Received: 31 May 2023 | Revised: 3 August 2023 | Accepted: 22 August 2023

DOI: 10.1002/1438-390X.12167

Population Ecology WILEY

SPECIAL FEATURE

Ecological Perspectives of Pedigree Reconstruction with Genome-wide Data

Estimating current effective sizes of large populations from a single sample of genomic marker data: A comparison of estimators by simulations

Jinliang Wang

Institute of Zoology, Zoological Society of London, London, UK

Correspondence
Jinliang Wang, Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, UK.
Email: jinliang.wang@ioz.ac.uk

Abstract

Genome-wide single nucleotide polymorphisms (SNPs) data are increasingly used in estimating the current effective population sizes (N_e) for informing the conservation of endangered species and guiding the management of exploited species. Previous assessments of sibship frequency (SF) and linkage disequilibrium (LD) estimators of N_e focused on small populations where genetic drift is strong and thus N_e is easy to estimate. Genomic single nucleotide polymorphism (SNP) data provide ample information and hold the potential for application of these estimators to large populations where genetic drift is rather weak and thus N_e is difficult to estimate. In this study, I simulated very large populations and sampled a widely variable number of individuals (genotyped at 10,000 SNPs) for estimating N_e by both SF and LD methods. I also considered the more realistic situation where a population experiences a bottleneck, and where marker data suffer from genotyping errors. The simulations show that both SF and LD methods can yield accurate N_e estimates of very large populations when sampled individuals are sufficiently numerous. When n is much smaller than N_e, however, N_e estimates are in a bimodal distribution with a substantial proportion of the estimates being infinitely large. For a population with a bottleneck, LD estimator overestimates and underestimates the N_e of the parental population from samples taken at and after the bottleneck, respectively. LD estimator also overestimates N_e substantially when applied to data suffering from allelic dropouts and false alleles. In contrast, SF estimator is unbiased and accurate when populations are changing in size or markers suffer from genotyping errors.

KEYWORDS

effective population size, genetic drift, genomic markers, inbreeding, linkage disequilibrium, sibship

1 | INTRODUCTION

Effective population size (Caballero, 1994; Wright, 1931) is a pivotal parameter in population genetics. It determines the strength of genetic stochasticity (measured by rates of inbreeding and genetic drift) of a population and affects the efficacy of all systematic evolutionary forces, such as selection and migration, acting on a population (Crow & Kimura, 1970). As a result, it determines the genetic diversity and the evolutionary potential of a population. Populations of conservation concern have a small N_e and are thus vulnerable to the loss of genetic diversity due to drift, to the decline of fitness due to inbreeding and inbreeding depression, to the accumulation of mutational load, and to the loss of adaptability (Frankham, 2005). Measuring and monitoring the N_e of an endangered species is thus critically important in assessing the risks of extinction both in the short terms and in the long run, and in gauging the effectiveness of conservation managements (Frankham et al., 2014). Similarly in the management of exploited species such as some marine fishes, monitoring N_e is helpful in assessing the population density and in making management or harvest decisions (Marandel et al., 2019).

Quite a few methods have been developed and applied to estimating N_e from data of various genetic markers (Leberg, 2005; Luikart et al., 2010; Wang, 2005; Wang et al., 2016), first enzyme polymorphisms, then microsatellites and now SNPs. For the contemporary or current N_e, which is relevant for conservation and management of populations (Luikart et al., 2010), the most well-studied and widely applied estimators are temporal (TM) methods (Krimbas & Tsakas, 1971; Nei & Tajima, 1981; Wang, 2001; Waples, 1989), sibship frequency (SF) methods (Wang, 2009), and linkage disequilibrium (LD) methods (Hill, 1981; Waples, 2006). TM methods exploit the changes in marker allele frequencies between two or more temporally spaced samples as information to estimate the genetic drift (N_e) occurred during the sampling period (Nei & Tajima, 1981). SF methods use the frequency of estimated full- and half-sibling pairs in a sample of individuals taken at random from a population as information for the N_e of the population at the sampling point of time (Akita, 2020; Wang, 2009). LD methods estimate the average LD between pairs of loci from the genotype data of a single sample of individuals as information for N_e (Hill, 1981). Compared with SF and LD methods which use a single sample, TM methods are more demanding because they require at least two samples taken from the same population separated by at least one generation. For this reason, single sample-based methods, especially LD methods, have gained more popularity over the recent years (Marandel et al., 2019).

The strength of genetic drift and inbreeding in a population and thus the strength of signal in a sample drawn from the population is reciprocally proportional to the N_e of the population. This is true no matter the signal is in the form of LD (Hill, 1981), sib frequency (Wang, 2009), or temporal allele frequency changes (Nei & Tajima, 1981). Therefore, while N_e can be estimated with ease (i.e., with low sampling effort) and with reasonably high accuracy for small populations, it becomes tremendously difficult to estimate with satisfactory accuracy for moderately large populations (say, N_e in the thousands) even when an extremely high sampling intensity (i.e., sampling many individuals and many markers) is applied (e.g., Marandel et al., 2019). Although these N_e estimation approaches are primarily developed for and applied to conservation populations which are typically small, they are also increasingly exploited for understanding the demography of managed or harvest populations which can be large (e.g., Delaval et al., 2022; Marandel et al., 2019; Nadachowska-Brzyska et al., 2021). Marine fish populations are usually very large. To understand their genetic status (e.g., population trends whether shrinking, stable, or increasing) and thus to inform management or harvest decisions, population N_e is increasingly measured from marker data using the LD approach (e.g., Hoarau et al., 2005; Laurent & Planes, 2007; Poulsen et al., 2005). Simulation studies (Macbeth et al., 2013; Marandel et al., 2019) confirm that large populations pose a serious challenge to the LD approach. To reduce negative N_e estimates (usually interpreted as infinitely large) to a reasonably low level, an untypically large sample of individuals (in the thousands) is necessary (Marandel et al., 2019). Unfortunately, no empirical or simulation studies are conducted to investigate how the SF approach performs in the challenging situation of large populations. Current knowledge of the behavior and performance of this approach is based on simulations (Wang, 2009, 2016; Waples, 2021) of small populations. It is urgently needed to know whether the SF approach is similar to LD approach or not in both power and accuracy for estimating the N_e of large populations.

The relative strength of signal of inbreeding and genetic drift occurring in a population depends not only on the N_e of the population, but also on the sampling intensity in terms of the numbers of markers (and polymorphisms) and individuals. Under certain situations, these two numbers could compensate each other for N_e inference accuracy. In the genomic era, millions of SNPs can be obtained virtually from any species by various Next Generation Sequencing (NGS) techniques. In contrast, due to logistics and cost, the sample size of individuals is usually limited to a few hundreds. Is it possible to obtain accurate N_e estimates of very large populations (say, N_e in millions)

by the LD and SF methods from the genomic marker data of a few hundreds of individuals? Although some studies investigated the possibility of applying LD methods to large populations, they assume a much smaller number of markers (Macbeth et al., 2013; Marandel et al., 2019). It is unclear whether the use of genomic marker data helps or not, and whether SF methods fare similarly to LD methods or not in the difficult situation of large populations.

