Relatedness coefficients and their applications for triplets and quartets of markers

Kermit Ritland

Biodiversity Research Center

University of British Columbia

RUNNING HEAD: Relatedness with marker triplets and quartets

ABSTRACT

Relatedness coefficients which seek the identity-by-descent of genetic markers are described. The markers are in groups of two, three or four, and if four, can consist of two pairs. It is essential to use cumulants (not moments) for four-gene probabilities, as the covariance of homozygosity, used in 4-marker applications, can only be described with cumulants. A covariance of homozygosity between pairs of markers arises when populations follow a mixture distribution. Also, the probability of four markers all identical-by-descent equals the normalized fourth cumulant. In this paper, a "genetic marker" generally represents either a gene locus or an allele at a locus.

Applications of three marker coefficients mainly involve conditional regression, and applications of four marker coefficients can involve identity disequilibrium. Estimation of relatedness using genetic marker data is discussed, with regards to both moment estimators and likelihood estimators. However, three- and four-marker estimators suffer from statistical and numerical problems, including higher statistical variance, complexity of estimation formula, and singularity at some intermediate allele frequencies.

INTRODUCTION

Relatedness is proportional to the fraction of alleles that individuals share due to common ancestry (MALÉCOT 1948; PAMILO AND CROZIER 1982). To quantify the degree of relatedness, gene identity coefficients are used, and these coefficients give the fraction of identity-by-descent due to common ancestry. The unit of observation is normally a pair of alleles and the coefficient of relationship is based on similarity within this pair. In this paper, the unit of observation is extended to triplets and quartets of alleles. This allows the characterization of "higher order" parameters of population structure.

Relatedness may be estimated with genetic markers (MORTON *et al.* 1971) and for pairs of marker genes, many computer programs are available for estimation of relatedness (WANG 2014), in particular for "pairwise relationship", such as the *r* of QUELLER AND GOODNIGHT (1989). However, the equations for pairwise relationship are not extendable to three or four alleles, as the covariances and higher moments need to be defined in new ways. Estimators for three and four allele measures of relatedness have recently been proposed. SAMANTA *et al.* (2009) provided the first estimator for three markers. ACKERMAN *et al.* (2017) described the estimation of seven of eight coefficients of relatedness. Multiallelic data have information about all eight coefficients, but they used a biallelic model which provides just seven degrees of freedom, constraining their space of estimates. In addition, they did not use cumulants and their moments are normalized differently than will be here.

Cumulants are of use in certain problems in quantitative genetics (BURGER 1991; TURELLI AND BARTON 1994). Compared to moments, cumulants have more useful theoretical properties (KENDALL *et al.* 1977). RITLAND (1987) found cumulants instead of moments

were an essential component of four-allele fixation indices. Fourth order cumulants are needed to specify the probability of gene identity for all four alleles, and to describe identity disequilibrium as the "covariance of covariances". As an example of the necessity of cumulants, for population allele frequency p, the fourth central moment for four alleles, denoted (X_1, X_2, X_3 and X_4), is $E[(X_1 - p)(X_2 - p)(X_3 - p)(X_4 - p)] = \sigma_{12}^2\sigma_{34}^2 + \sigma_{13}^2\sigma_{24}^2 + \sigma_{23}^2\sigma_{13}^2 + \kappa_4$ where σ_{ij}^2 is the covariance of X_i and X_j and κ_4 is a fourth-order cumulant which does not appear in moment-based treatments.

The probabilities developed in the paper are used for population genomic data These models and estimation procedures are readily applicable to the emerging mountains of genome data.

APPLICATIONS OF MARKERS

Definitions. Between pairs of genetic markers, the *coefficient of relationship* measures the degree of consanguinity (e.g., the probability that markers are identical by descent, termed ibd). The coefficient relationship equals twice the *kinship coefficient*. The *inbreeding coefficient* is the probability that a pair of markers within one individual are ibd. With more than two markers, the coefficient of relationship is more broadly defined with groups of markers (two, three or four). With four markers, there are nine *modes of Jacquard's gene identity,* with ibd genes connected by lines. The *normalized central moment* gives the probability of ibd of all markers. At the level of four markers, *cumulants* are necessary to describe *identity disequilibrium* (the excess of identity between marker pairs). The covariance of cumulants forms the machinery of higher order interactions.

Relatedness and two markers. The two-marker coefficient of relatedness is used for many inferences with genetic markers, mainly involving pairs of genes sampled between two individuals ("coefficient of relationship" or pairs of genes sampled within one individual (the "inbreeding coefficient"). Analysis of data with two marker measures are ubiquitous (WANG 2014) and the two-marker probabilities are often incorporated into probabilities of groups of three and four genes.

Regression and three markers. The three-marker relationship coefficient is the probability that the three marker loci have alleles all identical-by-descent (Figure 1a). This coefficient, *G*, is usually combined with two-allele coefficients for biological interpretable parameters, useful at least for problems involving mating systems or kin selection. In the theory of mating system estimation, the "effective selfing rate" is the genetically equivalent rate of selfing caused by all types of biparental inbreeding (RITLAND 1985); the effective selfing rate of individual *A* equals 2*R*-*G*, where *R* is the relatedness between mates and G the third moment involving the two maternal and single paternal allele (Figure 1b). In the theory of kin selection, the regression coefficient of relatedness is used and, properly, a three-gene model is needed (Figure 1c), as shown by MICHOD AND HAMILTON (1980), where their Equation 18 depends upon whether the reference genotype is homozygous (18a) vs. heterozygous (18b). Note that both the effective selfing rate and the regression coefficients of relationship can be asymmetrical when inbreeding coefficients differ between the two relatives.

Identity disequilibrium and four markers. Between two diploid individuals, there are 15 patterns of gene identity (LIU 2005). A pair of individuals can share two, three or four markers, and at each level, allelic similarities can describe aspects of relatedness. After

JACQUARD (1966); JACQUARD (2012), for four genes and two individuals, there are nine condensed identity modes, denoted as Δ_i (Figure 2a). There are eight independent parameters of relatedness: three pairwise measures (F_A, F_B, R), two three-way measures (G_A, G_B), and three four-marker measures (F_{AB}, R_{AB}, H). At the highest level, the measures are much different as the four-marker measures F_{AB}, R_{AB} are covariances (not identities). The four marker parameter, H, is the probability that four markers are identical. As well, the quantities must be defined as cumulants. Cumulants equal moments up to order three, but fourth-order moments do not equal fourth order cumulants.

