Ne Calculations

1. Linkage Disequilibrium Method

(1) Robin S. Waples, "*A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci*," Conserv Genet 7: 167–184 (2006).

(2) AT Jones, JR Ovenden and Y-G Wang, "*Improved confidence intervals for the linkage disequilibrium method for estimating effective population size*," Heredity (2016), 1–7.

(a) Review of LDNe (The calculations below are for random mating model.)

Let k be the number of polymorphic loci. For a pair of loci (i, j), $i < j$, let S_{ij} be the sample size at two loci (the number of individuals having data at both loci), then the expected $\hat{r}^2$-sample is calculated by

$$E(\hat{r}^2_{ij}) \;=\; \begin{cases} 1/S_{ij} \;+\; 3.19/S^2_{ij} & \text{if } S_{ij} \geq 30, \\ 0.0018 \;+\; 0.907/S_{ij} \;+\; 4.44/S^2_{ij} & \text{otherwise.} \end{cases} \tag{1.1}$$

Let

$$n_{ij} = (n_i - 1)(n_j - 1) \qquad (n_i,\, n_j \text{ are the numbers of alleles at loci } i, j, \text{ respectively}).$$

The n_{ij} are used as a weight for calculating the weighted harmonic sample size S of all S_{ij}:

$$\frac{N}{S} \;=\; \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{n_{ij}}{S_{ij}}, \qquad \text{where} \quad N = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_{ij}. \tag{1.2}$$

The expected $\hat{R}^2$-sample in the output for LDNe is calculated from S using formula (1.1).

$$E(\hat{R}^2) \;=\; \begin{cases} 1/S \;+\; 3.19/S^2 & \text{if } S \geq 30, \\ 0.0018 \;+\; 0.907/S \;+\; 4.44/S^2 & \text{otherwise.} \end{cases} \tag{1.3}$$

For the Burrows correlation $\hat{r}^2$ at pair loci (i, j), the weight is taken to be

$$w_{ij} = n_{ij} S^2_{ij}. \tag{1.4}$$

With these weights for r^2_{ij} calculated at pairs of loci (i,j), the overall $\hat{R}^2$ in the output is the weighted average:

$$\hat{R}^2 \;=\; \frac{1}{W}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} w_{ij}\, r^2_{ij}, \qquad \text{where} \quad W = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} w_{ij}. \tag{1.5}$$

With $E(\hat{R}^2)$ given in (1.3) and $\hat{R}^2$ given in (1.5), let

$$\hat{R}^{2\prime} \;=\; \hat{R}^2 \;-\; E(\hat{R}^2) \;=\; \begin{cases} \hat{R}^2 \;-\; 1/S \;-\; 3.19/S^2 & \text{if } S \geq 30, \\ \hat{R}^2 \;-\; 0.0018 \;-\; 0.907/S \;-\; 4.44/S^2 & \text{otherwise.} \end{cases} \tag{1.6}$$

This $\hat{R}^{2\prime}$ (called $\hat{R}^2$-drift) is used to produce the estimate $\hat{N}_e$ (assuming random mating model):

$$\hat{N}_e \;=\; \frac{1}{2\hat{R}^{2\prime}} \cdot \begin{cases} 1/3 \;+\; \sqrt{1/9 \;-\; (2.76)\hat{R}^{2\prime}} & \text{if } S \geq 30, \\ 0.308 \;+\; \sqrt{0.094864 \;-\; (2.08)\hat{R}^{2\prime}} & \text{otherwise,} \end{cases} \tag{1.7}$$

where the square roots are assigned 0 if the radicants are negative. This is the N_e output from LDNe.

In case of no missing data, all S_{ij} are the same, and equal to S, hence

$$E(\hat{r}^2_{ij}) \;=\; E(\hat{R}^2) \qquad \text{for all } i,j \qquad \text{(no missing data)} \tag{1.8}$$

(b) Confidence Intervals

For parametric CI, N in (1.2) is referred to as the degree of freedom, used in Chi-square distribution to get the lower and upper bounds for $\hat{R}^2$. These bounds, subtracted by $E(\hat{R}^2)$, give the lower and upper bounds for $\hat{R}^{2\prime}$, which are then used to produce CI for $\hat{N}_e$.

For CI obtained by jackknife on samples, we evaluate $\hat{R}^2$ of the population for each individual being removed. Those values of $\hat{R}^2$ will be used to obtain variance of $\hat{R}^2$, incorporated an empirical correction factor 0.84. The CI is obtained for $\hat{R}^{2\prime}$, and then for $\hat{N}_e$.