Genomic marker data contain not only ample information, but also copious noises (errors). SNP genotypes obtained from NGS data could suffer from a high rate of errors caused by multiple factors, including base-calling and alignment errors (Nielsen et al., 2011). Furthermore, nowadays NGS studies usually apply low-coverage sequencing to sequence a reasonably large sample of individuals at affordable cost. However, low-coverage sequencing results in a high probability that only one of the two chromosomes of a diploid individual is sampled at a specified site, producing a false homozygote while it is a heterozygote (Nielsen et al., 2011). These SNP genotyping errors are similar to allelic dropout errors of microsatellites generated during polymerase chain reaction, occurring often when DNA quantity and quality are low (Pompanon et al., 2005). It is apparently too naïve and unrealistic to ignore the abundant genotyping errors of genomic data in evaluating the performance of any N_e estimation methods. Among studies investigating the estimation of current N_e from genotype data, only Wang (2016) compared the robustness of different methods to the presence of allelic dropout errors. However, the study did not consider genomic marker data, and did not consider genotyping errors other than allelic dropouts (e.g., false alleles). It is unclear how different N_e estimation methods perform when genomic data with typing errors are used.

In this study, I use simulations to evaluate the accuracy of LD and SF methods coupled with genomic marker data in estimating the current effective size of very large populations. I investigate the minimal numbers of sampled individuals required by the two methods to obtain satisfactory N_e estimates from genomic marker data. I also compare the performances of the two methods in the realistic situation of the presence of allelic dropouts and other errors in genome-wide genotype data, and in the scenario of a population changing in effective size. The results are helpful in understanding the prospects of estimating N_e of large populations from genomic data, in optimizing sampling designs (or intensities) of marker-based N_e estimation studies, and in interpreting the analysis results from such studies.

2 | METHODS

In this section, I briefly describe the two widely used methods, LD and SF, for estimating the current N_e of a population from the multilocus genotypes of a single sample of individuals drawn from the population. I then introduce the procedures and parameter combinations used in simulations, and describe the methods used to assess the performance (biasness and accuracy) of the methods.

2.1 | Single-sample N_e estimators

Both LD and SF methods use the same data, the multilocus genotypes of a single sample of individuals drawn at random from a population, to estimate the N_e of the population at the sampling time point. More precisely, SF methods estimate the N_e of a population at the parental generation of the sampled individuals, while LD methods estimate the average N_e of a population at and a few generations before the parental generation of the sampled individuals. This is because the LD estimated from a sample of individuals reflects the cumulative contributions of drift occurred at the parental, grandparental, great grandparental generations, and so on. Although the estimated LD indicates predominantly the N_e at the parental generation, it also signifies the N_e at grandparental and more remote ancestral generations. Therefore, N_e estimated from LD by assuming a constant demography (e.g., Waples, 2006) is a weighted average of the effective population sizes at the parental and earlier generations, the exact number of generations involved and their weights on the final estimated N_e being unknown but dependent on the recombination rates between the two markers of each of many pairs of loci used for calculating LD. When N_e is indeed constant over generations, the linkage relationship among markers becomes irrelevant as the weighted average N_e is equivalent to the parental N_e, no matter how many generations are involved. However, when N_e is changing over the past few generations before sampling, then the LD-based N_e estimates become difficult to interpret, even in the simple case of estimates from unlinked markers. In all previous empirical and simulation studies of LD methods using unlinked markers, N_e was explicitly or implicitly assumed constant over time.

In this study, I use the LD-based N_e estimator derived by Hill (1981). It is improved by Waples (2006) for the special case of unlinked markers to reduce biasness caused by small sample sizes, and implemented in the software NeEstimator by Waples and Do (2008, 2010). The improved estimator was assessed by simulations for accuracy (e.g., Waples & Do, 2008) and for robustness to population subdivision (Waples & England, 2011) and overlapping generations (Waples et al., 2014). It was also compared with the SF methods for accuracy when populations were small (Wang, 2016; Waples, 2021). The present study investigates the biasness and accuracy of this improved estimator when populations are large and

changing in N_e, and when genomic markers realistically have genotyping errors. In simulating and analyzing the data by NeEstimator, markers are assumed unlinked and the default parameter setting was adopted. Markers with the minor allele frequency below 0.05 were removed as suggest by Waples and Do (2010) to minimize sampling bias, and negative estimates of N_e was interpreted as infinitely large populations. Because of this filtering, the number of markers actually used by the LD estimator is slightly smaller than that used by the SF estimator which uses all markers without filtering out those with rare alleles.

The SF estimator, as derived by Wang (2009), uses the frequency of sibling dyads inferred from a sample of individuals drawn at random from a population as information for estimating the N_e of the population. In the simple case of a random mating population, the estimator (eq. 10 of Wang, 2009) simplifies to

$$N_e = \frac{2n(n-1)}{n_{HS} + 2n_{FS}}, \quad (1)$$

where n_{HS} and n_{FS} are the number of half-sib pairs and the number of full-sib pairs found in a sample of n individuals. The SF estimator is essentially analogous to the mark–recapture (MR) method used in estimating population census size (Luikart et al., 2010; Wang, 2016; Waples & Feutry, 2022). The SF approach captures and recaptures nuclear families represented by siblings who share a single or a pair of parents, and the recapture rate (estimated by sibling frequency) informs the effective number of families or N_e of the population. Like the MR method for estimating census size, a low recapture rate means a population with many families or a large N_e, and a high recapture rate signifies a population with few families or a small N_e. The accuracy of SF method is determined by the sample size of individuals collected for genotyping, and by the sample size of markers (i.e., number of loci) genotyped for each sampled individual. The latter sample size affects sibship estimation errors (SEEs), with more markers yielding more accurate sibling inferences (Wang, 2004). With the use of genomic markers as is the case of this study, marker data are so informative that virtually a sibship analysis would be 100% accurate (as shown in this study). The former sample size determines the sibship sampling errors (SSEs), the errors of sample SF as an estimate of population SF. The larger is the number of sampled individuals, the smaller is the expected deviation of sample SF from the population SF.

SEEs are affected by the sample sizes of both individuals and markers, as well as by quite a few other factors such as mating system and data quality (i.e., genotyping error rates and data missing rates) (Wang, 2004). However, with the use of an increasing number of markers, SEEs should always decrease to zero, regardless of the other factors. SSEs are predominantly determined by the number of sampled individuals, n. Suppose the population sibling frequency is p, the number of sibling dyads included in a sample of n individuals, x, follows roughly a binomial distribution with parameters $n(n-1)/2$ and p. The variance of the relative sample sibling frequency estimate, $\widehat{p}/p = x/(pn(n-1)/2)$, is thus $(1/p-1)/(n(n-1)/2)$, which means this sampling variance decreases roughly quadratically with sample size n and increases with an increasing N_e (a correspondingly decreasing p). This means SSEs are highly sensitive to the sampling intensity of individuals, and large N_e (i.e., small p) is difficult to estimate accurately. A much larger sample size n is required for an accurate estimate of the N_e of a large population.

I use the SF method (Wang, 2009) implemented in the software package COLONY (Jones & Wang, 2010) to assign sibship and thus to estimate N_e from the simulated data (see below). There are many alternative parameter options built in the software for a fine-tuned sibship analysis, such as different sibship size priors. However, when marker information is sufficient, such as that of genomic SNPs as is the case of the present study, these alternative parameter options become irrelevant and sibship is always accurately assigned irrespective of the parameter settings. In this simulation study, I use the default parameter setting of the software.

2.2 | Simulation procedure

I simulated an ideal Wright–Fisher population (Fisher, 1930; Wright, 1931) with a constant size of N individuals at each discrete generation, of a monoecious diploid species, with random mating including selfing at the rate of $1/N$, with no selection (i.e., all parents have an equal chance of producing offspring such that the number of gametes contributed by a parent follows a binomial distribution B($2N$, $1/N$)) and with no mutations. For such an ideal population, we have $N_e = N$ (Wright, 1931). In most simulations, N is assumed constant over generations, but a bottleneck at which N is drastically reduced is also considered (below).

