OPEN ACCESS Freely available online

PLOS ONE

Confidence Intervals for Population Allele Frequencies: The General Case of Sampling from a Finite Diploid Population of Any Size

Tak Fung[1,2]*, **Kevin Keenan**[3]

1 National University of Singapore, Department of Biological Sciences, Singapore, Singapore, **2** Queen's University Belfast, School of Biological Sciences, Belfast, Northern Ireland, United Kingdom, **3** Queen's University Belfast, Institute for Global Food Security, School of Biological Sciences, Belfast, Northern Ireland, United Kingdom

Abstract

The estimation of population allele frequencies using sample data forms a central component of studies in population genetics. These estimates can be used to test hypotheses on the evolutionary processes governing changes in genetic variation among populations. However, existing studies frequently do not account for sampling uncertainty in these estimates, thus compromising their utility. Incorporation of this uncertainty has been hindered by the lack of a method for constructing confidence intervals containing the population allele frequencies, for the general case of sampling from a finite diploid population of any size. In this study, we address this important knowledge gap by presenting a rigorous mathematical method to construct such confidence intervals. For a range of scenarios, the method is used to demonstrate that for a particular allele, in order to obtain accurate estimates within 0.05 of the population allele frequency with high probability ($\geq$95%), a sample size of $>$30 is often required. This analysis is augmented by an application of the method to empirical sample allele frequency data for two populations of the checkerspot butterfly (*Melitaea cinxia* L.), occupying meadows in Finland. For each population, the method is used to derive $\geq$98.3% confidence intervals for the population frequencies of three alleles. These intervals are then used to construct two joint $\geq$95% confidence regions, one for the set of three frequencies for each population. These regions are then used to derive a $\geq$95% confidence interval for Jost's *D*, a measure of genetic differentiation between the two populations. Overall, the results demonstrate the practical utility of the method with respect to informing sampling design and accounting for sampling uncertainty in studies of population genetics, important for scientific hypothesis-testing and also for risk-based natural resource management.

Citation: Fung T, Keenan K (2014) Confidence Intervals for Population Allele Frequencies: The General Case of Sampling from a Finite Diploid Population of Any Size. PLoS ONE 9(1): e85925. doi:10.1371/journal.pone.0085925

Editor: Guy Brock, University of Louisville, United States of America

Received June 23, 2013; **Accepted** December 4, 2013; **Published** January 21, 2014

Funding: TF is being supported by the National University of Singapore start-up grant WBS R-154-000-551-133. In addition, TF acknowledges past funding support by a Ph.D. studentship from the Beaufort Marine Research Award in Ecosystem Approach to Fisheries Management, carried out under the Sea Change Strategy and the Strategy for Science Technology and Innovation (2006–2013), with the support of the Marine Institute, funded under the Marine Research Sub-Programme of the Irish National Development Plan 2007–2013 (http://www.marine.ie/home/research/MIFunded/BeaufortAwards/). KK is being supported by a Ph.D. studentship from the Beaufort Marine Research Award in Fish Population Genetics, funded by the Irish Government under the Sea Change Strategy as above. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: tfung01@qub.ac.uk

Introduction

Spatiotemporal patterns of genetic variation among populations are often used to test hypotheses about processes underlying the patterns, such as selection, migration and genetic drift (e.g., [1,2,3,4,5]). This genetic variation captures differences in genetic structure among populations, where the genetic structure of a population is determined by the distribution of alleles among individuals in the population. The allele distribution of a population is the result of a range of biological and environmental processes acting on the population and also on surrounding populations within the same geographical region, which result in non-random mixing of gametes among individuals in all populations. Patterns in allele distributions among populations are generally assessed by consideration of variations in allele frequencies (e.g., [6,7,8,9]).

Logistic and ethical constraints mean that in practice, a population is unlikely to be sampled in its entirety, such that the population allele frequencies have to be estimated using a subset of sampled individuals. For diploid organisms in a large population, it has been established that the frequency of an allele in a sample provides an unbiased estimate of the frequency in the population as a whole [10]. Thus, if samples of a given size are repeatedly taken from a large population, then the mean frequency of an allele in a sample converges to the population allele frequency as the number of samples increases. However, many studies only take a *single* sample from a population and present or use the resulting frequencies of alleles in the sample, without accounting for sampling uncertainty (e.g., [11,12,13,14,15,16,17,18]). Therefore, these studies implicitly assume that the sample allele frequencies are close to the population allele frequencies, which is by no means guaranteed. Interpretation of findings from these studies is therefore complicated by the potential for large sampling uncertainty. Ideally, in this single sample case, uncertainty bounds for the population allele frequencies would be quantified, based on the sample allele frequencies.

A recent study by Hale et al. [19] used computer simulations to draw samples from four diploid populations, with allele frequencies based on four empirical datasets. Subsequently, they used the sample data to derive the means, variances and ranges of allele frequencies in samples of varying size, as well as the means, variances and ranges of indicators that are a function of allele frequencies (specifically, heterozygosity and F_{ST}). Using these results, Hale et al. [19] concluded that a sample size of 30 is sufficient to accurately estimate population allele frequencies when using microsatellites. However, the authors did not quantify uncertainty bounds for the population allele frequencies using their sample data, such that their assessment of accuracy lacks a rigorous quantitative basis. Furthermore, for each sample size, sample frequencies were calculated using only 100 replicate samples, which may not give a good approximation of the true dispersion in the sample frequencies. This highlights a weakness of using a computational approach that lacks an underlying mathematical theory. Moreover, Hale et al. [19] did not consider the situation identified earlier, where one sample is taken and there is a need to quantify uncertainty of population allele frequencies using just the sample allele frequencies. For this situation, earlier studies have derived confidence intervals in order to capture uncertainty in the population allele frequencies. If a diploid population is of infinite size and is at Hardy-Weinberg equilibrium (HWE), then the frequency of an allele in a sample follows a binomial distribution [10]. Thus, Gillespie [20] proposed the use of the Wald confidence interval [21]. However, this interval only performs well for sufficiently large sample sizes – with small sample sizes, it is too short [22]. For the binomial distribution, other confidence intervals have been derived that do perform well for small sample sizes [22,23], notably the Clopper-Pearson interval that was derived eight decades ago [24]. However, these only apply in the limited cases when the population is at HWE. Weir [10] proposed a confidence interval that allows for deviations from HWE, analogous to the Wald confidence interval. However, like the latter, it is expected to perform well only for sufficiently large sample sizes [10]. This is problematic because the accuracy at a particular sample size is unknown, so "sufficiently large" cannot be rigorously quantified. Moreover, all the confidence intervals considered only apply to cases when the population can be assumed to be of infinite size, i.e. when the population size is much larger than the sample size. This does not reflect the range of scenarios encountered in empirical research (e.g., [13,25,26,27]).

In this study, we build on previous work by constructing confidence intervals for population allele frequencies for the general case where (i) the population is diploid, finite and can be of any size; (ii) the sample can take any size less than or equal to that of the population; and (iii) the population can deviate from HWE to any extent. These confidence intervals are guaranteed to contain the population allele frequency with a probability above a known threshold. The method derived for constructing these intervals is then used to calculate sample sizes required to achieve accurate estimates of population allele frequencies, under a range of scenarios. Here, accuracy is measured as the length of the confidence intervals. The sample sizes derived serve as a guide for determination of suitable sample sizes in future population genetic studies. In particular, we show that a sample size of 30 does not necessarily give accurate estimates, thus refining the conclusion of Hale et al. [19]. Lastly, we provide an example of how the method can be applied to microsatellite data for two populations of the checkerspot butterfly (*Melitaea cinxia* L.) [13], to derive confidence intervals for population allele frequencies and also for Jost's *D*, a measure of genetic differentiation between two populations that is a function of their population allele frequencies [9]. The data comes from a study [13] where it is unclear that the populations can be assumed to be of infinite sizes or at HWE. This example illustrates how the mathematical theory underlying our method can be used to quantify sampling uncertainty not only in population allele frequencies, but also in parameters that are functions of these frequencies, important for hypothesis-testing and also for natural resource management.

Overall, this study provides a rigorous mathematical quantification of sampling uncertainty in population allele frequencies, without the typically unrealistic constraints of assuming that a population has an infinite size and is at HWE. Thus, the results can be applied to a wide range of studies in population genetics.