Since the calculations of $\hat{R}^2$ is time consuming (most of run-time for the LD method lies on this calculation), the run time will be enormous if the calculation of $\hat{R}^2$ is carried out separately for each subpopulation. The implementation of this jackknife method, which avoids calculating those $\hat{R}^2$ separately, is described in another document.

(c) Revised Method for Missing Data

The weight as given in (1.4) will be assigned to $\hat{r}^2$-drift for each pair of loci, that is, $w_{ij} = n_{ij}S^2_{ij}$ is the weight for

$$\hat{r}^{2\prime}_{ij} = \hat{r}^2_{ij} - E(\hat{r}^2_{ij}), \tag{1.9}$$

where $E(\hat{r}_{ij}^2)$ is given in (1.1). Then the weighted average, denoted by $\hat{R}_0^2{}'$, is

$$\begin{aligned}
\hat{R}_0^2{}' &= \frac{1}{W}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} w_{ij}\,\hat{r}_{ij}^2{}', \qquad\qquad \text{where } W \text{ is given in (1.5)} \\
&= \frac{1}{W}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} w_{ij}\,\hat{r}_{ij}^2 \;-\; \frac{1}{W}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} w_{ij}\,E(\hat{r}_{ij}^2) \\
&= \hat{R}^2 \;-\; \frac{1}{W}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} w_{ij}\,E(\hat{r}_{ij}^2). \qquad (1.10)
\end{aligned}$$

The first term in (1.10) is (1.5). The second term is the weighted average of the expected $\hat{r}_{ij}^2$-sample. If there are no missing data, this term is $E(\hat{R}^2)$ as pointed out by (1.8), so $\hat{R}_0^2{}'$ is the same as $\hat{R}^2{}'$ given in (1.6).

Let $\hat{N}_e^0$ be the initial estimate of $\hat{N}_e$ calculated from $\hat{R}_0^2{}'$ as in (1.7). Note that when there are missing data, this value will be slightly different from $\hat{N}_e$ produced by LDNe because of the second term in (1.10), which may differ from $E(\hat{R}^2)$.

With the initial estimate $\hat{N}_e^0$, unless this value is negative or too large, we reassign the weight of pair (i, j) by

$$w_{ij}' \;=\; \frac{n_{ij}S_{ij}^2}{\left(S_{ij} + 3\hat{N}_e^0\right)^2}. \qquad (1.11)$$

Now, the pairs $(\hat{r}_{ij}^2,\ w_{ij}')$ produces the weighted average:

$$\hat{R}^2{}' \;=\; \frac{1}{W}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} w_{ij}'\,\hat{r}_{ij}^2{}', \qquad (1.12)$$

which will be used for the final estimate of $\hat{N}_e$ using equation (1.7).

If there is a recalculation of weights because of missing data, the weights apply to $\hat{r}_{ij}^2{}'$, therefore to both $\hat{r}_{ij}^2$ and $E(\hat{r}_{ij}^2)$. For parameter confidence interval, we find the confidence intervals for $\hat{R}^2$, subtract by the weighted average of $E(\hat{r}_{ij}^2)$ to obtain CI for $\hat{R}^2{}'$, then the CI for $\hat{N}_e$.

2. Heterozygote Excess Method

(1) A. I. Pudovkin, D. V. Zaykin and D. Hedgecock, "*On the Potential for Estimating the Effective Number of Breeders from Heterozygote-Excess in Progeny*," Genetics 144: 383–387 (1996).

(2) O. L. Zhadanova and A. I. Pudovkin, D. V. Zaykin, "*Nb_HetEx: A Program to Estimate the Effective Number of Breeders*," Journal of Heredity 99 (6): 694–695 (2008).

(3) A. I. Pudovkin, O. L. Zhadanova and D. Hedgecock, "*Sampling Properties of the Heterozygote-Excess Estimator of the Effective Number of Breeders*," Conserv Genet 11: 759–771 (2010).

(a) One locus

At each polymorphic locus j, for each allele i whose frequency is p_i, we calculate the *observed frequency* $H_j^{obs}(i)$ and the *expected frequency* $H_j^{exp}(i)$ of heterozygotes having allele i. The observed frequency $H_j^{obs}(i)$ is obtained by counting heterozygotes in the population based on N_j, the number of samples having data at this particular locus j. The expected frequency $H_j^{exp}(i)$ is given in Pudovkin's paper:

$$H_j^{exp}(i) = 2p_i(1-p_i)\left(1+\frac{1}{2N_j}\right).$$

For bias correction, the program follows the formula given in Zhdanova & Pudovkin's:

