Marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods

JINLIANG WANG

Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom

Left running head: J Wang

Right running head: Concepts and estimators of relatedness and inbreeding

Key words: inbreeding coefficient, relatedness, genetic markers, identical by descent

Corresponding author:

Jinliang Wang

Institute of Zoology

Regent's Park

London NW1 4RY

United Kingdom

Tel: 0044 20 74496620

Fax: 0044 20 75862870

Email: jinliang.wang@ioz.ac.uk

Abstract

Inbreeding (*F*) of and relatedness (*r*) between individuals are now routinely calculated from marker data in studies in the fields of quantitative genetics, conservation genetics, forensics, evolution and ecology. Although definable in terms of either correlation coefficient or probability of identity by descent (IBD) relative to a reference, they are better interpreted as correlations in marker-based analyses because the reference in practice is frequently the current sample or population whose *F* and *r* are being estimated. In such situations, negative estimates have a biological meaning, a substantial proportion of the estimates are expected to be negative, and the average estimates are close to zero for *r* and equivalent to F_{IS} for *F*. I show that while current *r* estimators were developed from the IBD-based concept of relatedness, some of them conform to the correlation-based concept of relatedness and some do not. The latter estimators can be modified, however, so that they estimate *r* as a correlation coefficient. I also show that *F* and *r* estimates can be misleading and become biased and marker dependent when a sample containing a high proportion of highly inbred and/or closely related individuals is used as reference. In analyses depending on the comparison between *r* (or *F*) estimates and *a priori* values expected under ideal conditions (e.g. for identifying genealogical relationship), the estimators should be used with caution.

Introduction

Knowledge of the degree of relatedness between individuals due to recent common ancestry is pivotal in many research areas in quantitative genetics, conservation genetics, forensics, evolution and ecology (Ritland, 1996; Lynch & Ritland, 1999). For natural populations in which pedigree records are usually lacking, methods have been proposed (e.g. Lynch, 1988; Queller & Goodnight, 1989; Li *et al*., 1993; Ritland, 1996; Lynch & Ritland, 1999; Wang, 2002; Thomas, 2010) and applied to estimating the genetic relatedness between a pair of individuals from their genotypes at marker loci. These simple estimators, based on allele frequency moments, were shown to provide unbiased albeit imprecise estimates of relatedness from a typical suit of microsatellite markers when the assumptions made in developing them were met (e.g. Lynch & Ritland, 1999; Van De Casteele *et al.*, 2001; Wang, 2002). Several likelihood estimators (Milligan, 2003; Wang, 2007; Anderson & Weir, 2007) were also proposed to estimate relatedness in more complicated situations involving inbred or structured populations and imperfect markers suffering from genotyping errors and mutations. Constraining estimates to their "legitimate" range of [0, 1], these likelihood estimators are biased but can be more precise than moment estimators in certain situations.

Relatedness (r) and inbreeding (F) have by definition an implicit reference population in which all homologous genes within and between individuals are assumed to be not identical by descent (IBD). Equivalently, the reference population is assumed to consist of non-inbred and unrelated individuals. The relatedness between and inbreeding of individuals are thus measured *relative* to this reference. In a pedigree based analysis in practice, founders who have no known parents included in the pedigree are assumed non-inbred and unrelated, and thus act effectively as reference although they may come from different generations. Relatedness between and inbreeding of any individuals in the pedigree are calculated relative to this reference by path analysis (Wright, 1922) or a recursive tabular method (Emik & Terrill, 1949). If the reference is moved a few generations backward into the past because the ancestors of some or all of the original founders are made known and used as founders, then some relatedness between and inbreeding of individuals will be increased. If the reference is moved a few generations forward because we are only interested in the most recent coalescences, then some relatedness between and inbreeding of individuals will be decreased. When we know the differentiation (F_{ST}) of the new reference relative to the old one, we can use it to adjust our estimates of relatedness and inbreeding calculated using the old reference so that they are relative to the new reference (Powell *et al*., 2010). However, not all relatedness and inbreeding coefficients are equally affected by a change of reference, and this F_{ST} based correction procedure works only as an approximation.

In a marker based analysis, r and F estimators are also defined and calculated *relative* to an underlying reference population (Anderson & Weir, 2007; Wang, 2011). In addition to the assumption of non-inbred and unrelated individuals in the reference, marker based r and F estimators assume that the marker allele frequencies in the reference are known. Strictly under these assumptions, various moment estimators mentioned above are truly unbiased, as checked by simulations (e.g. Van De Casteele *et al.*, 2001; Wang, 2002) and verified rigorously by analytical treatments (Wang, 2011). For example, the estimators give an average relatedness of 0.5 and 0.25 for non-inbred diploid full and half siblings respectively, when the allele frequencies used in simulating the genotypes of unrelated and non-inbred parents of the sampled individuals are assumed known and used in the estimation. In practice, however, allele frequencies of a population are rarely known and have to be estimated from a sample of individuals. With few exceptions as verified by a survey of the literature, a sample of individuals is used first for estimating allele frequencies assuming $r=F=0$, and then for estimating r and F using the estimated allele frequencies. This practice effectively assumes *a priori* non-inbred and unrelated individuals in the sample, which is used actually as reference. In such a situation, what do r and F measure by definition? What are the marker-based estimators actually estimating?

In this study, I will first clarify the definitions of relatedness and inbreeding when a sample of individuals is used for estimating both allele frequencies and *F* and *r*. This is important in understanding what *F* and *r* really mean, in answering elementary questions such as whether or not negative *F* and *r* values make biological sense and whether or not an individual with *F*=-0.1 is more inbred than an individual with *F*=-0.2. Clarifying the definitions is also important in designing properly an experiment for *r* and *F* analysis, in interpreting and applying *r* and *F* estimates correctly in downstream analyses, and in developing and comparing rightly different estimators. I will then investigate, by analytical and simulation analyses, the properties of several *r* and *F* moment estimators in the realistic situation of using the current sample or population as reference. I will modify several *r* estimators so that they estimate what are supposed to be estimating in the case of a sample being used as reference. Hereafter, I focus on the simple *r* and *F* estimators that are based on marker allele frequency moments, and the term "estimators" implicitly refer to these moment estimators except when explicitly preceded by the word "likelihood".

Definitions of *r* and *F*

The concept of inbreeding coefficient of an individual, *F*, was developed by Wright (1921). It was defined as the correlation between homologous genes of the two gametes (one from father and one from mother) uniting to form the individual, relative to the total array of such gametes in random derivatives of the foundation stock (or reference population). Later, Malecot (1948) introduced another definition of *F* as the probability of identity by descent (IBD) of the two homologous genes at a locus within an individual, where IBD is counted with respect to the reference population in which all homologous genes are assumed non-IBD. Genes IBD are copies of the same ancestral allele, and are thus identical in state (IIS) barring the rare events of mutations.

In both the correlation and IBD definitions, the *F* value of an individual is independent of locus specific properties such as the mutation rate and the number and frequencies of alleles at a locus, and is determined solely by the genealogical relationship or the shared ancestry of the individual's parents (Wright, 1965). Indeed, *F* is traditionally calculated by path analysis (Wright, 1922) of a pedigree without referring to any locus at all. For a given individual, all loci are expected to have the same *F* value because they have experienced the same genealogical process. For the same reason, different individuals with exactly the same pedigree (e.g. full siblings and twins) are also expected to have the same *F* value at any locus. Therefore, an individual's *F* value calculated from the pedigree or estimated (learned) from some marker loci can be used to make inference or explain observations at any loci, taking into account of locus specific properties (like mutations, selection, mistyping) of the latter loci if necessary.

