Theoretical Population Biology 79 (2011) 39–52

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Properties of Weir and Cockerham's F_{st} estimators and associated bootstrap confidence intervals

Sivan Leviyang [a,*], Matthew B. Hamilton [b]

[a] Georgetown University, Department of Mathematics, United States
[b] Georgetown University, Department of Biology, United States

ARTICLE INFO

Article history:
Received 11 July 2010
Available online 20 November 2010

Keywords:
F_{st}
Coalescent
Bootstrap

ABSTRACT

Weir and Cockerham introduced single locus and multiloci F_{st} estimators for the parameter θ. These estimators are commonly used, but little beyond their bias and variance is known. In this work, we develop formulas that allow us to describe how the underlying value of θ and the genetic diversity of sampled loci affect the distributions of these estimators. We show that in certain settings, these estimators are close to normal, while in others they are far from normal. We use these results to analyze confidence interval construction for θ, showing that the percentile-t bootstrap works well while the BCa bootstrap works poorly. Our results are derived using a novel coalescent based method.

1. Introduction

Sewall Wright introduced F_{st} to quantify the effect of structure on the genetics of a population (Wright, 1931). Later, Cockerham placed F_{st} on a firm statistical footing by identifying the parameter, which he labeled as θ, that Wright's formula for F_{st} was implicitly estimating. The parameter θ is important in many applications. (Weir, 1996) and the construction and analysis of a corresponding estimator $\hat{\theta}$ has occupied many authors (see Weir and Hill, 2002 for a review). In Weir and Cockerham (1984), Weir and Cockerham (hereafter WC) introduced estimators of θ for the cases of single locus and multiloci data which we label $\hat{\theta}_{\text{locus}}$ and $\hat{\theta}_{\text{loci}}$ respectively. Through analysis and simulation, WC showed that both estimators performed well relative to previously suggested $\hat{\theta}$ and today these estimators are commonly used.

The parameter θ is fairly well understood, having been derived under many different evolutionary models ranging from island models to more sophisticated stepping stone models (Slatkin, 1991; Rousset, 1996; Wilkinson-Herbots, 1998). In contrast, estimators of θ and specifically $\hat{\theta}_{\text{locus}}$ and $\hat{\theta}_{\text{loci}}$ have received less attention.

Some partial results do exist for $\hat{\theta}_{\text{locus}}$ and $\hat{\theta}_{\text{loci}}$. As the number of subpopulations becomes large, both estimators have been shown to converge to θ in a variety of population structure models Nei et al. (1977), Leviyang (in press, 2010) and Rottenstreich et al. (2007), but usually the number of subpopulations sampled is modest so asymptotic results do not apply. When θ is small, simulations and analytic approximations have shown that $\hat{\theta}_{\text{locus}}$ and $\hat{\theta}_{\text{loci}}$ are roughly chi-squared distributed Li (1996), Weir et al. (2005) and Samanta et al. (2009), although the generality of these results under different evolutionary models is unclear. For general values of θ, explicit formulas have been derived for the bias and variance of $\hat{\theta}_{\text{locus}}$ (Raufaste and Bonhomme, 2000; Weir and Cockerham, 1984), but little is known about higher order moments and no results exist for any moment of $\hat{\theta}_{\text{loci}}$. The situation for other estimators, for example the normal based estimators discussed in Weir and Hill (2002) and Samanta et al. (2009), is similar.

Without precise knowledge of the distributions of $\hat{\theta}_{\text{locus}}$ and $\hat{\theta}_{\text{loci}}$, confidence interval construction typically proceeds through resampling techniques. In Weir and Cockerham (1984), WC advocated bootstrapping over loci to construct confidence intervals for $\hat{\theta}_{\text{loci}}$ and indeed most software packages that estimate θ also construct such confidence intervals (e.g. fstat, genpop). Through simulation, bootstrapping approaches are seen to increase in accuracy as the number of loci and subpopulations sampled increase (Samanta et al., 2009), but no general results exist that allow us to understand when bootstrapping will work well.

In this paper we demonstrate some novel results concerning the distributions of $\hat{\theta}_{\text{locus}}$ and $\hat{\theta}_{\text{loci}}$. We examine the case in which many samples are taken from a small number of subpopulations, a common situation in application. In this setting, we are able to characterize the skewness of $\hat{\theta}_{\text{locus}}$, extending existing results for the mean and variance. We are also able to characterize the mean, variance, and skewness of $\hat{\theta}_{\text{loci}}$, where previously no results existed. With knowledge concerning the first three moments, we are able to clarify the role of θ and genetic diversity in determining

* Corresponding author.
E-mail address: sr286@georgetown.edu (S. Leviyang).

0040-5809/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2010.11.001

the distributions of $\hat{\theta}_{\text{locus}}$ and $\hat{\theta}_{\text{loci}}$. Our results are based on a novel coalescent approach that greatly simplifies computations associated with $\hat{\theta}_{\text{locus}}$. We apply our results to analyze the accuracy of bootstrap confidence intervals for $\hat{\theta}_{\text{loci}}$.

1.1. Definitions

We consider a population that is decomposed into distinct subpopulations. We assume that samples are taken from d subpopulations with samples taking the form of allelic data over L *unlinked* loci. We let A_ℓ be the number of alleles found at locus ℓ and we associate with each locus a collection of probabilities $p_{\ell,1}, p_{\ell,2}, \ldots, p_{\ell,A_\ell}$ such that $\sum_{u=1}^{A_\ell} p_{\ell,u} = 1$. We refer to a specific choice for the $p_{\ell,u}$ as the *allele model*, the meaning of this nomenclature will be made clear shortly. Letting n_k be the number of samples taken from sample deme k, we assume that $n_k \gg d$ for $k = 1, 2, \ldots, d$. We are interested in sampling regimes in which d is small, say 2 or 5, while L is modest, say 10 or 20, since these are most common in application. In our discussion below we set $n_k = n$ for all $k = 1, 2, \ldots, d$. Our results hold for general n_k as long as each $n_k \gg d$, but taking $n_k = n$ greatly simplifies notation and the essential ideas are not affected.

To specify the genetic state of each sample we first introduce a class of models. We will then study a specific model within this class. We do this to emphasize that our analytic methods apply to a large class of models, even though in this paper we consider a single model within this class.

Following ideas of Kingman, we specify the genetic state of each sample through the use of random partitions Kingman (1978). To explain this, consider the n samples taken from subpopulation k. We partition these n samples into B_k cohorts which we refer to as blocks (here we follow the nomenclature of Pitman (2002) and references therein). The blocks are numbered $1, 2, \ldots, B_k$ and the *fraction* of samples in block j is given by b_{kj}, so $\sum_{j=1}^{B_k} b_{kj} = 1$. B_k and the b_{kj} are random variables that specify a random partition, for now we leave these random variables unspecified allowing us to consider, as mentioned above, a class of models. To make our notation clear, suppose for example that $n = 10$ and we partition the 10 samples into three blocks of size 5, 4, and 1, then $B_k = 3$ and $b_{k1} = 0.5$, $b_{k2} = 0.4$, $b_{k3} = 0.1$. We apply such a partitioning for each $k = 1, 2, \ldots, d$. We let B represent the total number of blocks over all subpopulations sampled, i.e. $B = \sum_{k=1}^{d} B_k$.

With the partition variables B_k, b_{kj} in hand, we now specify the genetic state of the samples. All samples that share a block, share the same genetic state. That is, for each locus all samples within a block have the same allele. In this way we can speak about the genetic state of a block. For locus ℓ, the probability that a block has allele type u is $p_{\ell,u}$. This is why we refer to the $p_{\ell,u}$ as an allele model. For a given block, the allelic state of each locus is determined independently reflecting our assumption of unlinked loci. The genetic state of the B blocks are determined independently and in fact are identically distributed.

Cockerham defined θ through indicator variables. Let x_{kju} be an indicator variable for sample j from subpopulation k having allelic state u. Then θ can be defined as the correlation between x_{kju} and $x_{kj'u}$ for $j \neq j'$.

Many different types of population structure models fit into our general model. For instance, for the infinite island model with random mating and without mutation the B_k, b_{kj} will be determined by the Ewens sampling distribution and $p_{\ell,u}$ will be the frequency of allele type u at locus ℓ over the whole population. On the other hand, we can also include mutation in the infinite island model. If we choose a mutation rate that is much less than the migration rate then B_k, b_{kj} will still be determined by the Ewens sampling distribution, but now we can think of the $p_{\ell,u}$ as the equilibrium frequency of allele u at locus ℓ for a certain unspecified mutation model. If we choose a high mutation rate, our model still applies but B_k, b_{kj} will not be determined by the Ewens sampling distribution. The symmetry and infinite size of the island model are unnecessary. If one samples from a stepping stone model, say, with the d sampled demes sufficiently far away so that two lineages from separate demes will experience many mutations before coalescing, then our model will apply. Indeed, essentially all that is required is that *samples from different sampled demes have independent genetic states*.

Below, we will frame much of our arguments in the context of the general variables B_k, b_{kj}. But in order to derive specific results, we need to specify a choice for their distribution. We consider B_k, b_{kj} given by the Ewens sampling formula with parameter set to θ. This model is equivalent to assuming an infinite island model with scaled migration rate $\Gamma = \left(\frac{1}{\theta} - 1\right)$ (Γ is the usual 2 Nm for haploid and 4 Nm for diploid populations). However, we emphasize that even this specific choice of distribution for B_k, b_{kj} allows for many types of models.

Having specified the distribution of B_k, b_{kj} for a given θ, our model is completely determined by our choice for θ and the $p_{\ell,u}$. In some instances it will be more convenient to work with Γ rather than θ.

2. Results

We derive formulas for the bias, variance, and skewness of $\hat{\theta}_{\text{locus}}$ in the settings of $\theta \ll 1$ and $\theta \approx 1$. Our formulas for $\hat{\theta}_{\text{locus}}$ allow us to develop parallel formulas for $\hat{\theta}_{\text{loci}}$. The formulas we derive are asymptotic in that their precision increases as θ gets closer to 0 or 1 as well as when d grows larger. Our asymptotic formulas, combined with numerical simulations for θ in the intermediate range of [0, 1], provide a description of $\hat{\theta}_{\text{locus}}$ and $\hat{\theta}_{\text{loci}}$ for general θ. By developing formulas for skewness and by demonstrating through analysis and simulation that higher order moments should be small, we are able to show that in certain parameter regimes, $\hat{\theta}_{\text{loci}}$ is close to normal. This in turn allows us to gain deeper insight into confidence interval construction for $\hat{\theta}_{\text{loci}}$.

As mentioned, our model is determined by the parameters θ and the $p_{\ell,u}$ for each locus. For locus ℓ we define $H^{(\ell)} = \sum_{u=1}^{A_\ell} p_{\ell,u}^2$. We use $H^{(\ell)}$ as a diversity measure for locus ℓ in analogy with homozygosity levels in a panmicitic, diploid population under Hardy–Weinberg equilibrium. If H is near zero diversity is high, while H near 1 corresponds to low diversity.

Below, we examine the distributions of $\hat{\theta}_{\text{locus}}$ and $\hat{\theta}_{\text{loci}}$ as θ takes different values in its range [0, 1] with the cases $\theta \ll 1$ and $\theta \approx 1$ being the endpoints of the range. In this same way, we consider $\hat{\theta}_{\text{locus}}$ and $\hat{\theta}_{\text{loci}}$ over a range of diversity levels. Maximum diversity is achieved if each block has a different allelic state. More specifically, recall that our model forms B blocks. Maximum diversity occurs if each of the B blocks is assigned a different allelic state. We refer to this allele model as the infinite allele model, hereafter IAM. Formally, the IAM is not included under our definition of an allele model, rather the IAM is a limit of allele models as $H \to 0$. However, as a limiting model the IAM will be useful and so we will include it under the allele models we discuss. On the other end of the diversity spectrum, given any allele model we can consider a corresponding, less diverse, biallelic allele model, hereafter BAM, by grouping all alleles that are not allele type 1 into a single allelic class. The BAM and IAM will serve as endpoints for the diversity level of all possible allele models.

