Relatedness coefficients and their applications for triplets and quartets of genes

Kermit Ritland

Biodiversity Research Center

University of British Columbia

RUNNING HEAD: Relatedness with gene triplets and quartets

ABSTRACT

Relatedness coefficients which measure the genetic similarity of groups of genes (either three or four) are described. Cumulants and not moments are needed for four-gene probabilities, and also, the normalized fourth cumulant is the probability of identical-by-descent of all four genes. With four genes, a covariance of homozygosity between pairs of genes arises when populations follow a mixture distribution. Applications of three gene coefficients mainly involve conditional regression, and applications of four gene coefficients can involve identity disequilibrium. A novel estimator for identity disequilibrium emerges from calculations involving the cumulant. -Estimation of three- and four-gene relatedness using genetic marker data is briefly discussed. Three- and four-gene estimators explore new territory but suffer from statistical and numerical problems, including higher statistical variance, complexity of estimation formula, and non-estimability at some intermediate allele frequencies.

30 INTRODUCTION

31 Relatedness is a general term for the level of genetic similarity between individuals,
32 and is measured by the sharing alleles identical-by-descent (MALÉCOT 1948; PAMILO AND
33 CROZIER 1982). Relatedness is quantified with gene identity coefficients, which characterize
34 both the pattern and the frequency of identity-by-descent. The unit of observation is
35 normally a pair of genes, and the object of estimation is the kinship coefficient or the
36 coefficient of relationship. In this paper, the unit of observation is extended to triplets and
37 quartets of genes, allowing the opportunity to characterize additional parameters of
38 population structure.

39 Relatedness may be estimated with genetic markers (MORTON *et al.* 1971) and for
40 pairs of genes, many computer programs are available for estimation of relatedness (WANG
41 2014), in particular for "pairwise relationship", such as the *r* of QUELLER AND GOODNIGHT
42 (1989). However, the equations for pairwise relationship are not extendable to three or
43 four genes, as the covariances and higher moments need to be defined in new ways.

44 Estimators for three and four gene measures of relatedness have recently been
45 proposed. SAMANTA (2009) provided the first estimator for three genes. ACKERMAN *et al.*
46 (2017) examined all four genes and described the estimation of seven of eight coefficients
47 of relatedness. Multiallelic data have information about all eight coefficients, but they used
48 a biallelic model which provides just seven degrees of freedom, constraining their space of
49 estimates. In addition, they did not use cumulants and their moments are normalized
50 differently than will be here.

51 Cumulants are of use in certain problems in quantitative genetics (BURGER 1991;
52 TURELLI AND BARTON 1994) and compared to cumulants have more useful theoretical

properties (KENDALL *et al.* 1977), but ordinary moments are sufficient for two and three-gene fixation indices. RITLAND (1987) found cumulants instead of moments were an essential component of four-gene fixation indices. Fourth order cumulants are needed to specify the probability of gene identity for all four genes, and also to describe identity disequilibrium as the "covariance of covariances". 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 $E[(X_1 - p)(X_2 - p)(X_3 - p)(X_4 - p)] = \sigma_{12}^2\sigma_{34}^2 + \sigma_{13}^2\sigma_{24}^2 + \sigma_{23}^2\sigma_{13}^2 + \kappa_4$ 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 to estimate relatedness in a population with polymorphic genetic markers, with estimation procedures described in a companion paper (RITLAND 2023). Any numbers of alleles at a single locus are allowed in the equations, but practically one has two and at most alleles with SNP types of data, and for tractability, we use just a three-allele model to examine a special situation with four genes. These models and estimation procedures are readily applicable to the emerging mountains of genome data.

APPLICATIONS OF THREE AND FOUR GENE MEASURES

Between two diploid individuals, there are 15 patterns of gene identity, and these patterns can be parameterized in various ways to yield probabilities of pairs of individuals (LIU 2005). A pair of individuals can share two, three or four genes, and at each level, all genes are used to describes an aspect of relatedness.

Relatedness and two genes. The two-gene 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 sample within one individual (the "inbreeding coefficient"). Analysis of data with two gene measures are ubiquitous (WANG 2014) and the two-gene measures are often incorporated into probabilities of groups of three and four genes. 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).

