Received: 6 March 2023 | Revised: 7 June 2023 | Accepted: 5 July 2023

DOI: 10.1111/1755-0998.13837

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

An empirical test of the estimation of historical effective population size using *Drosophila melanogaster*

Irene Novo[1] | Noelia Pérez-Pereira[1] | Enrique Santiago[2] | Humberto Quesada[1] | Armando Caballero[1]

[1]Centro de Investigación Mariña, Universidade de Vigo, Facultade de Bioloxía, Vigo, Spain

[2]Departamento de Biología Funcional, Facultad de Biología, Universidad de Oviedo, Oviedo, Spain

Correspondence
Irene Novo, Centro de Investigación Mariña, Universidade de Vigo, Facultade de Bioloxía, Vigo 36310, Spain.
Email: irene.novo.gimenez@uvigo.es

Funding information
MCIN/AEI/10.13039/501100011033, Grant/Award Number: PID2020-114426GB-C21; Ministerio de Ciencia, Innovación y Universidades, Grant/Award Number: FPU18/04642; Ministerio de Educación, Cultura y Deporte, Grant/Award Number: FPU16/02299; Universidade de Vigo, Grant/Award Number: Open Access funding provided thanks to the CRUE-CS; Xunta de Galicia, Grant/Award Number: Centro singular de investigación de Galicia accre and ED431C 2020-05

Handling Editor: Kimberly J. Gilbert

Abstract

The availability of a large number of high-density markers (SNPs) allows the estimation of historical effective population size (N_e) from linkage disequilibrium between loci. A recent refinement of methods to estimate historical N_e from the recent past has been shown to be rather accurate with simulation data. The method has also been applied to real data for numerous species. However, the simulation data cannot encompass all the complexities of real genomes, and the performance of any estimation method with real data is always uncertain, as the true demography of the populations is not known. Here, we carried out an experimental design with *Drosophila melanogaster* to test the method with real data following a known demographic history. We used a population maintained in the laboratory with a constant census size of about 2800 individuals and subjected the population to a drastic decline to a size of 100 individuals. After a few generations, the population was expanded back to the previous size and after a few further generations again expanded to twice the initial size. Estimates of historical N_e were obtained with the software GONE both for autosomal and X chromosomes from samples of 17 individuals sequenced for the whole genome. Estimates of the historical effective size were able to infer the patterns of changes that occurred in the populations showing generally good performance of the method. We discuss the limitations of the method and the application of the software carried out so far.

KEYWORDS
effective population size, genetic drift, inbreeding, SNPs

1 | INTRODUCTION

Effective population size (N_e, Wright, 1931) is a key parameter to quantify the amount of genetic drift and inbreeding in a population, with great relevance in evolutionary biology, conservation genetics and animal and plant breeding (Charlesworth, 2009; Gilbert & Whitlock, 2015; Luikart et al., 2010; Caballero, 2020, Chap. 5; Wang et al., 2016). This parameter is considered one of the main indicators for biodiversity conservation for the "post-2020" framework by the Convention on Biological Diversity (CBD) (Frankham, 2021; Hoban et al., 2020). The availability of high-throughput genomic information allows the inference of the past demography of populations, which has an interest in human evolution and conservation genetics, to date and quantify the main events affecting population

size. A number of methods have been developed for this inference. While some of them are thought to provide estimates of effective size in the long run, usually for thousands of generations in the past (Kamm et al., 2020; Liu & Fu, 2020; Schiffels & Durbin, 2014; Speidel et al., 2019), others are intended to infer changes in N_e in the recent past, based on identity-by-descent of genomic fragments (Browning & Browning, 2015), allele-frequency spectrum and linkage disequilibrium (Boitard et al., 2016) or only linkage disequilibrium between markers (Barbato et al., 2015; Hayes et al., 2003; Hollenbeck et al., 2016; Santiago et al., 2020). The latest refinement of the latter method (Santiago et al., 2020) has been shown to accurately infer drastic changes that have occurred in the populations over the past 100–200 generations. This method has been shown to be robust to the interaction between positive or negative selection and recombination (Novo et al., 2022) so that it provides estimates of N_e only dependent on the variance of family size in the population.

The estimation of N_e from linkage disequilibrium (LD) between markers assumes that LD occurs by chance, that is, by pure genetic drift given the particular breeding system and variation in offspring numbers from parents (Hill, 1981). The relationship between the LD between markers, measured as the correlation (r) between alleles of pairs of loci, and the effective size, is given by the approximate relation $r^2 \approx 1/(1+4N_e c)$ (Sved, 1971), where c is the rate of recombination between loci. If independent markers ($c=0.5$) are considered in the analysis, estimates of the contemporary effective size can be obtained (Waples, 2006). However, linked markers also allow for the estimation of historical effective population sizes. This arises from the principle that LD between markers at a genetic distance of c Morgans provides information on the N_e at $1/(2c)$ generations in the past (Hayes et al., 2003).

The method developed by Santiago et al. (2020), which is implemented in the software GONE (available at https://github.com/esrud/GONE), has been evaluated with simulations under different scenarios (Novo et al., 2022; Reid & Pinsky, 2022; Santiago et al., 2020; Saura et al., 2021), generally showing a good ability to detect drastic changes in historical N_e. In fact, Reid and Pinsky (2022) showed that the method exhibits ≥90% accuracy for detecting fast severe declines in population size, outperforming other methods. The method has also been applied to different sets of data from different species (Atmore et al., 2022; Bird et al., 2023; Ding et al., 2022; Ferrette et al., 2023; Pacheco et al., 2022; von Seth et al., 2022; Wersebe & Weider, 2023). However, the simulation tests have always a restricted scope, as the models simulated cannot encompass all the complexities of real genomes. In addition, the empirical estimates are generally done without knowing the real demography of the populations, even though relevant changes in population size can be assumed from different sources of data or contrasted with historical data on some occasions (Waldman et al., 2022). Thus, in this study, we aimed to evaluate the performance of the Santiago et al. (2020) method with experimental data from a *Drosophila melanogaster* population subjected to laboratory controlled drastic changes in size over generations. Our objective was to assess GONE's ability to detect these changes and infer the patterns in population demography, as well as to discuss its limitations and the current application of the method.

