Estimating contemporary effective population size from SNP data while accounting for mating structure.

Enrique Santiago[1], Armando Caballero[2], Carlos Kopke[3], and Irene Novo[4]

[1]University of Oviedo
[2]Affiliation not available
[3]Plasma Labs Enterprises SL
[4]Universidad de Vigo

September 21, 2023

Abstract

A new method is developed to estimate the contemporary effective population size (N_e) from linkage disequilibrium between SNPs without information on their location, which is the usual scenario in non-model species. The general theory of linkage disequilibrium is extended to include the contribution of full-sibs to the measure of LD, leading naturally to the estimation of Ne in monogamous and polygamous mating systems, as well as in multiparous species, and non-random distributions of full-sib family size due to selection or other causes. The prediction of confidence intervals for N_e estimates was solved using a small artificial neural network trained on a dataset of over 10^5 simulation results. The method, implemented in a user-friendly and fast software (*currentNe*) is able to estimate N_e even in problematic scenarios with large population sizes or small sample sizes, and provides confidence intervals that are more consistent than parametric methods or resampling.

Estimating contemporary effective population size from SNP data while accounting for mating structure

Enrique Santiago[1], Armando Caballero[2], Carlos Köpke[3] and Irene Novo[2]

[1] Departamento de Biología Funcional, Facultad de Biología, Universidad de Oviedo, 33006 Oviedo, Spain

[2] Centro de Investigación Mariña, Universidade de Vigo, Facultade de Bioloxía, 36310 Vigo, Spain.

[3] Plasma Labs Enterprises SL, 33007 Oviedo, Spain.

Corresponding author: Enrique Santiago, Departamento de Biología Funcional, Facultad de Biología, Universidad de Oviedo, 33006 Oviedo, Spain. Email: esr@uniovi.es

Enrique Santiago: 0000-0002-5524-5641

Armando Caballero: 0000-0001-7391-6974

Irene Novo: 0000-0001-7187-1872

Running title: Contemporary N_e and mating structure.

Word count: 6,046

Abstract

A new method is developed to estimate the contemporary effective population size (N_e) from linkage disequilibrium between SNPs without information on their location, which is the usual scenario in non-model species. The general theory of linkage disequilibrium is extended to include the contribution of full-sibs to the measure of LD, leading naturally to the estimation of N_e in monogamous and polygamous mating systems, as well as in multiparous species, and non-random distributions of full-sib family size due to selection or other causes. The prediction of confidence intervals for N_e estimates was solved using a small artificial neural network trained on a dataset of over 10^5 simulation results. The method, implemented in a user-friendly and fast software (*currentNe*) is able to estimate N_e even in problematic scenarios with large population sizes or small sample sizes, and provides confidence intervals that are more consistent than parametric methods or resampling.

1- Introduction

The development of linkage disequilibrium (LD) between neutral sites is a cumulative process contributed by drift over generations (Hill and Robertson 1968). This accumulation is short-lived between loosely linked sites and between sites located in different chromosomes because chromosomal segregation and recombination rapidly remove LD generated by past drift. In contrast, the observed LD between closely linked sites is also due to drift events that occurred long ago. Thus, the demography and the recombination landscape shape the pattern of LD across the genome. Consequently, the observed pattern of LD between markers can be used to infer the demographic history of a population in terms of effective population size (N_e) if the genetic map of the markers is available (Hayes et al. 2003; Tenesa et al. 2007; Santiago et al. 2020). Although this requirement is not met for most species of interest in conservation biology, Waples (2006) and Waples et al. (2016) showed that contemporary N_e can be estimated from a set of unmapped markers. This solution relies on empirical modifications for sampling and linkage to improve the accuracy of the Weir and Hill (1980) equations for discrete generations in panmictic populations. In addition to random mating, specific equations for N_e under lifetime monogamy were also derived by Weir and Hill (1980) and Waples and Do (2008). A widely used software for estimating contemporary N_e from LD between markers accounting for random mating and monogamy is *NeEstimator* (Do et al. 2014). This software provides accurate estimates of N_e under these models in most scenarios. However, for small sample sizes it often produces estimates of N_e that are indistinguishable from infinity, especially when the population size is large. In addition, the accuracy of the method depends on the exclusion of rare alleles in the analyses, because these latter may bias the estimates.

