Supporting Information

Samples from subdivided populations yield biased estimates of effective size that overestimate the rate of loss of genetic variation by

Ryman N\textsuperscript{1}, Allendorf FW\textsuperscript{2}, Jorde PE\textsuperscript{3}, Laikre L\textsuperscript{1}, Hössjer O\textsuperscript{4}

Appendix S1

Derivations and mathematical expressions for expected estimates of variance effective population size ($N_{eV}$) when sampling from a population system assuming an island model of migration

1 Introduction

We consider an island model with $s$ islands of diploid size $N$. Our main focus in this appendix is to study the estimator $\hat{N}_{eV}$ of the variance effective size $N_{eV}$ due to Jorde and Ryman (2007). In particular, we study the bias of $\hat{N}_{eV}$ that results from ignoring the effect of migration between islands under non-uniform sampling schemes. We further ignore the effect of new mutations and assume selective neutrality.

For simplicity of presentation we first describe the model, concepts and results in sections 2-4 for a single biallelic locus such as a Single Nucleotide Polymorphism (SNP). We then demonstrate in section 5 that these results generalize to multilocus and multiallelic scenarios. Finally, we present the mathematical derivations in section 6.

\textsuperscript{1}Department of Zoology, Stockholm University
\textsuperscript{2}Division of Biological Sciences, University of Montana
\textsuperscript{3}Department of Biosciences, University of Oslo
\textsuperscript{4}Department of Mathematics, Stockholm University
2 Island Model

We number the islands $i = 1, \ldots, s$. Each of the $N$ individuals in island $i$ has two copies of a certain biallelic marker, henceforth referred to as a gene. We let $p_{it}$ denote the frequency of one of the two alleles in island $i$ at generation $t$, so that $2Np_{it}$ of the genes have this allele, and

$$P_t = \sum_{i=1}^{s} p_{it}/s$$

is the overall allele frequency of the whole population in generation $t$. Following the terminology of Sved and Latter (1977), our framework is stochastic migration with a fixed migration rate $m$. The reproduction scenario from generation $t$ to $t+1$ is described as follows:

1. Within each island $i$ of generation $t$, an infinite gamete pool is generated, with allele frequency $p_{it}$. These gamete pools are formed independently between islands, with no migration between islands involved.

2. The first $N(1-m)$ individuals of island $i$ at generation $t+1$ are obtained by sampling $2N(1-m)$ genes from gamete pool $i$ of generation $t$. This sampling is random, except that selfing is excluded, i.e. each of the $N(1-m)$ individuals is required to have their two gametes from different individuals of island $i$ and generation $t$.

3. The remaining $Nm$ individuals of island $i$ at generation $t+1$ are sampled from a merged pool, to which all $s$ gamete pools of generation $t$ contribute in equal proportions $1/s$, thus having allele frequency $P_t$. The $2Nm$ gametes of these $Nm$ individuals are drawn randomly, apart from the restriction that each individual should have gametes from different individuals in generation $t$ (i.e. no selfing).

4. The procedure in steps 2 and 3 is repeated independently for all islands $i = 1, \ldots, s$.

The construction above guarantees that the selfing probability in each generation is 0. Although we don’t explicitly distinguish between males and females in our diploid model, the analytical approximations of $E(\hat{N}_{eV})$ derived below agree well with our simulations in
the main text, using the EASYPOP program for diploid, dioecious species (Balloux, 2001). See also Hössjer et al. (2013), where dioecious extensions are discussed for reproduction scenarios related to the one above. It is also implicit from the proofs in section 6 that our approximate formulas for $E(\hat{N}_{eV})$ are also valid for a monoecious model where gametes are drawn randomly in steps 2 and 3, with a selfing probability of the order $1/N$. The difference from our model is asymptotically negligible for large populations.

When $m = 1$, the population corresponds to a Wright Fisher model with $sN$ individuals and $2sN$ genes that is modified to have a selfing probability of zero. This is equivalent to panmixia (random mating) of a diploid monoecious organism.

### 3 Fixation Index

The fixation index was introduced by Wright (1921, 1931) as a measure of subpopulation differentiation. It can be formulated in various ways. One definition (see for instance Nei, 1975, p. 123 and Cockerham and Weir, 1987) is

$$F_{ST} = \frac{f_S - f_T}{1 - f_T} = \frac{f_S - (f_S/s + (s-1)f_D/s)}{1 - (f_S/s + (s-1)f_D/s)},$$

(A1)

where $f_S$, $f_T$ and $f_D$ are the probabilities that two randomly chosen genes of generation $t$ are identical-by-state (IBS) when drawn from the same subpopulation, the total population, or different subpopulations, respectively. One may also express $F_{ST}$ as the difference between kinship coefficients of genes sampled from the same and different subpopulations respectively (Barbujani, 1987, Hardy and Vekemans, 1999, Rousset, 2002). However, in the context of $N_{eV}$, it is more convenient to use another, closely related, definition

$$F_{ST,t} = \frac{V_t}{P_t(1 - P_t)},$$

(A2)

equal to the variance $V_t = \sum_{i=1}^s (p_{it} - P_t)^2/s$ of allele frequencies between subpopulations, standardized by $P_t(1 - P_t)$, see for instance Wright (1951), equation 3.12.3 of Crow and Kimura (1970) or equation (12.13) of Nei and Kumar (2000).