Methods

In order to construct confidence intervals for the general case of taking samples of any size from a diploid population of any size (larger than or equal to the sample size) and with any degree of deviation from HWE, the sampling distribution of the allele frequencies in this general case is first exactly specified. This distribution is then used to derive formulae specifying confidence intervals that contain the population allele frequencies with probability above a known threshold. Using these formulae, confidence intervals are derived for a range of archetypal scenarios, which are then used to calculate sample sizes that permit accurate estimates of population allele frequencies in these scenarios. In addition, the formulae are applied to a real scenario where samples are taken from two butterfly populations [13], to construct confidence intervals for the population allele frequencies of these two populations. The intervals are then used to derive a corresponding confidence interval for Jost's *D* [9].

Derivation of sampling distribution of allele frequencies

Consider a population of M diploid individuals, from which a sample of N individuals is randomly drawn ($M \geq N$). At the locus of interest, there are $n > 1$ alleles, denoted by A_i, $i \in \{1, 2, \ldots, n\}$. Let the population allele frequencies be denoted by p_i, with the corresponding sample allele frequencies being denoted by $p_{i,N}$. Also, let P_{ij} be the frequency of individuals in the population with alleles A_i and A_j ($i, j \in \{1, 2, \ldots, n\}$), such that MP_{ij} is the number of corresponding individuals in the base population. p_i is related to P_{ij} by the formula $p_i = P_{ii} + \sum_{j=1,\, j \neq i}^{n} (P_{ij}/2)$. Here, P_{ii} is a measure of the homozygosity of individuals in the population with respect to allele A_i. $P_{ii} = p_i - \sum_{j=1,\, j \neq i}^{n} (P_{ij}/2)$, and thus $P_{ii} \leq p_i$. If $0 \leq p_i \leq 0.5$, then the minimum value of P_{ii} is 0, since all copies of allele A_i can be distributed among heterozygotes of allele A_i. However, if $p_i > 0.5$, then this is not possible – there must be at least one homozygote of allele A_i. In this case, the minimum number of homozygotes is realized when all heterozygotes have a copy of allele A_i, i.e. when $P_{ii} + \sum_{j=1,\, j \neq i}^{n} P_{ij} = 1$. Rearranging this for $\sum_{j=1,\, j \neq i}^{n} P_{ij}$ and substituting into $P_{ii} = p_i - \sum_{j=1,\ j \neq i}^{n} (P_{ij}/2)$ gives the minimum value of P_{ii} as $2p_i - 1$. Thus, overall, $\mathrm{Max}\{0, 2p_i - 1\} \leq P_{ii} \leq p_i$.

The frequency of allele A_i in the sample of size N is given by $p_{i,N} = Y_{i,N}/(2N)$, where $Y_{i,N}$ is the number of copies of allele A_i in the sample. Thus, the probability distribution of $p_{i,N}$ is the same as the probability distribution of $Y_{i,N}$ except with the x-axis scaled by a factor $1/(2N)$. Denote the probability mass function (pmf) for $Y_{i,N}$ by $P(Y_{i,N} = y_{i,N})$. The range of $Y_{i,N}$ is $[0, 2N]$, so $P(Y_{i,N} = y_{i,N}) = 0$ for $y_{i,N}$ outside this range. Thus, for the following calculations, only $y_{i,N} \in [0, 2N]$ are considered. Now, $Y_{i,N} = 2X_{ii} + X_{i\cdot}$, where $X_{i\cdot} = \sum_{j=1,\, j \neq i}^{n} X_{ij}$ and X_{ij} is the number

of individuals in the sample with alleles A_i and A_j. X_{ii}, $X_{i\cdot}$ and $Z_i = N - X_{ii} - X_{i\cdot}$ thus represent the number of individuals in the sample with two copies, one copy and no copies of allele A_i, respectively. X_{ii}, $X_{i\cdot}$ and Z_i follow a multivariate hypergeometric distribution with pmf given by:

$$P(X_{ii}=x_{ii},\, X_{i\cdot}=x_{i\cdot},\, Z_i=z_i) = \frac{\binom{MP_{ii}}{x_{ii}}\binom{M\sum_{j,j\neq i}^{n}P_{ij}}{x_{i\cdot}}\binom{M-MP_{ii}-M\sum_{j,j\neq i}^{n}P_{ij}}{z_i}}{\binom{M}{N}}, \quad (1)$$

where terms on the right-hand side in brackets are binomial coefficients. Since the number of individuals in the sample must equal N, $x_{ii}+x_{i\cdot}+z_i=N$. $P(Y_{i,N}=y_{i,N})$ is the sum of $P(X_{ii}=x_{ii},X_{i\cdot}=x_{i\cdot},Z_i=z_i)$ for all those biologically feasible combinations of x_{ii}, $x_{i\cdot}$ and z_i satisfying $y_{i,N}=2x_{ii}+x_{i\cdot}$. It will now be shown that this summation can be simplified to the sum of an expression that depends only on x_{ii} and $y_{i,N}$, over all x_{ii} between lower and upper bounds that only depend on $y_{i,N}$. This considerably eases calculation of $P(Y_{i,N}=y_{i,N})$. Firstly, the number of individuals in the sample with two copies, one copy or no copies of allele A_i cannot exceed the corresponding numbers in the sampled population. This gives rise to three inequalities:

$$x_{ii} \le MP_{ii}, \quad (2)$$

$$x_{i\cdot} \le M\sum_{j,j\neq i}^{n}P_{ij}, \quad (3)$$

$$N - x_{ii} - x_{i\cdot} \le M - MP_{ii} - M\sum_{j,j\neq i}^{n}P_{ij}. \quad (4)$$

Secondly, the number of individuals with two copies, one copy or no copies of allele A_i in the sample must be non-negative. This gives rise to three more inequalities:

$$x_{ii} \ge 0, \quad (5)$$

$$x_{i\cdot} \ge 0, \quad (6)$$

$$N - x_{ii} - x_{i\cdot} \ge 0. \quad (7)$$

Since $p_i = P_{ii} + \sum_{j=1,j\neq i}^{n}(P_{ij}/2)$ and $y_{i,N}=2x_{ii}+x_{i\cdot}$, then $\sum_{j=1,j\neq i}^{n}P_{ij} = 2(p_i - P_{ii})$ and $x_{i\cdot}=y_{i,N}-2x_{ii}$. Thus, the inequalities (2)–(7) can be rearranged to obtain the double inequality:

$$\mathrm{Max}\left\{\frac{y_{i,N}}{2} - Mp_i + MP_{ii},\, y_{i,N}-N,\, 0\right\} \le x_{ii} \le \mathrm{Min}\left\{MP_{ii},\, \frac{y_{i,N}}{2},\, M-N+y_{i,N}-2Mp_i+MP_{ii}\right\}. \quad (8)$$

It is noted that $x_{i\cdot}=y_{i,N}-2x_{ii}$, which means that $x_{i\cdot}\ge 0$ (inequality (6)) ensures that $x_{ii} \le y_{i,N}/2 \le N$, as required. Denote the upper and lower bounds in (8) by $L(y_{i,N})$ and $U(y_{i,N})$ respectively. Then $P(Y_{i,N}=y_{i,N})$ can be written as:

$$\begin{aligned} P(Y_{i,N}=y_{i,N}) &= \sum_{x_{ii}=\mathrm{ceiling}[L(y_{i,N})]}^{\mathrm{floor}[U(y_{i,N})]} P(X_{ii}=x_{ii},\, X_{i\cdot}=x_{i\cdot},\, Z_i=z_i) \\ &= \sum_{x_{ii}=\mathrm{ceiling}[L(y_{i,N})]}^{\mathrm{floor}[U(y_{i,N})]} P\begin{pmatrix} X_{ii}=x_{ii},\, X_{i\cdot}=y_{i,N}-2x_{ii}, \\ Z_i=N+x_{ii}-y_{i,N} \end{pmatrix}, \end{aligned} \quad (9)$$

where z_i has been rewritten using $z_i=N-x_{ii}-x_{i\cdot}$ and $x_{i\cdot}=y_{i,N}-2x_{ii}$, $\mathrm{ceiling}[a]$ is the smallest integer larger than a, and $\mathrm{floor}[a]$ is the largest integer smaller than a. The ceiling and floor functions are introduced to ensure that x_{ii} is an integer, which it must be to have a biological interpretation. $P(Y_{i,N}=y_{i,N})$ is equal to $P(p_{i,N}=y_{i,N}/(2N))$, the pmf for $p_{i,N}$, and thus defines the probability distribution of $p_{i,N}$. Equation (9) can be used to define the probability distribution of $p_{i,N}$ given only four parameters: M, p_i, P_{ii} and N.