Wright (1965) and others (e.g. Seger, 1981; Grafen, 1985) noted that the correlation and IBD concepts of *F* are identical in some cases, when the reference is a suitable population ancestral to the current population. They also pointed that, however, the correlation concept is more general than the IBD concept, and can give meaningful negative values in some situations. For example, the F1 hybrid individuals from crossing two differentiated parental populations will be expected to have a negative *F*, no matter the reference is the two parental populations combined or the current hybrid population. In a large population with mixed random selfing and outcrossing, the outbred individuals will have a negative *F* when the current population is used as reference. Similarly, for a population in which consanguine mating is avoided, individual *F* will tend to be negative on average if the current population is used as reference. These negative *F* values make biological sense, signifying that the probability of the two homologous genes within an individual being IBD is smaller than that of two homologous genes drawn at random from the reference population. In contrast, the IBD concept will never give a negative *F*, because it is a probability.

In principle, the correlation concept puts no constraint on which population can be used as a reference. One can use an ancestral, the current (focal), and even a descendant population as a reference, yielding in general a decreasing *F* value for a given individual. Pedigree analyses invariably use an ancestral population as reference, while marker analyses frequently use the current population from which a sample of individuals is taken for *F* analysis as the actual reference. There is neither methodological nor conceptual difficulty in using a descendant population as the reference in a marker-based analysis. In contrast, the IBD based *F* has to use an ancestral population as reference, because by definition negative values are prohibited and have no meaning. If the current or a descendent population were used as reference, the *F* of most or all individuals would be invariably zero.

The necessary but ambiguous and arbitrary nature of a reference in both the correlation and IBD concepts of *F* dictates that *F* values are always relative to an implicit reference population assumed to be composed of non-inbred and unrelated individuals such that all homologous genes in the reference are non-IBD. For any given individual, *F* can virtually take any value in the legitimate range [-1, 1] as a correlation coefficient, or in the range [0, 1] as an IBD coefficient, depending on the reference one chooses to measure *F* against. This relativity leads to the claims that *F* has something arbitrary in its definition (Maynard Smith, 1998, p141), to the so-called ‘inbreeding paradox’ (Seger, 1981), and the suggestion that relatedness (and *F* as well) is a measure of our information and not of anything real (Jacquard, 1974, p171). These claims are true to some extent, but they do not nullify the usefulness of *F* in population genetics theory and applications. So long as the reference is not extremely far away from the current population such that mutations and

selections become non-negligible compared with the genealogical process (inbreeding and drift), the *F* values suffice in most analyses such as regression and correlation analyses involving *F* as a variable. In these analyses, it is the relative *F* values of different individuals that matter and a linear transformation of *F* values does not alter the regression or correlation analysis result. For pedigree based analyses, however, a pedigree that is too deep or too shallow (i.e. the reference is too far away from and too close to the current population, respectively) will lead to *F* values close to 1 or 0, respectively, for all current individuals. Consequently, the variance of *F* would become much smaller than the maximum obtainable from a pedigree with an appropriate depth, resulting in under- or over-estimation of inbreeding effects in regression or correlation analyses. In contrast, marker based analyses are affected only when the reference is too far away into the past, and are little affected when the reference is or is close to the current population.

There are other definitions of *F* in the literature. Rousset (2002) noted the limitations of IBD-based concept of inbreeding, and gave a generic definition of *F* as ratios of differences of probabilities of genes identical in state (IIS). In ideal situations (e.g. the absence of locus specifics like mutations), it is equivalent to Wright's correlation definition when applied to markers. However, several difficulties arise with this IIS based definition. First, gene identities and thus IIS are more or less arbitrary. For example, classical genetics recognizes three alleles, A, B, and O that determine the compatibility of blood transfusions at the gene locus for the ABO blood type carbohydrate antigens in humans. It is now recognized that each of the three alleles is actually a class of multiple alleles having different DNA sequences and coding for different proteins with identical properties. More than 70 alleles are now identified at the ABO locus (Yip, 2002). A homozygote in the old 3-allele system may well be a heterozygote in the new +70-allele system, causing a huge drop in homozygosity or probability of IIS in an individual or a population. In contrast, *F* defined as correlation or IBD probability due to shared ancestry is unaffected by how alleles and loci are defined, and by the polymorphisms of markers. Second, the definition is not applicable to pedigree analysis. The IBD and correlation definitions of *F* are broad and coherent, and apply to both pedigree and marker analyses. Using the founders of a pedigree as reference, pedigree and markers should yield the same expected value of *F* for a given individual. These definitions make it possible to develop likelihood or Bayesian methods to use pedigree and marker data jointly in inferring realised (rather than expected) *F* and relatedness, and in estimating marker genotypes and allele frequencies from incomplete pedigree and marker information (e.g. Boehnke, 1991; Wang & Santure, 2009). Third, IIS based *F* depends not only on genealogy, but also on locus specifics. As a result, the expected *F* value of a given individual varies across loci, depending on locus specific properties like mutation rate and mistyping rate. In general, effects of mutations can

be negligibly small (Rousset, 2002), because in practice the time scale for F is usually much smaller than $1/u$ where u is the mutation rate. However, other locus properties may have a substantial effect on IIS and thus on F. In the imperfect world, genotyping errors are a rule rather than an exception (Bonin *et al*., 2004). Allelic dropouts and null alleles are particular common for microsatellite markers, and could cause an apparent increase in IIS and thus IIS-based F. It is true such mistypings can affect marker-based estimates of F in any concepts. However, under the correlation or IBD definition, F has the same expected value across loci such that a method can be developed to account for mistypings if the model and rate of their occurrences are known (e.g. Wang, 2007).

Closely related to F is the concept of coancestry coefficient or the coefficient of kinship, θ, between two individuals. In Wright's correlation definition, θ between two individuals is simply equal to the expected F of their (hypothetical) offspring, and F can be regarded as the coancestry coefficient between the male and female gametes that unite to form an individual. In terms of IBD, θ is the probability that two homologous genes, one taken at random from each individual, are identical by descent. Relatedness, r, is simply $r=2\theta$ if both individuals are non-inbred (Lynch & Ritland, 1999).

It is noticeable that most marker based r estimators are developed based on the IBD concept (e.g. Lynch, 1988; Li *et al*., 1993; Ritland, 1996; Lynch & Ritland, 1999; Wang, 2002; Thomas, 2010; Milligan, 2003; Wang, 2007; Anderson & Weir, 2007), using the full set or a subset of the 9 condensed IBD states for the 4 (2 in each individual) homologous genes and their probabilities (Harris, 1964; Jacquard, 1972). These estimators implicitly assume an appropriate ancestral population as the reference, and allele frequencies from the reference are known and are used in calculating the estimators. When these assumptions are met, these estimators are unbiased as checked by both simulations (e.g. Lynch & Ritland, 1999; Wang, 2002) and rigorous analytical treatments (Anderson & Weir, 2007; Wang, 2011). Negative values from the estimators are taken as due to sampling errors (e.g. Lynch & Ritland, 1999). In a similar vein, likelihood estimators (Milligan, 2003; Wang, 2007; Anderson & Weir, 2007) of r are constrained in the "legitimate" range of [0,1] based on the IBD concept, and as a result are upwardly biased when the assumptions are violated.

In practical applications, however, r and F are frequently estimated using allele frequencies calculated from the current sample of individuals whose F and r are being estimated. This practice effectively uses the current population (or sample) as the reference. A shift of reference from an ancestral population assumed in developing the estimators to the current population (from which the individuals are sampled) or sample assumed in applying the estimators alters imperceptibly and

insidiously the meanings of *r* and *F*. The estimates thus obtained can no longer be interpreted as probabilities of IBD of homologous genes between and within individuals relative to the reference, as is in developing the estimators. Rather, they should be understood as correlations of homologous genes between and within individuals (Hardy & Vekemans, 1999; Powell *et al*., 2010) due to shared ancestry, as Wright (1921) originally conceived. The shift in reference to the current sample causes some *F* values of and some *r* values between individuals to be legitimately negative, and so they obviously cannot be interpreted as probabilities and should not be simply dismissed as due to sampling errors. They can be understood, however, as the correlation of homologous alleles within and between individuals. The negative values imply that homologous genes within and between individuals are IIS at a lower probability than the average, because the shared ancestors are more distant or/and fewer than the average.

