# Effects of sampling close relatives on some elementary population genetics analyses

Jinliang Wang

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

*Left running head:* J Wang

*Right running head:* Effects of sampling close relatives

*Key words:* Genetic Variation, *F* statistics, Allele Frequency, Hardy-Weinberg Equilibrium, Linkage Equilibrium

*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**

Many molecular ecology analyses assume the genotyped individuals are sampled at random from a population and thus are representative of the population. Realistically, however, a sample may contain excessive close relatives (ECR) because, for example, localized juveniles are drawn from fecund species. Our knowledge is limited about how ECR affect the routinely conducted elementary genetics analyses, and how ECR are best dealt with to yield unbiased and accurate parameter estimates. This study quantifies the effects of ECR on some popular population genetics analyses of marker data, including the estimation of allele frequencies, *F*-statistics, expected heterozygosity ($H_e$), effective and observed numbers of alleles, and the tests of Hardy-Weinberg equilibrium (HWE) and linkage equilibrium (LE). It also investigates several strategies for handling ECR to mitigate their impact and to yield accurate parameter estimates. My analytical work, assisted by simulations, shows that ECR have large and global effects on all of the above marker analyses. The naïve approach of simply ignoring ECR could yield low-precision and often biased parameter estimates, and could cause too many false rejections of HWE and LE. The bold approach, which simply identifies and removes ECR, and the cautious approach, which estimates target parameters (e.g. $H_e$) by accounting for ECR and using naïve allele frequency estimates, eliminate the bias and the false HWE and LE rejections, but could reduce estimation precision substantially. The likelihood approach, which accounts for ECR in estimating allele frequencies and thus target parameters relying on allele frequencies, usually yields unbiased and the most accurate parameter estimates. Which of the four approaches is the most effective and efficient may depend on the particular marker analysis to be conducted. The results are discussed in the context of using marker data for understanding population properties and marker properties.

## Introduction

Genetic markers are an elementary and indispensable tool widely used in studies of evolutionary, ecological and conservation biology, and of human genetics and medicine (Sunnucks 2000; Selkoe & Toonen 2006). Rich and important information can be gleaned from a sample of multilocus marker genotypes, such as that about population demography and evolutionary history (Luikart *et al*. 2003), the basic biology of a species (like mating system/strategy/behaviour, migration), and the inheritance of quantitative traits including complex human diseases (Risch & Merikangas 1996).

In many studies, a population is represented by a sample of individuals drawn from it. To obtain unbiased and accurate estimates of population properties (e.g. subdivision and differentiation) or their underlying mechanisms (e.g. genetic drift and selection) leading to the observed properties, it is commonly assumed that the sampling is at random with regard to kinship. Biased sampling with too many or too few kin compared with that expected under random sampling could lead to biased and inaccurate parameter estimates. This is well recognized in human genetics studies, where unrecognised relatives cause allele frequency misspecifications and thus increased false positives in linkage analysis of inheritable disease (Ott 1992). Similarly in genome-wide association studies, cryptic relatedness would violate assumptions of statistical independence and introduce misclassification effects (McCarthy *et al*. 2008). It is now a common practice to use marker data to check self-reported pedigrees and identify cryptic relatedness. The uncovered close relatives are then removed (e.g. Wellcome Trust Case Control Consortium 2007) or explicitly accounted for by genomic-control approaches (Voight & Pritchard 2005; Zheng *et al*. 2006).

In the evolutionary and conservation studies of wild plant and animal populations, it is unclear how often excessive close relatives, ECR, are inadvertently included in a supposedly random sample. However, due to the lack of quality control measures such as self-reports in humans and because of the high fecundity, small population size and patchy, rather than even, spatial distribution of closely related individuals in many plant and animal species, close relatives could be rampant in samples. Hansen *et al*. (1997) showed that juvenile brown trout individuals sampled from a small area were represented by just a few families and thus provide biased allele frequency estimates. Goldberg & Waits (2010) sampled both tadpoles and adults from several Columbia spotted frog populations. They found that tadpole samples overall had a higher $F_{ST}$ between populations than adult samples. When the full-sib families in tadpoles were identified by using the marker data (Wang & Santure 2009) and all siblings but one per family were removed, the $F_{ST}$ estimates became closer to those estimated from adult samples (Goldberg & Waits 2010). However, these and other effects of ECR in a sample were not consistent across species (Goldberg & Waits 2010). For a Bayesian clustering analysis, it was shown (Anderson & Dunham 2008; Rodríguez-Ramilo & Wang 2012) that ECR in a sample could ruin the inference of the number of populations represented by the sample.

Theoretical studies on the sporadic effects of ECR have been concentrated on the estimation of allele frequencies (Boehnke 1991; Broman 2001; McPeek *et al*. 2004) and

expected heterozygosity (DeGiorgio & Rosenberg 2008). It is unclear whether other population parameters and analyses are also affected by the inclusion of ECR in a sample or not. If they are, how great are the effects of ECR and how can ECR be best dealt with for mitigating the effects? This study intends to fill the gaps of knowledge by investigating systematically the effects of ECR on the most commonly conducted population genetics analyses, including the estimation of allele frequencies, expected heterozygosity, *F* statistics, effective and observed numbers of alleles, and the tests for Hardy-Weinberg and linkage equilibrium. Based on analytical treatments assisted by simulations, I compare different approaches to handling a sample with ECR, and discuss the choice of the approaches in practice.

## Methods and Results

By close relatives, I mean relatives of the first (full sibling and parent-offspring) and second (avuncular, half sibling, grandparents and grand offspring) degrees. By excessive close relatives (ECR), I mean close relatives appearing in a sample at a higher proportion than that expected under complete random sampling. I use analytical treatments, where possible, to study the effects of ECR on each of a number of commonly conducted population genetics analyses. Some of the analytical results are checked and complemented by simulations (Appendix S1, Supporting information). The basic assumptions are diploid species and codominant markers. The general conclusions are however applicable to other species and other kinds of markers.

### *Allele frequency*

Some population properties, such as $F_{ST}$, can be inferred directly and solely from allele frequency information. Many others, such as $F_{IT}$, $F_{IS}$ and relatedness, however, require genotype as well as allele frequency information for inference. Regardless, allele frequencies provide elementary information for many population genetics analyses.

For a given sample size, inclusion of ECR reduces the effective sample size and thus the precision of allele frequency estimates. As an example, consider a sample of $m$ full siblings who share the same pair of parents. Although the sample size of genes is $2m$, these genes are not independent, coming from only 4 independent parental genes when the parents are non-inbred and unrelated and from fewer than 4 independent parental genes when

otherwise. The effective sample size of genes is thus 4 at the maximum. Let us now consider a locus with two codominant alleles, A and a, whose population frequencies are $p$ and 1-$p$, respectively. If $n$ diploid individuals are sampled at random from a large population at Hardy-Weinberg equilibrium (HWE), the number of copies of allele A, $n_A$, in the sample follows the binomial distribution

$$\Pr[n_A|p,2n] = \frac{(2n)!}{n_A!(2n-n_A)!}p^{n_A}(1-p)^{2n-n_A}. \quad (1)$$

Suppose $m$ diploid individuals are sampled at random from a single full-sib family in the same population. Among the 2$m$ sampled genes, half comes from the same father and half from the same mother. The probability that $m_A$ of the $m$ genes from a given parent are observed to be A allele is

$$\Pr[m_A|p,m] = p^2 0^{m-m_A} + 2p(1-p)\frac{m!}{m_A!(m-m_A)!}\left(\frac{1}{2}\right)^m + (1-p)^2 0^{m_A}, \quad (2)$$

where $0^i \equiv 1$ and 0 when $i$=0 and $i\neq 0$, respectively. (2) was derived by considering the probability of each possible parental genotypes, which is $p^2$, $2p(1-p)$, and $(1-p)^2$ for genotype AA, Aa and aa, respectively. For a given parental genotype, $m_A$ follows the binomial distribution with parameters $m$ and $q$, where $q$ =1, ½, 0 is the frequency of allele A in the parental genotype AA, Aa and aa, respectively.