The fourth central moment equals $\kappa_4 + F_{AB} + 2R_{AB}$ (note the covariances between second moments enter this expression) and the normalized cumulant κ_4 (equivalent to *H)* equals the probability of identity of all four genes. While the variance is a measure of the spread of the distribution, kurtosis is a measure of the "peakedness" of the distribution of random variables, and infrequent extreme deviations contributing excessively to this statistic (DE LA ROSA AND MORENO MUÑOZ 2008).

While applications of three and four gene measures are in their infancy, at least, the skew and kurtosis as measured by higher moments can help remove bias in DNA forensics caused by genotyping error (WEIR 1994).

In the four-marker model, many possibilities exist about attaching meaning to each of the Δ_i. One example is the progeny-pair model (RITLAND AND LEBLANC 2004) where *A* and *B* are two progeny of the same mother plant (Figure 2b). At another more abstract level, two of the genes can be markers and two are quantitative trait loci (Figure 2c). If identity disequilibrium is present, the regression of phenotypic similarity (QTL) on estimated relationship (markers) gives an estimate of heritability "in the field" (RITLAND 2000).

TWO, THREE AND FOUR MARKER PROBABILITIES

At any level of gene comparison, associations are measured as the frequency of a given configuration (allele "state") divided by the denominators in Table 1. These denominators are termed "normalization constants", and are the maximum possible value of the numerator. Some of these normalized measures of association arise naturally in the derivations below.

Probabilities of two-marker relationship

From Equation (7) of RITLAND (1987), which follows KENDALL *et al.* (1977 eq. 13.36), the frequency of gametes with allele i and with allele j is

$$f_{ij} = \boldsymbol{\kappa}_i\boldsymbol{\kappa}_j + \boldsymbol{\kappa}_{ij}$$

The two-marker coefficient of relationship can be estimated from the frequencies of each allele in a sample. For any given allele, say A_i, it derived by equating the observed frequency of homozygotes to that expected by the above equation

$$f_{ii} = \mathrm{E}[A_iA_i] = p_i^2 + p_i(1-p_i)R$$

$$f_{ij} = \mathrm{E}\left[A_iA_j\right] = 2p_ip_j(1-p_i)R$$

The likelihood of the data, given R, $L(R) = \prod_{ij} f_{ij}^{\,X_{ij}}$. Solving for R gives estimators based upon pairs of alleles A,

$$\hat{R}_{ii} = \frac{f_{ii}-p_i^2}{p_i(1-p_i)}$$

The estimate of R for allele i, estimates are combined across alleles as

$$\hat{R} = \textstyle\sum_i w_i\hat{R}_{ii} \qquad (1)$$

where the weights w_i sum to unity.

These weights are found by finding the w_i that minimize $\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}$, where $\mathbf{w}$ is an n element vector of weights, and $\mathbf{V}$ the nxn variance-covariance matrix of allele-specific estimates (for details see RITLAND (1996)). The weights require prior specification of true relatedness. With zero prior R, the weight for allele A_i is $w_i = \frac{1-p_i}{n-1}$. An m-allele locus receives the weight (n_m-1), giving the estimator for r given by equation 5 in RITLAND (1996).

Probabilities of three-marker relationship

The three-marker relationship coefficient is the probability that three sampled marker genes are all identical-by-descent. From Equation (7) of RITLAND (1987), which follows KENDALL *et al.* (1977 eq. 13.36), The joint frequency of markers i, j and k is

$$f_{ijk} = \boldsymbol{\kappa}_i\boldsymbol{\kappa}_j\boldsymbol{\kappa}_k + \boldsymbol{\kappa}_i\boldsymbol{\kappa}_j + \boldsymbol{\kappa}_i\boldsymbol{\kappa}_k + \boldsymbol{\kappa}_j\boldsymbol{\kappa}_k + \boldsymbol{\kappa}_{ijk}$$

This written in conventional population genetic terms as

$$f_{ijk} = p_ip_jp_k + p_ip_jv_f + (p_kp_l + p_jp_k)v_r + w_{ijk}$$

Where alleles i and j are from one individual and allele k from a second individual. The cumulants are written in bold face to emphasize they have a random component that may covary.

From Equation (7) of RITLAND (1987), there are three primary patterns

$$f_{iii} = \mathrm{E}[A_iA_iA_i] = p_i^3(1 - F - 2R + 2G) + p_i^2(F + 2R - 3G) + p_iG$$

$$f_{iij} = \mathrm{E}[A_iA_iA_j] = p_i^2p_j(1 - F - 2R + 2G) + p_ip_j(F - G)$$

$$f_{ijk} = \mathrm{E}[A_iA_jA_k] = p_ip_jp_k(1 - F - 2R + 2G) \qquad (2)$$

160 Where the order is irrelevant ($A_iA_iA_j$, $A_jA_iA_j$ and $A_iA_jA_j$ are equivalent). The genotype
161 frequencies are mixtures of gene identity: *G* is the probability that all three markers are ibd,
162 *R*-*G* is the probability of ibd of one pair of markers, $F + 2R - 3G$ for two pairs, and and 1-F-
163 $2R$+$2G$ is the probability of no ibd among the three markers).

164 Solving for *G* in Equation (7) gives three probabilities involving *G*,

$$\hat{G}_{iii} = \frac{f_{iii} - p_i^2(1-p_i)(F+2R) - p_i^3}{p_i(1-p_i)(1-2p_i)}$$

$$\hat{G}_{iij} = \frac{f_{iij} - p_i(1-p_i)p_j(F+2R) - p_i^2 p_j}{p_i p_j(2p_i - 3)}$$

$$\hat{G}_{ijk} = \frac{f_{ijk} - p_i p_j p_k(F+2R)}{2p_i p_j p_k}$$

165 (3)

166 *G* is a normalized third central moment and the normalization constant depends upon the
167 pattern of subscript.

168 Each allele can provide an estimate of *G*, denoted $\hat{G}_i$, and its weighted estimate
169 across possible alleles *i* is

170 $$\hat{G} = \sum_i w_i \hat{G}_i \,. \qquad (4)$$

171 This represents a "linear estimator" of G. The weights are derived in the appendix. The
172 best alternative to linear estimation is maximum likelihood.