Using the current sample as reference, *r* (or *F*) signifies the expected relative excess (when positive) or deficit (when negative) of the occurrences of homologous genes that are IIS between (or within) individuals due to the relative excess or deficit of shared ancestry. The mean estimate of *r* among pairs of individuals in a sample should be close to zero, because the probability of IBD of homologous genes between individuals is on average close to that of homologous genes taken at random from the sample except when it is extremely small. The mean estimate of *F* among individuals in a sample should be equivalent to Wright's F_{IS} by definition. Given the frequency of an allele, *p*, at a locus in the sample, an individual *i* with inbreeding coefficient F_i will be homozygous for the allele at a probability of $pF_i + p^2(1 - F_i)$. This probability is smaller than, equal to, and larger than the mean, p^2, when the individual has a negative, zero, and positive F_i, respectively. This interpretation of *F* is true across loci. For example, the probability of a multilocus homozygote for individual *i* is $\prod_{l=1}^{L}(p_l F_i + p_l^2(1 - F_i))$, where p_l is frequency of the allele at locus *l* (=1, …, *L*) that is homozygous for the individual. This interpretation of *F* is also true among individuals. For example, the frequency of a homozygote for an allele of frequency *p* in the sample is $\frac{1}{n}(\sum_{i=1}^{n}(pF_i + p^2(1 - F_i)))$, which reduces to p^2 because the average of F_i is zero in the sample of *n* individuals. Relatedness has a similar explanation.

Estimators of *r* and *F*

As shown above, *r* (or *F*) should be interpreted as correlations and should have an expected value that is equal or close to 0 (or F_{IS}) irrespective of the genealogy of the sample, when the current population or sample is actually used as the reference. Is this true with the estimators used currently in practical applications? Below I show by analytical and simulation approaches that while some *r* estimators can be construed as correlation coefficient, others are not. In the latter case, however, the

253 estimators can be modified so that they estimate *r* as a correlation coefficient. In contrast, all current
254 estimators of *F* can be interpreted as correlation coefficient.

255 I assume a single marker locus with *k* (>1) codominant alleles, A_i (*i*=1~*k*), is used in
256 estimating the *r* and *F* of a large sample of individuals taken from a half-sib family (The same
257 results are obtained from a full-sib family, and the derivations are available upon request). All
258 individuals in the sample share the same non-inbred parent of one sex but have distinctive non-
259 inbred and unrelated parents of the other sex. Both *r* and *F* can be defined and estimated using
260 either parental or current population as reference. In the former case, individuals in the reference are
261 non-inbred and unrelated, and the frequency of allele A_i, $\hat{p}_i$, used in calculating *r* and *F* is the
262 parental allele frequency p_i assumed known without error. In the latter case, individuals in the
263 reference are non-inbred half siblings, and $\hat{p}_i$ used in calculating *r* and *F* is estimated using the
264 genotypes of sampled individuals under the assumption of non-inbred and unrelatedness.

265 ***Relatedness estimators***

266 By the IBD or correlation definition using the parental population as reference, we have an expected
267 value of r=0.25 for each pair of individuals, and $\bar{r}$=0.25 across pairs. By the correlation definition
268 using the current population (sample) as reference, we have an expected value of r=0 for each pair
269 of individuals, and $\bar{r}$=0 across pairs. In the following, I investigate whether $\bar{r}$=0 is obtained from
270 each of a number of estimators when the current population is used as reference.

271 *Estimator by Queller and Goodnight (1989)*: There are a number of variants to this widely applied
272 estimator (denoted as QG), and I choose to use the symmetric one obtained by averaging the
273 estimates using each of the two individuals as reference. For individuals X and Y with genotypes
274 $\{a,b\}$ and $\{c,d\}$, respectively, at a locus (note that alleles A_i for *i*=1~*k* are denoted by *a*, *b*, *c*, *d* to
275 avoid subscripts), the estimator is

276 $$\hat{r} = (\hat{r}_{XY} + \hat{r}_{YX})/2, \quad (1)$$

277 where estimates using individual X and Y as references are

278 $$\hat{r}_{YX} = \hat{r}[c,d|a,b] = \frac{\delta_{ac}+\delta_{ad}+\delta_{bc}+\delta_{bd}-2(p_a+p_b)}{2(1+\delta_{ab}-p_a-p_b)}, \quad (2)$$

279 $$\hat{r}_{XY} = \hat{r}[a,b|c,d] = \frac{\delta_{ac}+\delta_{ad}+\delta_{bc}+\delta_{bd}-2(p_c+p_d)}{2(1+\delta_{cd}-p_c-p_d)}, \quad (3)$$

280 respectively, and the Kronecker delta variable δ_{ij} =1 if $i = j$ and $\delta_{ij} = 0$ otherwise. In some special
281 cases, equations (1-3) are undefined. For a monomorphic marker (*k*=1) or a biallelic marker (*k*=2)

with both X and Y being heterozygous, both (2) and (3) are undefined and as a results (1) is also undefined. In such a case, $\hat{r}$ is taken more or less arbitrarily as zero. When X and Y are a heterozygote and homozygote, respectively, at a biallelic locus, (2) is undefined and the estimator becomes $\hat{r} = \hat{r}_{XY}$. Similarly $\hat{r} = \hat{r}_{YX}$ when Y and X are a heterozygote and homozygote at a biallelic locus, respectively.

Under random mating, the genotypes of half siblings in the sample depend on the genotype of the shared parent, G_s, and allele frequencies of the parental population. G_s can be either a homozygote, $\{a,a\}$, or a heterozygote, $\{a,b\}$ ($a \neq b$). In the former case, the sibling genotypes are $\{a,x\}$, where $x=a, b, \ldots$, with a probability of p_x. The allele frequency calculated from the sample assuming outbred and unrelated individuals is $\hat{p}_x = (\delta_{ax} + p_x)/2$, where $\delta_{ax} = 1$ if $x = a$ and $\delta_{ax} = 0$ otherwise. Given $G_s = \{a,a\}$, the average relatedness between individuals of the sample is $\bar{r} = \sum_{b=1}^{k}\sum_{d=1}^{k} p_b p_d(\hat{r}[a,b|a,d] + \hat{r}[a,d|a,b])/2$. Substituting $\hat{r}$ by (2-3) and using sample allele frequencies $\hat{p}_x$ in place of p_x in the estimator, I obtain $\bar{r} \equiv 0$ for $k > 2$, and $\bar{r} \equiv -p_a^2$ for $k=2$.

Similarly, when the shared parent has a heterozygous genotype $G_s = \{a,b\}$ ($a \neq b$), the offspring genotypes, their frequencies, and the sample allele frequencies are listed in Table 1. Following the approach above, I obtain $\bar{r} \equiv 0$ for $k > 2$, and $\bar{r} \equiv (12p_1p_2 - 3)/(4p_1p_2 + 3)$ for $k=2$, when allele frequencies calculated from the sample assuming unrelated and non-inbred individuals are used in the estimation.

In summary, when the current population (sample) is used as reference (i.e. the allele frequencies estimated from the sample are used in *r* estimation), the average *r* between half siblings is zero, except when $k=2$. For a biallelic locus ($k=2$), $\bar{r} = 0$ only in the special case of a heterozygote of the shared parent and equal allele frequencies (i.e. $p_1=p_2=0.5$); otherwise, $\bar{r} < 0$. The negative $\bar{r}$ when $k=2$ occurs because the estimator is undefined with a heterozygous reference individual, and is set, more or less arbitrarily, a value of 0.