Here, we present an alternative method based on a combined approach of theory and neural networks to estimate the contemporary N_e and the corresponding confidence intervals. The theory is extended to include the contribution of full-sibs to the measure of LD, thus accounting, not only for random mating and monogamy, but for more complex mating systems. We show that the number of full-sibs in a sample is the ultimate source of bias in N_e estimates in populations with complex mating

systems, as the generation of gametes is restricted to compartments of related individuals. If this effect were ignored, N_e would be underestimated in monogamous, polygamous, and multiparous species, i.e. with more than a young at a birth, with increased bias caused by the effect of selection on the variance of family size, since the incidence of full siblings increases. All these effects are synthesized into a single parameter, the expected number of full siblings, which can be estimated from the genetic information in the sample data. A user-friendly software (*currentNe*) was developed to apply the method, which produces precise confidence intervals of the N_e estimates from a small artificial neural network, and is relatively accurate with small samples taken from large populations, a situation which can be common in many analyses of wild species. The software can be applied to scenarios with complex mating systems and does not require of minor allele frequency pruning to increase accuracy. Extensive simulations were performed to assess the impact of deviations from the assumptions of the theory.

2- Materials and Methods

In this study, we first develop a prediction equation for the N_e of the few most recent generations as a function of the observed average LD among all possible locus pairs from a genotyping analysis. This theoretical work involves new developments in the general theory of N_e, involving a change of perspective on how mating systems affect the measure of LD and hence N_e estimates. Where previously special equations for different mating systems were applicable, here we show that the whole problem can be reduced to considering the number of full siblings in the population. An artificial neural network (ANN) was designed to solve the problem of predicting the confidence interval of N_e estimates. During the training process, the ANN is told the true N_e along with other population parameters, and the difference between the squared difference between the true and observed N_e values given in the output and those observed in the simulations was used to adjust the weights of each neuron using backpropagation.

2.1- The Theory

Santiago et al. (2020) derived an equation for LD, measured as the squared correlation coefficient between markers weighted by the product of their genetic variances δ^2 (Rogers 2014), in terms of N_e and the recombination frequency c. The equation is valid for monoecious and dioecious populations assuming random pairing, i.e. when each offspring results from a new random pairing. Weir and Hill (1980) showed that lifetime mating has an effect on the measure of LD. Here we demonstrate that this effect is entirely due to the increase in frequency of full siblings above that expected from random pairing in a population of N_e reproducers (Sections 1 to 4 in the Appendix). While recombination removes LD, it also generates a small amount of LD due to recombinant gametes derived from full siblings. This effect can be included in the equation as

$$\delta_c^2 = \frac{1 + c^2 + c^2\frac{k}{4}}{2N_e(1-(1-c)^2) + 2.2(1-c)^2} \quad (1)$$

where c is the recombination frequency between two sites in the genome and k is the expected number of full siblings that a randomly selected individual will have among the reproducers. The equation is derived in Sections 1 to 4 of the Appendix and the connections to the equations of Weir and Hill (1980) are shown in Section 7 of the Appendix. The third term in the numerator ($c^2k/4$) corresponds to the contribution of full siblings to LD: two recombinant gametes from two full siblings have a 1/8 probability of matching each other in allele copies at both sites (Figure SF3 in Section 4 of the Appendix). This circumstance reduces the sampling space of allelic combinations and consequently increases the drift effect on LD compared to random pairing. This peculiarity does not occur to any significant extent for any other level of relatedness in a randomly mating population with discrete generations, except for an individual with itself: two recombinant gametes from the same individual will match each other in allele copies at both sites with probability 1/2, leading to an increase in LD already considered in the second term of the numerator (the c^2). Half siblings, which are common in

115 polygamy, share only one parent, while the other two are expected to be unrelated. Consequently, the
116 association of alleles at the two sites in recombinant gametes coming from two half siblings will not
117 match more often than two recombinant gametes from a random pair of unrelated individuals.