Derivation of confidence intervals for sample allele frequencies

For given population size (M), population allele frequency (p_i), population frequency of homozygotes with allele A_i (P_{ii}) and sample size (N), the probability distribution for the sample allele frequency ($p_{i,N}$) can be calculated exactly using equation (9). The mean value of this distribution is:

$$E[p_{i,N}] = E\left[\frac{Y_{i,N}}{2N}\right] = \frac{2E[X_{ii}]+E[X_{i\cdot}]}{2N} = \frac{2NP_{ii}+N\sum_{j=1,j\neq i}^{n}P_{ij}}{2N} = P_{ii} + \sum_{j=1,j\neq i}^{n}\frac{P_{ij}}{2} = p_i, \quad (10)$$

as in the case of sampling from an infinite diploid population [10]. In equation (10), the expectation $E[X_{ij}]=NP_{ij}$ has been used, which follows from the fact that the X_{ij}'s, with $i,j \in \{1, 2, \ldots, n\}$, follow a multivariate hypergeometric distribution with parameters M, N and MP_{ij} [28]. In addition, it can be proved that the variance of $p_{i,N}$ is

$$\sigma^2[p_{i,N}] = \frac{(M-N)(p_i - 2p_i^2 + P_{ii})}{2(M-1)N} \quad (11)$$

(see File S1). The variance thus includes a standard finite correction factor $(M-N)/(M-1)$ that tends to 1 as $M\to\infty$, as required. Therefore, as $M\to\infty$, $\sigma^2[p_{i,N}]\to(p_i-2p_i^2+P_{ii})/(2N)$, the variance in the case of an infinite population size [10]. The cumulative distribution function (cdf) for $p_{i,N}$ is specified by:

$$P(p_{i,N}\le w) = P\left(p_{i,N}\le\frac{y_{i,N}}{2N}\right) = P(Y_{i,N}\le y_{i,N}) = \sum_{k=0}^{y_{i,N}}P(Y_{i,N}=k), \quad (12)$$

where $y_{i,N}$ is an integer in the interval $[0, 2N]$. This cdf can be calculated using equation (9) and is a function of p_i. To construct a CI for p_i given M, N and P_{ii}, consider testing the null hypothesis $H_0 : p_i = p_{i,0}$ against the alternative $H_1 : p_i \neq p_{i,0}$ at significance level α for an observed value of $Y_{i,N}$, denoted by $\hat{y}_{i,N}$. The null hypothesis is not rejected if using $p_i = p_{i,0}$, $\hat{y}_{i,N}$ falls within an acceptance region defined by $B_\alpha = \left[y_{i,N,\alpha/2}, y_{i,N,1-(\alpha/2)}\right]$, where $y_{i,N,\alpha/2}$ is the largest integer for which $P\left(Y_{i,N} < y_{i,N,\alpha/2}\right) \leq \alpha/2$ and $y_{i,N,1-(\alpha/2)}$ is the smallest integer for which $P\left(Y_{i,N} > y_{i,N,1-(\alpha/2)}\right) \leq \alpha/2$. B_α can be calculated using equation (12). One method of constructing a $\geq 100(1-\alpha)\%$ CI for p_i is to determine the set of values of $p_{i,0}$ for which the null hypothesis is not rejected at significance level α, denoted by $P_{i,H_0,\alpha} = \left\{p_{i,0} : \hat{y}_{i,N} \in B_\alpha\right\}$, and defining the CI as

$$[L_\alpha, U_\alpha] = \left[\mathrm{Min}\left(P_{i,H_0,\alpha}\right), \mathrm{Max}\left(P_{i,H_0,\alpha}\right)\right] \quad (13)$$

[29]. This hypothesis-testing approach corresponds to the "test-method" described by Talens [30], who applied it to construct CI's for parameters of the univariate hypergeometric distribution. If $P_{i,H_0,\alpha} = \varnothing$, the empty set, then the acceptance region needs to be extended to include $\hat{y}_{i,N}$ for at least one value of $p_{i,0}$. Thus, in this case, α is decreased until $\hat{y}_{i,N}$ is in the acceptance region for at least one $p_{i,0}$ value.

The CI specified by equation (13) is not an exact $100(1-\alpha)\%$ CI because the distribution for $Y_{i,N}$ is discrete and because there may be some $p_{i,0}$ values between $\mathrm{Min}(P_{i,H_0,\alpha})$ and $\mathrm{Max}(P_{i,H_0,\alpha})$ for which $\hat{y}_{i,N} \notin B_\alpha$, since $P(Y_{i,N} \leq y_{i,N})$ is not guaranteed to be a monotonic function of p_i (see equations (9) and (12)). It is noted that for the probability parameter in a binomial distribution, Clopper and Pearson [24] derived $\geq 100(1-\alpha)\%$ CI's using an analogous method. Thus, for the case of a large population size relative to the sample size ($M >> N$) and HWE, where $p_{i,N}$ approximately follows a binomial distribution, CI's for p_i derived using our method would be virtually the same as Clopper-Pearson CI's. This can be verified by explicitly calculating and comparing CI's using the two methods – for example, if $M = 1{,}000{,}000 >> N = 30$ and $\hat{y}_{i,N} = 10$, then both methods give the CI $[0.083, 0.285]$ for p_i.

In deriving equation (13), it was assumed that P_{ii} is known. However, P_{ii} is generally unknown. In this case, a $\geq 100(1-\alpha)\%$ confidence region (CR) can be derived for p_i and P_{ii} in an analogous way, by considering the null hypothesis $H'_0 : p_i = p_{i,0}$, $P_{ii} = P_{ii,0}$. The CR can be derived by determining the set of vectors $(p_{i,0}, P_{ii,0})$ for which the null hypothesis is not rejected at significance level α. This set is denoted by $P_{i,H'_0,\alpha} = \left\{(p_{i,0}, P_{ii,0}) : \hat{y}_{i,N} \in B_\alpha\right\}$, where B_α is as defined before. Using the CR, $\geq 100(1-\alpha)\%$ CI's for p_i and P_{ii} can be defined as

$$\left[L'_{p_i,\alpha}, U'_{p_i,\alpha}\right] = \left[\mathrm{Min}\left(P_{i,H'_0,\alpha,1}\right), \mathrm{Max}\left(P_{i,H'_0,\alpha,1}\right)\right], \quad (14a)$$

and

$$\left[L'_{P_{ii},\alpha}, U'_{P_{ii},\alpha}\right] = \left[\mathrm{Min}\left(P_{i,H'_0,\alpha,2}\right), \mathrm{Max}\left(P_{i,H'_0,\alpha,2}\right)\right], \quad (14b)$$

where $P_{i,H'_0,\alpha,j}$ is the set of values consisting of the jth elements in the set of vectors $P_{i,H'_0,\alpha}$. In determining $P_{i,H'_0,\alpha}$, $p_{i,0}$ and $P_{ii,0}$ values over the biologically feasible ranges are tested. Since the number of copies of allele A_i in the population must be at least the number found in the sample, $\hat{y}_{i,N}/(2M) \leq p_{i,0}$. Also, the number of copies of all other alleles in the population must be at least the number found in the sample, $2N - \hat{y}_{i,N}$; this constrains the maximum value of $p_{i,0}$ according to $p_{i,0} \leq \left[2M - \left(2N - \hat{y}_{i,N}\right)\right]/(2M) = 1 - \left[\left(2N - \hat{y}_{i,N}\right)/(2M)\right]$. For a given value of $p_{i,0}$, $\mathrm{Max}\{0, 2p_{i,0} - 1\} \leq P_{ii,0} \leq p_{i,0}$, as determined earlier. If $P_{i,H'_0,\alpha} = \varnothing$, then the acceptance region is extended to include $\hat{y}_{i,N}$ for at least one pair of values of $p_{i,0}$ and $P_{ii,0}$, by decreasing α.

CI's specified by equation (13) can be calculated given M, N, P_{ii} and α, whereas those specified by equations (14a) and (14b) can be calculated given just M, N and α. In this paper, computation of any CI's using these equations, incorporating the underlying formulae specified by equations (1), (8), (9) and (12), was carried out using the software package *Mathematica* v5.0 [31]. Supporting Webpage 1 provides *Mathematica* code for computation of the CI's and can be viewed at http://rpubs.com/kkeenan02/Fung-Keenan-Mathematica/. However, other software packages such as *MATLAB* [32] and *R* [33] could also be used to implement the formulae; indeed, an *R* version of the code is provided on Supporting Webpage 2 and can be viewed at http://rpubs.com/kkeenan02/Fung-Keenan-R/. All source code for the composition of Supporting Webpages 1 and 2, including the raw *Mathematica* and *R* code used, can be accessed at https://github.com/kkeenan02/Fung-Keenan2013/.

