Estimating effective population size using RADseq: Effects of SNP selection and sample size

Florianne Marandel, Grégory Charrier, Jean-Baptiste Lamy, Sabrina Le Cam, Pascal Lorance, Verena M. Trenkel

Received: 21 November 2019 | Revised: 13 December 2019 | Accepted: 23 December 2019

DOI: 10.1002/ece3.6016

Ecology and Evolution Open Access WILEY

ORIGINAL RESEARCH

Estimating effective population size using RADseq: Effects of SNP selection and sample size

Florianne Marandel[1] | Grégory Charrier[2] | Jean-Baptiste Lamy[3] | Sabrina Le Cam[2,3] | Pascal Lorance[1] | Verena M. Trenkel[1]

[1]Ifremer, Ecologie et Modèles pour l'Halieutique, Nantes, France

[2]Laboratoire des Sciences de l'Environnement Marin (LEMAR, UMR 6539 CNRS/IRD/UBO/Ifremer), Université de Bretagne Occidentale, Institut Universitaire Européen de la Mer, Plouzané, France

[3]Ifremer, Génétique et Pathologie des Mollusques Marin (SG2M-LGPMM), La Tremblade, France

Correspondence
Verena M. Trenkel, rue de l'île d'Yeu, BP 21105, 44311 Nantes cedex 3, France.
Email: verena.trenkel@ifremer.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-14-CE02-0006-01; Fondation Total, Grant/Award Number: GenoPopTaille-Capsules; H2020 European Research Council, Grant/Award Number: 773713

Abstract

Effective population size (N_e) is a key parameter of population genetics. However, N_e remains challenging to estimate for natural populations as several factors are likely to bias estimates. These factors include sampling design, sequencing method, and data filtering. One issue inherent to the restriction site-associated DNA sequencing (RADseq) protocol is missing data and SNP selection criteria (e.g., minimum minor allele frequency, number of SNPs). To evaluate the potential impact of SNP selection criteria on N_e estimates (Linkage Disequilibrium method) we used RADseq data for a nonmodel species, the thornback ray. In this data set, the inbreeding coefficient F_{IS} was positively correlated with the amount of missing data, implying data were missing nonrandomly. The precision of $N_e$estimates decreased with the number of SNPs. Mean N_e estimates (averaged across 50 random data sets with2000 SNPs) ranged between 237 and 1784. Increasing the percentage of missing data from 25% to 50% increased N_e estimates between 82% and 120%, while increasing the minor allele frequency (MAF) threshold from 0.01 to 0.1 decreased estimates between 71% and 75%. Considering these effects is important when interpreting RADseq data-derived estimates of effective population size in empirical studies.

KEYWORDS

effective population size, linkage disequilibrium, NeEstimator, RADseq, skates and rays

1 | INTRODUCTION

Effective population size (N_e) is a valuable parameter in population genetics and conservation (Hamilton, 2009; Hare et al., 2011). This parameter is related to the number of individuals which actually participate to produce the next generation and thus informs on population viability (Soulé, 1987). However, estimating N_e can be challenging. Theoretically from a genetic point of view, N_e is defined as the size of an ideal population that would experience the same rate of change in allele frequencies or heterozygosity as the observed population (Beaumont, Boudry, & Hoare, 2010; Hamilton, 2009; Wright, 1931). Ideal populations are constituted of diploid organisms with sexual reproduction, nonoverlapping generations, random mating, no migration, no mutation, but also no natural selection and constant population size (Wright, 1931); census population size is equal to effective population size in an ideal population.

Two main approaches are employed for estimating N_e: demographic methods based on life history traits and genetic methods

based on genetic markers. Demographic approaches estimate N_e as a function of parameters such as mean and variance in offspring number, survival-at-age, and birth rate. However, demographic methods often rely on strong assumptions such as discrete generations (Caballero, 1994; Nomura, 2002) or if overlapping generations are admitted, stable age structure (Robin S. Waples, Do, & Chopelet, 2011). Only one demographic method allows demographic stochasticity and heterogeneity at the expense of challenging data demands such as individual-level information (Engen, Lande, Saether, & Gienapp, 2010). This might explain why the method has not been much used so far (but see Trask, Bignal, McCracken, Piertney, & Reid, 2017).

Genetic methods have gained in popularity and power due to recent advances in genotyping and sequencing technologies as well as in computer processing speed. They rely on the extraction of genetic signals (allele frequencies) which are theoretically known to be affected by population demography, mainly effective population size. Among the genetic methods available, single-sample approaches are appealing since they require sampling only at one point in time. The most popular N_e estimator is based on a measure of linkage disequilibrium (LD), that is, the nonrandom association of alleles at different loci. The LD method has been widely used during the last decade for a variety of organisms, including mammals (Cervantes, Pastor, Gutiérrez, Goyache, & Molina, 2011; Juarez et al., 2016), insects (Francuski & Milankov, 2015), reptiles (Bishop, Leslie, Bourquin, & O'Ryan, 2009), and fishes (Pilger, Gido, Propst, Whitney, & Turner, 2015; Wilson, McDermid, Wozney, Kjartanson, & Haxton, 2014).

Empirical estimates of N_e are often biased because all methods rely on strong assumptions which are likely violated in natural populations (R. S. Waples, Antao, & Luikart, 2014). Numerous recent genetic studies have documented how more realistic simulations or real data, which do not fulfill methods' assumptions, lead to biased N_e estimates (Gilbert & Whitlock, 2015; Hare et al., 2011; Luikart, Ryman, Tallmon, Schwartz, & Allendorf, 2010; Marandel et al., 2019; Robinson & Moyer, 2013; Russell & Fewster, 2009; R. S. Waples et al., 2014; Robin S. Waples & Do, 2010). Among the various sources of bias, the assumption of nonoverlapping generations is often violated. In this case, the amount and the direction of bias as well as the precision of estimates are highly dependent on life history traits, thus species-specific, but also on the sampling fraction (Marandel et al., 2019; R. S. Waples et al., 2014). Other factors such as unequal sex ratio, high level of inbreeding and high variance in family sizes have also been found to bias N_e estimates(Montarry et al., 2019).