Regression and three genes. The three-gene relationship coefficient is the probability that three sampled genes are all identical-by-descent (Figure 1a). This coefficient, G, is usually combined with two-gene 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 $2R\text{-}G_A$, which is the probability that either gene in A is identical to the single gene donated by its mate, B (RITLAND 1984) (Figure 1b). In the theory of kin selection, the strength of kin selection is proportional to the relatedness between the recipient and the altruist (QUELLER 1992; GARDNER *et al.* 2011). 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 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 genes. 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-gene measures (F_{AB}, R_{AB}, H). At the highest level, the measures are much different as the four-gene measures F_{AB}, R_{AB} are covariances (not identities) and H cannot be directly estimated (see below). 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 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).

In the four-gene 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 marker genes and two are quantitative trait loci genes (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- GENE 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 described in the "normalization constant" section in the discussion, 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-gene 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-gene 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

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

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

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

$$\hat{R} = \sum_i w_i \hat{R}_{ii} \tag{1}$$

where the weights w_{ii} sum to unity. 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-gene relationship

The three-gene relationship coefficient is the probability that three sampled genes are all identical-by-descent. Each possible combination allele triplets can be considered but it is simplest just to work with triplets identical for the same allele.

From Equation (7) of RITLAND (1987), which follows KENDALL *et al.* (1977 eq. 13.36), The joint frequency of genes 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 genes i and j are from one individual and gene 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

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

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

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

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

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

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

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

$$\hat{G} = \frac{f_{ijk} - p_i p_j p_k(F+2R)}{2p_i p_j p_k} \tag{3}$$

where the "hat" denotes an estimate. G is a normalized third central moment and note that the normalization constant depends upon the pattern of subscript. The above mainly shows that three separate estimators can be derived for G, and each provides information about G albeit in decreasing order.

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

$$\hat{G} = \sum_i w_i \hat{G}_i \,. \tag{4}$$

This represents a "linear estimator" of G. The weights and other topics about estimation are treated in RITLAND (2023).

Probabilities of four-gene relationship

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

$$\mathrm{E}[(X_1 - p_i)^2 (X_2 - p_i)^2] = \kappa_{22} + \kappa_{20}\kappa_{02} + 2\kappa_{11}^2$$

Whose form corresponds to $3\sigma_r^4$ for the fourth central moment with the difference that a the cumulant κ_{22} is added.

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 have a random component that may covary. The is written in conventional population genetic terms as

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

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, $E\left[\kappa_{i,m}\kappa_{j,m}\right] \neq \kappa_i\kappa_j$.

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

The first, termed R_{AB}, is the probability that both genes in the first relatives are identical to both genes in the second relative. The second, termed F_{AB}, is the likewise probability that both individuals have homologous genes identical-by-descent. Thus, the three unique four-gene 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*.

The general probability for the four genes is quite complex and cumbersome. One version is in the appendix but is difficult to work with. A second version is given below and used in this paper. This version is directly derived from Jacquard's identity modes and the line of derivation is clearer. The total probability is indexed by a vector F which contains the eight relatedness and a constant in the order $F = 1, F_A, F_B, R, G_A, G_B, H, F_{AB}, R_{AB})$. We multiply these vector elements by a function of gene frequency $Pr()$, which depends upon the identity modes (idmode) which have equal prior probability. Thus, the total genotypic probability is

$$f_{ijkl} = \sum_{idmode=1}^{9} \sum_{parameter=1}^{9} F(parameter)\, Pr(gene\ frequencies | parameter, idmode) \tag{7}$$

This equation follows the Equation (13) from (RITLAND 1987) in that we condition upon the identity mode (idmode). There are other ways to organize these calculations but here we start with Jacquard's identity modes. The appendix gives the values of $Pr()$ for each identity mode/coefficient combination.

In Equation (7), there are eight relationship coefficients which in principle will specify eight different classes of marker genotypes. This probability of four genes, f_{ijkl}, is then fitted to the observed frequencies in a sample, using some statistical procedure as described in RITLAND (2023).