Individuals in the founder generation were taken from an infinitely large base population such that they are non-inbred and unrelated. The genotype of a founder individual at each of L unlinked loci was drawn independently (i.e., no LD) from the base population, which is assumed to be in Hardy–Weinberg equilibrium at a diallelic locus with allele frequencies following a uniform Dirichlet distribution. Throughout this study, I assume $L = 10{,}000$ unlinked SNP markers are genotyped for estimating N_e. Of course, some of the $L = 10{,}000$ markers must be linked for any real species. However, sibship analysis and thus N_e estimation by SF method is barely affected by linkage and LD (Wang &

Santure, 2009), and unlinked marker pairs (say, loci on different chromosomes) can be selected for LD-based N_e estimation when the linkage groups of the markers are known.

Starting from the founder generation, a number of $T = 10$ generations were simulated before a sample of n offspring was taken at random from the population at generation $T + 1$ for genotype analysis and N_e estimation. At each generation t, a zygote (individual) is formed by combining a male gamete and a female gamete taken at random from a father and a mother at generation $t - 1$, respectively. The individual is thus from self-reproduction and outbreeding when the uniting gametes come from the same parent and different parents, respectively. The genotype of a gamete is generated from the parental genotype following Mendel's laws of segregation and independent assortment.

The simulation procedure described above was checked to ensure it works as expected. In a preliminary study, I also considered more than $T = 100$ burn-in generations and found that larger T value did not change the results. This is expected as both SF and LD methods do not rely on the equilibrium between mutations and drift which requires many generations, on the order of N or $1/u$ (where u is mutation rate) whichever is larger, to attain. I also recorded the simulated pedigrees and used them to calculate the realized pedigree-based estimate of N_e. As expected, the estimated N_e is equivalent to the theoretical value of $N_e = N$. I also calculated the F_{ST} from the multilocus genotype data at generations 0 and T, and found it is equal to the expected value of $F_{ST} = 1 - (1 - 1/(2N))^T$. All of these diagnostics verify that the simulation program works properly as expected.

2.3 | Simulation parameter combinations

Five simulations were designed, by choosing proper parameter combinations, and conducted to investigate the performances of SF and LD methods in using genomic marker data for estimating the current N_e under some neglected situations.

Simulation 1 considers a wide range of N_e (10,000, 20,000, 40,000, 60,000, 80,000, 100,000) and a wide range of sample size n (number of individuals) (20, 40, 80, ..., 1280; $= 20 \times 2^m$ for $m = 0, 1, 2,..., 6$) to understand the biasness and accuracy of SF and LD methods. In particular, the simulation is used to investigate the performances of the two methods when populations are extremely large and sample sizes are also large (but not deviating far from the reality).

Simulation 2 considers a relatively large N_e (1000) and a wide range of sample sizes n (50, 100, 200, ..., 3200; $= 50 \times 2^m$ for $m = 0, 1, 2,..., 6$) to investigate the performance of SF and LD estimators when sample sizes are larger than N_e. Previous simulation studies have not dealt with the situation in which N_e is large and n is larger than N_e. The incomplete parameter space considered in previous studies produced incomplete conclusions (see below) regarding the relative accuracies of SF and LD methods.

Simulation 3 checks the performances of SF and LD methods when a population experiences a bottleneck. The population has a N_e of 1000 at generations 1 to $T - 2$, then it crashes to $N_e = 500$ at generation $T - 1$, and recovers to the pre-bottleneck value of $N_e = 1000$ at generation T. A sample of $n = 100$, 200, 400, 800, 1600 individuals is sampled at generation $T - 1$ when the population is crashed and at generation T when the population has recovered. The sampled individuals are genotyped at 10,000 SNP loci for the estimation of N_e by SF and LD methods.

Simulation 4 investigates the impact of allelic dropouts, a common problem for SNP data derived from low-coverage NGS, on the LD and SF estimators of N_e. A wide range of allelic dropout rates (0, 0.01, 0.02, ..., 0.32) at each SNP locus is considered for its effects on the biasness and accuracy of SF and LD methods, when sample size n is fixed at 320 and N_e fixed at 10,000.

Simulation 5 examines the impact of false alleles, which may plague SNP genotypes from NGS due to low coverage or other causes such as low DNA quality and quantity, on the LD and SF estimators of N_e. Like allelic dropouts, I consider the effects of a wide range of false allele rates (0, 0.01, 0.02, ..., 0.32) at each SNP locus on the biasness and accuracy of SF and LD methods, when sample size n is fixed at 320 and N_e fixed at 10,000.

For each simulation, a number of 160 replicate datasets were generated and analyzed by SF and LD methods in parallel by 160 cores on a linux cluster, as both methods require substantial computational time for analyzing genomic data ($L = 10{,}000$). The use of parallelization and a large linux cluster makes this large-scale simulation study possible.

2.4 | Accuracy assessment

While SF estimates of N_e are always positive numbers, LD estimates of N_e can be negative. This happens most often with large populations where the actual N_e value is large and the signal of genetic drift is weak. In such a situation, the observed LD is so small that it can be accounted for by sampling alone and no drift needs to be incurred. Therefore, negative estimates of N_e by LD methods are converted to infinitely large values before being assessed for accuracy.

For a given simulated value of N_e, $m = 160$ replicate datasets were generated and analyzed by SF and LD methods to obtain m estimates of N_e, $\widehat{N}_e$, by each estimator. Ideally, these $\widehat{N}_e$ values should be identical to the simulated N_e value. Realistically, however, $\widehat{N}_e$ values deviate from the simulated N_e value due to sampling and estimation errors. The quality of an N_e estimator can be assessed by its bias, B, and variance, V, of $1/\widehat{N}_e$ rather than $\widehat{N}_e$ (Wang & Whitlock, 2003). I calculate B and V by

$$B = \frac{1}{\widehat{N}_e^H} - \frac{1}{N_e},$$

$$V = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{\widehat{N}_e^H} - \frac{1}{\widehat{N}_{ei}}\right)^2,$$

respectively, where $\widehat{N}_{ei}$ is the N_e estimate for the ith ($i = 1, 2, ..., m$) replicate dataset, N_e is the simulated parameter value, $\widehat{N}_e^H = \left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\widehat{N}_{ei}}\right)^{-1}$ is the harmonic mean of $\widehat{N}_{ei}$. Hereafter, the "mean" estimate of N_e always refers to this harmonic mean.

The overall performance (accuracy) of an estimator can be measured by the deviations of estimated values from the true (simulated) value of $1/N_e$, calculated by root mean squared error (RMSE),

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{\widehat{N}_{ei}} - \frac{1}{N_e}\right)^2}.$$

It is determined by both the bias, B, and the variance, V, of an estimator, as it can be shown that $RMSE = \sqrt{B^2 + V}$. Hereafter, "accuracy" refers to the level of agreements between estimated, $\frac{1}{\widehat{N}_{ei}}$, and true (simulated), $\frac{1}{N_e}$, parameter values, quantified by RMSE. The smaller is the value of RMSE, the more accurate is an estimator. The maximal accuracy occurs when the estimator is unbiased, $B = 0$, and highly precise such that $V = 0$, and therefore RMSE = 0. In this simulation study, I report $\widehat{N}_e^H$ (in comparison with simulated parameter value N_e) and RMSE to measure the bias and accuracy of each estimator for each simulated parameter combination with $m = 160$ replicates.