2 | METHODS

2.1 | Flies maintenance and demographic history

A laboratory population was established in 2009 and maintained with around 2800 (range 2500–3000) individuals, constant over generations (López-Cortegano et al., 2016; Pérez-Pereira et al., 2021). The population was maintained with constant temperature (25°C) and continuous lighting with circular mixing of 32 bottles each with 40–50 individuals of each sex and the feeding medium used in our laboratory (Pérez-Pereira et al., 2021). In each generation, male and female adults were placed into the corresponding bottle and were removed after a week, before the next generation emerged, so that there were no overlapping generations and the longevity of the flies in adult stage was of about 11 days. The expected effective size of the population was inferred to be about $N_e \approx 1400$ (López-Cortegano et al., 2016) (see Supplementary Material for more details on this inference and simulation results supporting it). At generation 208, this base population was subjected to different demographic scenarios (Figure 1). First, the population size was drastically reduced by sampling 100 individuals (half of each sex) and establishing them in a single bottle. This population was maintained for eight generations with this size. Then, over a period of two generations, the small

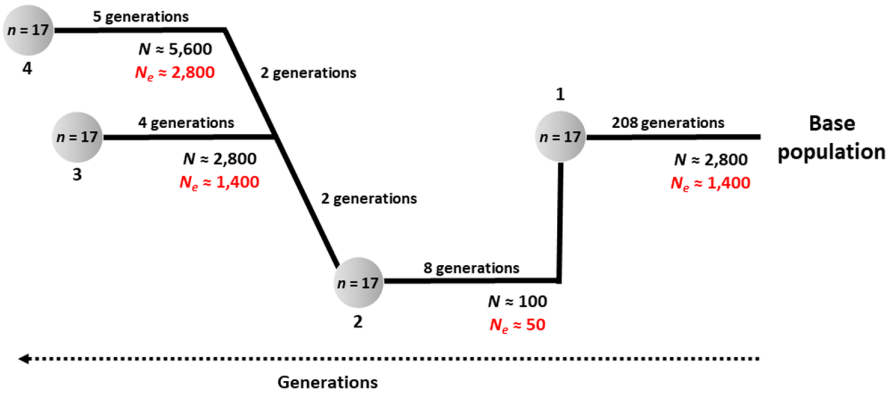

FIGURE 1 Scheme showing the experimental scenarios. The circles represent time points where a sample of $n=17$ individuals was taken for whole genome sequencing and they refer to: (1) a large base population with a constant size; (2) a severe population decline; (3) a severe bottleneck with recovering of the previous size and (4) a severe bottleneck and subsequent population expansion.

population was expanded back to the previous size (32 bottles with about 100 individuals each) and, after another two generations, to a size twice as large (64 bottles). The expanded population, first with 32 bottles, and later with 64 bottles, was maintained with its size for four and five generations respectively. The maintenance of the bottles in all cases was carried out with circular mixing between the bottles, in the same way as the original base population. We sampled $n=17$ individual males for the four points marked with numbered circles in Figure 1 in order to carry out genome sequencing of the individuals. Females were not considered to avoid offspring contamination due to the retention of fertilized eggs in their reproductive tracts. Thus, point 1 was used to estimate the historical effective size of the large initial base population; point 2 was used to estimate a sudden decline in census size in the population; and points 3 and 4 to estimate expansions in population size after the previous drop.

2.2 | DNA extraction and sequencing

The sampled individuals were frozen in liquid nitrogen and stored at −80°C. We performed an optimized protocol for DNA extraction from individual flies with the Gentra Puregene Cell Kit (Qiagen). All samples were sent to Macrogen (South Korea) for 2×150 bp paired-end whole genome sequencing on an Illumina Novaseq 6000 instrument using Nextera XT DNA libraries.

2.3 | SNP calling

The FASTQ sequencing files underwent a first quality control with FastQC (Andrews, 2010). Adapters were removed with Trimmomatic (Bolger et al., 2014) with Nextera adapter list. Quality and size trimming (minimum sequence length after trimming=36) were performed with ERNE-FILTER v2 (Del Fabbro et al., 2013). After a new quality control with FastQC, we mapped the resulting reads against the 6.14 *D.melanogaster* reference genome with BWA-MEM (Li, 2013). Using SAMtools v1 (Danecek et al., 2021), the resulting SAM files were converted to indexed BAM files, and chromosomes were isolated and indexed. SAMtools (Danecek et al., 2021) was used for PCR duplicate removal and filtering (minimum mapping quality=20). The quality control of the alignment was carried out with Qualimap v2 (Okonechnikov et al., 2016). 92.9% of the reference genome was covered, with an average coverage for autosomes of 49× (25× for the X chromosome), and an average mapping quality of 58.6. The HaplotypeCaller tool from GATK v4 (Van der Auwera & O'Connor, 2020) was used for raw variant calling (minimum phred-scaled confidence threshold=10). The resulting gVCF files were combined using the GATK CombineGVCFs tool. We kept only biallelic SNPs. Highly repeated sites, low information regions and SNPs within 10 bp of an indel were removed with BCFtools (Danecek et al., 2021). Finally, SNPs were also filtered with the GATK VariantFiltration tool applying the recommended presets. Then, PLINK (Purcell et al., 2007) was used for file conversion from VCF files to PLINK map/ped files. The final average number of SNPs in each sample was 1,135,122.

2.4 | Genetic map usage

For accurate historical N_e estimation, GONE requires a precise genetic map. The genetic distances in the *D.melanogaster* genetic map from Comeron et al. (2012) were originally calculated for females, since males do not recombine. However, the genomic consequences of recombination emerge from populations of both males and females, so for N_e estimation, we halved the genetic distances from Comeron's map in the case of autosomes and multiplied them by 2/3 for chromosome X. Flybase Coordinates Converter tool (Gramates et al., 2022) was used for conversion of the genetic map's coordinates to dm6. In order to obtain a similar density of SNPs for the four autosomal samples, these were matched to those appearing in the second sample (point 2 in Figure 1). The total number of SNPs available for analysis were 891,737, 942,323, 882,861 and 894,727 for the autosomal chromosomes and 143,742, 98,450, 108,026 and 85,767 for the X-chromosome, in the four samples, respectively.