173

174 *Probabilities of four-marker relationship*

175 From Equation (7) of RITLAND (1987), which follows KENDALL *et al.* (1977 eq. 13.36),

$$f_{ijkl} = \boldsymbol{\kappa}_i\boldsymbol{\kappa}_j\boldsymbol{\kappa}_k\boldsymbol{\kappa}_l + \boldsymbol{\kappa}_i\boldsymbol{\kappa}_j\boldsymbol{\kappa}_{kl} + \boldsymbol{\kappa}_i\boldsymbol{\kappa}_k\boldsymbol{\kappa}_{jl} + \boldsymbol{\kappa}_i\boldsymbol{\kappa}_l\boldsymbol{\kappa}_{jk} + \boldsymbol{\kappa}_j\boldsymbol{\kappa}_k\boldsymbol{\kappa}_{il} + \boldsymbol{\kappa}_j\boldsymbol{\kappa}_l\boldsymbol{\kappa}_{ik} + \boldsymbol{\kappa}_k\boldsymbol{\kappa}_l\boldsymbol{\kappa}_{ij}$$
$$+ \boldsymbol{\kappa}_i\boldsymbol{\kappa}_{jkl} + \boldsymbol{\kappa}_j\boldsymbol{\kappa}_{ikl} + \boldsymbol{\kappa}_k\boldsymbol{\kappa}_{ijl} + \boldsymbol{\kappa}_l\boldsymbol{\kappa}_{ijk} + \boldsymbol{\kappa}_{ij}\boldsymbol{\kappa}_{kl} + \boldsymbol{\kappa}_{ik}\boldsymbol{\kappa}_{jl} + \boldsymbol{\kappa}_{il}\boldsymbol{\kappa}_{jk} + \boldsymbol{\kappa}_{ijkl}$$

The cumulants $\boldsymbol{\kappa}_i$ are similar to moments and covariances and may have a random component that may covary with other cumulants. The subscripts indicate alleles. The recursion equation is

$$f_{ijkl} = p_i p_j p_k p_l + (p_i p_j + p_k p_l) v_f + \left(p_i p_l + p_j p_k + p_j p_l + p_k p_l\right) v_r + p_i \omega_{jkl} + p_j \omega_{ikl} + p_k \omega_{ijl} + p_l \omega_{jkl} + v_{ij}^2 v_{kl}^2 + v_{ik}^2 v_{jl}^2 + v_{il}^2 v_{jk}^2 + Cov\left(v_{ij}, v_{kl}\right) + Cov\left(v_{ik} v_{jl}\right) + Cov\left(v_{il} v_{jk}\right) + \boldsymbol{\kappa}_{ijkl} \quad (5)$$

Where the *v* terms are second order covariances. When there is a mixture model (which creates the covariances), each subpopulation *m*, contributes to the mean cumulant across pooled *m*. The term $\kappa_{i,m}\kappa_{j,m}\kappa_{k,m}\kappa_{l,m}$ contributes to all 18 population level moments, the term $\kappa_{i,m}\kappa_{j,m}\kappa_{kl,m}$ contributes to six population level moments, and so on. However, the quantitative extent of these contributions are complex and beyond treatment here. Regardless, that subpopulation cumulants "distill" to the same assortment of cumulants, albeit with perhaps slightly different values. where the covariance terms are across the mixture terms *m*. This is a finite mixture model, needed when a single component distribution is inadequate (McLachlan *et al.* 2019). These can get complex but Withers *et al.* (2015) does provide the first known expressions for cumulants used available computer technology (an equation solver), not available in 1987 to KR. Cumulants are allowed to vary across the mixture, and that this results in effective covariance between second-order cumulants. The expectations taken across *m* result in changes to the above expression due to associations among the *ijkl* across *m*, that causes the cumulants to be associated in a certain way, since for example, $E\left[\kappa_{i,m}\kappa_{j,m}\right] \neq \kappa_i\kappa_j$.

196 The associations between *pairs* of genes is termed identity disequilibrium. If one
197 pair of alleles is heterozygous, it is more likely the second pair is also heterozygous. This is
198 a four-gene marker measure that has been neglected due to inordinate attention to linkage
199 disequilibrium. Identity disequilibrium has classically been characterized as the excess of
200 homozygosity above that expected from the squared gene frequencies (HILL 1975) (OHTA
201 1980). The identity excess is closely correlated to the expectation of the total squared
202 linkage disequilibrium (TAKAHATA 1982). Some of the problem is that haploid gametes are
203 not directly assayed but rather imputed (VITALIS AND COUVET 2001b).

204 We can add a cumulant to the equation for the probability of identity-by-state. From
205 equation 3.78 in KENDALL *et al.* (1977), the moments about the mean for two squared
206 random variables (the genes present at each locus) equals

$$\mathrm{E}\left[p_i^2 p_j^2\right] = \kappa_{22} + \kappa_{20}\kappa_{02} + 2\kappa_{11}^2$$

207 Whose form corresponds to $3\sigma_r^4$ for the fourth central moment with the difference that a
208 the cumulant κ_{22} is added. VITALIS AND COUVET (2001a) and others have given estimator for
209 identity disequilibrium which omits this cumulant.

210 The four-allele case introduces higher-order associations and brings with it new
211 statistical problems. Among four alleles, two new measures arise. The first is termed H
212 and is the probability that all four alleles are identical-by-descent. The other two have not
213 been recognized in the literature, perhaps because they invoke the existence of cumulants,
214 which differ from the corresponding moments with products of four or more variates.

215 The first, termed R_{AB}, is the probability that both alleles in the first relatives are
216 identical-by-descent to both alleles in the second relative. The second, termed F_{AB}, is the

probability that both individuals have both genes identical-by-descent. Thus, the three unique four-allele measures are

$$H$$

$$F_{AB} = F_A F_B + Cov(F_A, F_B)$$

$$R_{AB} = R^2 + Cov(R, R^c) \tag{6}$$

the covariances between second moments, $Cov(F_A, F_B)$ and $Cov(R_{AB}, R'_A B')$, exist only when the distribution of gene frequency follows a mixture distribution where subpopulations vary for F and R.