Sved and Latter (1977) considered the equilibrium behaviour of $F_{ST,t}$ for the infinite island model ($s = \infty$) under several reproduction scenarios. In particular, for stochastic
migration with a fixed migration rate, their recursive formula (8) can be rewritten as

\[ F_{ST,t+1}^{s=\infty} = \frac{1}{2N} + \left( (1 - m)^2 - \frac{1 - m}{2N} \right) F_{ST,t}. \]  

(A3)

Putting \( F_{ST,t+1} = F_{ST,t} = F_{ST}^{eq} \) and solving for the equilibrium value \( F_{ST}^{eq} \) we obtain

\[ F_{ST}^{eq} \overset{s=\infty}{=} \frac{1}{2N(1 - (1 - m)^2) + 1 - m}. \]  

(A4)

When \( N \) is large and \( m \) small, (A4) is well approximated by the classical expression

\[ F_{ST}^{eq} \overset{s=\infty}{\approx} \frac{1}{4Nm + 1} \]

of Wright (1943). It is of interest to generalize (A4) to situations when \( s < \infty \). However, for finite \( s \) there is no strictly defined equilibrium value of \( F_{ST} \) when \( m > 0 \), since one allele will eventually be fixed in all subpopulations in absence of mutations. The time for this to happen is very long, though, and in subsection 6.1 we show that a quasi equilibrium approximation can be derived as

\[ F_{ST}^{eq} \approx \frac{1}{\frac{s}{s-1}2N(1 - (1 - m)^2) + 1 - m}. \]  

(A5)

Hössjer et al. (2013) derive approximations of \( F_{ST}^{eq} \) for the island model under a variety of different (monoeccious) reproduction schemes. For instance, when fertilization precedes migration, it is shown that

\[ F_{ST}^{eq} \approx \frac{1}{\frac{s}{s-1}2N(1 - (1 - m)^2) + 1 - m}, \]  

(A6)

whereas

\[ F_{ST}^{eq} \approx \frac{1}{\frac{s}{s-1}2N(1 - (1 - m)^2) + (1 - m)^2}, \]  

(A7)

is a good approximation when the order of migration and fertilization is reversed, referred to as stochastic migration with a stochastic migration rate by Sved and Latter (1977). Formula (A5), on the other hand, holds for the reproduction cycle described in section 2. It can be viewed as a hybrid of (A6) and (A7), since fertilization of the fraction \( m \) of all migrating genes takes place between two phases of gamete and individual migration. A common feature of (A5)-(A7) is that sampling for genetic analyses in the offspring generation \( t + 1 \) are taken after migration. See also Hössjer and Ryman (2012), where
$F_{ST}^{eq}$ is computed for more general population structures than the island model, allowing migration rates between pairs of subpopulations and the size of subpopulations to vary.

It is worth pointing out that for all equations (A5)-(A7) we have

$$F_{ST}^{eq} \approx \frac{s-1}{s} \cdot \frac{1}{2N} \text{ when } m = 1.$$  \hspace{1cm} (A8)

This is perhaps surprising, since $F_{ST}^{eq} = 0$ is often assumed under panmictia. However, in view of (A2), a positive value of the fixation index is to be expected, even when the subpopulations are fully connected, since they will still exhibit some stochastic variation in terms of allele frequencies, simply because of the randomness associated with assigning individuals to subpopulations. The alternative definition (A1) of $F_{ST}^{eq}$ in terms of IBS-probabilities (equivalent to identity by descent-probabilities if the infinite alleles model of Kimura, 1971, is used) gives different results depending on whether the two genes are drawn with or without replacement. The former approach has been used by Latter and Sved (1981) and Hössjer (2013). It is equivalent to (A2), with values of $F_{ST}^{eq}$ for $m = 1$ consistent with (A8). On the other hand, the latter approach, when genes from the same island are drawn without replacement, gives a slightly different definition of $F_{ST}^{eq}$, which ignores some (small) allele frequency variation between subpopulations, and in particular it yields $F_{ST}^{eq} = 0$ under panmictia.

In the sequel, we use short hand notation $F_{ST} = F_{ST,t}$ for any value of the fixation index, not necessarily in quasi equilibrium.

4 Variance Effective Population Size

There are several ways to define the global effective size $N_{e, tot}$ of a population. The variance effective size (Crow, 1954)

$$N_{eV, tot} = \frac{P_t(1 - P_t)}{2\text{Var}(P_{t+1} - P_t | P_t)}$$ \hspace{1cm} (A9)

refers to the number of individuals of the Wright-Fisher model with the same amount of genetic drift (quantified as variance of allele frequency change) as in the given population. Other definitions are the inbreeding effective size $N_{eI, tot}$ (Wright, 1931, 1939), the
eigenvalue effective size $N_{E,\text{tot}}$ (Crow, 1954, Ewens, 1982) or the coalescent effective size $N_{C,\text{tot}}$ (Wakeley, 1999, Nordborg and Krone, 2002, Sjödin et al., 2005, Hössjer, 2011). These four notions of global effective size are equivalent under a Wright-Fisher model but may differ for models with substructure or varying population size (Charlesworth, 2009). In particular, for models with high variation in individual reproductive success (Eldon and Wakeley, 2006) the coalescence effective population size is not even defined. For populations of varying size, $N_{e,\text{tot}}$ will also differ depending on which time scale that is used. This may explain fast adaptation in species due to anthropogenic changes, such as insecticide resistance (Karasov et al., 2010).