For nonmodel species, the absence of a reference genome challenges the development of genetic markers and the assessment of genomic ascertainment bias, and more generally the amount of expected species-specific bias for N_e estimates. A widely used method to develop de novo genetic markers and genotype individuals in one single step is the restriction associated DNA sequencing (RADseq), which provides thousands of sequenced SNP (single-nucleotide polymorphism) markers across many individuals at reasonable costs (Davey & Blaxter, 2010). A drawback of the method is the numerous sources of genotyping errors (Mastretta-Yanes et al., 2015) and missing data (information missing for certain individuals for certain markers). One case of genotyping errors is dropped alleles, that is, one allele is not typed making a heterozygous individual appearing homozygous (Bilton et al., 2018). Missing data and random allelic dropouts can bias LD estimates (Akey, Zhang, Xiong, Doris, & Jin, 2001; Bilton et al., 2018) and subsequently bias LD based N_e estimates and increase their variance (Nunziata & Weisrock, 2018; Russell & Fewster, 2009). The degree of bias in LD estimates depends on allele frequency (Akey et al., 2001). Further, rare alleles are known to cause positive bias in N_e estimates (e.g., Nunziata & Weisrock, 2018; Russell & Fewster, 2009). For microsatellites, this has led to the recommendation to select those with minor allele frequency (MAF) >0.01 if sample size >100 (Robin S. Waples & Do, 2010). For SNPs, rare alleles can be avoided by keeping only the SNPs with highest polymorphic content (Phillips et al., 2004). However, the effect of the MAF threshold remains poorly known, but see Nunziata and Weisrock (2018).

The aim of this study was to determine the effects of the MAF, the proportion of missing data and the number of SNPs on N_e estimates when applying the LD approach to RADseq data. These effects were explored using empirical data collected for the thornback ray (*Raja clavata*, Figure 1) in the Bay of Biscay.

2 | MATERIAL AND METHODS

2.1 | Sampling

Overall 159 thornback rays were sampled in the Bay of Biscay between 2011 and 2016 (half the samples were collected in 2015) (Figure 2). Sampled individuals were collected at sea (EVHOE and RaieJuve surveys carried out by Ifremer) and at landing ports from commercial fisheries. The sex ratio of the sample was close to 1:1 (78 females, 81 males). Total length varied from 12.5 to 96 cm.

FIGURE 1 Thornback ray *Raja clavata*

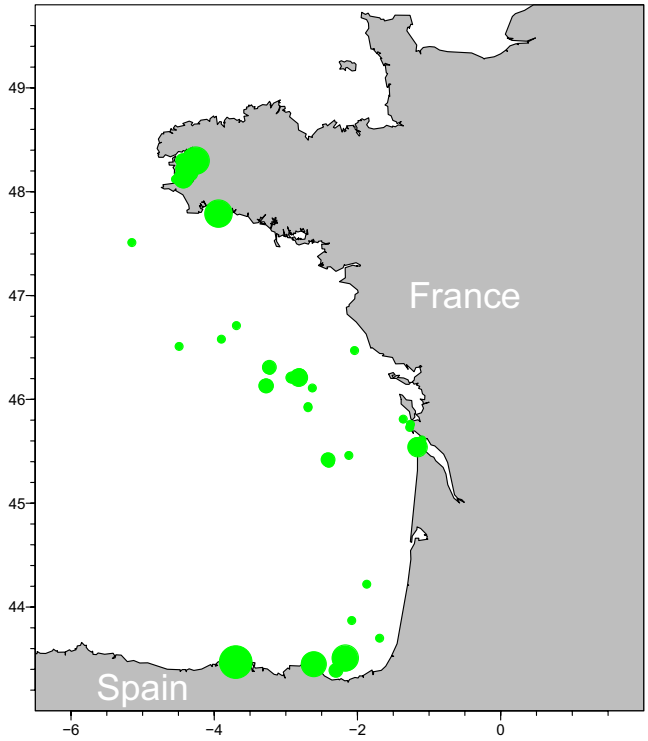

FIGURE 2 Sampling locations of thornback rays in the Bay of Biscay. Number proportional to bubble surface

2.2 | RAD-sequencing protocol and bioinformatics

All individuals were genotyped by sequencing using a RADseq protocol to effectively subsample the genome of multiple individuals at homologous genomic regions. The library construction followed the original protocol by Baird et al. (2008) with slight modifications. Briefly, 1 μg of genomic DNA from each individual was digested with the restriction enzyme *Sbf*I-HF (New England Biolabs), and then ligated to a P1 adapter labeled with a unique barcode. We used 16 barcodes of 5-bp and 16 barcodes of 6-bp length in our P1 adapters to build 32-plex libraries. The 159 individuals were part of a wider sample including individuals from other regions. From these, one pool of three individuals and seven pools of 32 individuals were made by mixing individual DNA in equimolar proportions and sheared to an average size of 500 and 350 bp, respectively, using a Covaris S220 sonicator (KBiosciences). A size-selected step was carried out on agarose gel to keep DNA fragments within the size range 500–1000 bp for the 3-plex and 300–700 pb for the 32-plexes. Each library was then submitted to end-repair, A-tailing, and ligation to P2 adapter before PCR amplification for 18 cycles. Amplification products from six PCR replicates were pooled for each library, gel-purified after size selection and quantified on a 2,100 Bioanalyzer using the High Sensitivity DNA kit (Agilent). The 3-plex library was sequenced in paired-ends 300 reads using Illumina Miseq technology. Each 32-plex library was sequenced on a separate lane of an Illumina Hiseq 2500 instrument by INTEGRAGEN, using 100-bp single reads.

We aligned *ca* 54.6M paired-end Miseq reads to the little skate (*Leucoraja erinacea*) genome assembly (Wang et al., 2012; Wyffels et al., 2014) using BWA-SW (version 0.7.12-r1039, default parameters) to build up thornback ray consensus sequences from high quality mapped reads (mapQ score = 60). The result was used as a reference for further analyses. Raw sequences from the 32-plexes were quality checked, trimmed to 95bp and demultiplexed using the process_radtags module of Stacks v1.32(Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013). Demultiplexed sequences were aligned to the custom reference genome from the mapped Miseq reads using BWA-SW version 0.7.12-r1039 (default parameters) for locus assembly and SNP calling was achieved with the reference mapping pipeline ref_map.pl (Stacks v1.32; (Catchen et al., 2013). Individuals were genotyped on 389 483 putative SNPs spread on 35 134 RAD loci (a sequence starting or ending with a restriction enzyme site). Given the variability due to laboratory work and sequencing protocols, we chose to retain only the loci with a calling rate percentage above 50% (i.e., maximum percentage of missing data (NA) of 50%) and a MAF above 0.01 for the 159 individuals to remove spurious SNPs. This raw data set had 43 088SNPs (Le Cam et al., 2019). Given the large number of amplification cycles (18), a preliminary analysis of the dependence of the heterozygote miscall rate on mean SNP read depth was carried out using the R package whoa (Anderson, 2019) as suggested by a reviewer. The potential heterozygote miscall rate was estimated from comparing the observed number of heterozygous individuals with the number expected given the allele frequency and assuming the SNP was in Hardy–Weinberg equilibrium (Hendricks et al., 2018). Based on this a data set containing only genotypes with read depths between 30 and 300 copies, MAF ≥0.01 and NAs ≤0.5 was created (referred to as full data, 17 843 SNPs). Removing genotypes with low and very high read depths (below 30 and above 300 copies) reduced the number of SNPs but also increased the proportion of missing data. A second data set with a maximum NA of 25% was therefore created (referred to as reduced data, 4,816 SNPs). The lower NA threshold value is more in line with common practices in empirical studies (e.g., 15% missing data in Pazmino, Maes, Simpfendorfer, Salinas-de-Leon, and van Herwerden (2017), 25% in Rodriguez-Ezpeleta et al. (2016)).