Determining minimum sample sizes for accurate estimation of population allele frequencies

In studies of population genetics, it is desirable to obtain a sample allele frequency close to the population allele frequency with high probability [19], i.e. a short CI of high probability. Thus, under three representative scenarios, we calculate the minimum sample sizes required to obtain $\geq 95\%$ CI's with lengths ≤ 0.2 and ≤ 0.1 across all possible values of the observed sample allele frequency, $\hat{p}_{i,N} = \hat{y}_{i,N}/(2N)$. These minimum sample sizes are denoted by $N_{\leq 0.2}$ and $N_{\leq 0.1}$ respectively. They are calculated by starting with $N = 10$, computing CI lengths for all possible values of $\hat{p}_{i,N}$ and then taking the maximum value. This is repeated for increasing N in increments of 10 until the maximum CI length becomes ≤ 0.1. The N value with maximum CI length closest to 0.2 is then chosen, and then increased or decreased as necessary to find $N_{\leq 0.2}$. $N_{\leq 0.1}$ is derived analogously. Maximum CI lengths of 0.2 and 0.1 represent small maximum absolute errors in the estimated allele frequency of 0.1 and 0.05 respectively, if the estimate is taken as the mid-point of the CI. The first scenario examined, Scenario 1, is sampling from a population of $M = 1{,}000$ at HWE. $M = 1{,}000$ is larger than or on the same order of magnitude as the upper bound of the range of size estimates for 15/24 (63%) species populations collated by Frankham et al. [25], covering mammals, birds, insects and plants. Thus, $M = 1{,}000$ is taken to represent a large population. Later, M is varied over two orders of magnitude to consider populations that range from small to very large. Since there is a HWE, $P_{ii} = p_i^2$. Also, in the sampled population, the number of homozygotes with allele i, MP_{ii}, must be an integer. This restricts the number of values that p_i can take to 11 equally spaced values in the interval $[0, 1]$ (starting from 0).

Scenarios 2 and 3 represent situations where HWE does not hold, such that $P_{ii} \neq p_i^2$. In Scenario 2, P_{ii} is assumed to take its minimum value, which is $\mathrm{Max}\{0, 2p_i - 1\}$, whereas in Scenario 3, P_{ii} is assumed to take its maximum value, which is p_i. In comparison with Scenario 1, p_i can now take a larger number of values – it can take the 2,001 equally spaced values in the interval $[0, 1]$ in Scenario 2 and the 1,001 equally spaced values in the interval $[0, 1]$ in Scenario 3. Since the variance of $p_{i,N}$ is an

increasing function of P_{ii} (equation (11)), CI's derived using $p_{i,N}$ are expected to increase in length with P_{ii} as well. This means that CI lengths and minimum sample sizes ($N_{\leq 0.2}$ and $N_{\leq 0.1}$) derived for Scenarios 2 and 3 are expected to encompass the entire ranges of possible values.

In the three scenarios examined, P_{ii} is specified as a function of p_i, such that equation (13) can be used to calculate a CI for p_i given a value of $\hat{p}_{i,N}$. Calculation of this CI involves considering all possible values of p_i and determining under which values $\hat{p}_{i,N}$ falls within the corresponding acceptance region (as described above). An alternative scenario, Scenario 4, is that the relationship between P_{ii} and p_i is unknown. In this case, equation (14a) has to be used to calculate a CI for p_i given a value of $\hat{p}_{i,N}$, which involves considering all possible values of p_i and P_{ii}. This results in a considerable increase in computation time; for example, when sampling $N=30$ individuals from a population of $M=1{,}000$, 2,001 values of p_i need to be considered but 1,002,001 combinations of p_i and P_{ii} need to be considered, representing an increase in computational time by two orders of magnitude. However, for a given value of p_i, the maximum value of P_{ii} gives the highest variance for $p_{i,N}$ and is thus expected to maximize the length of the acceptance region within which $\hat{p}_{i,N}$ could fall within. Thus, CI's for p_i derived under Scenario 4 are expected to closely match those derived under the case of maximum homozygosity (Scenario 3). This can be verified by explicit calculation of the CI's – for example, given $M=100$, $N=30$ and $\hat{p}_{i,N}=15$, the CI's derived under Scenario 4 and Scenario 3 are [0.14, 0.4] and [0.15, 0.39], representing a difference in length of only 0.02. Similar results hold for $\hat{p}_{i,N}=10$ and 5. Therefore, results from Scenario 3 are used to approximate those for Scenario 4, obviating the need for long computational runtimes to explore all possible combinations of p_i and P_{ii}.

Lastly, from equation (11), the variance of $p_{i,N}$ is an increasing function of M. Thus, CI lengths for p_i are expected to increase with M, which would result in increases in $N_{\leq 0.2}$ and $N_{\leq 0.1}$. To test this explicitly, for a population at HWE and with $N=30$, maximum lengths for the $\geq 95\%$ CI for p_i (across all values of $\hat{p}_{i,N}$) are calculated for $M=100$, 250, 500, 750, 1,000, 2,500, 5,000, 7,500 and 10,000. This range corresponds to populations that are small to very large [25].

Application to an empirical data set for checkerspot butterflies

To demonstrate how the theory developed can be used in practice, it is applied to microsatellite data for samples from two populations of the checkerspot butterfly (*Melitaea cinxia* L.) occupying meadows on the Åland Islands in Finland [13]. Specifically, for the *CINX1* locus, $\geq 100(1-2\alpha)\%$ CI's are calculated for the frequencies of alleles A, B and C for the Prästö and Finström populations, using the corresponding sample allele frequencies (Table 1 of Palo et al. [13]). For each population, a CI is not calculated for the population frequency of the fourth and final allele D, because this is fixed by the population frequencies of the first three alleles. To achieve consistency with the notation used in our study, henceforth, alleles A, B, C and D are referred to as alleles A_1, A_2, A_3 and A_4 respectively. The sample size for the Prästö population is $N_1=53$, whereas that for the Finström population is $N_2=74$ [13]. p_{i,N_1} and q_{i,N_2}, $i\in\{1,2,3,4\}$, are used to denote the sample allele frequencies of A_i in the Prästö and Finström populations respectively. Similarly, p_i and q_i are used to denote the population allele frequencies of A_i in the two populations, respectively. The two populations consist of two and seven subpopulations respectively. Thus, technically, the two populations can be referred to as metapopulations, although this terminology is not used in our study for clarity. The subpopulations form part of a total of about 536 subpopulations on the Åland Islands, with an estimated total size ranging from 35,000 to at least 200,000 [13]. Thus, the size of each subpopulation is assumed to be $[(35{,}000+200{,}000)/2]/536=219$, such that the Prästö and Finström populations are assumed to have a size of $M_1=2\times 219=438$ and $M_2=7\times 219=1533$ respectively. The observed sample allele frequencies in the two populations, denoted by $\hat{p}_{i,N_1}$ and $\hat{q}_{i,N_2}$ respectively, are taken directly from [13]. These are used to calculate the observed number of copies of allele A_i in each population, denoted by $\hat{y}_{i,N_1}$ and $\hat{z}_{i,N_2}$ respectively, using the formulae $\hat{y}_{i,N_1}=2N_1\hat{p}_{i,N_1}$ and $\hat{z}_{i,N_2}=2N_2\hat{q}_{i,N_2}$. Due to rounding error in $\hat{p}_{i,N_1}$ and $\hat{q}_{i,N_2}$ values from [13], $\hat{y}_{i,N_1}$ and $\hat{z}_{i,N_2}$ had to be rounded to the nearest integer. Since the true homozygosity of each allele in each population is unknown [13], conservative CI's are calculated assuming P_{ii} takes its maximum value of p_i (this is expected to maximize the lengths of the CI's, as explained in the previous section). In addition, $\alpha=0.025/3$ is chosen, such that a $\geq 98.3\%$ CI is derived for each of the three population allele frequencies in each of the two populations. The reason for this choice of α is because for each population, using the Bonferroni Inequality [34], the cubic region defined by the three CI's can be taken as a $\geq 100(1-2\beta)\%$ confidence region (CR) for the three population allele frequencies, where $3\alpha\geq\beta$. The choice of $\alpha=0.025/3$ allows $\geq 95\%$ CR's to be derived for each set of three population allele frequencies in the two populations.