Before stating our results, we define the following four quantities related to $\hat{\theta}_{\text{locus}}$:

$$\text{MSE}_{\text{locus}} = E[(\hat{\theta}_{\text{locus}} - \theta)^2] \tag{2.1}$$

$$\sigma^2_{\text{locus}} = E[(\hat{\theta}_{\text{locus}} - E[\hat{\theta}_{\text{locus}}])^2]$$
$$\kappa_{1,\text{locus}} = \frac{E[\hat{\theta}_{\text{locus}} - \theta]}{\sigma_{\text{locus}}}$$
$$\kappa_{3,\text{locus}} = \frac{E[(\hat{\theta}_{\text{locus}} - E[\hat{\theta}_{\text{locus}}])^3]}{\sigma^3_{\text{locus}}}.$$

We define corresponding quantities for $\hat{\theta}_{\text{loci}}$, e.g. MSE_{loci}. $\text{MSE}_{\text{locus}}$ is mean square error (MSE), σ_{locus} is variance, $\kappa_{1,\text{locus}}$ is a scaled bias, and $\kappa_{3,\text{locus}}$ is skewness. We refer to $\kappa_{1,\text{locus}}$ and $\kappa_{3,\text{locus}}$ collectively as cumulants. We use MSE as a measure of our estimator's error.

2.1. Results for $\hat{\theta}_{\text{locus}}$

In this section we consider only single locus data and hence $\hat{\theta}_{\text{locus}}$. Correspondingly, we write H, p_u, and A for $H^{(\ell)}$, $p_{\ell,u}$ and A_ℓ respectively. We derive the following formulas for MSE:

$$\text{MSE}_{\text{locus}} = \begin{cases} \dfrac{2\theta^2}{d}\left(\dfrac{H - 2I + H^2}{(1-H)^2}\right) & \text{if } \theta \ll 1 \\ \dfrac{1-\theta}{3d(1-H)} & \text{if } \theta \approx 1, \end{cases} \tag{2.2}$$

where $I = \sum_{u=1}^{A} p_u^3$.

From (2.2), we see that $\text{MSE}_{\text{locus}}$ collapses as θ gets close to 0 or 1, with a more severe collapse for $\theta \ll 1$ than $\theta \approx 1$ (our $\text{MSE}_{\text{locus}}$ formula for $\theta \ll 1$ is observed in Weir and Hill (2002) and references therein). When θ is small, (2.2) shows that θ can be estimated only to within an order of θ, a high relative error (see Samanta et al., 2009 for a similar observation). The situation is worse for $\theta \approx 1$ where $1-\theta$ can be estimated only to an order of $\sqrt{\frac{1}{1-\theta}}$.

The case $\theta = 0$ deserves special comment. (2.2) ignores expressions of $O\left(\frac{1}{n}\right)$ for $k = 1, 2, \ldots, d$ and as $\theta \to 0$ such expressions will dominate and our expansions will not be valid. The dominance of $O\left(\frac{1}{n}\right)$ expressions reflects the independence of $\hat{\theta}$ from θ for sufficiently low θ values. If $\theta = 0$ samples within a sampled deme are independent in their allelic state and $\hat{\theta}_{\text{locus}}$ corresponds to a standard F statistic. In fact, this scenario does not require $\theta = 0$. For sampled deme k, the condition $B_k = n$ corresponds to each block containing a single sample and, hence, $\hat{\theta}_{\text{locus}}$ reducing to a standard F statistic. Under our model, $P(B_k = n) \approx \prod_{j=1}^{n-1} \frac{1}{1+\theta j} \approx \exp\left[-\theta \frac{n^2}{2}\right]$. Given n then, once $\theta < \frac{1}{n^2}$, the distribution of $\hat{\theta}_{\text{locus}}$ will be weakly dependent on θ. A typical small value for θ is 0.1 and so one requires sample sizes with $n \gg \sqrt{10}$ for $\hat{\theta}_{\text{locus}}$ to exhibit significant θ dependence. For most data sets $n \gg \sqrt{10}$ is fulfilled.

To clarify the role of diversity, we first specify $\text{MSE}_{\text{locus}}$ for the IAM and BAM models. When $\theta \ll 1$, we find

$$\text{MSE}_{\text{locus}} = \begin{cases} \dfrac{2\theta^2}{d} & \text{under BAM} \\ \dfrac{2\theta^3}{d} & \text{under IAM,} \end{cases} \tag{2.3}$$

and when $\theta \approx 1$ we have,

$$\text{MSE}_{\text{locus}} = \begin{cases} \dfrac{1-\theta}{6dp(1-p)} & \text{under BAM} \\ \dfrac{1-\theta}{3d} & \text{under IAM.} \end{cases} \tag{2.4}$$

The IAM and BAM give minimum and maximum $\text{MSE}_{\text{locus}}$ values respectively over all possible allelic models. More precisely, for $\theta \ll 1$ and $\theta \approx 1$ the values given in (2.2) for $H \in (0, 1]$, fall between the corresponding values for the IAM and BAM given by (2.3) and (2.4). The case $H = 0$ corresponds to the IAM and $\text{MSE}_{\text{locus}}$ evaluates to zero in (2.2) due to the collapse of a first order term in the asymptotic expansion. The next order term in the expansion as seen through (2.3) is non-zero. This relationship between diversity and MSE holds for any value of θ. In Section 6.4, we prove that for any θ, $\text{MSE}_{\text{locus}}$ under the IAM is close to the minimum $\text{MSE}_{\text{locus}}$ over all allele models. Indeed, in our numerical results the $\text{MSE}_{\text{locus}}$ of the IAM is always minimum except in the case of $\theta \approx 1$ where it is very near the minimum. We direct the reader to Section 6.4 for a more precise explanation of this result.

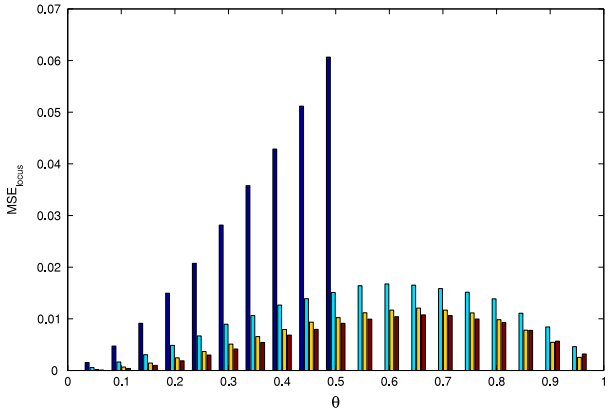

Fig. 1. $\text{MSE}_{\text{locus}}$ is graphed for different allele models. In all cases $d = 5$, $n = 50$ and 10 000 simulations were averaged. For each value of θ, the four bars represent from left to right a locus with the following allelic percentages in the total population: two alleles at 0.8 and 0.2, 5 alleles with equal percentages, 20 alleles with equal percentages, IAM. Results for the biallelic locus for $\theta > 0.5$ are not shown because the population is often homogeneous at the simulated locus.

Fig. 1 shows numerical results for $\text{MSE}_{\text{locus}}$ generated through simulation (C++ code available upon request). To explain our simulation approach, consider the n samples from the kth sampled subpopulation. To form the B_k, b_{kj} we run a coalescent simulation. Starting with n blocks each containing a single sample, we coalesce blocks or label them as 'migrant' blocks. In general, if there are k non-migrant blocks then we coalesce two randomly chosen non-migrant blocks with probability $\frac{\frac{k(k-1)}{2}}{\frac{k(k-1)}{2}+k\Gamma}$, otherwise we randomly select one the k blocks and label it as a migrant block. Migrant blocks are removed from the coalescent simulation. For each subpopulation, eventually we are left with a collection of blocks, only one of which is non-migrant. We take this collection and independently assign each block an allelic state according to the probability distribution specified by the p_u. For each subpopulation, the simulation results in blocks that satisfy the Ewens sampling distribution (Durrett, 2002).

In Fig. 1, four different diversity levels were simulated. Allowing θ to vary and fixing the diversity level, we see the small value of $\text{MSE}_{\text{locus}}$ predicted by (2.2) when $\theta \ll 1$ or $\theta \approx 1$. Between these two extreme ranges of θ, $\text{MSE}_{\text{locus}}$ reaches a maximum somewhere near $\theta = 0.6$. For a fixed value of θ, Fig. 1 shows the decrease in $\text{MSE}_{\text{locus}}$ as one moves from low diversity to high diversity except in the case of $\theta = 0.95$. In that case the $\text{MSE}_{\text{locus}}$ of IAM is not far from the minimum MSE of the four models.

We now consider the cumulants of $\hat{\theta}_{\text{locus}}$. We first present results for the BAM and IAM under $\theta \ll 1$ and $\theta \approx 1$. In Section 6 we obtain explicit formulas for any allele model under $\theta \ll 1$ or $\theta \approx 1$. From these explicit formulas, we can show that for $\theta \ll 1$ and $\theta \approx 1$ the extreme values for the cumulants are those under BAM and IAM.

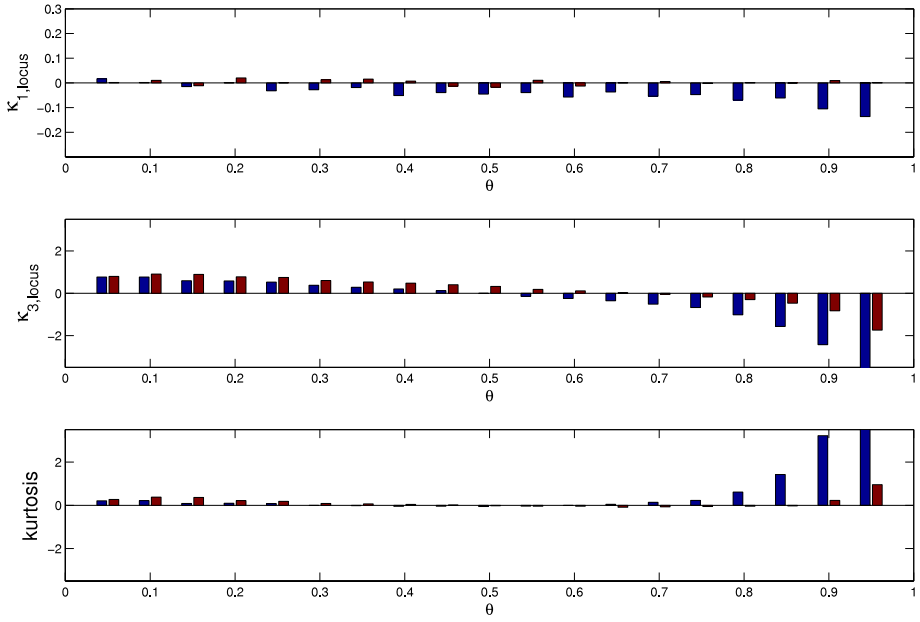

Fig. 2. $\kappa_{1,\text{locus}}$, $\kappa_{3,\text{locus}}$ and kurtosis are graphed for different allele models. In all cases $d = 5$, $n = 50$ and 10 000 simulations were averaged. For each value of θ, the first bar represents four locus model with allele percentages 0.2 0.2 0.3 0.3 and the second bar represents the IAM model. Note that $\kappa_{1,\text{locus}}$ is graphed on a much smaller scale.

We start by considering the $\theta \ll 1$ case,

$$\kappa_{1,\text{locus}} = \begin{cases} \dfrac{-(1-2p)^2}{p(1-p)}\left(\dfrac{\theta}{\sqrt{8d}}\right) & \text{under BAM and } \theta \ll 1 \\ 0 & \text{under IAM and } \theta \ll 1, \end{cases} \tag{2.5}$$

$$\kappa_{3,\text{locus}} = \begin{cases} \sqrt{\dfrac{8}{d}} & \text{under BAM and } \theta \ll 1 \\ (4\sqrt{\theta})\sqrt{\dfrac{8}{d}} & \text{under IAM and } \theta \ll 1. \end{cases} \tag{2.6}$$

For $\theta \ll 1$, (2.6) shows that skewness is bounded. Under the IAM model skewness collapses as θ gets smaller, but since the dependence is $O(\sqrt{\theta})$, for a reasonable θ the IAM and BAM model have the same level of skewness. Bias as measured through $\kappa_{1,\text{locus}}$ is zero in the IAM, but $O(\theta)$ under the BAM. Note that for a BAM with one dominant allele, bias can be quite large even for small θ. Taking these observations together, we see that raising diversity levels reduces bias but does not significantly effect skewness. Although, we do not develop analytic formulas for kurtosis, numerical results to be discussed shortly show that kurtosis does not depend strongly on diversity in the $\theta \ll 1$ regime.