2.5 | Historical N_e estimation

We estimated historical linkage-disequilibrium N_e using GONE (Santiago et al., 2020). We used all the default options of the software for the analysis of the autosomal chromosomes: unknown phase, Haldane's correction for genetic distances, no minor allele frequency filtering and use of all SNPs including those with missing data. We considered the four autosomal arms (2L, 2R, 3L and 3R) as chromosomes in the analysis, and used a randomly chosen set of 50,000 SNPs per arm in each of 20 replicates. We used a maximum value of recombination frequency of $c=0.05$, the default value recommended by Santiago et al. (2020). However, in order to test the relevance of this cut-off point, we also considered maximum values of $c=0.1$ and 0.01. We also made estimates without knowledge of the genetic map, assuming that the average recombination rate of the species was evenly distributed along the genome. The geometric mean of the estimated N_e values from the 20 replicates was obtained for each scenario.

Drosophila males are haploids for the X chromosome, so we analysed it separately. Since GONE requires diploid genomes, we combined the haplotypes of the X chromosome from pairs of individuals and used them to build $n=8$ diploid pseudo-genomes. Twenty replicates were made with random combinations of the 17 available chromosomes. The analysis with GONE assumed in this case known phase, with the rest of the parameters as for the autosomal analysis. Estimates assuming cut-off values of recombination rate $c=0.1$, 0.05 and 0.01, and lack of knowledge of the genetic map, were also considered, such as for the autosomal analyses. Again, the geometric mean of the estimated N_e values from the 20 replicates was obtained for each scenario.

The estimates of N_e for autosomal and X-chromosomes were combined as $N_e = (4/5)\ N_{e,AUTO} + (1/5)\ (4/3)\ N_{e,X}$, taking into account that the latter constitutes approximately 1/5 of the *Drosophila* genome length and the N_e for X-linked genes is expected to be 3/4 that of the autosomal N_e (Wright, 1933).

3 | RESULTS

Figure 2 shows the estimates of historical N_e combining autosomal and X-chromosome estimates. These are shown separately in Figure S1 and for each of the 20 replicates run in Figure S2. The average combined estimates of N_e for the most recent (1–4), intermediate (16–20) and most distant (56–60) generations are shown, along with their standard errors, in Table 1. For the base population scenario, there was a good estimation of the global N_e (Figure 2a), although this arises from a compensation between the underestimation from autosomal data and the overestimation from X-chromosome data (Figure S1a). There was a 31% underestimation of the true N_e in the most recent and most distant generations and a 35% overestimation in the intermediate generations (Table 1). The severe decline and subsequent expansions were generally well detected, although the timing of the events was not perfectly inferred and the ancestral N_e (before the bottleneck and expansion) were substantial overestimations. The estimated N_e in the most recent generations of the drop scenario (54.0 ± 1.1) was significantly smaller than the estimate in the ancestral generations (1005.1 ± 69.8) (Figure 2b; Table 1), thus showing the power of the method to detect a substantial drop in N_e. Likewise, the average N_e in the intermediate generations of the bottleneck and bottleneck and expansion scenarios were significantly lower than the corresponding values in the most recent and most distant generations (Figure 2c,d; Table 1). However, the expansion of population size to twice the initial size (Figure 2d) was not captured in full, although it was well predicted by the X-chromosome estimation (Figure S1d).

The estimations shown in Figure 2 assume a cut-off maximum value of the recombination rate of $c = 0.05$ and known genetic map. Analogous results assuming other cut-off c values (0.1 and 0.01) as well as a lack of knowledge of the genetic map are shown in Supplementary Figure S3. Using a maximum value of $c = 0.1$ can produce a huge bias in the estimates, with a high increase in the estimated N_e and a drastic drop in the most recent generations