A combined sample containing $n$ individuals drawn at random from a large population at HWE and $m$ individuals drawn at random from a single full sib family in the same population has a total number of 2($m$+$n$) genes. When $i$ genes from the random sample, $j$ genes from one parent and $k$ genes from the other parent of the full sib family are of allele A, the number of copies of A allele in the combined sample is $x=i+j+k$, with $x$ in the range [0, 2($m$+$n$)]. This leads to the probability

$$\Pr[x|p,2m,2n] = \sum_{i,j,k}\Pr[i|p,2n]\Pr[j|p,m]\Pr[k|p,m], \quad (3)$$

where the summation is over all possible $i$, $j$ and $k$ values, with constraints $i$=[0, 2$n$], $j$=[0, $m$], $k$=[0, $m$], and $i+j+k \equiv x$. In (3), $\Pr[i|p,2n]$, $\Pr[j|p,m]$, and $\Pr[k|p,m]$ are calculated by (1), (2) and (2), respectively. When the $m$ individuals are drawn at random from $M$ full sib families, with family $j$ ($j$=1~$M$) contributing $m_j$ individuals such that $\sum_{j=1}^{M} m_j = m$, the probability that $x$ of the 2($m$+$n$) genes are of allele A can be derived similarly to (3). The equation is however much more complicated, and is thus not shown herein (available upon

request). The computation of the probability quickly becomes infeasible with increasing values of $M$, $m$, and $n$.

The sample allele frequency distributions for some examples calculated by the equations above are shown in Figure 1. A marker with two codominant alleles of an equal frequency of $p$=0.5 is assumed at HWE in a large random mating population. A combined sample contains $n$=10 unrelated individuals drawn at random from the population, and $m$=10 individuals drawn at random from $M$=5 full-sib families in the population. The sample size distribution, $\{n, m_1, m_2, m_3, m_4, m_5\}$, is {10, 10, 0, 0, 0, 0}, {10, 8, 2, 0, 0, 0}, {10, 6, 2, 2, 0, 0}, {10, 4, 2, 2, 2, 0}, and {10, 2, 2, 2, 2, 2} for combined samples E1~E5, respectively. Removing all but one full sibling from each family, I obtain reduced samples e1~e5, which have sample size distributions {10, 1, 0, 0, 0, 0}, {10, 1, 1, 0, 0, 0}, {10, 1, 1, 1, 0, 0}, {10, 1, 1, 1, 1, 0}, and {10, 1, 1, 1, 1, 1}, respectively. Note, the total sample size is the same, $n+m$ =20, for samples E1~E5, but is variable, 11~15, for samples e1~e5. Figure 1 shows that the distribution of sample allele frequencies becomes less dispersed with a decreasing imbalance in full sib family size in samples E1~E5, and with an increasing sample size when all but one full siblings per family are removed in samples e1~e5. Note $\hat{p}$ distributions are discrete and the sample sizes are unequal for samples e1~e5, which create the illusion that e1~e5 become more dispersed in that order, but the opposite is true as shown in the lower panel of Figure 1.

All of the 10 samples (E1~E5, e1~e5) provide unbiased estimate of $p$, using the allele counting method

$$\hat{p} = \frac{1}{2n}\sum_{j=1}^{n} X_j, \quad (4)$$

where indicator variable $X_j$ is the number of copies of allele A carried by individual $j$. $X_j$ takes values 2, 1 and 0 for genotypes AA, Aa and aa respectively, with the expectation $E[X_j] = 2p$. In (4), $n$ is the individual sample size, which takes values 20 for samples E1~E5, and 11~15 for samples e1~e5. Although (4) is always unbiased, its precision is affected by the genetic structure of the sample. The inclusion of siblings has two effects on the precision of $\hat{p}$. On one hand, siblings do provide information about allele frequencies and thus act to increase the precision. On the other hand, however, they may also act to reduce estimation precision when full-sib family sizes are unbalanced. The overall effect depends on family size distribution. Full-sib families included in a sample can increase and decrease precision when these families have an even and uneven distribution of family size, respectively.

Figure 1 also shows the sampling variance of $\hat{p}$ obtained from samples E1~E5 and e1~e5. Removing siblings improves the precision of $\hat{p}$ when large sib families are sampled (E1~E3), but has the opposite effect when all sib families are small (E5). For the 10 samples, simulation results are almost identical to those calculated analytically, as shown in Figure 1.

Hereafter for easy reference, applying (4) by ignoring relatives is called the naïve approach, and applying (4) by identifying and removing all but one relative per family is called the bold approach.

For a sample containing close relatives, neither the naïve nor the bold approach is, in general, the best option that gives the most accurate $\hat{p}$. As discussed above, it all depends on the actual sibship size distribution in a sample. The best strategy in general is to use the genetic relationships (pedigree) of the sampled individuals in weighting their genotype information for allele frequency estimation. Several methods available (e.g. Boehnke 1991; McPeek *et al*. 2004) are applicable to any known genetic structure of the sampled individuals. For the case of a sample containing unknown close relatives including full-sib, half-sib and parent-offspring relationships, it is possible to use the genotype data to infer iteratively the relationships and allele frequencies jointly by a likelihood approach (Wang & Santure 2009). In the simplest case of all close relatives being full siblings, Broman (2001) proposed a simple estimator that calculates an allele frequency estimate for each sib family and then weights these estimates by the inverse of their sampling variances. Suppose a sample contains individuals from $M$ full-sib families, with family $i$ (=1~$M$) contributing $m_i$ individuals. The estimator for the frequency of an allele A is

$$\hat{p} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{m_i} X_{ji}/(m_i+1)}{2\sum_{i=1}^{M} m_i/(m_i+1)}, \qquad (5)$$

where the indicator variable $X_{ji}$ = 2, 1, 0 when individual $j$ from family $i$ has a genotype containing 2, 1, and 0 copies of allele A, respectively. When all sampled individuals are unrelated such that $m_i \equiv 1$, (5) reduces to (4) as expected. Note, estimator (5) uses information from all sampled individuals. However, it gives a lower weight, $1/(m_i + 1)$, to information from a larger family $i$ with $m_i$ individuals. From now on, (5) is called the weighted estimator.

A yet better estimator than (5) is obtained by accounting for the sample sibship structure in the likelihood framework. For each known or reconstructed sib family, the parental genotypes can be reconstructed probabilistically from sibling genotypes (Broman

2001; Wang & Santure 2009). The reconstructed parental genotypes, rather than the observed genotypes of sampled individuals, are then used in estimating allele frequencies. This estimator has no closed form. It is called likelihood estimator hereafter and is implemented in an expectation–maximization algorithm, as described in Broman (2001), for the simple case of full siblings. For the more general case of multiple types of close relatives (full and half siblings, parent-offspring), the likelihood estimator was implemented in a simulated annealing algorithm in Wang & Santure (2009).

Simulations (Appendix S1) were conducted to check the analytical results of the naïve and bold estimators, and to investigate their accuracies against those of weighted and likelihood estimators for the five example samples E1~E5. Both weighted and likelihood estimators are unbiased, and their sampling variances are shown in Figure 1. In all five cases of family size distributions, both estimators yield much better results than the naïve estimator (which ignores relatives) and the bold estimator (which removes all but one siblings per family). The likelihood estimator is slightly more precise than the weighted estimator, but the differences are almost imperceptibly small in all of the five examples. There seems to be little advantage of the likelihood estimator over the weighted estimator for the simple case of all ECR being full siblings.

*F*-statistics

Wright's (1965) *F*-statistics ($F_{IS}$, $F_{ST}$, and $F_{IT}$) are the traditional and most popular statistics used in assessing the distribution of genetic variation at different hierarchical levels of a population subdivision. Each of them measures the correlation between homologous genes drawn at one hierarchical level (i.e. individuals, subpopulations, total population) relative to genes drawn at another hierarchical level. Equivalently, it measures the inbreeding at one hierarchical level relative to another. Specifically, $F_{IS}$ and $F_{IT}$ are the inbreeding coefficients of an individual (I) relative to the subpopulation (S) and the total population (T) respectively to which the individual belongs, and $F_{ST}$ is the expected inbreeding coefficient of a hypothetical individual from random mating within a subpopulations (S) relative to the total population (T). The relationship among the three statistics is (Wright 1965)

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST}). \quad (6)$$

*F*-statistics can be estimated from a sample of individuals drawn at random from each of a number of subpopulations, using the pedigree or marker data of sampled individuals. In ideal conditions (e.g. when mutations have negligible effects relative to migration and drift, Whitlock 2011; Wang 2012, 2015), the two types of information should yield the same results in expectation. Herein I investigate the effect of close relatives on *F*-statistics by the genealogical approach, and the results were checked by simulations using the marker approach.