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 gene frequencies are p_i:

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

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

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

and homozygous for alternative allels A_i and A_j is

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

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

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

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

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

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

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

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 4 gene frequencies *f* given the 4 gene 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 $512p_i^8\, p_j^7 p_k^3(p_i - 1)\Big(96p_i^7 - 160p_i^6 - 4p_i^5\big(24p_j^2 - 16p_j - 9\big) + 3p_i^4\big(32p_j^2 - 80p_j + 47\big) - p_i^3\big(132p_j^2 - 302p_j + 147\big) + p_i^2\big(89p_j^2 - 154p_j + 59\big) + \mathrm{p}\big(1 - p_j\big)\big(23p_j - 11\big) + \big(p_j - 1\big)\big(2p_j - 1\big)\Big)$. 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

The associations between *pairs* of genes is termed identity disequilibrium. If one pair of alleles is heterozygous, it is more likely the second pair is also heterozygous. This is an inherent four-gene measure that has been neglected due to inordinate attention to linkage disequilibrium (which is not a four-gene measure).

To find the unique four-gene probabilities (*H*, *F*, *R*) we need to separate the three components in the fourth cumulant. The following equations are solved in parts a, b, and c below corresponding to the number of branches emanating from the cumulant

$$f_{AAAA} = p_i\lceil p_i^3 + p_i q_i^2(F_{AB} + 2R_{AB}) + q_i(1 - 6p_i q_i)H$$

$$f_{AAaa} = 2p_i p_j\lceil p_i p_j(1 + 2R_{AB}) + \big(1 - p_i - p_j + p_i p_j\big)F_{AB} - (1 - 2p_i - 2p_j + 6p_i p_j)H$$

$$f_{AaAa} = 4p_i p_j\lceil p_i p_j(1 - F_{AB}) + \big(1 - p_i - p_j + 2p_i p_j\big)R_{AB} + (1 - 2p_i - 2p_j + 6p_i p_j)H$$

$$f_{AAaa} + f_{AaAa} = 2p_ip_j\left(\left(1 - p_i - p_j + 3p_ip_j\right)R_{AB} - \left(1 - 2p_i - 2p_j + 6p_ip_j\right)H\right).$$

(8)

The following three cases represent the three patterns of that emanate from the cumulant

a. Three branches lead to cumulant

$$\begin{bmatrix} f_{AAAA,aaaa} \\ f_{AAaa,aaAA} \\ f_{AaAa,aAaA} \\ f_{AaaA,aAAa} \end{bmatrix} = \begin{bmatrix} pq(1-6p(1-p)) & p^2(1-p)^2 & p^2(1-p)^2 & p^2(1-p)^2 \\ -2pq(1-2p-2q+6pq) & 2pq(1-2p-2q+6pq) & 2p^2q^2 & 2p^2q^2 \\ -2pq(1-2p-2q+6pq) & 2p^2q^2 & 2pq(1-2p-2q+6pq) & 2p^2q^2 \\ -2pq(1-2p-2q+6pq) & 2p^2q^2 & 2p^2q^2 & -2pq(1-2p-2q+6pq) \end{bmatrix} \begin{bmatrix} H \\ R_{AB1} \\ R_{AB2} \\ R_{AB3} \end{bmatrix}$$

b. Two branches lead to cumulant

$$\begin{bmatrix} f_{AAAA} \\ f_{AaAa,AaaA} \\ f_{AAaa} \end{bmatrix} = \begin{bmatrix} pq(1-6p(1-p)) & p^2(1-p)^2 & 2p^2(1-p)^2 \\ -4pq(1-2p-2q+6pq) & 4p^2q^2 & 4pq((1-p)(1-q)+pq) \\ -2pq(1-2p-2q+6pq) & 2pq(1-p-q+pq) & 4p^2q^2 \end{bmatrix} \begin{bmatrix} H \\ R_{AB1} \\ R_{AB2} \end{bmatrix}$$

c. One branch only

$$\begin{bmatrix} f_{AAAA} \\ f_{AAaa,AaAa,AaaA} \end{bmatrix} = \begin{bmatrix} pq\left(1-6p(1-p)\right) & 3p_i^2(1-p_i)^2 \\ -6q_i(1-2p-2q+6pq) & 6p_iq_i(3p_iq_i - p_i - q_i + 1) \end{bmatrix} \begin{bmatrix} H \\ R_{AB} \end{bmatrix}$$