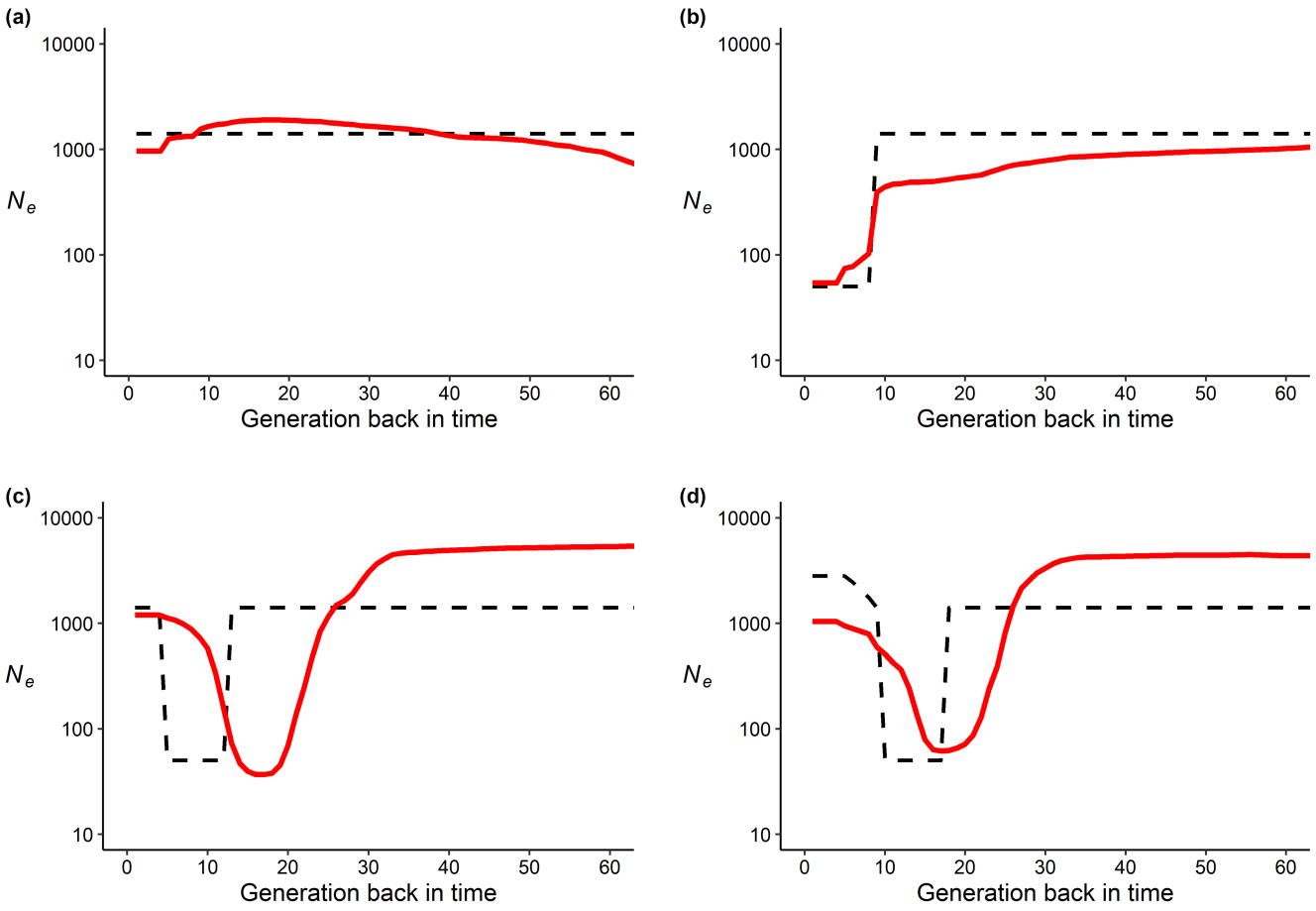

FIGURE 2 GONE's historical N_e estimates for the four demographic scenarios: (a) a large base population with a constant size; (b) a severe population decline; (c) a severe bottleneck with recovery of the previous size and (d) a severe bottleneck and subsequent population expansion. The estimates are obtained as a combination of autosomal and X-chromosome estimates, each obtained from the geometric mean of 20 replicates. The approximated true N_e is shown with a discontinuous black line.

TABLE 1 Mean and standard errors of the combined estimates of N_e at recent (1–4), intermediate (16–20) and late (56–60) generations back in time.

Generations	BP	DROP	BOTTLE	BOT + EXP
1–4	959.2 ± 441.4	54.0 ± 1.1	1196.3 ± 46.1	1034.0 ± 28.8
16–20	1892.5 ± 352.7	519.6 ± 22.9	44.98 ± 40.4	64.7 ± 1.4
56–60	946.7 ± 110.9	1005. ± 69.8	5293.0 ± 300.4	4403.4 ± 114.5

Abbreviations: BOT + EXP, severe bottleneck and subsequent population expansion; BOTTLE, severe bottleneck with recovering of the previous size; BP, base population with a constant size; DROP: severe population decline.

(Figure S3a). However, for some scenarios, such as the bottleneck and further expansion (Figure S3d), it better predicts the most recent N_e. As expected, using a too low cut-off value ($c < 0.01$) can miss the population size changes that occurred in the very recent past. If the genetic map remains unknown and the average recombination rate for the species is assumed to be uniform across the genome, the estimated N_e values are generally less precise than when the genetic map is known (Figure S3).

4 | DISCUSSION

Our results show that the method developed by Santiago et al. (2020) is useful to detect drastic changes in the demography of populations. *Drosophila* is, in some respects, a model species with some limitations for this purpose because of its short genome (about 180 Mb of about 1.25 Morgans) and because it only has two autosomal chromosomes. In addition, the sample size used in our study was relatively low (17 individuals). Even so, the method was able to infer the main changes that occurred in the population. The time for the bottleneck that occurred in the scenarios of Figure 2c,d was estimated to be a few generations before the real one, but the general pattern was well predicted. However, the fast expansion after the bottleneck was only partially captured, perhaps because the number of generations that elapsed after the expansion was relatively low to capture rapid changes in linkage disequilibrium. Furthermore, the method does not have enough power to detect the signal of a small difference in drift (the inverse of N_e) following a large perturbation in population size. In fact, the method predicts squared correlations between sites that are roughly proportional to $1/N_e c$. Given the small sample size, a difference between very small numbers, e.g., 1/(1400) vs 1/(2800), could not be detected.

The use of different cut-off values for the maximum recombination rate (c) considered in the analysis has also relevance for a proper estimation of recent changes in N_e (Figure S3). For example, the large expansion after the bottleneck can be better predicted considering a maximum value of $c = 0.1$ (Figure S3d) but, in contrast, using this cut-off value may produce substantial biases in the estimation of recent N_e for other scenarios (Figure S3a). In order to investigate this issue, we carried out computer simulations following the experimental design (Figure S4). The methods for these simulations and the corresponding results are detailed in the Supplementary Material. We considered a structured population such as that of the experiment, where individuals are mated in groups (like the experimental bottles) and there is mixing of groups across generations. Estimation of historical N_e was carried out assuming cut-off values of $c = 0.05$ and 0.5. The simulation results show that the most recent changes in N_e cannot be captured by using a cut-off value of 0.05 (Figure S5c,d) but are well captured if $c = 0.5$ (Figure S9c,d). However, because the population is somewhat structured, using this maximum value of c implies a bias in the most recent N_e for the constant size scenario (Figure S9a).

Because GONE currently does not account for sex chromosomes of haploid genomes, we opted for randomly pairing the chromosomes of different males producing eight pseudo-diploid genomes and applying the program considering a known phase. The results were very satisfactory taking into account that they were obtained with a single chromosome, suggesting that this procedure can be used if necessary. This was, in fact, the same procedure followed by Singh et al. (2021) for the analysis of the fungus *Zymoseptoria tritici*, a haploid species.

Our experimental populations show a ratio N_e/N of about 0.5 (N being the number of adult individuals of the population) (López-Cortegano et al., 2016). This ratio is coherent with other estimates of laboratory populations of *D. melanogaster* maintained in bottles or cages (average ratio of 0.34 for 10 estimates in Table S1 of Hoban et al., 2020), and it is rather common when there is male competition for female mating (Nunney, 1993). A value of $N_e/N \approx 0.5$ would reflect a variance of family size (V_k) about three times larger than that for an ideal population with a Poisson distribution of progeny number per individual (Wright, 1938). This variation must be the reflection on N_e of differences in fecundity and mating ability of the parents, as well as on the viability of the offspring. This N_{eVk} is the value estimated by the software GONE, which is not affected by further reductions in N_e due to the interaction between selection and recombination in genomic regions of tight linkage (Charlesworth et al., 1993; Hudson & Kaplan, 1995; Santiago & Caballero, 1998, 2016). Thus, as shown by Novo et al. (2022), GONE provides virtually unbiased estimates of the expected effective population size from the variance of progeny numbers (N_{eVk}).