We can rewrite Equation (5) as

$$\begin{aligned}
f_{ijkl} = {} & (2-\delta_{ij})(2-\delta_{kl})\,[p_i p_j p_k p_l \\
& + (\delta_{jl} p_i p_k p_j + \delta_{jk} p_i p_l p_j + \delta_{il} p_j p_k p_i + \delta_{ik} p_j p_l p_i - 4 p_i p_j p_k p_l) R \\
& + (\delta_{ik}\delta_{jl} p_i p_j + \delta_{il}\delta_{jk} p_i p_j - \delta_{jl} p_i p_k p_j - \delta_{jk} p_i p_l p_j - \delta_{il} p_j p_k p_i - \delta_{ik} p_j p_l p_i + 2 p_i p_j p_k p_l) R_{AB}. \\
& + p_k p_l (\delta_{ij} p_i - p_i p_j) F_A + p_i p_j (\delta_{kl} p_k - p_k p_l) F_B \\
& + (\delta_{ij}\delta_{kl} p_i p_k - \delta_{kl} p_i p_j p_k - \delta_{ij} p_i p_k p_l + p_i p_j p_k p_l) F_{AB} \\
& + 2(\delta_{ijl} p_k p_{ijl} + \delta_{ijk} p_l p_{ijk} - p_i p_k p_{jl} - p_i p_l p_{jk} - p_j p_k p_{il} - p_j p_l p_{ik} - 2 p_k p_l p_{ij} + 4 p_i p_j p_k p_l) G_A \\
& + 2(\delta_{jkl} p_i p_j + \delta_{ikl} p_i p_k - p_i p_k p_{jl} - p_i p_l p_{jk} - p_j p_k p_{il} - p_j p_l p_{ik} - 2 p_i p_j p_{kl} + 4 p_i p_j p_k p_l) G_B \\
& + \begin{matrix}(p_{ijkl} - p_i p_{jkl} - p_j p_{ikl} - p_k p_{ijl} - p_l p_{ijk} + 2 p_i p_j p_{kl} + 2 p_i p_k p_{jl} + 2 p_i p_l p_{jk} \\ + 2 p_j p_k p_{il} + 2 p_j p_l p_{ik} + 2 p_k p_l p_{ij} - p_{ij} p_{kl} - p_{ik} p_{jl} - p_{il} p_{jk} - 6 p_i p_j p_k p_l)]\end{matrix} H
\end{aligned} \tag{7}$$

where, for shorthand, $p_{ij} = \delta_{ij} p_i$.

In this expression, there are eight relationship coefficients (R_{AB}, R_{ABAB}, F_A, F_B, F_{AB}, G_A, G_B, H), which in principle will specify eight different classes of marker genotypes. This probability of four alleles, f_{ijkl}, is then fitted to the observed frequencies in a sample.

For equations that solve for all eight parameters, the choice is somewhat arbitrary but a natural set of eight classes, in which identity-by-state mirrors the identity-by-descent, is: $A_iA_iA_iA_i$ (all identical by state, or "ibs"), $A_iA_kA_iA_k$, $A_iA_iA_kA_k$, (two pairs ibs), $A_iA_iA_iA_k$, $A_iA_kA_kA_k$, (one triplet ibs) and $A_iA_iA_jA_k$, $A_iA_jA_kA_k$ and $A_iA_jA_jA_k$. (one pair ibs between A and B). Thus we seek the expected frequencies in the vector (f_{iiii}, f_{ijij}, f_{iijj}, f_{iiij}, f_{iijk}, f_{ijik}).

The frequency of $A_iA_iA_iA_i$, is obtained from Eq (7) where all δ=1 and all marker frequencies are p_i:

$$f_{iiii} = p_i\left[p_i^3 + p_i^2 q_i(F_A + F_B + 4R) + p_i q_i^2(F_{AB} + 2R_{AB}) + 2p_i q_i(1 - 2p_i)(G_A + G_B) + q_i(1 - 6p_i q_i)H\right.$$

Likewise, the frequency that A and B are both heterozygous for A_i and A_j is, irrespective of order or phase,

$$f_{ijij} = 4p_ip_j\left[p_ip_j(1 - F_A - F_B - F_{AB}) + (p_i + p_j - 4p_ip_j)R + (1 - p_i - p_j + 2p_ip_j)R_{AB} + (-p_i - p_j + 4p_ip_j)(G_A + G_B) - (1 - 2p_i - 2p_j + 6p_ip_j)H\right.$$

and homozygous for alternative alleles A_i and A_j is

$$f_{iijj} = 2p_ip_j\left[p_ip_j(1 - 4R + 2R_{AB}) + q_ip_jF_A + p_iq_jF_B + (1 - p_i - p_j + p_ip_j)F_{AB} - 2p_j(1 - 2p_i)G_A - 2p_i(1 - 2p_j)G_B - (1 - 2p_i - 2p_j + 6p_ip_j)H\right.$$

for triplets of identity-by-state f_{iiij}

$$f_{iiij} = 2p_ip_j\left[p_i^2 + p_iq_iF_A - p_i^2F_B - p_iq_iF_{AB} + (p_i + p_j - 4p_ip_j)R - (p_i + p_j - 2p_ip_j)R_{AB} + (1 - 4p_i + 4p_ip_j)G_A - 2p_i(1 - 2p_j)G_B - (1 - 6p_iq_i)H\right.$$

Finally, a single allele pair *ii* can be shared only within individual *A*,

$$f_{iijk} = 2p_ip_jp_k[p_i(1-4R+2R_{AB}-F_B+F_{AB})+q_iF_A-2(1-2p_i)G_A+4p_iG_B+2(1-3p_i)H$$

or shared only once between *A* and *B*:

$$f_{ijik} = 4p_ip_jp_k[p_i(1-F_A-F_B+F_{AB})+(1-4p_i)R-(1-2p_i)R_{AB}-(1-4p_i)(G_A+G_B) + 2(1-3p_i)H$$

The expressions for f_{ijjj} and f_{ijkk} are obtained by symmetry, and the expression for f_{ijjk} is summed over all four pairings of *j* between *A* and *B*: $f_{ijjk}, f_{ijkj}, f_{jijk}$, and f_{jikj}.

The appendix gives the 8x8 matrix of probabilities of observing the marker gene frequencies *f* given the relatedness coefficients. Of course, this depends upon the particular array of *f*'s used (there are others than the above). In this case, the determinant of the matrix is $512p_i^8\,p_j^7p_k^3(p_i-1)\left(96p_i^7-160p_i^6-4p_i^5(24p_j^2-16p_j-9)+3p_i^4(32p_j^2-80p_j+47)-p_i^3(132p_j^2-302p_j+147)+p_i^2(89p_j^2-154p_j+59)+\mathrm{p}(1-p_j)(23p_j-11)+(p_j-1)(2p_j-1)\right)$. That it is non-zero indicates all 8 parameters are jointly estimable, but a linear approach which uses residuals to simplify things is needed at this point.