(9)

Inversion of the above relations yields the following things moment solutions

a. Three branches

$$\widehat{H} = \frac{2q(1-p-q+3pq)f_{AAAA} - pq^2(f_{AAaa} + f_{AaAa,AaaA})}{2pq^2(1-p-q)(1-3p)}$$

$$\widehat{F_{AB}} = \frac{2q(1-2p-2q+6pq)f_{AAAA} + q_i(2p^2-4p+1)f_{AAaa} + p(1-2p)f_{AaAa,AaaA}}{2pq^2(1-p-q)(1-3p)}$$

$$\widehat{R_{AB}} = \frac{4q(1-2p-2q+6pq)f_{AAAA} + pq(1-2p)f_{AAaa} - (4p^2-5p+1)f_{AaAa,AaaA}}{2pq^2(1-p-q)(1-3p)}$$

b. Two branches

$$\widehat{H} = \frac{2q(1-p-q+3pq)f_{AAAA} - pq^2(f_{AAaa} + f_{AaAa,AaaA})}{2pq^2(1-p-q)(1-3p)}$$

$$\widehat{R_{AB}} = \frac{4q(1-2p-2q+6pq)f_{AAAA} + pq(1-2p)f_{AAaa} - (4p^2-5p+1)f_{AaAa,AaaA}}{2pq^2(1-p-q)(1-3p)}$$

c. One branch

$$\widehat{H} = \frac{f_{AAAA}2(3pq - p - q + 1) + pf_{AAaa,AaAa,AaaA}}{2pq(1 - p - q)(1 - 3p)}$$

$$\widehat{R_{AB}} = \frac{f_{AAAA}6(6pq - 2p - 2q + 1) - (1 - 6pq)f_{AAaa,AaAa,AaaA}}{6pq(1 - p - q)(1 - 3p)}$$

(10)

In the above, $\widehat{R_{AB}}$ is the identity disequilibrium ($\widehat{F_{AB}}$ also for the three branch case). This has never been characterized in terms of cumulants and requires a joint solution of H and ϕ_{XY}. The denominator shows that at least three alleles required in the population, and also that marker frequencies of p_i=1/3 are non-informative.

Identity disequilibrium has classically been characterized as the excess of homozygosity above that expected from the squared gene frequencies, as proposed by HILL (1975) and OHTA (1980). The identity excess is closely correlated to the expectation of the total squared linkage disequilibrium (TAKAHATA 1982). Some of the problem is that haploid gametes are not directly assayed but rather imputed (VITALIS AND COUVET 2001b).

The "classic" procedure for estimating identity disequilibrium involves comparing observed vs. expected double homozygotes. From VITALIS AND COUVET (2001a), one such estimator can be written in the form,

$$\widehat{R_{AB}} = \frac{af_{AAAA} - bf_{AA,aa}}{c}$$

where $a = 6(6pq - 2p - 2q + 1), b = (1 - 6pq),\ c = 6pq(1 - p - q)(1 - 3p)$. The numerator is positive as a is always larger than b, but denominator can be either positive or negative and negative values occurs when b is larger an a. Such is life with four genes. It is also amazing the four-gene frequencies f_{AAAA} and $f_{AA,aa}$ are directly used in an effective two gene estimator, as f_{AAAA} is the observed double homozygotes and $f_{AA,aa}$ represents

homozygotes expected with no zygotic association. Equally amazing is that VITALIS AND COUVET (2001a) and others have given estimators for identity disequilibrium which omits this cumulant.

DISCUSSION AND CONCLUSION

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

Another feature of higher-order relatedness is that the four-gene 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 gene 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.

