Pseudoreplication in genomics-scale datasets

Robin S. Waples[*], Ryan K. Waples[†], and Eric J. Ward[*]

[*]NOAA Fisheries, Northwest Fisheries Science Center
2725 Montlake Blvd. East, Seattle, WA 98112

[†]Department of Biology, Section for Computational and RNA Biology,
University of Copenhagen, Copenhagen, Denmark
Current address: Department of Biostatistics
University of Washington, Seattle WA

Corresponding author:
Robin S. Waples
robin.waples@noaa.gov
+1 206 860-3254

November 2020

Abstract

In genomics-scale datasets, loci are closely packed within chromosomes and hence provide correlated information. Averaging across loci as if they were independent creates pseudoreplication, which reduces the effective degrees of freedom (n') compared to the nominal degrees of freedom, n. This issue has been known for some time, but consequences have not been systematically quantified across the entire genome. Here we measured pseudoreplication (quantified by the ratio n'/n) for a common metric of genetic differentiation (F_{ST}) and a common measure of linkage disequilibrium between pairs of loci (r^2). Based on data simulated using models (*SLiM* and *msprime*) that allow efficient forward-in-time and coalescent simulations while precisely controlling population pedigrees, we estimated n' and n'/n by measuring the rate of decline in variance of mean F_{ST} and mean r^2 as more loci were used. For both indices, n' increases with N_e and genome size, as expected. However, even for large N_e and large genomes, n' for r^2 plateaus after a few thousand loci, and a variance components analysis indicates that the limiting factor is uncertainty associated with sampling individuals rather than genes. Pseudoreplication is less extreme for F_{ST}, but $n'/n \leq 0.01$ can occur in datasets using tens of thousands of loci. Commonly-used block-jackknife methods consistently overestimated var(F_{ST}), producing very conservative confidence intervals. Predicting n' based on our modeling results as a function of N_e, L, S, and genome size provides a robust way to quantify precision associated with genomics-scale datasets.

Keywords: effective degrees of freedom; linkage disequilibrium; F_{ST}; N_e; genome size; jackknife variance; simulations

INTRODUCTION

It is now relatively easy to generate data at tens or hundreds of thousands of single-nucleotide polymorphism (SNP) markers for non-model species, even in the absence of a reference genome (da Fonseca et al. 2016; Van Wyngaarden et al. 2017; Agguire et al. 2019; Choquet et al. 2019; Minias et al. 2019). This opens up vast new opportunities for researchers but also creates a host of analytical challenges, including ascertainment bias (Rosenblum and Novembre 2007; Albrechtsen et al. 2010), phylogenetic inference (Leaché et al. 2015), genotyping errors and missing data (Gautier et al. 2013; Graffelman et al. 2015; Huang and Knowles 2014), and effects of selection (Foll and Gaggiotti 2008; Wolf and Ellegren 2017).

One topic that has not received sufficient attention with respect to genomics data for non-model species is pseudoreplication—lack of independence that reduces the total information content. Statistical inference in biology is challenging because biological systems are complex, variable, and subject to measurement and sampling errors. Replication is generally necessary to ensure that apparently-interesting results are not due to small sample sizes and chance. But true replication (which produces multiple independent datapoints) is difficult to achieve, and flaws of experimental design and/or statistical inference that lead to pseudoreplication and overly-optimistic estimates of statistical significance have been found to be widespread in ecology (Hurlbert 1984; Ramage et al. 2013; Colegrave and Ruxton 2018) and biomedical studies (Aarts et al. 2014; Lin et al. 2019).

In the present context, we are interested in use of genetic data to draw inferences about key genetic parameters in real populations of non-model species. For decades, most studies were limited to a few dozen genetic markers, and it was routinely assumed that each marker was independent—or if not, that departures from independence were small enough to be safely ignored. This assumption is no longer tenable in the age of genomics. Most species have at most a few dozen chromosomes (Table 1), so in large genomics datasets a typical chromosome contains thousands of markers—a situation which guarantees that pseudoreplication occurs.

Here we focus on two widely-used genetic metrics: F_{ST}, a measure of differentiation among populations; and r^2, a measure of LD at pairs of loci. Inter-locus sampling variances for both of these metrics are large, so a high degree of replication is required to obtain reliable estimates of the mean. Let $\hat{F}_{ST(n1)}$ be the estimate of mean F_{ST} based on data for n_1 diallelic loci, and let $\hat{r}^2{}_{(n2)}$ be the two-locus analogue averaged over n_2 pairs of loci, where the '^' indicates an estimate. In statistical theory, if one has n independent observations of a random variable x with standard deviation σ, the standard error of the estimate of mean(x) is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. The genetic metrics of interest here also have the property that their variances are proportional to $1/n$, assuming the data points are independent (Lewontin and Krakauer 1973; Hill 1981).

In population genetic studies, two general sources of replication are available: one can sample multiple individuals from each population, and one can sample multiple genes from each individual. Often the number of individuals that can be sampled is limited by population size or logistical constraints, in which case the only feasible method to increase precision is by increasing the number of loci. Easy access to genomics-scale datasets has made it possible to vastly increase the number of genes sampled per individual—but to what extent do these genes provide independent information about population-level parameters of interest?

We encounter two kinds of pseudoreplication we need to be concerned with. First, if loci do not freely recombine, information they provide is correlated, so (for example) var($\hat{F}_{ST(n1)}$) does not decline as fast with addition of more loci as would be the case if the loci were

independent. The total information content of a dataset is constrained by the amount of recombination. This constraint means, for example, that even with many millions of SNPs available in human genetic datasets, it is not possible to confidently resolve distant familial relationships with genetic data alone (Thompson 2013). Pseudoreplication due to lack of independent assortment applies both to $\hat{F}_{ST}$ and $\hat{r}^2$. Analyses of LD also generate a second kind of pseudoreplication, caused by overlapping pairs of loci. For L loci, there are $L(L\text{-}1)/2 \approx L^2/2$ different pairs, but the pairs are not independent because each locus occurs in L-1 pairwise comparisons.

Problems related to this lack of independence have not been quantified in any systematic way for non-model species. For such species, a typical experimental design involves orders of magnitude more SNP loci than individuals, and if any detailed genomic mapping information is available, loci typically can only be placed on short genomic scaffolds rather than full chromosomes. In this study we investigate how much precision is reduced by pseudoreplication, compared to what it would be if the assumption of independence were completely satisfied. Assuming one has data for L diallelic (SNP) loci, the number of datapoints (and the nominal degrees of freedom associated with the overall estimate) is $n_1 = L$ for F_{ST} and $n_2 \approx L^2/2$ for the LD analyses. [Hereafter, for simplicity we drop the subscript for n, while recognizing that it takes different values for the two metrics.] Pseudoreplication causes the actual (effective) degrees of freedom (n') to be less than n. The key question thus can be framed as follows: How much smaller is n' than n, and how does this depend on user-specified aspects of experimental design (number of individuals (S) and loci sampled) and uncontrollable parameters of the population(s) of interest (genome size or number of chromosomes, *Chr*, and effective population size, N_e)? If one knew how to estimate n', that would allow an unbiased evaluation of precision and facilitate placing accurate confidence bounds on estimates of key population genetic parameters.

For species (including humans) for which detailed linkage information is available, these dependencies have been addressed to some extent, particularly with respect to multiple testing when the tests are not independent (Nyholt 2004; Pe'er et al., 2008; Galwey 2009). A popular approach is to use a weighted, block-jackknife method that breaks up the genome into blocks of contiguous loci and sequentially leaves each block out in jackknife fashion (Busing et al. 1999). LD pruning (Purcell et al. 2007)—excluding loci to reduce LD—can also be used to generate sets of loci that act more independently and thus and reduce prereplication. These approaches, however, require detailed mapping information, and even so they only deal with correlations of the original variables. To evaluate the second type of pseudoreplication noted above, it is necessary to consider second-order correlations—that is, the degree to which r^2 (locus 1 × locus 2) is correlated with r^2 (locus 1 × locus 3) and r^2 (locus 2 × locus 3). In theory n' could be calculated this way, but it would require one to specify the relevant covariance matrix, and with $\approx L^2/2$ pairwise correlations of loci the covariance matrix of these correlations has $\approx L^4/8$ elements. For a dataset with 1.5 million SNPs, calculating n' therefore would require one to specify over 6.3×10^{23} elements in the covariance matrix—more than Avogadro's number! Not surprisingly, we are not aware of any attempts to do this.

The objectives of this paper are to develop model-based approximations of effective degrees of freedom, and provide general guidance – given known or measurable covariates (*Chr*, L, S, N_e). Our approach to quantifying the degree of pseudoreplication involves simulating a large number of replicate populations, and for each replicate we calculate mean values for both of the genetic indices ($\hat{r}^2$, $\hat{F}_{ST}$). Observed variances of the multilocus metrics across replicates

allow us to calculate the effective degrees of freedom and the ratio n'/n associated with samples of individuals and gene loci. This process was repeated for several evolutionary scenarios involving different combinations of *Chr* and N_e.

We have the following general expectations:

1) As more loci are packed into a fixed number of chromosomes, the number of loci per chromosome increases, minimum distance between a new locus and existing loci shrinks (Figure S1), and the marginal increase in information content provided by each new locus declines. Therefore, we expect that as the ratio L/Chr increases, pseudoreplication will increase and the ratio n'/n will decrease. In addition, the rate of decay of LD with distance between loci increases with effective population size (Figure S2). So, we expect that, for a given L/Chr ratio, n' will be smaller for populations with smaller N_e. These effects, however, have not been quantified across the entire genome.
2) It seems intuitive that the magnitude of LD increases with the number of loci, but that is not actually the case. Each new locus increases opportunities for that locus to be in LD with existing loci, but this is balanced by new pairwise comparisons with loci on different chromosomes. As a result, the probability that any two randomly-chosen loci will be in any particular LD association is independent of the number of loci, but it does depend on genome size and N_e (Waples et al. 2016). Therefore, we expect that the first type of pseudoreplication for LD analyses will be inversely correlated with both N_e and *Chr*. The second type of pseudoreplication arising from LD analyses—that caused by overlapping pairs of the same loci—has received little study. However, we note that, of the $\sim L^2/2$ pairwise comparisons of L loci, only $L/2$ are completely independent (non-overlapping), so the proportion of independent comparisons is $1/L$. Hence, we expect that this type of pseudoreplication will increase with the number of loci.

We hope to find one or both of the following:

a. Patterns in empirical variances across simulated datasets that allow us to provide general guidance for users interested in predicting n'/n, based on measurable or estimable covariates (*Chr*, L, S, N_e).
b. Existing jackknife methods prove to be reliable at estimating precision for a given dataset.

METHODS

Below we provide an overview of methods used; for more details, see Supplemental Information. Table 2 summarizes notation.

Conceptual framework

Our focus is on actual populations, each of which has a single, realized population pedigree (Wakeley et al. 2012; Ralph 2019). To mimic sampling of individuals and genes from real populations, we combine efficient coalescent and forward simulation programs (*SLiM*, Messer 2013; and *msprime*, Kelleher et al. 2016) that allow control over multigenerational pedigrees in replicate populations (Haller et al. 2019). Our experimental design explicitly models two major sources of uncertainty in estimating population-level parameters: sampling of individuals and sampling of genes. Both processes are important when evaluating uncertainty around parameter estimates. If replication only occurs across genes, estimates will converge on

values determined by the pedigree of the observed individuals (Waples and Faulkner 2009; Wakeley et al. 2012; King et al. 2018). In general, however, one wants to draw inference about the entire population, and uncertainty related to sampling individuals from the population cannot be eliminated by intensive replication across genes.

Modeled scenarios and simulations

For 16 different evolutionary scenarios (combinations of effective size [N_e=50,200,800,3200] and number of chromosomes [*Chr*=1,4,16,64]), we simulated four separate ancestral populations of N_e=N diploid individuals for 10N_e generations under Wright-Fisher (WF) reproduction and ensured that each ancestral population fully coalesced. Each ancestral population then split into four daughter populations of size N_e, which evolved in isolation for enough generations that expected F_{ST}≈0.05-0.1, common values for natural populations of many species. From this point, the models differed slightly for analyses of F_{ST} and LD.