Joint values of H and identity disequilibrium

$$\begin{bmatrix} f'_{AAAA} \\ f'_{AAaa} \end{bmatrix} = \begin{bmatrix} p(1-p)(1-6p(1-p)) & p^2(1-p)^2 \\ -2pq(1-2p-2q+6pq) & 2p^2q^2 \end{bmatrix} \begin{bmatrix} H \\ F_{ab} \end{bmatrix}$$

whose solution is

$$\widehat{H} = \frac{q^2 f'_{AAAA} - (1-p)^2 f'_{AAaa}}{pq(1-p)(1-2p)(1-p-q)}$$

$$\widehat{F_{ab}} = \frac{q(1-2p-2q+6pq)f'_{AAAA} - (1-p)\big(1-6p(1-p)\big)f'_{AAaa}}{pq(1-p)(1-2p)(1-p-q)}$$

259 Joint estimates of H and joint identity disequilibrium

260 $$\begin{bmatrix} f_{AAAA} \\ f_{AaAa} \\ f_{AAaa} \end{bmatrix} = \begin{bmatrix} p(1-p)\big(1-6p(1-p)\big) & p^2(1-p)^2 & 2p^2(1-p)^2 \\ -4pq(1-2p-2q+6pq) & 4p^2q^2 & 4pq((1-p)(1-q)+pq \\ -2pq(1-2p-2q+6pq) & 2pq(1-p-q+pq) & 4p^2q^2 \end{bmatrix} \begin{bmatrix} H \\ F \\ R \end{bmatrix}$$

$$\widehat{H} = \frac{2q(1-p-q+3pq)f'_{AAAA} - p(1-p)^2(f'_{AAaa} + f'_{AaAa})}{2pq(1-p)(1-p-q)(1-3p)}$$

$$\widehat{F_{ab}} = \frac{2q(1-2p-2q+6pq)f'_{AAAA} + (1-p)(2p^2-4p+1)f'_{AAaa} + p(1-2p)f'_{AaAa}}{2pq(1-p)(1-p-q)(1-3p)}$$

$$\widehat{R_{abab}} = \frac{4q(1-2p-2q+6pq)f'_{AAAA} + (1-p)p(1-2p)f'_{AAaa} - (4p^2-5p+1)f'_{AaAa}}{2pq(1-p)(1-p-q)(1-3p)}$$

261

262 The denominator shows that at least three alleles required in the population, and marker
263 frequencies of p_i=1/3 are non-informative.

264

265 ## DISCUSSION AND CONCLUSION

266 A main feature of higher-order relatedness is the covariance of homozygosity
267 between pairs of genes, this is effectively a covariance of second moments. Such a
268 "covariance of covariance" arises when pedigrees occur in a mixture distribution
269 (McLachlan *et al.* 2019). Such a distribution generates the genomic variation of
270 homozygosity necessary for the existence of covariance of homozygosity at individual loci.
271 The simplest mixture distribution is that of two populations with gene frequency $p+a$ and
272 $p-a$; in this case covariance of heterozygosity, after mixing in equal proportions, equals
273 $a^4+6a^2p^2$.

Another feature of higher-order relatedness is that the four-marker coefficient of gene identity must be described with cumulants and not moments. As an example of the necessity of cumulants, for population gene frequency p, the fourth central moment for four genes, denoted (X_1, X_2, X_3 and X_4), is $\sigma_{12}^2\sigma_{34}^2 + \sigma_{13}^2\sigma_{24}^2 + \sigma_{23}^2\sigma_{13}^2 + \kappa_4$ where σ_{ij}^2 is the covariance of X_i and X_j and κ_4 is a fourth-order cumulant which does not appear in the moment. Some type of term (not involving the product of variances) is needed for κ_4 and it could be any rational number. In summary, incorporating cumulants into four marker measures only requires some value X in the expansion of the fourth central moment $\sigma_{12}^2\sigma_{34}^2 + \sigma_{13}^2\sigma_{24}^2 + \sigma_{23}^2\sigma_{13}^2 + X$ and this X is numerically estimated in the same way as the lower order cumulant terms.

Cumulants do have useful properties for models of quantitative traits, the most important is that the cumulant of the sum of two random variables X+Y is M(X+Y)=M(X)+M(Y); differential equations for models of selection on quantitative traits that involve cumulants are simpler than models involving moments (BURGER 1991; TURELLI AND BARTON 1994). This cumulant will also be key in deriving a marker-based estimator for Q_{st} (RITLAND IN PREP) and for a portrayal of higher order population structure that separately accounts for both the correlation of relationship and the squared linkage disequilibrium (RITLAND IN PREP).

We give probabilities of relationship for a homogenous population of just one generation. Such populations are most commonly assayed in genomics, however it should be noted that the levels of nucleotide variation (for SNPs) is not high and loci with more than two alleles are uncommon; in fact, only about 5% of human SNPs are triallelic (CAO *et al.* 2015) although microsatellites and other types of repeat markers show greater

variation. Reconstruction of pedigree relationship has traditionally involved cumbersome graph-tracing algorithms, and simpler recursive methods which require at least two generations of records (KARIGL 1981; THOMPSON 1988; WHITTEMORE AND HALPERN 1994). Also, current recursive methods (ZHENG *et al.* 2018) assume a known pedigree (KIRKPATRICK *et al.* 2018). This is somewhat like estimating Q_{st} with current methods, where aspects of the pedigree must be known.

Normalization constants

Relatedness coefficients are obtained by calculating the pairwise covariance of relatives and dividing it by a normalization constant that converts the covariance into a correlation. This constant is the maximum possible value that the covariance can take. For cases where pairwise comparisons involve the frequency of identical genotypes, it is simple to calculate as a binomial variance. For the two-marker relationship coefficient as described in Equation (1), the maximum covariance between two genotypes, conditioned upon observing allele i, is $\mathrm{E}[A_iA_i] - \mathrm{E}[A_i]^2 = p_i(1\text{-}p_i)$ when R=1. Likewise, the three marker coefficient has a normalization constant of $p_i(1\text{-}p_i)(1\text{-}2p_i)$ for $A_iA_iA_i$., and the four-marker coefficients are normalized by $p_i(1\text{-}p_i)(1\text{-}6p_i(1\text{-}p_i))$ for $A_iA_iA_iA_i$. The normalization constants for combinations of alleles falls out of the analyses.