316 Cumulants do have useful properties for models of quantitative traits, the most
317 important is that the cumulant of the sum of two random variables X+Y is
318 M(X+Y)=M(X)+M(Y); differential equations for models of selection on quantitative traits
319 that involve cumulants are simpler than models involving moments (BURGER 1991; TURELLI
320 AND BARTON 1994). This cumulant will also be key in deriving a marker-based estimator for
321 Q_{st} (RITLAND *in prep*).

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

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

334 In natural populations, individual heterozygosity and fitness are often observed to
335 be correlated in natural populations with inbreeding depression (as opposed to heterosis).
336 Such correlations can be caused by biparental inbreeding combined with loss of fitness
337 with homozygosity (BALLOUX *et al.* 2004). To this end, DAVID *et al.* 2007 presented how the
338 statistic "g_k " measures the covariance of heterozygosity at k loci $E(h_1 h_2 \ldots h_k) = E(h_1)E(h_2)\ldots$

339 E(h_k) (1 + g_k). Levels in natural species are moderate; MILLER AND COLTMAN (2014) assessed
340 the magnitude of ID (as measured by the g^2 statistic) from 50 previous studies, in a list of
341 103 estimates the average g_2 was 0.0577 of which 15 were significantly positive.
342
343 *Normalization constants*
344 Relatedness coefficients are obtained by calculating the pairwise covariance of
345 relatives and dividing it by a normalization constant that converts the covariance into a
346 correlation. This constant is the maximum possible value that the covariance can take. For
347 cases where pairwise comparisons involve the frequency of identical genotypes, it is simple
348 to calculate as a binomial variance. For the two-gene relationship coefficient, at a locus
349 with allele A with frequency p, the maximum covariance between two genotypes is
350 $\mathrm{E}[AA] - \mathrm{E}[A]^2 = p(1\text{-}p)$, as E(AA)=$p$ when R=1. Likewise, the three gene coefficient has a
351 normalization constant of $p(1\text{-}p)(1\text{-}2p)$ for AAA, and the four-gene coefficient is normalized
352 by $p(1\text{-}p)(1\text{-}6p(1\text{-}p))$ for $AAAA$. The normalization constants for combinations of alleles do
353 not need to be specified at the model building stage, but rather naturally appear at various
354 places in the analyses, especially for four genes. Table 1 summarizes the motifs that were
355 normalized at each level of analysis. The four-gene coefficients are particularly complex.
356 ACKERMAN *et al.* (2017) provided a different set of normalization constants for three
357 and four gene measures. In their Equation 6, they normalized the third central moment by
358 the geometric mean gene frequency of the three central moments (rather than by $p(1\text{-}p)(1\text{-}$
359 $2p)$ as done here). Their justification was a similarity of this "third moment correlation" to
360 a bivariate correlation formula. Their normalization constant for the four-gene coefficient
361 (Equation 8) involves a parameter α that mixes the unknown proportions of the two types

362 of higher order identity (all identical vs. 2 pairs identical), resulting in an inference that
363 may be subject to biases.
364
365 *Estimation*
366 Table 2 gives example estimates of *R*, *G* and *H* at a single locus. As expected, the
367 variances decrease with numbers of alleles. While the variances for *R* are reasonable, those
368 for *H* are and the variances for the identity disequilibrium R_{AB} are quite large. Interestingly,
369 and not remarkably, the variances for G are all over the place and its is not even estimate in
370 one case (p=0.5). More details will be provided in a companion paper, in which the
371 estimation properties of method of moments and likelihood estimators are examined with
372 Monte-Carlo simulations (RITLAND 2023).
373 Likelihood requires numerical solutions which introduces complications, as the
374 numerical solution is normally iterated until convergence. When using likelihood, RITLAND
375 (2023) found the number of marker loci needed for adequate convergence was about 20
376 loci for three-gene coefficients and roughly double at 30-50 loci for four gene coefficients.
377 Interestingly, it was found that loci with fewer alleles are more likely to give convergent
378 estimates because the problem with non-convergence arises when relatives do not share
379 the same marker allele.
380 Calculating the probabilities of higher-order relationship poses an interesting set of
381 obstacles. Equation solvers such as Derive or Mathematica help with the derivation and
382 interpretation of with the complicated formulae. The coefficients can also undefined at
383 certain intermediate frequencies (p=1/2 or 1/3) and show high statistical variance about
384 those frequencies (RITLAND 1987).