In the LD analyses, for each of the 4x4=16 population pedigrees, we created eight non-overlapping sets of loci by adding mutations to gene trees present in the population pedigree (Figure 1). Gene trees reflect the evolutionary history of genes and haplotypes back in time. For each scenario, the 4x4x8 hierarchical design produced 128 replicate populations with up to 75K diallelic loci (up to 100K loci for N_e=3200). From each replicate population, we took four random subsamples of S=25,50,100 diploid individuals (only S≤50 for N_e=50), and data for each subsample were analyzed for L=100-75K loci. Except for N_e=3200, we also took exhaustive samples of the population (S=N_e). The experimental design allowed us to calculate two separate variance components: V_1 (same individuals, different loci) and V_2 (same loci, different but potentially overlapping samples of individuals) (Table 3). Data from cells on the diagonal (different individuals and different loci) were used to calculate the effective degrees of freedom, n', as described below.

In the F_{ST} analyses, for each of the four ancestral populations, the four daughter populations shown in Figure1 allowed 6 pairwise comparisons among populations, for a total of 24 two-population pedigrees (Figure S3). Six mutational replicates of 200K loci were created for each pedigree, and eight samples for each of several different sizes were taken from each mutational replicate.

The resulting distributions of minor allele frequency (MAF) followed the familiar U-shaped pattern expected at mutation-drift equilibrium (Figure S4). In addition to the *SLiM*/*msprime* simulations that organize the genome into chromosomes, we also modeled scenarios with unlinked loci and infinite N_e to provide insights into asymptotic behavior.

Genetic indices

For each replicate in each evolutionary scenario, we calculated $\hat{F}_{ST(n)}$ and $\hat{r}^2{}_{(n)}$ across variable numbers of loci or locus pairs. LD analyses were conducted using both all pairs of loci, and only pairs on different chromosomes. For each pair, the sample estimate of the squared correlation coefficient ($\hat{r}^2$) was computed using the Pearson product-moment correlation of alleles at different loci. After applying the sample-size adjustments implemented in *LDNE* (Waples and Do 2008), mean $\hat{r}^2$ can be used to estimate N_e. Because this method assumes all loci are unlinked, use of all locus pairs leads to a predictable pattern of downward bias in $\widehat{N_e}$ that is inversely proportional to log(*Chr*) (Waples et al. 2016), whereas $\widehat{N_e}$ is unbiased when

computed only using pairs of loci on different chromosomes (Figure S5). To reduce effects of rare alleles, within each sample we omitted loci with two or fewer copies of the minor allele.

We used two methods for estimating the standardized variance of allele frequency among populations, F_{ST}. Nei's (1973) gene diversity method calculates $G_{ST} = (H_T - H_s)/H_T$, where H_s is the expected within-population heterozygosity (averaged across both populations) and H_T is expected total heterozygosity, based on mean allele frequencies across both populations, without an adjustment for sample size. Calculated this way, and with just two alleles at each locus, G_{ST} is identical to Wright's (1951) F_{ST} (Nei and Chakravarti 1977); we refer to this estimator as $\hat{F}_{ST}^{Nei}$.

Another widely-used measure of population differentiation is the coancestry coefficient, θ (Cockerham 1969; Reynolds et al. 1983); the relationship between the two parameters is given by (Cockerham and Weir 1987)

$$\theta = \frac{sG_{ST}}{G_{ST}+s-1}, \qquad (1)$$

where s is the number of subpopulations. For s=2 (as considered here) this reduces to $\theta = 2G_{ST}/(G_{ST}+1)$, which is close to $2G_{ST}$ if G_{ST} is small. We considered Hudson's (1992) estimator $\hat{F}_{ST}^{Hudson} = 1 - H_W/H_B$, where H_W and H_B are the mean numbers of allelic differences within and between populations, respectively; the formula we used was Equation 10 in Bhatia et al. (2013). Under conditions modeled here (two populations, equal sample sizes), $\hat{F}_{ST}^{Hudson}$ is identical to Weir and Cockerham's (1984) $\hat{\theta}$ or nearly so (Bhatia et al. 2013).

Because performance of multilocus F statistics is generally better when they are computed as ratios of multi-locus means rather than means of single-locus ratios (Jorde and Ryman 2007; Bhatia et al. 2013), we evaluated both methods for calculating $\hat{F}_{ST}^{Nei}$ ($\hat{F}_{ST}^{Hudson}$ already is a ratio of averages). We considered three ascertainment schemes for identifying variable loci: 1) loci variable in at least one of the two populations; 2) loci variable in the ancestral population; 3) loci with overall MAF≥0.05 in the two samples combined.

Effective degrees of freedom

We compared the nominal degrees of freedom (df, n) with the estimated effective degrees of freedom (n'). For analysis of population differentiation, n' was calculated from observed variances of $\hat{F}_{ST(n)}$ using the relationship (Lewontin and Krakauer 1973) $\phi = Var(\hat{F}_{ST})/E^2(\hat{F}_{ST})$ = squared coefficient of variation of $\hat{F}_{ST}$. If $\hat{F}_{ST(n)}$ is based on n independent loci, the following relationship should hold: $\phi = 2/n$. Therefore, we computed the effective degrees of freedom from the empirical variance of $\hat{F}_{ST(n)}$ as $n' = 2/\phi$. n' can be interpreted as the number of independent data points that would be expected to produce the value of ϕ observed in the data (Cox 1984; Giesbrecht 2006). The ratio n'/n therefore provides an index of how much nonindependence has increased variance of the estimator. Hill (1981) showed that the relationship $\phi = 2/n$ also applies to $\hat{r}^2{}_{(n)}$, under the assumption that all pairwise comparisons of loci are independent.

Many commonly-used estimators of F_{ST} and related quantities (Weir and Cockerham 1984; Hudson et al. 1992; Patterson et al. 2012) include adjustments for sampling individuals and/or populations, and this complicates comparisons with theoretical expectations of the variance-to-mean ratio that depend on raw F_{ST} values. This is why, for estimating n', we used Nei's (1973) gene diversity method, which does not include a sample-size adjustment. For Hudson's estimator, we measured the rate of decline in the variance of $\hat{F}_{ST(n)}^{Hudson}$ as more loci were used in the analysis, and this provided an alternative way to quantify the degree of pseudoreplication.

Model fitting

We considered a wide range of covariates (N_e, *Chr*, *S*, *L*, and *n*) as predictors of *n*'. To account for potential non-linearities, we also considered transformed responses of these original variables (Table S1). Because of asymptotic relationships between *n*' and the predictor variables (Figure S9), we focused inference on fitting statistical models with asymptotes, rather than using linear models. The general function we used to parameterize the models is

$$f(x) = \frac{x}{\left[\frac{1}{p^r}+\frac{x^r}{q^r}\right]^{\frac{1}{r}}}\ , \qquad (2)$$

where parameters *p*, *q*, and *r* control the shape of the function. When shape parameter *r*=1, this function takes the form of the familiar Michaelis–Menten equation (also known in fisheries as the Beverton–Holt stock recruitment model; Beverton and Holt 1957). We considered models with both covariates and transformed functions of covariates; these latter models treated each parameter (*p*, *q*, *r*) as linear functions of other predictors; e.g., $p = b\mathbf{X}$, where **X** is a design matrix of predictors and *b* are estimated coefficients. All models were fit in R (R Core Team 2020) using maximum likelihood. We used Akaike's Information Criterion (AIC) to evaluate which combinations of parameters were best supported. Additional details about model fitting are in Supplemental Information.

Confidence intervals

To evaluate accuracy of our estimates of *n*', for selected scenarios we generated many (>1000) new samples of individuals and loci, and for each sample we calculated $\hat{F}_{ST(n)}$ or $\hat{r}^2{}_{(n)}$ averaged over data for 500–5000 SNPs (for LD) or 5000–50000 SNPs (for F_{ST}). We calculated confidence intervals (CIs) around these means using standard statistical theory (Equations S4 and S5). Width of the CIs was calculated two ways, using: 1) our modeled estimates of *n*'; 2) a published jackknife method (Busing et al. 1999 for F_{ST}, Jones et al. 2016 for $\widehat{N_e}$ based on r^2).

Data availability

This manuscript did not generate new empirical data. Simulation scripts will be made available on Github following acceptance of the manuscript.

RESULTS

Linkage disequilibrium

The four different ancestral populations produced modest differences in mean r^2 when averaged across all descendant populations, and within each daughter population the eight mutational replicates also produced relatively small differences in mean r^2 (Figure S6). In contrast, daughter populations descended from the same ancestral population varied substantially in the mean magnitude of LD. This result emphasizes the importance of accounting for recent population pedigrees in assessing variability of genetic indices and shows the sensitivity of LD-based estimates of N_e to recent effective population size (Waples and Faulkner 2009).

Effective degrees of freedom: Table S2 gives our estimates of n' for r^2 for every scenario we simulated. We expected substantial effects of genome size and N_e on n', and those expectations were borne out (Figure 2). For a given number of loci, n' increased smoothly as the number of chromosomes increased from 1 to 64, with results for $Chr = 64$ being largely indistinguishable from those for unlinked loci (Figure 2, top). The influence of effective size was even stronger: n' increased systematically as N_e increased from 50 to 3200 (Figure 2, bottom); however, for large numbers of loci, n' for N_e=3200 was still substantially lower than the $n' \approx n$ that was found for $N_e=\infty$. With 5000 loci, var(r^2) decreased by a factor of 5× when N_e increased four-fold from 50 to 200, and the variance decreased by a factor of 68× when N_e increased 64-fold to 3200 (using data from Table S2 with S=25 and Chr=4). In contrast, quadrupling the number of chromosomes (from 1 to 4, with N_e fixed at 50) decreased var(r^2) only by 2.5×, and a 64-fold increase in Chr reduced var(r^2) only by 10×.

Relatively speaking, sample size of individuals has less influence on n': with Chr=16, a four-fold increase in S reduced var(r^2) by only 1.9× for N_e=200 and by only 6% for N_e=800. There was one notable exception, however: we found a qualitative difference in n' between scenarios in which the entire population was sampled and those in which only a subset of individuals was analyzed (Figure 3). With exhaustive sampling ($S=N=N_e$), n' continues to rise, albeit increasingly slowly, as larger and larger numbers of loci are used to compute $\hat{r}^2{}_{(n)}$. In contrast, for incomplete sampling ($S<N$), n' rapidly plateaus once about 1000 loci are used, after which additional loci do little to increase precision. As a consequence, for 10,000 SNPs n' is about 3 orders of magnitude smaller than the number of pairwise comparisons. Furthermore, as long as the entire population is not exhaustively sampled, sample size has relatively little effect on n' (Figure 3). Samples of individuals were all drawn from the $N=N_e$ individuals in the final simulated generation, and the variance associated with replicate hypergeometric samples is proportional to $(N-S)/N$. *A priori*, therefore, we expected that n' would increase smoothly with sample size. Two factors likely explain the patterns actually observed.

First, larger samples produce less sampling error (Hill 1981; Waples 2006), which reduces both var($\hat{r}^2$) and mean($\hat{r}^2$). Because n' depends on ϕ which is the ratio of var($\hat{r}^2$) to [mean($\hat{r}^2$)]^2, the net effect of increasing S is only a modest change in n'. Second, the variance components analysis (Figure 4) shows that whereas V_1 (variance among replicate sets of loci assayed on the same individuals) continues to decline rapidly with increasing numbers of loci, V_2 (variance among different samples of individuals assayed for the same loci) does not decline much after about 500-1000 loci are used. As a consequence, except for small numbers of loci, V_2 dominates overall var($\hat{r}^2{}_{(n)}$) and hence this component largely determines n'. Although sampling a larger fraction of the population does reduce V_2, the resulting variance still is much larger than V_1 and still dominates n'. Note also that the actual variance of $\hat{r}^2{}_{(n)}$ is smaller than the sum of V_1 and V_2 (Figure 4), which indicates that the two variance components must be negatively correlated to some extent. In what follows, we focus on $S<N$, which is the most common scenario in studies of natural populations.