118 The generalization is that when genomes are arranged in pairs within diploid individuals or in
119 larger groups of full-sib families, the expectation of LD increases because recombination is restricted
120 to fixed pairs of haplotypes, leading to a higher probability of identical combinations of alleles in
121 recombinant gametes compared to the random haploid model. In this model, each haploid offspring is
122 generated by meiosis from a new random mating of two haploid parents, so recombination is not
123 restricted to paired genomes in diploid individuals, and the numerator of equation (1) equals "1" (see
124 Appendix in Santiago et al. 2020).

125 If there were full lifetime monogamy, i.e. lifetime monogamy for the whole population, any
126 individual selected from a population with random distribution of family size is expected to have two
127 full siblings (i.e., $k = 2$) while the expectation is of the order of N_e^{-1} if each offspring is generated by a
128 new random mating. Although a more formal derivation for monogamy is given in Section 3 of the
129 Appendix, there is a simple demonstration that applies to all the scenarios discussed below: If we
130 sample a single reference offspring among $N_e/2$ full-sib families, each of the subsequent offspring
131 sampled from the same population will be a full sibling of the reference offspring with probability
132 $2/N_e$; therefore, the expected number of full siblings that the reference offspring has in the entire
133 population is the product of the population size and the probability of siblings: $k = (N_e - 1) \cdot$
134 $2/N_e \approx 2$. However, the expected size of his family is three full siblings (including himself), but not
135 two, because large families are sampled more often. In the case of lifetime polygamy, the expected
136 number of brothers is different from the number of sisters in the parental group. However, scaling the
137 number of fathers (N_m) and mothers (N_f) to the theoretical $N_e = 4N_mN_f/(N_m + N_f)$ (Wright 1933),
138 the applicable value of k for lifetime polygyny results $k = 4N_m/(N_m + N_f)$, since the probability of
139 two random offspring coming from the same full-sib family is $1/N_f$ (Section 3 in the Appendix). In
140 the theoretical condition of lifetime polyandry, N_m and N_f must be swapped in the above equations,

because $N_m > N_f$. The expectation k decreases proportionally with the rate of lifetime pairings. If the population is considered to be a mixture of lifetime pairings (say with a proportion of m) and random pairings, then k is reduced proportionally to m, i.e. the k values given above should be multiplied by m. Full siblings are also produced in multiparous species. A reproductive scheme with two litters per female of equal size, both sired by the same father, is equivalent to monogamy ($k = 2$), but if sired by different fathers $k = 1$. In general, $k = 2 / L_e$ where L_e is the effective number of litters per female:

$$L_e = \frac{L^2}{\sum_{i=1}^{s} L_i^2}, \text{ such that } \sum_{i=1}^{s} L_i = L$$

and L is the number of litters per female, s is the number of sires per female and L_i is the number of litters sired by father i. For example, an average of $L = 4$ litters per female, most of which (say 3.2 litters) are sired on average by a single male and the rest sired by another male, gives about L_e=1.47. Selection also affects the frequency of full siblings, as a few families contribute the most to the next generation. A simple equation for the effect of selection on k (Section 3 in the Appendix) is

$$k = \frac{V}{M} + M - 1$$

where V and M are the variance and the mean, respectively, of the full-sib family size. With random contribution of families to the next generation (i.e. Poisson distribution of family size), V is equal to M and the equation reduces to $k = M$. However, under selection, V is expected to be greater than M and k may even be greater than 2, which is the expected value for constant population size and full lifetime monogamy.

The parameter k can also be estimated empirically from the observed frequency of full siblings in the sample k_s, regardless of the mating structure and the way in which selection increases the variance of family size (Section 3 in the Appendix):

$$\hat{k} = \frac{N_e}{n-1} k_s$$

(2)

where $k_s = \frac{\sum_{i=1}^{n} h_i}{n}$ is the particular value of k in the sample, n is the sample size and h_i is the number of full siblings that individual i has in the sample. The estimation of k requires the value of N_e, which is usually the unknown in equation (1). However, k can be estimated recursively together with N_e in the system of equations (1) and (2), provided that δ_c^2, i.e. the LD in the population, is known. This is likely to be the method of choice in many scenarios, as the details of the mating system and how selection operates are almost always not well known.