Now we turn to the $\theta \approx 1$ case.

$$\kappa_{1,\text{locus}} = \begin{cases} \dfrac{-(1-2p)^2}{\sqrt{p(1-p)}}\sqrt{\dfrac{3(1-\theta)}{8d}} & \text{under BAM and } \theta \approx 1 \\ 0 & \text{under IAM and } \theta \approx 1, \end{cases} \tag{2.7}$$

$$\kappa_{3,\text{locus}} = \begin{cases} -\dfrac{\sqrt{3}}{5}\sqrt{\dfrac{1}{d(1-\theta)p(1-p)}} & \text{under BAM and } \theta \approx 1 \\ -\dfrac{\sqrt{12}}{5}\sqrt{\dfrac{1}{d(1-\theta)}} & \text{under IAM and } \theta \approx 1. \end{cases} \tag{2.8}$$

In comparison to the regime $\theta \ll 1$, (2.8) demonstrates that skewness is large in the $\theta \approx 1$ regime. Note also that under BAM if one allele is dominant the skewness can grow arbitrarily large. For $\theta \approx 1$ then, diversity does impact skewness. (2.7) demonstrates a large bias when $\theta \approx 1$, at least in low diversity settings, as compared to bias under $\theta \ll 1$. Numerical results show that kurtosis is much larger in the $\theta \approx 1$ setting than $\theta \ll 1$.

Fig. 2 shows numerical results comparing the cumulants and kurtosis of a locus with four alleles to the IAM model. Notice that $\kappa_{3,\text{locus}}$ is quite similar for both loci when $\theta = 0.05$ since $4\sqrt{\theta} \approx 0.9$. When $\theta \approx 1$ the cumulants are much larger than in the $\theta \ll 1$ case. From Fig. 2 we see that skewness and kurtosis are minimized between the two extremes of $\theta \ll 1$ and $\theta \approx 1$ at approximately $\theta = 0.55$.

Finally, we note that in the $\theta \ll 1$ regime we are able to approximate the distribution of θ_{locus} rather than simply the moments. At the end of Section 6.2.2, we show that for all BAM models and for allele models with all alleles having the same frequency, $\hat{\theta}_{\text{locus}}$ is approximately chi-squared distributed, thereby recovering the results of previous authors (Li, 1996). However, as diversity rises, this approximation breaks down.

2.2. Results for $\hat{\theta}_{\text{loci}}$

Now we turn to $\hat{\theta}_{\text{loci}}$. In Section 7, we show that when $\theta \ll 1$ or the loci considered have high diversity, then

$$\theta_{\text{loci}} \approx \sum_{\ell=1}^{L} x_\ell \theta_{\text{locus}}^{(\ell)} \tag{2.9}$$

with

$$x_\ell = \frac{1 - H^{(\ell)}}{\sum_{\ell'=1}^{L}(1 - H^{(\ell')})} \tag{2.10}$$

and where $\theta_{\text{locus}}^{(\ell)}$ and $H^{(\ell)}$ are the single locus estimator and H value associated with locus ℓ. In the case of identical loci (certainly not a realistic situation) (2.9) simplifies to,

$$\theta_{\text{loci}} \approx \frac{1}{L}\sum_{\ell=1}^{L} \theta_{\text{locus}}^{(\ell)}. \tag{2.11}$$

(2.9) and (2.11) reveal several important points. First, if the loci are similar in diversity then MSE for θ_{loci} will be a factor of $\frac{1}{L}$ less

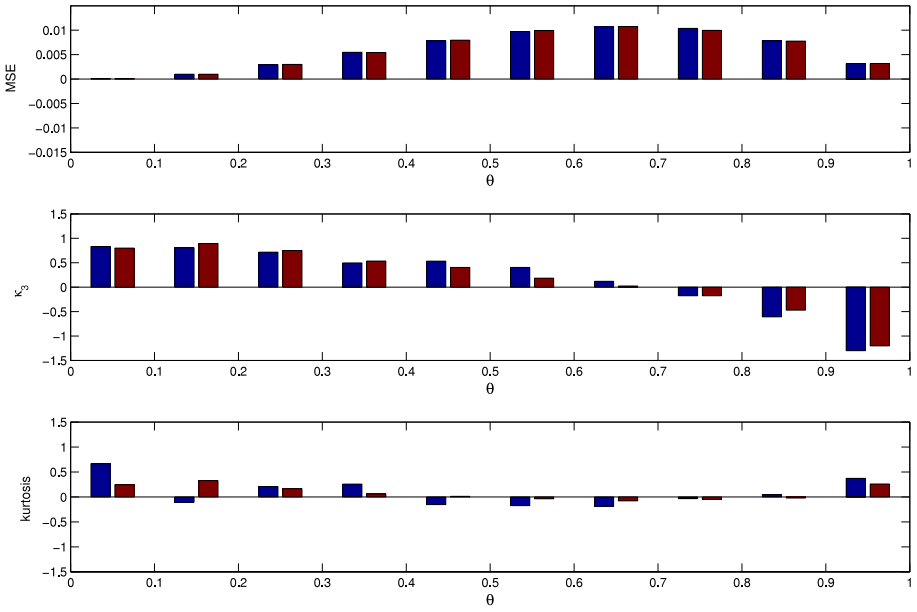

Fig. 3. Cumulants and MSE of $\hat{\theta}_{\text{loci}}$ formed from 20 identical IAM loci are compared to the cumulants and MSE of $\hat{\theta}_{\text{locus}}$ for a single IAM locus. In all cases $d = 5$, $n = 50$ and 10 000 simulations were averaged. Figures from top to bottom compare $L(\text{MSE}_{\text{loci}})$ to $\text{MSE}_{\text{locus}}$; $\sqrt{L}(\kappa_{3,\text{loci}})$ to $\kappa_{3,\text{locus}}$; and $\sqrt{L}$ times the kurtosis of $\hat{\theta}_{\text{loci}}$ to the kurtosis of $\hat{\theta}_{\text{locus}}$ divided by $\sqrt{L}$. $\kappa_{1,\text{loci}}$ and $\kappa_{1,\text{locus}}$ are not graphed, both have values that are negligible compared to the other cumulants. The scaling of kurtosis serves to show kurtosis and skewness on an appropriate scale for assessment of normalcy.

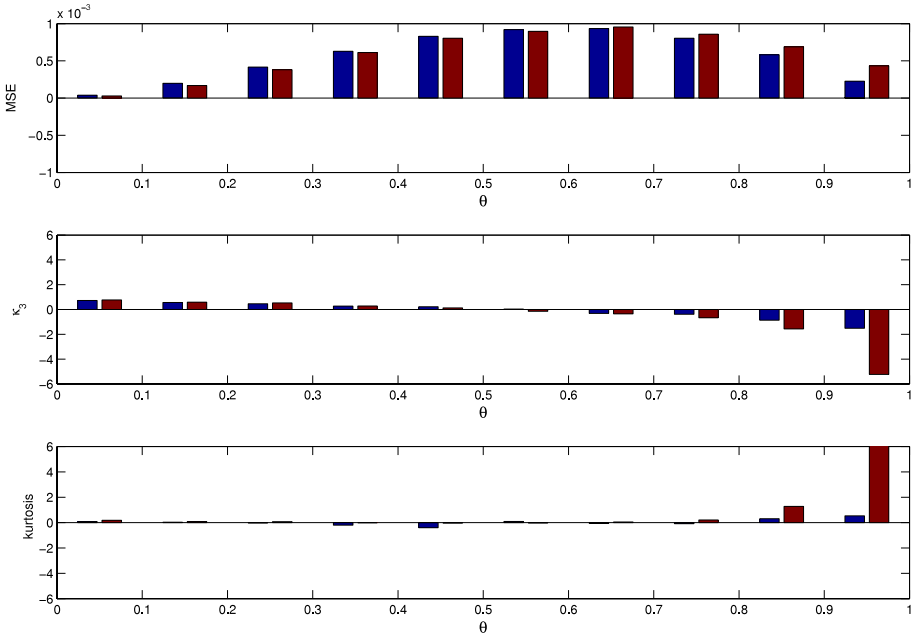

Fig. 4. Cumulants and MSE of $\hat{\theta}_{\text{loci}}$ formed from 20 loci are compared to the cumulants and MSE of $\hat{\theta}_{\text{locus}}$ for a single locus, all loci have 4 alleles with percentages 0.2, 0.2, 0.3, 0.3. See Fig. 3 for all other information.

than that of θ_{locus}. Similarly, $\kappa_{3,\text{loci}}$ will be $\frac{1}{\sqrt{L}}$ less than $\kappa_{3,\text{locus}}$ while kurtosis will see a $\frac{1}{L}$ factor decrease. The situation for $\kappa_{1,\text{loci}}$ is more complex and we direct the reader to Section 7 for a discussion.

In practice, the loci will not have identical diversity. The effect of x_ℓ in (2.9) will be to give diverse loci more weight. In general this will produce a lower MSE_{loci}, (see Section 7 for conditions that make $\hat{\theta}_{\text{loci}}$ the optimal estimator) but in such settings MSE_{loci} will not enjoy a $\frac{1}{L}$ factor reduction with respect to $\text{MSE}_{\text{locus}}$.

Fig. 3 compares $\hat{\theta}_{\text{loci}}$ to $\hat{\theta}_{\text{locus}}$ in the case of 20 identical loci all obeying the IAM model, i.e. a high diversity setting. The figure demonstrates the accuracy of (2.11) over all values of θ. Further, skewness and kurtosis are seen to be quite low. Fig. 4 compares $\hat{\theta}_{\text{loci}}$ to $\hat{\theta}_{\text{locus}}$ in the case of 20 identical with low diversity. As can be seen, the relations predicted by (2.11) hold except when $\theta \approx 1$. These numerical results and others not shown suggest that (2.9) holds for any θ away from $\theta \approx 1$, thereby loosening the assumptions of $\theta \ll 1$ or high diversity loci needed for our analytic justifications of (2.9). In the parameter regimes of Figs. 3 and 4, $\kappa_{1,\text{loci}}$ (not graphed) is of a much lower order than skewness. Finally, Fig. 5 considers a case in which $\hat{\theta}_{\text{loci}}$ is formed by one IAM locus and 19 loci with very low diversity. MSE_{loci} is shown to be of the same order as $\text{MSE}_{\text{locus}}$ rather than $\frac{1}{L}\text{MSE}_{\text{locus}}$, demonstrating the effect of the x_ℓ weighting discussed above.

Assuming (2.9) applies, standard arguments involving Edgeworth expansions show that $\hat{\theta}_{\text{loci}}$ will be close to normal with mean

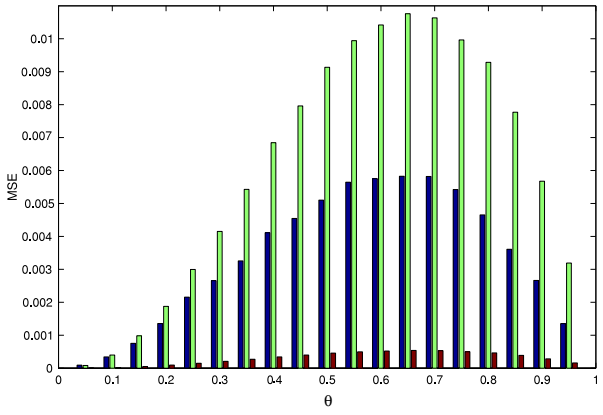

Fig. 5. MSE of $\hat{\theta}_{\text{locus}}$ in mixed loci setting is graphed. 20 loci form $\hat{\theta}_{\text{locus}}$. 1 locus is under IAM and 19 loci have 2 alleles with percentages 0.95, 0.05. For each value of θ, the first bar is the actual value of MSE_{loci}, the second and third bars are $\text{MSE}_{\text{locus}}$ and $\frac{1}{L}\text{MSE}_{\text{locus}}$ for a single IAM locus respectively. See Fig. 3 for all other information.

θ if the bias, skewness and kurtosis of $\hat{\theta}_{\text{locus}}$ are relatively small and L is reasonably large. Our results for $\hat{\theta}_{\text{locus}}$ show that the absolute values of the three cumulants are minimum at intermediate values of θ and maximum at $\theta \approx 1$, with values for $\theta \ll 1$ occupying a middle range. For all values of θ, raising the diversity lowers the absolute value of the cumulants. Putting all this together, we expect $\hat{\theta}_{\text{loci}}$ to be farthest from normal for $\theta \approx 1$ followed by $\theta \ll 1$. Further, the more diverse our loci, the closer to normal $\hat{\theta}_{\text{loci}}$ will become.