3 | RESULTS

3.1 | Simulation 1: Large populations

Both SF and LD estimates of N_e are biased when sample size n is smaller than 320, but become almost unbiased as n increases above 320 (Figure 1b). This pattern is consistent across the range of simulated N_e from 10,000 to 100,000. Overall, SF is less biased than LD. While SF always underestimates N_e when it is biased at a small n, LD may make both underestimates and overestimates of N_e depending on sample sizes. It underestimates N_e when n is below 40 and overestimates N_e when n is in the range of 40–320. This bias pattern is consistent for all simulated N_e values.

A close examination of N_e estimates shows that, when n is small and thus both LD and SF estimators are biased, the estimates are in a bimodal distribution (Figure 2). These estimates are either infinitely large or much smaller than the simulated value of N_e. For the case of $N_e = 40{,}000$ and $n = 80$ as an example, the LD and SF

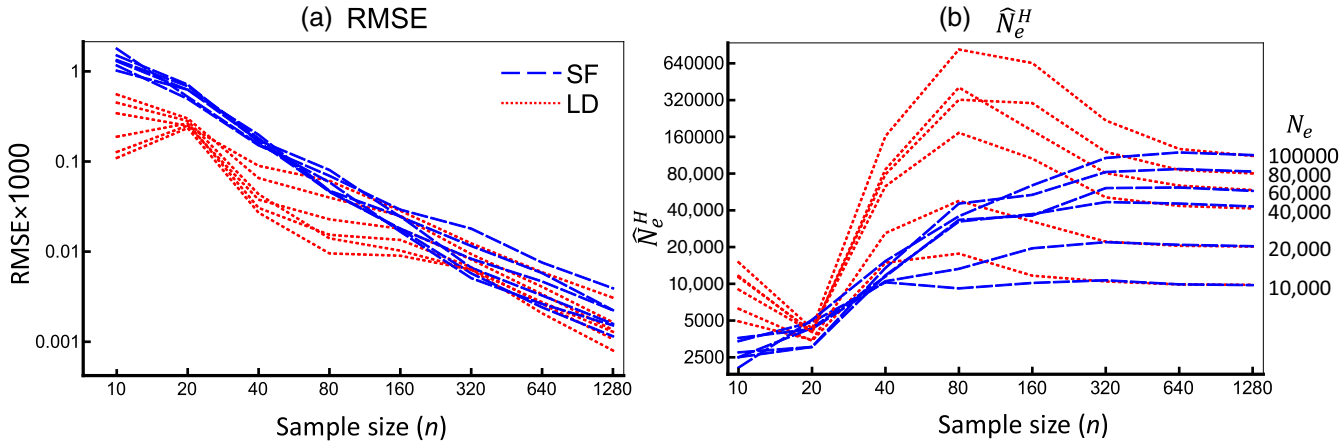

FIGURE 1 RMSE (a) and $\widehat{N}_e^H$ (b) of estimators SF and LD for populations with $N_e = 10{,}000$, 20,000, 40,000, 60,000, 80,000, and 100,000 using a sample with $n = 10$, 20, 40, 80, 160, 320, 640, and 1280 individuals. Each individual is genotyped at 10,000 diallelic loci. RMSE or $\widehat{N}_e^H$ of each estimator for each simulated N_e is represented by a line (dotted line in red color for LD and dashed line in blue color for SF). While these lines are identifiable for each simulated N_e for the plot of $\widehat{N}_e^H$ (shown on the right vertical frame line), they are unidentifiable for the plot of RMSE. [Color figure can be viewed at wileyonlinelibrary.com]

estimators yield 105 and 90 infinitely large estimates of N_e (out of a total number of 160 replicate estimates), respectively. For all of the 90 replicate datasets where SF estimator gives an infinite N_e estimate, the LD estimator also produced infinite N_e estimates. LD estimator also generated infinite N_e estimates in an additionally 15 replicate datasets. Part of the reason that SF yields on average a lower $\widehat{N}_e^H$ than LD when n is small is because SF produces fewer infinite estimates than LD. As n increases, the frequency of infinite N_e estimates declines rapidly for both SF and LD estimators (Figure 2). When $n = 320$, no infinite N_e estimates are observed for both estimators.

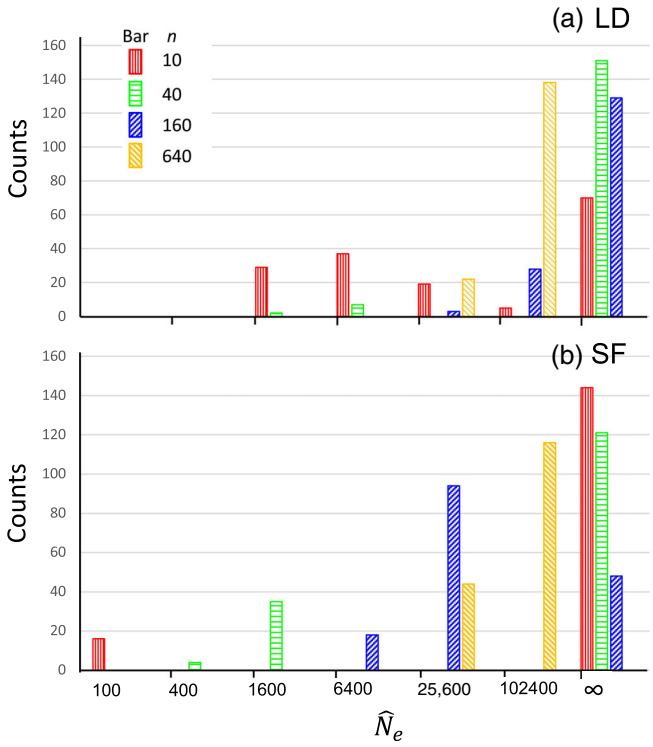

FIGURE 2 Frequency distribution of effective population size ($\widehat{N}_e$) estimated by (a) the LD method and (b) the SF method from 160 replicate datasets simulated with true $N_e = 10^6$. The sample size (number of sampled individuals, n) is 10 (red color), 40 (green color), 160 (blue color), or 640 (orange color), and each sampled individual is genotyped at 10,000 diallelic SNP loci. [Color figure can be viewed at wileyonlinelibrary.com]

When n is small (i.e., $n < 160$) and thus both LD and SF estimators are biased, LD estimator is more accurate than SF estimator (Figure 1a) because it gives less dispersed estimates of N_e (Figure 2). With an increasing n such that both estimators are rapidly becoming unbiased, they tend to have similar overall accuracy as measured by RMSE. Numerically the RMSE of LD estimator is still slightly smaller than that of SF estimator at the largest sample size of $n = 1280$, but the difference is hardly detectable (Figure 1a).

3.2 | Simulation 2: Large sample sizes

Simulation 2 considered a much larger range of sample sizes, n, with the lower bound of n much smaller than N_e and the upper bound of n much larger than N_e, which is assumed to be 1000. Similar to the results shown in Figure 1b, LD estimator is more biased than SF estimator when sample size n is small (Figure 3b). LD estimator becomes almost unbiased when $n > 200$, while SF estimator is unbiased even with n as small as 50.