Let us consider a subdivided population, and denote the probabilities of identical by descent (PIBD) for the two homologous genes within an individual, for the two homologous genes drawn at random within a subpopulation and within the total population by $\alpha$, $\beta$, and $\gamma$, respectively. By definition, we have

$$F_{IS} = \frac{\alpha-\beta}{1-\beta}, F_{ST} = \frac{\beta-\gamma}{1-\gamma}, F_{IT} = \frac{\alpha-\gamma}{1-\gamma}, \quad (7)$$

which satisfy (6) as expected.

Suppose the average PIBD between individuals within a sample drawn from a subpopulation is increased from $\beta$ to $\beta'$, because ECR are (intentionally or unintentionally) included in the sample. As a result, $F_{IS}$ would be expected to be decreased to

$$F'_{IS} = \frac{\alpha-\beta'}{1-\beta'}. \quad (8)$$

$F'_{IS}$ is always smaller than $F_{IS}$, because $\beta' > \beta$ and $\alpha$ is unaffected by ECR. The larger the increase in PIBD between individuals drawn from within a subpopulation due to the inclusion of a greater proportion of close relatives, $\beta' - \beta$, the smaller will be $F'_{IS}$ relative to $F_{IS}$. Let us consider a numerical example. For a large subpopulation in HWE, we have $\alpha = \beta = 0$ and $F_{IS} = 0$. When a sample drawn from the subpopulation contains a proportion of $\delta$ full-sib pairs (PIBD=1/4) and a proportion of $1 - \delta$ non-sib pairs (PIBD=0), its average PIBD is $\beta' = \delta \times \frac{1}{4} + (1 - \delta) \times 0 = \delta/4$. Inserting $\beta' = \delta/4$ into (8) leads to $F'_{IS} = -\delta/4$, which is smaller than the actual value of $F_{IS} = 0$ of the population. When the frequency of full sib pairs in a sample is a quarter ($\delta = 1/4$), then $F'_{IS} = -0.0625$, much smaller than its real value of 0. The inclusion of ECR (i.e. full siblings) in a sample results in negative inbreeding, leading possibly to the false conclusion that the subpopulation is affected by admixture (hybridization) and/or avoids close relative mating.

Because of the inclusion of ECR, $\beta' > \beta$, and the differentiation between subpopulations becomes

$$F'_{ST} = \frac{\beta' - \gamma}{1 - \gamma}. \quad (9)$$

We have always $F'_{ST} > F_{ST}$, because $\beta' > \beta$ and $\gamma$ is unaffected by ECR. In a subdivided population with a high migration rate ($m$) between subpopulations and/or a large effective size ($N_e$) of subpopulations, we have $\beta = \gamma = 0$ and $F_{ST} = 0$. When a sample containing a proportion of $\delta$ full-sib pairs is drawn from each subpopulation and is used in estimating $F_{ST}$, we have $\beta' = \delta/4$ (as derived above) and thus $F'_{ST} = \beta' = \delta/4$ from (9). If a quarter of the pairs of individuals drawn from a subpopulation are full sib pairs ($\delta = 1/4$), then $F_{ST}$ would be estimated as 0.0625, much larger than its real value of 0. Sampling ECR results in an increase in estimated differentiation, leading to a false conclusion that the subpopulations are small and are isolated (i.e. $mN_e$ is small).

It is clear from definition (7) that ECR in a sample drawn from a subpopulation do not affect $F_{IT}$ in expectation, because they cause an increase in $\beta$ but no changes in $\alpha$ and $\gamma$. ECR act to decrease $F_{IS}$ and to increase $F_{ST}$ to the same extent. These effects cancel out exactly and thus ECR in a sample do not affect $F_{IT}$ estimates in expectation. This conclusion can also be reached from (6), given that $F_{IS}$ and $F_{ST}$ are affected to the same extent but in opposite directions by non-random sampling such that $(1 - F'_{IS})(1 - F'_{ST}) = (1 - F_{IS})(1 - F_{ST})$ and thus $F'_{IT} = F_{IT}$. However, it should be noted that ECR do cause an increased sampling variance and thus a decreased accuracy of $F_{IT}$ when it is estimated from marker data, because ECR result in less precise estimates of allele frequencies (above) which must be used in estimating $F_{IT}$.

Simulations were conducted to check the above analytical results, and to investigate the effects of different approaches to estimating allele frequencies on the bias and accuracy of $F_{ST}$ estimates. I assumed the simple situation of $R$=20 discrete populations in Wright's (1931) island migration model. The populations had reached equilibrium among mutations, drift and migration, with an equilibrium $F_{ST}$ value in the range [0, 0.16]. A number of $n$=50 unrelated individuals and a number of $m$=50 individuals from 5 full sib families were drawn from each population. The family size distributions $\{n, m_1, m_2, m_3, m_4, m_5\}$, where $m_i$ is the number of full siblings from family $i$, are {50, 50, 0, 0, 0, 0}, {50, 40, 10, 0, 0, 0}, {50, 30, 10, 10, 0, 0}, {50, 20, 10, 10, 10, 0}, {50, 10, 10, 10, 10, 10} in the five simulated samples E1~E5. Each

sampled individual was genotyped for a locus with $k$=4 codominant alleles. The ancestral allele frequencies, $\boldsymbol{p_0} = \{p_{10}, p_{20}, \ldots, p_{k0}\}$, were drawn from a uniform Dirichlet distribution $\mathcal{D}(1,1, \ldots,1)$. Conditional on $\boldsymbol{p_0}$, the allele frequencies of a population $j$ ($j$=1~$R$), $\boldsymbol{p_j} = \{p_{1j}, p_{2j}, \ldots, p_{kj}\}$, were drawn from the Dirichlet distribution $\mathcal{D}(fp_{10}, fp_{20}, \ldots, fp_{k0})$, where $f = \frac{1}{F_{ST}} - 1$ (Nicholson *et al*. 2002). Given $\boldsymbol{p_j}$, genotype data of an unrelated individual or of full siblings from population $j$ were obtained (Appendix S1, Supporting information).

The simulated genotype data were then used to calculate Nei & Chesser's (1983) nearly unbiased estimator of $G_{ST}$ (equivalent to Wright's $F_{ST}$, see Wang 2012, 2015)

$$\hat{G}_{ST} = 1 - \hat{H}_S/\hat{H}_T, \tag{10}$$

where

$$\hat{H}_S = \frac{2\tilde{M}}{(2\tilde{M}-1)R}\sum_{j=1}^{R}\left(1 - \sum_{i=1}^{k}\hat{x}_{ij}^2\right), \tag{11}$$

$$\hat{H}_T = 1 - \sum_{i=1}^{k}\left(\frac{1}{R}\sum_{j=1}^{R}\hat{x}_{ij}\right)^2 + \frac{\hat{H}_S}{2\tilde{M}R}. \tag{12}$$

In (11) and (12), $\hat{x}_{ij}$ is the estimated frequency of allele $i$ in the sample from population $j$, $k$ is the number of alleles observed in the set of samples from the $R$ populations, and $\tilde{M}$ is the harmonic mean sample sizes. $\hat{x}_{ij}$ was estimated by the naïve, bold and weighted approaches and was used in (10-12) to obtain the corresponding $F_{ST}$ estimates. The means and RMSEs (root mean squared error, $=\sqrt{\text{variance} + \text{bias}^2}$) of $F_{ST}$ estimates from the naïve, bold and weighted approaches are compared in Figure 2. The likelihood estimator is computational intensive but has only slight accuracy improvement over the weighted estimator (Figure 1), and so it is not considered in estimating $F_{ST}$.

Ignoring relatedness by the naïve estimator leads to an overestimate of $F_{ST}$ across the samples with different proportions of full siblings, and in the entire range of actual $F_{ST}$ values (0~0.16). For samples E1~E5 from an undifferentiated (true $F_{ST}$ =0) population, the proportions of sibling pairs, calculated by $\frac{\sum_{i=1}^{5} m_i(m_i-1)/2}{(n+m)(n+m-1)/2}$ where $n = m = 50$, are 0.247, 0.167, 0.106, 0.066 and 0.045 respectively, and the predicted $F_{ST}$ values by (9) are 0.062, 0.042, 0.027, 0.016 and 0.011 respectively. These predicted values by the pedigree approach are almost identical to the simulated values by the marker approach as shown in Figure 2. The naïve estimator is not only biased, but also imprecise. Overall, it has a much larger

RMSE than the bold and weighted estimators. The latter two estimators give nearly indistinguishable results for all samples and true $F_{ST}$ values.

