ESTIMATION OF POPULATION ALLELE FREQUENCIES FROM SMALL SAMPLES CONTAINING MULTIPLE GENERATIONS

DMITRY A. KONOVALOV

School of Mathematics, Physics and Information Technology, James Cook University, Townsville, Queensland 4811, Australia

DIK HEG[†]

Department of Behavioural Ecology, Zoological Institute, University of Bern, Hinterkappelen, Switzerland

Estimations of population genetic parameters like allele frequencies, heterozygosities, inbreeding coefficients and genetic distances rely on the assumption that all sampled genotypes come from a randomly interbreeding population or sub-population. Here we show that *small* cross-generational samples may severely affect estimates of allele frequencies, when a small number of progenies dominate the next generation or the sample. A new estimator of allele frequencies is developed for such cases when the kin structure of the focal sample is unknown and has to be assessed simultaneously. Using Monte Carlo simulations it was demonstrated that the new estimator delivered significant improvement over the conventional allele-counting estimator.

1 Introduction

The estimation of population frequencies of codominant genetic markers (*e.g.* microsatellites) from samples with unknown kin structures is of paramount importance to the population genetic studies, since they form the foundation for downstream genetic analyses.[1] The frequencies can be used to estimate, for instance, the genetic distance between two populations, or the effective population size. Similarly, deviations from Hardy-Weinberg Equilibrium (HWE) of these alleles can be used to assess past effects on the genetic structure of the population due to, for instance, genetic drift, inbreeding, and genetic bottlenecks. The population frequencies are normally estimated from a large sample of assumed to be unrelated individuals.[2] In practice, it may be difficult to acquire genotypes from free-living individuals fulfilling this basic assumption of sampling population frequencies and often samples contain a mixture of related genotypes from multiple generations.[3] Currently, it is unknown how the population allele frequencies can be reliably estimated when actual pedigrees within data sets are unknown and have to be assessed simultaneously. Although this may not matter for large sample sizes within a randomly interbreeding population, where all individuals contribute equally to the next generation, this certainly will matter for small samples from populations wherein some individuals are more productive than others.[1] For example, if a sample of 100 individuals consists of 40 full-sibs and 60 unrelated individuals,[1] it is very likely that the sample will

[†] Work partially supported by SNF grant 3100A0-108473.

fail the exact test for HWE,[4] *e.g.* calculated via the GENEPOP program.[5] Such a case is the focus of this study, when the null hypothesis of HWE is rejected (*e.g.* $P < 0.05$), but the sample may still contain sufficient information for the estimation of the population allele frequencies in the HWE sense. That is, the 60 unrelated individuals in the considered example is commonly deemed a "large" sample.[6]

Methods for estimating allele frequencies do exist but they are mostly a by-product of sibship reconstruction.[7-13] However, it is not known if such frequencies could be obtained effectively for a multi generational population sample which could contain any kin groups, such as cousins, half and full sibs including or excluding parental genotypes.[3]

In addition, the generic pedigree reconstruction problem[14] is clearly more difficult than the problem of detecting all unrelated individuals (to be used for allele frequency estimates). Hence there is a much higher chance that the allele frequencies obtained this way would be affected by the pedigree reconstruction errors. Moreover, the population allele frequencies must be estimated iteratively during the sibship reconstruction,[9] thus frequencies' errors feeding into the reconstruction procedure. If incorrectly done, they reduce the reconstruction accuracy drastically, *e.g.* when the frequencies are estimated from the population sample containing a large family of full sibs as in data sets with family sizes of 40,5,2,2, and 1.[11,12,15]

It is important to differentiate the problem at hand from the problem of estimating population allele frequencies when the pedigree of the sampled individuals is known or assumed to be known, in which case population allele frequencies can be calculated exactly.[2,16] In this preliminary study we report for the first time that a robust method for estimation of the outbred population allele frequencies may be possible even when sample genotypes contain individuals from multiple generations and when the actual pedigree is assessed simultaneously using the same genetic markers.