ACKERMAN *et al.* (2017) provided a different set of normalization constants for three and four marker measures than given here. In their Equation 6, they normalized the third central moment by the geometric mean gene frequency of the three central moments (rather than by $p(1\text{-}p)(1\text{-}2p)$ as done here). Their justification was a similarity of this "third moment correlation" to a bivariate correlation formula. Their normalization

constant for the four-marker coefficient (Equation 8) involves a parameter α that mixes the unknown proportions of the two types of higher order identity (all identical vs. 2 pairs identical), resulting in an inference that may be subject to biases.

Estimation

Calculating the probabilities of higher-order relationship poses an interesting set of obstacles. For determining the exact probability of pairwise relatedness, we must consider that gene frequencies are simultaneously estimated along with relatedness. By ignoring this we assume the population gene frequencies are estimated from an effectively infinite sample. Although formula can describe how sample size bias might be corrected, in practice, this bias can be eliminated by excluding the pair of relatives under consideration from the estimation of population gene frequency. Such probabilities are can also undefined at certain intermediate frequencies (p=1/2 or 1/3) and about these frequencies give low information (RITLAND 1987). The algebra of higher-order relationships is also cumbersome and the statistical uncertainty of estimates much greater. With four markers, equation solvers such as Derive or Mathematica can help with the complicated formulae.

The estimation properties of method of moments and likelihood estimators were both examined with Monte-Carlo simulations. Some observations are when genotypes are identical, that pairwise estimates of G and H are high (above 1 for rarer alleles), while common alleles give lower and even negative values. Likelihood requires numerical solutions which introduces complications, as the numerical solution is normally iterated until convergence. In examining likelihood, I found the number of marker loci needed for adequate convergence was about 20 loci for three-marker coefficients and roughly double

at 30-50 loci for four marker coefficients. Interestingly, it was found that loci with fewer alleles are more likely to give convergent estimates because the problem with non-convergence arises when relatives do not share the same marker allele. Software such as PLINK (PURCELL *et al.* 2007) should incorporate higher order relatedness coefficients.

Possible approaches

The complications of correctly estimating population structure are discussed and treated by WEIR AND GOUDET (2017). They developed two-marker moment estimators that can describe the "relativity" and this requires an explicit reference population. They develop their estimator in a multilevel approach (within individuals, between individual within populations, and between populations) which promoted a unified treatment of relatedness and population structure. Clearly, further progress will depend upon adequate definitions and applications of models.

"Relatedness mapping" (ALBRECHTSEN *et al.* 2009) uses relatedness to identifying causative mutations, using the principle that affected individuals share higher relatedness about the mutation. A somewhat related activity is "IBD mapping", in which segments of identity-by-descent (IBD) present in high-density genomic data are used to map casual variants (BROWNING AND BROWNING 2012). However, the data by itself only reveals the presence of the variant.

Other fields have adopted the use of cumulants, which may show new approaches that population genetics can undertake. The central 4^{th} cumulant has been used to detect early stages of termite infestation, as it can separate termite alarm signals from background noise (DE LA ROSA AND MORENO MUÑOZ 2008). Advances in electrophysiological

and imaging techniques are used to study the synchrony of neuron cell firings in the brain (STAUDE AND ROTTER 2010), and have highlighted the need for correlation measures that go beyond simple pairwise analyses, taking advantage of the "interaction property" of higher order cumulants as measures of correlation (STAUDE *et al.* 2010). In information systems, a "covariance of covariance" approach for individual pixels has been developed for image description and classification (SERRA *et al.* 2014).

ACKNOWLEDGMENTS Joe Felsenstein sponsored the sabbatical and provided discussion in the early stages of this work. This work was supported by ongoing grants from NSERC to KR, a Sloan Sabbatical fellowship to KR in 1991

Table 1. Examples of the statistical variances of relationship coefficients when estimates are based upon a single locus and when the true values are zero ("x" denotes not estimable). See equations 3 and 6 for definitions of G and H.

Array of p	F	G_A	G_B	ϕ_{XY}	H
0.6,0.3,0.1	2.93	73.5	74.3	12.96	7.63
0.5,0.3,0.2	2.32	x	x	11.66	3.66
0.4,0.3,0.2,0.1	1.58	0.19	7.79	13.05	2.18

Table 2. Allele states and denominators of cumulants.

Allele states	Expectation	Denominato
Two alleles		
$i = j$	$p_i^2 + p_i(1-p_i)F$	$p_i(1-p_i)$
$i \neq j$	$2p_ip_j(1-F)$	$-2p_ip_j$
Three alleles		
$i = j = k$	$p_i^3 + p_i^2(1-p_i)(F+2R) + p_iG$	$p_i(1-p_i)(1-$
$i = j \neq k$	$p_i^2p_k(2R-2G)$	$-p_ip_k(1-2p$
$i \neq j \neq k$	$p_ip_jp_k(1-F-2R+2G)$	$2p_ip_jp_k$
Four alleles		
$i = j = k = l$	$p_i^2q_i^2(F_{AB}+2R_{AB}) + q_i(1-6p_iq_i)H$	$p_i(1-p_i)\big(1-6p_i(1-$
$i = j = k \neq l$	$2p_ip_j(F_{AB} + \big(p_i+p_j-2p_ip_j\big)R_{AB} - (1-6p_iq_i)H)$	$p_ip_k(1-2p_i-2p_k+6$
$i = j \neq k = l$	$2p_ip_j\big(2p_ip_jR_{AB} + \big(1-p_i-p_j+p_ip_j\big)F_{AB} - (1-2p_i-2p_j+6p_ip_j)H\big)$	$p_ip_k(1-2p_i-2p_k+6$
$i = j \neq k \neq l$	$2p_ip_jp_k([p_i(2R_{AB}+F_{AB}) + 2(1-3p_i)H)$	$2p_ip_jp_k(1-3p_i)$
$i \neq j \neq k \neq l$	$p_ip_jp_kp_l(1-4R-2F+4G-8H)$	$-6p_ip_jp_kp_l$