The randomness of missing data in the full data set was tested by estimating the Spearman rank correlation between the proportion of missing data and the inbreeding index $F_{IS} = 1-H_{obs}/H_{exp}$, where H_{obs} is the observed proportion of heterozygous individuals and H_{exp} the expected proportion under Hardy–Weinberg equilibrium for a given SNP. We also tested the correlation between the proportion of missing data and the proportion of heterozygous individuals (H_{obs}).

Seven individuals including five from outside the Bay of Biscay were genotyped twice. This replicate data set was used to explore genotyping error and allelic dropout. Allelic dropout corresponded to one of the replicate genotypes being heterozygous but not the other one or one being homozygous for the major allele and the other for the minor allele. The correlation between the proportion of replicated individuals exhibiting dropout and the inbreeding coefficient for all 159 individuals was tested.

2.3 | Effective population size

The single point estimation method linkage disequilibrium (LD) is based on linkage disequilibrium due to the nonrandom association of alleles at different gene loci. LD is measured at one point in time by the covariance between loci. We used NeEstimatorV2.1 (Do et al., 2014) for estimating effective population size $\hat{N}_e$. All samples from different years and cohorts (size classes) were pooled for estimation.

First, the effect of the number of SNPs was evaluated by drawing randomly (without replacement) 500 to 4,000 SNPs from the reduced data set with MAF ≥0.01 and percent missing data NA ≤25%. Fifty replicate data sets were created, and the mean and coefficient of variation (standard deviation/mean) of replicate $\hat{N}_e$ estimates were calculated.

The effects of four thresholds for the minimum MAF were then evaluated for the full data set: 0.01, 0.02, 0.05, and 0.1 where a value of 0.01 means that the selected SNPs had a MAF in the range 0.01 to 0.5. This rather wide range of threshold values was chosen to explore the shape of the relationship between the MAF filter and $\hat{N}_e$ estimates. High threshold values (>0.05) might be unsuitable for practical applications (see discussion). The MAF filter was combined with a filter for NA for each SNP which had six levels between 25% and 50% (5% steps). Combining MAF and NA thresholds led to 24 empirical genetic datasets for which $\hat{N}_e$ was estimated. The number of available SNPs varied between data sets from 1549 to 17 842 (Table 1). For standardization, 2000 SNPs were randomly selected (without replacement) from each data set, except for the smallest data set for which it was 1,000 (NA ≤25%, MAF ≥0.1). This number of SNPs was sufficient to stabilize estimates (see results). An ANOVA was fitted to replicate log-transformed $\hat{N}_e$ estimates for comparing the effects of MAF and missing data filters, as well as their interaction. Residuals were checked for normality.

The effect of the sample size on $\hat{N}_e$ estimates was evaluated with the reduced data set (MAF ≥0.01; NA ≤25%) by creating random data sets with the number of individuals ranging from 25 to 150. A rarefaction curve analysis was calculated as a function of sample size. A parametric model (Michaelis-Menten) was fitted using nonlinear least-squares to estimate $\hat{N}_e$ free of sample size effects, which corresponds to the model asymptote.

All data handling and analysis of results were carried out in R (R Development Core Team, 2008).

TABLE 1 Number of SNPs available for different data selection thresholds for minor allele frequency (MAF) and missing data

	MAF lower threshold			
Missing data (%)	0.01	0.02	0.05	0.1
25	4,816	3,849	2,374	1549
30	7,072	5,718	3,497	2,238
35	9,388	7,620	4,751	3,030
40	11,913	9,682	5,979	3,754
45	14,782	11,958	7,368	4,566
50	17,842	14,315	8,788	5,401

3 | RESULTS

3.1 | Exploratory analysis

In the raw data, the estimated mean miscall rate decreased strongly as the minimum read depth increased (Figure 3). It was around 0.75 considering all SNPs in the raw data set. Therefore, further analyses were restricted to genotypes with read depth in the range 30 to 300 copies (full data set).

The distribution of MAF values in the full data set was nonuniform with most of the SNPs displaying MAF <0.1 (Figure 4a) and NA >25% (Figure 4b). The distribution of missing individuals was nonuniform across individuals with 33 individuals missing more than 50% of SNPs (Figure 4c).

A significant positive correlation (Spearman's rho = 0.61, *p*-value < .001) was found between the percentage of missing data and the inbreeding coefficient F_{IS} of a given SNP (Figure 5), while the correlation between the proportion of missing data and the proportion of heterozygous individuals was significantly negative (Spearman's rho = −0.28, *p*-value < .001). Thus, data were not missing at random: Individuals with missing data were more likely to be heterozygous.

For the seven individuals genotyped twice, on average 11% of SNPs (all SNPs genotyped twice) had a different genotype (median 8%, range 4%–19%).When genotypes differed, in 85% (median 89%, range 67%–98%) of cases one replicate was heterozygote and the other homozygote. Further, the proportion of individuals exhibiting allelic dropout for a given SNP was significantly negatively correlated with the inbreeding coefficient for all individuals for the same SNP (Spearman's rho = −0.12, *p*-value < .001, *n* = 17 842 SNPs).

3.2 | Ne estimation

The mean estimate of $\hat{N}_e$ across the 50 random data sets stabilized at around 1,500 SNPs and uncertainty decreased with the number of SNPs (Figure 6). The coefficient of variation (CV) decreased from 0.29 for 500 SNPs to 0.07 for 2,000 SNPs and 0.02 for 4,000 SNPs. This indicates that the 2000 SNPs used for exploring the effects of missing data and MAF thresholds were sufficient for obtaining reliable estimates.