When genetic model parameters are estimated from real data, it is known that population substructure, migration and sampling schemes may induce bias, not only for the effective population size (Charlesworth, 2009), but also for statistics obtained from allele frequency spectra (Wakeley, 1999, Städler et al, 2009, Cutter et al, 2012) or population size changes (Chiki et al., 2010). Here we focus on these confounding aspects of sampling and migration for the island model when using the temporal method (Waples, 1989) to estimate the variance effective size by $\hat{N}_{eV}$. Varying reproductive success between islands and correlated migration patterns (so called collective dispersal) may also introduce biased estimates of $N_{eV}$, especially when the local effective size $N$ is small (Broquet et al., 2013), although this will not be treated here.

Depending on context, $\hat{N}_{eV}$ can be interpreted as estimating either the global variance effective size or the local variance effective size, defined similarly as in (A9), but from the allele frequency change of a single isolated island without migration ($m = 0$).

The estimate $\hat{N}_{eV}$ is based on sampling $n_t$ individuals from a subset $I_t \subset \{1, \ldots, s\}$ of $k$ islands in generation $t$ and $n_{t+1}$ individuals from another subset $I_{t+1} \subset \{1, \ldots, s\}$ of $k$ islands in generation $t + 1$. We let $0 \leq l \leq k$ denote the number of overlapping sampled subpopulations in generation $t$ and $t + 1$, and

$$
p_t = \frac{\sum_{i \in I_t} p_{i,t}}{k},
$$

$$
p_{t+1} = \frac{\sum_{i \in I_{t+1}} p_{i,t+1}}{k},
$$

are the average allele frequencies of the islands sampled from in generation $t$ and $t + 1$
respectively. The corresponding estimated frequencies of the \( k \) islands sampled from are denoted by \( \hat{p}_t \) and \( \hat{p}_{t+1} \) respectively. Before bias-correction, the estimator of genetic drift proposed by Jorde and Ryman (2007) is

\[
\hat{F} = \frac{(\hat{p}_{t+1} - \hat{p}_t)^2}{\hat{p}_{t+1} + \hat{p}_t} \left( 1 - \frac{\hat{p}_{t+1} + \hat{p}_t}{2} \right).
\]

Individuals may be sampled with or without replacement, randomly from all \( k \) subpopulations or in a stratified manner - from each subpopulation separately. In the case of random sampling with replacement (Plan II of Waples, 1989), the bias-corrected estimator of genetic drift in Jorde and Ryman (2007) is

\[
F^* = \frac{\hat{F} (1 - 1/(4\tilde{n})) - 1/\tilde{n}}{(1 + \hat{F}/4)(1 - 1/(2n_{t+1}))},
\]

where \( \tilde{n} \) is the harmonic mean of \( n_t \) and \( n_{t+1} \). The accompanying estimate of effective population size is

\[
\hat{N}_{eV} = \frac{1}{2F^*} = \frac{\left( \frac{\hat{p}_{t+1} + \hat{p}_t}{2} \left( 1 - \frac{\hat{p}_{t+1} + \hat{p}_t}{2} \right) + \frac{1}{4}(\hat{p}_{t+1} - \hat{p}_t)^2 \right) \left( 1 - \frac{1}{2n_{t+1}} \right)}{2(\hat{p}_{t+1} - \hat{p}_t)^2 \left( 1 - \frac{1}{4\tilde{n}} \right) - \frac{2}{\tilde{n}} \frac{\hat{p}_{t+1} + \hat{p}_t}{2} \left( 1 - \frac{\hat{p}_{t+1} + \hat{p}_t}{2} \right)}. \tag{A10}
\]

Our main goal is to find explicit expressions for

\[
E(\hat{N}_{eV}) \approx \frac{E \left( \frac{\hat{p}_{t+1} + \hat{p}_t}{2} \left( 1 - \frac{\hat{p}_{t+1} + \hat{p}_t}{2} \right) + \frac{1}{4}(\hat{p}_{t+1} - \hat{p}_t)^2 \right) \left( 1 - \frac{1}{2n_{t+1}} \right)}{2E \left( (\hat{p}_{t+1} - \hat{p}_t)^2 \right) \left( 1 - \frac{1}{4\tilde{n}} \right) - \frac{2}{\tilde{n}} E \left( \frac{\hat{p}_{t+1} + \hat{p}_t}{2} \left( 1 - \frac{\hat{p}_{t+1} + \hat{p}_t}{2} \right) \right)} \tag{A11},
\]

using the approximation \( E(X/Y) \approx E(X)/E(Y) \), which will be justified in section 5. We regard \( I_t \) and \( I_{t+1} \) as randomly chosen subsets of \( k \) islands, subject to the restriction of \( l \) overlapping islands. When averaging over \( I_t \) and \( I_{t+1} \) this yields very explicit expressions for \( E(\hat{N}_{eV}) \) that depend on \( s \), \( k \), \( l \), \( N \) and \( F_{ST} \) only. We can motivate this in terms of the multilocus extension of section 5, when the effect of genetic drift at several loci is averaged in a similar manner, although \( I_t \) and \( I_{t+1} \) are kept fixed.