$$H_j^{exp}(i) = 2p_i(1-p_i)\left(1+\frac{1}{2N_j-1}\right). \tag{2.13}$$

The D index for excess or deficiency of heterozygote excess is given by

$$D_j(i) = \frac{H_j^{obs}(i) - H_j^{exp}(i)}{H_j^{exp}(i)}.$$

Let n_j be the number of alleles at locus j. Then, the average D_j taken over all allele a at locus j is

$$D_j = \frac{1}{n_j}\sum_{i=1}^{n_j} D_j(i).$$

(b) Multiple loci

Suppose there are k loci. The D index is taken as the weighted mean of all $D_j(i)$. The weight for each allele at locus j, as given in Zhdanova & Pudovkin's, is $w_{ij} = \sqrt{N_j}\ \frac{n_j-1}{n_j}$. Since all alleles of the same locus have the same weight, we can represent D in terms of the weighted mean of D_j, $j = 1, \ldots, k$, where the weight W_j at locus j is the total weight of alleles at the locus:

$$D = \frac{1}{W}\sum W_j\, D_\ell, \qquad W = \sum_{j=1}^{k} W_j, \quad W_j = \,(n_\ell - 1)\sqrt{N_j}. \tag{2.14}$$

Then the effective number of breeders is

$$N_b = \frac{1}{2D} + \frac{1}{2(D+1)} = \frac{2D+1}{2D(D+1)}. \tag{2.15}$$

The following are printed in the "Frequency Data" additional output file:

- The weighted and unweighted means, taken over all loci, of effective number of breeders $N_b(\ell)$.
- The effective number of breeders based on the D index as the unweighted mean of D_ℓ, and as the mean of all $D_\ell(i)$ in all loci.

(c) Confidence Intervals

In calculating confidence intervals, we follow Pudovkin et al. paper (Conserv. Genet. (2010) 11: 759 – 771), using formula (3) in the paper for standard error, which can be written as

$$\mathrm{SE} \;=\; \sqrt{\frac{(D_{[2]} - D^2)}{\mathfrak{I}} \cdot \frac{W^2}{W^2 - W_{[2]}}}, \tag{2.16}$$

where D, W are given in (2.14), $W_{[2]}$ is the sum of all w_{ij}^2:

$$W_{[2]} \;=\; \sum_{j=1}^{k}\sum_{i=1}^{n_j} \frac{(n_j-1)^2 N_j}{n_j^2} \;=\; \sum_{j=1}^{k} \frac{(n_j-1)^2 N_j}{n_j} \;=\; \sum_{j=1}^{k} \frac{W_j^2}{n_j},$$

and

$$D_{[2]} \;=\; \frac{1}{W}\sum_{j=1}^{k} W_j D_{j,2}, \qquad D_{j,2} \;=\; \frac{1}{n_j}\sum_{i=1}^{n_j} d_{ij}^2,$$

($D_{j,2}$ is the average of d_{ij}^2, $D_{[2]}$ is the weighted average of $D_{j,2}$),

$$\mathfrak{I} \;=\; \sum_{j=1}^{k}(n_j - 1) \qquad (\mathfrak{I} \text{ is the total number of independent alleles across loci}).$$

Then the 95% confidence intervals are calculated using the t-distribution.

(d) Restrictions on Frequencies

Let q be a number between 0 and 0.5. We want to only consider alleles whose frequencies are at least (including) q. The number of alleles K_ℓ now will exclude those whose frequencies are less than q. Thus, if at a locus ℓ, there is only one allele having frequency at least q, then its weight is zero, the locus will be dropped from consideration. An allele i at locus ℓ whose frequency p_i less than q will be dropped, so $D^\ell(i)$ will not be calculated as a part of D_ℓ. However, a sample is not dropped even if it contains only dropped alleles across all loci. (If samples having only dropped alleles are to be dropped, then the frequencies of all alleles must be recalculated, and for those alleles that are not present in the newly dropped samples, their frequencies will increase and then some dropped alleles may now have frequencies at least q!) All the dropped alleles are lumped together as one allele. Thus, if a genotype consists of two dropped alleles, then it is now considered as a homozygote.

3. Molecular Coancestry Method

Tetsuro Nomura, "*Estimation of Effective Number of Breeders from Molecular Coancestry of Single Cohort Sample*," Evolutionary Applications (2008) pp 462–474.