3. Application to confidence intervals

Confidence interval (CI) construction methods for $\hat{\theta}_{\text{loci}}$ usually fall into two categories: normal and bootstrapped CI. Normal CI intervals are constructed under the assumption that $\hat{\theta}_{\text{loci}}$ is normal (or at least very close) and requires a variance estimator which we label $\hat{\sigma}_{\text{loci}}$. Once $\hat{\sigma}_{\text{loci}}$ has been constructed, the α-CI is found as,

$$[\hat{\theta}_{\text{loci}} - z_\alpha \hat{\sigma}_{\text{loci}}, \hat{\theta}_{\text{loci}} + z_\alpha \hat{\sigma}_{\text{loci}}] \tag{3.12}$$

where z_α is the familiar 1.96 for the case of $\alpha = 0.95$. $\hat{\sigma}_{\text{loci}}$ is often determined through jackknife methods. Alternatively, our theoretical results in Section 7 suggest a formula for $\hat{\sigma}_{\text{loci}}$, see (7.9).

Bootstrap confidence interval construction can proceed in many ways. We consider three popular approaches: percentile, percentile-t, and BCa bootstrapping (see DiCiccio and Efron, 1996; Efron and Tibshirani, 1986; Hall, 1992 for detailed descriptions). In any of these three methods, the current practice is to bootstrap over the loci. The percentile-t bootstrap requires a variance estimator while the BCa bootstrap requires both a variance and skewness estimator to form the 'acceleration' term and an estimator of $P(\hat{\theta}_{\text{loci}} < 0)$ to form the 'bias correction' term. Typically, jackknifing is used to produce the variance and skewness estimators, however as suggested above our theoretical formula could be used to estimate variance while an analogous formula can be used for the third moment and hence to form skewness, see (7.10) for the third moment formula. We implemented bootstrap methods through our own C++ code (available upon request) and through the *R* boot package. Both implementations gave the same results.

We let I_{loci} represent the confidence interval constructed from the multilocus data used to form $\hat{\theta}_{\text{loci}}$. For a given set of data, the different methods of CI construction; i.e. normal, percentile bootstrap, percentile-t, etc.; produce different I_{loci}. For a given construction method, I_{loci} will vary depending on the allelic data of the L loci. We define the coverage value of I_{loci} by,

$$\text{Coverage Value} = P(\theta \in I_{\text{loci}}), \tag{3.13}$$

Table 1
Coverage values under low diversity levels.

θ	Normal methods		Bootstrap methods		
	Jackknife	Formula	Perc.	Perc.-t	BCa
0.05	0.82 (0.92)	0.79 (0.91)	0.84 (0.91)	0.94 (0.95)	0.81 (0.91)
0.35	0.84 (0.93)	0.78 (0.91)	0.79 (0.92)	0.93 (0.96)	0.79 (0.92)
0.65	0.86 (0.94)	0.80 (0.93)	0.82 (0.93)	0.96 (0.96)	0.81 (0.92)
0.95	0.61 (0.90)	0.60 (0.88)	0.82 (0.90)	0.59 (0.97)	0.68 (0.88)

Table 2
Coverage values under high diversity levels.

θ	Normal methods		Bootstrap methods		
	Jackknife	Formula	Perc.	Perc.-t	BCa
0.05	0.90 (0.93)	0.87 (0.92)	0.90 (0.92)	0.94 (0.94)	0.90 (0.92)
0.35	0.92 (0.94)	0.89 (0.92)	0.89 (0.93)	0.95 (0.95)	0.89 (0.92)
0.65	0.94 (0.94)	0.91 (0.93)	0.88 (0.93)	0.94 (0.96)	0.88 (0.93)
0.95	0.82 (0.93)	0.79 (0.91)	0.85 (0.92)	0.87 (0.97)	0.86 (0.92)

where the probability is taken for a fixed CI construction method. We will consider only 95% CI, so the optimal coverage value is 0.95. For CI constructed through normal approximations, we would expect the coverage error to increase with $\kappa_{1,\text{loci}}$ and $\kappa_{3,\text{loci}}$. Tables 1 and 2 give coverage error results for the normal approximation using the jackknife and formula variance estimators. In the tables, the numbers in parentheses correspond to a model with $d = 5, L = 20$ while the unparenthesized numbers are for $d = 2, L = 10$, both tables use 50 samples from each sampled deme. 10 000 simulations were run and the results averaged to form the coverage values given. Table 1 assumes a BAM model for each locus with percentages 0.8 and 0.2. Table 2 assumes equal allele percentages over twenty alleles for each locus.

Examining Tables 1 and 2 we see, as expected from our cumulant analysis and numerical simulations, that coverage values drop when θ approaches 0 or 1, with a more severe drop for $\theta \approx 1$. Further, we see that coverage is better under high diversity levels, reflecting the drop in the cumulants caused by increased diversity. Increasing d and L lowers the cumulant values, and indeed we see coverage values are quite good for all values of θ in the case $d = 5, L = 20$.

When a statistic is close to normally distributed, the percentile method gives the same coverage values as the normal approximation (Efron and Tibshirani, 1986; Hall, 1992). Except in the case $\theta \approx 1$, the cumulants are relatively small and $\hat{\theta}_{\text{loci}}$ is close to normal, correspondingly for $\theta \neq 0.95$ the percentile and normal coverage values are close. However, when $\hat{\theta}_{\text{loci}}$ is far from normal the percentile method can significantly improve on normal approximations (Efron and Tibshirani, 1986; DiCiccio and Efron, 1996) and for $\theta = 0.95$ this seems to be the case.

When a distribution is close to normal, both percentile-t and BCa bootstrapped should improve on the normal approximation (Hall, 1992). Examining Tables 1 and 2 for $\theta \neq 0.95$ we see that percentile-t significantly improves on the normal construction methods while BCa does not. Numerical simulations shown in Section 6 suggest that both the jackknife and formula estimators for skewness preform quite poorly, especially in the low diversity setting. Poor estimation of skewness leads to errors in the BCa construction method. The variance estimators perform well, so the percentile-t failure when $\theta = 0.95$ and diversity is low may be attributable to $\hat{\theta}_{\text{loci}}$ being very far from normal in this setting.

4. Discussion

Table 3 compares error between $\hat{\theta}_{\text{locus}}$ (single locus) and $\hat{\theta}_{\text{loci}}$ with a 20 loci sample. Error is measured as the square root of MSE. In all cases 50 samples were taken from each of two sampled

Table 3
Estimation error.

L	Allele model	θ			
		0.05	0.35	0.65	0.95
1	IAM	0.01	0.11	0.16	0.09
	INT	0.04	0.2	0.3	–
	BAM	0.06	0.26	0.48	–
20	IAM	0.003	0.03	0.03	0.02
	INT	0.01	0.04	0.05	0.03
	BAM	0.02	0.08	0.1	0.04

subpopulations. We consider three allelic models: IAM, INT corresponding to 4 alleles with percentages 0.2, 0.2, 0.3, 0.3, and BAM with percentages 0.8 and 0.2. For $\hat{\theta}_{\text{loci}}$, all 20 loci have the same allelic distribution. Errors for INT and BAM in the single locus, $\theta = 0.95$ case are not shown because the locus is often homogeneous in that setting. The results demonstrate the importance of using multiple loci for θ estimation. Further, error is seen to substantially increase as we move from high diversity in the IAM to low diversity in the BAM, with INT occupying a middle ground. Good estimation requires many, diverse loci.

Along these same lines, the results of Section 3 show that accurate confidence intervals require a reasonable level of diversity at the loci sampled, especially in cases where we suspect θ to be relatively large. Our results suggest that the BCa bootstrap should not be used to construct confidence intervals, at least until the role and precision of skewness estimators is better understood. The percentile-t performed well in the context of our model. Our results also show that for values of θ in which error is largest, around $\theta = 0.6$, the estimators are closest to normal. This is a fortuitous situation. In essence, when we need confidence intervals the most, they will be the most accurate. Further, for such values of θ a normal approximation for the confidence interval will be quite accurate.

If we care more about sensitivity rather than specificity, the confidence interval estimation issues we have discussed are not as important. Indeed if we want an interval that contains θ with high probability, we can estimate error through either jackknifing over loci or using our variance estimator and then build a corresponding interval that contains θ with as high a probability as we like. If enough diverse loci are used such intervals will be quite narrow.

5. Methods

In this section we explain the methods we use to analyze $\hat{\theta}_{\text{locus}}$. Recall the definition of x_{kju} from Section 1.1. Define $\bar{p}_{k,u}$ to be the sample frequency of allele u in subpopulation k. More precisely, $\bar{p}_{k,u} = \frac{1}{n}\sum_{j=1}^{n} x_{kju}$. Then we further define,

$$\bar{p}_u = \frac{1}{d}\sum_{k=1}^{d} \bar{p}_{k,u}, \tag{5.1}$$

$$\Delta p_{k,u} = \bar{p}_{k,u} - p_u, \tag{5.2}$$

$$\Delta p_u = \frac{1}{d}\sum_{k=1}^{d} \Delta p_{k,u}. \tag{5.3}$$

Note that $\bar{p}_u$ is simply the sample frequency of allele u over all d sampled subpopulations.

Before explaining our methods, we describe the approach employed by WC and Raufaste and Bonhomme (hereafter WC–RB) in deriving the bias and variance of $\hat{\theta}_{\text{locus}}$ (Weir and Cockerham, 1984; Raufaste and Bonhomme, 2000). $\hat{\theta}_{\text{locus}}$ is a function of the $\Delta p_{k,u}$ and so computing the moments of $\hat{\theta}_{\text{locus}}$ can be reduced to computing the moments of $\Delta p_{k,u}$. Indeed, this is the approach used by WC–RB. For clarity, we explain the complexities of computing such moments through the WC–RB approach in the context of an infinite island model with scaled migration rate Γ. Consider $E[\Delta p_{k,u}^2]$.

$$\begin{aligned} E[\Delta p_{k,u}^2] &= E[(x_{kju} - p_u)(x_{kj'u} - p_u)] + O\left(\frac{1}{n}\right) \\ &= E[x_{kju}x_{kju'}] - p_u(E[x_{kju}] \\ &\quad + E[x_{kju'}]) + p_u^2 + O\left(\frac{1}{n}\right). \end{aligned} \tag{5.4}$$

To solve for $E[\Delta p_{k,u}^2]$ we need to find $E[x_{kju}x_{kju'}]$ and $E[x_{kju}]$. Since a single lineage simply samples from the full population we have $E[x_{kju}] = p_u$. We have $x_{kju}x_{kju'} = 1$ in two cases: samples j, j' coalesced prior to migration and their shared line of descent is of allele type u or the two samples did not coalesce prior to a migration and both lines of descent are of allele type u. Then we have,

$$E[x_{kju}x_{kju'}] = \left(\frac{1}{1+\Gamma}\right)p_u + \left(\frac{\Gamma}{1+\Gamma}\right)p_u^2. \tag{5.5}$$

Plugging these results into (5.4) and using $\theta = \frac{1}{1+\Gamma}$ gives

$$E[\Delta p_{k,u}^2] = \theta p_u(1-p_u) + O\left(\frac{1}{n}\right), \tag{5.6}$$

as it must by definition of θ. Now consider computing $E[\Delta p_{k,u}^3]$.

$$\begin{aligned} E[\Delta p_{k,u}^3] &= E[(x_{kju} - p_u)(x_{kj'u} - p_u)(x_{kj''u} - p_u)] + O\left(\frac{1}{n}\right) \\ &= E[x_{kju}x_{kju'}x_{kju''}] - 3p_uE[x_{kju}x_{kju'}] \\ &\quad + 3p_u^2E[x_{kju}] - p_u^3 + O\left(\frac{1}{n}\right). \end{aligned} \tag{5.7}$$

Notice first that a higher power moment requires the summation of more terms. Of all the terms to the right of the equality directly above, we know all but $E[x_{kju}x_{kju'}x_{kju''}]$. In three different cases we have $x_{kju}x_{kju'}x_{kju''} = 1$, compared to two for $x_{kju}x_{kju'} = 1$. The three cases are as follows: samples j, j', j'' coalesce prior to a migration, a pair of j, j', j'' coalesce and then a migration occurs, and two migrations occur before any of j, j', j'' coalesce. Computing the probabilities of these cases gives,

$$\begin{aligned} E[x_{kju}x_{kju'}x_{kj''u}] &= \left(\frac{2}{2+\Gamma}\right)\left(\frac{1}{1+\Gamma}\right)p_u \\ &\quad + \left(\frac{2}{2+\Gamma}\right)\left(\frac{\Gamma}{1+\Gamma}\right)p_u^2 \\ &\quad + \left(\frac{\Gamma}{2+\Gamma}\right)\left(\frac{\Gamma}{1+\Gamma}\right)p_u^3. \end{aligned} \tag{5.8}$$

At this point we hope the reader is convinced that higher moments are complex to compute. To estimate the variance of $\hat{\theta}_{\text{locus}}$ one needs to consider moments up to $E[\Delta p_{k,u}^4]$, a feat accomplished in WC–RB but resulting in complex formulas. To find the skewness of $\hat{\theta}_{\text{locus}}$, we would need $E[\Delta p_{k,u}^6]$ and so correspondingly $E[\prod_{i=1}^{6} x_{kj_iu}]$. The combinatorics of the different migration-coalescent combinations become difficult to enumerate. Indeed, this is the main obstacle in applying the WC–RB methods to analyze the skewness of $\hat{\theta}_{\text{locus}}$.