The following is the outline of this study: (1) given the difficulty of inferring allele frequencies and kin structure from the same sample simultaneously, a pair-wise relatedness estimator is developed, which does not require allele frequencies; (2) the structure of the pair-wise relatedness matrix is examined when the sample kin structure is known exactly; (3) using the properties of the relatedness matrix, a new approach is proposed for searching for the largest sample subset, which resembles a set of unrelated individuals; (4) and finally the new approach is tested via Monte Carlo simulations on three different data sets.

2 Method

2.1 *Estimation of Pairwise Relatedness*

Following in some respects Broman[2] and McPeek *et al.*,[16] let a diploid population sample consists of n genotype vectors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ at a single locus with k codominant alleles. The i'th genotype is defined via the number of observed alleles:[17] $\mathbf{x}_i = (..., x_{mi}, ..., x_{m'>mi}, ...)^T = (..., 1, ..., 1, ...)^T$ for heterozygotes and

$\mathbf{x}_i = (...,x_{mi},...)^T = (...,2,...)^T$ for homozygotes, where the rest of the values are zero, and where 'T' denotes the transpose. For example, a genotype (A_1, A_2) is encoded as (1,1,0,0) at a locus with four alleles $\{A_1, A_2, A_3, A_4\}$. Each diploid genotype contains exactly two alleles, $(\mathbf{1}\cdot\mathbf{x}_i) = \sum_{m=1}^{k} x_{mi} = 2$, where $\mathbf{1}$ is the vector of 1's of length k and where the dot-product notation is used for summations when the summation index and range is clear by context, *i.e.* $(\mathbf{x}\cdot\mathbf{y}) = \sum_{m=1}^{k} x_m y_m$.

Let an outbred population (or sub-population) be in HWE and described by the population allele frequencies $\mathbf{p} = (p_1, p_2, ..., p_k)^T$. Then each observed (sample) genotype $\mathbf{x}_i$ could be represented as a sum of two statistically independent gamete vectors $\mathbf{x}_i = \boldsymbol{\varepsilon}_i + \boldsymbol{\varepsilon}'_i$, *i.e.* $x_{mi} = \varepsilon_{mi} + \varepsilon'_{mi}$, obtaining $E(\varepsilon^2_{mi}) = E(\varepsilon_{mi}) = p_m$, $\mathrm{var}(\varepsilon_{mi}) = p_m(1-p_m)$, $E(x_{mi}) = 2p_m$, $E(x^2_{mi}) = 2p_m(1+p_m)$, $E(\mathbf{x}_i\cdot\mathbf{x}_i) = 2(1+\gamma)$, and $\mathrm{var}(x_{mi}) = 2p_m(1-p_m)$.[2] The pairwise relatedness matrix could be defined in the identity-by-descent (IBD) sense[18] via $\mathbf{x}_j = r_{ij}\mathbf{x}_i + (1-r_{ij})\mathbf{z}_{ij}$, where $r_{ii} = 1$, and $\mathbf{z}_{ij}$ is statistically independent of $\mathbf{x}_i$. Then $\mathrm{cov}(x_{mi}, x_{mj}) = 2r_{ij}p_m(1-p_m)$, $E(x_{mi}x_{mj}) = 2[r_{ij}p_m(1-p_m) + 2p_m^2]$ and $E(\mathbf{x}_i\cdot\mathbf{x}_j) = 2(r_{ij}h + 2\gamma)$, where $\gamma = (\mathbf{p}\cdot\mathbf{p}) = \sum_{m=1}^{k} p_m^2$ and $h = 1-\gamma$ are the population homozygosity and heterozygosity of the given locus, respectively.