To demonstrate how the theory developed in this paper can be used to derive CI's for genetic indicators that are a function of the population allele frequencies, the $\geq 95\%$ CR's are used to calculate a $\geq 95\%$ CI for Jost's D for the *CINX1* locus and the Prästö and Finström butterfly populations. Jost's D is a measure of genetic distance between populations [9]. For the case of two populations, it is given by

$$D_{\text{Jost}}=2\left[1-\left(\frac{J_T}{J_S}\right)\right], \tag{15}$$

where, using the notation in this example,

$$\begin{aligned} J_T &= \sum_{i=1}^{4}\left(\frac{p_i+q_i}{2}\right)^2=\frac{1}{4}(p_4+q_4)^2+\frac{1}{4}\sum_{i=1}^{3}(p_i+q_i)^2 \\ &= \frac{1}{4}\left(2-\sum_{i=1}^{3}p_i-\sum_{i=1}^{3}q_i\right)^2+\frac{1}{4}\sum_{i=1}^{3}(p_i+q_i)^2 \end{aligned} \tag{16}$$

and

$$\begin{aligned} J_S &= \frac{\sum_{i=1}^{4}p_i^2+\sum_{i=1}^{4}q_i^2}{2}=\frac{1}{2}\left(p_4^2+q_4^2\right)+\frac{1}{2}\sum_{i=1}^{3}\left(p_i^2+q_i^2\right) \\ &= \frac{1}{2}\left[\left(1-\sum_{i=1}^{3}p_i\right)^2+\left(1-\sum_{i=1}^{3}q_i\right)^2\right]+\frac{1}{2}\sum_{i=1}^{3}\left(p_i^2+q_i^2\right). \end{aligned} \tag{17}$$

A lower limit to a $\geq 95\%$ CI for D_{Jost} can be derived by minimizing the function in (15) given the constraints that the six population allele frequencies are contained within the two corresponding $\geq 95\%$ CR's, and the two constraints $1-\sum_{i=1}^{3}p_i=p_4\geq 0$ and $1-\sum_{i=1}^{3}q_i=q_4\geq 0$. This lower limit is denoted by L_D. Similarly, the upper limit to a $\geq 95\%$ CI for D_{Jost} can be derived by maximizing the function in (15) under the

same constraints. This is denoted by U_D. In this study, the "Minimize" and "Maximize" functions in *Mathematica* v5.0 [31] are used to compute L_D and U_D, but corresponding functions may be used in other software packages, such as the "solnp" function in the "Rsolnp" *R* package. A $\geq 95\%$ CI for D_{Jost} can then be defined as the interval $[L_D, U_D]$. Supporting Webpage 1 provides *Mathematica* code that can be used to calculate $[L_D, U_D]$ for the butterfly case study examined (http://rpubs.com/kkeenan02/Fung-Keenan-Mathematica/). Corresponding *R* code is presented on Supporting Webpage 2 (http://rpubs.com/kkeenan02/Fung-Keenan-R/).

Results

Maximum length of $\geq 95\%$ confidence interval with increasing sample size

For Scenario 1, where the sampled diploid population of size $M = 1{,}000$ was at HWE, the maximum length of the $\geq 95\%$ CI for the population frequency of allele A_i, p_i, considering all possible observed values of the sample allele frequency, $\hat{p}_{i,N}$, decreased non-linearly with sample size N (Figure 1). The minimum N required to achieve a maximum length ≤ 0.2, $N_{\leq 0.2}$, was 22, whereas the minimum N required to achieve a length ≤ 0.1, $N_{\leq 0.1}$, was 49 (Figure 1).

For given N and $\hat{p}_{i,N}$, the CI for p_i consists of all possible values of p_i for which $\hat{p}_{i,N}$ lies in the corresponding acceptance region, as described in *Methods*. This acceptance region is expected to increase with the variance of $p_{i,N}$, $\sigma^2[p_{i,N}]$. At HWE, the frequency of homozygotes of allele A_i, P_{ii}, is equal to p_i^2; thus, according to equation (11), $\sigma^2[p_{i,N}]$ takes its highest values at intermediate values of p_i. As a result, intermediate values of $\hat{p}_{i,N}$ are likely to fall within the acceptance regions of more values of p_i, such that the corresponding CI lengths are generally longer. Simulation results for Scenario 1 were broadly in agreement with these theoretical expectations (Figure 2).

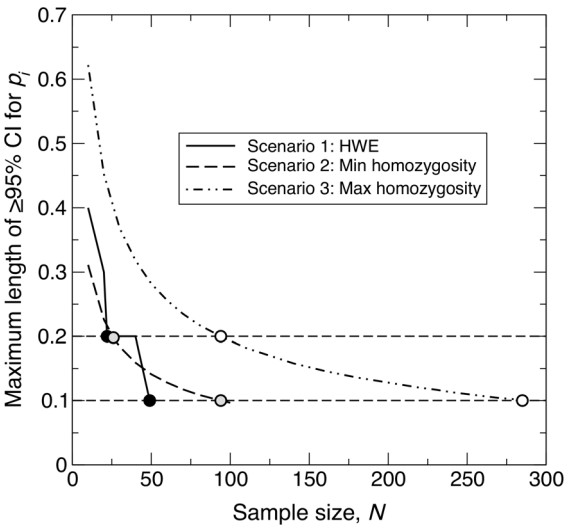

Figure 1. Change in maximum length of ≥95% confidence interval with increasing sample size. Graph showing how the maximum length of the $\geq 95\%$ confidence interval (CI) for the population frequency of an allele A_i (p_i) changes with increasing sample size *N*, when sampling from a diploid population of size $M = 1{,}000$. For a given *N*, the maximum CI length was derived by calculating CI lengths for all possible values of the observed sample allele frequency and then taking the maximum length. The three curves correspond to three scenarios where the population is (1) at Hardy-Weinberg equilibrium (HWE), (2) attains its lowest homozygosity value with respect to A_i, and (3) attains its highest homozygosity value with respect to A_i. For each curve, the two filled circles represent the minimum *N* values required for the maximum CI length to reach values of ≤ 0.2 and ≤ 0.1. For visual guidance, the two dashed horizontal lines mark maximum CI lengths of 0.2 and 0.1. The exact values of *N* tested are described in *Methods*, and equation (13) was used to calculate the CI lengths.
doi:10.1371/journal.pone.0085925.g001

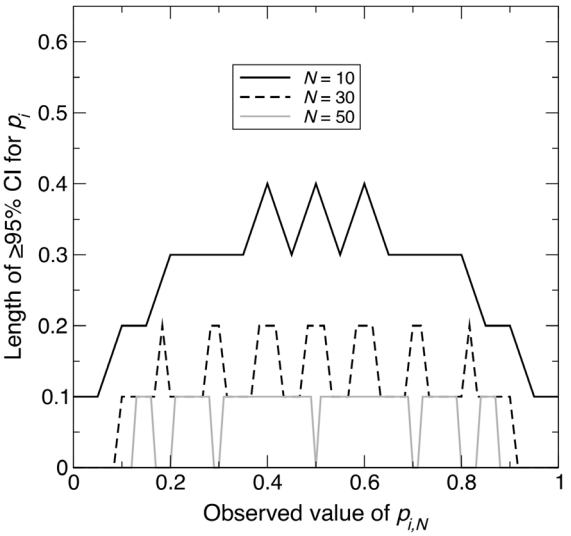

Figure 2. Change in length of $\geq 95\%$ confidence interval across different observed values, under Hardy-Weinberg equilibrium. For a sample of size *N* taken from a population of size $M = 1{,}000$ at Hardy-Weinberg equilibrium, graph showing the length of the $\geq 95\%$ confidence interval (CI) for the population frequency of an allele A_i (p_i) across all possible observed values of the sample allele frequency ($p_{i,N}$). The three curves correspond to $N = 10$, 30 and 50. Equation (13) was used to calculate the length of each CI.
doi:10.1371/journal.pone.0085925.g002

In Scenario 2, where the population was no longer at HWE but attained its lowest P_{ii}, the maximum CI length also decreased non-linearly with increasing N (Figure 1). Compared with Scenario 1, $N_{\leq 0.2}$ was slightly larger and $N_{\leq 0.1}$ was larger by about a factor of two, taking values of 26 and 94 respectively (Figure 1). This is contrary to expectations that minimum homozygosity would give smaller $N_{\leq 0.2}$ and $N_{\leq 0.1}$ (see *Methods*). The reason is that although $\sigma^2[p_{i,N}]$ is smaller for given N and p_i compared with the case of HWE, which would be expected to result in fewer p_i values for which a given $\hat{p}_{i,N}$ falls within the corresponding acceptance regions and hence a shorter CI, there are more possible values of p_i (2,001 compared with 11 – see *Methods*), which increased the number of p_i values for which $\hat{p}_{i,N}$ falls within the corresponding acceptance regions. For Scenario 2, $P_{ii} = \text{Max}\{0, 2p_i - 1\}$, such that $\sigma^2[p_{i,N}]$ attains its highest values at p_i values close to 0.25 and 0.75 (equation (11)). Thus, for given N, values of $\hat{p}_{i,N}$ near 0.25 and 0.75 are expected to generally exhibit the longest CI lengths. Simulation results for Scenario 2 were broadly in agreement with these theoretical expectations (Figure 3).