In the present paper we confine ourselves to large sample sizes, \( n_t, n_{t+1} \to \infty \) (which is equivalent to sampling, without replacement, the whole finite population). The effect of ignoring variation due to finite samples is small for moderate or high migration rates. On the other hand, it can be more severe for small migration rates, when different alleles are
close to fixation in different subpopulations, so that the genetic drift gets smaller than the ignored sampling variation. Moreover, the time to reach quasi equilibrium increases for small migration rates, so that occurrence of new mutations cannot be neglected. We plan to treat the effect of finite samples in a follow-up paper, but assume large sample sizes here, with \( \hat{p}_t = p_t, \hat{p}_{t+1} = p_{t+1} \) and terms \( 1/\tilde{n} \) and \( 1/n_{t+1} \) ignored. Then

\[
\hat{N}_{eV} = \frac{\frac{p_t + p_{t+1}}{2} (1 - \frac{p_t + p_{t+1}}{2}) + \frac{1}{3} (p_{t+1} - p_t)^2}{2(p_{t+1} - p_t)^2}, \tag{A12}
\]

and (A11) can be rewritten as

\[
E(\hat{N}_{eV}) \approx \frac{P_t(1 - P_t) - E((p_t - P_t)(\bar{p} - P_t))}{2E((p_{t+1} - p_t)^2)}, \tag{A13}
\]

where

\[
\bar{p} = E(p_{t+1} | p_t).
\]

For the numerator of (A13), we used the fact that \( E(p_{t+1} - \bar{p} | p_t) = 0 \) to deduce

\[
E\left(\frac{p_t + p_{t+1}}{2} (1 - \frac{p_t + p_{t+1}}{2}) + \frac{1}{3} (p_{t+1} - p_t)^2\right) = E\left(\frac{\tilde{p} + m}{2} (1 - \frac{\tilde{p} + m}{2}) + \frac{1}{3} E(\bar{p} - p_t)^2\right)
\]

\[
= E\left((P_t + (\frac{\tilde{p} + m}{2} - P_t)) (1 - P_t - (\frac{\tilde{p} + m}{2} - P_t))\right) + \frac{1}{4} E(\bar{p} - p_t)^2
\]

\[
= P_t(1 - P_t) - E\left((\frac{\tilde{p} + m}{2} - P_t)^2 + \frac{1}{4} E(\bar{p} - p_t)^2\right)
\]

\[
= P_t(1 - P_t) - E((p_t - P_t)(\bar{p} - P_t)).
\]

When \( k = l = s \), i.e. when samples are taken from all islands of generations \( t \) and \( t + 1 \), we notice that \( \bar{p} = p_t = P_t \) and \( p_{t+1} = P_{t+1} \). Then the right hand side of (A13) equals \( N_{eV,\text{tot}} \), so that \( \hat{N}_{eV} \) is an approximately unbiased estimator of \( N_{eV,\text{tot}} \), as noted by Jorde and Ryman (2007).

### 4.1 Results for \( k = l \)

In this case (A13) has the form

\[
E(\hat{N}_{eV}) ^{s=1} \approx N,
\]

as it should be for an unbiased estimator of the variance effective population size. It is proved in subsection 6.2 that

\[
E(\hat{N}_{eV}) ^{s>1} \approx kN \cdot \frac{1 - \frac{(1-m)(s-k)}{k(s-1)} F_{ST}}{1 - (1-m) F_{ST} + \frac{m^2(s-k)}{s-1} 2N F_{ST}}, \tag{A14}
\]

8
which holds for any value of $F_{ST}$, although in practice one often assumes quasi equilibrium and inserts (A5). Formula (A14) further simplifies to

\begin{align}
E(\hat{N}_{ev}) \overset{m=0}{\approx} kN \cdot \frac{1 - \frac{s-k}{k(s-1)} F_{ST}}{1 - F_{ST}},
\end{align}

(A15)

\begin{align}
E(\hat{N}_{ev}) \overset{m=1}{\approx} kN \cdot \frac{1}{1 + \frac{s-k}{s-1} 2NF_{ST}},
\end{align}

(A16)

\begin{align}
E(\hat{N}_{ev}) \overset{s=\infty}{\approx} kN \cdot \frac{1 - \frac{1-m}{k} F_{ST}}{1 - (1-m)F_{ST} + m^2 2NF_{ST}},
\end{align}

(A17)

\begin{align}
E(\hat{N}_{ev}) \overset{s-k>1}{\approx} N_{ev,tot} = sN \cdot \frac{1}{1 - (1-m)F_{ST}}.
\end{align}

(A18)

The last equation provides $N_{ev,tot}$ for any value of $F_{ST}$ in generation $t$, without need to assume quasi equilibrium. It differs slightly from the classical formula

\begin{align}
N_{ev,tot} = \frac{sN}{1 - F_{ST}}
\end{align}

(A19)

of Wright (1943), see also Wang and Caballero (1999, eqn. 15) and Rousset (2004, p. 161). However, it is shown in Hössjer and Ryman (2012) and Appendix D of Hössjer et al. (2013), that the discrepancy between (A18) and (A19) is due to the definition of the reproduction cycle. Formula (A19) holds very generally (even for other population structures than the island model) when fertilization precedes migration, whereas

\begin{align}
N_{ev,tot} = \frac{sN}{1 - (1-m)^2 F_{ST}}
\end{align}

(A20)

holds for the island model when migration precedes fertilization. Formula (A18), on the other hand, is based on stochastic migration with a deterministic migration rate. As discussed in section 3, this reproduction cycle is a hybrid between the models underlying (A19) and (A20).