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.

Different levels of the population (within individuals, between individual within populations, and between populations) were considered by WEIR AND GOUDET (2017), in terms of both pairwise relatedness and population structure. They developed two-gene moment estimators that describe the "relativity" with an explicit reference population. Other two-gene estimators are available in the software package PLINK (PURCELL *et al.* 2007), but this and other packages should incorporate three- and four-gene coefficients, to the extent they are used in the context of the examples of Figures 1 and 2.

Possible approaches in the future

"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). It might be possible to use higher moments to detect such variants. For example, central 4th cumulant of termite sounds 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).

Higher order relatedness can inform topics in systems biology. 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). We have also found that a “covariance of covariance” portrays the interactions of correlations in genetic data.

ACKNOWLEDGMENTS Joe Felsenstein provided many helpful comments in the early stages of this work. This work was supported by ongoing grants from NSERC to KR.

Table 1. Allele states and denominators of cumulants.

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

Table 2. 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).

Array of p	R	G_A	G_B	R_{AB}	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

447

448 FIGURE LEGENDS

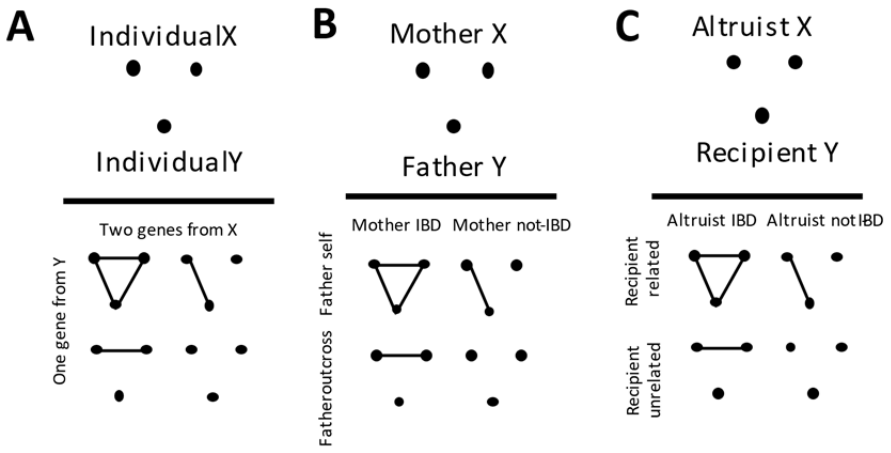
449

450 Figure 1. Two cases where three-gene modes of gene identity are used. (A) Effective selfing
451 model involves two genes of one parent, (B) Altruist-recipient. Identical genes are linked
452 by lines.