When n is small, LD estimator is more accurate than SF estimator (Figure 3a). The difference in RMSE between the two estimators decreases with an increasing

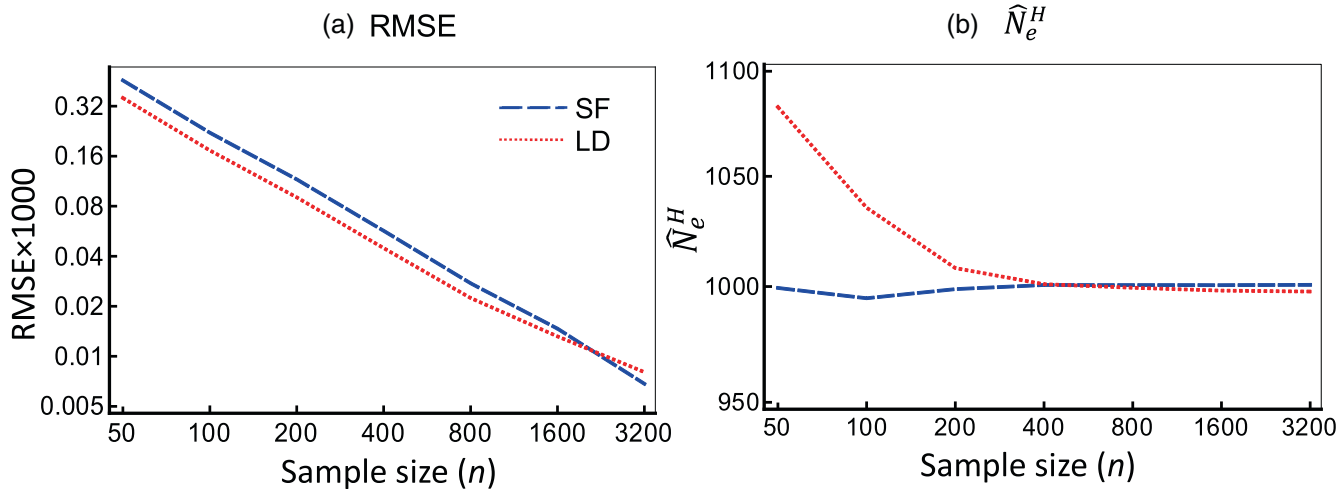

FIGURE 3 RMSE (a) and $\widehat{N}_e^H$ (b) of estimators SF and LD for a population with $N_e = 1000$ using a sample with $n = 50$, 100, 200, 400, 800, 1600, and 3200 individuals. Each individual is genotyped at 10000 diallelic loci. RMSE or $\widehat{N}_e^H$ is represented by a red dotted line for LD and by a blue dashed line for SF. [Color figure can be viewed at wileyonlinelibrary.com]

sample size n. When $n > 2500$, SF estimator becomes more accurate than LD estimator.

3.3 | Simulation 3: Population bottleneck

LD estimator overestimates N_e at a population bottleneck and underestimates N_e after a population bottleneck (Figure 4b), as expected. In both cases, the bias is substantial and is consistent across different sample sizes. In contrast, SF estimator yields unbiased N_e estimates for the parental population both at the bottleneck and after the bottleneck.

For a population with a nonconstant and varying N_e, the SF estimator gives an unbiased and accurate estimate of N_e at any generation. The accuracy advantage of SF over LD estimator increases with an increasing sample size n (Figure 4a). While the accuracy of SF estimator increases log linearly with n, the accuracy of LD estimator asymptotes at about $n = 400$. Above $n = 400$, the accuracy of LD estimator no longer increases with n. This is because the LD estimator's RMSE is mainly determined by the bias rather than the sampling variance when n is above 400. While a larger n leads to a smaller sampling variance, it has no effects on bias when $n > 400$.

3.4 | Simulation 4: Allelic dropouts

SF estimator is almost unbiased in the whole range of allelic dropout rate, from 0.01 to 0.32 (Figure 5b). In contrast, LD estimator tends to overestimate N_e substantially when dropout rate is not small (i.e., >0.04). Similarly, the accuracy as measured by RMSE of SF estimator is nearly constant with an increasing allelic dropout rate (Figure 5a), but the accuracy of LD

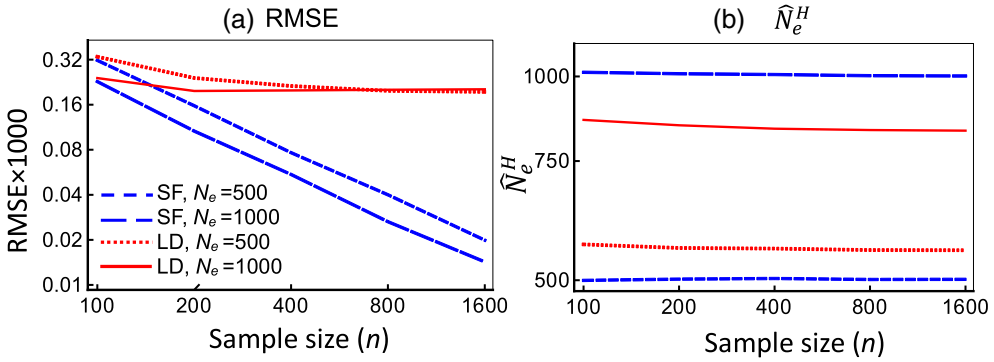

FIGURE 4 RMSE (a) and $\widehat{N}_e^H$ (b) of estimators SF and LD using a sample with $n = 100$, 200, 400, 800, 1600 individuals taken from a population at a bottleneck (at generation $T-1$, $N_e = 500$) and immediately after the bottleneck (at generation T, $N_e = 1000$). Each individual is genotyped at 10,000 diallelic loci. RMSE and $\widehat{N}_e^H$ are represented by red dotted lines (at bottleneck, $N_e = 500$) or red solid lines (after bottleneck, $N_e = 1000$) for LD, and by blue short-dashed lines (at bottleneck, $N_e = 500$) or blue long-dashed lines (after bottleneck, $N_e = 1000$) for SF. [Color figure can be viewed at wileyonlinelibrary.com]

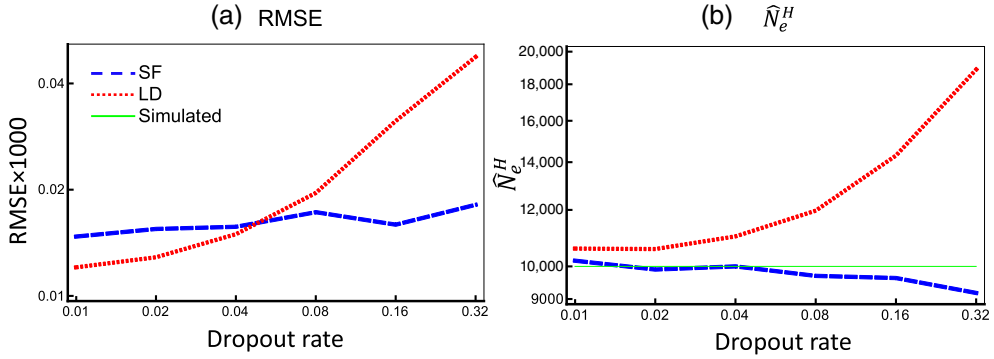

FIGURE 5 RMSE (a) and $\widehat{N}_e^H$ (b) of estimators SF and LD as a function of allelic dropout rate at each of 10,000 SNP loci. A number of $n = 320$ individuals were taken from a population with a constant effective size of $N_e = 10{,}000$, and each individual was genotyped at 10,000 diallelic loci. RMSE and $\widehat{N}_e^H$ are represented by red dotted lines for LD, and by blue dashed lines for SF. The simulated $N_e = 10{,}000$ is represented by a thin green line. [Color figure can be viewed at wileyonlinelibrary.com]

estimator decreases consistently with an increasing allelic dropout rate.