Estimator by Ritland (1996): This estimator (denoted as R), derived by Li & Horvitz (1953) and *Ritland* (1996), is

$$\hat{r} = \frac{2}{k-1}\left[\left(\sum_{i=1}^{k}\frac{S_i}{p_i}\right) - 1\right], \quad (4)$$

where S_i gives the similarity for allele *i* between individuals X and Y. S_i has 4 possible values, which are 0, 0.25 and 1 when both X and Y have exactly 0, 1 and 2 *i* alleles, and 0.5 when X and Y have a total of 3 *i* alleles.

Using the genotype and estimated allele frequencies of half sib families listed in Table 1, the estimator always gives an average relatedness of 0, irrespective of the genotype of the shared parent and the number and frequencies of alleles at a locus.

Estimator by Lynch and Ritland (1999): The estimator (denoted as LR) of relatedness between individuals X and Y with genotypes {a,b} and {c,d} respectively is given by (1), where the estimates using X and Y as references are

$$\hat{r}_{YX} = \hat{r}[c,d|a,b] = \frac{p_a(\delta_{bc}+\delta_{bd})+p_b(\delta_{ac}+\delta_{ad})-4p_ap_b}{(1+\delta_{ab})(p_a+p_b)-4p_ap_b}, \quad (5)$$

$$\hat{r}_{XY} = \hat{r}[a,b|c,d] = \frac{p_c(\delta_{da}+\delta_{db})+p_d(\delta_{ca}+\delta_{cb})-4p_cp_d}{(1+\delta_{cd})(p_c+p_d)-4p_cp_d}, \quad (6)$$

respectively. Applying the estimator to a large half sib family as listed in Table 1 yields an average relatedness of 0, irrespective of the genotype of the shared parent, except for the special case of a biallelic locus with equal allele frequencies. In this special case, the LR estimator becomes undefined when the reference individual is a heterozygote (Lynch & Ritland, 1999).

Estimator by Lynch (1988) and Li et al. (1993): This estimator (denoted as LL), proposed by Lynch (1988) and improved by Li *et al*. (1993), estimates *r* using a similarity index S_{XY}. This index is defined as the average fraction of alleles at a locus in a reference individual, *X* or *Y*, for which there is another allele in the other individual, *Y* or *X*, that is IIS. Thus, S_{XY} has a value of 1 for genotype pairs {A_iA_i, A_iA_i} or {A_iA_j, A_iA_j}, 0.75 for {A_iA_i, A_iA_j}, 0.5 for {A_iA_j, A_iA_k}, and 0 for {A_iA_j, A_kA_l}, where different subscripts *i, j, k, l* indicate distinctive alleles. The estimator for individuals X and Y is

$$\hat{r} = \frac{S_{XY}-S_0}{1-S_0}, \quad (7)$$

where $S_0 = 2a_2 - a_3$ (with $a_m = \sum_{i=1}^{n} p_i^m$ for *m* = 2, 3) is the expected similarity index for unrelated individuals.

Applying the estimator to a large half-sib family (Table 1), I obtain, after tedious algebra, an average relatedness $\bar{r}[i,i] = \frac{1-p_i-p_i^2+a_3}{5-5p_i+3p_i^2-4a_2+a_3}$ and $\bar{r}[i,j] = \frac{1-(p_i+p_j)-2(p_i^2+p_j^2)+4a_3}{25-13(p_i+p_j)+6(p_i^2+p_j^2)-16a_2+4a_3}$ when the shared parent has a homozygote genotype {A_i,A_i} and a heterozygote genotype {A_i,A_j} ($j \neq i$=1~k), respectively. It can be shown that $\bar{r} > 0$ in both cases, and the magnitude depends on the number and frequencies of alleles. This means that LL estimator does not estimate *r* as a correlation

coefficient when the current sample (population) is used as reference. Otherwise, the expected *r* should be zero, like the QG, R, and LR estimators.

To understand how much the LL estimator deviates from the expected value of $\bar{r} = 0$ if it were a correlation estimator, let's consider the simple case of a biallelic locus. Combining the three possible genotypes of the shared parent, I obtain an overall average relatedness of $\bar{r} = \sum_{i=1}^{2} p_i^2 \bar{r}[i,i] + 2p_1p_2\bar{r}[1,2]$, which simplifies to $\bar{r} = \frac{p_1p_2(7-4p_1p_2)}{(1+p_1)(1+p_2)(1+2p_1)(1+2p_2)}$. Figure 1 plots $\bar{r}$ as a function of allele frequency p_1 (=1-p_2), and shows simulation values for comparison. As expected, simulation and analytical values agree very well. Except when allele frequency is close to zero or one such that the marker gives little information, $\bar{r}$ is substantially higher than 0. The maximal value of $\bar{r}$ is 1/6 when $p_1 = p_2 = 0.5$. It is clear that the LL estimator applies to the IBD definition of relatedness only, and becomes meaningless when the current sample contains a high proportion of related individuals and is used as the reference because in such a case the estimates depend heavily on allele frequencies. It also implies that LL relatedness estimates for pairs of individuals are incomparable if these individuals have missing data at different loci.

It is possible to modify LL estimator so that, like QG, LR and R estimators, it applies to the more general definition of relatedness in terms of correlation (Wright, 1921). The original LL estimator is calculated using a constant S_0, which is the *expected* similarity for unrelated individuals. For a reference population (such as an appropriate ancestral population) of non-inbred and unrelated individuals, S_0 can be calculated as $S_0 = 2a_2 - a_3$ from allele frequencies. For a more general reference that may contain related and inbred individuals, S_0 should be replaced by the average *observed* similarity over all possible pairs of individuals, S_a. When the reference is a large random mating ancestral population as assumed in deriving the LL estimator, we have $S_a = \sum_{a=1}^{k}\sum_{b=1}^{k}\sum_{c=1}^{k}\sum_{d=1}^{k} p_a p_b p_c p_d S_{(a,b),(c,d)}$ at a locus with *k* codominant alleles, where $S_{(a,b),(c,d)}$ is the same as S_{XY} in (7) and denotes the similarity index for a genotype {a,b} and a genotype {c,d}. It can be shown, after some algebra, that S_a reduces to $S_0 = 2a_2 - a_3$ as expected. When the reference is the current sample of *n* individuals being calculated for relatedness, then

$$S_a = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} S_{ij}, \quad (8)$$

where S_{ij} is defined similarly to S_{XY} in (7).

Replacing S_0 by S_a, (7) gives relatedness estimates relative to a reference chosen by a researcher. When the reference is an ancestral, the current, and a descendent population, the average

relatedness across pairs of individuals in a sample tends to be greater than, equal to, and smaller than zero respectively, independent of markers and their allele frequencies.

Consider the half sib family listed in Table 1 as an example. When the shared parent has a homozygote genotype $\{A_i,A_i\}$ at a locus with k alleles, the half siblings have an average observed similarity index $S_a = \sum_{j=1}^{k}\sum_{l=1}^{k} p_j\, p_l(1+\delta_{jl})/2$ which, after some algebra, reduces to $S_a = (1 + a_2)/2$. The average relatedness is $\bar{r} = \sum_{j=1}^{k}\sum_{l=1}^{k} p_j\, p_l \left(\frac{1+\delta_{jl}}{2} - S_a\right)/(1 - S_a)$, which reduces to $\bar{r} \equiv 0$. It can be shown similarly that $\bar{r} \equiv 0$ when the shared parent has a heterozygote genotype $\{A_i,A_j\}$ ($j \neq i$).