453

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

458



459

460 Figure 1

461

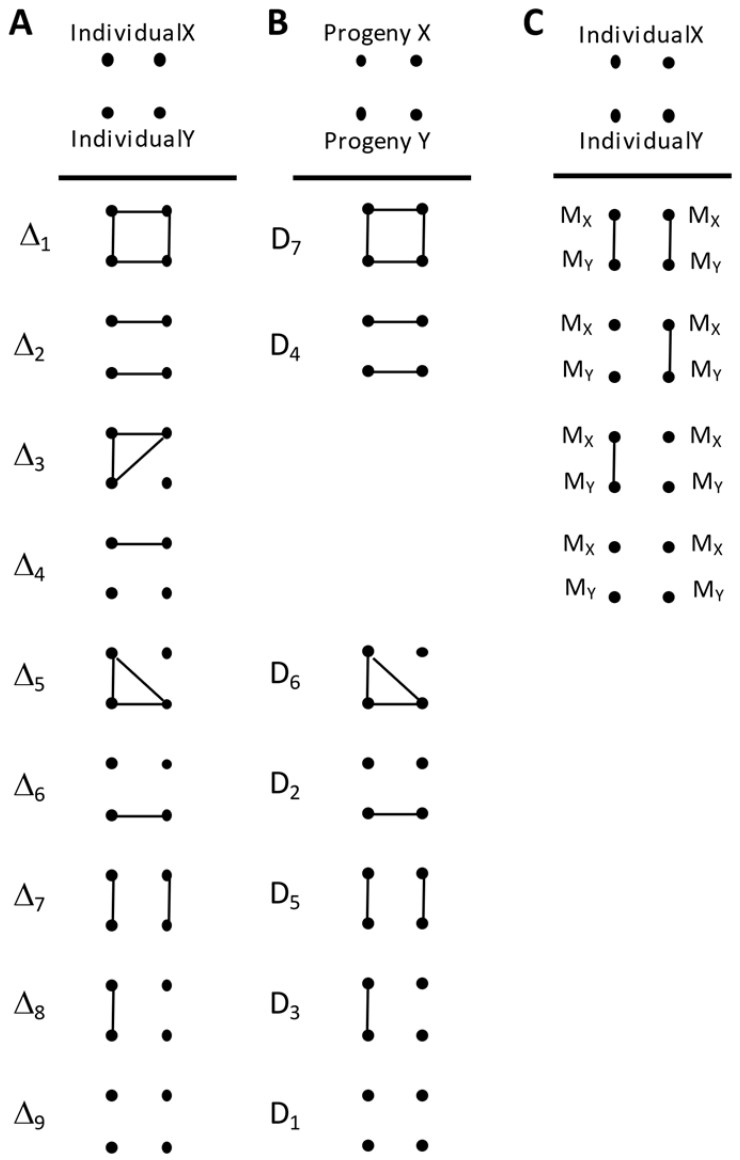

A
IndividualX
IndividualY
Δ1
Δ2
Δ3
Δ4
Δ5
Δ6
Δ7
Δ8
Δ9
B
Progeny X
Progeny Y
D7
D4
D6
D2
D5
D3
D1
C
IndividualX
IndividualY
MX
MY

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.

Balloux, F., W. Amos and T. Coulson, 2004 Does heterozygosity estimate inbreeding in real populations? Molecular Ecology 13.

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.

Miller, J. M., and D. W. Coltman, 2014 Assessment of identity disequilibrium and its relation to empirical heterozygosity fitness correlations: a meta-analysis. Mol Ecol 23**:** 1899-1909.

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., 2023 Estimation of three and four gene coefficients of relatedness. In_Prep.

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., 2009 A Statistical Characterization of the Genetic Structure of Populations, pp. 250. University of North Carolina, North Carolina.

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.

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

Appendix 1. Coefficients that multiply the relationship coefficients , conditioned upon identity mode.