Equation (1) shows that the effect of full siblings on LD is rather small, especially under tight linkage. However, full siblings also bias the measure of LD from samples of unphased genotypes because they become more similar in terms of allele association at both loci, since the phase of the alleles in the gametes inherited from their common parents cannot be distinguished (Section 5 in the Appendix). The result is an increase in the measure of LD. If both effects on population LD and on sampling were ignored, N_e could be severely underestimated, especially under loose linkage (Figure S1), which is the dominant condition in the derivations below.

Equation (1) is valid for the whole range of recombination frequencies from 0 to 0.5 (Tables S1 and S2), and therefore a genome-wide prediction of LD can be obtained by integrating all possible site pairs. Suppose an imaginary species has v chromosomes, each of length l Morgans. Two random sites in this genome are in different chromosomes with probability $(v-1)/v$ and for them c is 0.5. If two random sites are in the same chromosome (probability = $1/v$), the density of their genetic distance follows a triangular distribution with the highest frequency corresponding to tight linkage (i.e., $c = 0$) and the lowest frequency corresponding to the maximum distance l in the chromosome. Assuming a random distribution of recombination events (i.e., Haldane's map function $c = (1-e^{-2x})/2$, where x is the genetic distance in Morgans), the expected LD between two random sites ($\overline{\delta^2}$) in the population can be calculated as the average over all pairs of sites:

$$\overline{\delta^2} = \frac{v-1}{v} \cdot \delta_{0.5}^2 + \frac{1}{v} \cdot \frac{1}{l/2} \cdot \int_0^l \frac{l-x}{l} \cdot \delta_c^2 \cdot dx \quad (3)$$

The first term on the right-hand side corresponds to pairs of sites in different chromosomes and the second term corresponds to pairs in the same chromosome; the former are expected to be much more common than the latter in most cases unless the number of chromosomes is very small. After including the effect of sampling, the resulting equation (S8) (Section 6 in the Appendix) predicts the LD from a sample taken from a population with a particular incidence of full siblings. The predictions are remarkably similar regardless of how the genome is distributed among the chromosomes: v chromosomes of length l Morgans each are effectively equivalent to a single chromosome of length vl Morgans, unless the chromosomes are very small. Thus, knowing the number of chromosomes (the average size is close to 1 Morgan per chromosome across species - Otto and Payseur 2019) or the genetic size of the genome of the species of interest, equation (3) can be solved for N_e from the observed LD ($\overline{\delta^2}$) in a sample. Strictly speaking, the estimate is not for a particular generation, but it is a kind of N_e averaged over the most recent generations in the past. Figure 1 shows how past generations contribute to the current LD: N_e estimates for genomes with large genetic sizes correspond mainly to the two generations just before sampling, with the highest contribution from the most recent one, but estimates for small genomes are contributed by a long sequence of past generations. For the latter, demographic changes in the past could bias the estimate of the contemporary N_e.

The prediction method based on the above theory has been implemented in the *currentNe* software, which receives the genotyping information in *ped* format (Chang et al. 2015) and the expected number of full siblings k as an optional modifier. Figure 2 illustrates how N_e estimates using this software change when different k values are considered under full monogamy. When monogamy is ignored (i.e., k is set to 0) N_e is underestimated as indicated above. This bias is eliminated by using either the k value for full lifetime monogamy (i.e., k=2) or the k value estimated from the observed incidence of full siblings in the sample using equation (2). That is, no prior information is needed

about the particular mating system of the species or about how the differences in family size are distributed in the population, since the only parameter required is k, which can be estimated from the distribution of full siblings in the sample. The exclusion of full siblings from the sample, which may seem to be an option to avoid the bias caused by a supposed excess of relatives, leads on average to an overestimation of N_e by almost a factor of two, probably due to the distortion of the randomness of the sample. Selection increases the variance of family size, ultimately leading to a reduction in N_e (Wright 1938). However, selection also increases the expectation of k, leading to a further reduction in N_e estimates and complicating theoretical predictions. Nevertheless, the joint estimation of N_e and k takes into account the combined effects of both aspects (Figure 2).

