the two-loci haploid data, the ML method, and linkage equilibrium. Two-loci relatedness values are set at $\kappa_{11} = 0.3$, $\kappa_{10} = 0.2$, and $\kappa_{01} = 0.2$.

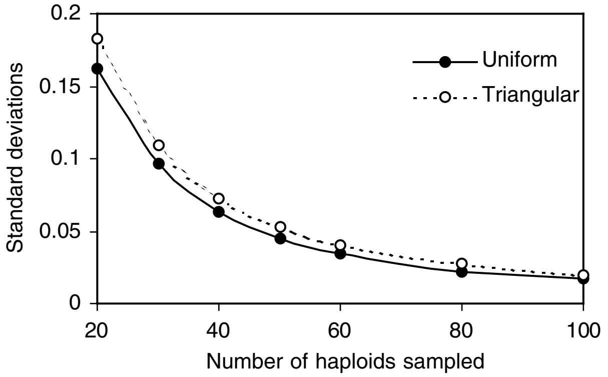

Figure 3 Effects of the uniform *versus* triangular distributions of allele frequencies. The results are obtained from 5000 independent runs with the two-loci haploid data (two alleles per locus) and linkage equilibrium. Two-loci relatedness values are set at $\kappa_{11} = 0.3$, $\kappa_{10} = 0.2$, and $\kappa_{01} = 0.2$.

unknown parameters (R_1, R_2, and R_3) can be estimated with a good accuracy and precision in the case of four alleles per locus (Figure 6a–f). Both the weighted LS and the ML methods have a very similar performance.

Discussion

In this paper, I have shown that the correlation of pairwise relatedness among loci along chromosomes can be estimated from the approach of partitioning the joint probabilities of IIS into the probabilities of IBD at two or three loci. Such an approach of using two- or three-loci probabilities of IBD enables the estimation of individual pairwise relatedness and its correlation among loci simultaneously, and allows us to study the picture of 'landscape' relatedness along chromosomes and to infer the naturally occurring pattern of co-evolution. Unlike traditional kinship mapping, which pictures the 'static' relationships between the physical position of markers and IBD (eg Morton and Simpson, 1983), the map of the correlation of relatedness implies 'dynamic' relationships among linked loci.

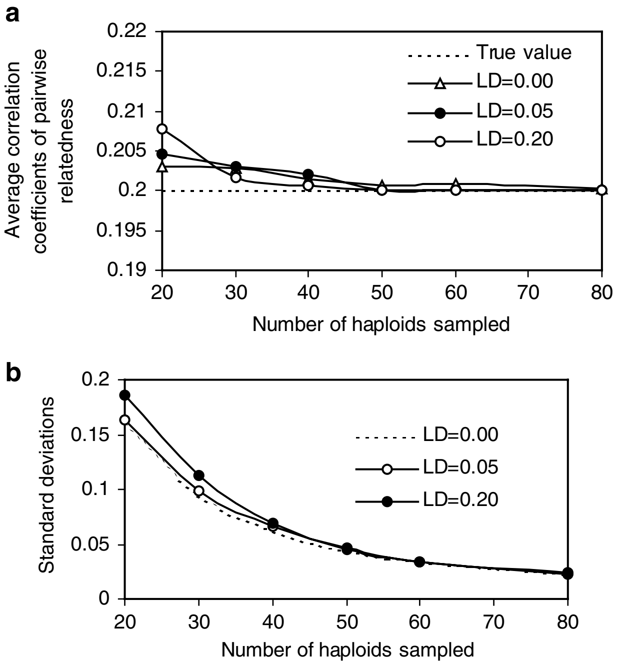

Figure 4 Effects of LD: (**a**) average correlation coefficients; (**b**) standard deviations. The results are obtained from 5000 independent runs under the uniform distribution of allele frequencies, with the two-loci haploid data (two alleles per locus), and the weighted LS method. Two-loci relatedness values are set at $\kappa_{11} = 0.3$, $\kappa_{10} = 0.2$, and $\kappa_{01} = 0.2$.

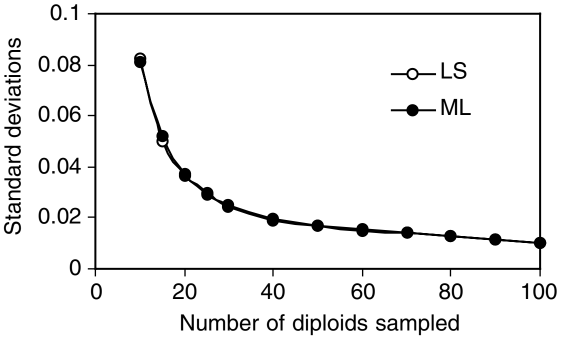

Figure 5 Comparison of the weighted LS *versus* the ML methods in terms of the standard deviation of estimates. The results are obtained from 5000 independent runs under the uniform distribution of allele frequencies, with the two-loci diploid data (four alleles per locus), and linkage equilibrium. Two-loci relatedness values are set as $\kappa_{11} = 0.3$, $\kappa_{10} = 0.2$, and $\kappa_{01} = 0.2$.