Now we explain our approach. We use a coalescent, implicitly characterized through the B_k, b_{kj}, formed from all n samples in a given sampled deme. In contrast, the approach of WC–RB, considers a coalescent of m samples if one wants the mth moment of $\Delta p_{k,u}$. Characterizing the distribution of B_k, b_{kj} is a difficult problem, but one that has been solved in several important settings (see Pitman, 2002; Durrett, 2002 and references therein). The

payoff of using B_k, b_{kj} is an increase in our ability to analyze $\Delta p_{k,u}$. Indeed, we can express $\Delta p_{k,u}$ as a function of the B_k, b_{kj}.

Define the random variables A_{kj} as follows,

$$A_{kj}^{(u)} = \begin{cases} 1 & \text{with probability } p_u \\ 0 & \text{otherwise.} \end{cases} \tag{5.9}$$

The $A_{kj}^{(u)}$ are indicator variables for the allelic state of the jth block in sampled deme k. With this in mind we have,

$$\Delta p_{k,u} = \sum_{j=1}^{B_k} (A_{kj}^{(u)} - p_u) b_{kj}. \tag{5.10}$$

Characterizing $\Delta p_{k,u}$ through (5.10) has two advantages. First, rather than simply considering the moments of $\Delta p_{k,u}$ we can also consider its distribution. Second, the form of (5.10) allows us to separate the stochasticity caused by allelic state from that of the coalescent and this greatly simplifies computations by eliminating the combinatorics of the WC–RB approach. To be more specific, consider computing $E[\Delta p_{k,u}^3]$. We have,

$$\begin{aligned} E[\Delta p_{k,u}^3] &= E\left[\left(\sum_{j=1}^{B_k} (A_{kj}^{(u)} - p_u) b_{kj}\right)^3\right] \\ &= \sum_{j=1}^{B_k} E[(A_{kj}^{(u)} - p_u)^3 b_{kj}^3] \\ &= \left(E[(A_{kj}^{(u)} - p_u)^3]\right) E\left[\sum_{j=1}^{B_k} b_{kj}^3\right] \\ &= \left(p_u(1-p_u)(1-2p_u)\right) E\left[\sum_{j=1}^{B_k} b_{kj}^3\right]. \end{aligned} \tag{5.11}$$

The complicated combinatorics of (5.8) are gone, instead we must deal with moments of the b_{kj} such as $E\left[\sum_{j=1}^{B_k} b_{kj}^3\right]$.

Working with the B_k, b_{kj} also provides an intuition for the form of $\hat{\theta}_{\text{locus}}$. For large n, we have

$$\theta = E\left[\sum_{j=1}^{B_k} b_{kj}^2\right]. \tag{5.12}$$

Through (5.12), ideally we would estimate θ using $\hat{\theta}$ given by,

$$\hat{\theta} = \frac{1}{d} \sum_{k=1}^{d} \left(\sum_{j=1}^{B_k} b_{kj}^2\right). \tag{5.13}$$

The trouble with this estimator is that we don't know the block sizes b_{kj} because different blocks may share an allelic type and hence be indistinguishable. But note that if we assume that each block has a different allelic type, exactly the case for the IAM model, then $\hat{\theta}$ is a well defined estimator. Indeed, *under IAM* $\hat{\theta}_{\text{locus}}$ *is precisely* $\hat{\theta}$ *defined in* (5.13). If we are not in the IAM then we must estimate the b_{kj}^2, doing so gives the general form of $\hat{\theta}_{\text{locus}}$.

(5.10) does not apply to IAM because the $A_{kj}^{(u)}$ cannot be properly defined. Indeed, under IAM every $A_{kj}^{(u)} = 0$ with probability 1 because each block has zero probability of being one of infinitely many alleles. Several approaches to analyzing IAM are possible. Perhaps simplest, one could consider finite allele models and take $A \to \infty$. However, as A gets large some of the first order terms in our expansion go to zero and the second order terms dominate. As a result, using finite allele models to understand IAM is technically difficult. Instead, when considering the IAM we use (5.13) as the basis for our computations.

Before proceeding we collect some formulas that will be useful. In all cases we ignore terms of $O\left(\frac{1}{n}\right)$.

$$\begin{aligned} E[\Delta p_u^2] &= \frac{\theta}{d} p_u(1-p_u), \\ E[\Delta p_u \Delta p_{u'}] &= -\frac{\theta}{d} p_u p_{u'}, \\ E[\Delta p_{k,u}^2] &= \theta p_u(1-p_u). \end{aligned} \tag{5.14}$$

Up to now we have not discussed the form of $b_{k,j}$. Indeed, this reflects the flexibility of our approach. However, to actually derive results we must understand the distribution of the $b_{k,j}$. For the case of our model, it is known that b_{kj} have the following form (Donnelly and Tavare, 1986; Pitman, 2002),

$$b_{kj} = W_{kj} \prod_{i=1}^{j-1} (1 - W_{ki}), \tag{5.15}$$

where the W_{kj} are all independent and Beta$(1, 2\Gamma)$ distributed. Defining $\gamma = E\left[\sum_{j=1}^{B_k} b_{kj}^3\right]$ and $\delta = E\left[\sum_{j=1}^{B_k} b_{kj}^4\right]$ straightforward computations using (5.15) give,

$$\gamma = \left(\frac{1}{1+\Gamma}\right)\left(\frac{1}{1+2\Gamma}\right), \tag{5.16}$$

$$\delta = \left(\frac{1}{1+\frac{2}{3}\Gamma}\right)\left(\frac{1}{1+\Gamma}\right)\left(\frac{1}{1+2\Gamma}\right). \tag{5.17}$$

Finally, another approach to considering $\hat{\theta}_{\text{locus}}$ is through analysis of the Dirichlet distribution. Assuming an infinite island model, the Dirichlet distribution gives the distribution of the $p_{k,u}$ and hence can be applied to the analysis of $\hat{\theta}_{\text{locus}}$ (Durrett, 2002). However, in spirit this is equivalent to the indicator function approach of Cockerham as the Dirichlet distribution specifies $p_{k,u}$ without separating coalescent and allelic stochasticity. Our approach is, in our opinion, more intuitive and should generalize better to cases where there is no explicit characterization of the $p_{k,u}$ distribution.

6. Analysis for single locus estimator

This section provides the analysis that underlies the results stated in Section 2. In Section 6.1 we precisely define $\hat{\theta}_{\text{locus}}$. Then in Sections 6.2 and 6.3 we derive asymptotic expansions for the first three moments of $\hat{\theta}_{\text{locus}} - \theta$ in the cases of $\theta \ll 1$ and $\theta \approx 1$. These moment expansions allow us to find $\kappa_{1,\text{locus}}$, $\kappa_{3,\text{locus}}$ and σ_{locus}. Finally in Section 6.4 we analyze the effect of diversity on the distribution of $\hat{\theta}_{\text{locus}}$.

Throughout this section we consider only single locus data and our notation follows that introduced at the start of Section 5.

6.1. Basic formulas

In Weir and Cockerham (1984), WC introduced an estimator $\hat{\theta}_u$ for each allele u associated with a given locus. In its simplest form the WC estimator can be written as,

$$\hat{\theta}_u = \frac{s_u^2}{\bar{p}_u(1-\bar{p}_u)}, \tag{6.1}$$

where

$$s_u^2 = \frac{1}{d-1} \sum_{k=1}^{d} (\bar{p}_{k,u} - \bar{p}_u)^2. \tag{6.2}$$

However, WC corrected for bias due to finite n and d. In our model, since we have assumed a constant sample size from each subpopulation, the WC bias correction takes the form,

$$\hat{\theta}_u = \frac{s_u^2 - \frac{1}{n-1}\left[\bar{p}_u(1-\bar{p}_u) - \frac{d-1}{d}s_u^2\right]}{\bar{p}_u(1-\bar{p}_u) + \frac{s_u^2}{d}}. \tag{6.3}$$

The bias correction above includes the $O\left(\frac{1}{n}\right)$ order bias correction terms derived by WC. We assume that $n \gg d$ and so in all our asymptotic expansions below we will ignore the bracketed term in the numerator of (6.3). However, in our numerical results we use (6.3) exactly as written. From this point on we take (6.3) as our definition of $\hat{\theta}_u$.

One would like to combine the $\hat{\theta}_u$ in some way to form an estimator $\hat{\theta}$. The simplest approach is to consider a linear combination of the $\hat{\theta}_u$ and define,

$$\hat{\theta} = \sum_u \bar{w}_u \hat{\theta}_u. \tag{6.4}$$

Ideally, $\bar{w}_u$ should be chosen to minimize variance and $\sum_u \bar{w}_u = 1$ to maintain a low bias. WC derived a formula for the optimal $\bar{w}_u$, but this formula depends on unknown parameters that are difficult to estimate. WC suggested that for θ relatively large, a good (but not optimal) choice is $\bar{w}_u = \frac{\bar{p}_u(1-\bar{p}_u)}{\sum_{u'}\bar{p}_{u'}(1-\bar{p}_{u'})}$. We define $\hat{\theta}_{\text{locus}}$ as $\hat{\theta}$ for this particular choice of $\bar{w}_u$. Other options for the $\bar{w}_u$ exist. The Robertson and Hill estimator, $\hat{\theta}_{\text{RH}}$, corresponds to the choice $\bar{w}_u = \frac{1-\bar{p}_u}{A-1}$ and is suggested for use when θ is small (Robertson and Hill, 1984).

In this section we develop a Taylor expansion for $\hat{\theta}_{\text{locus}}$ in powers of $O\left(\frac{1}{\sqrt{d}}\right)$. We assume a finite allele model, recalling that we analyze IAM using (5.13). Since we are interested in samples with d small this is seemingly troublesome, however in the following sections we show this expansion is appropriate for the cases $\theta \ll 1$ and $\theta \approx 1$. To start, notice first that since samples from separate subpopulations are independent by the central limit theorem (CLT), $\Delta p_u = O\left(\frac{1}{\sqrt{d}}\right)$. To understand the order of s_u^2, we first express s_u^2 as follows,

$$s_u^2 = \frac{d}{d-1}\left(\theta p_u(1-p_u) - \Delta p_u^2\right) + \Delta s_u \tag{6.5}$$

where

$$\Delta s_u = \frac{1}{d-1}\sum_{k=1}^{d}(\Delta p_{k,u}^2 - \theta p_u(1-p_u)). \tag{6.6}$$

It will be useful to define,

$$\Delta s_{k,u} = \Delta p_{k,u}^2 - \theta p_u(1-p_u). \tag{6.7}$$

From (5.14) we have $E[\Delta p_{k,u}^2] = \theta p_u(1-p_u)$ and so by the CLT $\Delta s_u = O\left(\frac{1}{\sqrt{d}}\right)$. By replacing $\bar{p}_u$ by $p_u + \Delta p_u$ and expanding (6.4) in Taylor series we can arrive at the following expansion,

$$\begin{aligned}\hat{\theta}_{\text{locus}} - \theta = &\left[\sum_u \frac{\Delta s_u}{1-H} + 2\theta\sum_u p_u\frac{\Delta p_u}{1-H}\right] \\ &- \left[\frac{2}{(1-H)^2}\left(\sum_u p_u\Delta p_u\right)\left(\sum_{u'}\Delta s_{u'}\right)\right. \\ &+ \left(\frac{1-\theta}{1-H}\right)\sum_u \Delta p_u^2 + \frac{4\theta}{(1-H)^2}\left(\sum_u p_u\Delta p_u\right)^2 \\ &\left.+ \frac{(\theta^2-\theta)}{d}\right] + O\left(\frac{1}{d^{3/2}}\right).\end{aligned} \tag{6.8}$$

The terms in the first and second brackets of (6.8) are $O\left(\frac{1}{\sqrt{d}}\right)$ and $O\left(\frac{1}{d}\right)$ respectively. The expansion for $\hat{\theta}_{\text{locus}}$ in (6.8) allows us to demonstrate the asymptotic normality of $\sqrt{d}(\theta_{\text{locus}} - \theta)$. Indeed,

$$\begin{aligned}&\lim_{d\to\infty}\sqrt{d}(\hat{\theta}_{\text{locus}} - \theta) \\ &= \lim_{d\to\infty}\left[\frac{\sqrt{d}\sum_u \Delta s_u}{1-H} + 2\theta\frac{\sqrt{d}\sum_u p_u\Delta p_u}{1-H}\right],\end{aligned} \tag{6.9}$$

and our previous comment that the CLT applies to $\sqrt{d}\Delta s_u$ and $\sqrt{d}\Delta p_u$ demonstrates the asymptotic normality of $\sqrt{d}(\theta_{\text{locus}} - \theta)$.