*Expected heterozygosity*

An important measurement of within population genetic variation at a marker locus is expected heterozygosity ($H_e$), or gene diversity (Nei 1973). It is defined as the probability that two homologous genes taken at random (with replacement) from a population at a given locus are not identical in state. Equivalently, it is the expected frequency of a heterozygote individual if the population is at HWE. Denoting the population frequencies of alleles at a locus as $p_i$ (where $i$=1~$k$, $0 < p_i < 1, \sum_{i=1}^{k} p_i \equiv 1$), $H_e$ is calculated by

$$H_e = \sum_{i=1}^{k} \sum_{j=i+1}^{k} 2p_i\, p_j = 1 - \sum_{i=1}^{k} p_i^2. \quad (13)$$

$H_e$ varies in the range 0~1, determined by the number of alleles, $k$, and the evenness of the frequencies of the $k$ alleles. More alleles and/or more even allele frequencies lead to a higher $H_e$.

As a measure of genetic variation, $H_e$ is informative about the demographic history and genetic structure of a population. If a population is found to have a low $H_e$ relative to other populations of the same species at the same marker locus, then it is likely to have a low effective population size, to have experienced a recent bottleneck, and to be isolated from other populations. $H_e$ is also informative about population structure when it is compared with the observed heterozygosity, $H_o$. If $H_e > H_o$, the apparent deviation from HWE could be caused by inbreeding, hidden population subdivision, and genotyping artefacts (e.g. null alleles, allelic dropouts). If $H_e < H_o$, the apparent deviation from HWE could indicate the presence of hybridization or admixture in the population.

Now let us consider the effect of sampling ECR on $H_e$. Suppose $n$ diploid individuals are drawn from a population at HWE for a locus with $k$ codominant alleles. An unbiased allele frequency estimate, $\hat{p}_i$, can be made from the sampled $n$ individuals by eqn (4). Given $\hat{p}_i$, an unbiased estimator of $H_e$ is (Nei & Roychoudhury 1974)

$$\widehat{H}_e = \frac{2n}{2n-1}\left(1 - \sum_{i=1}^{k} \hat{p}_i^2\right). \quad (14)$$

Estimator (14) will become biased when sampled individuals are either inbred or related, although $\hat{p}_i$ from (4) is always unbiased. Suppose the presence of ECR in a sample

results in an average PIBD of $\theta$, which is greater than 0 as expected for the corresponding PIBD value for a randomly drawn sample. With no inbreeding, we have $F_{IS} = -\theta/(1-\theta)$ from (8), and the observed heterozygosity is expected to be

$$\widehat{H}'_e = \frac{2n}{2n-1}\sum_{i=1}^{k}\sum_{j=i+1}^{k} 2\hat{p}_i\,\hat{p}_j\,(1-F_{IS}) = \frac{2n}{(2n-1)(1-\theta)}\left(1-\sum_{i=1}^{k}\hat{p}_i^2\right). \quad (15)$$

When $\theta$ =0, (15) reduces to (14) as expected. Except for the small sample correction factor $2n/(2n-1)$, (15) is identical to DeGiorgio & Rosenberg's (2008) equation (8) derived by a much more complicated approach.

Hereafter, estimators like (15) which use allele frequencies estimated by the naïve approach (i.e. assuming no relatedness) and account for sample relatedness in estimating target parameters are called cautious estimators.

To understand the impact of ECR on $H_e$ estimates, let us consider a numerical example. For simplicity, consider a locus with two codominant alleles, A and a, whose population frequencies are $p$ and $q$=1-$p$, respectively. If $n$ individuals are sampled at random from a large population at HWE, the counts $n_2$, $n_1$ and $n_0$ of genotypes AA, Aa, and aa, respectively, follow the multinomial distribution

$$\Pr[n_2, n_1, n_0 | p, n] = \frac{n!}{n_2! n_1! n_0!}(p^2)^{n_2}(2pq)^{n_1}(q^2)^{n_0}. \quad (16)$$

Suppose $m$ individuals are sampled at random from a single full-sib family in the population. The counts $m_2$, $m_1$ and $m_0$ of genotypes AA, Aa, and aa, respectively, in the sample of $m$ full siblings have the probability

$$\begin{aligned}\Pr[m_2, m_1, m_0 | p, (m, \mathrm{FS})] &= \frac{m!}{m_2!\,m_1!\,m_0!}\left[p^4 0^{m-m_2} + 4p^3 q\left(\frac{1}{2}\right)^{m} 0^{m_0} + 2p^2q^2 0^{m-m_1}\right. \\ &\left. + 4p^2q^2\left(\frac{1}{2}\right)^{2m-m_1} + 4pq^3\left(\frac{1}{2}\right)^{m} 0^{m_2} + q^4 0^{m-m_0}\right], \end{aligned} \quad (17)$$

derived by considering the six possible types of full-sib families (characterized by parental genotype combinations) and the probabilities of obtaining the sample from these families (Appendix S2, Supporting information). In (17), $0^x \equiv 0$ when $x \neq 0$ and $0^x \equiv 1$ when $x = 0$.

In the combined sample of $n$+$m$ individuals, the probability of the counts $x_2$, $x_1$ and $x_0$ (where $x_2, x_1, x_0 \geq 0$ and $x_2 + x_1 + x_0 \equiv m + n$) of genotypes AA, Aa, and aa, respectively, is

$$\Pr[x_2, x_1, x_0 | p, n, (m, \mathrm{FS})] = \sum_{n_i, m_i} \Pr[n_2, n_1, n_0 | p, n] \Pr[m_2, m_1, m_0 | p, (m, \mathrm{FS})], \quad (18)$$

where the summation is over all possible $n_i$ and $m_i$ values (where $i = 2, 1, 0$), with constraints $n_2 + n_1 + n_0 \equiv n$, $m_2 + m_1 + m_0 \equiv m$, $n_i$=[0, $x_i$] and $m_i \equiv x_i - n_i \geq 0$ for $i = 2, 1, 0$. In (18), $\Pr[n_2, n_1, n_0 | p, n]$ and $\Pr[m_2, m_1, m_0 | p, (m, \mathrm{FS})]$ are calculated by (16) and (17), respectively.

Given $x_2, x_1, x_0$, $H_e$ can be estimated by (4) and (14) when all sampled individuals are assumed unrelated, and by (4) and (15) when the relatedness among individuals, $\theta$, is known and accounted for. In the combined sample containing $n$ unrelated individuals and $m$ full siblings, the frequency of full sib pairs is $Q[n, m] = \frac{m(m-1)/2}{(n+m)(n+m-1)/2}$. The coancestry is ¼ and 0 for a full sib pair and an unrelated pair, respectively. The average coancestry of the combined sample is thus $\theta = Q[n, m]\left(\frac{1}{4}\right) + (1 - Q[n, m])(0) = \frac{m(m-1)}{4(n+m)(n+m-1)}$. The distributions of the two $H_e$ estimators are calculated by (16-18).

The distributions of $H_e$ estimates calculated by (14) and (15) are shown in Figure 3 (upper panel) for a combined (C) sample with $n$=$m$=20, and an unrelated (U) sample obtained by removing all but one siblings from sample C (i.e. $n$=20 and $m$=1). For both samples, the population allele frequency is $p$=0.5 and thus the actual value of $H_e$ is 0.5. With sample C, both estimators (14) and (15) are highly dispersed. In general, accounting for relatedness by (15) gives unbiased $H_e$ estimates, while ignoring relatedness by (14) underestimates $H_e$. The U sample has only 21 individuals, roughly half of that of C sample, 40. However, it yields much less dispersed $H_e$ estimates around the true value of 0.5 than C sample.