For the last scenario, Scenario 3, the population attained its highest P_{ii}, representing the opposite extreme to Scenario 2. As for the previous two scenarios, the maximum CI length decreased non-linearly with increasing N (Figure 1), but this time, $N_{\leq 0.2}$ and $N_{\leq 0.1}$ took values of 94 and 285 respectively. These values were approximately four and six times as large as the corresponding values in Scenario 1 (Figure 1). In Scenario 3, $P_{ii} = p_i$, and

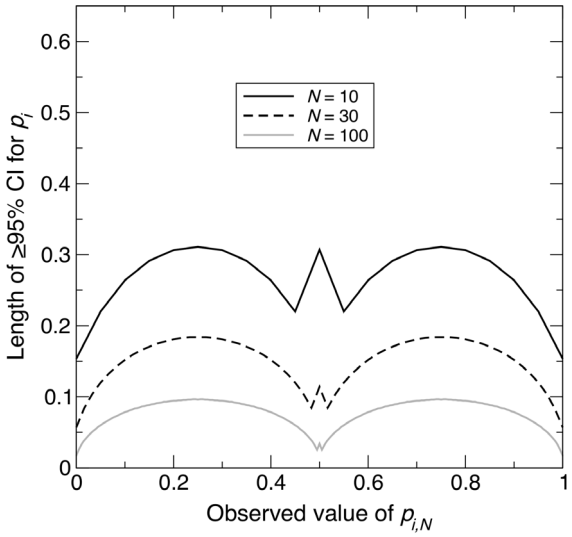

Figure 3. Change in length of ≥95% confidence interval across different observed values, under minimum homozygosity. For a sample of size *N* taken from a population of size $M = 1{,}000$ with the minimum homozygosity possible for an allele A_i, graph showing the length of the ≥95% confidence interval (CI) for the population frequency of A_i (p_i) across all possible observed values of the sample allele frequency ($p_{i,N}$). The three curves correspond to $N = 10$, 30 and 100. Equation (13) was used to calculate the length of each CI.
doi:10.1371/journal.pone.0085925.g003

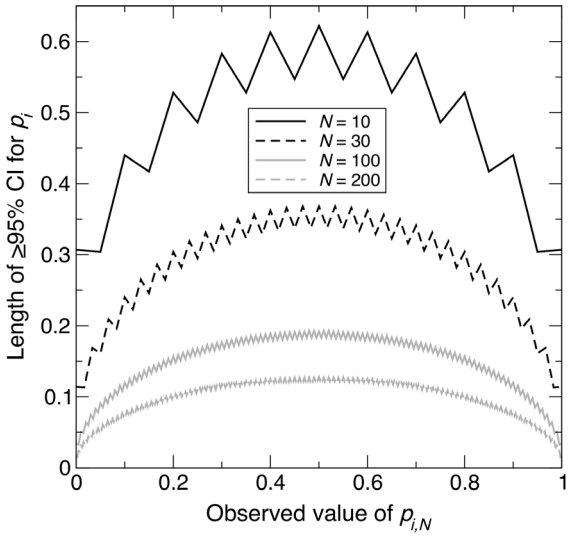

Figure 4. Change in length of ≥95% confidence interval across different observed values, under maximum homozygosity. For a sample of size *N* taken from a population of size $M = 1{,}000$ with the maximum homozygosity possible for an allele A_i, graph showing the length of the ≥95% confidence interval (CI) for the population frequency of A_i (p_i) across all possible observed values of the sample allele frequency ($p_{i,N}$). The four curves correspond to $N = 10$, 30, 100 and 200. Equation (13) was used to calculate the length of each CI.
doi:10.1371/journal.pone.0085925.g004

equation (11) shows that intermediate values of p_i give the highest values of $\sigma^2[p_{i,N}]$. Therefore, as in Scenario 1, intermediate values of $\hat{\boldsymbol{p}}_{i,N}$ are expected to generally exhibit the longest CI lengths for given *N*. Again, simulation results were consistent with these expectations (Figure 4).

Maximum length of ≥95% confidence interval with increasing population size

When taking a sample of size $N = 30$ from a diploid population at HWE, the maximum length of the ≥95% CI for p_i, across all $\hat{\boldsymbol{p}}_{i,N}$, increased with population size *M* (Figure 5). As *M* was increased from a small value of 100 to 1,000, the maximum CI length remained the same at 0.2 (this is possible because there is nothing to prevent different values of *M* giving the same maximum CI length, using equation (13)). The maximum CI length increased when *M* was increased from 1,000 to a large value of 2,500, but only by 0.04. Thereafter, the length remained constant up to a very large value of $M = 7{,}500$, and only increased by a small amount of 0.02 with a further increase in *M* to 10,000. Thus, simulation results confirm the theoretical expectation that CI length increases with *M* (as explained in *Methods* – this expectation arises because $\sigma^2[p_{i,N}]$ increases with *M*, according to equation (11)). However, the increase in the CI length was modest as *M* was increased over two orders of magnitude.

≥95% confidence interval for Jost's D between two checkerspot butterfly populations

For the Prästö checkerspot butterfly population, the ≥98.3% CI's for the population frequencies for three of the four alleles at the *CINX1* locus were computed using equation (13) and data from Palo et al. [13], as described in *Methods*. The three alleles are denoted by A_1, A_2 and A_3 respectively, with the population frequencies denoted by p_1, p_2 and p_3 respectively. The three corresponding ≥98.3% CI's derived were [0.002, 0.080], [0.598, 0.872] and [0.114, 0.381] respectively. Using the same methodology, for the Finström population, the ≥98.3% CI's for the population frequencies of A_1, A_2 and A_3 were calculated as[0.003, 0.109], [0.714, 0.914] and [0.003, 0.109] respectively.

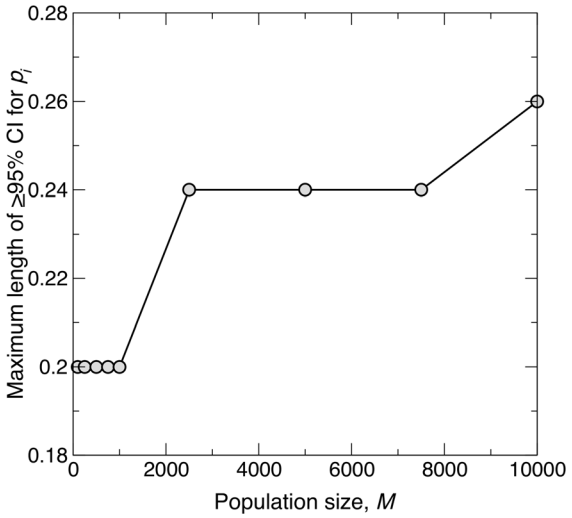

Figure 5. Change in maximum length of ≥95% confidence interval with increasing population size. Graph showing how the maximum length of the ≥95% confidence interval (CI) for the population frequency of an allele A_i (p_i) changes with increasing population size *M*, when taking samples of size $N = 30$. The population is at Hardy-Weinberg equilibrium. For a given *M*, the maximum CI length was derived by calculating CI lengths for all possible values of the observed sample allele frequency and then taking the maximum length. *M* values of 100, 250, 500, 750, 1,000, 2,500, 5,000, 7,500 and 10,000 were tested, as indicated by the filled circles.
doi:10.1371/journal.pone.0085925.g005

The three $\geq 98.3\%$ CI's for the Prästö population were used to form a cubic region, which corresponds to a $\geq 95\%$ CR for A_1, A_2 and A_3 [34]. In the same way, the three $\geq 98.3\%$ CI's for the Finström population were used to form a $\geq 95\%$ CR for A_1, A_2 and A_3. Within these two CR's, the maximum and minimum values of D_{Jost} (D_{Jost} is given by equation (15)) were calculated and used to derive a $\geq 95\%$ CI for D_{Jost}, as described in *Methods*. This CI for D_{Jost} was derived as $\left[2.34 \times 10^{-5}, 0.186\right]$. If D_{Jost} was calculated simply using the sample allele frequencies and equation (15), then only one value would have been obtained: 0.043.