The estimations of historical N_e presented in Figure 2 and Figure S1 are obtained from the average of 20 replicates of each analysis (Figure S2). In the case of autosomal estimations, each replicate is obtained by sampling (with replacement) a different random set of 50,000 SNPs per chromosomal arm, which is the procedure followed when the software is run at different times and there is a

sufficiently large number of SNPs. As noted in Figure S2, the variation observed among replicates was rather small in the case of autosomal chromosomes. For the X-chromosomes the observed variation across replicates was higher, as the variation depended not only on the set of 50,000 SNPs analysed but also on the random combination of chromosomes considered. In addition, the number of sampled pseudo-diploid genomes was only eight in this case. Other authors applying the method have suggested to perform a block-jackknife approach by setting non-overlapping blocks of 10,000 consecutive SNPs and removing one of these sets in each replicated analysis (Magnier et al., 2022). Yet, others have opted for making replicates from subsamples of half of the chromosomes available (Pacheco et al., 2022). These procedures seem appropriate, but we do not believe that they will offer a large difference with that followed here. In general, the limiting factor is uncertainty associated with sampling individuals rather than the number of SNPs (Waples et al., 2022), provided these are sufficiently numerous and are uniformly distributed across the genome, and the inferences obtained refer to the particular pedigree of the individuals sampled (Ralph, 2019).

The method of Santiago et al. (2020) has now been applied to different species, particularly in the last year, including insects, such as honeybees (Sang et al., 2022); birds, such as Black Robin (von Seth et al., 2022); fishes, such as turbot, seabream and seabass (Saura et al., 2021), Baltic herring (Atmore et al., 2022), pikeperch (De Los Ríos-Pérez et al., 2022), coho salmon (Martinez et al., 2022), catfish (Coimbra et al., 2023) and sailfish (Ferrette et al., 2023); wild mammals, such as grey wolf (Pacheco et al., 2022), killer whales (Kardos et al., 2023), sika deer (Iijima et al., 2023), scimitar-horned oryx (Humble et al., 2023) and gorilla (Alvarez-Estape et al., 2023); humans (Bird et al., 2023); domestic species, such as pigs (Krupa et al., 2022), cattle (Jin et al., 2022; Magnier et al., 2022), sheep (Djokic et al., 2023; Drzaic et al., 2022), horse (Criscione et al., 2022) and chicken (Gao et al., 2023; Liu et al., 2023); plants, such as walnut (Ding et al., 2022); crustaceans, such as Daphnia (Wersebe & Weider, 2023) and fungi (Singh et al., 2021). As suggested by Santiago et al. (2020), the method is generally reliable for about 200 generations in the past, although the software provides values up to about 600 generations. The analyses carried out to date have followed this recommendation considering between 20 and 200 generations, or up to 300–400 generations in a couple of studies (Atmore et al., 2022; von Seth et al., 2022) and up to 700–800 in others (Gao et al., 2023; Singh et al., 2021). It is important to note that the linkage disequilibrium signal is diminished over time so that very far events cannot generally be inferred with precision. Furthermore, the method is expected to be more reliable when a genetic map is available for the SNPs considered. However, it can also be used assuming a constant average recombination rate for the species, although with somewhat less precision (Figure S3). In fact, three quarter of the studies mentioned above assumed this latter approach, as on most occasions a genetic map for the species is not available or it is not too reliable.

As mentioned above, Santiago et al. (2020) recommended to ignore the information arising from pairs of SNPs showing a very large recombination frequency because the information from these pairs of SNPs is very sensitive to possible complications in the sampling of individuals, such as non-random sampling or recent admixture in the population. For example, if there is a recent admixture, a typical artefact observed is a recent huge increase in the estimated N_e followed by a sudden drop in the last four generations (see Figure 2f of Santiago et al., 2020). Our experimental design implies some structuring of the large base population. This structuring involves the above-mentioned artefact and it is enhanced when the analyses are made with a high cut-off c value (see Figure S3 and results from simulations in the Supplementary Material). Note that, because the estimation process is made considering blocks of generations, the last four have always an identical value and some authors have opted for disregarding these initial generations (Atmore et al., 2022) or even the first 25 (Sang et al., 2022). However, the artefact produced by admixture can be ameliorated at least partially by ignoring pairs of SNPs with a recombination frequency higher than 0.05 (Figure 2f of Santiago et al., 2020, and Figures S3 and S5–S10 of the Supplemental Material). This is the recommended and default option of the software GONE. Previous studies have followed this recommendation and some of them assumed an even lower threshold ($c < 0.01$; Ding et al., 2022; Alvarez-Estape et al., 2023; or $c < .02$, Kardos et al., 2023). Even so, this artefact may have occurred in some of the estimations made so far, for example, in two or three Croatian sheep breeds analysed by Drzaic et al. (2022) or one of the horse breeds analysed by Criscione et al. (2022). Note, however, that it is possible that a great decline in size can be expected to occur in the last few generations, for example, when a selection program has started within that period (Saura et al., 2021). Therefore, it is sometimes difficult to disentangle an artefact from a bona fide population change. The use of a threshold for c will depend on the number of SNPs available at short genetic distances, and it has to be borne in mind that disregarding SNPs at long genetic distances will also decrease the power to detect very recent changes in N_e, as shown in Figure S3.

For the analyses carried out so far, most of them have suggested declines in N_e at different times and extents. This is something expected, as most analyses have been carried out for populations which are known to have declined in size for different reasons, or for populations that have entered a selection program, such as those maintained in aquaculture settings (e.g., turbot, seabream and seabass, Saura et al., 2021; coho salmon, Martinez et al., 2022; or pikeperch, De Los Ríos-Pérez et al., 2022). One exception is the historical expansions detected by Magnier et al. (2022) for Mayotte and Madagascar cattle breeds started about 80 generations ago, and coincident with the arrival of the Europeans to the islands and the massive use of cattle. Another is the recent increase in population size detected for most of the honeybee populations analysed by Sang et al. (2022).

Our results show that when a drastic bottleneck is detected by the method, the ancestral N_e is usually overestimated. This was already noted by simulations carried out by Saura et al. (2021); see

their Additional File 3, Figure S2. Thus, caution should be taken about the inferences made for the ancestral N_e even though those of the recent N_e are generally reliable. In addition, it is sometimes easier to detect abrupt changes in N_e than constant population sizes when these are large. In fact, our estimates for the base population N_e showed a nearly constant size over generations, as expected, but the actual N_e value was probably underestimated by the autosomal data and clearly overestimated by the X-chromosome data.

Although the method experimentally examined in this work has been shown to be reliable, several issues must be further studied. The most important is the effect of admixture or non-random sampling in the analyses, which must be studied in more detail. This has implications particularly in the estimates of the most recent generations, which can show substantial artefacts depending on the *c*-value cut-off considered. Another issue is that we have applied the current method for haploid data with a heuristic approach, but the method should be more formally extended to sex chromosome or haploid data, and current work is in progress in this regard.