(a) Formulas

As in Nomura's paper, let n be the number of samples, L be the number of loci, n_P be the total number of distinct sample pairs (x, y), $n_P = \frac{1}{2}n(n-1)$, $f_{M,xy,\ell}$ be the molecular coancestry between samples x and y at locus ℓ as given in formula (4) of the paper, w_ℓ be the "weight" at locus ℓ, W be their sum, $\hat{s}_\ell$ be the average molecular coancestry over putative nonsib pairs at locus ℓ. The coefficient $\hat{f}_{1,xy}$ as given in formula after (6) in the paper:

$$\hat{f}_{1,xy} = \frac{1}{W}\sum_{\ell=1}^{L} w_\ell \, \frac{f_{M,xy,\ell} - \hat{s}_\ell}{1 - \hat{s}_\ell}. \tag{3.1}$$

The weight w_ℓ is given by

$$w_\ell = \frac{(1-\hat{s}_\ell)^2}{\left(\sum_{i=1}^{m} p_i^2\right)\left(1 - \sum_{i=1}^{m} p_i^2\right)}, \tag{3.2}$$

where p_i is the frequency of allele i, $i = 1, \ldots, m$ are all alleles at locus ℓ.

Formula for $\hat{f}_1$ preceding formula (7) on Nomura's, p. 464 (which gives the effective number of breeders as $\frac{1}{2\hat{f}_{1,xy}}$) can be written as

$$\begin{aligned}
\hat{f}_1 &= \frac{1}{n_P}\sum_{x<y}\hat{f}_{1,xy} = \frac{1}{n_P}\sum_{x<y}\frac{1}{W}\sum_{\ell=1}^{L} w_\ell \, \frac{f_{M,xy,\ell} - \hat{s}_\ell}{1-\hat{s}_\ell} = \frac{1}{n_P W}\sum_{\ell=1}^{L}\sum_{x<y} w_\ell \, \frac{f_{M,xy,\ell} - \hat{s}_\ell}{1-\hat{s}_\ell} \\
&= \frac{1}{n_P W}\sum_{\ell=1}^{L}\frac{w_\ell}{1-\hat{s}_\ell}\sum_{x<y}(f_{M,xy,\ell} - \hat{s}_\ell) = \frac{1}{n_P W}\sum_{\ell=1}^{L}\frac{w_\ell}{1-\hat{s}_\ell}\left[\sum_{x<y} f_{M,xy,\ell} \; - \; \sum_{x<y}\hat{s}_\ell\right] \\
&= \frac{1}{n_P W}\sum_{\ell=1}^{L}\left[\frac{w_\ell}{1-\hat{s}_\ell}\sum_{x<y} f_{M,xy,\ell}\right] - \frac{1}{n_P W}\sum_{\ell=1}^{L}\left[\frac{w_\ell}{1-\hat{s}_\ell}(n_P\hat{s}_\ell)\right] \qquad \left(\textstyle\sum_{x<y}\hat{s}_\ell = n_P\hat{s}_\ell\right) \\
&= \frac{1}{W}\sum_{\ell=1}^{L} w_\ell\left[\frac{1}{1-\hat{s}_\ell}\left(\frac{1}{n_P}\sum_{x<y} f_{M,xy,\ell} - \hat{s}_\ell\right)\right]
\end{aligned} \tag{3.3}$$

The last expression can be interpreted as the weighted average of

$$\hat{f}_{\ell,1} = \frac{1}{1-\hat{s}_\ell}\left(\frac{1}{n_P}\sum_{x<y} f_{M,xy,\ell} \; - \; \hat{s}_\ell\right) \tag{3.4}$$

across loci, where the first term in the parentheses is the average of molecular indices at locus ℓ of all sample pairs. Both terms in the parentheses are printed in the auxiliary file. The term in (3.4) will be viewed as the "$\hat{f}_1$" value at locus ℓ with weight w_ℓ for finding the adjusted variance of the overall $\hat{f}_1$ by Jackknife method.

At each locus ℓ, the estimate of $\hat{s}_\ell$ is to be found by determining distinct putative nonsib pairs; the algorithm is described in the last paragraph on p. 464 in Nomura's paper. Basically, the process is to choose among eligible pairs, the one that yields the smallest average value of coancestry indices taken across loci that differ from ℓ. A maximum 20 putative nonsib pairs are listed in the auxiliary file.

(b) Missing data

When determining putative pairs (x, y) at a locus, both x and y must have data at that locus. In the process of determining if a pair (x, y) can be taken as a putative nonsib at locus ℓ, we only take average of the coancestry indices of (x, y) at other loci $\ell' \neq \ell$ where the pair have full data.

Also, for the average $\frac{1}{n_P} \sum_{x<y} f_{M,xy,\ell}$, the summation includes only $f_{M,xy,\ell}$ where the pair (x, y) have full data. Thus, the denominator n_P is replaced by the number of (x, y), $x < y$, that have full data.