Estimator by Wang (2002): This estimator (denoted by W) uses the similarity index of Lynch (1988) and Li *et al*. (1993) but can estimate both two- and four-gene relatedness, and thus the total relatedness r. Using the same similarity index as LL estimator, W estimator is similar to LL estimator and applies to the IBD definition of relatedness only. When the current sample is used as reference, W estimator gives an average relatedness larger than 0 when relatives are included in the sample. However, unlike LL estimator, W estimator is complicated and it is difficult to derive its $\bar{r}$ even for the simple case of a sample of individuals having the same relationship, such as a half siblings. Simulations showed that W estimator has a $\bar{r}$ similar to LL estimator, as shown in Figure 1 for a biallelic locus.

To modify W estimator such that it is relative to a reference no matter the reference is an ancestral or current population (sample), I transform the original 2- or 4-gene relatedness or total relatedness estimates, w, from W estimator to $(w - \bar{w})/(1 - \bar{w})$, where $\bar{w}$ is the average of the original estimates across all dyads.

Inbreeding estimators

In the IBD or correlation definition using the parental population as reference, we have an expected value of F=0 for each individual in the sample and thus $\bar{F}$=0. In the correlation definition using the current population (sample) as reference, we have an expected value of F<0 for each individual and thus $\bar{F}$<0 because the two homologous genes within an individual have a lower IBD probability than two genes taken at random from the sample (i.e. individuals are more heterozygous than expected at Hardy-Weinberg equilibrium, F_{IS}<0).

A number of estimators (Li & Horvitz, 1953; Ritland, 1996; Wang, 2011) have been developed to estimate F from marker data. Herein I choose to analyze a few. I show that these estimators estimate F as a correlation coefficient (Wright, 1921), and the average F among

individuals is expected to be smaller than zero when the current sample (population) containing highly related individuals is used as reference. However, these estimators may give misleading results in such a case because the estimates become dependent on allele frequencies of the markers.

Estimator by Li & Horvitz (1953) and Ritland (1996): This estimator (denoted as LHR) was derived based on the proportion of alleles in homozygous condition at a single locus, $\sum_{i=1}^{k}\frac{z_{ii}}{p_i} = 1 + F(k-1)$, where $z_{ii} = (1-F)p_i^2 + Fp_i$ is the proportion of homozygotes for allele A_i and p_i is the frequency of allele A_i. In the expression for z_{ii}, F can be interpreted as correlation and can take a negative value for an individual having less homozygosity than an individual expected in the reference population under Hardy-Weinberg equilibrium. Solving for F gives an estimator

$$F = \frac{1}{k-1}\sum_{i=1}^{k}\frac{S_i - p_i^2}{p_i}, \quad (9)$$

where S_i =1 if the individual is homozygous for allele i and S_i = 0 if otherwise. For the half sib family considered in Table 1, all individuals have an expected F=0 because their parents are unrelated. Estimator (9) gives indeed F=0 when the allele frequencies of the parental population are known without error and are used in the estimation. For a shared parent with a homozygous $\{A_i,A_i\}$ and heterozygous $\{A_i,A_j\}$ genotype, the averages of individual F values calculated by (9) are $\frac{-1}{k-1}(1-p_i) + \frac{\frac{1}{p_i}-1}{k-1}(p_i)$ and $\frac{-1}{k-1}\left(1-\frac{p_i}{2}-\frac{p_j}{2}\right) + \frac{\frac{1}{p_i}-1}{k-1}\left(\frac{p_i}{2}\right) + \frac{\frac{1}{p_j}-1}{k-1}\left(\frac{p_j}{2}\right)$, respectively. Both reduce to zero as expected, regardless of the number and frequencies of alleles at a locus.

However, when the observed allele frequencies in the sample are used in the estimation, (9) gives $F = \frac{-(1-p_i)}{(k-1)(1+p_i)}$ and $F = \frac{-(1-4p_ip_j)}{(k-1)(1+2p_i)(1+2p_j)}$ when the shared parent is a homozygote $\{A_i,A_i\}$ and heterozygote $\{A_i,A_j\}$, respectively. In both cases $F < 0$ in general, and $F = 0$ only when the shared parent has a heterozygous genotype at a biallelic locus with equal allele frequencies. Figure 2 plots the average F when the shared parent has a homozygous and heterozygous genotype, and has the two kinds of genotypes at frequencies under Hardy-Weinberg equilibrium. As is clear, F is negative in general, and its magnitude depends on parental allele frequencies. This means different markers with different numbers and frequencies of alleles will yield different expected F estimates. This negative and marker-dependent F is caused by using allele frequencies calculated from the current sample which is assumed to contain unrelated individuals.

Estimator by Li & Horvitz (1953) and Carothers et al. (2006): This estimator (denoted as LHC), based on the consideration of expected heterozygosity h, is

$$\hat{F} = \frac{h-1+S}{h}, \quad (10)$$

where $S = 1$ if the individual is a homozygote and $S = 0$ if otherwise. Similar to (9), (10) is an unbiased estimator of F as a correlation coefficient when individuals in the reference population are non-inbred and unrelated (Carothers *et al*., 2006). If some individuals in the reference are related, however, the expected value of (10) is greater and smaller than zero when the actual inbreeding is higher and lower than average relatedness in the reference, respectively. With a significant level of relatedness among individuals in the reference, (10) becomes marker dependent and does not reflect purely the level of inbreeding.

Consider the half sib case of Table 1 and use the current population (sample) as reference. When the shared parent is a homozygote, $\{A_i,A_i\}$, and heterozygote, $\{A_i,A_j\}$, the expected heterozygosity of the sample can be obtained from Table 1 as $h = (3 - 2p_i - a_2)/4$ and $h = (7 - 2p_i - 2p_j - 2a_2)/8$, respectively. Using these and (10), I obtain the average F of the sample

$$\bar{F} = -\sum_{i=1}^{k} \frac{p_i^2(1+a_2-2p_i)}{3-a_2-2p_i} - \sum_{i=1}^{k}\sum_{j=i+1}^{k} \frac{2p_ip_j(1+2a_2-2p_i-2p_j)}{7-2a_2-2p_i-2p_j}.$$

For a biallelic locus, this is identical to the average F from estimator (9). For a locus with k equifrequent alleles, the average F values calculated by (10) and (9) are plotted as a function of k in Figure 3. As can be seen, both estimators are negative and marker-dependent when the current sample containing related individuals is used as reference.

The magnitude of *r* and *F* values

The above analytical treatment considered a sample containing a single large family, and all sampled individuals have the same expected inbreeding and relatedness. When a sample containing individuals of variable relatedness and inbreeding coefficients is used as reference, the magnitude of r and F estimates should be taken with caution, because they are not determined purely by the actual relatedness between and inbreeding of individuals involved, but also dependent on the actual relatedness and inbreeding of other individuals in the sample, and may also be affected by the allele frequencies of markers.

Let's consider a simple example. Suppose a sample containing N individuals taken at random from n half-sib families in a population, with each family contributing $m=N/n$ (integer) half siblings who share the same father but have distinctive mothers. All parents of the half sib families are non-inbred and unrelated. When the current sample is used as reference (i.e. its allele frequencies are calculated assuming $F=r=0$ and used in the estimation), the average estimated

459 relatedness $q\bar{r}_{hs} + (1-q)\bar{r}_{ns} = 0$, where $q = \frac{nm(m-1)/2}{N(N-1)/2}$ is the proportion of half-sib dyads and
460 $\bar{r}_{hs}$ and $\bar{r}_{ns}$ are the average relatedness for half-sib and non-sib dyads, respectively. $\bar{r}_{hs}$ and $\bar{r}_{ns}$ are
461 smaller than 0.25 and 0 respectively, the expected values when the parental population is used as
462 reference or when the reference does not contain related and inbred individuals. The values of $\bar{r}_{hs}$
463 and $\bar{r}_{ns}$ depend on the genetic structure of the sample (*n* and *m*), and the estimator and markers used.