Hössjer and Ryman (2012), Whitlock and Barton (1997), Wang and Caballero (1999) and Waples (2002) give expressions for the global effective population size for other subdivided population genetic models. See also Wakeley (1999), where the structured coalescent is used to compute $N_{eC,tot}$ for the island model in the limit of small migration rates $m$, large $N$ and constant $Nm$. 

9
Whereas \( \hat{N}_{eV} \) is an approximately unbiased estimator of \( N_{eV,\text{tot}} \) when samples are taken from all islands, the expected value of \( \hat{N}_{eV} \) is of order \( kN \) for any \( k \) and thus tends to underestimate \( N_{eV,\text{tot}} \) when \( k < s \). This can be explained as an upward bias of the estimated amount of genetic drift, since 1) migration from other islands is erroneously interpreted as genetic drift and 2) the allele frequency change between generations \( t \) and \( t + 1 \) is only averaged over \( k \) rather than \( s \) islands.

Wang and Whitlock (2003) consider a structured model with \( s = 2 \) islands, either one focal island and an infinite source, or two finite islands of possibly different size. In the latter case a total of four samples (\( s = k = l = 2 \)) is required in order to estimate jointly the size of and migration rate to each island (see Figure 5 in their paper). In this way they are able to separate the effects of genetic drift and migration, so that migration does not confound estimates of genetic drift. It would be of interest to generalize their method to \( s > 2 \). However, it requires that all subpopulations can be identified. As discussed in the main text, for marine applications, this is often difficult.

### 4.2 Results for \( l = 0 \)

In this case \( (A13) \) has the form

\[
E(\hat{N}_{eV}) \approx kN \cdot \frac{1 + \frac{1-m}{s-1} F_{ST}}{1 - (1 - m) F_{ST} + \frac{((1-m)^2 + 1)(s-k) + 2k(1-m)}{s-1} 2N F_{ST}},
\]

whether quasi equilibrium \( (A5) \) has been attained or not. Some special cases of \( (A21) \) are

\[
E(\hat{N}_{eV}) \overset{m=0}{\approx} kN \cdot \frac{1 + \frac{F_{ST}}{s-1}}{1 - F_{ST} + \frac{2s}{s-1} 2N F_{ST}},
\]

\[
E(\hat{N}_{eV}) \overset{m=1}{\approx} kN \cdot \frac{1}{1 + \frac{s-k}{s-1} 2N F_{ST}}
\]

and

\[
E(\hat{N}_{eV}) \overset{s=\infty}{\approx} \frac{k}{(1 - (1 - m) F_{ST})/N + ((1 - m)^2 + 1) 2F_{ST}} = \frac{k}{2(F_{ST} + F_{ST,t+1})},
\]

where the last step follows from the recursion formula \( (A35) \) for \( F_{ST} \) derived in subsection 6.1.
5 More than two alleles and multiple loci

We have so far assumed a single biallelic locus, and our derived expressions for \( E(\hat{N}_{eV}) \) rely on the approximation \( E(X/Y) \approx E(X)/E(Y) \) used in (A11). We will now extend the argument to multiple, possibly multiallelic, loci, and argue that the right-hand side of (A11) provides a good approximation not only of \( E(\hat{N}_{eV}) \), but also of \( \hat{N}_{eV} \), when the number of loci \( J \) is large.

Let \( a_j \) denote the number of possible alleles at locus \( j \). For SNPs we have \( a_j \equiv 2 \), but markers with higher values of \( a_j \), such as microsatellites (see for instance Payseur and Cutter, 2006) can also be incorporated within our framework. Put \( p_t^{jq} \) for the average allele frequency in generation \( t \) (over all islands sampled from) of allele \( q \) at locus \( j \). In the limit of infinite samples, the allele frequency estimator \( \hat{p}_j^{iq} \) equals \( p_t^{jq} \). Then the natural extension of (A10) is

\[
\hat{N}_{eV} = \frac{\sum_{j=1}^{J} \sum_{q=1}^{a_j} \left( \frac{p_t^{jq} + p_t^{iq}}{2} \left( 1 - \frac{p_t^{jq} + p_t^{iq}}{2} \right) + \frac{1}{4}(p_t^{jq} - p_t^{iq})^2 \right)}{\sum_{j=1}^{J} \sum_{q=1}^{a_j} 2(p_t^{jq} - p_t^{iq})^2}.
\]  

(A25)

Unless all loci are in strong linkage disequilibrium (LD), we can justify that the approximations

\[
\hat{N}_{eV} \approx E(\hat{N}_{eV}) \approx \frac{\sum_{j=1}^{J} \sum_{q=1}^{a_j} E \left( \frac{p_t^{jq} + p_t^{iq}}{2} \left( 1 - \frac{p_t^{jq} + p_t^{iq}}{2} \right) + \frac{1}{4}(p_t^{jq} - p_t^{iq})^2 \right)}{\sum_{j=1}^{J} \sum_{q=1}^{a_j} 2E \left( (p_t^{jq} - p_t^{iq})^2 \right)}
\]  

(A26)

are increasingly accurate the larger \( J \) is. Indeed, the coefficient of variation of the numerator and denominator of (A25) will both tend to zero with increasing \( J \) under mild conditions of average LD decay between the \( J(J-1)/2 \) pairs of loci.