In practice, the pedigree of a sample is often not known *a priori* and hence the relatedness matrix must be estimated together with the allele frequencies. This could be done by using the following estimators of heterozygosity and relatedness, which do not require allele frequencies. An estimator h' of heterozygosity at a locus (and hence homozygosity via $\gamma' = 1-h'$) is given by $h' = \sum_{i=1}^{n} u_i h_{ii}$, where the weights $(u_1, u_2, ..., u_n)^T$ are normalized by $\sum_{i=1}^{n} u_i = 1$, and where $h_{ii} = 1$ and $h_{ii} = 0$ for heterozygotes and homozygotes, respectively. If the relatedness matrix $\mathbf{r} = \{r_{ij}\}$ were known, the most optimal weights could be found by minimising $\mathrm{var}(h')$. Since $\mathbf{r}$ is not known, the equal weights $u_i = 1/n$ are used, which yield an unbiased, but not necessarily the most efficient, estimator of heterozygosity in the absence of allele frequencies. The estimate at a locus simply equals to the number of observed heterozygotes averaged over the sample size *n*. Assuming unlinked loci, for multilocus genotypes $\mathbf{X}_i = \{\mathbf{x}_i(1), \mathbf{x}_i(2), ..., \mathbf{x}_i(L)\}$, the $h' \equiv h'(l)$ estimator is averaged across loci obtaining $H = \sum_{l=1}^{L} h(l)/L$ and $H' = \sum_{l=1}^{L}\sum_{i=1}^{n} h_{ii}(l)/(nL)$, where $E(H') = H$ and $\mathrm{var}(H') = \sum_{l=1}^{L} \mathrm{var}[h'(l)]/L^2$, *i.e.* the estimate equals to the number of observed heterozygotes averaged over the sample size *n* and number of loci *L*. An estimator for relatedness is given by $r'_{ij}(h) = 1 - d^2_{ij}/H'$, where $d^2_{ij} = \sum_{l=1}^{L} d^2_{ij}(l)/L$ and $d^2_{ij}(l) = [\mathbf{x}_i(l) - \mathbf{x}_j(l)]^2/4$.

2.2 *Estimation of Allele Frequencies from Known Pedigree*

Following McPeek *et al.*[16] the class of best linear unbiased estimators (BLUE) of allele frequencies is given by

$$\mathbf{q} = \tfrac{1}{2}\sum_{i=1}^{n} w_i \mathbf{x}_i \,, \qquad (1)$$

where the weights $\mathbf{w}^T = (w_1, w_2, ..., w_n)^T$ are normalized by $\sum_{i=1}^{n} w_i = 1$ and hence $E(q_m) = p_m$. The *sample* allele frequencies $\mathbf{s} = (s_1, ..., s_k)^T$ are obtained via $w_i = 1/n$,[6]

$$\mathbf{s} = \sum_{i=1}^{n} \mathbf{x}_i /(2n), \tag{2}$$

which specifies the conventional *allele-counting* estimator. In general, the weights are found by minimizing the variance of each resulting frequency q_m, $\mathrm{var}(q_m) = \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \,\mathrm{cov}(x_{mi}, x_{mj})$. Treating each allele with equal weight at the locus, the problem is transformed into finding the weights that minimize

$$V = \sum_{m=1}^{k} \mathrm{var}(q_m) = \tfrac{1}{2} h \sum_{i=1}^{n}\sum_{j=1}^{n} w_i r_{ij} w_j \,, \tag{3}$$

where the same weights minimize both the absolute and the relative variances, $\sum_{m=1}^{k} \mathrm{var}(q_m)/p_m$. If all individuals are unrelated ($r_{ij} = \delta_{ij}$, $w_i = 1/n$ and $V = h/(2n)$), the commonly used heterozygosity estimator is obtained $h_{\mathrm{Nei}} = 2n(1 - \sum_{m=1}^{k} q_m^2)/(2n-1)$,[6] where δ_{ij} is the Kronecker delta defined by $\delta_{ii} = 1$ and $\delta_{i\neq j} = 0$. The estimator is also known as the gene diversity and is bias corrected for the sample size[19] but not for the sample kin structure.