(c) Restrictions on Frequencies

In this method, all alleles should be accepted; there is no restriction based on their frequencies.

(d) Confidence Intervals

We determine the confidence interval for $\hat{f}_1$, the weighted average of $\hat{f}_{\ell,1}$ across loci as given in (3.3) and (3.4), where the weight w_ℓ is given in (3.2), by Jackknife method on loci as mentioned after (3.4). The confidence interval for $\hat{f}_1$ is then translated to the confidence interval for N_e.

4. Temporal Method

(1) Edward Pollak, " *A New Method for Estimating the Effective Population Size From Allele Frequency Changes*," Genetics 1041: 531–548 (1983).

(2) Robin S. Waples, " *A Generalized Approach for Estimating Effective Population Size From Temporal Changes in Allele Frequency*," Genetics 121: 379–391 (1989).

(3) Masatoshi Nei and Fumio Tajiama, "*Genetic Drift and Estimation of Effective Population Size*," Genetics 98: 625–640 (1981).

(4) Per Eric Jorde and Nils Ryman, "*Unbiased Estimator for Genetic Drift and Effective Population Size*," Genetics 177: 927–935 (2007).

Unlike other methods, this method will require at least two population samples. The word "sample" used in other methods to describe one member of the population. In this method, it will be called an "individual."

In other methods, only polymorphic loci are calculated. In this case, there may be a locus which is monomorphic in one generation but polymorphic in the other; such locus will be in the calculation.

In fact, only locus that is monomorphic in both samples for the *same allele* will be dropped. This is the minimum requirement so that the terms used on the calculation of some measure F are all well-defined.

The following notations will be used:

- Two samples: sample 1 and sample 2 are taken at generations t_1 and t_2.
- L is the number of accepted loci. (The accepted loci, besides the minimum requirement, may also depend on frequency criteria.)
- ℓ ($\ell = 1, \dots, L$) is an accepted locus.
- m_ℓ, n_ℓ are the numbers of individuals having data at locus ℓ in sample 1 and sample 2, respectively.
- K_ℓ is the number of alleles at locus ℓ (all alleles that appear in either sample are counted).
- $i = 1, \dots, K_\ell$ is an allele at locus ℓ. (The same index i may represent different alleles if cited at different loci.)
- x_i, y_i are the frequencies of allele i at two samples: x_i at generation t_1 and y_i at generation t_2.
- z_i, $\overline{z}_i$ are the unweighted and weighted means of x_i, y_i respectively:
$$z_i = \tfrac{1}{2}(x_i + y_i), \qquad \overline{z}_i = \frac{m_\ell x_i + n_\ell y_i}{m_\ell + n_\ell}.$$
($2m_\ell\, x_i$ is the total alleles i in sample 1, and $2n_\ell\, y_i$ is the total alleles i in sample 2.)

(a) Nei & Tajima

The change of frequencies of allele i in two samples is taken as

$$\frac{(x_i - y_i)^2}{z_i - x_i y_i}. \tag{4.1}$$

The measure F (at locus ℓ) is stated in Waples, formula (8):

$$F_c^\ell \;=\; \frac{1}{K_\ell} \sum_{i=1}^{K_\ell} \frac{(x_i - y_i)^2}{z_i - x_i y_i}. \tag{4.2}$$

The overall F_c is calculated as the weighted average of all F_c^ℓ, where the weight of each locus is K_ℓ, the number of alleles in that locus. Then

$$F_c \;=\; \frac{1}{\sum_{\ell=1}^{L} K_\ell} \sum_{\ell=1}^{L} \sum_{i=1}^{K_\ell} \frac{(x_i - y_i)^2}{z_i - x_i y_i}. \tag{4.3}$$

This is the average of all terms in (4.1) taken for all alleles in all loci.

(b) Pollak

The change of frequencies of allele i in two samples is taken as

$$\frac{(x_i - y_i)^2}{z_i}. \tag{4.4}$$

The measure F is given as formula (9) in Waples:

$$F_k^\ell = \frac{1}{K_\ell - 1} \sum_{i=1}^{K_\ell} \frac{(x_i - y_i)^2}{z_i}. \tag{4.5}$$

$K_\ell - 1$ is the total number of independent alleles at locus ℓ. The overall F_k is calculated as the weighted average of all F_k^ℓ, where the weight of each locus is $K_\ell - 1$. Then

$$F_k = \frac{1}{\sum_{\ell=1}^L K_\ell \; - \; L} \sum_{\ell=1}^{L} \sum_{i=1}^{K_\ell} \frac{(x_i - y_i)^2}{z_i}. \tag{4.6}$$