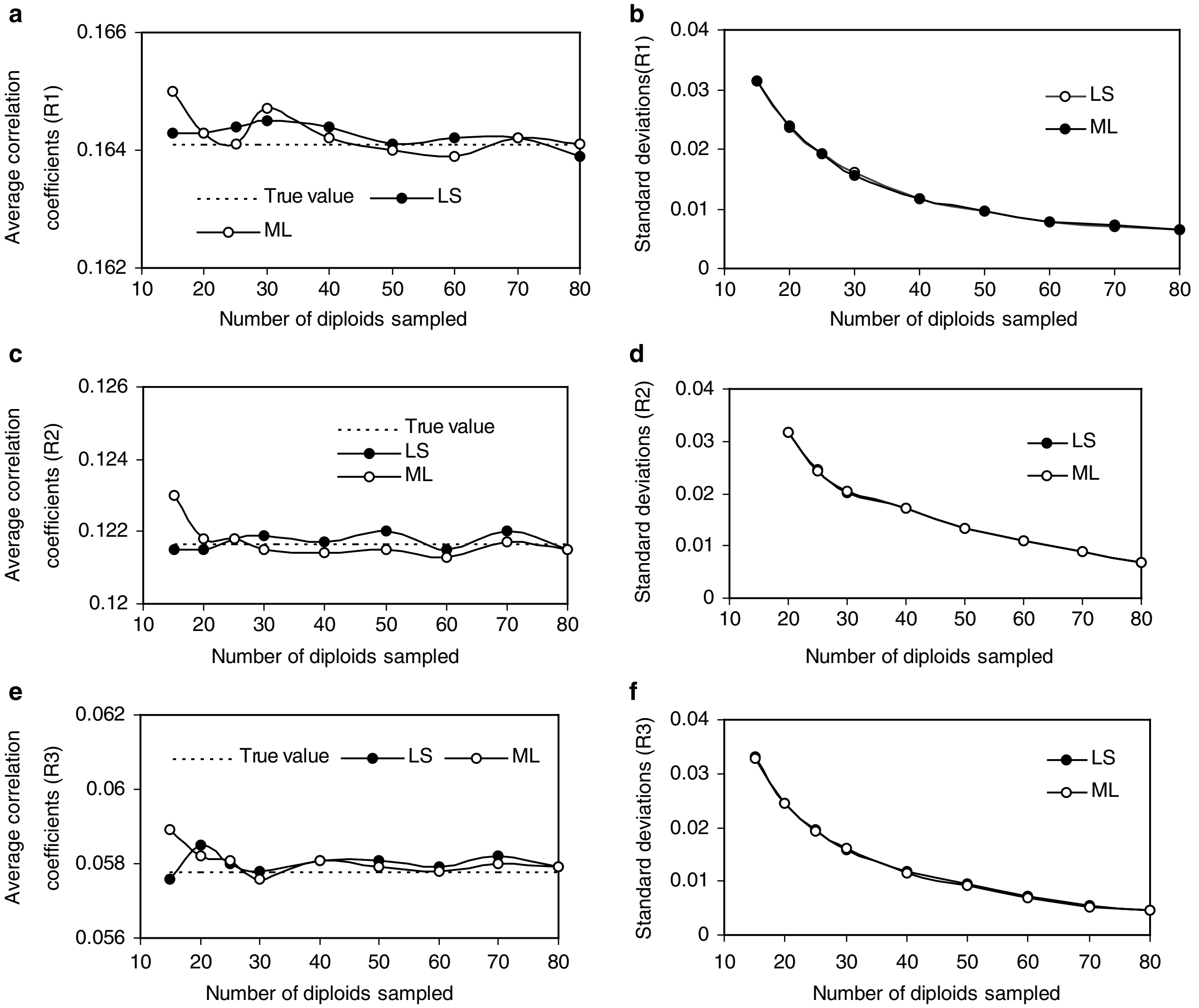

Figure 6 Effects of sample size on three-loci relatedness analysis: (**a**) average correlation coefficients between *A* and *B* loci (R_1); (**b**) standard deviations (R_1); (**c**) average correlation coefficients between *B* and *C* (R_2); (**d**) standard deviations (R_2); (**e**) average correlation coefficients between *A* and *C* loci (R_3); and (**f**) standard deviations (R_3). The results are obtained from 5000 independent runs under the uniform distribution of allele frequencies, with the three-loci diploid data (four alleles per locus), and linkage equilibrium. Three-loci relatedness values are set as $\kappa_{111}=0.15$, $\kappa_{110}=0.1$, $\kappa_{101}=0.05$, $\kappa_{011}=0.08$, $\kappa_{100}=0.08$, $\kappa_{010}=0.05$, and $\kappa_{001}=0.1$.