AUTHOR CONTRIBUTIONS

Design of experiments: AC and HQ; Experimental work: NPP and HQ; Analysis of sequencing data: IN and HQ; Estimation procedure: IN, ES, HQ and AC; Writing – review and editing: IN, NPP, ES, HQ and AC.

ACKNOWLEDGEMENTS

We thank Joachim Mergeay and two anonymous referees for their useful comments on the manuscript. This work was funded by grants PID2020-114426GB-C21 (MCIN/AEI/10.13039/501100011033), Xunta de Galicia (ED431C 2020-05), Centro singular de investigación de Galicia accreditation 2019-2022 and "ERDF A way of making Europe". N.P.-P. was founded by a predoctoral (FPU16/02299) grant from Ministerio de Educación, Cultura y Deporte (Spain). I.N. was funded by a predoctoral (FPU18/04642) grant from Ministerio de Ciencia, Innovación y Universidades (Spain). Funding for open access charge: Universidade de Vigo/CISUG. Open Access funding provided thanks to the CRUE-CSIC agreement with Wiley. We thank Raquel Sampedro for technical help with the flies and Mary Riádigos for administrative support.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Computer scripts used in the analysis are available at Github address https://github.com/irene-novo-g/Historical-Ne-Drosophila. FASTQ files have been deposited in the NCBI Sequence Read Archive SRA database under accession numbers SAMN35298180-SAMN35298247 (BioProject accession: PRJNA974804).

BENEFIT SHARING STATEMENT

Our group is committed to international scientific partnerships, as well as institutional capacity building.

ORCID

Irene Novo https://orcid.org/0000-0001-7187-1872
Noelia Pérez-Pereira https://orcid.org/0000-0002-4731-3712
Enrique Santiago https://orcid.org/0000-0002-5524-5641
Humberto Quesada https://orcid.org/0000-0002-2927-0915
Armando Caballero https://orcid.org/0000-0001-7391-6974

REFERENCES

Alvarez-Estape, M., Pawar, H., Fontsere, C., Trujillo, A. E., Gunson, J. L., Bergl, R. A., Bermejo, M., Linder, J. M., McFarland, K., Oates, J. F., Sunderland-Groves, J. L., Orkin, J., Higham, J. P., Viaud-Martinez, K. A., Lizano, E., & Marques-Bonet, T. (2023). Past connectivity but recent inbreeding in Cross River gorillas determined using whole genomes from single hairs. *Genes*, *14*(3), 743. https://doi.org/10.3390/genes14030743

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Atmore, L. M., Martínez-García, L., Makowiecki, D., André, C., Lõugas, L., Barrett, J. H., & Star, B. (2022). Population dynamics of Baltic herring since the Viking age revealed by ancient DNA and genomics. *Proceedings of the National Academy of Sciences*, *119*(45), e2208703119. https://doi.org/10.1073/pnas.2208703119

Barbato, M., Orozco-terWengel, P., Tapio, M., & Bruford, M. W. (2015). SNeP: A tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Frontiers in Genetics*, *6*, 109. https://doi.org/10.3389/fgene.2015.00109

Bird, N., Ormond, L., Awah, P., Caldwell, E. F., Connell, B., Elamin, M., Fadlelmola, F. M., Matthew Fomine, F. L., López, S., MacEachern, S., Moñino, Y., Morris, S., Näsänen-Gilmore, P., Nketsia, V. N. K., Veeramah, K., Weale, M. E., Zeitlyn, D., Thomas, M. G., Bradman, N., & Hellenthal, G. (2023). Dense sampling of ethnic groups within African countries reveals fine-scale genetic structure and extensive historical admixture. *Science Advances*, *9*(13), eabq2616. https://doi.org/10.1126/sciadv.abq2616

Boitard, S., Rodríguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data—An approximate Bayesian computation approach. *PLoS Genetics*, *12*(3), e1005877. https://doi.org/10.1371/journal.pgen.1005877

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Browning, S. R., & Browning, B. L. (2015). Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *American Journal of Human Genetics*, *97*(3), 404–418. https://doi.org/10.1016/j.ajhg.2015.07.012

Caballero, A. (2020). *Quantitative Genetics*. Cambridge University Press.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*, 195–205.

Charlesworth, B., Morgan, M. T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, *134*, 1289–1303.

Coimbra, M. R. M., Farias, R. S., da Silva, B. C. N. R., Blanco, A., Hermida, M., Caballero, A., Bekaert, M., & Martinez, P. (2023). A genetic linkage map of the threatened catfish Lophiosilurus alexandri: Inferences on effective population size. *Aquaculture and Fisheries*, *8*(6), 689–694. https://doi.org/10.1016/j.aaf.2023.02.003

Comeron, J. M., Ratnappan, R., & Bailin, S. (2012). The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genetics*, *8*(10), e1002905. https://doi.org/10.1371/journal.pgen.1002905

Criscione, A., Mastrangelo, S., D'Alessandro, E., Tumino, S., Di Gerlando, R., Zumbo, A., Marletta, D., & Bordonaro, S. (2022). Genome-wide

survey on three local horse populations with a focus on runs of homozygosity pattern. *Journal of Animal Breeding and Genetics*, *139*(5), 540–555. https://doi.org/10.1111/jbg.12680

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

De Los Ríos-Pérez, L., Druet, T., Goldammer, T., & Wittenburg, D. (2022). Analysis of Autozygosity using whole-genome sequence data of full-sib families in pikeperch (*Sander lucioperca*). *Frontiers in Genetics*, *12*, 786934. https://doi.org/10.3389/fgene.2021.786934

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One*, *8*(12), e85024. https://doi.org/10.1371/journal.pone.0085024

Ding, Y. M., Cao, Y., Zhang, W. P., Chen, J., Liu, J., Li, P., Renner, S. S., Zhang, D. Y., & Bai, W. N. (2022). Population-genomic analyses reveal bottlenecks and asymmetric introgression from Persian into iron walnut during domestication. *Genome Biology*, *23*(1), 145. https://doi.org/10.1186/s13059-022-02720-z

Djokic, M., Drzaic, I., Shihabi, M., Markovic, B., & Cubric-Curik, V. (2023). Genomic diversity analyses of some indigenous Montenegrin sheep populations. *Diversity*, *15*(5), 640. https://doi.org/10.3390/d15050640