Restricting LD analyses to pairs of loci on different chromosomes reduces the number of pairwise comparisons by the proportion $1/Chr$, but this has little effect on n', which is essentially the same regardless whether same-chromosome comparisons are allowed or not (Figure S7). Unless otherwise noted, all results presented are for all pairwise comparisons.

The fact that results for unlinked loci approximate those for simulations with 64 chromosomes, and that both scenarios show $n' \ll n$ for large numbers of loci (Figure 2, top), indicate that physical linkage is not the primary factor that reduces the effective degrees of

freedom for analyses involving LD. The second factor relevant to two-locus analyses involves overlapping pairs of the same loci. We expect that, if we have four unlinked loci (w,x,y,z) and compute all six pairwise correlations, the correlation [w,x] will be independent of the correlation [y,z], but to what extent does the correlation [w,x] provide independent information to the correlation [w,y], since one locus is shared?

To evaluate this factor, we simulated unlinked loci in many replicate populations and calculated r^2 for each pair of loci in each replicate, across all N_e individuals in the population (see Detailed Methods in Supplemental Information). Then, across all replicates, we computed the squared correlation between r^2 values that did and did not share one locus. Results show that correlations between pairs of r^2 values that did not share a locus were essentially 0, regardless how large or small N_e was (Figure S8). This is also the result one obtains if one compares correlations between pairs of vectors of *i.i.d.* random variables, regardless whether the pairwise correlations being compared share one variable or not (data not shown). However, when the data being compared are generated by a process mediated by a pedigree (as occurs during reproduction in finite populations), then the pairwise correlations are not independent when they share one variable, and the degree of non-independence is inversely related to N_e (Figure S8).

Confidence intervals: For datasets with more than 1000 loci, the number of pairwise comparisons of loci is of order 10^6 or higher, in which case parametric CIs for $\widehat{N}_e$ that assume all datapoints are independent become vanishingly small (Figure 5). Based on estimates of the effective degrees of freedom obtained in this study, actual CIs for large numbers of loci are much wider, and this difference is important to understand to avoid misleading conclusions about precision.

Our results provide a basis for users to develop robust CIs for their own datasets. The best model to predict n' included several sets of covariates for each parameter in the Michaelis-Menten asymptotic function (p,q,r). Parameter estimates, standard errors, and more details are included in Supplemental Information. When fit to the original data, the correlation between log(predicted n') and log(true n') was 0.997 (Figure S10). When we evaluated performance of CIs for $\widehat{N}_e$ using Equation S5 based on these predicted n' values, the fraction of 90% CIs that contained the true N_e was close to the expected 0.9 (data not shown). But a typical user will know covariate values only for the numbers of individuals and loci in their samples and will have to estimate N_e (from the genetic data, or elsewhere) and perhaps *Chr* (e.g., from a related species), and finally estimate n' based on our modeling results. Accounting for these additional sources of uncertainty reduced CI performance only slightly, such that overall coverage was 90% for S=100, 88% for S=50, and 87% for S=25 (Table 4).

Performance of the Jones et al. (2016) jackknife method varied with sample size (Table 4), which is not surprising given that this method jackknifes over individuals. For S=100 the method produced conservative 90% CIs that included the true N_e>93% of the time, indicating that n' was generally underestimated. Overall coverage was close to the expected 90% for S=50 but only 87% for S=25, indicating a tendency of the jackknife method to overestimate precision for small samples. Across all scenarios, most of the jackknife CIs that did not contain the true value were too high, meaning that the lower bound was larger than true N_e, and this effect was much stronger for smaller samples.

Scatterplots of estimated n' vs mean r^2 for individual datasets (Figure S11) help to explain performance of Jones's jackknife method. First, $\hat{n}$' varied considerably across replicates, spanning one order of magnitude for S=100 and more than two orders for S=25. Second, for all

scenarios we found a strong, negative correlation between $\hat{n}$' and mean r^2. Datasets with below-average mean r^2 consistently were estimated to have a relatively high n' and hence relatively narrow CIs. Relatively small r^2 translates into a relatively large estimate of N_e, around which the CI was relatively narrow. In combination, these factors produce an excess of large $\widehat{N}_e$ estimates whose lower CI bound is larger than true N_e, and this effect is exacerbated for small S. Because n' is an increasing function of N_e, estimates of n' based on modeling results in this study also show a negative correlation between $\hat{n}$' and mean r^2, but the range of $\hat{n}$' for the current study was about half that of the Jones jackknife method (Figure S11).

F_{ST}

An initial sensitivity analysis (Figures S12-S13 and Supplemental Information) showed that Nei's weighted $\hat{F}_{ST}^{Nei}$ performed better than the unweighted version, and ascertainment method 3 performed better than the other options. Therefore, results that follow apply to $\hat{F}_{ST}^{Nei}$ and $\hat{F}_{ST}^{Hudson}$ with a MAF cutoff of 0.05.

With no pseudoreplication, variance of $\hat{F}_{ST(n)}$ should be inversely proportional to n; the actual rate of decline in var($\hat{F}_{ST(n)}$) as the number of SNPs increases is shown in Figure 6. For a scenario with N_e=200, Chr=4, and S=25, the decline is approximately log-linear up to $n \approx 2\text{x}10^3$, after which point addition of more loci does little to further reduce the variance. For a scenario with larger N_e, S, and genome size, the rate of decline in var($\hat{F}_{ST(n)}$) does not start to plateau until the number of loci is an order of magnitude larger (Figure S14). Notably, in both scenarios var($\hat{F}_{ST}^{Hudson}$) declines at the same rate as var($\hat{F}_{ST}^{Nei}$).

For a given sample size, mean values of the Nei and Hudson estimators are perfectly correlated with a common slope and sample-size specific intercepts (Figure S18); the slopes and intercepts, however, vary with mean $\hat{F}_{ST}$. This linear relationship explains why the rate of decline in var($\hat{F}_{ST}^{Hudson}$) with increasing numbers of loci is nearly identical to that for var($\hat{F}_{ST}^{Nei}$) (Figure 6).

Effective degrees of freedom: Table S4 gives our estimates of n' for F_{ST} for every scenario we simulated. As with LD, we found that n' for F_{ST} depends strongly on both N_e and genome size (Figure 7). In contrast to the two-locus results, however, we found a clear F_{ST} asymptote for n' only for smaller values of N_e and Chr; for larger values, n' was still increasing after bringing 200K loci into the analysis. As also with LD, we found a more modest effect on n' of sample size of individuals, which became more pronounced for large numbers of loci (Figure S15).

For a given number of loci, n'/n was considerably higher for F_{ST} than for LD (Figure S17). Nevertheless, in populations with small N_e and few chromosomes that are assayed for large numbers of loci, the effective degrees of freedom associated with mean F_{ST} can be two orders of magnitude or more smaller than the number of SNPs. The variance components analysis for F_{ST} (Figure S16) produced a general pattern similar to that for LD (Figure 4), with one important difference. For the same evolutionary scenario (N_e=200, Chr=16, S = 50), whereas V_2 for LD rapidly diverges from V_1 and starts to plateau after ~500-1000 loci, the divergence comes much later for F_{ST}, and V_2 does not begin to level off until 10^4–10^5 loci are included.

Figure 8 provides a detailed look at variation in $\hat{F}_{ST}^{Hudson}$ across 48 samples from each of two 2-population pedigrees that produced substantially different mean $\hat{F}_{ST(n)}$ values. Within each pedigree, for each of eight samples of individuals, variation among six replicate samples of

n loci reflects variance component V_1.

Our estimates of n' for $\hat{F}_{ST(n)}$ can be used to estimate realistic variances for $\hat{F}_{ST(n)}^{Hudson}$ as follows: (1) Given that $n' = 2/\phi$ and $\phi = \text{var}(\hat{F}_{ST(n)}^{Nei})/[\text{mean}(\hat{F}_{ST(n)}^{Nei})]^2$, the variance of Nei's multilocus estimator declines according to $\text{var}(\hat{F}_{ST(n)}^{Nei}) = \text{var}(\hat{F}_{ST(1)}^{Nei})/n'$, where $\text{var}(\hat{F}_{ST(1)}^{Nei})$ is the variance among single-locus $\hat{F}_{ST}$ values and $\text{var}(\hat{F}_{ST(1)}^{Nei}) = 2*[\text{mean}(\hat{F}_{ST(n)}^{Nei})]^2$. (2) Because $\text{var}(\hat{F}_{ST(n)}^{Hudson})$ declines at the same rate as $\text{var}(\hat{F}_{ST(n)}^{Nei})$, $\text{var}(\hat{F}_{ST(n)}^{Hudson}) = \text{var}(\hat{F}_{ST(1)}^{Hudson})/n'$, where $\text{var}(\hat{F}_{ST(1)}^{Hudson})$ is the variance of single-locus estimates and can be calculated from empirical data.

Confidence intervals*:* The best model to predict n' for F_{ST} was similar in form to that found for LD (Supplemental Information). When fit to the original data, the correlation between log(predicted n') and log(true n') was >0.99 (Figure S10). CIs based on modeled estimates of $\text{var}(\hat{F}_{ST(n)}^{Nei})$ obtained in this study contained the true value 89-95% of the time (mean 91.5%), and of those that did not, roughly equal numbers were too high and too low (Figure 9 and Table 5).

Across all scenarios [N_e,*Chr,L,S*], we found three consistent patterns in block jackknife results, and each applied equally to the Nei and Hudson estimators. (1) Block jackknife estimates of variance are positively correlated with mean $\hat{F}_{ST(n)}^{Nei}$ and mean $\hat{F}_{ST(n)}^{Hudson}$ (Figure S19); (2) Virtually all block jackknife estimates exceeded the actual variances we calculated from our simulations (Figure 9 and Table 5); (3) Replicate samples generated using the same parameters produced block jackknife estimates of variance that ranged widely in magnitude (Figure S19), and this variation was greater for the larger block size (one chromosome) and smaller samples of individuals. Because the 5Mb blocks showed less variation among replicates, we focus on those results. Upward bias in the block jackknife estimates of variance produced overly conservative confidence intervals that almost always contained the true values of $\hat{F}_{ST(n)}^{Nei}$ and $\hat{F}_{ST(n)}^{Hudson}$ (98-100% coverage for 90% CIs for both Hudson and Nei estimators; Figure 9 and Tables 5 and S6).

R code that allows users to predict n' for their own data or other scenarios, for both F_{ST} and r^2, is provided in Supplemental Information.

DISCUSSION

The common scenario considered in this paper involves a researcher who has collected data for large numbers of genetic markers in one or a few actual populations. All real populations have a single multigeneration pedigree, and a typical goal is to use genetic methods to draw inferences about evolutionary process that helped shape that population pedigree. Our primary interest is on quantifying uncertainty arising from two sources: sampling genes, and sampling individuals. We do this by simulating many replicate populations and measuring how fast variances of key genetic parameters decline as more loci are used and individuals are sampled.

A substantial complication arises from the fact that the Wright-Fisher reproduction process is inherently stochastic and has many possible realizations. As a consequence, replicate WF populations have different multi-generational pedigrees and different mean values for genetic indices like r^2 and F_{ST} (Cockerham and Weir 1983; Waples and Faulkner 2009), and averaging across this sort of demographic variance would inflate our estimates of $\text{var}(\hat{F}_{ST})$ and $\text{var}(\hat{r}^2)$ and bias results. To avoid this complication, we used WF reproduction for the forward-in-time component of our simulations, but we took advantage of recently-developed methods

(Haller et al. 2019) to generate many replicate samples of genes and individuals drawn from the same population pedigree. This eliminated the demographic component of variance so we could focus on uncertainty associated with sampling of genes and individuals from a single population (or pair of populations). Collectively, our results demonstrate the importance of accounting for differences between pedigrees of the population as a whole and pedigrees of sampled individuals. The variance components analysis showed that for both r^2 and F_{ST}, the primary factor limiting precision in genomics-scale datasets is uncertainty associated with sampling individuals, and this uncertainty cannot be eliminated by sampling arbitrarily large numbers of genes for the same individuals. This is an important consideration for experimental design in genomics-scale datasets, as often it is faster, easier, and cheaper to assay large numbers of genes for a small number of individuals, rather than the reverse.