464 Simulations were conducted to check the above analytical predictions. I fixed *m* at 50, and
465 varied *n* between 2 and 10. Ten markers, each having *k*=3~10 alleles in a triangular frequency
466 distribution of $p_i = i/(2k(k+1))$ in the parental population were simulated. Allele frequencies at
467 each locus were calculated from the sample assuming unrelated non-inbred individuals and were
468 used in calculating the LR, R, and QG estimators. Values of $\bar{r}_{hs}$ and $\bar{r}_{ns}$ across 100 replicate runs
469 are shown in Figure 4. As can be seen, with an increase in *n*, $\bar{r}_{hs}$ and $\bar{r}_{ns}$ for each estimator increase
470 towards to the expected values of 0.25 and 0 when the reference contains no related individuals.
471 Different estimators give different values of $\bar{r}_{hs}$ and $\bar{r}_{ns}$, the difference being large between QG and
472 the other estimators. $\bar{r}_{hs}$ and $\bar{r}_{ns}$ are also marker dependent. Markers with a higher polymorphism
473 tend to give higher values of $\bar{r}_{hs}$ and lower values of $\bar{r}_{ns}$, especially for R and LR estimators. The
474 estimate of average relatedness across all possible pairs of individuals (data not shown) is very
475 close to zero, regardless of the estimators, the family structure of the sample, and the markers.

476 **Discussions**

477 Although marker based relatedness estimators are developed using the IBD concept of relatedness,
478 they are better interpreted in terms of Wright's (1921) original correlation concept of relatedness.
479 This is because the IBD definition has to use an appropriate ancestral population as the reference,
480 and assume non-inbred and unrelated individuals in the reference. In practice, this definition poses
481 no problem when a pedigree of sufficient depth is analysed for relatedness. However, when marker
482 data are analysed for relatedness, frequently genotype or allele frequency data are unavailable from
483 an ancestral population, and allele frequencies used in calculating relatedness have to be estimated
484 from the current sample in which relatedness between individuals is being calculated. This practice
485 effectively uses the current population (sample) as reference, and an estimator conforming to the
486 correlation concept of relatedness should give an average estimate of zero. This is true regardless of
487 the actual relatedness among individuals in the sample, as shown by simulation and analytical
488 results in this study. Relatedness between two individuals can be understood as the probability of
489 IBD between two genes, one taken at random from each individual, relative to the probability of
490 IBD between two genes taken at random from the reference population. A negative value signifies

that the individuals are less related in ancestry than the average, and as a result have genotypes less similar in expectation than the average.

The shift of reference from an ancestral to the current population also entails that the constraint of IBD coefficients in the range of [0,1] used by likelihood estimators of *r* (Milligan, 2003; Wang, 2007; Anderson & Weir, 2007) is not justified, and may lead to biased *r* estimates. This bias is caused by the presence of related or/and inbred individuals in a sample which are assumed absent in calculating allele frequencies, and persists even if genomic data with millions of SNPs are used. For a sample taken at random from a large outbred population, most individuals will be unrelated or only loosely related (Csillery *et al*., 2006), and the bias of likelihood estimators should be small and could be negligible compared with the typically large sampling variance of *r*. For small or inbreeding (e.g. partial selfing) populations, however, the bias can be substantial. In general, the higher the variance in actual relatedness and/or inbreeding in a sample, the higher the bias will the likelihood estimators yield. Operationally it is simple to extend the legitimate range of *r* to [-1,1] in searching for the maximum likelihood estimate of *r* (Konovalov & Heg, 2008), and such a procedure will undoubtedly reduce estimation bias. However, it is unclear how to determine the exact range of values for each of the 9 IBD coefficients for a pair of possibly inbred individuals, and how to ensure *r* estimates are constrained in the range [-1,1] as a result. More work is needed in this direction.

The present study shows that the practice of using the current sample as reference causes two difficulties in the estimation and interpretation of *r*. The first difficulty is that *r* should be defined and interpreted as correlation as conceived originally by Wright (1921), rather than a probability of IBD as currently widely perceived. As correlation, the average *r* across pairs of individuals in the entire sample is always close to zero, and negative *r* values have biological meanings. Accordingly, *r* estimators should be estimating *r* as a correlation coefficient rather than a probability of IBD. I showed that indeed some estimators (e.g. QG, LR and R) can be interpreted as such, while others using similarity index (e.g. LL and W) cannot. The latter estimators, however, can be modified to conform to the correlation definition of relatedness. The second difficulty comes from the assumption of unrelated individuals in the current sample (inbreeding has negligible effect compared with relatedness because it is the latter that predominantly determines the probability of IBD of genes taken at random from the sample), which is necessary for estimating allele frequencies. The use of the same sample for estimating relatedness and allele frequencies introduces circularity, and violates the basic assumption of independence of *r* and allele frequencies in all estimators. Simulations show that, in the presence of a high proportion of related individuals in a sample, *r* estimates should be treated with caution because they depend on the actual genetic

structure and allele frequencies of the sample as well as on relatedness estimators. However, when most individuals are unrelated, the problem is minor and can be ignored as a good approximation. In practice, random sampling from a large outbred population is expected to produce a sample containing only a small fraction of highly related individuals (e.g. Csillery *et al*., 2006). However, for some species, family members (especially juveniles) tend to cluster spatially and sampling without realising and accounting for this family structure may lead to a sample containing just a few large families, as exemplified for a brown trout population (Hansen *et al*., 1997).

It is tempting to estimate *r* and allele frequencies jointly to solve the 2nd problem. However, a proper account of the genetic structure in a sample in estimating allele frequencies requires a full pedigree of all individuals in the sample, not just the pairwise relatedness (Boehnke, 1991; Ritland, 1996). For a sample of individuals with some simple genetic structures such as a 2-generation pedigree, it proves to be possible and effective to estimate both relationship and allele frequencies iteratively (Wang, 2004). Algorithms have also been developed to estimate allele frequencies and inbreeding jointly, assuming unrelated individuals within a population (Hill *et al*., 1995) or a subpopulation (Gao *et al*., 2007). However, no accurate method is available that allows for the joint estimation of pairwise relatedness and allele frequencies from the same sample. As a rough approximation, one may take a 3-step approach. First, *r* is calculated using crude allele frequencies estimated by assuming all individuals in a sample are unrelated. Second, a group of sampled individuals that are mutually unrelated or lowly related are identified using the crude *r* estimates, and is used for refining allele frequencies. Third, the refined allele frequencies are then used for calculating *r*. There are however several difficulties with this approach. First, *r* is a continuous quantity and it is unclear which threshold value should be used in selecting "unrelated" or "lowly related" individuals. Second, it can be difficult in practice to choose sufficiently many mutually unrelated individuals for accurate estimates of allele frequencies. Due to genuine genealogical relationships or merely sampling errors, the crude *r* estimates may indicate that individual X_1 is related to X_2, X_2 to X_3, …, X_{n-1} to X_n, while the other pairs of the *n* individuals may be unrelated as indicated by the *r* estimates. In such a case, one has to discard *n*-1 individuals in calculating allele frequencies, which may become very inaccurate because of a small sample size when *n* is large. Third, simply discarding related individuals throws away information for allele frequencies.

Another problem caused by the practice of using the current sample as reference is the sampling errors of allele frequencies due to a finite sample size. Using the same individuals for estimating relatedness and allele frequencies introduces a negative covariance between them (Ritland, 1996). Effectively, the relatedness between two individuals is estimated by using the sample, including the two individuals, as reference. As a result, relatedness is underestimated by an

559 amount in the order of $1/N$, where N is the sample size. This bias can be removed by excluding the
560 focal individuals in calculating allele frequencies used in estimating their relatedness (Queller &
561 Goodnight, 1989; Ritland, 1996). However, the frequency of an allele present only in the focal
562 individuals will be estimated to be zero by this exclusion procedure, which causes some estimators
563 to become undefined.