The effects on $\hat{N}_e$ estimates of the thresholds for MAF and NA were important (Figure 7). Mean $\hat{N}_e$ estimates ranged between 237 and 1,784 corresponding to a factor of 7.5. Mean values decreased by 71% to 75% with increasing MAF threshold and increased by 82% to 120% with NA. For example, for the smallest NA (25%), mean $\hat{N}_e$ (averaged across 50 replicates) decreased by 76% from 982 to 237 as the MAF threshold increased from 0.01 to 0.1. In contrast, for the smallest MAF threshold (0.01), the mean $\hat{N}_e$ increased by 82% from 982 to 1784 when NA increased from 25% to 50%.

The ANOVA revealed that the effect of the MAF threshold value was eight times larger than that of NA (Table 2). There was a weak but significant interaction between the two factors.

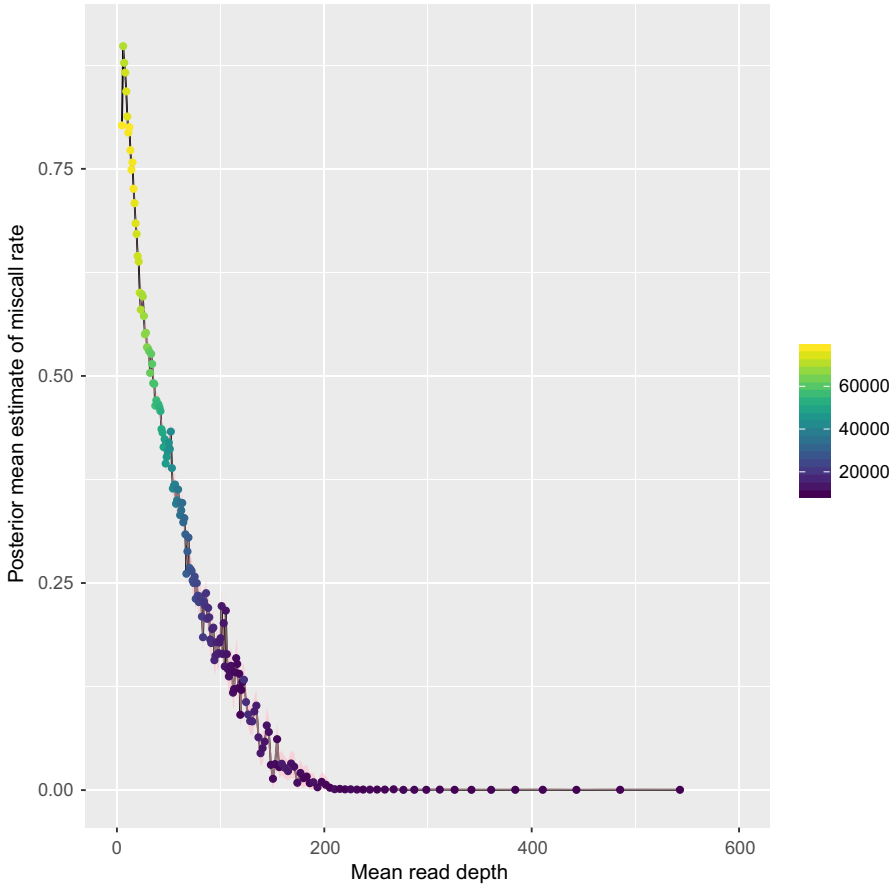

FIGURE 3 Estimated miscall rate of SNPs as function of mean read depth of each SNP for raw data set for thornback ray in the Bay of Biscay. The color scale indicates the number of data points (number of individuals * number of SNPs)

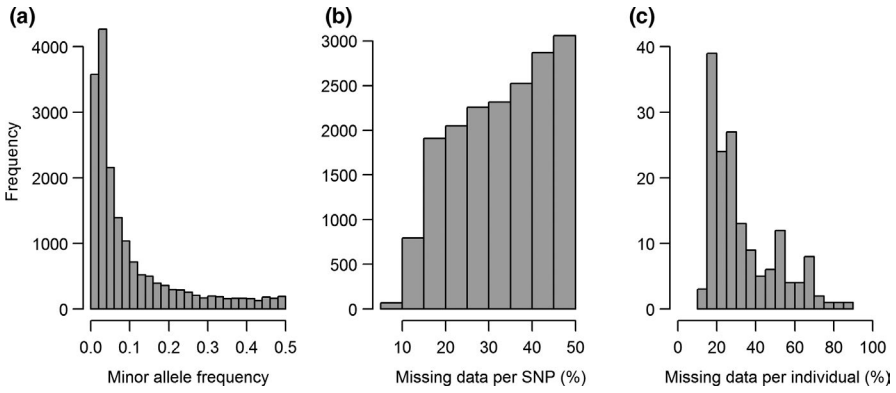

FIGURE 4 (a) Histogram of minor allele frequencies of SNPs with percentage missing data ≤50%. (b) Histogram of percent missing data for SNPs with minor allele frequency ≥0.01. (c) Percent missing SNPs per individuals for percentage missing data ≤50% and minor allele frequency ≥0.01

Estimates of N_e were negative for a sample size of 25 individuals and decreased somewhat from an average of 1,165 for 50 individuals to 977 for 150 individuals (Figure 8). The asymptote of the fitted model ignoring negative estimate was 903 (SE 21.6) which can be interpreted as the estimate that would have been obtained with a sufficient sample size, implying that the 159 individuals were insufficient.

4 | DISCUSSION

Empirical genetic data from thornback rays sampled in the Bay of Biscay were used to explore the effects of data selection on N_e estimates. Genetic markers were obtained from a RADseq protocol in which individuals with missing data for a given SNP are common (Nunziata & Weisrock, 2018) and genotyping errors frequent (Mastretta-Yanes et al., 2015). In the thornback ray data, for 11% of SNPs the genotype differed between the two replicates on average across the seven individuals genotyped twice retaining only genotypes with read depth 30–300 copies. Unfortunately, genotyping errors for RADseq data are seldom reported in the literature. Higher disagreement rates have been found for oyster (J.B. Lamy pers. comm.) while lower error rates (2%–12%) have been reported for the plant *Berberis alpina* (Mastretta-Yanes et al., 2015). Further, SNPs were missing nonrandomly with the amount of missing data increasing with the inbreeding coefficient while, the proportion of replicate individuals with allelic dropout was lower for SNPs with higher inbreeding coefficient. For microsatellites Soulsbury, Iossa,