Drzaic, I., Curik, I., Lukic, B., Shihabi, M., Li, M. H., Kantanen, J., Mastrangelo, S., Ciani, E., Lenstra, J. A., & Cubric-Curik, V. (2022). High-density genomic characterization of native Croatian sheep breeds. *Frontiers in Genetics*, *13*, 940736. https://doi.org/10.3389/fgene.2022.940736

Ferrette, B. L. D. S., Coimbra, R. T. F., Winter, S., De Jong, M. J., Williams, S. M., Coelho, R., Rosa, D., Rotundo, M. M., Arocha, F., Mourato, B. L., Mendonça, F. F., & Janke, A. (2023). Seascape genomics and Phylogeography of the sailfish (Istiophorus platypterus). *Genome Biology and Evolution*, *15*(4), evad042. https://doi.org/10.1093/gbe/evad042

Frankham, R. (2021). Suggested improvements to proposed genetic indicator for CBD. *Conservation Genetics*, *22*, 531–532. https://doi.org/10.1007/s10592-021-01357-y

Gao, C., Du, W., Tian, K., Wang, K., Wang, C., Sun, G., Kang, X., & Li, W. (2023). Analysis of conservation priorities and runs of homozygosity patterns for Chinese indigenous chicken breeds. *Animals*, *13*(4), 599.

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

Gramates, L. S., Agapite, J., Attrill, H., Calvi, B. R., Crosby, M. A., dos Santos, G., Goodman, J. L., Goutte-Gattat, D., Jenkins, V. K., Kaufman, T., Larkin, A., Matthews, B. B., Millburn, G., & Strelets, V. B. (2022). The FlyBase consortium. FlyBase: A guided tour of highlighted features. *Genetics*, *220*(4), iyac035. https://doi.org/10.1093/genetics/iyac035

Hayes, B. J., Visscher, P. M., McPartlan, H. C., & Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research*, *13*(4), 635–643. https://doi.org/10.1101/gr.387103

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetical Research*, *38*(3), 209–216. https://doi.org/10.1017/S0016672300020553

Hoban, S., Bruford, M., Jackson, J. D. U., Lopes-Fernandes, M., Heuertz, M., Hohenlohe, P. A., Paz-Vinas, I., Sjögren-Gulve, P., Segelbacher, G., Vernesi, C., Aitken, S., Bertola, L. D., Bloomer, P., Breed, M., Rodríguez-Correa, H., Funk, W. C., Grueber, C. E., Hunter, M. E., Jaffe, R., … Laikre, L. (2020). Genetic diversity targets and indicators in the CBD post-2020 global biodiversity framework must be improved. *Biological Conservation*, *248*, 108654. https://doi.org/10.1016/j.biocon.2020.108654

Hollenbeck, C. M., Portnoy, D. S., & Gold, J. R. (2016). A method for detecting recent changes in contemporary effective population size from linkage disequilibrium at linked and unlinked loci. *Heredity*, *117*(4), 207–216. https://doi.org/10.1038/hdy.2016.30

Hudson, R. R., & Kaplan, N. L. (1995). Deleterious background selection with recombination. *Genetics*, *141*, 1605–1617.

Humble, E., Stoffel, M. A., Dicks, K., & Ogden, R. (2023). Conservation management strategy impacts inbreeding and mutation load in scimitar-horned oryx. *PNAS*, *120*(18), e2210756120.

Iijima, H., Nagata, J., Izuno, A., Uchiyama, K., Akashi, N., Fujiki, D., & Kuriyama, T. (2023). Current sika deer effective population size is near to reaching its historically highest level in the Japanese archipelago by release from hunting rather than climate change and top predator extinction. *The Holocene*, *33*(6), 718–727. https://doi.org/10.1177/09596836231157063

Jin, H., Zhao, S., Jia, Y., & Xu, L. (2022). Estimation of linkage disequilibrium, effective population size, and genetic parameters of phenotypic traits in Dabieshan cattle. *Genes*, *14*(1), 107. https://doi.org/10.3390/genes14010107

Kamm, J., Terhorst, J., Durbin, R., & Song, Y. S. (2020). Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, *115*(531), 1472–1487. https://doi.org/10.1080/01621459.2019.1635482

Kardos, M., Zhang, Y., Parsons, K. M., Yunga A., Kang, H., Xu, X., Liu, X., Matkin, C. O., Zhang, P., Ward, E. J., Hanson, M. B., Emmons, C., Ford, M. J., Fan, G., & Li, S. (2023). Inbreeding depression explains killer whale population dynamics. *Nature Ecology and Evolution*, *7*, 675–686. https://doi.org/10.1038/s41559-023-01995-0

Krupa, E., Moravčíková, N., Krupová, Z., & Žáková, E. (2022). Assessment of the genetic diversity of a local pig Breed using pedigree and SNP data. *Genes*, *12*(12), 1972. https://doi.org/10.3390/genes12121972

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. https://doi.org/10.48550/arXiv.1303.3997

Liu, C., Wang, D., He, Y., Liang, W., Li, W., Wang, K., Li, D., Li, Z., Tian, Y., Kang, X., & Sun, G. (2023). Population structure and genetic diversity analysis of "Yufen 1" H line chickens using whole-genome resequencing. *Life*, *13*(3), 793. https://doi.org/10.3390/life13030793

Liu, X., & Fu, Y. X. (2020). Stairway plot 2: Demographic history inference with folded SNP frequency spectra. *Genome Biology*, *21*, 280. https://doi.org/10.1186/s13059-020-02196-9

López-Cortegano, E., Vilas, A., Caballero, A., & García-Dorado, A. (2016). Estimation of genetic purging under competitive conditions. *Evolution*, *70*(8), 1856–1870. https://doi.org/10.1111/evo.12983

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. https://doi.org/10.1007/s10592-010-0050-7

Magnier, J., Druet, T., Naves, M., Ouvrard, M., Raoul, S., Janelle, J., Moazami-Goudarzi, K., Lesnoff, M., Tillard, E., Gautier, M., & Flori, L. (2022). The genetic history of Mayotte and Madagascar cattle breeds mirrors the complex pattern of human exchanges in Western Indian Ocean. *G3 (Bethesda)*, *12*(4), jkac029. https://doi.org/10.1093/g3journal/jkac029