Effects of pedigrees on statistical inference have been noted previously (e.g., Laurie and Weir 2003; Wakeley et al. 2012, 2016; Ralph 2019). The standard coalescent treats every independent gene as if it were produced on a different pedigree, but in real populations all genes have to percolate through the single, fixed pedigree that captures the genealogical relationships among individuals in the current population and their ancestors, and this pedigree-dependence creates correlations among alleles at different gene loci (Bhaskar and Song 2009). This effect is strongest for analyses that are sensitive to pedigrees from the most recent generations, such as relatedness, admixture, population differentiation, and LD (King et al. 2018; Nelson et al. 2020). Furthermore, departures from the standard coalescent model increase when sample size is more than a small fraction of effective size (Bhaskar et al. 2014), a condition commonly encountered in real-world applications.

Linkage disequilibrium

Our simulations show that, except in populations with large N_e and large genomes, once a few thousand diallelic loci are used to estimate multilocus r^2, adding more loci does little to further reduce var($\hat{r}^2$). As a consequence, for datasets with 10^4 or more SNPs, n' can be many orders of magnitude smaller than the number of pairs of loci. This in turn means that confidence intervals for $\widehat{N}_e$ are much wider than they would be if all the pairwise comparisons were independent. The modeling results find significant effects of N_e, *Chr*, *S* and their interactions on n', with N_e having the strongest influence.

The fact that n' depends heavily on N_e even for non-syntenic loci indicates that physical linkage is not the major factor creating lack of independence of pairwise r^2 values; instead, most pseudoreplication arises from overlapping pairs of the same loci in multiple pairwise comparisons. Surprisingly (but conveniently), n' differs very little whether all pairwise comparisons are used or only those on different chromosomes. Restricting comparisons to non-syntenic loci reduces n but simultaneously increases n'/n, and the two factors effectively offset each other.

It is somewhat ironic that the degree of physical linkage has relatively little effect on pseudoreplication in analyses of LD, but this result can be understood when one considers what happens as more loci are packed into a fixed number of chromosomes. Any new locus will inevitably be linked with many existing loci on whatever chromosome it ends up on, but for every such pairing there are (*Chr*-1)/*Chr* new comparisons with existing loci on other chromosomes, and these dilute any effects of linkage. Here we quantify for the first time the large amount of pseudoreplication that occurs because each locus appears in many pairwise

comparisons. This overlapping-pairs-of-loci effect is modulated through the population pedigree and is strongest when N_e is small.

Values of n' estimated by the Jones et al. (2016) jackknife method were close to, but on average slightly smaller than, overall n' we calculated from simulated data. For replicate datasets simulated using the same fixed parameters, jackknife estimates of n' varied widely and were negatively correlated with mean r^2, which produced relatively narrow CIs when N_e was estimated to be relatively large. Collectively, these features cause results of the jackknife method to be somewhat unpredictable, being on average slightly conservative for larger sample sizes and the opposite for S=25, with the latter producing a large excess of CIs that were higher than true N_e. Overall coverage of CIs based on n' values estimated by our modeling results was close to the target 90%. However, because of the strong positive correlation between n' and N_e, our modeling results also produce CIs that are relatively narrow when $\widehat{N}_e$ is relatively high, an effect that can be reduced by setting an upper limit to $\widehat{N}_e$ when estimating n'.

F_{ST}

Consequences of pseudoreplication for precision of $\hat{F}_{ST}$ are less dramatic than those for r^2, but nonetheless not trivial. n' approaches an asymptote for relatively small genomes and small N_e, but for much larger numbers of loci (~10-20K) than is the case for r^2. For relatively large effective and genome sizes, n' was still increasing after 200K loci, indicating that in those circumstances precision of $\hat{F}_{ST}$ can be enhanced by very large numbers of loci.

Increasing *Chr* from 1 to 4 increased n' more than any comparable increases in N_e. However, most higher organisms have *Chr*≥4 (Table 1), in which case comparable increases in genome size and effective size produce roughly similar increases in precision. As is the case with LD, increases in sample size have less effect on n' for $\hat{F}_{ST}$. Although ascertainment and estimation methods we evaluated affect mean values of the index, these differences do not affect the rate at which var($\hat{F}^{Nei}_{ST(n)}$) or var($\hat{F}^{Hudson}_{ST(n)}$) declines with addition of more loci. In related evaluations, we found that changes in the variance of temporal F (Nei and Tajima 1981; Jorde and Ryman 2007) parallel those for F_{ST} (unpublished data). This means that results obtained here can be applied broadly to predict realized precision of F statistics and related measures in genomics-scale datasets.

We found that widely-used block jackknife methods consistently overestimate var($\hat{F}_{ST}$), leading to CIs that are much too conservative. In addition, for a given parameter set, the estimated jackknife variance varied several-fold across replicates and was positively correlated with the mean. According to Busing et al. (1999), block size should be large enough to encompass all non-independence, which suggests the appropriate block size should be one chromosome. Although the common block size of 5Mb might be large enough to capture correlations involving sites near the center of the block, this approach arbitrarily divides a continuous system of correlated loci in a way that guarantees that many tightly-linked pairs will be in different blocks. Despite this apparent drawback, the undesirable attributes mentioned above were less extreme for the 5Mb blocks. CIs for $\hat{F}_{ST}$ based on n' estimated according to results of this study performed well, even after accounting for uncertainty associated with estimating N_e and *Chr*.

Experimental design and practical applications

Our simulation and modeling results demonstrate that robust estimates of n' can be obtained as a function of numbers of loci and individuals, genome size, and N_e. The first two

covariates are under control of the investigator, and the third can generally be approximated reasonably well, even for non-model species. Dependence of n' on N_e introduces a complication, but with even moderate amounts of genetic data one can obtain a fairly precise estimate of N_e using either single-sample or two-sample (temporal) methods. Even after having to estimate N_e to predict n', LD-based confidence intervals for $\widehat{N}_e$ performed at least as well as those obtained using the Jones et al. (2016) jackknife method, and with less variability among replicates (Table 4; Figure S11). Why the block-jackknife method consistently overestimates var($\widehat{F}_{ST}$) and produces CIs that are too wide is not clear, but it might be related to the fact that blocks with arbitrary boundaries within chromosomes do not capture all dependencies among loci. In any case, using our modeling results to predict n' produced robust CIs for $\widehat{F}_{ST}$. Our results should be particularly useful in planning and experimental design, as expected precision for a wide range of scenarios can be evaluated quickly and easily.

The simulation framework we used, which combines coalescent simulations of the distant past with fast and efficient Wright-Fisher forward simulations of the recent past, provides more realistic results than can be achieved by either process alone (Nelson et al. 2020). Nevertheless, as is inevitable our model required simplifying assumptions, so some caveats are in order. We assumed closed populations and did not evaluate potential consequences of migration for var($\hat{r}^2$) or var($\widehat{F}_{ST}$). We modeled non-random associations of neutral genes and did not attempt to account for correlations due to selective pressures on linked or unlinked sites, so in that respect our estimates might be considered upper limits to actual n'. We did not explicitly model variation in recombination rate, which is known to be common both across the genome and between sexes (Ritz et al. 2017; Sardell and Kirkpatrick 2020), so our results reflect a generic genome-wide average. Although the genome sizes we simulated (1-64 chromosomes of 50 Mb) encompassed the range of mean values reported for higher organisms (Table 1), all chromosomes we modeled were the same size. For real species, which have chromosomes of varying sizes generally smaller or larger than 50 Mb, the following procedure can be used to convert genome size into an index comparable to that used in this study. (1) Calculate the effective number of chromosomes as $C_e=1/\Sigma x_i^2$, where x_i is the relative length of the ith chromosome, standardized such that $\Sigma x_i=1$. C_e is analogous to the effective number of alleles at multiallelic loci (see Supplemental Information). (2) Calculate total equivalent genome size as $\bar{L}C_e$, where $\bar{L}$ is the mean chromosome length in Mb. (3) Divide the result by 50 to obtain the equivalent number of equal-sized chromosomes as modeled in this study.

We modeled discrete generations, and our sampling design assumed that individuals were sampled randomly from the entire adult population. Forward simulations used Wright-Fisher dynamics with a constant number of ideal adults (N), so $N_e \approx N$. We found a qualitative difference in n' for samples that included the entire population ($S=N_e$), but for real populations (typically with $N_e<N$) the relevant criterion is whether all individuals have been sampled ($S=N$). n' values reported in Table S2 for $S=N_e$ would provide a robust estimate of expected precision for the special case where it is possible to assay the entire population, thus eliminating the large variance component associated with sampling individuals. Finally, for many species it is most convenient to sample juvenile offspring rather than adults. The variance associated with juvenile samples approximates that for a very small sample of the parents (see Supplemental Information). Therefore, an approximate value for n' for such samples can be obtained by using the predicted n' from this study for small S.

Acknowledgments

We thank Eric Anderson, Nicholas Galwey, Marty Kardos, Peter Ralph, and John Wakeley for useful discussions and Martin Liermann for suggestions regarding model fitting.

Author contributions

RSW conceived the study; RSW and RKW designed the study and simulated and analyzed the data. EW conducted the analyses to model *n*' as a function of key covariates. RSW wrote the manuscript with input from RKW and EW, and all authors edited the manuscript.

Literature Cited

Aarts, E., Verhage, M., Veenvliet, J.V., Dolan, C.V. and Van Der Sluis, S., 2014. A solution to dependency: using multilevel analysis to accommodate nested data. *Nature Neuroscience*, *17,* 491.

Aguirre, N.C., Filippi, C.V., Zaina, G., Rivas, J.G., Acuña, C.V., Villalba, P.V., García, M.N., González, S., Rivarola, M., Martínez, M.C. and Puebla, A.F., 2019. Optimizing ddRADseq in non-model species: A case study in Eucalyptus dunnii Maiden. *Agronomy*, *9*(9), p.484.

Albrechtsen, A., Nielsen, F.C. and Nielsen, R., 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*, *27*(11), 2534-2547.

Beverton, R. J. H.; Holt, S. J. (1957), On the Dynamics of Exploited Fish Populations, Fishery Investigations Series II Volume XIX, Ministry of Agriculture, Fisheries and Food.

Bhaskar, A., Song, Y.S., 2009. Multi-locus match probability in a finite population: a fundamental difference between the Moran and Wright-Fisher models. Bioinformatics, 25, i187–i195.

Bhaskar, A., Clark, A.G. and Song, Y.S., 2014. Distortion of genealogical properties when the sample is very large. *Proceedings of the National Academy of Sciences*, *111*(6), pp.2385-2390.

Bhatia, G., Patterson, N., Sankararaman, S. and Price, A.L., 2013. Estimating and interpreting F_{ST}: the impact of rare variants. *Genome research*, *23*(9), pp.1514-1521.

Busing, F.M., Meijer, E. and Van Der Leeden, R., 1999. Delete-m jackknife for unequal m. *Statistics and Computing*, *9*(1), 3-8.

Choquet, M., Smolina, I., Dhanasiri, A.K., Blanco-Bercial, L., Kopp, M., Jueterbock, A., Sundaram, A.Y. and Hoarau, G., 2019. Towards population genomics in non-model species with large genomes: a case study of the marine zooplankton Calanus finmarchicus. *Royal Society open science*, *6*(2), p.180608.

Cockerham, C. C. 1969. Variance of gene frequencies. Evolution 23:72-84.

Cockerham, C.C. and Weir, B.S., 1983. Variance of actual inbreeding. *Theoretical population biology*, *23*(1), pp.85-109.

Cockerham, C.C. and Weir, B.S., 1987. Correlations, descent measures: drift with migration and mutation. *Proceedings of the National Academy of Sciences*, *84*(23), pp.8512-8514.

Colegrave, N. and Ruxton, G.D., 2018. Using biological insight and pragmatism when thinking about pseudoreplication. *Trends in ecology & evolution*, *33*(1), pp.28-35.