Since the relatedness matrix r_{ij} is symmetric and positive definite $(V > 0)$, its eigenvectors can always be found and defined as orthonormal $(\boldsymbol{\xi}_\alpha \cdot \boldsymbol{\xi}_\beta) = \delta_{\alpha\beta}$ and sorted by the corresponding real positive eigenvalues $\{0 < \lambda_1 \le \lambda_2 \le ... \le \lambda_n\}$, where $\mathbf{r}\boldsymbol{\xi}_\alpha = \lambda_\alpha \boldsymbol{\xi}_\alpha$. The weights vector in its most generic form is then given by $\mathbf{w} = \sum_{\alpha=1}^{n} C_\alpha \boldsymbol{\xi}_\alpha$ obtaining $V = \frac{1}{2} h \sum_{\alpha=1}^{n} C_\alpha^2 \lambda_\alpha$, subject to the original normalization of the weights $\sum_{i,\alpha=1}^{n} C_\alpha \xi_{i\alpha} = 1$. The minimum is found via Lagrange multiplier obtaining $C_\alpha = \zeta_\alpha/(\eta\lambda_\alpha)$ and $\min(V) = h/(2\eta)$, where $\eta = \sum_{\alpha=1}^{n} \zeta_\alpha^2/\lambda_\alpha$ and $\zeta_\alpha = \sum_{i=1}^{n} \xi_{i\alpha}$. Observing that the inverse matrix of r_{ij} can be written as $(r^{-1})_{ij} = \sum_{\alpha=1}^{n} \xi_{i\alpha}\xi_{j\alpha}/\lambda_\alpha$, the solution can also be expressed via $w_i = \sum_{j=1}^{n} (r^{-1})_{ij}/\eta$, $\eta = \sum_{i,j=1}^{n} (r^{-1})_{ij}$. For multiple loci the resulting formulas for the weights are locus independent, hence the same weights are used to estimate allele frequencies at all loci. The obtained weights (and hence frequencies) provide the *exact* solution to the problem of finding an unbiased estimator of frequencies, which is the most efficient in terms of achieving the smallest possible (absolute and relative) variance of the frequencies in Eq. (3).

When the above formulas are applied (results not shown) to the samples from the *unrelated* data set (see Results section, below) a solution is normally found in the form $w_{i\in U} = 1/u$ and $w_{i\notin U} = 0$ (ignoring rounding errors), where *u* is the number of elements in the subset *U* of all unrelated parents in the sample. Note that the weights represent the *theoretical limit* of the allele frequency inference from a single sample, *i.e.* a biologist would select the same weights if he or she knew which individuals are unrelated parents and which are offspring.

2.3 *Unknown Pedigree*

The population allele frequencies could be calculated exactly from a given relatedness matrix but only if the matrix is positive definite. A sample instance of the r'_{ij} matrix may not be positive definite (regardless of which estimator of *r* is used) and hence it cannot be used directly to infer frequencies. If used, it yields meaningless weights and frequencies essentially amplifying its eigenvectors with near zero eigenvalues (some of them could even be negative; results not shown). This could explain why it was reported that an iterative procedure for estimating relatedness and frequencies yielded worse estimates of relatedness values (and hence the frequencies).[20,21]

This study proposes a new approach where the weights $\{w_i\}$ in Eq. (1) are found by searching for a subset *U* of unrelated individuals in the sample, $\mathbf{q}(U) = \sum_{i\in U} w_i \mathbf{x}_i /(2u)$, where $w_{i\in U} = 1$ and $w_{i\notin U} = 0$, and where *u* is the number of elements in the subset *U*. As indicated earlier, a subset of all unrelated individuals (including unrelated parents) in the sample would give the best theoretically possible estimation of population allele frequencies. If a parent or parents of one or more offspring are missing from the sample, the best one or more representatives of the sibship genotypes should be selected.