564 Understanding the concepts of relatedness and inbreeding, especially their relative nature
565 defined by the reference, is pivotal in correctly interpreting and applying the estimates in practice.
566 First, relatedness and inbreeding should be understood as correlations between gametes between
567 and within individuals caused by *recent* coancestry (coalescent). Essentially any two organisms are
568 related and any individual is inbred on the earth because of the existence of recent or remote
569 common ancestors. However, the relevant time scale for relatedness and inbreeding is the recent
570 past (i.e. $\ll 1/u$ generations where u is the mutation rate). This relatively short time scale was not
571 explicitly spelt out by Wright (1921, 1922), but is necessary for relatedness and inbreeding to be
572 useful in most practical applications. For example, an individual with inbreeding coefficient F is
573 expected to be homozygous for an allele with frequency p (in the reference) at a probability of $pF +$
574 $p^2(1-F)$. This function applies when mutations are unimportant relative to drift and inbreeding,
575 implying the most distant reference should be much smaller than $1/u$. Otherwise, mutations have to
576 be accounted for in this probability. In practice, the time scale is invariably much shorter than $1/u$,
577 no matter in pedigree or marker based analyses. Within this time scale, how many generations as a
578 minimum should we trace back for relatedness and inbreeding estimation? Obviously, the further
579 the genealogy is traced back into the past, the higher the r and F estimates for all individuals in the
580 current generation. However, for most applications, it is the relative values of r and F of the current
581 focal individuals that are important. So long as the variance of r and F estimates becomes constant,
582 then there is no need to trace pedigree further back. For a population with a mating system that
583 allows well mixing of the genes (i.e. random mating), it is necessary to trace just ~5 ancestral
584 generations (e.g. Balloux *et al*., 2004) to obtain genealogical F and r values that correlate highly
585 with estimates obtained from a much deeper pedigree. This is understandable because a more
586 remote ancestor will tend to contribute more evenly to all current descendants (Wray & Thompson,
587 1990), and thus has smaller effect on the variance of r and F. However, for a population with a
588 mating system that does not allow quick and extensive mixing of genes, such as subdivision with
589 little migration, then a deeper pedigree with many more ancestral generations might be needed to
590 provide a reliable description of the relative levels of inbreeding and relatedness. For example, Toro
591 *et al*. (2002) showed that genealogical r estimates from a shallow pedigree of 5 generations are less
592 correlated with molecular r estimates than those from a deep pedigree of 19~20 generations,

because the 62 pigs in the analysis were taken from two stains that were isolated. Assuming non-inbred and unrelated founders in a shallow pedigree may lead to distorted *r* and *F* estimates when the assumption is violated.

Second, it is the relative values of *r* and *F* that are relevant in most applications. For example, *r* and *F* estimates from pedigree or marker analyses are usually correlated with or regressed to a phenotype of a fitness component in investigations of inbreeding depression (Nielson *et al.*, 2012; Brekke *et al.*, 2010) and of a quantitative trait in estimating its heritability (Ritland, 2000). The estimates are also compared between groups of individuals, such as between sexes or age classes, in studying the social and population structures. For example, Surridge *et al*. (1999) found that the average relatedness is negative between males and is positive between females in a European wild rabbit population, and interpreted the result as indicating male biased migration among social groups and female philopatry. In conservation management of endangered species, *r* and *F* estimates can be used to optimise the selection and mating scheme for maximising the genetic diversity (e.g. Fernández *et al*., 2003). In all these applications, the magnitute of *r* and *F* values is irrelevant, and a linear transformation of the estimates (by adding or multiplying a constant non-zero value) does not affect a downstream analysis. This means that, in a pedigree-based analysis, any reference generation suffices so long as the pedigree is sufficiently deep and thus variation of *r* and *F* is close to its maximum. In a marker based analysis, allele frequencies at any reference generation can be used in *r* and *F* estimation if the estimators conform to the correlation definitions.

Third, caution must be exercised in applications in which the magnitudes of *r* and *F* values have more definite biological meanings. One such application is to classify pairs of individuals into well-separated relationship categories such as first- and second-degree relationships (e.g. Blouin *et al*., 1996; Glaubitz *et al*., 2003; van Dan *et al*., 2008) from pairwise relatedness estimates. If a dyad has an estimated *r* of 0.52 and 0.28, for example, it is classified as first (e.g. parent-offspring, full-sib) and second (e.g. half-sib, avuncular) degree relationship, respectively. However, the misclassification rate is generally very high even many markers are used (Blouin *et al*., 1996; Glaubitz *et al*., 2003; van Dan *et al*., 2008; Csillery *et al*., 2006), because of the high sampling variance of *r* and thus the wide overlap in distributions of possible *r* values between even well-separated relationships. This study shows further that the magnitudes of *r* values are more or less arbitrary, depending on the reference allele frequencies. When the current sample is used as reference, *r* is usually underestimated such that the average value of *r* for the sample is zero. These biases depend on the actual fine genetic structure of the sample, and the markers being used (Figure 4). A better approach is to estimate relationships directly from marker data with a pairwise (e.g. Marshall *et al*., 1998; Goodnight & Queller, 1999) or full (e.g. Wang & Santure, 2009) likelihood

627 method. This direct approach is much more robust to misspecifications of reference allele
628 frequencies, and has the option to jointly estimate relationship and allele frequencies.

629 In this study, I investigated a few *F* and *r* estimators that are developed from population
630 genetics models. When the underlying assumptions are met, they provide unbiased and marker-
631 independent estimates of *F* and *r*. It is noticeable that some marker-based surrogate statistics are
632 also proposed and applied in indicating the levels of inbreeding and relatedness. These include, for
633 example, multilocus heterozygosity (MLH) or its complement for indicating inbreeding (e.g.
634 Hansson & Westerberg, 2002) and similarity indexes (including the one used in (7)) (e.g. Ellegren,
635 1999) for indicating relatedness. Compared with model-based estimators, these non-model based
636 measurements may have a similar correlation coefficient with genealogical *F* and *r* estimates in
637 some circumstances (Wang, 2011). However, these surrogate statistics are undesirable in several
638 aspects. First, they do not estimate, although correlate with, *F* and *r*, and as a result have limited
639 uses in practice. For example, MLH or its complement calculated from a set of markers as a
640 surrogate for *F* cannot be used directly in predicting the probability of a genotype or the
641 heterozygosity at another locus with given allele frequencies. Second, they are highly marker
642 dependent. For the same individual, MLH is always higher for highly (e.g. microsatellites) than
643 lowly (e.g. SNPs) polymorphic markers. For the same two individuals, similarity indexes and
644 molecular coancestry are always lower for highly (e.g. microsatellites) than lowly (e.g. SNPs)
645 polymorphic markers. This causes problems in comparing estimates involving individuals with
646 missing data at different loci. An individual with data missing at highly polymorphic loci will tend
647 to have a lower MLH, and higher similarity indexes and molecular coancestry with another
648 individual, than an individual with no missing data or with missing data at lowly polymorphic loci.
649 This marker-dependency also causes difficulties in comparisons within and across studies. Third,
650 being empirical statistics lacking an underlying population genetics model, they have difficulty in
651 weighing information among loci. In contrast, *F* and *r* estimators can weigh the information from
652 different loci properly, using for example the inverse of the expected sampling variance of a locus
653 (e.g. Ritland, 1996; Lynch & Ritland, 1999). The weighting becomes important when markers vary
654 substantially in polymorphism. In view of these shortcomings, these surrogate statistics should be
655 discouraged in practical applications.
656

657 **Acknowledgments**

658 I thank the editor and two anonymous reviewers for helpful comments on earlier versions of this
659 manuscript.

660

661

References

Anderson, A.D. & Weir, B.S. 2007. A maximum likelihood method for estimation of pairwise relatedness in structured populations. *Genetics* **176:** 421–440.