Using (6.8) to consider the first through third moments of $\hat{\theta}_{\text{locus}} - \theta$ leads to the following first order approximations,

$$E[\hat{\theta}_{\text{locus}} - \theta] \approx \frac{2}{d(1-H)^2}\sum_u p_u Z_u, \tag{6.10}$$

$$E[(\hat{\theta}_{\text{locus}} - \theta)^2] \approx \frac{1}{d(1-H)^2}Y_2, \tag{6.11}$$

$$\begin{aligned}E[(\hat{\theta}_{\text{locus}} - \theta)^3] = &\frac{1}{d^2(1-H)^3}\left(Y_3 - \left(\frac{18Y_2}{1-H}\right)\right. \\ &\left.\times\sum_u p_u Z_u - 6(1-\theta)\sum_u Z_u^2\right)\end{aligned} \tag{6.12}$$

where

$$Y_\ell = E\left[\left(\sum_u(\Delta s_{k,u} + 2\theta p_u\Delta p_{k,u})\right)^\ell\right], \tag{6.13}$$

$$Z_u = E\left[\left(\sum_{u'}(\Delta s_{k,u'} + 2\theta p_{u'}\Delta p_{k,u'})\right)\cdot\Delta p_u\right].$$

We emphasize that (6.10) and (6.11) are equivalent to formulas found in Weir and Cockerham (1984) and Raufaste and Bonhomme (2000). However, those authors expand around the mean values of the numerator and denominator, i.e. they use the Delta method, while we are expanding in $\Delta p_{k,u}$.

6.2. Small θ

In this section we take $\theta \ll 1$ or equivalently, by (5.16), $\Gamma \gg 1$. In Section 6.2.1, we show that as θ becomes small $\Delta p_{k,u}$ approaches a normal distribution. We then use this fact in Section 6.2.2 to derive moments and cumulants for $\hat{\theta}_{\text{locus}} - \theta$. Further, we provide a chi-squared approximation for the distribution of $\hat{\theta}_{\text{locus}}$, recovering a result noted in Li (1996) and Samanta et al. (2009). Many authors have considered θ estimators assuming that $\Delta p_{k,u}$ is normally distributed and this is often contrasted against the assumption of a Dirichlet distribution (see Weir and Hill, 2002 and references therein). As our results show, the two assumptions overlap for small θ.

6.2.1. Distribution of $\Delta p_{k,u}$

We want to show that $\sqrt{\frac{1}{\theta}}\Delta p_{k,u}$ is approximately normally distributed when θ is small or equivalently, from (5.16), when Γ is large. Recall the form of b_{kj} in (5.15). Roughly, normality occurs for θ small because the each W_{ki} is small and so $\prod_{i=1}^{j-1}(1-W_{ki}) \approx 1$. In turn, $b_{kj} \approx W_{kj}$ and our blocks become independent and identically distributed (i.i.d.). More precisely, we argue

$$b_{kj} = W_{kj} \exp\left[\sum_{i=1}^{j-1} \log(1 - W_{ki})\right]$$
$$= W_{kj} \exp\left[-(j-1)E[W] - \sum_{i=1}^{j-1} \Delta W_{ki} + O\left(\frac{1}{\Gamma^2}\right)\right], \quad (6.14)$$

where $\Delta W_{ki} = W_{ki} - E[W_{ki}]$. Applying the CLT gives $\sum_{\ell=1}^{j-1} \Delta W = O\left(\frac{\sqrt{j}}{\Gamma}\right)$ and evaluating $E[W_{kj}]$ leads to,

$$\Delta p_{k,u} = \sum_j (A_j^u - p_u) W_j \exp\left[-\frac{(j)}{2\Gamma}\right]\left(1 + O\left(\frac{\sqrt{j}}{\Gamma}\right)\right)$$
$$= \left(\sum_j (A_j^u - p_u) W_j \exp\left[-\frac{(j)}{2\Gamma}\right]\right)\left(1 + O\left(\frac{1}{\sqrt{\Gamma}}\right)\right). \quad (6.15)$$

(6.15) expresses $\Delta p_{k,u}$ as a sum of independent random variables. Note however that due to the expression $\exp\left[-\frac{(j)}{\Gamma}\right]$, the variables are not identical. We may, however, apply a generalized CTL argument (see for example Durrett (2000)) to arrive at,

$$\sqrt{\frac{1}{\theta}}\Delta p_{k,u} \approx \mathcal{N}(0, p_u(1 - p_u)). \quad (6.16)$$

To find the moments of $\hat{\theta}_{\text{locus}}$, we will also require the covariance of the $\Delta p_{k,u}$ over u. For $u \neq u'$,

$$E[\Delta p_{k,u} \Delta p_{k,u'}] = -\theta p_u p_{u'}. \quad (6.17)$$

The precision of (6.16) has three important caveats. First, as seen in (6.15), $\frac{1}{\sqrt{\Gamma}}$ or equivalently $\sqrt{\theta}$ must be small. Second, our expression for block size, (5.15), is accurate only if the number of blocks is small relative to the sample size for which we require $\frac{\log(n)}{n\theta} \ll 1$ (Hoppe, 1984). Finally, high levels of diversity will destroy the normal approximation. To see this, note that under the IAM, only a single block can be of type u and so $\Delta p_{k,u}$ cannot possibly be normally distributed. Under low levels of diversity we can consider only the first two contributions. We then find an error which is at least $O\left(\left(\frac{\log(n)}{n}\right)^{\frac{1}{3}}\right)$. To arrive at errors under 10% requires $n \approx 10000$, an unreasonable sample size. For $n = 50$, the value we use in our simulations, we find errors of approximately 20%.

6.2.2. Moments

From (6.10), we can determine the first moment, $E[\hat{\theta}_{\text{locus}} - \theta]$. Using the methods of Section 5, we can find

$$Z_u = \frac{2}{d(1-H)^2}(\theta^2 - \gamma)p_u(p_u - H), \quad (6.18)$$

which gives $\sum_u p_u Z_u = \frac{2}{d}(\theta^2 - \gamma)\left(\frac{I - H^2}{(1-H)^2}\right)$ (this formula is equivalent to (21) in Raufaste and Bonhomme (2000)). Referencing (5.16) allows us to find the order of $\theta^2 - \gamma$ and we have the following approximation for bias,

$$E[\theta_{\text{locus}} - \theta] \approx \frac{-\theta^2}{d}\left(\frac{I - H^2}{(1-H)^2}\right). \quad (6.19)$$

Now we consider $E[(\hat{\theta}_{\text{locus}} - \theta)^2]$. One approach would be to use the expression in (6.11) without modification, we would square the binomial inside the expectation given by Y_2 and then take the expected value of the four resultant terms. While this is not too difficult, indeed this is what is done by WC, when $\theta \ll 1$ a simplification occurs. Using (5.14) and similar computations, we have,

$$E[(\theta \Delta p_u)^2] = O\left(\frac{\theta^3}{d}\right), \quad (6.20)$$

$$E[\Delta s_u^2] = O\left(\frac{\theta^2}{d}\right). \quad (6.21)$$

So $\theta \Delta p_u \ll \Delta s_u$. We can then simplify (6.11) by writing,

$$\hat{\theta}_{\text{locus}} - \theta \approx \sum_u \frac{\Delta s_u}{1 - H} \quad (6.22)$$

and then approximate

$$E[(\hat{\theta}_{\text{locus}} - \theta)^2] \approx \frac{1}{d}\left(\frac{1}{(1-H)^2}\right) E\left[\left(\sum_u \Delta s_{k,u}\right)^2\right]. \quad (6.23)$$

Recalling $\Delta s_{k,u} = (\Delta p_{k,u})^2 - \theta p_u(1 - p_u)$ and plugging in our asymptotic result gives $\Delta p_{k,u} = \sqrt{\theta p_u(1-p_u)}\mathcal{N}(0,1)$. Letting $\mathcal{N}_{k,u}$ be standard normals that are independent over k and for which $E[\mathcal{N}_{k,u}\mathcal{N}_{k,u'}] = -\sqrt{\frac{p_u p_u'}{(1-p_u)(1-p_{u'})}}$ we have,

$$E\left[\left(\sum_u \Delta s_{k,u}\right)^2\right]$$
$$\approx \theta^2 E\left[\left(\sum_u p_u(1-p_u)(\mathcal{N}_{k,u}^2 - 1)\right)^2\right]. \quad (6.24)$$

A standard multivariate normal computation gives,

$$E[(\hat{\theta}_{\text{locus}} - \theta)^2] \approx \frac{2\theta^2}{d}\left(\frac{H - 2I + H^2}{(1-H)^2}\right). \quad (6.25)$$

As an aside we can also use the above methods to consider the variance of $\hat{\theta}_{\text{RH}}$,

$$E[(\hat{\theta}_{\text{RH}} - \theta)^2] \approx \frac{2\theta^2}{(A-1)d}. \quad (6.26)$$

As WC points out, in the $\theta \ll 1$ setting, $\hat{\theta}_{\text{RH}}$ is a better estimator than θ_{locus}. However, while $E[(\hat{\theta}_{\text{locus}} - \theta)^2] > E[(\hat{\theta}_{\text{RH}} - \theta)^2]$, the above formulas show that $\frac{E[(\hat{\theta}_{\text{locus}} - \theta)^2]}{E[(\hat{\theta}_{\text{locus}} - \theta)^2]} \to 1$ as $H \to 0$. So in the case of high diversity, $\hat{\theta}_{\text{locus}}$ is close to optimal.

Now we turn to the third moment using (6.12). As we did in the case of the first moment, we can show that terms in (6.12) involving Z_u are $O\left(\frac{\theta^4}{d^2}\right)$ which leads to,

$$E[(\hat{\theta}_{\text{locus}} - \hat{\theta})^3] \approx E\left[\left(\sum_u \frac{\Delta s_u}{1 - H}\right)^3\right]. \quad (6.27)$$

By the same arguments that gave us (6.25) we find

$$E[(\hat{\theta}_{\text{locus}} - \hat{\theta})^3] \approx \frac{8\theta^3}{d^2}\frac{(I - 3J + 3HJ - H^3)}{(1-H)^3}, \quad (6.28)$$

where $J = \sum_u p_u^4$.

Using (6.22) and the expression for $\Delta s_{k,u}$ given directly above (6.24) we can approximate the distribution of $\hat{\theta}_{\text{locus}}$. We find,

$$\hat{\theta}_{\text{locus}} \approx \frac{\theta}{1-H}\frac{1}{d}\sum_{k=1}^{d}\sum_{u=1}^{A} p_u(1-p_u)\mathcal{N}_{k,u}^2. \quad (6.29)$$

(6.29) shows that $\hat{\theta}_{\text{locus}}$ is well approximated by summing the squares of normals, but for the general allelic model the normals are multiplied by unequal factors. A simplification occurs in the case of equal allelic states, i.e. $p_u = \frac{1}{A}$ for all u. In this setting the covariance matrix of the $\mathcal{N}_{k,u}$ over u for fixed k has entries of 1 along the diagonal and $\frac{-1}{A-1}$ for all other entries. This covariance matrix has one eigenvalue equal to 0 and $A-1$ eigenvalues equal to $\frac{A}{A-1}$. Since $p_u(1-p_u)$ in (6.29) is now constant, we are considering the sum of $\mathcal{N}_{k,u}^2$. Standard multivariate normal arguments then give,

$$\hat{\theta}_{\text{locus}} \approx \frac{\theta}{d}\sum_{k=1}^{d}\frac{1}{A}\left(\frac{A}{A-1}\chi^2_{A-1}(k)\right) = \theta\frac{\chi^2_{d(A-1)}}{d(A-1)} \tag{6.30}$$

where $\chi^2_{A-1}(k)$ is the kth i.i.d. version of a chi-squared random variable with $A-1$ degrees of freedom. The case $A=2$ corresponds to the BAM, in this setting (6.30) holds for any allelic model, not just the symmetric case $p_1 = p_2 = \frac{1}{2}$.