The following criterion for selecting *U* is proposed. The weights could be used to estimate average (over loci) heterozygosity via the standard formula $H(U) = 2u[1 - \frac{1}{L}\sum_{l=1}^{L} q_m^2(U)]/(2u-1)$, which is bias corrected for sample size but not for the sample kin structure. The expected value of the estimate is given by $E\big[H(U)\big] = H - R(U)$, where $R(U) = \sum_{i\in U}\sum_{j\neq i\in U} R_{ij}/[u(2u-1)]$, $R_{ij} = Hr_{ij}$, and where $E[H(U)] = H$ if *U* consists of only unrelated individuals, *i.e.* $r_{ij} = \delta_{ij}$ and $R(U) = 0$. Using the unbiased estimator $R'_{ij} = r'_{ij}H'$, the problem is reduced to finding the minimum of

$$R'(U) = \sum_{i\neq j\in U} R'_{ij}/[u(2u-1)]. \qquad (4)$$

The new approach searches for the largest subset of the sample which best resembles a group of mutually unrelated genotypes. In the above analysis, it is implied that *U* with the largest size *u* should be preferred. This condition is specified by the denominator $u(2u-1)$ in Eq. (4). However, a large subset would only be preferred if the resulting *R* increases slower than $u(2u-1)$, *e.g.* if the sample consists of only full siblings, *R* becomes $R(u) = 0.5(u-1)/(2u-1)$ and $R(u-1) < R(u)$ hence the number of selected full sibs will be minimized (subject to the observed R'_{ij}). While the proposed approach minimizes the number of full-sibs, the approach should also maximize the number of mutually unrelated individuals. This is achieved by using $|r'_{ij}|$ instead of r'_{ij}, which prevents the algorithm from achieving zero in Eq. (4) on not the largest subset *U*. If r'_{ij} is used, potentially a small number of negative *r* estimates[1] could cancel out contributions from an equally small number of positive *r* estimates.

Once a solution is obtained, an exact test[4] for HWE could be used via available software programs[5,10] to assess the solution by verifying that the *P* value does not reject

the HWE null hypothesis. If the original sample does not pass the test for some or all of the loci (*e.g.* $P \le 0.05$), the new approach offers a practical alternative if it obtains the subset U that passes the test (*e.g.* $P > 0.05$). Note that the proposed solution could be viewed as an approximation for a more general formulation of the problem: "Find the largest subset U that passes such an exact HWE test as the test of Guo & Thompson",[4] where it is assumed that complete sample does not pass the test.

2.4 *Algorithm*

The above approach, when the kin structure of the sample is not known, could be viewed as partitioning of the given sample into two groups: the group of putative unrelated individuals (the subset U) and the rest of the sample. A set of n elements could be partitioned into the two groups $2^n - 1$ ways, where the single case when all individuals are excluded from U is omitted from consideration. Even though the search space for this problem is "smaller" than the space of the sibship reconstruction problem,[12] it is still non-polynomial and the exhaustive search is possible only for trivially small samples. Moreover, if the relatedness matrix R'_{ij} is viewed as a complete undirected graph (omitting the additional complexity of the dependency on u), the problem of finding a complete sub-graph (*clique*) with the minimum (equivalent to maximum) sum of weights is known to be *NP*-hard,[22] *i.e.* an exact algorithm with polynomial complexity $O(n^{\alpha<\infty})$ does not exist.

Since an exact solution may not be possible, a heuristic approximation is required. One such heuristic for traversing the search space is the simulated annealing technique[23] which was shown to be effective for such related (and more difficult) problems as the sibship[9] and pedigree[14] reconstruction problems. The following algorithm is proposed, where the issue of rare alleles[21] is addressed by ensuring that each putative set U contains at least one instance of every allele observed in the sample. We recognise that the R'_{ij} matrix has a special structure and further study could be done to investigate if a more efficient algorithm exists.

Regarding the design of the algorithm: the main purpose of this study is to develop an algorithm that is implemented in a readily available software program (KINGROUP[10] in our case) so that it could be used by biologists. A typical geneticist/biologist is neither an expert programmer nor a computer scientist, hence we totally agree with the comment of Pearse and Crandall[24] who emphasised that "*improving software usability is essential*". Even though *usability* is often a personal preference, we believe that an algorithm should have as few "magic" numbers controlling the algorithm as possible. Hence the proposed algorithm is controlled by a single parameter, the number of iterations N. The number is set to $N = 100 \times n$, *i.e.* each sample genotype is considered 100 times for inclusion or exclusion (on average). User's access to the computing power controls the quality of the solution, *i.e.* the higher the number N the higher is the probability of finding the optimal solution. When working with a real sample, the

algorithm should be run a number of times with larger N each time to verify that the obtained solution is convergent in N.