	id mode 1	id mode 2	id mode 3	id mode 4	id mode 5	id mode 6	id mode 7	id mode 8	id mode 9
One	pi**4	pi**3*pj	pi**3*pj	pi**2*pj**2	pi**2*pj**2	pi**2*pj*pk	pi**2*pj*pk	pi**2*pj*pk	pi*pj*pk*pl
F_A	pi**3*qi	pi**2*qi*pj	-pi**3*pj	pi*qi*pj**2	-pi**2*pj**2	pi*pj*pk*qi	-pi**2*pj*pk	-pi**2*pj*pk	-pi*pj*pk*pl
F_B	pi**3*qi	-pi**3*pj	pi**2*qi*pj	pj*qj*pi**2	-pi**2*pj**2	-pi**2*pj*pk	pi*pj*pk*qi	-pi**2*pj*pk	-pi*pj*pk*pl
R	4*pi**3*qi	2*pi**2*pj*(1.-2*pi)	2*pi**2*pj*(1.-2*pi)	-4*pi**2*pj**2	pi*pj*(pi+pj-4*pi*pj)	-4*pi**2*pj*pk	-4*pi**2*pj*pk	pi*pj*pk*(1.-4*pi)	-4*pi*pj*pk*pl
G_A	2*pi**2*qi*(1.-2*pi)	pi*pj*(1.-2*pi)**2	-2*pi**2*pj*(1.-2*pi)	-2*pi*pj**2*(1.-2*pi)	-pi*pj*(pi+pj-4*pi*pj)	-2*pi*pj*pk*(1.-2*pi)	4*pi**2*pj*pk	-pi*pj*pk*(1.-4*pi)	4*pi*pj*pk*pl
G_B	2*pi**2*qi*(1.-2*pi)	-2*pi**2*pj*(1.-2*pi)	pi*pj*(1.-2*pi)**2	-2*pi**2*pj*(1.-2*pj)	-pi*pj*(pi+pj-4*pi*pj)	4*pi**2*pj*pk	-2*pi*pj*pk*(1.-2*pi)	-pi*pj*pk*(1.-4*pi)	4*pi*pj*pk*pl
H	pi*qi*(1.-6*pi*qi)	-pi*pj*(1.-6*pi*qi)	-pi*pj*(1.-6*pi*qi)	-pi*pj*((1.-2*pi)*(1.-2*pj)+2*pi*pj)	-pi*pj*((1.-2*pi)*(1.-2*pj)+2*pi*pj)	2*pi*pj*pk*(1.-3*pi)	2*pi*pj*pk*(1.-3*pi)	2*pi*pj*pk*(1.-3*pi)	-6*pi*pj*pk*pl
F_{AB}	pi**2*qi**2	-pi**2*pj*qi	-pi**2*pj*qi	pi*qi*pj*qj	pi**2*pj**2	-pi*qi*pj*pk	-pi*qi*pj*pk	pi**2*pj*pk	pi*pj*pk*pl
R_{AB}	2*pi**2*qi**2	-2*pi**2*pj*qi	2*pi**2*pj*qi	2*pi**2*pj**2	pi*pj*(qi*qj+pi*pj)	2*pi**2*pj*pk	2*pi**2*pj*pk	-pi*pj*pk*(1.-2*pi)	2*pi*pj*pk*pl

Appendix 2.

The equation for genotypic frequencies in terms of *cumulants* is rewritten to formula in terms of *relationship coefficients*, as follows. From Equation 13 in RITLAND (1987), and for $p_{ij} = \delta_{ij} p_i = \delta_{ij} p_j$, the frequency of the quartet f_{ijkl} is the behemoth

$$\begin{aligned} f_{ijkl} = (2-\delta_{ij})(2-\delta_{kl})\big[p_i p_j p_k p_l + \big(p_i p_k p_{jl} + p_i p_l p_{jk} + p_j p_k p_{il} + p_j p_l p_{ik} - 4p_i p_j p_k p_l\big)R_{AB} \\ + \big(p_{ik} p_{jl} + p_{il} p_{jk} - p_i p_k p_{jl} - p_i p_l p_{jk} - p_j p_k p_{il} - p_j p_l p_{ik} + 2p_i p_j p_k p_l\big)R_{ABAB}. \\ + p_k p_l \big(p_{ij} - p_i p_j\big)F_A + p_i p_j (p_{kl} - p_k p_l)F_B \\ + \big(p_{ij} p_{kl} - p_i p_j p_{kl} - p_k p_l p_{ij} + p_i p_j p_k p_l\big)F_{AB} \\ + 2\big(p_k p_{ijl} + p_l p_{ijk} - p_i p_k p_{jl} - p_i p_l p_{jk} - p_j p_k p_{il} - p_j p_l p_{ik} - 2p_k p_l p_{ij} + 4p_i p_j p_k p_l\big)G_A \\ + 2\big(p_i p_{jkl} + p_j p_{ikl} - p_i p_k p_{jl} - p_i p_l p_{jk} - p_j p_k p_{il} - p_j p_l p_{ik} - 2p_i p_j p_{kl} + 4p_i p_j p_k p_l\big)G_B \\ + (p_{ijkl} - p_i p_{jkl} - p_j p_{ikl} - p_k p_{ijl} - p_l p_{ijk} + 2p_i p_j p_{kl} + 2p_i p_k p_{jl} + 2p_i p_l p_{jk} + 2p_j p_k p_{il} + \\ 2p_j p_l p_{ik} + 2p_k p_l p_{ij} - p_{ij} p_{kl} - p_{ik} p_{jl} - p_{il} p_{jk} - 6p_i p_j p_k p_l)\big]H \end{aligned} \tag{13}$$