Cox, D.R., 1984. Effective degrees of freedom and the likelihood ratio test. *Biometrika*, *71*(3), pp.487-493.

da Fonseca RR, Albrechtsen A, Themudo GE, Ramos-Madrigal J, Sibbesen JA, Maretty L, Zepeda-Mendoza ML, Campos PF, Heller R, Pereira RJ. 2016. Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine Genomics* 30, 3-13.

Foll, M. and Gaggiotti, O.E., 2008. A genome scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180:977:993.

Galwey NW. 2009. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology* 33(7):559-68.

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet JM, Estoup A. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between

populations. *Mol Ecol*. 22:3165–3178.
Giesbrecht, FG. 2006. Degrees of freedom, effective. Encyclopedia of Statistical Sciences. Available at https://onlinelibrary.wiley.com/doi/abs/10.1002/0471667196.ess0539.pub2.
Graffelman J, Nelson S, Gogarten SM, Weir BS. 2015. Exact Inference for Hardy-Weinberg Proportions with Missing Genotypes: Single and Multiple Imputation. G3: Genes| Genomes| Genetics 5(11):2365-73.
Haller, B.C., Galloway, J., Kelleher, J., Messer, P.W. and Ralph, P.L., 2019. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular ecology resources*, *19*(2), pp.552-566.
Hill, W.G. 1981. Estimation of effective population size from data on linkage disequilibrium. *Genetical Research* 38:209–216.
Huang H, Knowles LL. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. 2014. *Systematic Biology* 4:syu046.
Hudson, R.R., Slatkin, M. and Maddison, W.P., 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*(2), 583-589.
Hurlbert, S.H., 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, *54*, 187-211.
Jones AT, Ovenden JR, Wang Y-G. 2016. Improved confidence intervals for the linkage disequilibrium method for estimating effective population size Heredity *117*(4), 217-223.
Jorde, P.E., and N. Ryman. 2007. Unbiased estimator for genetic drift and effective population size. *Genetics* 177: 927–935.
Kelleher J, AM Etheridge, G McVean (2016), *Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes*, *PLoS Comput Biol* 12(5): e1004842. doi: 10.1371/journal.pcbi.1004842
King, L., Wakeley, J. and Carmi, S., 2018. A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci. *Theoretical Population Biology* 122:22-29.
Laurie, C., Weir, B.S., 2003. Dependency effects in multi-locus match probabilities. *Theor. Popul. Biol*. 63, 207–219.
Leaché AD, Banbury BL, Felsenstein J, de Oca AN, Stamatakis A. 2015. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. Systematic Biology, 29:syv053.
Lewontin, R.C. and Krakauer, J., 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, *74*(1), 175-195.
Li, X., Zhu, C., Lin, Z., Wu, Y., Zhang, D., Bai, G., Song, W., Ma, J., Muehlbauer, G.J., Scanlon, M.J. and Zhang, M., 2011. Chromosome size in diploid eukaryotic species centers on the average length with a conserved boundary. *Molecular biology and evolution*, *28*(6), pp.1901-1911.
Lin, Y., Ghazanfar, S., Wang, K.Y., Gagnon-Bartsch, J.A., Lo, K.K., Su, X., Han, Z.G., Ormerod, J.T., Speed, T.P., Yang, P. and Yang, J.Y.H., 2019. scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proceedings of the National Academy of Sciences*, *116*(20), pp.9775-9784.
Messer PW. 2013. SLiM: simulating evolution with selection and linkage. Genetics 1944:1037–1039.

Minias, P., Dunn, P.O., Whittingham, L.A., Johnson, J.A. and Oyler-McCance, S.J., 2019. Evaluation of a Chicken 600K SNP genotyping array in non-model species of grouse. *Scientific reports*, *9*(1), pp.1-10.

Nei M. 1973. Analysis of gene diversity in subdivided populations. Proc Nat Acad Sci, USA. 70:3321-3323.

Nei, M. and Chakravarti, A., 1977. Drift variances of F_{ST} and G_{ST} statistics obtained from a-finite number of isolated populations. Theoretical Population Biology 11:307-325.

Nei, M. and Tajima, F. 1981. Genetic drift and estimation of effective population size. 40 *Genetics* **98:**625–640.

Nelson, D., Kelleher, J., Ragsdale, A.P., Moreau, C., McVean, G. and Gravel, S., 2020. Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLoS genetics*, *16*(5), p.e1008619.

Nyholt DR. 2004. A simple correction for multiple testing for single nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765–769.

Patterson, N. J., P. Moorjani, Y. Luo, S. Mallick, N. Rohland et al., 2012 Ancient admixture in human history. Genetics 192: 1065–1093.

Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. and Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), pp.559-575.

R R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Ralph, P.L., 2019. An empirical approach to demographic inference with genomic data. *Theoretical population biology*, *127*:91-101.

Ramage, B.S., Sheil, D., Salim, H.M., Fletcher, C., Mustafa, N.Z.A., Luruthusamay, J.C., Harrison, R.D., Butod, E., Dzulkiply, A.D., Kassim, A.R. and Potts, M.D., 2013. Pseudoreplication in tropical forests and the resulting effects on biodiversity conservation. *Conservation Biology*, *27*(2), 364-372.

Reynolds, J., B. S. Weir, And C. C. Cockerham. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767-779.

Ritz, K. R., M. A. Noor, and N. D. Singh. 2017. Variation in recombination rate: adaptive or not? Trends in Genetics 33:364–374.

Rosenblum, E.B. and Novembre, J., 2007. Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. *Journal of Heredity*, *98*(4), pp.331-336.

Sardell JM, Kirkpatrick M. 2020. Sex differences in the recombination landscape. The American Naturalist 195:361–379.

Thompson, EA. 2013. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* 194:301-326.

Van Wyngaarden, M., Snelgrove, P.V., DiBacco, C., Hamilton, L.C., Rodríguez-Ezpeleta, N., Jeffery, N.W., Stanley, R.R. and Bradbury, I.R., 2017. Identifying patterns of dispersal, connectivity and selection in the sea scallop, Placopecten magellanicus, using RAD seq-derived SNPs. *Evolutionary Applications*, *10*(1), 102-117.

Wakeley, J., King, L., Low, B.S., Ramachandran, S., 2012. Gene genealogies within a fixed

pedigree, and the robustness of Kingman's coalescent. Genetics 190, 1433–1435.
Wakeley, J., King, L. and Wilton, P.R., 2016. Effects of the population pedigree on genetic signatures of historical demographic events. Proceedings of the National Academy of Sciences, 113(29), pp.7994-8001.
Waples RK, Larson WA, and Waples RS. 2016. Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. Heredity 117:233-240; doi:10.1038/hdy.2016.60
Waples, R.S. 2006. A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conservation Genetics 7:167-184.
Waples, R.S., and C. Do. 2008. *LDNe*: A program for estimating effective population size from data on linkage disequilibrium. Mol. Ecol. Resources 8:753-756.
Waples, R.S., and J.R. Faulkner. 2009. Modeling evolutionary processes in small populations: Not as ideal as you think. Molecular Ecology 18:1834-1847.
Weir, B.S. and Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. E*volution*, *38*(6), 1358-1370.
Wright 1951. The genetical structure of populations. Annals of Eugenics 15: 323-354.
Wolf, J.B. and Ellegren, H., 2017. Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, *18*(2), p.87.

Table 1. Summary of information on haploid (1*n*) chromosome number and mean chromosome size and genome size for major taxonomic groups (data from Li et al. 2011).

	Chromosome number				Size (Mb)	
	Mean	Min	Max	sd	Chromosome	Genome
Prokaryotes	2.2	2	4	0.50	2.5	5.7
Unicellular eukaryotes	16.6	2	35	9.33	1.7	23.6
Invertebrates	11.1	2	16	5.18	21.7	168.4
Vascular plants	13.4	5	20	5.66	47.7	558.0
Vertebrates	25.2	8	38	6.27	85.4	1933.1

Table 2. Notation used in this study

N	Number of individuals in each population (we assumed constant size with discrete generations)
N_e	Effective population size. Because we modeled Wright-Fisher reproduction, each generation $E(N_e) = N$; due to random demographic stochasticity, however, each replicate population has a different realized, multigeneration N_e
Chr	Number of chromosomes, each of length 50 Mb
S	Number of individuals sampled, drawn from the $N_e = N$ individuals in the final generation
L	Number of diallelic (SNP) loci
n	Nominal degrees of freedom = number of datapoints. For F_{ST}, $n = L$; for LD analyses, n = number of pairwise comparisons of loci = $L(L\text{-}1)/2 \approx L^2/2$
$\hat{F}_{ST(n)}$	An estimate of the standardized variance of allele frequency between two populations, based on data for n diallelic loci
θ	The coancestry coefficient, which is related to F_{ST} as shown in Equation 1
$\hat{F}_{ST}^{Nei}$	A weighted version of Nei's (1973) estimator of F_{ST}, commonly referred to as G_{ST}. For two populations and diallelic loci, G_{ST} is identical to Wright's (1951) F_{ST}.
$\hat{F}_{ST}^{Hudson}$	Hudson et al.'s (1992) estimator of F_{ST}, as implemented by Bhatia et al. (2013). For two populations and diallelic loci, $\hat{F}_{ST}^{Hudson}$ is identical to Weir and Cockerham's (1984) $\hat{\theta}$.
$\hat{r}^2{}_{(n)}$	An estimate of the mean squared correlation of alleles at different loci, based on data for n pairwise comparisons
ϕ	The squared coefficient of variation of $\hat{F}_{ST}$ or $\hat{r}^2$; ϕ = var/mean^2
n'	Effective degrees of freedom; $n' = 2/\phi$
V_1	The component of var($\hat{F}_{ST}$) or var($\hat{r}^2$) that reflects uncertainty associated with sampling replicate sets of genes for the same individuals
V_2	The component of var($\hat{F}_{ST}$) or var($\hat{r}^2$) that reflects uncertainty associated with taking replicate samples of individuals assayed for the same genes
CI	Confidence interval

Table 3. Experimental design for sampling individuals and genes for analyses of LD. From each population, subsets of individuals (rows) were drawn without replacement from the N_e total individuals in the final generation. Subsets of L loci (columns) were non-overlapping. Mean r^2 was calculated for each cell. The variance among mean r^2 within rows was used to estimate variance component V_1 (same individuals, different loci), and the variance within columns was used to estimate variance component V_2 (same loci, different individuals). A measure of pseduoreplication, ϕ, was calculated across mean r^2 for the cells (in **bold**) along the diagonal (different sets of individuals and loci). This sampling design was repeated for different numbers of loci (L), sampled individuals (S), chromosome number (Chr), and N_e. See the text and Figure S3 for details of a modified sampling design for F_{ST} that involved pairwise comparisons of daughter populations.

		Sets of loci				
		1	2	3	4	V_1
	A	$\mathbf{r^2_{A1}}$	r^2_{A2}	r^2_{A3}	r^2_{A4}	$var(r^2_{A*})$
Samples of	B	r^2_{B1}	$\mathbf{r^2_{B2}}$	r^2_{B3}	r^2_{B4}	$var(r^2_{B*})$
individuals	C	r^2_{C1}	r^2_{C2}	$\mathbf{r^2_{C3}}$	r^2_{C4}	$var(r^2_{C*})$
	D	r^2_{D1}	r^2_{D2}	r^2_{D3}	$\mathbf{r^2_{D4}}$	$var(r^2_{D*})$
	V_2	$var(r^2_{*1})$	$var(r^2_{*2})$	$var(r^2_{*3})$	$var(r^2_{*4})$	

Table 4. Effects of N_e, number of chromosomes (*Chr*), and number of individuals sampled (*S*) on coverage of confidence intervals (CIs) around mean r^2. Results are shown for CIs based on the Jones et al. (2016) jackknife method and *n*' estimated using the modeling results from this study, which required estimating N_e and *Chr*. Shown are the percentages of 1024 replicate samples whose 90% CIs included the true value ("In"), were entirely above the true value ("Above"), or were entirely below ("Below"). Each cell represents results averaged over simulations with data for 500, 1000, and 5000 SNPs, and the bottom set of rows averages results across all scenarios, by sample size and method. Results shown used data for pairs of loci on different chromosomes.