The estimator (A26) can be written as a weighted sum (cf. (A11))

\[
\sum_{j=1}^{J} \sum_{q=1}^{a_j} w_{jq} \frac{E \left( \frac{p_t^{jq} + p_t^{iq}}{2} \left( 1 - \frac{p_t^{jq} + p_t^{iq}}{2} \right) + \frac{1}{4}(p_t^{jq} - p_t^{iq})^2 \right)}{2E \left( (p_t^{jq} - p_t^{iq})^2 \right)},
\]  

(A27)

where \( w_{jq} \) are weights proportional to \( E(p_t^{jq} - p_t^{iq})^2 \), satisfying \( \sum_{jq} w_{jq} = 1 \). But (A27) is a weighted sum of terms as in the right-hand sides of (A14) and (A21), with \( F_{ST}^{jq} \), the
locus and allele specific version of Wright’s fixation index, instead of $F_{ST}$. The right hand sides of (A14) and (A21) have the form $kN(1 + UF_{ST})/(1 + VF_{ST})$, for some constants $U$ and $V$ depending on $s$, $k$, $m$ and $N$. Hence, in the multilocus and multiallelic case, we find that

$$E(\hat{N}_{eV}) \approx kN \sum_{jq} w_{jq} \frac{1 + UF_{ST}^{jq}}{1 + VF_{ST}^{jq}} \approx kN \frac{1 + U \bar{F}_{ST}}{1 + V \bar{F}_{ST}},$$

(A28)

where

$$\bar{F}_{ST} = \sum_{jq} w_{jq} F_{ST}^{jq}$$

(A29)

can be viewed as a multiallelic multilocus summary value of $F_{ST}$. The second approximation in (A28) follows by a Taylor expansion argument and is accurate when all $UF_{ST}^{jq}$ and $VF_{ST}^{jq}$ are small.

The conclusion is that the right-hand sides of (A14) and (A21) provide good approximations not only of $E(\hat{N}_{eV})$, but also of $\hat{N}_{eV}$, when the number of loci $J$ is large, whether these are biallelic or not, provided we replace $F_{ST}$ by (A29). However, if quasi equilibrium (A5) has been attained for all alleles at all loci, we can replace the multilocus $\bar{F}_{ST}$ by the right hand side of (A5).

It is interesting to compare $\bar{F}_{ST}$ with $G_{ST}$, the most commonly used multilocus and multiallele extension of $F_{ST}$ (Nei, 1973). Ryman and Leimar (2008) have obtained an approximation of $G_{ST}$ in presence of migration and/or mutation (see equation (A15) in their appendix) which, in absence of mutation, is quite close to the quasi-stationarity approximation (A5).

It turns out that $\bar{F}_{ST}$ is related to $G_{ST}$ even when quasi stationarity cannot be assumed. Indeed, it is shown in chapter 3 of Wright (1978) that $G_{ST}$ can be written in the form (A29) with weights $w_{jq}$ proportional to $P^{jq}_{t}(1 - P^{jq}_{t})$, where $P^{jq}_{t}$ is the frequency of allele $q$ at locus $j$ over the whole population. See also Nei (1977) and Chakraborty and Leimar (1987) for related discussions. However, it follows from (A40) and (A42) that our weights satisfy

$$w_{jq} \propto E(p^{jq}_{t+1} - p^{jq}_{t})^2 = P^{jq}_{t}(1 - P^{jq}_{t}) \left(1/(2kN) + O(F_{ST}^{jq})\right).$$

(A30)

The main difference between $\bar{F}_{ST}$ and $G_{ST}$ is thus the extra $O(F_{ST}^{jq})$ terms appearing in
the weights (A30) of $\bar{F}_{ST}$. They can be neglected only when $k = l = s$ and $F_{ST}$ is small, cf. (A40) and (A42).

6 Derivations

6.1 Derivation of (A5)

Our strategy will be to approximate the recursion formula (A3), replacing $F_{ST,t+1}$ by an approximation $F_{ST,t+1}^{\text{appr}}$, and then put $F_{ST,t+1}^{\text{appr}} = F_{ST,t}$. The solution of this equation is $F_{ST}^{\text{eq}}$. To this end, write $Q_t = \sum_{i=1}^{s}(p_{it} - P_t)^2/s$ and $Q_{t+1} = \sum_{i=1}^{s}(p_{i,t+1} - P_{t+1})^2/s$, so that

$$F_{ST,t} = Q_t / (P_t(1 - P_t)),$$

$$F_{ST,t+1} = Q_{t+1} / (P_{t+1}(1 - P_{t+1})).$$

Our approximation

$$F_{ST,t+1}^{\text{appr}} = \frac{E(Q_{t+1})}{P_t(1 - P_t)} \quad (A31)$$

is increasingly accurate the larger $N$ is, since then the random fluctuations of $Q_{t+1}$ around $E(Q_{t+1})$ and of $P_{t+1}$ around $P_t$ will decrease. The allele frequency of island $i$ at generation $t+1$ can be written as