When the linkage maps of all the markers assayed are available beforehand, the relatedness at each locus and the correlation of relatedness among linked loci can be readily mapped. Since the present method involves only one-generation data randomly sampled from the population with unknown pedigree, the recombination fraction between linked loci cannot be estimated. The physical linkage map of relatedness cannot be directly constructed. However, the following properties are likely applicable to constructing the map of pairwise relatedness and its correlation among loci. First, a significant difference of κ_{11} from zero indicates that the two loci are likely to be linked. Second, the joint probability of two-loci IBD (κ_{11}) is negatively correlated with the mapping distance between the two loci, while the joint probabilities of one-locus IBD and one-locus non-IBD (κ_{10} and κ_{01}) are positively correlated with mapping distance. Thus, a larger κ_{11} and a smaller κ_{10} or κ_{01} indicate a shorter distance between the two loci. Third, a larger positive correlation of relatedness indicates a shorter physical distance, while a larger negative correlation of relatedness indicates a larger physical distance. These properties can be combined for ordering markers. Among all possible linkage maps for a given set of markers, the optimal one should have the largest sum of all κ_{11}'s among adjacent markers.

The analytical formulae presented here are only suitable for the case of two- and three-loci relatedness, where each locus has an arbitrary number of alleles. When many loci on a chromosome (more than three loci) are analyzed, these loci can be analyzed in terms of two- or three-loci as a unit, similar to the procedure of classical linkage mapping analysis. The individual two- or three-loci results are then jointly analyzed to map the correlation of pairwise relatedness.

Statistically, the critical problem for the weighted LS method with haploid data is that the number of distinguishable pairs is substantially increased with the number of alleles. Only when the number of sampled haplotype pairs is much greater than the number of distinguishable haplotype pairs can an appropriate estimate be obtained. Thus, the optimal sample size varies with the number of alleles. The present simulations suggest sample sizes of 100–200 haploids for moderate numbers of alleles, the lower bound for the two-allele case, the upper bound for the eight-allele case. However, such situations are not often met in the single locus case (Ritland, 1996; Lynch and Ritland, 1999; Wang, 2002), where the sample size is much larger than the expected number of distinguishable two-gene pairs.

Compared with the haploid case, the weighted LS method with the diploid data has a better performance, and the optimal sample size is also smaller (eg 40 diploids for two loci, each with four to eight alleles). The reason for the better performance of the weighted LS method with the diploid than with the haploid data is a small number of 'units of observations' – the heterozygous types, that is, four in the two-loci case and eight in the three-loci case. These 'condensed' variables contain all possible sampled haplotype pairs, and display a 'robust' property even when the sample size is small. Clearly, the ML method for the haploid data is suggested when highly polymorphic markers (eg, ≥ 4 alleles per locus) and a small sample size are used. However, either the weighted LS or ML method can be applied with diploid data.

The advantage of the ML over the weighted LS method with the haploid data is that all information is utilized, including different sampling variances for individual pairs and the correlation between different pairs. This can be seen from the Fisher information matrix ($\mathbf{F}(\boldsymbol{\kappa})$). One of the assumptions underlying the weighted LS method is the independence among different observations of distinguishable gamete-pairs, which is actually violated.

Another striking result is the same performance for the ML method with either the haploid or diploid data. The condensed H variable does not affect the accuracy and precision of estimation, compared with the analysis with the haploid data. The reason for such robust behavior is that the score function ($\mathbf{s}(\boldsymbol{\kappa})$) and the Fisher information matrix ($\mathbf{F}(\boldsymbol{\kappa})$) are essentially the same with either approach under the assumption of random association of gametes, and this can be shown algebraically. However, the advantage of using diploid over haploid data is significant in practice.

There are two distinctions between the present two-loci haplotype approach and the previous 'four-gene' pairs at a single locus in pairwise relatedness analysis. There are three parameters in the former (κ_{11}, κ_{10}, and κ_{01}), but only two parameters in the latter (eg Ritland, 1996; Lynch and Ritland, 1999). There are 10 informative pairs in the former for a four-gene case (two alleles per locus), but six in the latter (four alleles per locus). Such distinctions imply that a larger sample size is required in the two-loci analysis, compared with the case of a single locus.