Values of $H_e$ estimated from sample U or from sample C by estimator (15) are always unbiased, irrespective of the actual parameter values of $p$ and $H_e$ (Figure 3, middle panel). The RMSE for sample U is always much smaller than that for sample C obtained by either estimator (14) or estimator (15) (Figure 3, lower panel). The maximum difference occurs when true $H_e$ = 0.269 (which is realised when $p = 0.16$ or $p = 0.84$), while the minimum difference occurs when true $H_e$ is either the maximal, 0.5 (i.e. when $p = 0.5$) or close to the minimal, 0 (i.e. when $p \to 0$ or $p \to 1$). The large accuracy advantage of sample U (or the

bold approach) over sample C (or the naïve and cautious approaches) is due to largely its much smaller sampling variance (Figure 3, upper panel) and secondarily its smaller bias (Figure 3, middle panel). This is remarkable considering that sample U is only half as large as sample C.

Although the cautious estimator (15) is unbiased, it has a larger sampling variance and is thus less accurate than the naïve estimator (14) in most of the parameter range of $p$=[0, 1]. Parameters like $H_e$ are mainly determined by allele frequencies. As long as inappropriately estimated (e.g. by the naïve approach) allele frequencies are used in estimating these parameters, they cannot be estimated accurately, despite they can be estimated without bias by accounting for the relatedness structure of the sample.

Simulations were also conducted to check the analytical results shown in Figure 3, and to investigate how much $H_e$ estimates can be improved by using the weighted allele frequency estimator, (5). In addition to sample C which had a single large full-sib family, the simulation also considered a sample containing several small sib families, with structure E5=$\{n, m_1, m_2, m_3, m_4, m_5\}$ = {20, 4, 4, 4, 4, 4}. The simulation results were almost identical with the analytical results (not shown Figure 3 for clarity) for samples C and U. Furthermore, applying the weighted estimator to C samples yields the most accurate results in all cases. Confirming the analytical results, the cautious estimator (15) that accounts for relatedness but uses inappropriately (by the naïve approach) estimated allele frequencies is unbiased but often yields the worst estimates with the highest RMSE.

The numerical example shows that, when ECR are sampled but are ignored, allele frequency estimate $\hat{p}_i$ by (4) is unbiased but $\hat{p}_i^2$ is biased. The bias leads to an underestimate of $H_e$. As a result of this and a higher sampling variance, the naïve estimator usually has a high RMSE, especially when large full-sib families are sampled. If the relatives in a sample can be identified and thus the average relatedness of all sampled individuals is calculated and used in estimator (15), then unbiased $H_e$ can be estimated. However, most often (15) has a lower accuracy than (14) because of its larger variance. When a sample is dominated by a few large sib families, a better option is the bold approach which identifies and removes all but one sibling from each family, and uses only unrelated individuals in estimating $H_e$. Such a strategy reduces sample size substantially, but still yields unbiased and more accurate estimates of $H_e$. However, when a sample contains numerous small sib families, the naïve method yields more accurate $H_e$ estimates than the bold method. Over widely different

distributions of family sizes in a sample, the best estimates of $H_e$ are obtained by identifying and accounting for (not removing) close relatives in calculating allele frequencies, which are then used in calculating $H_e$.

*Effective number of alleles*

Effective number of alleles, $A_e$, was first proposed by Kimura and Crow (1964) and has been widely used as a measurement of genetic variation at a locus. It is defined as the number of equally frequent alleles it would take to achieve the same $H_e$ as in a study population where the allele frequencies are not equal. The mathematical expression is

$$A_e = \frac{1}{1-H_e} = \frac{1}{\sum_{i=1}^{k} p_i^2}, \quad (19)$$

where $p_i$ is frequency of allele $i$ (1~$k$) at a locus with $k$ alleles.

Sampling but ignoring ECR results in an underestimate of $H_e$, and thus also an underestimate of $A_e$ as is clear from (19). For a sample containing ECR such that the average PIBD among sampled individuals is $\theta$, the gene diversity can be estimated by (14) and (15) when relatedness is ignored and taken into account respectively. Replacing $H_e$ in (19) by estimator (14) and (15) yields estimator $\hat{A}_{e(14)}$ which ignores relatedness and estimator $\hat{A}_{e(15)}$ which accounts for relatedness, respectively. For both estimators of $\hat{A}_{e(14)}$ and $\hat{A}_{e(15)}$, the allele frequencies are estimated by the naïve estimator (4) which ignores ECR.

The distributions, means and RMSEs of $\hat{A}_{e(14)}$ and $\hat{A}_{e(15)}$ for the same numerical example considered in Figure 3 for $H_e$ mirror those for $H_e$ shown by Figure 3, as expected. When close relatives are included in a sample but are ignored, the naïve estimator $\hat{A}_{e(14)}$ is downwardly biased. It is however still more accurate than $\hat{A}_{e(15)}$ due to its lower sampling variance. Identifying and removing full siblings leads to a much smaller sample, but gives both unbiased and much more accurate estimates of $A_e$. Not surprisingly, the same conclusion as that for $H_e$ can be reached. Over widely different distributions of family sizes in a sample, the most accurate estimates of $A_e$ are obtained by identifying and accounting for close relatives in calculating allele frequencies in a likelihood framework, which are then used in calculating $A_e$.

*Observed number of alleles*

The number of distinctive alleles observed at a locus in a sample of individuals, $A_o$, is another important measure of genetic variation. Also called allelic richness, $A_o$ is more sensitive to population demographic changes than $H_e$. A population bottleneck can cause a drastic reduction in $A_o$ with little effect on $H_e$, especially for a highly polymorphic locus with many rare alleles. It thus creates an apparent excess of $H_e$ compared with the heterozygosity expected from $A_o$ if the population were at mutation and drift equilibrium. The excess calculated from a set of markers can then be used as signal to infer population bottlenecks (e.g. Cornuet & Luikart 1996). $A_o$ is also believed to be a measure of genetic diversity more appropriate than $H_e$ for indicating a population's long-term potential for adaptability and persistence (Allendorf 1986; Caballero & García-Dorado 2013).

$A_o$ reflects both the population and sample properties. It is much more affected by sampling intensity than $H_e$, and is a non-decreasing function of sample size $n$. Consider a locus with $k$ alleles of frequencies $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ in a population at HWE. When $n$ diploid individuals are drawn at random from the population, the counts of different alleles, $\mathbf{x}=(x_1, x_2, \ldots, x_k)$ where $x_i \geq 0$ and $\sum_{i=1}^{k} x_i = 2n$, in the sample follow the multinomial distribution

$$\Pr[\mathbf{x}|2n, \mathbf{p}] = \frac{(2n)!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} . \quad (20)$$

Using (20), we can calculate the probability that $A_o$ (=1, 2, …, $k$) is observed in a sample of $n$ individuals. As an example, herein I consider the events that not all $k$ known alleles are included in the sample, $A_o < k$. The probability of the events is

$$\Pr[A_o < k|2n, \mathbf{p}] = 1 - \sum_{\mathbf{x}} \frac{(2n)!}{\prod_{i=1}^{k} x_i!} \prod\nolimits_{i=1}^{k} p_i^{x_i}, \quad (21)$$

where the summation is over $\mathbf{x}=(x_1, x_2, \ldots, x_k)$ with the constraints $x_i > 0$ and $\sum_{i=1}^{k} x_i = 2n$.

A sample with ECR would have a lower $A_o$ and a higher $\Pr[A_o < k|2n, \mathbf{p}]$ than a sample containing the same number of unrelated individuals. Effectively, close relatives provide correlated information about $A_o$ and reduce the effective sample size, $n_e$. Let us consider a simple example. Suppose a sample contains $n$ individuals drawn at random from a large population, and $m$ individuals drawn at random from a full sib family in the same population. The total number of genes in the combined sample is $2(n + m)$, but the effective

number of genes in the sample, $n_e$, is smaller. The number of independent (i.e. not IBD) genes in the $m$ full siblings is $x$ with probability

$$\Pr[x|q,2] = \frac{2!}{(4-x)!(x-2)!} q^{4-x}(1-q)^{x-2}, \quad (22A)$$

where $q = 2^{1-m}$ (Appendix S3, Supporting information) and $x$=2, 3 and 4. For example, the probabilities of $x$=2, 3, 4 independent genes are 0.25, 0.5, and 0.25 respectively in a sample of $m$=2 full siblings, and are 0.004, 0.117 and 0.879 respectively in a sample of $m$=5 full siblings. Therefore, the combined sample with $2(n+m)$ genes has $n_e = 2n+2$, $2n+3$ and $2n+4$ with the same corresponding probabilities. Considering the three cases together, I obtain the probability of $A_o < k$

$$\Pr[A_o < k|n,m,\mathbf{p}] = \sum_{x=2}^{4} \Pr[x|q,2] \Pr[A_o < k|2n+x,\mathbf{p}], \quad (22B)$$

where $\Pr[x|q,2]$ and $\Pr[A_o < k|2n+x,\mathbf{p}]$ are calculated by (22A) and (21), respectively, for $x$=2, 4, and 4.