The maximum value for both terms, $\frac{(x_i-y_i)^2}{z_i}$ and $\frac{(x_i-y_i)^2}{z_i - x_i y_i}$ is 2. We first look at the latter. From $x_i^2 \le x_i,\ y_i^2 \le y_i$ (since $x_i, y_i \le 1$),

$$x_i + y_i - 2x_i y_i \ge x_i^2 + y_i^2 - 2x_i y_i = (x_i - y_i)^2 \ge 0 \quad \Rightarrow \quad \tfrac{1}{2}\,(x_i + y_i) - x_i y_i \ge \; \tfrac{1}{2}(x_i - y_i)^2 \ge 0,$$

we see that the denominator $z_i - x_i y_i$ of $\frac{(x_i-y_i)^2}{z_i - x_i y_i}$ is ≥ 0. It can be zero only if $x_i = y_i = 0$ or $x_i = y_i = 1$, each of which is excluded since allele i must be present in at least one sample, and it is not the only allele at both. Thus, the term is always well-defined. The first of the above inequalities shows that $\frac{(x_i-y_i)^2}{z_i - x_i y_i} \le 2$. The equality happens, i.e. $x_i + y_i - 2x_i y_i = x_i^2 + y_i^2 - 2x_i y_i$ (equivalently, $z_i - x_i y_i = \frac{1}{2}(x_i - y_i)^2$), only if $x_i + y_i = x_i^2 + y_i^2$, which implies $x_i = x_i^2,\ y_i = y_i^2$, and then $x_i, y_i = 0$ or 1. Since x_i and y_i cannot be both 0 or both 1, this implies that one of them should be 0 and the other is 1. Therefore, from

$$0 \le \frac{(x_i - y_i)^2}{z_i} \le \frac{(x_i - y_i)^2}{z_i - x_i y_i} \le 2,$$

both terms attain maximum value 2 when $x_i = 1$ and $y_i = 0$, or when $y_i = 1$ and $x_i = 0$ (locus ℓ is monomorphic with allele i in one sample and contains no allele i in the other). These are the only cases that either term can take value 2. From the above inequalities, the numerator in the definition of F_k is less than that of F_c. However, the denominator in F_k (which is $\sum_{\ell=1}^L K_\ell \; - \; L$) is also less than that of F_c (which is $\sum_{\ell=1}^L K_\ell$); so there is no direct comparison of F_k and F_c.

(c) Jorde & Ryman

The change of frequencies of allele i in two samples in this method is taken as

$$\frac{(x_i - y_i)^2}{z_i(1 - z_i)}. \tag{4.7}$$

The measure F is given as in formula (9) in Jorde & Ryman's paper:

$$F_s^\ell \;=\; \frac{\sum_{i=1}^{K_\ell}(x_i - y_i)^2}{\sum_{i=1}^{K_\ell} z_i(1-z_i)}. \tag{4.8}$$

For the overall F_s, the sums in both numerator and denominator are extended across all loci, that is,

$$F_s \;=\; \frac{\sum_{\ell=1}^{L}\sum_{i=1}^{K_\ell}(x_i - y_i)^2}{\sum_{\ell=1}^{L}\sum_{i=1}^{K_\ell} z_i(1-z_i)}. \tag{4.9}$$

From $z_i^2 = \frac{1}{4}(x_i + y_i)^2 = \frac{1}{4}(x_i - y_i)^2 + x_i y_i$, we have

$$\frac{(x_i - y_i)^2}{z_i(1-z_i)} \;=\; \frac{(x_i - y_i)^2}{z_i - z_i^2} \;=\; \frac{(x_i - y_i)^2}{z_i - x_i y_i - \frac{1}{4}(x_i - y_i)^2} \;\geq\; \frac{(x_i - y_i)^2}{z_i - x_i y_i}.$$

The rightmost term is a term in F_c. Strict inequality should hold unless $x_i = y_i$. However, the overall F_s and F_c may not be compared.

From $z_i - x_i y_i \geq \frac{1}{2}(x_i - y_i)^2$ shown earlier (part (b)), we have $z_i - x_i y_i - \frac{1}{4}(x_i - y_i)^2 \;\geq\; \frac{1}{4}(x_i - y_i)^2$, so if $x_i - y_i \neq 0$,

$$\frac{(x_i - y_i)^2}{z_i(1-z_i)} \;\leq\; \frac{(x_i - y_i)^2}{\frac{1}{4}(x_i - y_i)^2} \;=\; 4.$$

Equality holds (that is, $\frac{(x_i - y_i)^2}{z_i(1-z_i)} = 4$) if and only if $z_i - x_i y_i = \frac{1}{2}(x_i - y_i)^2$. This is the condition that the terms in F_c and F_k attain maximum value 2 as seen in part (b). Therefore, all three terms defining those measures F's attain their maximum values at the same time.