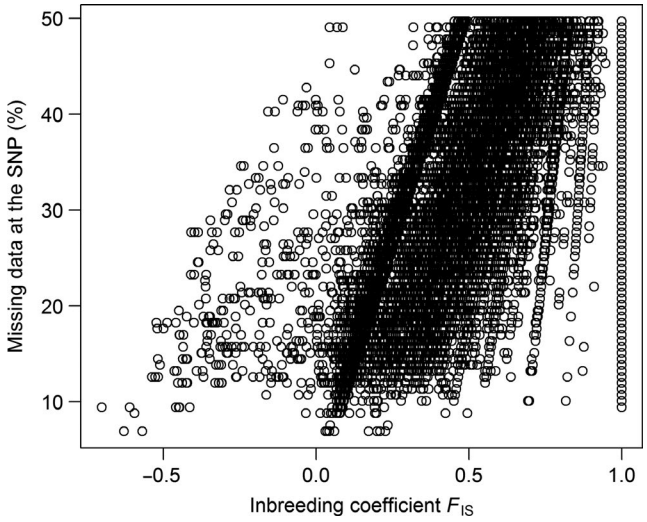

FIGURE 5 Relationship between the missing data threshold and the inbreeding coefficient of selected SNPs (minor allele frequency ≥0.01; percent missing data ≤50%) for thornback ray in the Bay of Biscay

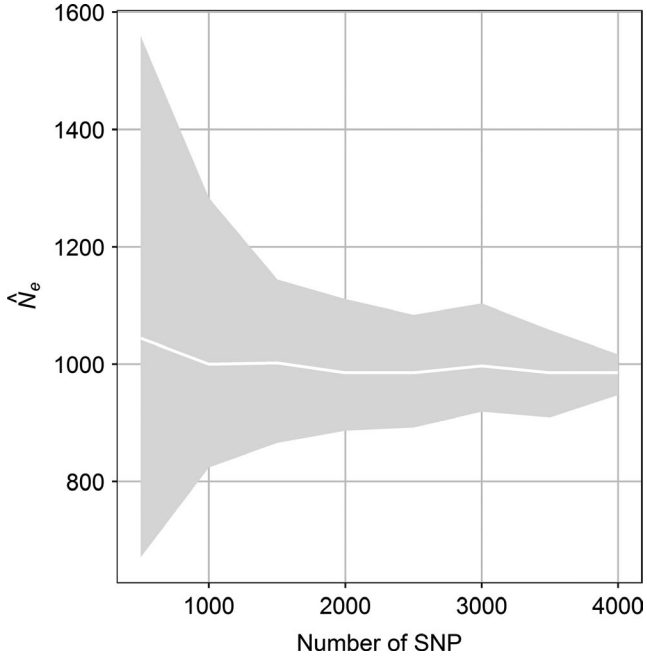

FIGURE 6 Relationship between Ne estimates and the number of SNPs for thornback ray in the Bay of Biscay (minor allele frequency ≥0.01; percent missing data ≤25%). White line is mean of 50 random data sets and shaded area central 90% percentile band

Edwards, Baker, and Harris (2007) also found a relationship between allelic dropout and departure from Hardy–Weinberg equilibrium, that is, the inbreeding coefficient.

Contrary to other ecological studies of nonmodel species a reference genome was used here (assembly from a related species) that allowed us to identify SNPs with greater power and avoid common problems encountered with the *de novo* RADseq analysis such as merging paralogous loci as alleles (e.g., Diaz-Arce & Rodriguez-Ezpeleta, 2019).

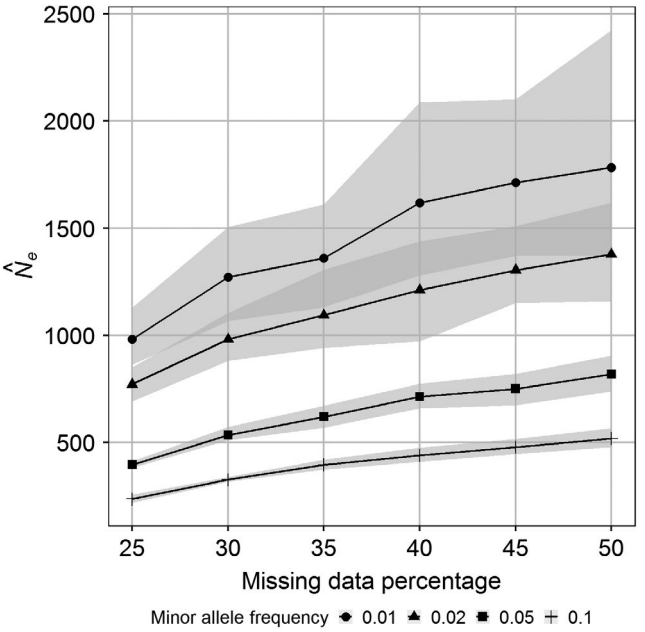

FIGURE 7 Relationship between Ne estimates and missing data percentage threshold for different threshold levels of the minor allele frequency for thornback ray in the Bay of Biscay. Continuous lines are mean values for 50 random data sets with 2000 SNPs and shaded areas central 90% percentile bands

The uncertainty of N_e estimates obtained with the linkage disequilibrium method decreased strongly with the number of SNPs. Mean values stabilized at around 1,500 SNPs for the full data set. In comparison, Pazmino et al. (2017) used 8,103 neutral SNPs (MAF ≥0.02; missing data ≤15%; replication error ≤5%) for a shark species while Montes et al. (2016) used 349 neutral SNPs for estimating effective population size for an anchovy population and Diaz-Arce and Rodriguez-Ezpeleta (2019) 96 SNPs (missing data ≤9%) for salmon.

Subsampling the data with different thresholds for missing data and minimum minor allele frequency permitted us to evaluate the effects of these two factors on N_e estimates. Depending on the combination of threshold values, N_e estimates varied by up to a factor of 7.5. In comparison, the well-known effect of underestimation of N_e due to ignoring overlapping generations is only around 30% in the thornback ray, independent of the census population size (Marandel et al., 2019). Further, the MAF threshold value (tested range 0.01 to 0.1) had a larger effect compared with the NA (tested range 25 to 50%). This is not surprising given NeEstimator accounts for missing data (NeEstimator V2.1 online documentation at http://www.molecularfisherieslaboratory.com.au/neestimator-software). The effect of the polymorphism on N_e estimates using LD was previously addressed by Russell and Fewster (2009) for ideal populations, and we agree with these authors that researchers should be aware of the effects of the MAF threshold applied for SNP selection.