$$p_{i,t+1} = (1 - m)(p_{it} + \Delta_{1i}) + m(P_t + \Delta_{2i}), \quad (A32)$$

where $\Delta_{1i}$ quantifies the amount of genetic drift, i.e. change in allele frequency, within subpopulation $i$ between generations $t$ and $t + 1$ and $\Delta_{2i}$ is the amount of genetic drift among individuals of subpopulation $i$ and generation $t + 1$ that are sampled from whole population of generation $t$. We will show that

$$E(\Delta_{1i}^2) = \frac{p_{it}(1 - p_{it})}{2N(1 - m)} + O\left(\frac{1}{N^2(1 - m)}\right),$$

$$E(\Delta_{2i}^2) = \frac{P_t(1 - P_t)}{2Nm} + O\left(\frac{1}{N^2 m}\right). \quad (A33)$$

Whereas the two remainder terms in (A33) vanish for a monoecious Wright Fisher model with binomial variation, we will motivate that they are added for a population with no selfing, as described in the four steps of the reproduction scenario in section 2. For
simplicity, we do this for the lower part of (A33), since the argument for the upper part is analogous. We can write
\[
\Delta_{2i} = \frac{1}{2Nm} \sum_{j=1}^{2Nm} A_j - P_t,
\]
where \( A_j \in \{0, 1\} \) is an allele indicator, referring to the allele of gamete \( j \) that is transferred from the merged gamete pool to subpopulation \( i \) in Step 3 of the reproduction scheme.

Assume further that these \( 2Nm \) gametes are numbered so that \( A_{2a-1} \) and \( A_{2a} \) form the genotype of individual \( a = 1, \ldots, Nm \).

Clearly, \( E(A_j) = P_t \) and \( \text{Var}(A_1) = P_t(1 - P_t) \). This implies \( E(\Delta_{2i}) = 0 \) and
\[
E(\Delta_{2i}^2) = \text{Cov}(\Delta_{2i})
\]
\[
= \frac{1}{(2Nm)^2} (2Nm \text{Var}(A_1) + 2Nm \text{Cov}(A_1, A_2))
\]
\[
= \frac{P_t(1-P_t)}{2Nm} + \frac{\text{Cov}(A_1, A_2)}{2Nm},
\]
using the fact that \( \text{Cov}(A_j, A_k) = 0 \) when \( j \) and \( k \) correspond to gametes of different individuals. We thus need to prove that \( \text{Cov}(A_1, A_2) \) is of order \( 1/(sN) \). To this end, we write \( sN = N_{00} + N_{01} + N_{11} \), where \( N_{xy} \) refers to the number of individuals in generation \( t \) with genotype \( xy \). Then, conditioning on the kind of parental genotype from which \( A_1 \) was inherited, we find that
\[
\text{Cov}(A_1, A_2) = E(A_1A_2) - P_t^2
\]
\[
= \frac{N_{00}}{sN} \cdot 0 + \frac{N_{01}}{sN} \cdot \frac{1}{2} \frac{2sNP_t - 1}{2sN - 2} + \frac{N_{11}}{sN} \cdot \frac{2sNP_t - 2}{2sN - 2} - P_t^2
\]
\[
= O \left( \frac{1}{sN} \right),
\]
regardless of the values of \( N_{00} \), \( N_{01} \) and \( N_{11} \), since \( P_t = 0.5N_{01}/(Ns) + N_{11}/(Ns) \).

We can use (A32) and obtain an expansion
\[
Q_{t+1} = \frac{1}{s} \sum_{i=1}^{s} ((1-m)(p_i - P_t) + (1-m)(\Delta_{1i} - \Delta_1) + m(\Delta_{2i} - \Delta_2))^2,
\]
of \( Q_{t+1} \), with \( \Delta_1 = \sum_{i=1}^{s} \Delta_{1i}/s \) and \( \Delta_2 = \sum_{i=1}^{s} \Delta_{2i}/s \). Taking expectation, we find that
\[
E(Q_{t+1}) = (1-m)^2 Q_t + \frac{N_{2s}}{s^2} \sum_{i=1}^{s} E(\Delta_{1i}^2) + \frac{(s-1)m^2}{s^2} \sum_{i=1}^{s} E(\Delta_{2i}^2)
\]
\[
= (1-m)^2 Q_t + \frac{P_t(1-P_t)(s-1)}{2Ns} ((1-m)(1-F_{ST}) + m) + O \left( \frac{1}{N^2} \right), \tag{A34}
\]
using (A33) in the last step, and in particular the $O(1/N^2)$ term originates from the remainder terms of (A33). Dividing both sides of (A34) by $P_t(1 - P_t)$ gives a recursion formula

$$F_{ST,t+1}^{\text{appr}} = \frac{s - 1}{2Ns} + \left(1 - m\right)^2 - \frac{(1 - m)(s - 1)}{2Ns} F_{ST,t} + O\left(\frac{1}{N^2}\right). \quad (A35)$$

Finally, we put $F_{ST,t+1}^{\text{appr}} = F_{ST,t}$ and ignore the remainder term $O(1/N^2)$ in (A35), which is asymptotically negligible for large populations. The solution of the resulting equation is (A5).