Finally, one must acknowledge the assumptions underlying the present method of estimating the correlation of pairwise relatedness. First, the method is based on the assumption of the availability of accurate and precise estimates of gamete and allele frequencies (eg Ritland, 1996; Lynch and Ritland, 1999; Ritland, 2000). In practice, gamete and allele frequencies will probably be estimated from the same data sets used for relatedness analysis. Clearly, biased estimates of these frequencies can bring about biased estimates of the correlation coefficients of pairwise relatedness. Second, the present diploid approach is based on the assumption of random association of gametes. Extension to partially selfing populations is clearly needed in future study, as selfing populations show greater linkage disequilibria (Wright, 1969) and likely enhance the correlation of pairwise relatedness along chromosomes.

Acknowledgements

I sincerely appreciate John Brookfield and three anonymous referees for very constructive comments and editorial assistance, and Kermit Ritland for useful comments and discussions during the preparation of this article.

References

Brookes AJ (1999). The essence of SNPs. *Gene* **234**: 177–186.

Charlesworth B, Morgan MT, Charlesworth D (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.

Cotterman CW (1940). *A calculus for statistico-genetics*. Unpublished thesis, Ohio State University, Columbus, OH.

Dempster AP, Laird NM, Bubin DB (1977). Maximum likelihood from incomplete data via EM algorithm. *J R Stat Soc Ser B* **39**: 1–38.

Falconer DS, Mackay TFC (1996). *Introduction to Quantitative Genetics*, 3rd edn. Longman Sci and Tech: Harlow, UK.

Hartl DL, Clark AG (1989). *Principles of Population Genetics*, 2nd edn. Sinauer Associates Inc.: Sunderland.

Hill WG (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33**: 229–239.

Hudson RR (1991). Gene genealogies and the coalescent process. In: Futuyma D, Antonovics J (eds) *Oxford Surveys in Evolutionary Biology*, Vol. 7. Oxford Unversity Press: Oxford, pp 1–44.

Jacquard A (1974). *The Genetic Structure of Populations*. In: Krickeberg K, Lewontin RC, Neyman J, Schreiber M (eds) *Biomathematics*. Springer-Verlag: Berlin. Vol. 5.

Kalinowski ST, Hedrick PW (2001). Estimation of linkage disequilibrium for loci with multiple alleles: basic approach and application using data from bighorn sheep. *Heredity* **87**: 698–708.

Laurie C, Weir BS (2003). Dependency effects in multi-locus match probabilities. *Theor Pop Biol* **63**: 207–219.

Lynch M (1988). Estimation of relatedness by DNA fingerprinting. *Mol Biol Evol* **5**: 584–599.

Lynch M, Ritland K (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152**: 1753–1766.

Maynard Smith J, Haigh J (1974). The hitch-hiking effect of a favorable gene. *Genet Res* **23**: 23–35.

Milligan BG (2003). Maximum-likelihood estimation of relatedness. *Genetics* **163**: 1153–1167.

Morton NE, Simpson SP (1983). Kinship mapping of multilocus systems. *Hum Genet* **64**: 103–104.

Morton NE, Yee S, Harris DE, Lew R (1971). Bioassay of kinship. *Theor Pop Biol* **2**: 507–524.

Pamilo P, Crozier RH (1982). Measuring genetic relatedness in natural populations: methodology. *Theor Pop Biol* **21**: 171–193.

Queller DC, Goodnight KF (1989). Estimating relatedness using genetic markers. *Evolution* **43**: 258–275.

Ritland K (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet Res* **67**: 175–185.

Ritland K (2000). Marker-inferred relatedness as a tool for detecting heritability in nature. *Mol Ecol* **9**: 1195–1204.

Rousset F (2002). Inbreeding and relatedness coefficients: what do they measure? *Heredity* **88**: 371–380.

Thompson EA (1975). The estimation of pairwise relationships. *Ann Hum Genet* **39**: 173–188.

Vitalis R, Couvet D (2001a). Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population. *Genet Res* **77**: 67–81.

Vitalis R, Couvet D (2001b). Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* **157**: 911–925.

Wang J (2002). An estimator for pairwise relatedness using molecular markers. *Genetics* **160**: 1203–1215.

Weir BS (1996). *Genetic Data Analysis II*. Sinauer Associates Inc.: Sunderland.

Whitlock MC, Phillips PC, Wade MJ (1993). Gene interaction affects the additive genetic variance in subdivided populations with migration and extinction. *Evolution* **47**: 1758–1769.

Wright S (1922). Coefficients of inbreeding and relationship. *Am Nat* **56**: 330–338.

Wright S (1969). *Evolution and the Genetics of Populations, Vol. 2. The Theory of Gene Frequencies*. University of Chicago Press: Chicago.

Yang RC (2002). Analysis of multilocus zygotic associations. *Genetics* **161**: 435–445.