In contrast to SNP selection criteria, the sample size was found to impact estimates only slightly, given at least 50 individuals were used. Negative estimates are expected when sample size is

TABLE 2 Analysis of variance for testing the effects of threshold values for percent of missing data (NA) and minimum minor allele frequency (MAF) on log-transformed effective population size (N_e) estimates

Name	*df*	MS	F	*P*-value
NA	5	12.21	1,590.25	<.001
MAF	3	101.45	13,215.66	<.001
NA:MAF	15	0.08	10.73	<.001
Residuals	1,176	0.01		

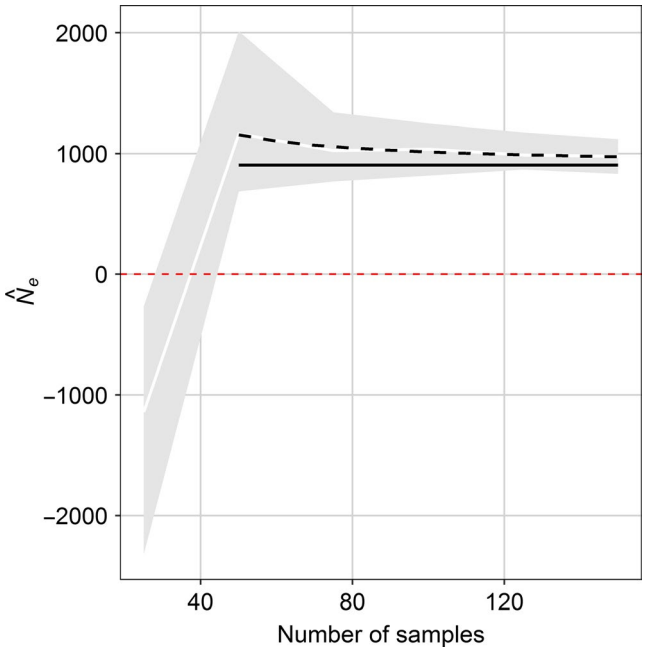

FIGURE 8 Relationship between Ne estimates and sample size for thornback ray in the Bay of Biscay (minor allele frequency ≥0.01; percent missing data ≤25%). Continuous white line is mean value for 50 random data sets with 2000 SNPs and shaded areas central 90% percentile bands. Black dotted line is fitted model whose asymptote is plotted as continuous horizontal black line

insufficient (Marandel et al., 2019). Further, the 159 individuals were probably not enough to obtain stabilized estimates. This might not be surprising given that simplified genetic simulations for a thornback ray like species indicated that around 1% of the population needed to be sampled to obtain reliable estimates (Marandel et al., 2019) and the Bay of Biscay population is potentially large (Marandel, Lorance, & Trenkel, 2016). The rarefaction analysis with the reduced data set indicated a stabilized effective population size of 903 (asymptote of fitted model). For the thornback ray population in the Irish Sea and Bristol Channel Chevolot, Ellis, Rijnsdorp, Stam, and Olsen (2008) estimated N_e as being 283 using five microsatellites and a temporal estimation method with samples from two time periods. Given the difference in approach, it is unknown whether their sample of 363 individuals and number of microsatellites was sufficient and hence whether the two effective population size estimates can be compared. If they are comparable, the Bay of Biscay populations would be the larger one.

Other genetic simulations for thornback ray populations in European waters using contrasted assumptions for migration rates suggested a stable large scale population structure with little exchange (Marandel et al., 2018). This could mean that migration might not be expected to impact much allele frequencies in European thornback ray population, hence effective population size estimates. Selection can also cause nonrandom association of alleles within and across loci which will again be interpreted as genetic drift (underestimation of N_e) by the linkage disequilibrium estimator (Waples & Do, 2010). Contrary to genetic drift that affects all loci in the genome, selection only affects certain loci (depending on the genetic architecture) but its effect should be diluted when using a large number of SNPs (»100) as done here). Depending on the genetic determinism of the selected traits (monogenic to polygenic) and the intensity of the selective process, the effect on N_e estimates is hard to predict.

Further, physically unlinked SNPs were assumed, which is clearly unrealistic (Waples, Larson, & Waples, 2016). The number of truly independent SNPs is equal to the number of chromosomes, which is 98 for thornback ray (Nygren, Nilsson, & Jahnke, 1971) times the average number of crossing-over per chromosome in thornback ray. The finite number of chromosomes will create linkage disequilibrium (more precisely gametic linkage disequilibrium) purely due to physical linkage between SNPs, rather than true N_e changes (Waples et al., 2016).

Based on our results as a guide for practitioners we recommend to use the lowest feasible percentage of missing data, though the precise threshold value will depend on the overall sample size and the expected effective population size. The main principle is to maintain a sufficiently large sample size (in terms of genotyped individuals) for all SNPs included in the analysis. It is more difficult to make recommendations regarding the threshold value for the minor allele frequency. It is important to keep in mind, that in a perfect Fisher-Wright population, thresholds on MAF values are nonsense since any filtration will remove important genetic information to infer N_e. However, empirical datasets will always contain loci with alleles of spurious low or very low frequencies. There are a growing number of methods to discard spurious SNPs with a low MAF within the bioinformatics pipeline by taking conservative filters on minimum read depth of the loci. Here, a read depth of at least 30 was required to reduce replication error. At the least, the lowest possible MAF filter should be chosen (compromise between loosing relevant genetic information and noise) and the results for different threshold values should be compared.

In conclusion, for nonmodel species special attention should be paid to the interpretation of N_e estimates as large bias in estimates might occur when using the LD method. For thornback ray, we found that nonrandomly missing data, allele frequency filters and sample size had much larger effects than the expected bias due to ignoring overlapping generations (Marandel et al., 2019). We expect these

findings to hold for other nonmodel species though we recommend further studies to confirm this.

ACKNOWLEDGMENTS

We acknowledge funding from the French "Agence Nationale de la Recherche" (ANR) for the GenoPopTaille project (ANR-14-CE02-0006-01), from the Fondation Total for the GenoPopTaille-Capsules project and the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773713 (PANDORA). FM thanks Ifremer for a PhD studentship. The authors thank Jennifer Ovenden and Robin Waples for comments on an earlier version of this manuscript and Florence Cornette for help with the laboratory work. We thank three anonymous referees for their comments and suggestions. We acknowledge the projects RAIECOAM and RAIEbeca, the volunteers of APECS, Eric Stephan, Graham Johnston and Guzman Diez for thornback ray samples. The authors thank the UMR 8199 LIGAN-PM Genomics platform (Lille, France) which belongs to the "Federation de Recherche" 3508 Labex EGID (European Genomics Institute for Diabetes; ANR-10-LABX-46) and were supported by the ANR Equipex 2010 session (ANR-10-EQPX-07-01; "LIGAN-PM"). The LIGAN-PM Genomics platform (Lille, France) is also supported by the FEDER and the Region Nord-Pas-de-Calais-Picardie.