3.5 | Simulation 5: False alleles

When applied to data with false alleles, the SF and LD estimators show behaviors (Figure 6) similar to those observed with data having allelic dropouts (Figure 5). However, at the same rate, false alleles have a larger impact than dropouts on both the bias (Figure 6b) and the accuracy (Figure 6a) of LD estimator. In the whole range of false allele rate from 0.01 to 0.32, the SF estimator remains unbiased and has a RMSE almost constant, showing that SF is highly robust to the prevalence of false alleles.

4 | DISCUSSION

Large populations experience, by definition, very weak genetic drift. As signals and measurements of genetic drift, the squared correlation of allele frequencies at two diallelic loci is expected to be in the order $1/(2N_e)$ (Hill, 1981), and the frequency of sibling pairs is also in this order (Wang, 2009). It is easy to understand that it becomes very difficult to capture such weak signals by any marker-based N_e estimator. The current effective sizes of large populations are therefore challenging to estimate accurately even with genomic data, no matter which method is adopted. Picking up the weak drift signal and thus obtaining high-quality estimates of N_e requires large samples sizes of both loci (markers) and individuals. In this study, I showed by simulations that both SF and LD methods can be used to estimate the effective sizes of large populations (with N_e up to 10^5) unbiasedly and accurately using genomic marker data (10,000 SNPs), provided the sample sizes of individuals are sufficiently large ($n > 320$). However, when the number of sampled and genotyped individuals is not large (i.e., $n < 320$), both SF and LD estimators are biased and have a poor accuracy (Figure 1). The LD estimators looks more biased but seems to have a lower variance than SF estimator when sample size is small (relative to the N_e of the population). Both estimators show a bimodal distribution (Figure 2), with a substantial proportion of the estimates being infinitely large while the remaining proportion of estimates being in general much smaller than N_e. The same bimodal distribution was also observed for the LD estimator in the simulation studies of Marandel et al. (2019) and Waples et al. (2016) when much fewer markers (100 or 200 SNPs) were used in the estimation. In such a situation of bimodal distribution, pursuing the unbiasedness of an estimator is meaningless as any single one estimate will depart from the truth substantially.

Why does an estimator produce a bimodal distribution of N_e estimates when n is much smaller than the true N_e? This is easily understood by considering the SF estimator in a numerical example. For a monogamous species (i.e., no half-siblings) with $N_e = 10{,}000$ and a number of $n = 5$ sampled individuals, for example, there are two possibilities of the sibship configuration in the five sampled individuals. With a high probability, p_1, is the possibility that the $n = 5$ individuals contain no individuals sharing a pair of parents (i.e., $n_{FS} = 0$). In such a case, the estimated N_e (Equation 1) is $\widehat{N}_e = \frac{2\times5\times4}{0+2\times0} = \infty, \widehat{N}_{e1} = \infty$. With a small probability, $p_2 = 1 - p_1$, is the possibility that the $n = 5$ individuals contain a single pair of full siblings (i.e., $n_{FS} = 1$) which results in $\widehat{N}_e = \frac{2\times5\times4}{0+2\times1} = 20$ calculated by Equation (1), $\widehat{N}_{e2} = 20$. In this example with extremely small n relative to N_e, the bimodal distribution reduces to 2 possible estimates, one estimate is $\widehat{N}_{e1} = \infty$ with a

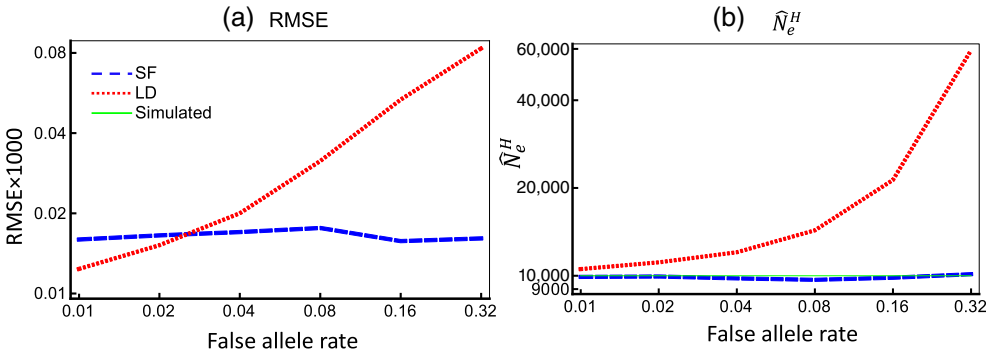

FIGURE 6 RMSE (a) and $\widehat{N}_e^H$ (b) of estimators SF and LD as a function of false allele rate at each of 10,000 SNP loci. A number of $n = 320$ individuals were taken from a population with a constant effective size of $N_e = 10{,}000$, and each individual was genotyped at 10,000 diallelic loci. RMSE and $\widehat{N}_e^H$ are represented by red dotted lines for LD, and by blue dashed lines for SF. The simulated $N_e = 10{,}000$ is represented by a thin green line. [Color figure can be viewed at wileyonlinelibrary.com]

frequency of p_1 and the other estimate is $\widehat{N}_{e2} = 20$ with a frequency of $p_2 = 1\text{-}\ p_1$. When the estimator is unbiased, we have $\frac{p_1}{\infty} + \frac{1-p_1}{20} = \frac{1}{10000}$, which leads to $p_1 = 499/500$ and $p_2 = 1/500$ which mean 499 and 1 out of 500 estimates are expected to be infinitely large and 20, respectively.

I considered a maximal $N_e = 100{,}000$ in the simulations. Is it large enough to represent the real large marine species? In a review of 26 studies, published between years 2000 and 2017, on estimating the effective sizes of marine species by Marandel et al. (2019), they found the census sizes of the species range from a few thousands in zebra shark in southern Queensland Australia (Dudgeon & Ovenden, 2015) to several billions in European anchovy in the Bay of Biscay (Montes et al., 2016). Considering the now well recognized small N_e/N ratio, on the order of $10^{-2} \sim 10^{-6}$, in high fecund marine fish species (e.g., Frankham, 1995; Hauser & Carvalho, 2008) caused perhaps by the high variance in reproductive success (Hedgecock, 1994; Hedrick, 2005), this simulated N_e may not deviate far from that of large populations in marine fishes. The review by Marandel et al. (2019) also showed that sample sizes vary widely in these studies, from 19 to 4063. Although in most studies sample sizes are small (less than 500), there do exist a few studies with sample sizes larger than 1000. My simulations indicate that large samples of individuals and loci are required for satisfactory estimates of N_e when populations are large.

A few simulation studies (e.g., Wang, 2016; Waples, 2021) have compared the biasedness and accuracy of SF and LD estimators in somewhat ideal situations (constant N_e, no mistypings, no missing data, no linkage, ...), and have reached inconclusive conclusions. Wang (2016) showed that SF is more accurate than LD when genotype data at a few microsatellites are used to estimate the N_e of small populations. In that study, a suitable sibship size prior was used by the SF estimator to reduce SEEs and thus to improve N_e estimates when marker information is not ample. Waples (2021) compared the maximal accuracy of SF (when sibship can be estimated without errors) with the accuracy of LD. He found that, compared with LD, SF is more accurate when N_e is small (say, $N_e < 100$) and less accurate when otherwise. In this study, however, the sample size n relative to the true N_e decreases rapidly with an increasing N_e. In the situation of a small population with $N_e = 50$, n varies between 20 and 35 (i.e., $n = 40\%$ or 70% of N_e), while in the situation of a large population with $N_e = 1000$, n varies between 50 and 150 (i.e., $n = 5\%$ or 15% of N_e). My simulations (Figure 3) showed that, indeed SF is less accurate than LD when $N_e = 1000$ and $n < 2000$. However, with a large enough sample size ($n > 2000$), SF becomes more accurate than LD. In general, the relative accuracies of the two estimators for any given true N_e depend on, among many other factors, n. SF estimator relies on a larger sample size of individuals than the LD estimator to yield a precise and accurate estimate of N_e. SF performs better than LD for small populations even when n is small. However, it becomes more accurate than LD for large populations only when n is very large, at least larger than N_e.