Figure 4 shows the effects of allele frequency distribution, sample size, and the inclusion of full siblings on $A_o$, quantified by equation (22B). A locus with 4 alleles of frequencies $\mathbf{p_1}$=(0.1, 0.2, 0.3, 0.4) and $\mathbf{p_2}$=(0.01, 0.02, 0.03, 0.94) is considered. The 2nd frequency distribution is much more skewed than the first. Two samples of diploid individuals are used to obtain $A_o$. The first, unrelated sample, contains $n$ individuals drawn at random from a large population at HWE. The second, related sample, contains $n/2$ individuals drawn at random from the large population and $n/2$ individuals drawn at random from a full-sib family in the same population. While both samples have the same size of $2n$ genes, the related sample has an effective size, $n_e$, which is only slightly larger than half of that of the unrelated sample ($n_e = 2n$). As is clear from Figure 4, the related sample has a much larger $\Pr[A_o < k]$ than the unrelated sample, except when $n$ is trivially small (i.e. < 4). The maximal difference occurs when sample size $n$ is intermediate. For a given number of individuals, fewer than $k$=4 alleles are more likely to be observed when close relatives are included in the sample, and when rare alleles exist at the locus. Some simulations confirm the analytical results and are also shown in the figure.

*Hardy-Weinberg equilibrium (HWE)*

At HWE, the two alleles at a locus in any diploid individual are independent. As a result, the frequency of a genotype is equal to the product of the frequencies of the two alleles in the genotype. These genotype frequencies are called Hardy-Weinberg proportions, which are realized in a large population by one generation of random mating if there is no difference in allele frequencies between males and females and there is no selection at the locus. A test of the deviation from HWE is most often one of the first analyses conducted on marker genotype data. Such a test is revealing not only for population properties (such as non-random mating and admixture), but also for marker properties or genotyping abnormalities (e.g. allelic dropouts, null alleles, see Bonin *et al*. 2004).

As analysed above, ECR in a sample cause a reduction in the estimated $F_{IS}$ and $H_e$, and correspondingly an increase in expected frequencies of homozygotes given the observed allele frequencies. Therefore, when a sample containing ECR is drawn from a large population at HWE and is tested for HWE, the equilibrium is likely to be falsely rejected, leading to the possible false conclusion that the population is sub-structured, avoids close relative mating, and/or the markers have genotyping problems such as allelic dropouts and null alleles.

It is not easy to quantify analytically the effect of ECR on HWE tests. A simulation was conducted to investigate how often HWE was falsely rejected when ECR were included in a sample drawn from a population under HWE. I assumed a large, monoecious, random mating (including selfing) population at HWE. A sample of 80 unrelated individuals (called unrelated sample hereafter) were drawn at random. Additionally, a number of 5, 10, 20, 40 or 80 full siblings were also drawn from a single family in the same population, and were included in the unrelated sample to form the combined sample. The combined sample and the unrelated sample were then independently tested for HWE at a marker locus with simulated allele frequencies **p**=(0.1, 0.2, 0.3, 0.4) or **p**=(0.01, 0.02, 0.03, 0.94). For each sample, exact test for HWE was conducted using the permutation approach (Guo & Thompson 1992; Weir 1996) with $10^5$ replicates. The proportion of replicate datasets (total = 10000) detected to depart from HWE at the 5% significance level was reported separately for combined samples and unrelated samples (Figure 5, upper panel). It is clear from the graph that full siblings included in the combined sample cause an elevated frequency at which HWE is falsely rejected. The larger the proportion of full siblings in a sample, the greater is the rate of rejection of HWE. The effect of full siblings is greater with a more even allele frequency

distribution. When full sibs are removed and only unrelated individuals are used for HWE tests, the rejection rate is always close to the expected value of 0.05.

*Linkage disequilibrium*

The association of alleles within an individual at a single locus measures the deviation from HWE. Alleles within an individual at different loci may also be associated, a phenomenon called linkage disequilibrium (LD), because of several mutually nonexclusive factors such as admixture, random genetic drift, non-random mating, and selection. The loci are not necessarily linked physically to become associated, although physical linkage leads to a higher chance and a greater extent of association (Hill & Robertson 1968). LD is defined and measured by the deviation of the observed genotype or gametic frequency at two or more loci from the product of allele frequencies. The degree of LD can be estimated and tested for statistical significance directly from the genotype frequencies in a sample of individuals taken from the population, whether the gametic phase is known or not.

In some analyses, the degree of LD is estimated from a sample of genotypes and is used for inferring population parameters such as effective population size (e.g. Hill 1981). In many other analyses, however, the degree of LD is of no direct relevance but is tested for statistical significance. If two loci are found to deviate significantly from linkage equilibrium (LE), then they could be suspected to be linked, under selection, affected by genotyping errors, or the population is under non-random mating or is affected by admixture (Slatkin 2008). To avoid increased type I errors in downstream analyses, it has been suggested to discard one locus in a pair of loci with significant LD (e.g. Selkoe & Toonen 2006). Herein I show non-random sampling with respect to kin can also lead to significant LD across loci.

I conducted simulations to investigate the effect of sampling excessive full siblings on LD tests. A fixed number of 80 diploid individuals were drawn at random from a large population at HWE and LE at two loci with identical allele frequency distribution $\mathbf{p}_1$=(0.1, 0.2, 0.3, 0.4) or $\mathbf{p}_2$=(0.01, 0.02, 0.03, 0.94). These allele frequency distributions were chosen to represent even ($\mathbf{p}_1$) and skewed ($\mathbf{p}_2$) distributions. These 80 individuals constitute the unrelated sample. A number of 5, 10, 20, 40, or 80 full siblings were taken from a single full-sib family in the same population and were added to the unrelated sample to constitute the combined sample. Both unrelated and combined samples were tested for LD using the exact test with the permutation approach (Weir 1996) with $10^5$ replicates. The proportion of replicate datasets (total = 1000) in which LD was significant at the 5% significance level was

compared between the unrelated and combined samples (Figure 5, lower panel). The pattern in LD test results mirrors that in HWE. In short, the inclusion of full siblings yields false significant LD. The more evenly distributed the allele frequencies are and the more full siblings included in a sample are, the larger is the frequency at which LD is tested significant.

**Discussion**

In almost all population genetics analyses, sampling of individuals is assumed to be at random such that the population is adequately represented by the sample and its properties, or the underlying genetic mechanisms leading to the properties, can be inferred reliably from the sample. For marker-based genetic studies and many others, random sampling is often implicitly with respect to kinship. Including too much or too little kin in a sample compared with that under random sampling could lead to biased and/or low-precision estimates of population parameters, as shown in this study. My results show that, given the large and wide effects of non-random sampling, every effort should be made to ensure random sampling of individuals in the experimental design and implementation stages. Once a sample is obtained and genotyped, effort should also be made to detect possible cryptic close relatives included in the sample by analysing the marker data. This becomes a routine of data quality control in human genetics, but awaits applications to other organisms in molecular ecology studies. When ECR are detected, one needs to consider how best to deal with them. Different analyses are affected differently by ECR in a sample, and should ideally be conducted using different strategies to deal with non-random sampling. The optional strategies are ignoring close relatives (naïve estimator), identifying and removing close relatives (bold estimator), estimating allele frequencies by ignoring relatives but then estimating the target parameter (e.g. $H_e$) by accounting for relatives (cautious estimator), and estimating allele frequencies by accounting for relatives, which are then used in calculating the target parameters (weighted and likelihood estimators).