Martinez, V., Dettleff, P. J., Galarce, N., Bravo, C., Dorner, J., Iwamoto, R. N., & Naish, K. (2022). Estimates of effective population size in commercial and hatchery strains of Coho Salmon (*Oncorhynchus kisutch* (Walbaum, 1792)). *Animals (Basel)*, *12*(5), 647. https://doi.org/10.3390/ani12050647

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. https://doi.org/10.1371/journal.pgen.1009764

Nunney, L. (1993). The influence of mating system and overlapping generations on effective population size. *Evolution*, *47*(5), 1329–1341. https://doi.org/10.1111/j.1558-5646.1993.tb02158.x

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2016). Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), 292–294. https://doi.org/10.1093/bioinformatics/btv566

Pacheco, C., Stronen, A. V., Jędrzejewska, B., Plis, K., Okhlopkov, I. M., Mamaev, N. V., Drovetski, S. V., & Godinho, R. (2022). Demography and evolutionary history of grey wolf populations around the Bering Strait. *Molecular Ecology*, *31*(18), 4851–4865. https://doi.org/10.1111/mec.16613

Pérez-Pereira, N., Pouso, R., Rus, A., Vilas, A., López-Cortegano, E., García-Dorado, A., Quesada, H., & Caballero, A. (2021). Long-term exhaustion of the inbreeding load in *Drosophila melanogaster*. *Heredity*, *127*, 373–383. https://doi.org/10.1038/s41437-021-00464-3

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Ralph, P. L. (2019). An empirical approach to demographic inference with genomic data. *Theoretical Population Biology*, *127*, 91–101. https://doi.org/10.1016/j.tpb.2019.03.005

Reid, B. N., & Pinsky, M. L. (2022). Simulation-based evaluation of methods, data types, and temporal sampling schemes for detecting recent population declines. *Integrative and Comparative Biology*, *62*(6), 1849–1863. https://doi.org/10.1093/icb/icac144

Sang, H., Li, Y., & Sun, C. (2022). Conservation genomic analysis of the Asian honeybee in China reveals climate factors underlying its population decline. *Insects*, *13*(10), 953. https://doi.org/10.3390/insects13100953

Santiago, E., & Caballero, A. (1998). Effective size and polymorphism of linked neutral loci in populations under directional selection. *Genetics*, *149*, 2105–2117. https://doi.org/10.1093/genetics/149.4.2105

Santiago, E., & Caballero, A. (2016). Joint prediction of the effective population size and the rate of fixation of deleterious mutations. *Genetics*, *204*(3), 1267–1279. https://doi.org/10.1534/genetics.116.188250

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. https://doi.org/10.1093/molbev/msaa169

Saura, M., Caballero, A., Santiago, E., Fernández, A., Morales-González, E., Fernández, J., Cabaleiro, S., Millán, A., Martínez, P., Palaiokostas, C., Kocour, M., Aslam, M. L., Houston, R. D., Prchal, M., Bargelloni, L., Tzokas, K., Haffray, P., Bruant, J. S., & Villanueva, B. (2021). Estimates of recent and historical effective population size in turbot, seabream, seabass and carp selective breeding programmes. *Genetics, Selection, Evolution*, *53*(1), 85. https://doi.org/10.1186/s12711-021-00680-9

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925. https://doi.org/10.1038/ng.3015

Singh, N. K., Karisto, P., & Croll, D. (2021). Population-level deep sequencing reveals the interplay of clonal and sexual reproduction in the fungal wheat pathogen *Zymoseptoria tritici*. *Microbial Genomics*, *7*(10), 000678. https://doi.org/10.1099/mgen.0.000678

Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, *51*(9), 1321–1329. https://doi.org/10.1038/s41588-019-0484-x

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology*, *2*(2), 125–141. https://doi.org/10.1016/0040-5809(71)90011-6

Van der Auwera, G. A., & O'Connor, B. D. (2020). Genomics in the cloud: Using Docker. In *GATK, and WDL in Terra* (1st ed.). O'Reilly Media.

von Seth, J., van der Valk, T., Lord, E., Sigeman, H., Olsen, R. A., Knapp, M., Kardailsky, O., Robertson, F., Hale, M., Houston, D., Kennedy, E., Dalén, L., Norén, K., Massaro, M., Robertson, B. C., & Dussex, N. (2022). Genomic trajectories of a near-extinction event in the Chatham Island black robin. *BMC Genomics*, *23*(1), 747. https://doi.org/10.1186/s12864-022-08963-1

Waldman, S., Backenroth, D., Harney, É., Flohr, S., Neff, N. C., Buckley, G. M., Fridman, H., Akbari, A., Rohland, N., Mallick, S., Olalde, I., Cooper, L., Lomes, A., Lipson, J., Cano Nistal, J., Yu, J., Barzilai, N., Peter, I., Atzmon, G., … Reich, D. (2022). Genome-wide data from medieval German Jews show that the Ashkenazi founder event predated the 14th century. *Cell*, *185*(25), 4703–4716.e16. https://doi.org/10.1016/j.cell.2022.11.002

Wang, J., Santiago, E., & Caballero, A. (2016). Prediction and estimation of effective population size. *Heredity*, *117*, 193–206. https://doi.org/10.1038/hdy.2016.43

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. https://doi.org/10.1007/s10592-005-9100-y

Waples, R. S., Waples, R. K., & Ward, E. J. (2022). Pseudoreplication in genomic-scale data sets. *Molecular Ecology Resources*, *22*, 503–518. https://doi.org/10.1111/1755-0998.13482

Wersebe, J., & Weider, L. J. (2023). Resurrection genomics provides molecular and phenotypic evidence of rapid adaptation to salinization in a keystone aquatic species. *Proceedings of the National Academy of Sciences*, *120*(6), e2217276120. https://doi.org/10.1073/pnas.2217276120

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*(2), 97–159. https://doi.org/10.1093/genetics/16.2.97

Wright, S. (1933). Inbreeding and homozygosis. *Proceedings of the National Academy of Sciences*, *19*(4), 411–420. https://doi.org/10.1073/pnas.19.4.411

Wright, S. (1938). Size of population and breeding structure in relation to evolution. *Science*, *87*, 430–431.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Novo, I., Pérez-Pereira, N., Santiago, E., Quesada, H., & Caballero, A. (2023). An empirical test of the estimation of historical effective population size using *Drosophila melanogaster*. *Molecular Ecology Resources*, *00*, 1–9. https://doi.org/10.1111/1755-0998.13837