(d) Restriction on Frequencies

Let $0 < q < 0.5$. The measures of F_c, F_k, F_s are based on alleles whose frequencies are at least (and including) q. Since an allele i may have frequency $< q$ in one sample but $\geq q$ in the other, there are different ways to interpret if this allele has frequency at least q. The following are possibilities:

(1) Allele i has frequency at least q if $x_i,\ y_i \geq q$.
(2) Allele i has frequency at least q if either $x_i \geq q$, or $y_i \geq q$.
(3) Allele i has frequency at least q if $z_i \geq q$.
(4) Allele i has frequency at least q if $\overline{z}_i \geq q$.

It is clear that criterion (1) is more restricted than the rest. Criteria (3) and (4) are more restricted than (2). Also, loci that are "nearly" monomorphic for allele i will be dropped, that is, if the frequency of allele i is at least $1-q$ using one of the criteria listed above. In the program, we use criterion (4). This means that an allele i has frequency at least q (resp. $1-q$) if its frequency in the combined samples is at least q (resp. $1-q$).

All alleles whose frequencies are less than q will be lumped together as one allele in the calculations of F's. Then K_ℓ, representing the total number of alleles at locus ℓ, is the number of alleles whose frequencies $>= q$, plus one, if there are alleles (no matter how many) whose frequencies are $< q$. It should be noted that under this convention, if there is exactly one allele whose frequency $< q$, then the calculation at this locus is the same as if there is no allele being dropped at all!

(e) Weighted harmonic mean sample size

The sample size that best characterizes the effects of random sampling error in the overall estimator must reflect two factors: variation in sample size among loci (due to missing data), and variation in information content at each locus. The appropriate measure is a weighted harmonic mean of the single-locus sample sizes, with the weights proportional to the numbers of alleles.

First, at each locus, we find harmonic mean s_ℓ of individuals having data in two samples:

$$\frac{1}{s_\ell} = \frac{1}{2}\left(\frac{1}{m_\ell} + \frac{1}{n_\ell}\right).$$

Then find the weighted harmonic mean S of s_ℓ ($\ell = 1, \ldots, L$). For Nei/Tajima and Jorde/Ryman methods, the weight w_ℓ of a locus is the number of its alleles ($w_\ell = K_\ell$); it is also the weight assigned to F_c^ℓ in the calculation of overall F_c in the Nei/Tajima method. For Pollak method, $w_\ell = K_\ell - 1$, is the number of its independent alleles at locus ℓ used in the calculation of the overall F_k:

$$\frac{1}{S} = \frac{1}{\sum_{\ell=1}^{L} w_\ell} \sum_{\ell=1}^{L} \frac{w_\ell}{s_\ell}. \tag{4.10}$$

(f) Estimating $\widehat{Ne}$, Plan II

For Pollak and Nei/Tajima temporal methods, there are two approaches for estimating Ne using multi-allelic data. Both use the following general equations:

$$\widehat{Ne} = \frac{|t_2 - t_1|}{2F'}, \qquad \text{where} \quad F' = F - \frac{1}{S}. \tag{4.11}$$

(1) In the first approach, $F^{\ell\,\prime}$ and $\widehat{Ne}^{\ell}$ are calculated separately for each locus ℓ,

$$F^{\ell\,\prime} = F^\ell - \frac{1}{s_\ell}, \qquad \widehat{Ne}^{\ell} = \frac{|t_2 - t_1|}{2\,F^{\ell\,\prime}},$$

and overall $\widehat{Ne}$ is computed as the weighted harmonic mean of the single-locus $\widehat{Ne}^{\ell}$:

$$\frac{1}{\widehat{Ne}} = \frac{1}{\sum_{\ell=1}^{L} w_\ell} \sum_{\ell=1}^{L} \frac{w_\ell}{\widehat{Ne}^{\ell}}. \tag{4.12}$$

The weights w_ℓ (at locus ℓ) is the number of alleles (K_ℓ; Nei and Tajima) or the number of independent alleles ($K_\ell - 1$; Pollak).