			Jackknife			This study		
N_e	*Chr*	*S*	Above	Below	In	Above	Below	In
200	4	25	9.4	1.7	88.9	12.7	1.8	85.5
200	4	50	7.1	2.3	90.6	8	3.8	88.2
200	4	100	4.9	2.5	92.7	4.6	3.6	91.8
200	16	25	9.5	1.3	89.1	9.4	2.4	88.2
200	16	50	5.8	1.6	92.6	7.1	5.0	87.9
200	16	100	3.5	1.7	94.8	4.7	4.4	90.9
800	4	25	14.8	0.6	84.6	11.7	0.8	87.5
800	4	50	8.4	1.2	90.4	10.1	1.5	88.4
800	4	100	6.1	2.7	91.2	7.4	4.5	88.1
800	16	25	17.1	0.5	82.4	11.4	1.2	87.5
800	16	50	8.9	1.0	90.1	9.7	2.9	87.4
800	16	100	5	1.9	93.1	6.1	3.7	90.1
		25	12.7	1.0	86.2	11.3	1.6	87.2
	Means	50	7.5	1.5	90.9	8.7	3.3	88.0
		100	4.9	2.2	92.9	5.7	4.1	90.2

Table 5. Effects of N_e, number of chromosomes (*Chr*), and number of individuals sampled (*S*) on coverage of confidence intervals (CIs) around $\hat{F}_{ST}$. Results are shown for CIs based on block jackknife estimates of var($\hat{F}_{ST}^{Hudson}$) and *n*' for $\hat{F}_{ST}^{Nei}$ estimated using the modeling results from this study, which required estimating N_e and *Chr*. Shown are the percentages of 1152 replicate samples whose 90% CIs included the true value ("In"), were entirely above the true value ("Above"), or were entirely below ("Below"). Each cell represents results averaged over simulations with data for 5000, 20000, and 50000 SNPs, and the bottom set of rows averages results across all scenarios, by sample size and method. These results used a block size of 5 Mb. Very similar results were found for block sizes of 1 chromosome and for block-jackknife estimates of var($\hat{F}_{ST}^{Nei}$) (see Table S6).

			Block jackknife			This study		
N_e	*Chr*	*S*	Above	Below	In	Above	Below	In
200	4	25	0.7	0.6	98.7	2.4	3.9	93.7
200	4	50	0.0	0.0	100.0	2.3	2.7	94.9
200	4	100	0.0	0.0	100.0	4.0	5.4	90.6
200	16	25	0.2	0.8	99.0	2.4	3.8	93.8
200	16	50	0.2	0.2	99.6	2.9	4.0	93.1
200	16	100	0.1	0.1	99.8	3.8	4.6	91.6
800	4	25	0.5	0.3	99.2	3.0	5.6	91.3
800	4	50	0.1	0.5	99.3	4.4	6.6	89.0
800	4	100	0.1	0.1	99.8	4.7	6.4	88.9
800	16	25	0.7	1.2	98.1	3.8	4.8	91.4
800	16	50	0.5	1.2	98.4	3.7	5.7	90.6
800	16	100	0.4	0.8	98.8	5.1	5.4	89.5
		25	0.5	0.7	98.7	2.9	4.5	92.5
	Means	50	0.2	0.5	99.3	3.3	4.8	91.9
		100	0.2	0.3	99.6	4.4	5.4	90.2

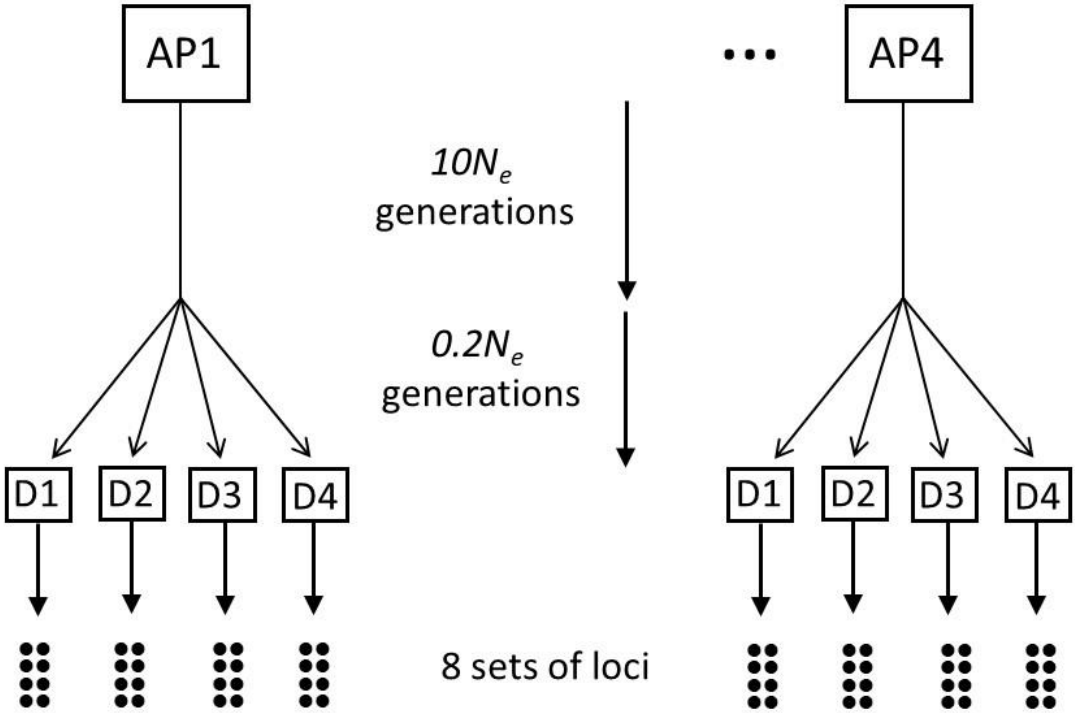

Figure 1. Experimental design for simulations. For each evolutionary scenario (combination of N_e and genome size), four ancestral populations (AP1–AP4) were simulated to ensure coalescence ($10N_e$ generations), at which point each ancestral population split into four daughter populations (D1-D4). The 4x4 = 16 daughter populations then evolved independently under isolation for $t = 0.2N_e$ generations. Subsequently, the model differed slightly for the F_{ST} and LD analyses. In the latter (as depicted in the figure), for each daughter population, eight mutational replicates (different set of loci) were generated based on the same pedigree, producing a total of 128 replicates for each evolutionary scenario. For F_{ST}, each set of four daughter populations allowed six pairwise comparisons of populations, and for each two-population pedigree six mutational replicates were generated (Figure S4).

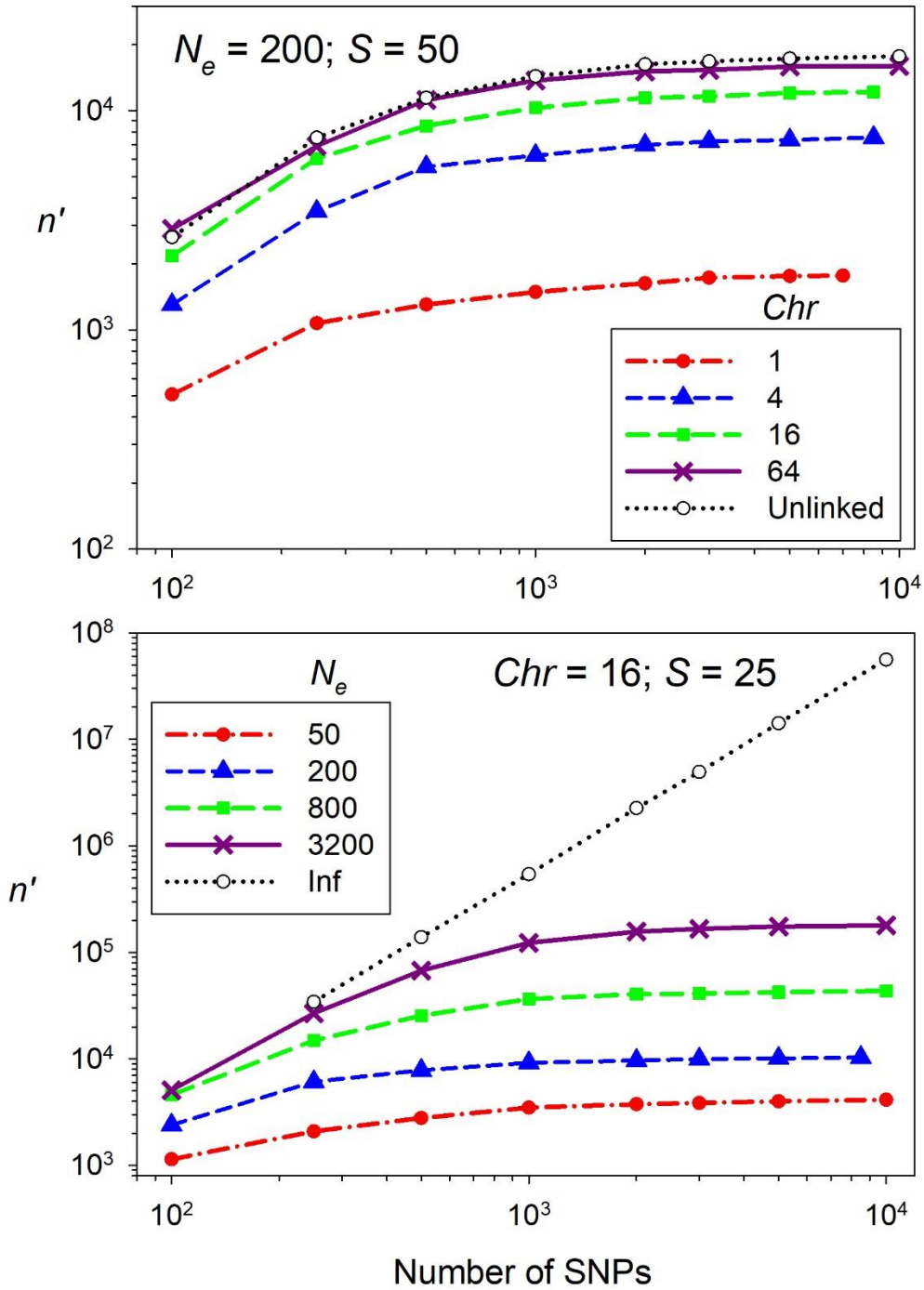

Figure 2. Effective degrees of freedom (n') for mean r^2 as a function of the number of diallelic (SNP) loci. Top: Effect of number of chromosomes (Chr), with N_e = 200 and S = 50. Bottom: effect of N_e, with Chr = 16 and S = 25. Mean r^2 was calculated across all pairs of loci.