My analysis shows that the choice of the four estimators depends on the parameter being estimated, as well as the actual genetic structure (family size and distribution) of a sample. Since $A_o$ estimate is a non-decreasing function of sampled individuals (Figure 4) regardless of their relatedness, it is best to disregard relatedness and use all sampled individuals (i.e. the naïve estimator) in $A_o$ analysis. In contrast, inclusion of close relatives distorts genotype and allele frequencies at a single locus and at multiple loci and thus leads to the sporadic rejections of HWE and LE (Figure 5). The best option is to identify and remove

close relatives before conducting HWE and LE tests. In the case that a large proportion of sampled individuals are relatives, it may be infeasible to remove all relatives as the sample size would become too small. If the pedigree of the sampled individuals is available, HWE can be tested by accounting for relatedness (Bourgain *et al*. 2004).

Many parameters, such as *F* statistics, $H_e$ and $A_e$, are highly dependent on allele frequency estimates. All of the 4 options for handling ECR are applicable to estimating these parameters from a sample containing relatives. The naïve approach, which ignores relatedness completely and estimates the parameters as if all sampled individuals were unrelated, leads to unbiased (but imprecise, see Figure 1) allele frequency *p* estimates. However, it results in biased and imprecise estimates of higher order terms of *p*, such as $p^2$, and thus inaccurate estimates of parameters (e.g. *F* statistics, $H_e$ and $A_e$) involving these terms. The bold approach, which identifies and removes all but one relative in a family before conducting an analysis, always yields unbiased estimates of allele frequency *p* and target parameters (e.g. $F_{ST}$). However, due to the reduction in sample size, it could lead to more accurate and less accurate parameter estimates than the naïve approach when family sizes are highly unbalanced and are small and balanced, respectively. The cautious approach, which estimates allele frequencies by ignoring relatedness (i.e. naïve approach) but estimates the focal parameters by accounting for relatedness, yields unbiased but usually imprecise parameter estimates. The likelihood approach, which accounts for relatedness in the likelihood estimates of allele frequencies and uses the frequencies in calculating focal parameters, usually performs the best for those parameters that are dependent allele frequencies and their higher order terms (such as *F* statistics, $H_e$ and $A_e$). Some of the results comparing the above four options are shown in Figures 1-3.

In this paper, I focused on the most popular and elementary population parameters analysed from marker data. However, markers enable the analyses of so many parameters that it is impossible to consider all of them herein. The effects of ECR on these parameters are likely to be qualitatively similar to what I have shown in this study. For example, marker data can be collected from 2 (or more) samples taken from the same population and separated by a few generations. These temporal samples can be used to calculate a variant of Wright's standardized allele frequency variance, *F*. The differences between *F* and $F_{ST}$ are that *F* measures the differentiation between temporally spaced samples from the same population while $F_{ST}$ measures the spatial differentiation between populations at the same time, and that *F* is determined by both samples and populations while $F_{ST}$ is expected to be determined by

populations only. $F$ estimates can be used to calculate the average effective size ($N_e$) of the population during the sampling interval (e.g. Nei & Tajima 1981; Waples 1989). Like many other analyses, $F$ and $N_e$ are estimated by assuming random sampling for all temporal samples. When ECR are included in any of the temporal samples, $F$, like $F_{ST}$, will be overestimated and $N_e$ underestimated if the relatives are ignored. Similarly, $N_e$ can be estimated from the LD observed in a single sample of individuals taken at the same time from a population (Hill 1981; Waples & Do 2008). Excessive close relatives inflate the estimates of LD, as shown in Figure 5, and thus deflate $N_e$ estimates. Ideally for an accurate estimate of $N_e$, excessive relatives included in a sample should be identified and accounted for in calculating allele frequencies and thus $F$ and $N_e$ in the temporal method. Because it is unclear how to accommodate relatives in calculating LD, it is unclear how best to deal with ECR in estimating $N_e$ from LD. The simulations of Waples and Anderson (2017) provide some insight into this problem.

The estimation of different population parameters varies in sensitivity to non-random sampling, and is thus affected differently by ECR. One parameter critically dependent on random sampling for unbiased and accurate estimation is $N_e$ when it is estimated from the sibship frequency approach (Wang 2009). As shown by Waples and Anderson (2017), removing all siblings from a sample will lead to an overestimated $N_e$. This is true even when a population is very large such that the expected sibship frequency in the population or in a randomly drawn sample is very small. The result is not surprising, because sibship frequency in a randomly drawn sample is inversely proportional to $N_e$ of the population (Wang 2009). Non-random sampling with too many or too few siblings, or manipulating sibling frequencies in a truly randomly drawn sample by removing some identified siblings, will result in a biased estimate of sibling frequency and thus of $N_e$. When all siblings are identified and removed from a sample, the estimate of $N_e$ from the sibship frequency approach is always infinite, irrespective of the actual $N_e$ and sample size. For the estimation of parameters that are critically dependent on the relatedness structure of a sample, such as $N_e$ estimated from the sibship frequency approach, it is difficult or impossible to use any of the strategies (e.g. removing siblings) investigated in this study to eliminate the bias. The only effective strategy is to ensure, in experimental design and implementation, that sampling is indeed at random with regard to relatedness.

It is not abnormal that a sample contains close relatives. A sample of individuals drawn truly at random from a population could contain close relatives. The larger the sample

and the smaller the population is, the more relatives a random sample is expected to contain. For example, siblings would occur at an expected frequency in the order of $1/N_e$ in both the population and in a random sample from the population, where $N_e$ is the effective size of the sampling population (Wang 2009). A sample becomes abnormal when it contains too many or too few relatives than a truly randomly drawn sample of the same size. Such a sample is unrepresentative of the source population, and as a result provides biased and imprecise parameter estimates as shown in this study. However, as long as the frequency of relatives in a (non-randomly drawn) sample does not deviate much from that expected when sampling is at random, the estimates of the parameters considered in this study should not be biased much. Unfortunately, it is difficult to judge whether a sample is abnormal (i.e. containing too few or too many relatives) or not, because the expected frequency of relatives in a natural population is seldom known. In this difficult but realistic situation, should we always endeavour to identify and deal with close relatives in a sample no matter how frequent they are and how frequent they should be under random sampling? Is it possible that removing or accounting for relatives in a sample could cause the opposite and serious effects (e.g. decreasing $F_{ST}$)? I have not considered the effects of deficient close relatives (because they are under-sampled or are identified and removed) in a sample drawn from a small population (i.e. with a substantial expected frequency of close relatives). However, the effects are likely to be small and negligible on the analyses of those population parameters considered in this study, because the frequency of close relatives in a population is expected to be small, in the order of $1/N_e$. Except for extremely small populations (say, $N_e < 50$), blindly identifying and accounting for (or removing) close relatives, few or many, in a sample has a high probability of beneficial effects and a low probability of small and negligible adverse effects on estimating those population parameters shown in this study. For a population with $N_e \geq 50$, the maximal difference in sibling frequency is $\leq 0.02$ between the population and a sample without siblings (because siblings are not sampled, or removed), but is $\geq 0.98$ between the population and an inadvertently acquired sample with ECR. The maximal difference in the latter case is realized when all sampled individuals come from a single sib family, which is possible for fecund species with clustered distributions of relatives, regardless of $N_e$.

In all numerical examples of this study, I used full siblings to represent close relatives because they are highly related, and are frequent in juveniles of highly fecund species (Hansen *et al*. 1997; Goldberg & Waits 2010) and could constitute a large proportion of sampled individuals. Loosely speaking, the impact of non-random sampling of relatives

depends on the product of the relatedness and the frequency of the relatives in a sample, or the average relatedness between sampled individuals. Other kinds of close relatives have similar effects to full siblings. They are however expected to be either less frequent or less related, and thus to be less important than full siblings. For example, although parent-offspring pairs have the same relatedness as full-sib pairs, they are expected to be less frequent in natural populations and in samples. Half sibs could be more abundant than full sibs in some species, but their relatedness is only half of that of full sibs. Similarly, cousins can also be more copious than full sibs, but their relatedness is a small fraction of that of full sibs. This does not mean full sibs are always the most important relatives that determine the effect of non-random sampling. For certain species under certain situations, half sibs, cousins or other types of relatives singly or collectively may overwhelm the impact of full sibs. However, the general patterns and conclusions about the effects of non-random sampling obtained in this study using full siblings still apply.