(2) In the second approach, overall $\widehat{Ne}$ is computed from the overall F', which is computed from an overall weighted harmonic mean S in (4.10) and an overall weighted mean F, i.e.,

$$F \;=\; \frac{1}{\sum_{\ell=1}^{L} w_\ell} \sum_{\ell=1}^{L} w_\ell F^\ell, \qquad \frac{1}{S} \;=\; \frac{1}{\sum_{\ell=1}^{L} w_\ell} \sum_{\ell=1}^{L} \frac{w_\ell}{s_\ell},$$

$$F' = F - \frac{1}{S}, \qquad \widehat{Ne} \;=\; \frac{|t_2 - t_1|}{2\,F'}.$$

It is easy to see that F' in the second approach is the weighted mean of ${F^\ell}'$, and $\widehat{Ne}$ in both approaches are identical.

For Jorde & Ryman method, F' is calculated from F based on formula (13) in Jorde & Ryman's paper:

$$F_s' \;=\; \frac{F_s\,[1 \;-\; 1/(4S)] \;-\; 1/S}{(1 \;+\; F_s/4)\;\,[1 \;-\; 1/(2S_2)]}, \tag{4.13}$$

where S is the harmonic mean of sample sizes of the two samples, and S_2 is the sample size of sample 2. Under the notations stated at the beginning, where sample sizes at a locus ℓ for samples 1 and 2 are denoted m_ℓ and n_ℓ, respectively, then S_2 is n_ℓ, and S will be the harmonic mean of m_ℓ and n_ℓ, which is s_ℓ. With the possibility of missing data, m_ℓ (resp. n_ℓ, s_ℓ) may not be identical across loci, so we will take S as the weighted mean of s_ℓ as in (4.10), and S_2 as the weighted mean of n_ℓ.

(g) Adjusted for Plan I

For Pollak and Nei & Tajima temporal methods, the measure F0 is adjusted by adding the term

$$\frac{1}{N}, \qquad \text{where } N \text{ is the census size at the first time of sampling.}$$

For Jorde & Ryman method, F_s' is calculated by

$$F_s' \;=\; \frac{F_s\,[1 \;-\; 1/(4S) \;+\; 1/(4N)] \;-\; 1/S \;+\; 1/N}{(1 \;+\; F_s/4)\;\,[1 \;-\; 1/(2S_2)]}, \tag{4.14}$$

(h) Confidence Intervals

For Pollak and Nei&Tajima temporal methods, we find parameter confidence intervals for F using Chi-square distribution where the degree of freedom is taken to be the total number of independent alleles across loci. Then the confidence intervals for $\hat{N}_e$ are derived.

As we see in Pollak and Nei & Tajima methods, the overall estimate of $\widehat{Ne}$ can be calculated as the weighted average of single-locus $\widehat{Ne}^{\,\ell}$, so the Jackknife method on loci may be applied to obtain confidence interval for F, then CI for F' (by subtracting the inverse of weighted harmonic mean of sample sizes), and finally, CI for $\widehat{Ne}$.

For Jorde & Ryman method, the measure F_s' was calculated from both F_s and the harmonic sample sizes in a non-linear way, so we find confidence intervals for F_s' in both parameter and jackknife methods. For parameter CI, we use the total number of independent of alleles as the degree of freedom in Chi-square distribution. For Jackknife method, we follow the authors' recommendations, using equation (4.13) where references to each locus are sequentially removed, to obtain standard error of the overall F'. Then we apply normal distribution to obtain confidence intervals.

(i) Missing Data: Recalculation of Weights in Pollak Method

In the case of missing data, the weight at locus ℓ for measure ${F^{\ell}}'$ is initially taken as

$$w_\ell \;=\; (K_\ell - 1)\cdot s_\ell^2,$$

which is used to calculate the weighted average of F', and then the initial $\widehat{N}_{e0}$. Unless this value is negative or too large, we recalculate the weight by

$$w_\ell \;=\; \frac{(K_\ell - 1)\cdot s_\ell^2}{\left(s_\ell\cdot t + 2\widehat{N}_{e0}\right)^2}, \qquad \text{where } t \text{ is the time gap, } \; t = |t_1 - t_2|.$$

Then the rest of the calculations go the same way as before. Note that with this weight calculation, and in the case of *no missing data*, the factor

$$\frac{s_\ell^2}{\left(s_\ell\cdot t + 2\widehat{N}_{e0}\right)^2}$$

is invariant across loci, so setting the weights as above is the same as setting $w_\ell = (K_\ell - 1)$.

As for Jackknife method, removing a locus will cause an estimate of $\widehat{N}_e$ for the remaining loci with initial weights at those loci, and then a possibility of recalculation of weights to arrive at the estimate of F_k and F_k'. We use values of F_k with one locus being removed to obtain the variance of F_k and then the CI for F_k. Subtracted by the inverse of weighted harmonic mean across loci (the recalculation weights) to obtain CI for F_k' and then CI for $\widehat{N}_e$.