As mentioned at the end of Section 6.2.1, the normal approximation for $\Delta p_{k,u}$ breaks down under high diversity. Correspondingly, our comments regarding the distribution of $\hat{\theta}_{\text{locus}}$ do not apply in cases of high diversity, most notably under IAM.

6.3. The $\theta \approx 1$ case

The development of this section parallels that of Section 6.2, except that here we take $\theta \approx 1$. As we shall show, when $\theta \approx 1$, $\Delta p_{k,u}$ is not normally distributed.

6.3.1. Distribution of $\Delta p_{k,u}$

Algebraic rearrangement of (5.10) gives,

$$\Delta p_{k,u} = (A^{(u)}_{k1} - p_u) + \sum_{i=1}^{\infty}(A^{(u)}_{k(i+1)} - A^{(u)}_{ki})\left(\prod_{i'=1}^{i}(1-W_{ki'})\right). \tag{6.31}$$

When $\theta \approx 1$ we have $(1-W_{ki}) = O(1-\theta)$. Since the W_{ki} are i.i.d., $\prod_{i'=1}^{i}(1-W_{ki'})$ is $O((1-\theta)^i)$ and so from (6.31) we have,

$$\Delta p_{k,u} = (A^{(u)}_{k1} - p_u) + (A^{(u)}_{k2} - A^{(u)}_{k1})(1-W_{k1}) + O((1-\theta)^2). \tag{6.32}$$

An important element to the computations below is the relation $E[(1-W_{ki})^h] = \frac{2\Gamma}{h} + O((1-\theta)^2)$ which reflects the heavy tails of the W_{ki}.

6.3.2. Moments

For the first moment, using the same arguments as in Section 6.2.2, we find

$$E[\hat{\theta}_{\text{locus}} - \theta] \approx \frac{-(1-\theta)}{d}\left(\frac{I-H^2}{(1-H)^2}\right). \tag{6.33}$$

Now we consider $E[(\hat{\theta}_{\text{locus}} - \theta)^2]$. As in the previous section, we start with the expression found in (6.11). However, in contrast to Section 6.2.2, here we need to consider the full expression $\sum_u \Delta s_u + 2\theta\sum_u p_u\Delta p_u$. Let v_{kj} equal the allelic type of the jth sample from the kth sampled subpopulation and define q_{kj} by $q_{kj} = p_{v_{kj}}$. Some algebraic manipulation leads to the following formulas,

$$\sum_u \Delta s_u + 2\theta\sum_u p_u\Delta p_u\frac{1}{d-1}\sum_{k=1}^{d}\Big\{-2\delta(v_{k1}\neq v_{k2})((1-W_{k1}) - (1-W_{k1})^2) + (1-\theta)(H+1-q_{k1}) + O((1-\theta)^2)\Big\}. \tag{6.34}$$

Computing the second moment is now straightforward,

$$\begin{aligned} E[(\hat{\theta}_{\text{locus}} - \theta)^2] &\approx \frac{4}{(1-H)^2}\frac{1}{d}E[\delta(v_{k1}\neq v_{k2})] \\ &\quad\times E[\left((1-W_{k1}) - (1-W_{k1})^2\right)^2] \\ &= \frac{1-\theta}{3d(1-H)}. \end{aligned} \tag{6.35}$$

For the third moment, we start by examining (6.12) and dropping all terms that are lower order in $(1-\theta)$. As in Section 6.2 we find that all terms involving Z_u are lower order which leads to

$$E[(\hat{\theta}_{\text{locus}} - \theta)^3] \approx \frac{1}{(1-H)^3}E\left[\left(\sum_u \Delta s_u + 2\theta\sum_u p_u\Delta p_u\right)^3\right]. \tag{6.36}$$

Evaluating (6.36) is not as arduous as it might seem. The third moment of $\sum_u \Delta s_u + 2\theta\sum_u p_u$ will be $O(1-\theta)$. This means that we may drop all terms involving $1-\theta$ in the expansion of $\sum_u \Delta s_u + 2\theta\sum_u p_u$ and simplify as follows,

$$\begin{aligned} &E\left[\left(\sum_u \Delta s_u + 2\theta\sum_u p_u\right)^3\right] \\ &\approx E\left[\left(-2\delta(v_{k1}\neq v_{k2})\left((1-W_{k1}) - (1-W_{k1})^2\right)\right)^3\right] \\ &= 4E[\delta(v_{k1}\neq v_{k2})]E\left[\left((1-W_{k1}) - (1-W_{k1})^2\right)^3\right]. \end{aligned} \tag{6.37}$$

The expression directly above is easy to evaluate and we can arrive at,

$$E[(\hat{\theta}_{\text{locus}} - \theta)^3] \approx \frac{-2(1-\theta)}{15d^2(1-H)^2}. \tag{6.38}$$

6.4. Diversity

The formulas in Sections 6.2 and 6.3 depend in simple ways on θ but in complex ways on diversity. In this section, we show that for any fixed θ (we are not restricted to $\theta \ll 1$ or $\theta \approx 1$) the MSE under the IAM is at most $O\left(\frac{1}{d^2}\right)$ greater than the minimum MSE under all models. A CLT argument shows MSE to be $O\left(\frac{1}{d}\right)$, so the result becomes more meaningful as d grows larger. Since the IAM corresponds to maximum diversity, this result shows that high diversity leads to lowered MSE.

Intuitively, higher diversity leads to lower MSE because we are able to more accurately estimate the blocks, i.e. the b_{kj}, in our sample. For the IAM we can exactly estimate the blocks, but for a low diversity model such as the BAM, many blocks will share the same allelic state and hence be indistinguishable. Our numerical results suggest that MSE is minimized under the IAM model for even small d, i.e. $d = 2$ or $d = 5$, except in the case of large θ. We speculate that in this setting, high diversity raises the error associated in approximating p_u by $\bar{p}_u$ more than it lowers the error of block estimation. Indeed, Δp_u is precisely the error of p_u estimation and we have $E\left[\sum_u(\Delta p_u)^2\right] = \frac{\theta}{d}(1-H)$, showing that p_u estimation error is substantial when diversity and θ are both large.

The MSE can be decomposed into allelic and coalescent associated variance. To do this, let $\mathcal{B}$ be the set of all possible choices for the B_k, b_{kj} over all k. Given $\mathbf{b} \in \mathcal{B}$ we let $\mathcal{M}_{\mathbf{b}}$ be the set of all possible allelic assignments to the blocks specified by $\mathbf{b}$. Then we define a probability space Ω such that $\omega \in \Omega$ can be

identified by $\omega = (\mathbf{b}, \mathbf{m})$ for some $\mathbf{b} \in \mathcal{B}$ and $\mathbf{m} \in \mathcal{M}_{\mathbf{b}}$. Each ω has an associated probability and $\hat{\theta}_{\text{locus}}$ is a r.v. on Ω. Define,

$$\hat{\theta}_{\text{locus}}(\mathbf{b}, \mathbf{m}) - \theta = \phi_{\text{mut}}(\mathbf{b}, \mathbf{m}) + \phi_{\text{coal}}(\mathbf{b}, \mathbf{m}) \tag{6.39}$$

where,

$$\phi_{\text{mut}}(\mathbf{b}, \mathbf{m}) = \hat{\theta}_{\text{locus}}(\mathbf{b}, \mathbf{m}) - E[\hat{\theta}_{\text{locus}} \mid \mathbf{b}], \tag{6.40}$$

$$\phi_{\text{coal}}(\mathbf{b}, \mathbf{m}) = E[\hat{\theta}_{\text{locus}} \mid \mathbf{b}] - \theta. \tag{6.41}$$

Intuitively $E[\theta_{\text{locus}} \mid \mathbf{b}]$ is formed by choosing a set of blocks, $\mathbf{b}$, and then taking the expectation over all possible allelic choices for the blocks. By standard properties of conditional expectation, we have

$$E[\phi_{\text{coal}}\phi_{\text{mut}}] = E[E[\phi_{\text{mut}} \mid \mathbf{b}]\phi_{\text{coal}}] = 0, \tag{6.42}$$

and then,

$$E[(\hat{\theta}_{\text{locus}} - \theta)^2] = E[\phi^2_{\text{mut}}] + E[\phi^2_{\text{coal}}]. \tag{6.43}$$

To understand ϕ_{mut} and ϕ_{coal} we return to (6.8). Ignoring higher order terms and using $E[\Delta p_u \mid \mathbf{b}] = 0$ leads to,

$$\phi_{\text{mut}} = \frac{1}{1-H}\sum_u (\Delta s_u + 2\theta p_u \Delta p_u - E[\Delta s_u \mid \mathbf{b}]) + O\left(\frac{1}{d}\right), \tag{6.44}$$

$$\phi_{\text{coal}} = \frac{1}{1-H}\sum_u E[\Delta s_u \mid \mathbf{b}] - \theta + O\left(\frac{1}{d}\right). \tag{6.45}$$

Using (5.14) we can find,

$$\phi_{\text{coal}} = \frac{1}{d}\sum_{k=1}^{d}\sum_j b^2_{kj} - \theta + O\left(\frac{1}{d}\right). \tag{6.46}$$

Further, we can rewrite (6.44) as,

$$\phi_{\text{mut}} = \frac{1}{d-1}\sum_{k=1}^{d}\left(\frac{1}{1-H}\sum_u \Delta s_{k,u} + 2\theta p_u \Delta p_{k,u} - E[\Delta s_{k,u} \mid \mathbf{b}]\right) + O\left(\frac{1}{d}\right). \tag{6.47}$$

Under varying H, $E[\phi^2_{\text{coal}}]$ will remain constant up to $O\left(\frac{1}{d^2}\right)$, while $E[\phi^2_{\text{mut}}]$ will vary with H on $O\left(\frac{1}{d}\right)$. If we can show that $E[\phi^2_{\text{mut}}] = 0$ under the IAM we will have proved our claim. Under the IAM we can form $\hat{\theta}_{\text{locus}}$ simply from our knowledge of the b_{kj}. That is, given a $\mathbf{b}$, $\hat{\theta}_{\text{locus}}$ is determined. This gives $\hat{\theta}_{\text{locus}} = E[\hat{\theta}_{\text{locus}} \mid \mathbf{b}]$ and so $\phi_{\text{mut}} = 0$ under the IAM.

7. Analysis for multiple loci estimator

Our analysis for $\hat{\theta}_{\text{loci}}$ essentially follows from our results for $\hat{\theta}_{\text{locus}}$. In Section 7.1, we express $\hat{\theta}_{\text{loci}}$ in terms of the $\hat{\theta}_{\text{locus}}$, one for each locus sampled. Then in Section 7.2, we find the moments and cumulants of $\hat{\theta}_{\text{loci}}$ in terms of the moments of the $\hat{\theta}_{\text{locus}}$.

To simplify notation, define $\hat{\theta}^{(\ell)}_{\text{locus}}$ to be the $\hat{\theta}_{\text{locus}}$ estimator associated with locus $\ell \in [1, 2, \ldots, L]$. Throughout this section, we use the (ℓ) tag to specify that the variable or parameter being considered corresponds to the ℓth locus. For example, we write $H^{(\ell)}$ for H corresponding to locus ℓ.

7.1. Basic formulas

In the case of a single locus estimator WC linearly combined the $\hat{\theta}_u$ to form $\hat{\theta}_{\text{locus}}$. Here again, in the multiple loci case, one can linearly combine the $\hat{\theta}^{(\ell)}_{\text{locus}}$ to form a multiple loci estimator $\hat{\theta}$. For a collection of weights $\bar{x}_\ell$ summing to 1 we define,

$$\hat{\theta} = \sum_{\ell=1}^{L} \bar{x}_\ell \hat{\theta}^{(\ell)}_{\text{locus}}. \tag{7.1}$$

To minimize variance one should define $\bar{x}_\ell \sim \frac{1}{V[\hat{\theta}^{(\ell)}_{\text{locus}}]}$. As we have seen in Section 6, $V[\hat{\theta}^{(\ell)}_{\text{locus}}]$ will depend on θ and diversity levels in a complex way. For $\theta \ll 1$ we should define $\bar{x}_\ell \sim \frac{(1-H^{(\ell)})^2}{H^{(\ell)}-2I^{(\ell)}+(H^{(\ell)})^2}$ while for $\theta \approx 1$ we have $x_\ell \sim (1 - H^{(\ell)})$.