Boehnke, M. 1991. Allele frequency estimation from data on relatives. *Am. J. Hum. Genet.* **48**: 22-25.

Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C. & Taberlet, P. 2004. How to track and assess genotyping errors in population genetics studies. *Mol. Ecol.* **13**: 3261-3273.

Brekke, P., Bennett, P.M., Wang, J., Pettorelli, N. & Ewen, J.G. 2010. Sensitive males: inbreeding depression in an endangered bird. *Proc. R Soc. Lond. B Biol. Sci*. **277**: 3677-3684.

Csillery, K., Johnson, T., Beraldi, D., Clutton-Brock, T., Coltman, D. *et al*. 2006 .Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. *Genetics* **173:** 2091–2101.

Ellegren, H. 1999. Inbreeding and relatedness in Scandinavian grey wolves *Canis lupus*. *Hereditas* **130**: 230–244.

Emik, L.O. & Terrill, C.R. 1949. Systematic procedures for calculating inbreeding coefficients. *J. Hered*. **40**: 51-55.

Fernández, J., Toro, M.A. & Caballero, A. 2003. Fixed contributions designs versus minimization of global coancestry to control inbreeding in small populations. *Genetics* **165**: 885 – 894.

Gao, H., Williamson S. & Bustamante, C.D. 2007. A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics* **176**: 1635–1651.

Goodnight, K. & Queller, D. 1999. Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Mol. Ecol*. **8**: 1231–1234.

Grafen, A. 1985. A geometric view of relatedness. *Oxford Surv. Evol. Biol*. **2**: 28–89.

Hansen, M.M., Nielsen, E.E. & Mensberg, K.L. 1997. The problem of sampling families rather than populations: relatedness among individuals in samples of juvenile brown trout *Salmo trutta L*. *Mol. Ecol*. **6**: 469–474.

Hansson, B. & Westerberg, L. 2002. On the correlation between heterozygosity and fitness in natural populations. *Mol. Ecol.* **11**: 2467–2474.

Hardy, O.J. & Vekemans, X. 1999. Isolation by distance in a continuous population: reconciliation between spatial autocorrelation and population genetics models. *Heredity* **83**: 145–154.

Harris, D.L. 1964. Genotypic covariances between inbred relatives. *Genetics* **50**: 1319-1348.

695 Hill, W.G., Babiker, H.A., Ranford-Cartwright, L.C. & Walliker, D. 1995. Estimation of inbreeding
696 coefficients from genotypic data on multiple alleles, and application to estimation of clonality in
697 malaria parasites. *Genet. Res*. **65**: 53–61.
698 Jacquard, A. 1972. Genetic information given by a relative. *Biometrics* **28**: 1101-1114.
699 Jacquard, A. 1974. *The Genetics Structure of Populations*. Springer, New York.
700 Konovalov, D.A. & Heg, D. 2008. A maximum-likelihood relatedness estimator allowing for
701 negative relatedness values. *Mol. Ecol. Res*. **8**: 256–263.
702 Lynch, M. & Ritland, K. 1999. Estimation of pairwise relatedness with molecular markers. *Genetics*
703 **152**: 1753–1766.
704 Malecot, G. 1948. *Les mathematiques de 1'heredite*. Masson et Cie, Paris, 63 pp.
705 Marshall, T., Slate, J., Kruuk, L. & Pemberton, J. 1998. Statistical confidence for likelihood-based
706 paternity inference in natural populations. *Mol. Ecol.* **7:** 639–655.
707 Maynard Smith, J. 1998. *Evolutionary Genetics*. 2nd edn. Oxford University Press: Oxford.
708 Milligan, B. G., 2003. Maximum-likelihood estimation of relatedness. *Genetics* **163:** 1153–1167.
709 Nielsen, J.F., English, S., Goodall-Copestake, W.P., Wang, J., Walling, C.A., Bateman, A.W., *et al*.
710 2012. Inbreeding and inbreeding depression of early life traits in a cooperative mammal. *Mol.*
711 *Ecol.* **21**: 2788-2804.
712 Powell, J.E., Visscher, P.M. & Goddard, M.E. 2010. Reconciling the analysis of IBD and IBS in
713 complex trait studies. *Nat. Rev. Genet.* **11**: 800–805.
714 Queller, D. C. & Goodnight, K. F. 1989. Estimating relatedness using molecular markers. *Evolution*
715 **43**: 258–275.
716 Ritland, K. 1996. Estimators for pairwise relatedness and inbreeding coefficients. *Genet. Res.* **67**:
717 175–186.
718 Ritland, K. 2003. Marker-inferred relatedness as a tool for detecting heritability in nature. *Mol. Ecol*.
719 **9**: 1195–1204.
720 Rousset, F. 2002. Inbreeding and relatedness coefficients: what do they measure? *Heredity* **88**: 371–
721 380.
722 Seger, J. 1981. Kinship and covariance. *J. Theoret. Biol*. **91**: 191-213.
723 Surridge, A.K., Ibrahim, K.M., Bell, D.J., Webb, N.J., Rico, C. & Hewitt, G.M. 1999. Fine-scale
724 genetic structuring in a natural population of European wild rabbits (*Oryctolagus cuniculus*).
725 *Mol. Ecol*. **8**: 299-307.
726 Thomas, S.C. 2010. A simplified estimator of two and four gene relationship coefficients. *Mol. Ecol*.
727 *Res*. **10**: 986-994.

728 Toro, M., Barragán, C., Óvilo, C., Rodrigañez, J., Rodriguez, C. & Silió, L. 2002. Estimation of
729 coancestry in Iberian pigs using molecular markers. *Conserv. Genet*. **3**: 309–320.
730 van Horn, R., Altmann, J. & Alberts, S. 2008. Can't get there from here: inferring kinship from
731 pairwise genetic relatedness. *Anim. Behav*. **75**: 1173–1180.
732 Van de Casteele, T., Galbusera, P. & Matthysen, E. 2001. A comparison of microsatellite-based
733 pairwise relatedness estimators. *Mol. Ecol*. **10:** 1539–1549.
734 Wang, J. 2002. An estimator for pairwise relatedness using molecular markers. *Genetics* **160**: 1203–
735 1215.
736 Wang, J. 2007. Triadic IBD coefficients and applications to estimating pairwise relatedness. *Genet.*
737 *Res*. **89**: 135–153.
738 Wang, J. & Santure, A.W. 2009. Parentage and sibship inference from multilocus genotype data
739 under polygamy. *Genetics* **181**: 1579–1594.
740 Wang, J. 2011. Unbiased relatedness estimation in structured populations. *Genetics* **187**: 887–901
741 Wright, S. 1921. Systems of mating. *Genetics* **6**: 111-178.
742 Wright, S. 1922. Coefficients of inbreeding and relationship. *American Naturalist* **56**: 330–338.
743 Wright, S. 1965. The interpretation of population structure by *F*-statistics with special regard to
744 systems of mating. *Evolution* **19**: 395-420.
745 Yip, S. P. 2002. Sequence variation at the human ABO locus. *Annals of Human Genetics* **66**: 1–27.
746
747

748 **Table 1** Genotypes and frequencies of a large half-sib family

Shared parent			Half-sib offspring			Offspring sample allele frequency, $\hat{p}_x$
Genotype	Allelic state	Frequency	Genotype	Allelic state	Frequency	
ii	$\forall i$	p_i^2	*ix*	$\forall i, \forall x$	p_x	$\frac{1}{2}(\delta_{ix} + p_x)$
ij	$\forall i, \forall j \neq i$	$2p_i p_j$	{*ix*,*jx*}	$\forall i, \forall j \neq i, \forall x$	$\{\frac{1}{2}p_x, \frac{1}{2}p_x\}$	$\frac{1}{4}(\delta_{ix} + \delta_{jx}) + \frac{1}{2}p_x$

749