A population in the real world rarely keeps a constant N_e as assumed by the LD estimator (but not by the SF estimator). Most often it is exactly because a population is changing in demography that we are interested in knowing its N_e and the temporal changes in N_e (Schwartz et al., 2007) to inform the management for conservation or harvest. The LD information extractable from a sample of multilocus genotypes comes from (and thus signals) the cumulative contribution of genetic drift occurred in the parental and more remote ancestral generations, and thus indicates a complicated average of the effective population sizes at these generations. This is true no matter the markers used in estimating LD are linked or unlinked. Therefore, when a population is changing in size over generations, the LD estimator assuming a constant N_e would be difficult to interpret. If it is taken as the estimated N_e for the parental generation, it would be highly biased. This is confirmed in my simulations (Figure 4b) which shows that the N_e at and after a bottleneck is substantially overestimated and underestimated, respectively. When using the LD estimator for monitoring population demographic changes, therefore, caution must be exercised in interpretating the analysis results because any drastic changes in N_e would be largely "smoothed" by the estimator. In contrast, the sibling frequency in individuals sampled at random from a population is affected by the parental N_e only, and has nothing to do with the grand parental and more remote ancestral generations. Therefore, the SF estimator always infers the parental N_e, regardless of the demographic changes occurred in any generations.

With the rapid development of next generation sequencing (NGS) technology, genome-wide SNP data can be generated virtually for any species with or without a reference genome. This is great for evidence-based conservation and management of populations, as such SNP data provide ample information about, among others, the current and historical demography. However, NGS-based SNP data could also contain a lot of noises which, if not adequately dealt with, may ruin an analysis. Genotyping errors can be copious in SNP data especially when DNA quantity and quality are low (when, e.g., DNA is extracted from noninvasive samples such as feces or ancient samples) or when low-depth sequencing is applied due to reasons such as cost control. In a sibship

reconstruction analysis, genotyping errors such as allelic dropouts and false alleles can be accounted for by adopting proper error models (Wang, 2004) with rough estimates of the error rates. Loosely speaking, the likelihood of a set of individuals being siblings is calculated under these error models such that it is still sufficiently high to approve the relationship when the genotype data at only a few loci reject (due to mistyping) but the genotype data at many more loci support the relationship. Using these error models, sibship can be inferred accurately from genotype data at many loci even when each locus has a high mistyping rate (Wang, 2019), as shown in the present simulation study (Figures 5 and 6). The LD estimator, on the other hand, assumes perfect data without genotyping errors. When such errors do exist, they would ruin the LD and cause a reduced estimate of LD and thus a biased estimate of N_e, as observed in this study (Figures 5 and 6). Unlike the strict parentage or sibship exclusion analysis which does not tolerate mistyping because true parentage or sibship would be excluded even when mismatched genotypes (due to mistyping) are observed at a single locus, the LD estimator of N_e could tolerate low levels of false alleles and allelic dropouts. When the mistyping rates are substantial, say >5%, then LD estimator yields highly inaccurate and overestimated N_e values. In practice, only high-quality SNP data with few mistyping errors should be analyzed by LD estimator. Sved et al. (2013) introduced a permutation correction to the LD estimator of N_e to remove the bias caused by the presence of null alleles. The correction is not yet implemented in software available for LD based N_e estimation. Furthermore, whether the same or a similar permutation correction can be applied to data with allelic dropouts and false alleles warrants further study.

In this study, I assume the markers are physically unlinked even though they are numerous. This assumption of no linkage is unnecessary for the SF estimator, because sibship inference is hardly affected at all by the linkage of markers (Wang & Santure, 2009) sampled at random from a genome of a typical genetic map length (say, 20 Morgans). In contrast, the extent of linkage (measured by the recombination rate, c) for a pair of markers has a functional relationship with the extent of LD between the markers (Hill, 1981). As a measurement of LD, the squared correlation of allele frequencies, r^2, at a pair of loci with recombination rate c is expected to be roughly inversely proportional to $N_e c$ (Weir & Hill, 1980). Therefore, a pair of loci with tighter linkage (smaller c) are expected to have a larger r^2. If linked markers are assumed unlinked, the LD based estimator would underestimate N_e, as observed in a simulation study (Waples et al., 2016) and in an empirical study of Chinook salmon (Larson et al., 2014). The use of unlinked pairs of markers required by the LD estimator of N_e (Waples, 2006; Waples & Do, 2008) can be achieved when the linkage structure (chromosomal locations) of the markers is known and only pairs of markers from different linkage groups (chromosomes) are chosen for N_e estimation. For many nonmodel species where reference genome is unavailable, some SNPs from NGS must be linked but such linked pairs of loci cannot be reliably identified and removed from a LD based N_e analysis. In such a situation, some empirically derived rough corrections can be made to reduce the estimation bias, using information of the number of chromosomes or the size of the genome (Waples et al., 2016). When marker locations in the genetic map are known, the linkage relationships of the markers, together with the genotype information, can be used to infer the trajectory of N_e in the recent past (Santiago et al., 2020).

Similar to the issue of physical linkage of markers, some SNPs generated from NGS might be under directional or balancing selection and thus produce distorted LD and biased estimates of N_e. Again, the SF estimator should be robust to the presence of selection on some of the SNPs, as sibship is an individual level quantity while LD is a finer locus-level quantity. One might naively use some methods to identify and remove the markers under selection before conducting a LD based N_e analysis. However, while this approach might be effective for markers under strong selection, it becomes powerless for markers under weak selection. Some simulations show that LD based N_e estimates are robust to the presence of selection on linked markers (Novo et al., 2022). Whether the conclusion is extrapolatable to LD based N_e analysis of unlinked marker data by LDNe methods (Waples & Do, 2008) deserves further analysis in the future.

CONFLICT OF INTEREST STATEMENT

The author declares no conflicts of interest.

ORCID

Jinliang Wang https://orcid.org/0000-0002-8467-5448

REFERENCES

Akita, T. (2020). Nearly unbiased estimator of contemporary effective mother size using within-cohort maternal sibling pairs incorporating parental and nonparental reproductive variations. *Heredity*, *124*(2), 299–312.

Caballero, A. (1994). Developments in the prediction of effective population size. *Heredity*, *73*, 657–679.

Crow, J. F., & Kimura, M. (1970). *An introduction to population genetics theory*. Harper and Row.

Delaval, A., Frost, M., Bendall, V., Hetherington, S. J., Stirling, D., Hoarau, G., Jones, C. S., & Noble, L. R. (2022). Population and seascape genomics of a critically endangered benthic elasmobranch, the blue skate *Dipturus batis*. *Evolutionary Applications*, *15*, 78–94.