The following algorithm was implemented:

1. Generate an initial configuration by placing all available individuals into the group of putative unrelated individuals, $U_{\text{curr}} = \{1, 2, ..., n\}$. Calculate the current cost function $Z_{\text{curr}} = R'(U_{\text{curr}})$, which is always positive due to the use of $|r'_{ij}|$ and plus H' being non-negative by definition.
2. Generate a new configuration by randomly selecting an individual $1 \le i \le n$. If $i \in U$ and the individual can be taken out of the group, *i.e.* each observed allele at each locus appears at least once in U, the individual is removed $(U_{\text{new}} = U_{\text{curr}} - i)$. If $i \in U$ and the individual can not be taken out of the group, another individual is randomly selected. If $i \notin U$, the individual is added $(U_{\text{new}} = U_{\text{curr}} + i)$. Calculate Z_{new} from U_{new}.
3. Calculate relative change via $\Delta Z = (Z_{\text{new}} - Z_{\text{curr}})/Z_{\text{new}}$. If $\Delta Z \le 0$, the new configuration is accepted becoming "current". If $\Delta Z > 0$, accept the new configuration with the probability $\Pr(\Delta Z) = \exp(-\Delta Z/(k_{\text{B}} T_\alpha))$, where T_α is the annealing temperature, k_{B} is originally the Boltzmann's constant which becomes just a scaling constant, and where the original Boltzmann distribution is used as per Kirkpatrick *et al*.[23]
4. Repeat steps 2 and 3 with $T_\alpha = (N - \alpha + 1)/N$, where α is the iteration count. Since $0 < \Delta Z \le 1$, Boltzmann's constant $k_{\text{B}} = 1/\ln 2 = 1.4427$ is selected to achieve $\Pr(\Delta L = 1) = \exp(-1/k_{\text{B}}) = 0.5$, *i.e.* there is at least 50% chance in accepting the new configuration with larger cost value at the beginning of the annealing process.

3 Results and Discussion

Following Wang[21] a triangular population allele frequency distribution was considered, $p_m(l) = 2m/[(1+k)k]$, yielding the locus heterozygosity of $h = 1 - 2(2k+1)/[3(k+1)k]$. The effect of multiple generations was studied by Monte Carlo simulation using f full-sibs in a sample of n individuals. A population sample of n individuals was generated by firstly generating $n - f$ unrelated individuals based on the given population allele frequencies, **p**. Then, two of the individuals were randomly selected and used to generate f full-sibs according to the Mendelian rules of inheritance. The generated set of samples was labelled the *single-family* data set. The theoretically *best* possible estimation of allele frequencies was calculated using only the $n - f$ unrelated individuals, $\mathbf{b} = \sum_{i=1}^{n-f} \mathbf{x}_i / [2(n-f)]$, where, without loss of generality, the unrelated genotypes where labelled from $\mathbf{x}_1$ to $\mathbf{x}_{n-f}$. Assuming the absence of the pedigree information, the frequencies were estimated via the proposed algorithm obtaining the **q** frequencies. The mean squared error (MSE) was used to measure the estimation error, where MSE was averaged across loci, $\text{MSE}(\mathbf{q}) = \sum_{m=1}^{k} \sum_{l=1}^{L} [p_m(l) - q_m(l)]^2 / (kL)$.

The second data set was chosen to contain n unrelated individuals. For this *unrelated* data set the best possible estimator of allele frequencies is identical to the allele-counting estimator, *i.e.* $\mathbf{b} \equiv \mathbf{s}$.