AUTHOR CONTRIBUTIONS

GC, PL, and VT designed research, SL carried out laboratory work and bioinformatics, FM and VT analyzed the data and wrote the first draft, all authors critically revised the manuscript.

DATA AVAILABILITY STATEMENT

Data were be archived at Seanoe https://doi.org/10.17882/70648.

ORCID

Florianne Marandel https://orcid.org/0000-0001-8140-0599
Jean-Baptiste Lamy https://orcid.org/0000-0002-6078-0905
Pascal Lorance https://orcid.org/0000-0002-6453-2925
Verena M. Trenkel https://orcid.org/0000-0001-7869-002X

REFERENCES

Akey, J. M., Zhang, K., Xiong, M. M., Doris, P., & Jin, L. (2001). The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *American Journal of Human Genetics*, *68*(6), 1447–1456. https://doi.org/10.1086/320607

Anderson, E. C. (2019). *Evaluation of genotyping error in genotype-by-sequencing data. R package 'whoa'*. https://CRAN.R-project.org/package=whoa

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Beaumont, A. R., Boudry, P., & Hoare, K. (2010). *Biotechnology and genetics in fisheries and aquaculture*, 2nd. ed. Chichester,UK: Blackwell.

Bilton, T. P., McEwan, J. C., Clarke, S. M., Brauning, R., van Stijn, T. C., Rowe, S. J., & Dodds, K. G. (2018). Linkage disequilibrium estimation in low coverage high-throughput sequencing data. *Genetics*, *209*(2), 389–400. https://doi.org/10.1534/genetics.118.300831

Bishop, J. M., Leslie, A. J., Bourquin, S. L., & O'Ryan, C. (2009). Reduced effective population size in an overexploited population of the Nile crocodile (*Crocodylus niloticus*). *Biological Conservation*, *142*(10), 2335–2341. https://doi.org/10.1016/j.biocon.2009.05.016

Caballero, A. (1994). Developments in the prediction of effective population size. *Heredity*, *73*(6), 657–679. https://doi.org/10.1038/hdy.1994.174

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Cervantes, I., Pastor, J. M., Gutiérrez, J. P., Goyache, F., & Molina, A. (2011). Computing effective population size from molecular data: The case of three rare Spanish ruminant populations. *Livestock Science*, *138*(1–3), 202–206. https://doi.org/10.1016/j.livsci.2010.12.027

Chevolot, M., Ellis, J. R., Rijnsdorp, A. D., Stam, W. T., & Olsen, J. L. (2008). Temporal changes in allele frequencies but stable genetic diversity over the past 40 years in the Irish Sea population of thornback ray. *Raja Clavata Heredity*, *101*(2), 120–126. https://doi.org/10.1038/hdy.2008.36

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: Next-generation population genetics. *Briefings in Functional Genomics*, *9*(5–6), 416–423. https://doi.org/10.1093/bfgp/elq031

Diaz-Arce, N., & Rodriguez-Ezpeleta, N. (2019). Selecting RAD-seq data analysis parameters for population genetics: The more the better? *Frontiers in Genetics*, *10*, https://doi.org/10.3389/fgene.2019.00533

Do, C., Waples, R. S., Peel, D., Macbeth, G. M., Tillett, B. J., & Ovenden, J. R. (2014). NeEstimator V2: Re-implementation of software for the estimation of contemporary effective population size Ne from genetic data. *Molecular Ecology Resources*, *14*(1), 209–214. https://doi.org/10.1111/1755-0998.12157

Engen, S., Lande, R., Saether, B.-E., & Gienapp, P. (2010). Estimating the ratio of effective to actual size of an age-structured population from individual demographic data: Effective population size of age-structured populations. *Journal of Evolutionary Biology*, *23*(6), 1148–1158. https://doi.org/10.1111/j.1420-9101.2010.01979.x

Francuski, L., & Milankov, V. (2015). Assessing spatial population structure and heterogeneity in the dronefly: Spatial population structure in the dronefly. *Journal of Zoology*, *297*, 286–300. https://doi.org/10.1111/jzo.12278

Gilbert, K. J., & Whitlock, M. C. (2015). Evaluating methods for estimating local effective population size with and without migration: Estimating N_e in the presence of migration. *Evolution*, *69*(8), 2154–2166. https://doi.org/10.1111/evo.12713

Hamilton, M. B. (2009). *Population genetics*. Chichester, UK; Hoboken, NJ: Wiley-Blackwell.

Hare, M. P., Nunney, L., Schwartz, M. K., Ruzzante, D. E., Burford, M., Waples, R. S., … Palstra, F. (2011). Understanding and estimating effective population size for practical application in marine species management: Applying effective population size estimates to marine species management. *Conservation Biology*, *25*(3), 438–449. https://doi.org/10.1111/j.1523-1739.2010.01637.x

Hendricks, S., Anderson, E. C., Antao, T., Bernatchez, L., Forester, B. R., Garner, B., … Luikart, G. (2018). Recent advances in conservation and population genomics data analysis. *Evolutionary Applications*, *11*(8), 1197–1211. https://doi.org/10.1111/eva.12659

Juarez, R. L., Schwartz, M. K., Pilgrim, K. L., Thompson, D. J., Tucker, S. A., Smith, J. B., & Jenks, J. A. (2016). Assessing temporal genetic variation in a cougar population: Influence of harvest and neighboring populations. *Conservation Genetics*, *17*, 379–388. https://doi.org/10.1007/s10592-015-0790-5

Le Cam, S., Bidault, A., Charrier, G., Cornette, F., Lamy, J.-B., Lapegue, S., …Trenkel, V. (2019). RADSeq-derived SNP genotypes for 159 thornback ray Raja clavata from the Bay of Biscay. SEANOE. https://doi.org/10.17882/70648

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*(2), 355–373. https://doi.org/10.1007/s10592-010-0050-7

Marandel, F., Lorance, P., Andrello, M., Charrier, G., Le Cam, S., Lehuta, S., & Trenkel, V. M. (2018). Insights from genetic and demographic connectivity for the management of rays and skates. *Canadian Journal of Fisheries and Aquatic Sciences*, *75*, 1291–1302. https://doi.org/10.1139/cjfas-2017-0291