### 6.2 Derivation of (A14)

We will make use of the fact that

$$E \left( (p_t - P_t)^2 \right) = P_t(1 - P_t) \frac{s - k}{k(s - 1)} F_{ST}, \quad (A36)$$

a standard result for sampling without replacement (in this case sampling $k$ islands out of $s$, i.e. randomly choosing $I_t$). We notice further that

$$\bar{p} = (1 - m)p_t + mP_t,$$

so that $\bar{p} - P_t = (1 - m)(p_t - P_t)$. In conjunction with (A36) this implies

$$E \left( (\bar{p} - P_t)(p_t - P_t) \right) = (1 - m)E \left( (p_t - P_t)^2 \right) = P_t(1 - P_t) \frac{(1 - m)(s - k)}{k(s - 1)} F_{ST}. \quad (A37)$$

Using (A32), the allele frequency change of island $i$ can be written as

$$p_{i,t+1} - p_{it} = (1 - m)\Delta_{1i} + m\Delta_{2i} - m(p_{it} - P_t). \quad (A38)$$

Averaging (A38) over all $k$ sampled islands we get

$$p_{t+1} - p_t = \frac{1}{k} \sum_{i \in I_t} (p_{i,t+1} - p_{it})$$

$$= \frac{(1 - m)}{k} \sum_{i \in I_t} \Delta_{1i} + \frac{m}{k} \sum_{i \in I_t} \Delta_{2i} - m(p_t - P_t). \quad (A39)$$
We then square $(A39)$, take expectation, use $(A33)$ (with the asymptotically negligible remainder terms ignored) and $(A36)$. This yields

$$E((p_{t+1} - p_t)^2) = \frac{(1-m)^2}{k^2} E \left( \sum_{i \in I_t} E(\Delta_i^2 | p_t) \right) + \frac{m^2}{k^2} E \left( \sum_{i \in I_t} E(\Delta_i^2) \right) + m^2 E (p_t - P_t)^2$$

$$= \frac{(1-m)^2}{k^2} E \left( \sum_{i \in I_t} \frac{p_t(1-p_t)}{2N(1-m)} \right) + \frac{m^2}{k^2} E \left( \sum_{i \in I_t} \frac{P_t(1-P_t)}{2Nm} \right) + m^2 E (p_t - P_t)^2$$

$$= \frac{P_t(1-P_t)}{2kN} ((1-m)(1-F_{ST}) + m) + m^2 P_t(1-P_t) \frac{s-k}{k(s-1)} F_{ST}$$

$$= P_t(1-P_t) \left( \frac{1-(1-m)F_{ST}}{2kN} + m^2 \frac{s-k}{k(s-1)} F_{ST} \right). \quad (A40)$$

Inserting $(A37)$ and $(A40)$ into $(A13)$ we arrive at $(A14)$.

### 6.3 Derivation of $(A21)$

In this case

$$\bar{p} = P_t - \frac{(1-m)k}{s-k}(p_t - P_t)$$

and hence $(A36)$ implies

$$E((\bar{p} - P_t)(p_t - P_t)) = -\frac{(1-m)k}{s-k} E((p_t - P_t)^2) = -P_t(1-P_t) \frac{(1-m)}{s-1} F_{ST}. \quad (A41)$$

In order to study $E((p_{t+1} - p_t)^2)$, we introduce

$$\bar{p} = \frac{1}{k} \sum_{i \in I_{t+1}} p_{it},$$

and notice that

$$E(p_{t+1} | \bar{p}) = (1-m)\bar{p} + mP_t.$$

This yields

$$E((p_{t+1} - p_t)^2) = E((p_{t+1} - p_t)^2 | \bar{p})$$

$$= E((p_{t+1} - (1-m)\bar{p} - mP_t)^2) + E(((1-m)\bar{p} + mP_t - p_t)^2)$$

$$= \frac{P_t(1-P_t)}{2Nk} \left[ (1-(1-m)F_{ST}) + E(((1-m)(\bar{p} - P_t) - (p_t - P_t))^2) \right]$$

$$= \frac{P_t(1-P_t)}{2Nk} \left[ (1-(1-m)F_{ST}) + ((1-m)^2 + 1) E((p_t - P_t)^2) \right.$$

$$- 2(1-m)E((\bar{p} - P_t)(p_t - P_t))$$

$$= \frac{P_t(1-P_t)}{2Nk} \left[ (1-(1-m)F_{ST}) \right.$$  

$$+ P_t(1-P_t) F_{ST} \left[ (((1-m)^2 + 1) \frac{s-k}{k(s-1)} + \frac{2(1-m)}{s-1} \right), \quad (A42)$$
using a similar argument as in (A40) for \( E ((p_{t+1} - (1 - m)\tilde{p} - mP_t)^2) \) in the third equality of (A42) and

\[
E ((\tilde{p} - P_t)(p_t - P_t)) = -\frac{P_t(1 - P_t)F_{ST}}{s - 1}
\]

in the fifth equality. Formula (A43) is a standard result for sampling with replacement, and is proved similarly as (A36). Inserting (A41) and (A42) into (A13) we finally arrive at (A21).

**References**


Kimura M (1971) Theoretical foundations of population genetics at the molecular level.