Discussion

In scientific studies that use sample data to estimate unknown population parameters, the sampling uncertainty in the estimates needs to be quantified in order to make reliable inferences on population processes captured by the parameters. This forms an essential part of scientific hypothesis-testing. Therefore, in studies of population genetics, it is essential to quantify the sampling uncertainty of key population parameters used to infer past and present evolutionary processes. These include allele frequencies, which are often used to quantify genetic variation among populations, thereby allowing hypotheses on processes driving this variation to be tested (e.g., [1,2,3,4,5]). However, many studies do not include sampling uncertainty for allele frequencies, instead presenting and/or using single point estimates based on one sample per population (e.g., [11,12,13,14,15,16,17,18]). Thus, it is not possible to assess the accuracy of any inferences from these studies. This hinders not only the advance of scientific knowledge but also decision-making based on this knowledge, such as for sustainable management and conservation of natural resources. In this context, the work presented in this paper is valuable in that it provides a method of quantifying sampling uncertainty in allele frequencies for diploid populations, in the form of confidence intervals (CI's) containing true values with probability equal to or greater than a desired threshold.

The method presented pertains to the general case of a locus with n alleles, with a sample of size N taken from a population of size M and any degree of homozygosity with respect to the n alleles. In this case, the method allows construction of a CI for the population frequency of each allele, which can then be combined to create a joint confidence region (CR) for all the population allele frequencies at a given locus. It is noted that if more than one locus is considered simultaneously, then a joint CR for all population allele frequencies at all loci can be calculated by combining CI's in an analogous way. For the subcase of an infinite population size ($M \to \infty$) and a large sample size of $N \geq 30$, Weir [10] had proposed an approximate $100(1-2\alpha)\%$ CI for the population allele frequency of an allele A_i, p_i. This is $[p_{i,N} - z_{1-\alpha}\hat{\sigma}[p_{i,N}], p_{i,N} - z_{\alpha}\hat{\sigma}[p_{i,N}]]$, where $p_{i,N}$ is the sample allele frequency; z_α satisfies $\Phi(z_\alpha) = \alpha$, with Φ being the cumulative distribution function (cdf) of the standard Normal distribution; and $\hat{\sigma}[p_{i,N}]$ is an estimate of the standard deviation of $p_{i,N}$ using sample data. $\hat{\sigma}[p_{i,N}]$ is specified by equation (11) with $p_{i,N}$ replacing p_i and $P_{ii,N}$, the sample frequency of homozygotes with allele A_i, replacing P_{ii}, the corresponding population frequency of homozygotes. However, the accuracy of this CI depends both on how close $\hat{\sigma}[p_{i,N}]$ is to the true standard deviation $\sigma[p_{i,N}]$ (specified by equation (11)) and how close the cdf of $p_{i,N}$ is to a Normal distribution with the same mean and variance. This accuracy has not been quantified [10] and thus, it is not known whether the CI actually contains p_i with a probability of at least $100(1-2\alpha)$, rendering its use problematic. In this study, we have rectified this problem by constructing a CI for p_i with probability coverage of at least $100(1-2\alpha)$, for the more general case where the population size can take any value larger than or equal to the sample size.

The method we derived was used to show that the sampling uncertainty in p_i, measured as the maximum length of the $\geq 95\%$ CI for p_i across all possible values of the observed sample allele frequency $\hat{p}_{i,N}$, decreased non-linearly with N when sampling from a large population ($M = 1{,}000$) under three archetypal scenarios. These three scenarios represent the cases where the population (1) is at Hardy-Weinberg equilibrium (HWE), (2) has the lowest value of P_{ii} and (3) has the highest value of P_{ii}. As expected from theory, for any given N, the maximum CI lengths for Scenario 3 were always greater than corresponding lengths in Scenarios 1 and 2. However, the maximum CI lengths for Scenario 2 was unexpectedly greater than that in Scenario 1 for some values of N, reflecting the greater possible number of values p_i can take in a finite population with minimum homozygosity compared to one at HWE. This illustrates how the finite size of a population can give opposite trends to those obtained under an assumption of infinite size. According to theory and simulations, Scenario 3 gives CI lengths that closely approximate those in the case where P_{ii} is unknown (see *Methods*). Thus, if P_{ii} is unknown, CI lengths derived under Scenario 3 should be used. This was the approach used in the application of our method to sample data for two butterfly populations, discussed further below. On the other hand, if there is evidence that the population is at HWE, then the shorter CI's derived under Scenario 1 could be used.

Under the three scenarios examined, the non-linear decreases in sampling uncertainty with increasing N are consistent with results from the simulation study of Hale et al. [19], who found that the average difference between $p_{i,N}$ and p_i also exhibited non-linear decreases with N. However, Hale et al. [19] did not use their simulation results to quantify sampling uncertainty for the realistic situation where only one sample is taken from a population; this situation was considered in our study. Furthermore, as mentioned in the Introduction, for given N, they only used 100 samples to numerically construct the distribution for $p_{i,N}$, resulting in an incomplete distribution that may not closely reflect the true distribution. This highlights a weakness of a simulation-based approach without a rigorous mathematical underpinning, which is present in our approach. Hale et al. [19] concluded that $N = 25$–30 is sufficient to give accurate estimates of p_i, but this conclusion has to be interpreted in light of the limitations identified. Our results refine this conclusion by showing that across the three scenarios examined, $N = 49$–285 is required to ensure that, with a high probability of ≥ 0.95, an estimate for p_i can be derived from any one sample that is within 0.05 of the true value; this corresponds to a CI of length ≤ 0.1. To ensure that the estimate is within 0.1 rather than 0.05, corresponding to a CI of length ≤ 0.2, our results show that $N = 22$–94 is required. Thus, $N = 30$ is not guaranteed to give "accurate" estimates of p_i under all or most scenarios, and N values up to 10 times larger could be required. Decreasing the population size M from 1,000 would help to decrease sampling uncertainty, but results showed that decreasing M over two orders of magnitude from 10,000 to 100 only resulted in modest decreases in the maximum CI length of ≤ 0.06, with no decrease when M was decreased from 1,000 to 100. Thus, the overall conclusion is that $N = 30$ is often insufficient to guarantee accurate estimates of p_i, in the sense that p_i is within 0.05 or 0.1 of the estimate. Considering that alleles at highly polymorphic loci, such as microsatellites, often occur at population frequencies of < 0.05 [35], it might be desirable to derive CI's for p_i that are of length < 0.1. Thus, sample sizes even larger than the values found in our simulations might be required under some circumstances.

The application of our method to empirical data for two populations of the checkerspot butterfly [13] demonstrated how the underlying theory can be applied to construct joint $\geq 95\%$ CR's for the population frequencies of multiple alleles at a single locus. These CR's were then used to construct a $\geq 95\%$ CI for Jost's D, which measures genetic differentiation between the two populations. This illustrates how our method can be used to quantify sampling uncertainty in genetic indicators that are a function of population allele frequencies, thus facilitating hypothesis-testing and also risk-based natural resource management. In the example considered, the single point estimate of Jost's D using the sample allele frequencies was about four times lower than the upper bound of its $\geq 95\%$ CI. Thus, use of the single point estimate without accounting for sampling uncertainty could lead to misleading conclusions. The effectiveness of any management measures based on such conclusions would be compromised, hindering the achievement of objectives related to conservation or sustainable use. Our example therefore highlights the practical utility of the method that we have derived.

In conclusion, we have presented a rigorous mathematical method for quantifying sampling uncertainty in estimates of population allele frequencies, for a general case that has hitherto not been analyzed. In addition, we have demonstrated its practical application in informing sampling design and determining uncertainty in genetic indicators. Thus, the method derived advances both theory and practice, with broad implications for a range of disciplines, including: conservation genetics, evolutionary genetics, genetic epidemiology, genome wide association studies (GWAS), forensics and medical genetics. In particular, the method provides exact answers to the question of how many individuals need to be sampled from a population in order to achieve a given level of accuracy in estimates of population allele frequencies. This is a question that has rarely been studied before [19], despite its important practical implications. Previous studies have derived sample sizes required to sample, with high probability, at least one copy of all alleles at a locus above a given frequency [35,36,37], but these do not correspond to sample sizes required to achieve accurate estimates of population allele frequencies. Derivation of the latter requires explicit quantification of sampling uncertainty, as we have done in this study.