It should be emphasized that excessive relatives in a sample have *universal* effects on population parameter estimates. In other words, the effects are expected to be the same for all markers, and thus cannot be reduced or removed by simply increasing the number of markers. In fact, the effects on parameter estimation accuracy become more prominent with an increasing number of markers. This is because accuracy (measured by RMSE) is determined by both bias and sampling variance. More markers usually lead to a smaller sampling variance, but cannot reduce bias caused by non-random sampling. Therefore, with an increase in marker number, the accuracy is increasingly determined by bias rather than sampling variance. In other words, with an increase in marker number, bias (due to non-random sampling of individuals) becomes more important in determining parameter estimation accuracy. Therefore, with the use of more markers, unbiased estimators (e.g. those removing relatives or accounting for relatedness) become better than biased but precise estimators (e.g. naïve estimator). In human genetics where now hundreds of thousands markers are routinely used to map genes responsible for inheritable diseases, precision is no longer an issue but bias is. Therefore, it is almost certain that the bold approach of removing close relatives always yields better results than the naïve approach, except in the case where the frequency of relatives in a sample is used directly as the information in estimating parameters such as $N_e$ (Waples & Anderson 2017). On the other hand, more markers mean higher power and accuracy in identifying relatives, and thus potentially better parameter estimates by accounting for the identified relatives.

Excessive relatives in a sample also affect the estimates of marker specific quantities, such as genotyping error rates. Indeed, the deficiency of heterozygotes compared with those under HWE at a marker locus is regarded as signalling the presence of null alleles or allelic dropouts in PCR (Bonin *et al*. 2004) at the locus, and has been used to estimate null allele frequencies (e.g. Brookfield 1996) and dropout rates (e.g. Johnson & Haydon 2007). A critical assumption in these estimation methods is that the sample would be at HWE if there were no marker genotyping problems. Thus the excess of observed homozygosity is due solely to null alleles or allelic dropouts and can be used for estimating their frequencies. With ECR included in a sample, the apparent homozygosity at all marker loci would be reduced as shown in this study. If the sample is naively assumed to be taken at random from a population at HWE, an underestimate of null allele or dropout frequencies would result. In this context, more general and powerful methods have been developed to use known pedigrees for inferring marker genotyping errors (including dropouts, null alleles, false alleles) or for inferring pedigrees and typing errors jointly (e.g. Sobel *et al*. 2002; Wang 2004; Wang & Santure 2009). In these likelihood frameworks, known or unknown relatedness between individuals is not noise but information for identifying erroneous genotypes at particular loci and in particular individuals, and for inferring genotyping error rates at each locus.

**Acknowledgements**

I thank Robin Waples for showing me his unpublished manuscript. I also thank Robin Waples, Eric Anderson, Stephen Spear and two anonymous referees for helpful comments and discussions on some of the topics covered in this study.

## References


Allendorf FW (1986) Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology* **5**, 181–190.

Anderson EC, Dunham KK (2008) The influence of family groups on inferences made with the program Structure. *Molecular Ecology Resources* **8**, 1219-1229.

Boehnke M (1991) Allele frequency estimation from data on relatives. *American Journal of Human Genetics* **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. *Molecular Ecology* **13**, 3261-3273.

Bourgain C, Abney M, Schneider D, Ober C, McPeek MS (2004) Testing for Hardy-Weinberg equilibrium in samples with related individuals. *Genetics* **168**, 2349-2361.

Broman KW (2001) Estimation of allele frequencies with data on sibships. *Genetic Epidemiology* **20**, 307-315.

Brookfield J (1996) A simple new method for estimating null allele frequency from heterozygote deficiency. *Molecular Ecology* **5**, 453-455.

Caballero A, García-Dorado A (2013) Allelic diversity and its implications for the rate of adaptation. *Genetics* **195**, 1373–1384.

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**, 2001-2014.

DeGiorgio M, Rosenberg NA (2009) An unbiased estimator of gene diversity in samples containing related individuals. *Molecular Biology and Evolution* **26**, 501-512.

Goldberg CS, Waits LP (2010) Comparative landscape genetics of two pond-breeding amphibian species in a highly modified agricultural landscape. *Molecular Ecology* **19**, 3650-3663.

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**, 361-372.

Hansen MM, Nielsen EE, Mensberg KL (1997) The problem of sampling families rather than populations: relatedness among individuals in samples of juvenile brown trout *Salmo trutta L*. *Molecular Ecology* **6**, 469-474.

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226.

Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium. *Genetical Research* **38**, 209-216.

Johnson PC, Haydon DT (2007) Maximum-likelihood estimation of allelic dropout and false allele error rates from microsatellite genotypes in the absence of reference data. *Genetics* **175**, 827-842.

Kimura M, Crow J (1964) The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725-738.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* **4**, 981-994.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**, 356-369.

McPeek MS, Wu X, Ober C (2004) Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics* **60**, 359-367.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* **70**, 3321-3323.

Nei M, Chesser RK (1983) Estimation of fixation indices and gene diversities. *Annals of Human Genetics* **47**, 253-259.

Nei M, Roychoudhury AK (1974) Sampling variances of heterozygosity and genetic distance. *Genetics* **76**, 379-390.

Nei M, Tajima F (1981) Genetic drift and estimation of effective population size. *Genetics* **98**, 625-640.

Nicholson G, Smith AV, Jónsson F, Gústafsson O, Stefansson K, *et al*. (2002) Assessing population differentiation and isolation from single nucleotide polymorphism data. *Journal of the Royal Statistical Society Series B* **64**, 695–715

Ott J (1992) Strategies for characterizing highly polymorphic markers in human gene mapping. *American Journal of Human Genetics* **51**, 283-290.

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517.

Rodríguez-Ramilo ST, Wang J (2012) The effect of close relatives on unsupervised Bayesian clustering algorithms in population genetic structure analysis. *Molecular Ecology Resources* **12**, 873-884.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* **9**, 615-629.

Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* **9**, 477-485.

Sobel E, Papp JC, Lange K (2002) Detection and integration of genotyping errors in statistical genetics. *The American Journal of Human Genetics* **70**, 496-508.

Sunnucks P (2000) Efficient genetic markers for population biology. *Trends in Ecology & Evolution* **15**, 199-203.

Voight BF, Pritchard JK (2005) Confounding from cryptic relatedness in case–control association studies. *PLoS Genetics* **1**, e32.

Wang J (2004) Sibship reconstruction from genetic data with typing errors. *Genetics* **166**, 1963-1979.

Wang J (2009) A new method for estimating effective population sizes from a single sample of multilocus genotypes. *Molecular Ecology* **18**, 2148-2164.

Wang J (2012) On the measurements of genetic differentiation among populations. *Genetics Research* **94**, 275-289.

Wang J (2015) Does $G_{ST}$ underestimate genetic differentiation from marker data? *Molecular Ecology* **24**, 3546-3558.

Wang J, Santure AW (2009) Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics* **181**, 1579-1594.

Waples RS (1989) A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* **121**, 379-391.

Waples RS, Do CHI (2008) LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources* **8**, 753-756.

Waples RS, Anderson EC (2017) Purging putative siblings from population genetic data sets: a cautionary view. *Molecular Ecology* **26**, 1211-1224.

Wellcome Trust Case Control Consortium (2007) Genome wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.

Weir BS (1996) *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Associates: Sunderland, MA, USA.

Whitlock MC (2011) $G_{ST}$ and $D$ do not replace $F_{ST}$. *Molecular Ecology* **20**, 1083–1091.

Wright S (1931) Evolution in Mendelian populations. *Genetics* **16**, 97-159.

Wright S (1943) Isolation by distance. *Genetics* **28**, 114–138.

Wright S (1965) The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution* **19**, 395-420.

Zheng G, Freidlin B, Gastwirth JL (2006) Robust genomic control for association studies. *American Journal of Human Genetics* **78**, 350–356.

---

J. Wang is interested in developing population genetics models and methods for analysis of empirical data to address issues in evolutionary and conservation biology.

---

**Supporting Information**

Additional Supporting Information may be found in the online version of this article:

**Appendix S1.** Simulation procedures.

**Appendix S2.** Derivation of equation (17).

**Appendix S3.** The probability of the number of independent genes drawn from a full sib family.

**Data accessibility**

The computer programs (Fortran source codes and executables) for simulating genotype data and investigating the impacts of sampling excessive full sibs on genetic parameter estimation: Dryad DOI: 10.5061/dryad.4pv20