408 FIGURE LEGENDS

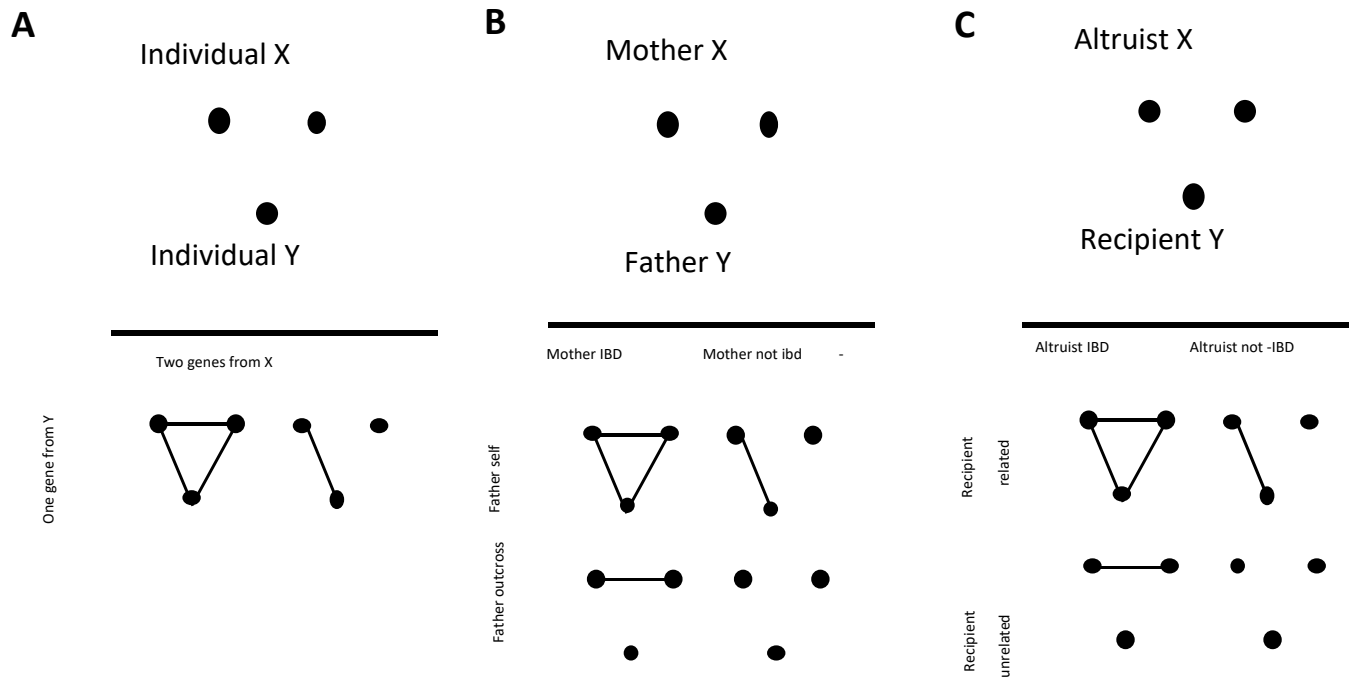
409

410 Figure 1. Three cases where three-gene modes of gene identity are used. (A) Effective
411 selfing model involves two genes of one parent, (B) progeny pair model, (C) Altruist-
412 recipient. Identical genes are linked by lines. In the effective selfing model, two of the
413 genes are from the maternal parent and the other gene is the paternal contribution.

414

415 Figure 2. Two cases where four-gene modes of identity are used. (A) The general case
416 where each of the nine identity modes are inferred (B) A model to fit progeny pairs to
417 mating system parameters (C) The inference of heritability in the field, where M is the
418 marker and Q the quantitative trait.

419



420

421 Figure 1

422

423

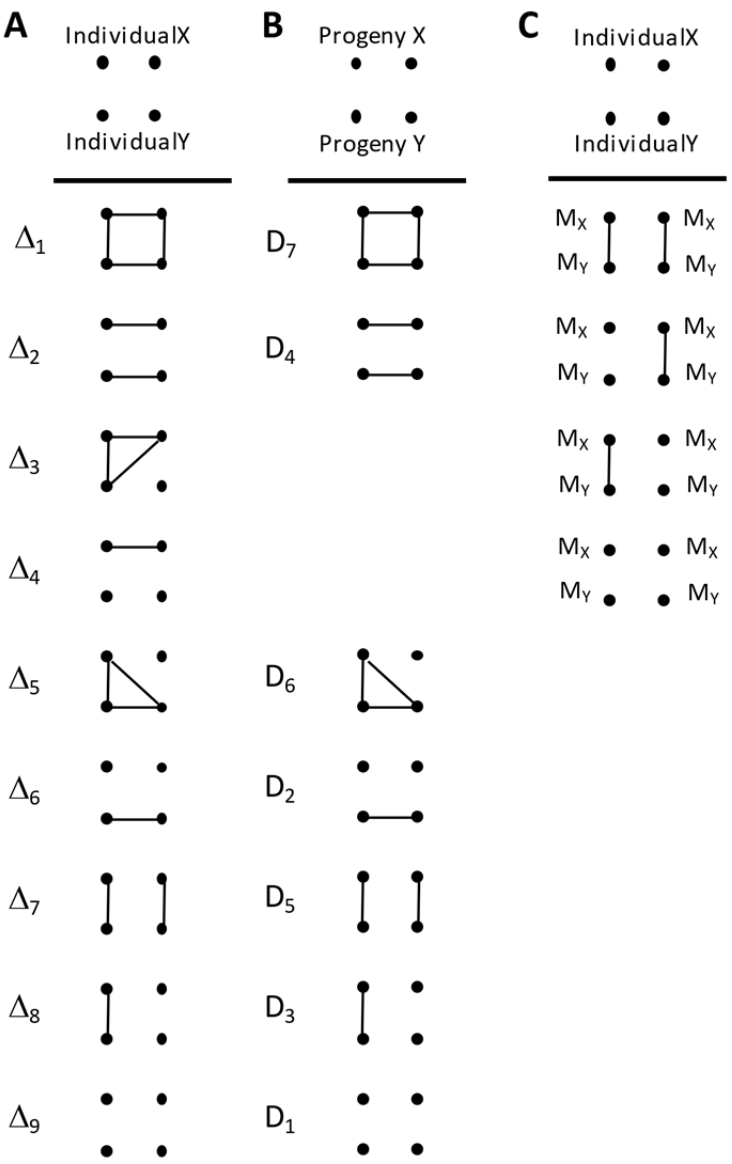

A
IndividualX
IndividualY
B
Progeny X
Progeny Y
C
IndividualX
IndividualY
Δ1
Δ2
Δ3
Δ4
Δ5
Δ6
Δ7
Δ8
Δ9
D7
D4
D6
D2
D5
D3
D1
MX
MY

Figure 2.

Literature cited

Ackerman, M. S., P. Johri, K. Spitze, S. Xu, T. G. Doak *et al.*, 2017 Estimating Seven Coefficients of Pairwise Relatedness Using Population-Genomic Data. Genetics 206**:** 105-118.