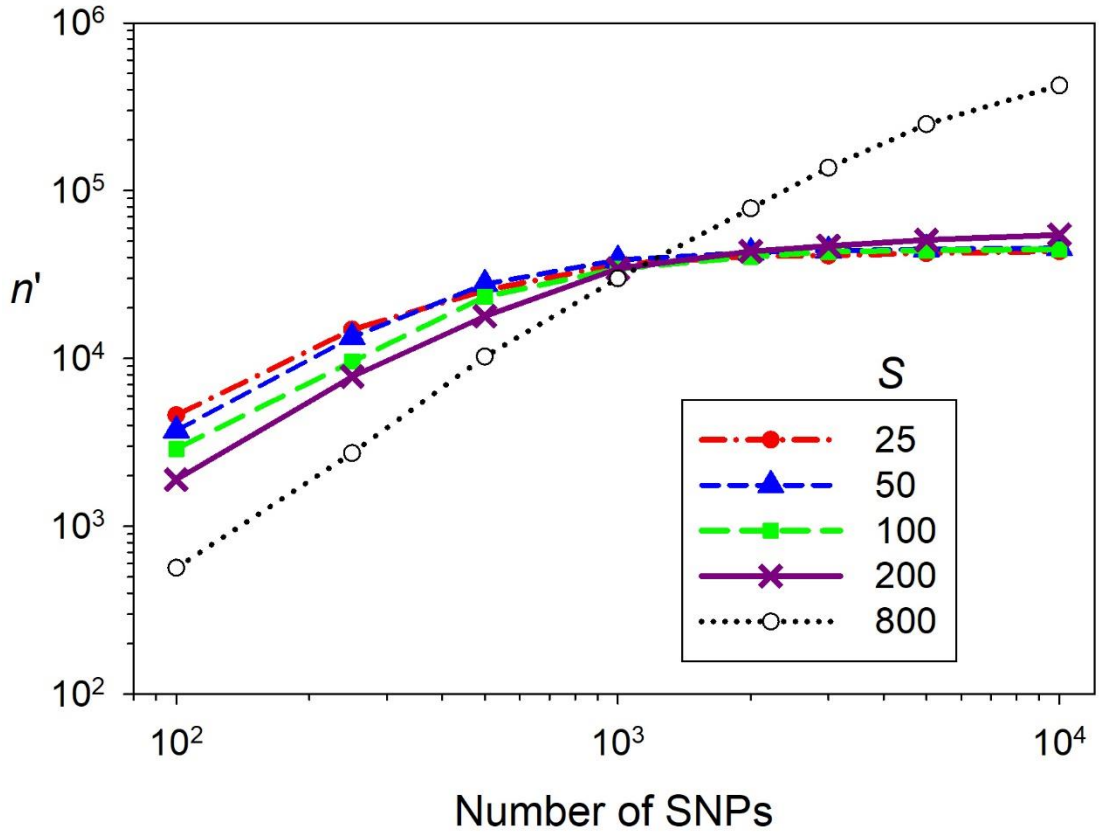

Figure 3. Effective degrees of freedom (*n*') for mean r^2 as a function of the sample size of individuals (*S* = 25-800) and the number of diallelic (SNP) loci. Results are for N_e = 800, *Chr* = 16, and using all pairwise comparisons of loci.

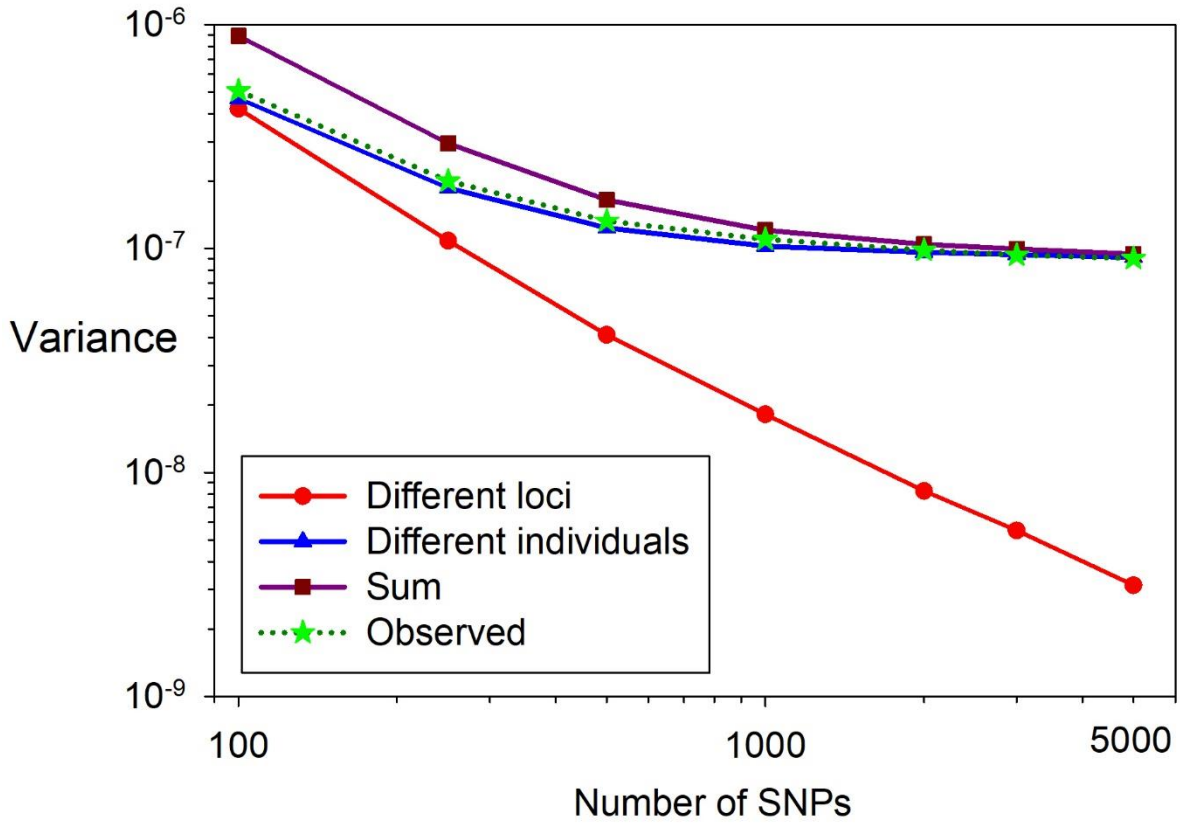

Figure 4. Variance components analysis for mean r^2. As depicted in Table 3, V_1 is the variance of mean r^2 for the same individuals assayed for different, non-overlapping sets of loci, and V_2 is the variance of mean r^2 for different (potentially overlapping) sets of individuals assayed for the same loci. "Sum" = V_1+V_2 and "Observed" is the total observed variance of mean r^2. Results are for N_e = 200, Chr = 16, S = 50, and using all pairwise comparisons of loci.

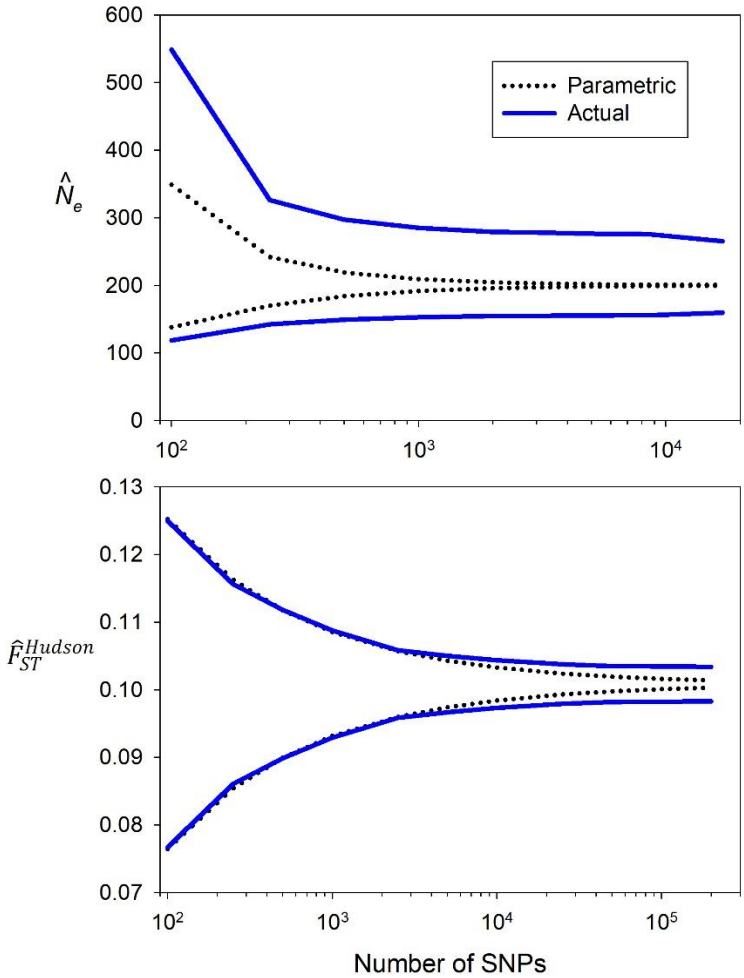

Figure 5. Comparison parametric and actual 90% confidence intervals for $\hat{N}_e$ based on LD (top) and $\hat{F}_{ST(n)}^{Hudson}$ (bottom). Parametric CIs use the nominal degrees of freedom (n = the number of diallelic (SNP) loci, L, for $\hat{F}_{ST(n)}^{Hudson}$; $n = L(L\text{-}1)/2$ for LD); actual CIs use the effective degrees of freedom calculated in this study. Results are for simulations with $N_e = 200$, $Chr = 16$, and $S = 50$. Note the different X-axis scales in the two panels.

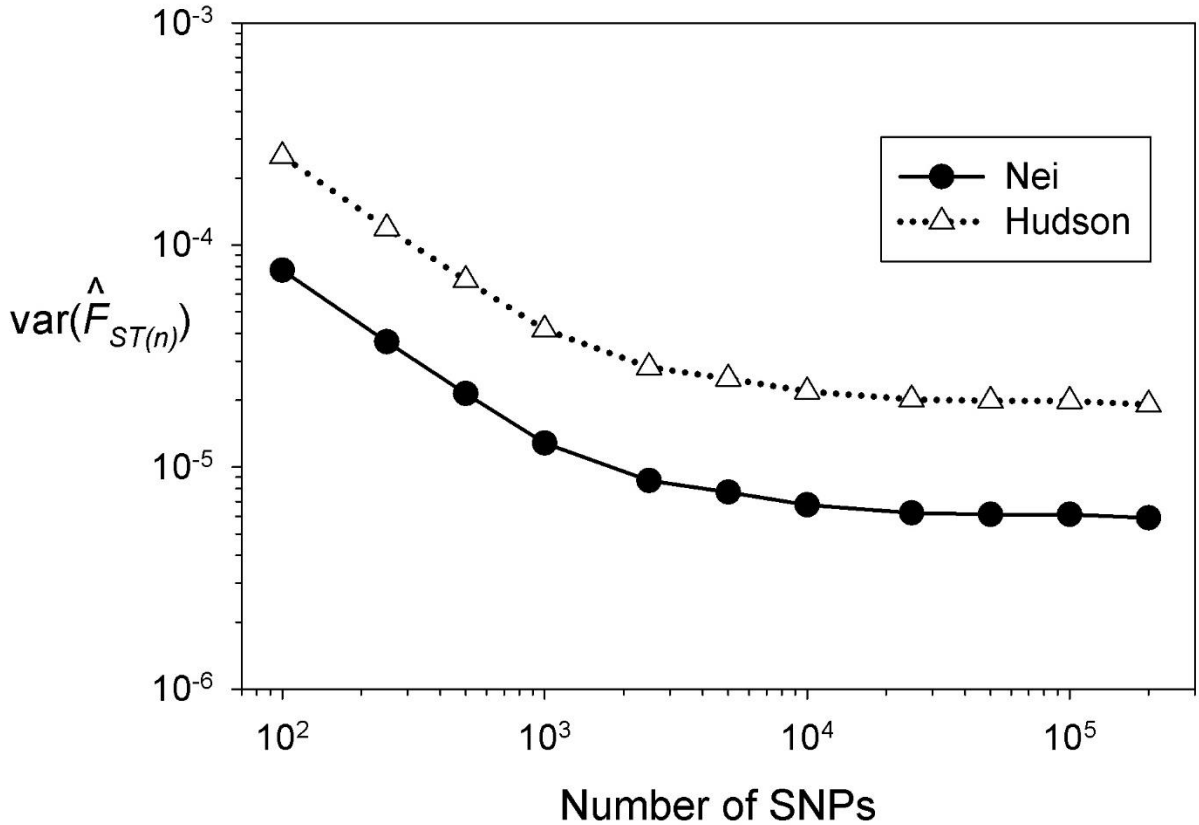

Figure 6. Rate of decline in the variance of multilocus $\hat{F}_{ST(n)}$ as more diallelic loci (SNPs, n) were used in the analysis. Results are for simulations with $N_e = 200$, $Chr = 4$, and $S = 25$ and are shown for the estimators of Nei ($\hat{F}_{ST}^{Nei}$) and Hudson ($\hat{F}_{ST}^{Hudson}$).