Possible future extensions

The CI's and CR's constructed using our method are conservative in the sense that they contain the true values with a probability equal to *or* greater than a desired threshold. This conservative property is useful in hypothesis-testing if there is a need to decrease the probability of obtaining a false positive below a certain threshold. However, researchers would ideally like to construct CI's and CR's containing true values with a known probability, not just with probability at or above a known threshold. Therefore, future research could attempt to tighten the intervals and regions that we have derived, ideally until they cover a known probability. For example, Cai and Krishnamoorthy [23] devised a method (using their "Combined Test" approach) that was shown to give shorter CI's for the probability parameters of both the binomial and univariate hypergeometric distributions, when compared with a hypothesis-testing approach analogous to the one used in this paper. If their method could be extended to the population allele frequency parameter for the more complex distribution used in this paper (equation (9)), based on a multivariate hypergeometric distribution (equation (1)), then this might result in tighter CI's for population allele frequencies when sampling from a finite diploid population.

In addition, the CI's derived in our paper were designed to quantify the uncertainty in population allele frequencies that arises from taking a random sample of a finite diploid population, which can exhibit any degree of relatedness. The "random" refers to equal probability of choosing individuals that may be of any type, which does not imply that once an individual of a particular type has been sampled, the next individual sampled is equally likely to belong to any of the types (i.e., does not imply individuals in the population or sample are unrelated). Relatedness among individuals in the population is implicitly included within the population frequencies of the different genotypes, and is thus accounted for when calculating the sampling distribution for the population frequency of an allele A_i, as specified by equation (1). However, this representation does not give an explicit quantification of the degree of relatedness among individuals in the population, for example using kinship coefficients. DeGiorgio and Rosenberg [38] used such coefficients to derive an unbiased estimator of heterozygosity ($H = 1 - \sum_{i=1}^{n} p_i^2$, for a locus with n alleles) in the case of sampling from a population of diploid individuals that could be related, and DeGiorgio et al. [39] extended these results to the case of individuals with arbitrary ploidy. In their calculations, the number of copies of allele A_i in each sampled individual k was treated as a random variable and the covariances of these variables were then related to the kinship coefficients. This is different to calculations in our paper, where the number of sampled individuals with alleles A_i and A_j was treated as a random variable (see *Methods*). The method in this paper might be revised by considering the number of copies of allele A_i for each sampled individual instead, following [38,39]. This could allow explicit quantification of relatedness in the context of deriving CI's for the population allele frequencies.

Supporting Information

File S1 Derivation of the variance of the sample allele frequency.
(PDF)

Acknowledgments

We would like to thank two anonymous reviewers for helpful and insightful comments, which have led to great improvements in this manuscript.

Author Contributions

Conceived and designed the experiments: TF KK. Performed the experiments: TF. Analyzed the data: TF KK. Contributed reagents/materials/analysis tools: TF KK. Wrote the paper: TF KK.

References

1. Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, et al. (1991) Drift, admixture, and selection in human evolution: A study with DNA polymorphisms. Proc Natl Acad Sci USA 88: 839–843.
2. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12: 1805–1814.
3. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. Nature 4: 981–994.
4. de Kovel CGF (2006) The power of allele frequency comparisons to detect the footprint of selection in natural and experimental situations. Genet Sel Evol 38: 3–23.

5. Friedlander SM, Herrmann AL, Lowry DP, Mepham ER, Lek M, et al. (2013) *ACTN3* allele frequency in humans covaries with global latitudinal gradient. PLoS ONE 8(1):e52282.
6. Nei M (1972) Genetic distance between populations. Am Nat 106: 283–292.
7. Nei M, Chesser RK (1983) Estimation of fixation indices and gene diversities. Ann Hum Genet, 47: 253–259.
8. Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. Evolution 38: 1358–1370.
9. Jost L (2008) G_{ST} and its relatives do not measure differentiation. Mol Ecol 17: 4015–4026
10. Weir BS (1996) Genetic data analysis II. Sunderland, USA, Sinauer Associates. 445 p.
11. Eanes WF, Koehn RK (1978) An analysis of genetic structure in the Monarch butterfly, *Danaus plexippus* L. Evolution 32: 784–797.
12. Forbes SH, Hogg JT, Buchanan FC, Crawford AM, Allendorf FW (1995) Microsatellite evolution in congeneric mammals: domestic and bighorn sheep. Mol Biol and Evol 12: 1106–1113.
13. Palo J, Varvio S-L, Hanski I, Väinölä R (1995) Developing microsatellite markers for insect population structure: complex variation in a checkerspot butterfly. Hereditas 123: 295–300.
14. Luikart G, Allendorf FW, Cournuet J-M, Sherwin WB (1998) Distortion of allele frequency distributions provides a test for recent population bottlenecks. J Hered 89:238–247.
15. Rank NE, Dahlhoff EP (2002) Allele frequency shifts in response to climate change and physiological consequences of allozyme variation in a montane insect. Evolution 56: 2278–2289.
16. Hugnet C, Bentjen SA, Mealey KL (2004) Frequency of the mutant MDR1 allele associated with multidrug sensitivity in a sample of collies from France. J Vet Pharmacol Ther 27: 227–229.
17. Seider T, Fimmers R, Betz P, Lederer T (2010) Allele frequencies of the five miniSTR loci D1S1656, D2S441, D10S1248, D12S391 and D22S1045 in a German population sample. Forensic Sci Int Genet 4: e159–e160.
18. Petrejčíková E, Soták M, Bernasovská J, Bernasovský I, Rębała K, et al. (2011) Allele frequencies and population data for 11 Y-chromosome STRs in samples from Eastern Slovakia. Forensic Sci Int Genet 5: e53–e62.
19. Hale ML, Burg TM, Steeves TE (2012) Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. PLoS ONE 7(9): e45170.
20. Gillespie JH (2004) Population genetics: a concise guide, second edition. Baltimore, USA, The John Hopkins University Press. 217 p.
21. Laplace PS (1812) Théorie analytique des probabilités. Paris, France, Courcier. 464 p.
22. Agresti A, Coull BA (1998) Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat 52: 119–126.
23. Cai Y, Krishnamoorthy K (2005) A simple improved inferential method for some discrete distributions. Comput Stat Data Anal 48: 605–621.
24. Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26: 404–413.
25. Frankham R (1996) Relationship of genetic variation to population size in wildlife. Conserv Biol 10: 1500–1508.
26. Green DM (2003) The ecology of extinction: population fluctuation and decline in amphibians. Biol Conserv 111: 331–343.
27. Vredenberg VT, Knapp RA, Tunstall TS, Briggs CJ (2010) Dynamics of an emerging disease drive large-scale amphibian population extinctions. Proc Natl Acad Sci USA 107: 9689–9694.
28. Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions. Hoboken, USA, Wiley-Blackwell. 328 p.
29. Cox DR, Hinkley DV (1974) Theoretical statistics. London, UK, Chapman and Hall. 511 p.
30. Talens E (2005) Statistical auditing and the AOQL-Method. Ridderkerk, The Netherlands, Labyrint Publications. 164 p.
31. Wolfram Research Inc. (2003) Mathematica Edition: Version 5.0. Champaign, Illinois, , USA, Wolfram Research Inc.
32. The MathWorks Inc. (2012) MATLAB version 8.0. Natick, Massachusetts, , USA, The MathWorks Inc.
33. R Development Core Team (2010) R: a language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing. 1731 p.
34. Rice JA (1995) Mathematical statistics and data analysis, second edition. Belmont, USA, Duxbury Press. 651 p.
35. Chakraborty R (1992) Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. Hum Biol 64: 141–159.
36. Gregorius H-R (1980) The probability of losing an allele when diploid genotypes are sampled. Biometrics 36: 643–652.
37. Sjögren P, Wyöni P-I (1994) Conservation genetics and detection of rare alleles in finite populations. Conserv Biol 8: 267–270.
38. DeGiorgio M, Rosenberg NA (2009) An unbiased estimator of gene diversity in samples containing related individuals. Mol Biol Evol 26: 501–512.
39. DeGiorgio M, Jankovic I, Rosenberg NA (2010) Unbiased estimation of gene diversity in samples containing related individuals: exact variance and arbitrary ploidy. Genetics 186: 1367–1387.