Albrechtsen, A., T. Sand Korneliussen, I. Moltke, T. van Overseem Hansen, F. C. Nielsen *et al.*, 2009 Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 33**:** 266-274.

Browning, S. R., and B. L. Browning, 2012 Identity by Descent Between Distant Relatives: Detection and Applications. Annual Review of Genetics 46**:** 617-633.

Burger, R., 1991 Moments, cumulants, and polygenic dynamics. Journal of mathematical biology 30**:** 199-213.

Cao, M., J. Shi, J. Wang, J. Hong, B. Cui *et al.*, 2015 Analysis of human triallelic SNPs by next-generation sequencing. Ann Hum Genet 79**:** 275-281.

de la Rosa, J. J. G., and A. Moreno Muñoz, 2008 Higher-order cumulants and spectral kurtosis for early detection of subterranean termites. Mechanical Systems and Signal Processing 22**:** 279-294.

Gardner, A., S. A. West and G. Wild, 2011 The genetical theory of kin selection. Journal of Evolutionary Biology 24**:** 1020-1043.

Hill, W. G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor Popul Biol 8**:** 117-126.

Jacquard, A., 1966 Logique du calcul des coefficients d'identité entre deux individus. Population (french edition)**:** 751-776.

Jacquard, A., 2012 *The genetic structure of populations*. Springer Science & Business Media.

Karigl, G., 1981 A recursive algorithm for the calculation of identity coefficients. Annals of Human Genetics 45**:** 299-305.

Kendall, M. G., A. Stuart and J. K. Ord, 1977 *The advanced theory of statistics.* , London, Griffin.

Kirkpatrick, B., S. Ge and L. Wang, 2018 Efficient computation of the kinship coefficients. Bioinformatics 35**:** 1002-1008.

Liu, W. W. B., 2005 Genotypic probabilities for pairs of inbred individuals. Philosophical Transactions of the Royal Society B 360**:** 1379-1385.

Malécot, G., 1948 *Mathématiques de l'hérédité*. Masson, Paris.

McLachlan, G. J., S. X. Lee and S. I. Rathnayake, 2019 Finite mixture models. Annual review of statistics and its application 6**:** 355-378.

Michod, R. E., and W. D. Hamilton, 1980 Coefficients of relatedness in sociobiology. Nature 288**:** 694-697.

Morton, N. E., S. Yee, D. E. Harris and R. Lew, 1971 Bioassay of kinship. Theoretical Population Biology 2**:** 507-524.

Ohta, T., 1980 Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families. Genetical research 36**:** 181-197.

Pamilo, P., and R. H. Crozier, 1982 Measuring genetic relatedness in natural populations: Methodology. Theoretical Population Biology 21**:** 171-193.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics 81**:** 559-575.

Queller, D. C., 1992 Quantitative Genetics, Inclusive Fitness, and Group Selection. The American Naturalist 139**:** 540-558.

Queller, D. C., and K. F. Goodnight, 1989 Estimating relatedness using genetic markers. Evolution 43**:** 258-275.

Ritland, K., 1984 The effective proportion of self-fertilization with consanguineous matings in inbred populations. Genetics 106**:** 139-152.

Ritland, K., 1985 The genetic mating structure of subdivided populations I. Open-mating model. Theoretical Population Biology 27**:** 51-74.

Ritland, K., 1987 Definition and Estimation of Higher-Order Gene Fixation Indices. Genetics 117**:** 783-793.

Ritland, K., 1996 Estimators for pairwise relatedness and individual inbreeding coefficients. Genetics Research 67**:** 175-185.

Ritland, K., 2000 Marker-inferred relatedness as a tool for detecting heritability in nature. Molecular Ecology 9**:** 1195-1204.

Ritland, K., and M. Leblanc, 2004 Mating system of four inbreeding monkeyflower (Mimulus) species revealed using ‘progeny-pair’analysis of highly informative microsatellite markers. Plant Species Biology 19**:** 149-157.

Samanta, S., Y. J. Li and B. S. Weir, 2009 Drawing inferences about the coancestry coefficient. Theoretical Population Biology 75**:** 312-319.

Serra, G., C. Grana, M. Manfredi and R. Cucchiara, 2014 Covariance of covariance features for image classification, pp. 411-414 in *Proceedings of International Conference on Multimedia Retrieval*.

Staude, B., S. Grün and S. Rotter, 2010 Higher-order correlations and cumulants, pp. 253-280 in *Analysis of parallel spike trains*. Springer.

Staude, B., and S. Rotter, 2010 Higher-order correlations in non-stationary parallel spike trains: statistical modeling and inference. Frontiers in computational neuroscience 4**:** 16.

Takahata, N., 1982 Linkage disequilibrium, genetic distance and evolutionary distance under a general model of linked genes or a part of the genome. Genetical Research 39**:** 63-77.

Thompson, E. A., 1988 Two-locus and Three-locus Gene Identity by Descent in Pedigrees. Mathematical Medicine and Biology: A Journal of the IMA 5**:** 261-279.

Turelli, M., and N. H. Barton, 1994 Genetic and statistical analyses of strong selection on polygenic traits: what, me normal? Genetics 138**:** 913-941.

Vitalis, R., and D. Couvet, 2001a Estimation of Effective Population Size and Migration Rate From One- and Two-Locus Identity Measures. Genetics 157**:** 911-925.

Vitalis, R., and D. Couvet, 2001b Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population. Genetics Research 77**:** 67-81.

Wang, J., 2014 Marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods. Journal of Evolutionary Biology 27**:** 518-530.

Weir, B. S., 1994 The effects of inbreeding on forensic calculations. Annu Rev Genet 28**:** 597-621.

Weir, B. S., and J. Goudet, 2017 A unified characterization of population structure and relatedness. Genetics 206**:** 2085-2103.

Whittemore, A. S., and J. Halpern, 1994 Probability of Gene Identity by Descent: Computation and Applications. Biometrics 50**:** 109-117.

Withers, C. S., S. Nadarajah and S. H. Shih, 2015 Moments and Cumulants of a Mixture. Methodology and Computing in Applied Probability 17**:** 541-564.

511 Zheng, C., M. P. Boer and F. A. van Eeuwijk, 2018 Recursive algorithms for modeling genomic ancestral origins in
512 a fixed pedigree. G3: Genes, Genomes, Genetics 8**:** 3231-3245.

513