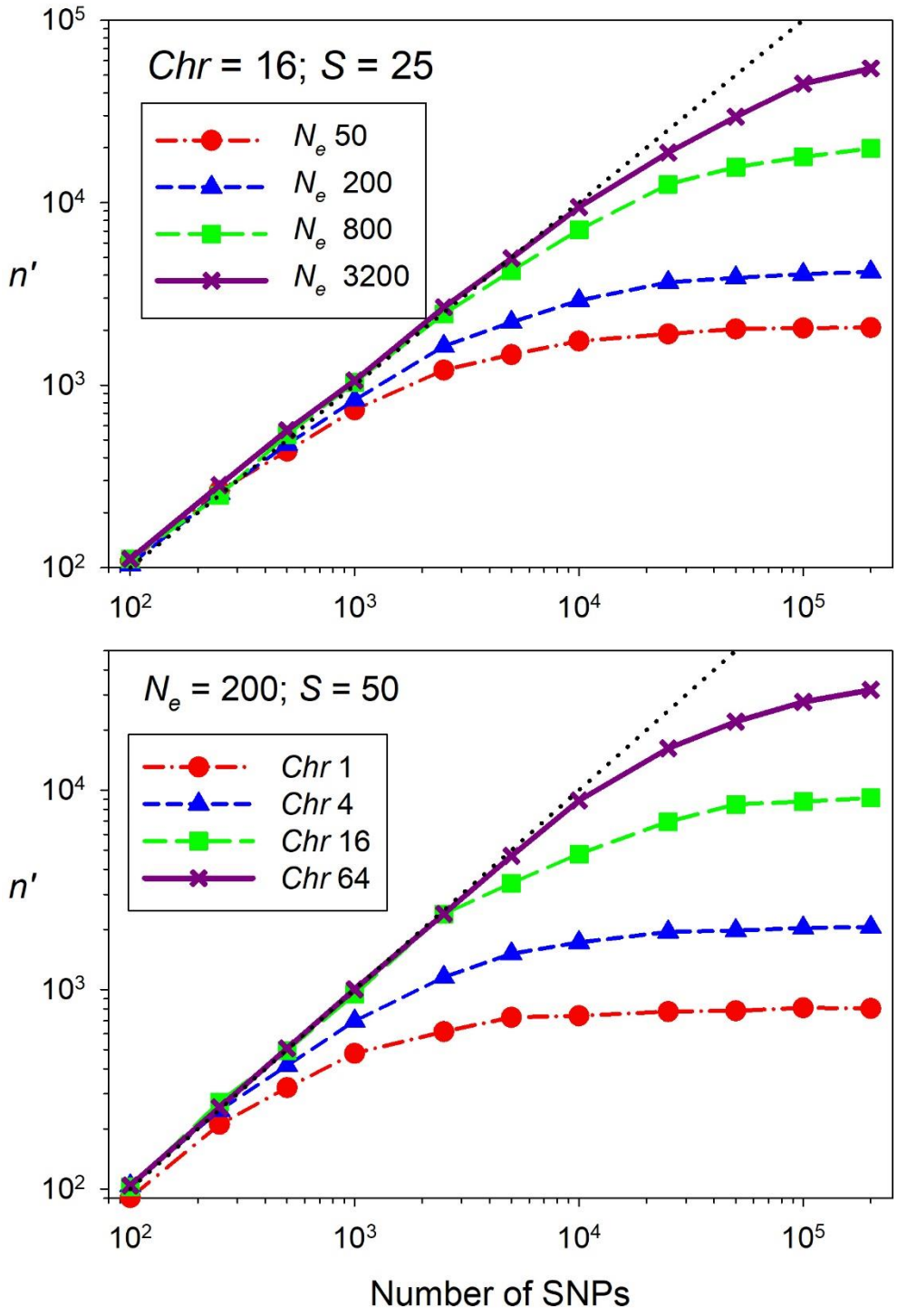

Figure 7. Effects of N_e (top panel, with number of chromosomes set at 16) and *Chr* (bottom panel, with N_e set at 200) on the effective degrees of freedom (*n*') for $\hat{F}_{ST(n)}^{Nei}$ computed between pairs of populations. Black dotted line represent *n*' = *n* = the number of SNPs.

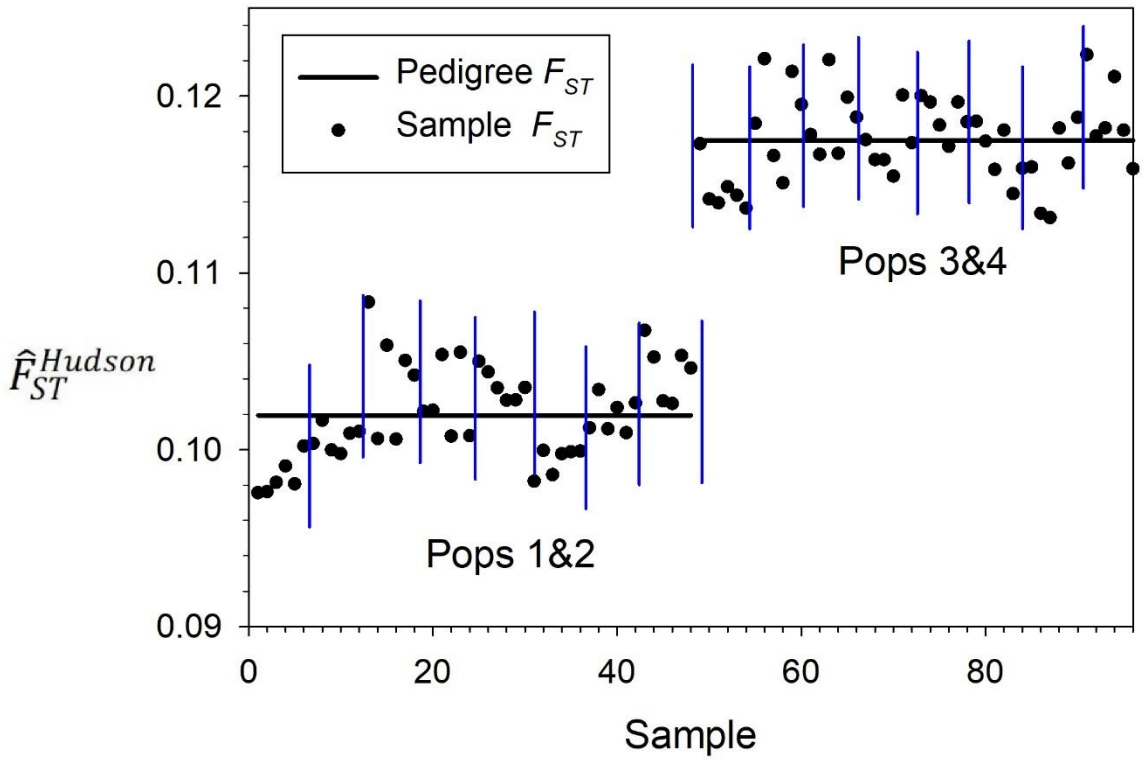

Figure 8. Effects of pedigree on variation in mean $\hat{F}_{ST(n)}^{Hudson}$. For each of two, 2-population pedigrees, 8 replicate samples (demarcated by vertical blue lines) were taken of 100 individuals. These results are for simulations with N_e = 200 and 4 chromosomes. Sampled individuals were drawn hypergeometrically from the N_e individuals in the final generation. For each sample, six mutational replicates generated non-overlapping sets of n = 5000 SNP loci. Solid horizontal lines ("Pedigree F_{ST}") represent mean $\hat{F}_{ST}^{Hudson}$ across all 8x6 = 48 replicates within each pedigree. The first set of samples shows results for comparison of daughter populations 1 and 2 and the second set of samples shows results for comparison of daughter populations 3 and 4, all derived from the same ancestral population.

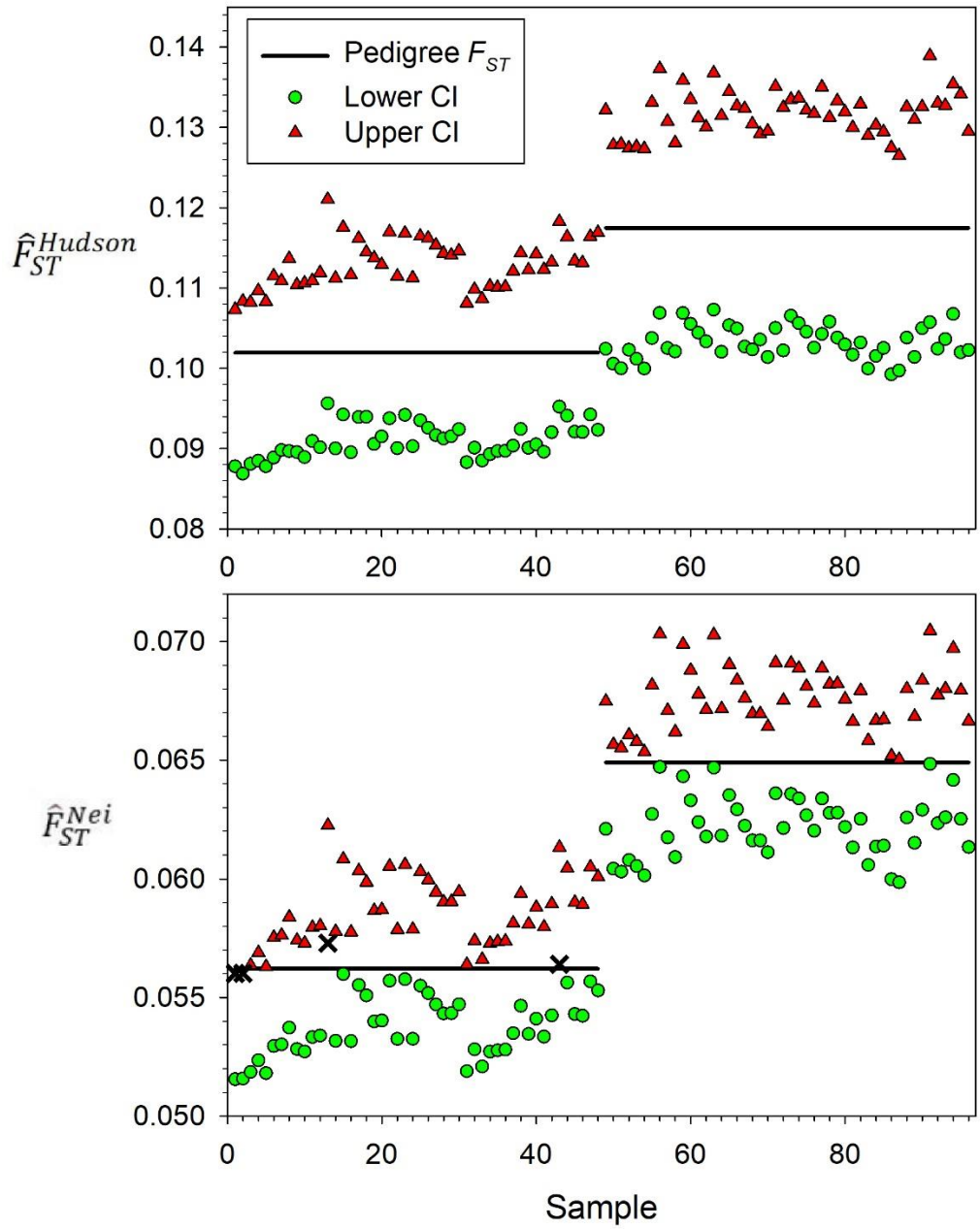

Figure 9. Coverage of 90% confidence intervals (CIs) around F_{ST} estimators for the population pedigrees and samples shown in Figure 8 (N_e = 200; *Chr* = 4; *n* = 5000; *S* = 100). Top: CIs generated from block-jackknife estimates of var($\hat{F}_{ST}^{Hudson}$). Bottom: CIs generated based on *n*' for $\hat{F}_{ST}^{Nei}$ estimated from this study. CI coverage is evaluated with respect to mean $\hat{F}_{ST}^{Hudson}$ or mean $\hat{F}_{ST}^{Nei}$ across all replicates within each pedigree ("Pedigree F_{ST}", horizontal lines). The black X symbols indicate an upper (or lower) bound that was below (or above) the mean pedigree F_{ST}.