The third data set was based on the experimentally observed allele frequencies from a real biological sample of a cooperatively breeding Lake Tanganyika cichlid (*Neolamprologus pulcher*).[3] The cichlid frequencies are specified at $L = 5$ loci with $\{k_1,...,k_5\} = \{39,34,28,17,10\}$ alleles and corresponding locus heterozygosities $\{h^{(1)},...,h^{(5)}\}=\{0.929, 0.937, 0.847, 0.478, 0.537\}$. This *cichlid* data set is denoted by $G(u,g,s)$, where u is the number of unrelated individuals, g is the number of parental pairs (*i.e.* families), s is the number of full-sibs in the first family. The set is obtained by generating $u+g+1$ *unrelated* genotypes $\{\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_{u+g+1}\}$ according to the specified allele frequencies. Then the $s+i-1$ full-sibs of the i'th group are generated from the $(\mathbf{X}_i,\mathbf{X}_{i+1})$ parental pair.

Fig. 1 presents the root mean square error (RMSE) simulation results: RMSE(**b**), RMSE(**q**) and RMSE(**s**). Fig. 1(a) displays the results for the single-family data set, where $n = 50$ individuals were genotyped with $L = 10$, $k = 10$ ($h = 0.8727$) and variable number of full-sibs f. The results for the unrelated data set are displayed in Fig. 1(b), where each sample contained a variable number of individuals genotyped with $L = 5$ and $k = 20$ ($h = 0.9349$). The cichlid data set was generated as $G(u = 10, g, s = 5)$ with a variable number of families. Each point in Fig.1 was obtained by averaging MSE obtained from 100 independent simulation trials and displaying the square root of the average MSE (RMSE).

The results in Fig. 1 are very encouraging as they clearly demonstrate that the new estimator is more accurate than the conventional allele-counting estimator for "dirty" samples with high level of cross-generational contamination, *e.g.* when 20 or more individuals belong to the next generation. Interesting questions still remain for future studies: (1) How much of the RMSE is due to simulated annealing not being able to find the global optimum, and how much is due to the inaccuracy of the relatedness estimates? (2) How robust is the new frequency estimator to the presence of genotyping errors and/or inbreeding? Note that the new estimator is comparable to or even less accurate than the allele-counting estimator for "clean" population samples (Fig. 1(b)) where the level of cross-generational contamination is small. However such clean samples are likely to pass the HWE test anyway and hence the question of a "better" estimation of population allele frequencies would not arise.

And finally, since the exact HWE test of Guo and Thompson[4] played such an important conceptual role in this study, we would like to comment on the two versions of the HWE test. The first HWE test uses the conventional Monte Carlo (CMC) method and is relatively easy to implement (implemented in KINGROUP[10] and used in this study). This method guarantees P values to within 0.01 with 99% confidence by selecting 17000 simulations regardless of the sample size or the number of observed alleles, hence no "guessing" is required from a software user. Moreover, even Guo and Thompson[4] themselves remarked that the "*method is most suitable for data with a large number of*

alleles but small sample size", which is the focus of this study. The second method uses the Markov Chain (MC) estimation. The main argument in favour of the MC method was that it is faster than CMC when the sample size is moderate or large. This argument does not hold in practice since a diligent user would have to run MC a number of times to ensure that the obtained *P* values are converged, *i.e.* stable to the variations in the three input parameters (dememorization number, number of batches and iterations per batch). In fact, the first method should always be preferred to the second MC method, which is controlled by the three input parameters, which input values are, arguably, meaningless for a typical biologist and can not be deduced easily.

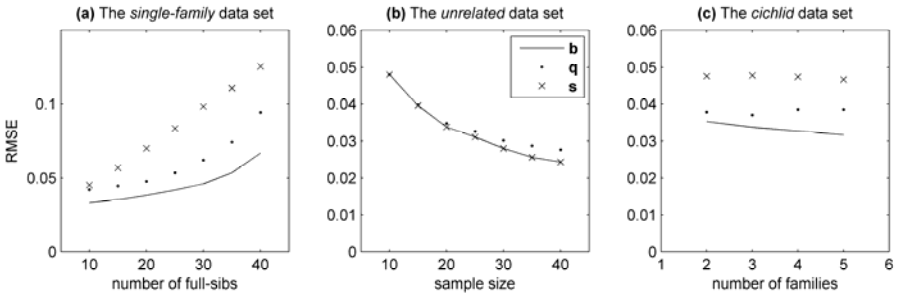

Figure 1. Root mean square error of population allele frequency estimates, where **b** denotes the best possible estimates due to the limited sample size; **q** denotes this study; **s** denotes the allele-counting estimates.