WC suggested the use of a $\hat{\theta}$ corresponding to the choice $\bar{x}_\ell = \left(1 - \bar{H}^{(\ell)} + \sum_u \frac{(s^{(\ell)}_u)^2}{d}\right)$ where $\bar{H}$ is the estimator of H, i.e. $\bar{H}^{(\ell)} = \sum_u (\bar{p}^{(\ell)}_u)^2$. (WC did not make the choice of $\bar{x}_\ell$ explicit, rather they characterized this estimator as the ratio of the sums of the numerators to the sums of the denominators found over all the $\hat{\theta}^{(\ell)}_{\text{locus}}$.) We label this estimator $\hat{\theta}_{\text{loci}}$ and note that it is approximately optimal in the case $\theta \approx 1$.

We define $x_\ell = \frac{1-H^{(\ell)}}{\sum_{\ell'}(1-H^{(\ell')})}$ and set $\Delta x_\ell = \bar{x}_\ell - x_\ell$. Notice that x_ℓ is deterministic. We can write,

$$\hat{\theta}_{\text{loci}} - \theta = \sum_\ell (x_\ell + \Delta x_\ell)(\theta^{(\ell)}_{\text{locus}} - \theta). \tag{7.2}$$

If we had $\Delta x_\ell = 0$, then the moments of $\hat{\theta}_{\text{loci}}$ would be simple functions of the moments of θ_{locus}. Since this is not the case, computing the moments of $\hat{\theta}_{\text{loci}}$ is not straightforward. However, as we now show, for $\theta \ll 1$ or in the case of high diversity, we have $\frac{\Delta x_\ell}{x_\ell} \ll 1$ and Δx_ℓ can be ignored. Letting $S = \sum_{\ell'}(1 - H^{(\ell')})$, a Taylor expansion gives the following approximation

$$\Delta x_\ell = -\frac{\Omega_\ell}{S} + \frac{(1-H_\ell)\sum_{\ell'}\Omega_{\ell'}}{S^2} \tag{7.3}$$

where,

$$\Omega_\ell = 2\sum_u p_{u,\ell}\Delta p_{u,\ell}. \tag{7.4}$$

Using (5.14) we find $E[\Omega^2_\ell] = \frac{\theta}{d}(I^{(\ell)} - J^{(\ell)})$ and then $E[(\sum_{\ell'}\Omega_{\ell'})^2] = \sum_{\ell'}\frac{\theta}{d}(I^{(\ell')} - J^{(\ell')})$ since the loci are unlinked. Plugging these two relations into (7.3) leads to

$$\Delta x_\ell \approx O\left(\frac{1}{S}\sqrt{\frac{\theta}{d}(I^{(\ell)} - J^{(\ell)})}\right) + O\left(\frac{1}{S^2}\sum_{\ell'}\sqrt{\frac{\theta}{d}(I^{(\ell')} - J^{(\ell')})}\right). \tag{7.5}$$

The terms to the right of the equality directly above are small relative to x_ℓ if $\theta \ll 1$ or there is a high level of diversity at each locus.

7.2. Moments and cumulants

In this section, in order to compute moments and cumulants we assume that Δx_ℓ can be ignored. In other words, we are considering situations of high diversity or $\theta \ll 1$. Having made this assumption, the MSE and cumulants of $\hat{\theta}_{\text{loci}}$ follow easily from our results for $\hat{\theta}_{\text{locus}}$. Rather than give the general formulas, we state the

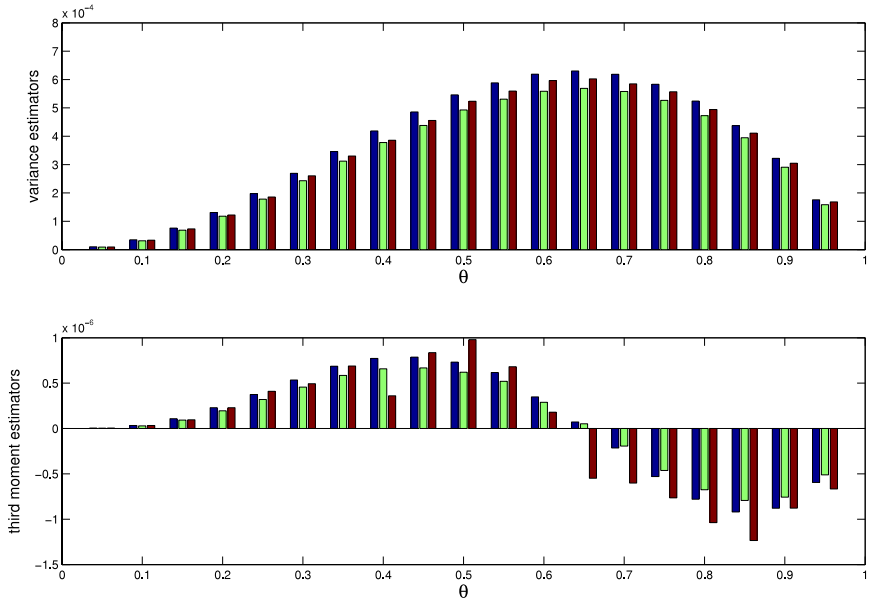

Fig. 6. Graph comparing estimators defined in (7.9) and (7.10) to actual variance and the third moment in the case of loci with high diversity allele models. In top graph for each θ the first bar is the jackknife variance estimator, the second bar is the variance formula estimator, (7.9), and the third bar is the actual variance. In bottom graph for each θ the first bar is the jackknife third moment estimator, the second bar is the third moment formula estimator, (7.10), and the third bar is the actual third moment. $\hat{\theta}_{\text{loci}}$ is formed from 20 loci each with 20 alleles of equal percentages. $d = 5$, $n = 50$. 10 000 simulations were run and the results averaged.

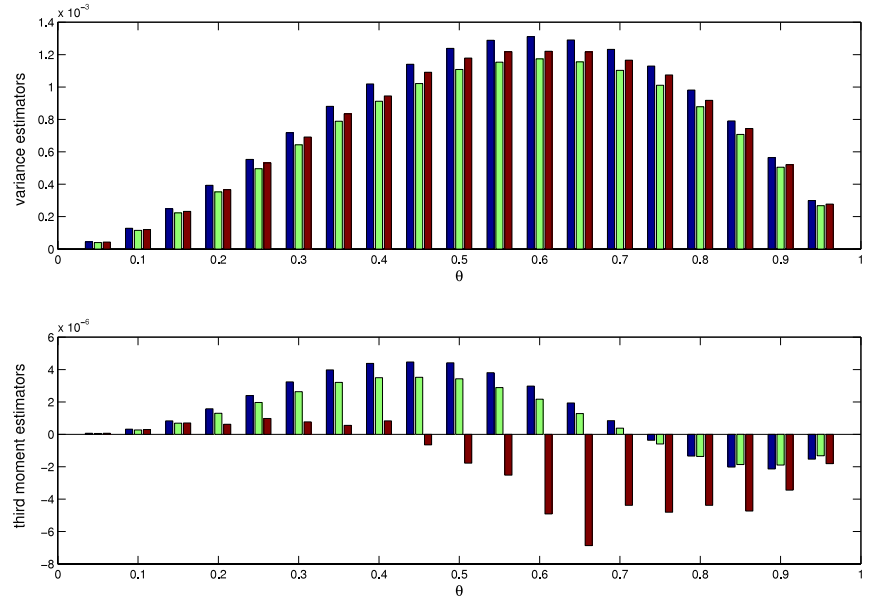

Fig. 7. Graph comparing estimators defined in (7.9) and (7.10) to actual variance and the third moment in the case of loci with low diversity allele models. $\hat{\theta}_{\text{loci}}$ is formed from 20 loci each with 5 alleles of percentages 0.6, 0.1, 0.1, 0.1, 0.1. All other information is the same as Fig. 6.

formulas in the case where all loci have the same number of alleles and allelic percentages. While this is not realistic, it does reveal the impact of moving from $\hat{\theta}_{\text{locus}}$ to $\hat{\theta}_{\text{loci}}$.

$$\text{MSE}_{\text{loci}} \approx \frac{1}{L}\sigma^2_{\text{locus}}\left(1 + L\kappa^2_{1,\text{locus}}\right) \tag{7.6}$$

$$\kappa_{1,\text{loci}} \approx \sqrt{L}(\kappa_{1,\text{locus}}), \tag{7.7}$$

$$\kappa_{3,\text{loci}} \approx \frac{\kappa_{3,\text{locus}}}{\sqrt{L}}. \tag{7.8}$$

Notice that as L becomes large, the MSE_{loci} does not collapse to zero as $L \to \infty$ due to bias. But recall that the above formulas only hold for $\theta \ll 1$ or high diversity loci. In both these cases $\kappa_{1,\text{locus}}$ is very small and so for reasonable values of L, say less than 100, $\kappa_{1,\text{loci}}$ will be quite small.

The discussion in Section 7.1 also leads to estimates for the variance and skewness of $\hat{\theta}_{\text{loci}}$. Indeed, we can suggest as a variance estimator:

$$\hat{\sigma}^2_{\text{loci}} = \sum_{i=1}^{L} \bar{x}_i^2 \left(\hat{\theta}^{(i)}_{\text{locus}} - \hat{\theta}_{\text{loci}}\right)^2, \tag{7.9}$$

and as a third moment estimator,

$$\sum_{i=1}^{L} \bar{x}_i^3 \left(\hat{\theta}^{(i)}_{\text{locus}} - \hat{\theta}_{\text{loci}}\right)^3. \tag{7.10}$$

Figs. 6 and 7 suggest that the variance estimator (7.9) and a jackknife variance estimator are accurate both for low and high

diversity settings, but the third moment estimators (7.10) and a jackknife third moment estimator both fail in the low diversity setting.

Acknowledgments

We would like to thank two anonymous reviewers whose comments significantly improved this paper. We also thank F. Rousset for pointing out serious flaws in an earlier version of this work.

References

DiCiccio, T., Efron, B., 1996. Bootstrapping confidence intervals. Statist. Sci. 11 (3), 189–212.
Donnelly, P., Tavare, S., 1986. The ages of alleles and a coalescent. Adv. Appl. Probab. 18, 1–18.
Durrett, R., 2000. Probability: Theory and Examples. Cambridge University Press.
Durrett, R., 2002. Probability Models for DNA Sequence Evolution. Springer.
Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statist. Sci. 1 (1), 54–77.
Hall, P., 1992. The Bootstrap and Edgeworth Expansion. In: Springer Series in Statistics.
Hoppe, F., 1984. Polya-like Urns and the Ewens' sampling formula. J. Math. Biol. 20, 91–94.
Kingman, J., 1978. Random partitions in population genetics. Proc. R. Soc. Lond. Ser. A 361 (1704), 1–20.
Leviyang, S., 2010. The distribution of F_{st} for the Island model in the large population, weak mutation limit. Stoch. Anal. Appl. 28 (4), 577–601.
Leviyang, S., The distribution of F_{st} and other genetic statistics for a class of population structure models, (in press and currently available online through J. Math. Biol.).
Li, Y.-J., Characterizing the structure of genetic populations. Ph.D. Thesis. North Carolina State Univ. 1996.
Nei, M., et al., 1977. Mean and variance of F_{st} in a finite number of incompletely isolated populations. Theor. Popul. Biol. 11, 291–306.
Pitman, J., 2002. Combinatorial Stochastic Processes. In: St. Flour XXXII Lecture Notes, Springer.
Raufaste, N., Bonhomme, F., 2000. Properties of bias and variance of two multiallelic esimtators of F_{st}. Theor. Popul. Biol. 57, 285–296.
Robertson, A., Hill, W., 1984. Deviation from Hardy–Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics 107, 703–718.
Rottenstreich, S., et al., 2007. Steady state of homozygosity and G_{st} for the Island model. Theor. Popul. Biol. 72, 231–244.
Rousset, F., 1996. Equilibrium values of meausures of population subdivision for stepwise mutation processes. Genetics 142, 1357–1362.
Samanta, S., et al., 2009. Drawing inferences about the coancestry coefficient. Theor. Popul. Biol. 75, 312–319.
Slatkin, M., 1991. Inbreeding coefficients and coalescence times. Genet. Res. Camb. 58, 167–175.
Weir, B., 1996. Genetic Data Analysis, vol. II. Sinauer Associates.
Weir, B., Cockerham, C., 1984. Estimating F statistics for the analysis of population structure. Evolution 38 (6), 1358–1370.
Weir, B., Hill, W., 2002. Estimating F-statistics. Annu. Rev. Genet. 36, 721–750.
Weir, B., et al., 2005. Measure of human population structure show heterogeneity among genomic regions. Genome Res. 15, 1468–1476.
Wilkinson-Herbots, H., 1998. Genealogy and subpopulation differentiation under various models of population structure. J. Math. Biol. 37, 535–585.
Wright, S., 1931. The genetical structure of populations. Ann. Eugenics 15, 323–354.