Dudgeon, C. L., & Ovenden, J. R. (2015). The relationship between abundance and genetic effective population size in elasmobranchs: An example from the globally threatened zebra shark *Stegostoma fasciatum* within its protected range. *Conservation Genetics*, *16*(6), 1443–1454.

Fisher, R. A. (1930). *The genetical theory of natural selection*. Clarendon Press.

Frankham, R. (1995). Effective population size/adult population size ratios in wildlife: A review. *Genetics Research*, *66*(2), 95–107.

Frankham, R. (2005). Genetics and extinction. *Biological Conservation*, *126*, 131–140.

Frankham, R., Bradshaw, C. J., & Brook, B. W. (2014). Genetics in conservation management: Revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biological Conservation*, *170*, 56–63.

Hauser, L., & Carvalho, G. R. (2008). Paradigm shifts in marine fisheries genetics: Ugly hypotheses slain by beautiful facts. *Fish and Fisheries*, *9*(4), 333–362.

Hedgecock, D. (1994). Does variance in reproductive success limit effective population sizes of marine organisms. *Genetics and Evolution of Aquatic Organisms*, *122*, 122–134.

Hedrick, P. (2005). Large variance in reproductive success and the N_e/N ratio. *Evolution*, *59*(7), 1596–1599.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetical Research*, *38*, 209–216.

Hoarau, G., Boon, E., Jongma, D. N., Ferber, S., Palsson, J., van der Veer, H. W., Rijnsdorp, A. D., Stam, W. T., & Olsen, J. L. (2005). Low effective population size and evidence for inbreeding in an overexploited flatfish, plaice (*Pleuronectes platessa L.*). *Proceedings of the Royal Society B: Biological Sciences*, *272*, 497–503.

Jones, O. R., & Wang, J. (2010). COLONY: A program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, *10*(3), 551–555.

Krimbas, C. B., & Tsakas, S. (1971). Genetics of *dacus-oleae*. 5. Changes of esterase polymorphism in a natural population following insecticide control-selection or drift. *Evolution*, *25*, 454–460.

Larson, W. A., Seeb, L. W., Everett, M. V., Waples, R. K., Templin, W. D., & Seeb, J. E. (2014). Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications*, *7*, 355–369.

Laurent, V., & Planes, S. (2007). Effective population size estimation on *Sardina pilchardus* in the Bay of Biscay using a temporal genetic approach: Effective population sizes of sardines. *Biological Journal of the Linnean Society*, *90*, 591–602.

Leberg, P. (2005). Genetic approaches for estimating the effective size of populations. *The Journal of Wildlife Management*, *69*(4), 1385–1399.

Luikart, G., Ryman, N., Tallmon, D. A., Schwartz, M. K., & Allendorf, F. W. (2010). Estimation of census and effective population sizes: The increasing usefulness of DNA-based approaches. *Conservation Genetics*, *11*, 355–373.

Macbeth, G. M., Broderick, D., Buckworth, R. C., & Ovenden, J. R. (2013). Linkage disequilibrium estimation of effective population size with immigrants from divergent populations: A case study on Spanish Mackerel (*Scomberomorus commerson*). *G3*, *3*(4), 709–717.

Marandel, F., Lorance, P., Berthelé, O., Trenkel, V. M., Waples, R. S., & Lamy, J. B. (2019). Estimating effective population size of large marine populations, is it feasible? *Fish and Fisheries*, *20*, 189–198.

Montes, I., Iriondo, M., Manzano, C., Santos, M., Conklin, D., Carvalho, G. R., Irigoien, X., & Estonba, A. (2016). No loss of genetic diversity in the exploited and recently collapsed population of Bay of Biscay anchovy (*Engraulis encrasicolus, L.*). *Marine Biology*, *163*, 1–10.

Nadachowska-Brzyska, K., Dutoit, L., Smeds, L., Kardos, M., Gustafsson, L., & Ellegren, H. (2021). Genomic inference of contemporary effective population size in a large Island population of collared flycatchers (*Ficedula albicollis*). *Molecular Ecology*, *30*, 3965–3973.

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

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*, 443–451.

Novo, I., Santiago, E., & Caballero, A. (2022). The estimates of effective population size based on linkage disequilibrium are virtually unaffected by natural selection. *PLoS Genetics*, *18*(1), e1009764.

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: Causes, consequences and solutions. *Nature Reviews Genetics*, *6*, 847–859.

Poulsen, N. A., Nielsen, E. E., Schierup, M. H., Loeschcke, V., & GrøNkjaer, P. (2005). Long-term stability and effective population size in North Sea and Baltic Sea cod (*Gadus morhua*): Effective population size in Atlantic cod. *Molecular Ecology*, *15*, 321–331.

Santiago, E., Novo, I., Pardiñas, A. F., Saura, M., Wang, J., & Caballero, A. (2020). Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. *Molecular Biology and Evolution*, *37*(12), 3642–3653.

Schwartz, M. K., Luikart, G., & Waples, R. S. (2007). Genetic monitoring as a promising tool for conservation and management. *Trends in Ecology & Evolution*, *22*(1), 25–33.

Sved, J. A., Cameron, E. C., & Gilchrist, A. S. (2013). Estimating effective population size from linkage disequilibrium between unlinked loci: Theory and application to fruit fly outbreak populations. *PLoS One*, *8*(7), e69078.

Wang, J. (2001). A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical Research*, *78*, 243–257.

Wang, J. (2004). Sibship reconstruction from genetic data with typing errors. *Genetics*, *166*(4), 1963–1979.

Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*, 1395–1409.

Wang, J. (2009). A new method for estimating effective population size from a single sample of multilocus genotypes. *Molecular Ecology*, *18*, 2148–2164.

Wang, J. (2016). A comparison of single-sample estimators of effective population sizes from genetic marker data. *Molecular Ecology*, *25*, 4692–4711.

Wang, J. (2019). Pedigree reconstruction from poor quality genotype data. *Heredity*, *122*(6), 719–728.

Wang, J., Santiago, E., & Caballero, A. (2016). Prediction and estimation of effective population size. *Heredity*, *117*, 193–206.

Wang, J., & Santure, A. W. (2009). Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics*, *181*(4), 1579–1594.

Wang, J., & Whitlock, M. C. (2003). Estimating effective population size and migration rates from genetic samples over space and time. *Genetics*, *163*, 429–446.

Waples, R. K., Larson, W. A., & Waples, R. S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity*, *117*(4), 233–240.

Waples, R. S. (1989). A generalised approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, *121*, 379–391.

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. (2021). Relative precision of the sibship and LD methods for estimating effective population size with genomics-scale datasets. *Journal of Heredity*, *112*, 535–539.

Waples, R. S., Antao, T., & Luikart, G. (2014). Effects of overlapping generations on linkage disequilibrium estimates of effective population size. *Genetics*, *197*, 769–780.

Waples, R. S., & Do, C. (2008). LDNe: A program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, *8*, 753–756.

Waples, R. S., & Do, C. (2010). Linkage disequilibrium estimates of contemporary N_e using highly variable genetic markers: A largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, *3*, 244–262.

Waples, R. S., & England, P. R. (2011). Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. *Genetics*, *189*, 633–644.

Waples, R. S., & Feutry, P. (2022). Close-kin methods to estimate census size and effective population size. *Fish and Fisheries*, *23*, 273–293.

Weir, B. S., & Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics*, *95*, 477–488.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*, 97–159.

How to cite this article: Wang, J. (2023). Estimating current effective sizes of large populations from a single sample of genomic marker data: A comparison of estimators by simulations. *Population Ecology*, 1–13. https://doi.org/10.1002/1438-390X.12167