Acknowledgments

This study was partly undertaken when D.K. was on sabbatical leave at the University of Bern. We thank: the University of Bern and James Cook University and, in particular, Michael Taborsky and Bruce Litow for supporting this collaborative project; Peter Stettler for his hospitality; Ross Crozier and Dean Jerry for helpful comments and discussions; and three anonymous reviewers for the thorough review of the earlier version of this manuscript.

References

1. D. A. Konovalov and D. Heg. A maximum-likelihood relatedness estimator allowing for negative relatedness values *Molecular Ecology Notes*, in press, 2007.
2. K. W. Broman. Estimation of allele frequencies with data on sibships. *Genetic Epidemiology*, 20:307-315, 2001.
3. P. Dierkes, D. Heg, M. Taborsky, E. Skubic and R. Achmann. Genetic relatedness in groups is sex-specific and declines with age of helpers in a cooperatively breeding cichlid. *Ecology Letters*, 8:968-975, 2005.
4. S. W. Guo and E. A. Thompson. Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles. *Biometrics*, 48:361-372, 1992.

5. M. Raymond and F. Rousset. Genepop (Version-1.2) - Population-Genetics Software for Exact Tests and Ecumenicism. *Journal of Heredity*, 86:248-249, 1995.
6. M. Nei. Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. *Genetics*, 89:583-590, 1978.
7. S. C. Thomas and W. G. Hill. Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics*, 155:1961-1972, 2000.
8. B. R. Smith, C. M. Herbinger and H. R. Merry. Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics*, 158:1329-1338, 2001.
9. J. Wang. Sibship reconstruction from genetic data with typing errors. *Genetics*, 166:1963-1979, 2004.
10. D. A. Konovalov, C. Manning and M. T. Henshaw. KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Molecular Ecology Notes*, 4:779-782, 2004.
11. D. A. Konovalov. Accuracy of four heuristics for the full sibship reconstruction problem in the presence of genotype errors. *Series on Advances in Bioinformatics and Computational Biology*, 3:7-16, 2006.
12. D. A. Konovalov, N. Bajema and B. Litow. Modified SIMPSON $O(n^3)$ algorithm for the full sibship reconstruction problem. *Bioinformatics*, 21:3912-3917, 2005.
13. D. A. Konovalov, B. Litow and N. Bajema. Partition-distance via the assignment problem. *Bioinformatics*, 21:2463-2468, 2005.
14. A. Almudevar. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theoretical Population Biology*, 63:63-75, 2003.
15. J. Beyer and B. May. A graph-theoretic approach to the partition of individuals into full-sib families. *Molecular Ecology*, 12:2243-2250, 2003.
16. M. S. McPeek, X. D. Wu and C. Ober. Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics*, 60:359-367, 2004.
17. J. M. Olson. Robust Estimation of Gene-Frequency and Association Parameters. *Biometrics*, 50:665-674, 1994.
18. K. F. Goodnight and D. C. Queller. Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Molecular Ecology*, 8:1231-1234, 1999.
19. M. Nei. Analysis of Gene Diversity in Subdivided Populations. *Proceedings of the National Academy of Sciences of the United States of America*, 70:3321-3323, 1973.
20. K. Ritland. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, 67:175-185, 1996.
21. J. Wang. An estimator for pairwise relatedness using molecular markers. *Genetics*, 160:1203-1215, 2002.
22. M. Locatelli, I. M. Bomze and M. Pelillo. The combinatorics of pivoting for the maximum weight clique. *Operations Research Letters*, 32:523-529, 2004.
23. S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220:671-680, 1983.
24. D. E. Pearse and K. A. Crandall. Beyond F-ST: Analysis of population genetic data for conservation. *Conservation Genetics*, 5:585-602, 2004.