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. https://doi.org/10.1111/faf.12338

Marandel, F., Lorance, P., & Trenkel, V. M. (2016). A Bayesian state-space model to estimate population biomass with catch and limited survey data: Application to the thornback ray (*Raja clavata*) in the Bay of Biscay. *Aquatic Living Resources*, *29*(2), 209. https://doi.org/10.1051/alr/2016020

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Pinero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, *15*(1), 28–41. https://doi.org/10.1111/1755-0998.12291

Montarry, J., Bardou-Valette, S., Mabon, R., Jan, P. L., Fournet, S., Grenier, E., & Petit, E. J. (2019). Exploring the causes of small effective population sizes in cyst nematodes using artificial Globodera pallida populations. Proceedings of the Royal Society B-Biological Sciences, 286(1894). https://doi.org/10.1098/rspb.2018.2359

Montes, I., Iriondo, M., Manzano, C., Santos, M., Conklin, D., Carvalho, G. R., … 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*(5), https://doi.org/10.1007/s00227-016-2866-2

Nomura, T. (2002). Effective size of populations with unequal sex ratio and variation in mating success. *Journal of Animal Breeding and Genetics*, *119*(5), 297–310. https://doi.org/10.1046/j.1439-0388.2002.00347.x

Nunziata, S. O., & Weisrock, D. W. (2018). Estimation of contemporary effective population size and population declines using RAD sequence data. *Heredity*, *120*(3), 196–207. https://doi.org/10.1038/s41437-017-0037-y

Nygren, A., Nilsson, B., & Jahnke, M. (1971). Cytological studies in Hypotremata and Pleurotremata (Pisces). *Hereditas*, *67*, 275–282. https://doi.org/10.1111/j.1601-5223.1971.tb02380.x

Pazmino, D. A., Maes, G. E., Simpfendorfer, C. A., Salinas-de-Leon, P., & van Herwerden, L. (2017). Genome-wide SNPs reveal low effective population size within confined management units of the highly vagile Galapagos shark (Carcharhinus galapagensis). *Conservation Genetics*, *18*(5), 1151–1163. https://doi.org/10.1007/s10592-017-0967-1

Phillips, C., Lareu, M., Sanchez, J., Brion, M., Sobrino, B., Morling, N., … Carradeco, A. (2004). Selecting single nucleotide polymorphisms for forensic applications. *International Congress Series*, *1261*, 18–20. https://doi.org/10.1016/j.ics.2003.12.001

Pilger, T. J., Gido, K. B., Propst, D. L., Whitney, J. E., & Turner, T. F. (2015). Comparative conservation genetics of protected endemic fishes in an arid-land riverscape. *Conservation Genetics*, *16*(4), 875–888. https://doi.org/10.1007/s10592-015-0707-3

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna: Austria.

Robinson, J. D., & Moyer, G. R. (2013). Linkage disequilibrium and effective population size when generations overlap. *Evolutionary Applications*, *6*(2), 290–302. https://doi.org/10.1111/j.1752-4571.2012.00289.x

Rodriguez-Ezpeleta, N., Bradbury, I. R., Mendibil, I., Alvarez, P., Cotano, U., & Irigoien, X. (2016). Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers: Effects of sequence clustering parameters and hierarchical SNP selection. *Molecular Ecology Resources*, *16*(4), 991–1001. https://doi.org/10.1111/1755-0998.12518

Russell, J. C., & Fewster, R. M. (2009). Evaluation of the Linkage Disequilibrium method for estimating effective population size. In D. L. Thomson, E. G. Cooch, & M. J. Conroy (Eds.), *Modeling Demographic Processes In Marked Populations* (pp. 291–320). Boston, MA: Springer, US.

Soulé, M. (1987). *Viable populations for conservation*. Cambridge, UK: Cambridge University Press.

Soulsbury, C. D., Iossa, G., Edwards, K. J., Baker, P. J., & Harris, S. (2007). Allelic dropout from a high-quality DNA source. *Conservation Genetics*, *8*(3), 733–738. https://doi.org/10.1007/s10592-006-9194-x

Trask, A. E., Bignal, E. M., McCracken, D. I., Piertney, S. B., & Reid, J. M. (2017). Estimating demographic contributions to effective population size in an age-structured wild population experiencing environmental and demographic stochasticity. *Journal of Animal Ecology*, *86*(5), 1082–1093. https://doi.org/10.1111/1365-2656.12703

Wang, Q., Arighi, C. N., King, B. L., Polson, S. W., Vincent, J., Chen, C., … Wu, C. H. (2012). Community annotation and bioinformatics workforce development in concert–Little Skate Genome Annotation Workshops and Jamborees. *Database*, *2012*, bar064–bar064. https://doi.org/10.1093/database/bar064

Waples, R. S., Antao, T., & Luikart, G. (2014). Effects of overlapping generations on Linkage Disequilibrium estimates of effective population size. *Genetics*, *197*(2), 769–780. https://doi.org/10.1534/genetics.114.164822

Waples, R. S., & Do, C. (2010). Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: A largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, *3*(3), 244–262. https://doi.org/10.1111/j.1752-4571.2009.00104.x

Waples, R. S., Do, C., & Chopelet, J. (2011). Calculating N_e and N_e / N in age-structured populations: A hybrid Felsenstein-Hill approach. *Ecology*, *92*(7), 1513–1522. https://doi.org/10.1890/10-1796.1

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. https://doi.org/10.1038/hdy.2016.60

Wilson, C. C., McDermid, J. L., Wozney, K. M., Kjartanson, S., & Haxton, T. J. (2014). Genetic estimation of evolutionary and contemporary effective population size in lake sturgeon (Acipenser fulvescens Rafinesque, 1817) populations. *Journal of Applied Ichthyology*, *30*(6), 1290–1299. https://doi.org/10.1111/jai.12615

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

Wyffels, J., King, L. B., Vincent, J., Chen, C., Wu, C. H., & Polson, S. W. (2014). SkateBase, an elasmobranch genome project and collection of molecular resources for chondrichthyan fishes. *F1000Research*. *3*, 191–https://doi.org/10.12688/f1000research.4996.1.

How to cite this article: Marandel F, Charrier G, Lamy J-B, Le Cam S, Lorance P, Trenkel VM. Estimating effective population size using RADseq: Effects of SNP selection and sample size. *Ecol Evol*. 2020;10:1929–1937. https://doi.org/10.1002/ece3.6016