2.2- The Neural Network

There is no consistent theory for predicting the confidence intervals of N_e estimates from LD data. Intuitively, the main factors affecting the sampling variance are the number of individuals in the sample, the number of markers, the genetic size of the genome and the effective population size to be estimated. However, the relationship between these factors appears to be complex. The raw data for LD analysis is typically a table of a number of individuals by a number of markers but the effect of increasing the number of individuals in the sample is not proportional to the effect of increasing the number of markers (Waples et al. 2022). Markers are transmitted from generation to generation in chromosomal blocks that are broken by recombination events, so markers are correlated in genealogies, especially if they are close together. Therefore, the genetic size of the genome also appears to be relevant.

ANNs are machine learning computational systems loosely modelled on biological neurons and are universal function approximators. They have been used successfully in many different fields, but have not been as widely adopted in population genetics. Here we used a supervised learning algorithm, in which the network is told the expected response and the difference between this expectation and the output of the ANN is used to adjust the weights of each neuron using

backpropagation. This ANN was designed and trained to approximate the error in N_e estimates using a dataset of 128,692 combinations of population size (from 10 to 100,000), sample size (from 8 to 107), number of markers (from 1,000 to 32,000) and genome size (from 5 to 30 Morgans) using simulations. The combinations were approximately equally distributed on the logarithmic scales of the three first variables and on the linear scale of the last one. N_e was estimated for each combination of the dataset using the *currentNe* software based on the LD between all markers.

The dataset was randomly split into two sets, an 80% for training and the remaining 20% for evaluation. A simple network architecture provided the best results and made it computationally efficient to train with minimal resources. This ANN consists of an input layer, two hidden layers and an output layer, and its architecture is shown in Figure 3. Its inputs are the genome size in Morgans and the common logarithms of the population size estimate, the sample size and the number of markers. The output layer approximates the squared difference between the logarithms of the estimate of N_e by *currentNe* and the true simulated N_e. None of the hidden layers have a bias component; they are the simplest form of multi-layer perceptron. During training, pruning of the connections between the input and the first hidden layer was used to improve the results while reducing the computational complexity during the training phase. It is important to note that the input parameters are normalized before training the network, which can lead to a loss of precision in the extreme cases. A set of combined ANNs specialized in different population/sample sizes may provide more accurate results, although this is left for future research.

Figure 3 shows the approximation of the sampling variance by the ANN as a function of the number of individuals in the sample and the number of SNPs: doubling the number of individuals is much better than doubling the number of markers when the number of markers is not small (say > 1,000) (Waples et al. 2022). The calculation of confidence intervals by ANN has also been included in the *currentNe* software.

2.3- Computer simulations

To test and compare the accuracy of N_e predictions, computer simulations were generally performed using the software SLiM (Haller and Messer 2019). This software simulates an individual-based forward Wright-Fisher model of reproduction. Random mating populations of constant population size (N = 100, 1000 and 10000 individuals) with discrete generations were run for up to 10,000 generations. A scenario with 20 chromosomes of 100 Mb length each was considered where the rate of recombination between nucleotides was assumed to be 10^{-8}, implying a total chromosome length of one Morgan. The mutation rate per nucleotide was assumed to be 10^{-8}, 0.8×10^{-9} and 2.0×10^{-10} for the three population sizes, respectively, in order to obtain a total number of SNPs for analysis between around 20,000 and 50,000. Two sample sizes were considered (n = 10 and 100 individuals). The number of simulation replicates varied between 300 and 600, depending on the population and sample size. A custom program was developed to generate the large dataset for training and testing the ANN. This program simulates discrete generations in an evolving randomly mating population. Genotypes were initially assigned according to a neutral distribution of allele frequencies. The population was then run for thousands of generations in order to achieve an approximate stable spectrum of LD across the genome prior to sampling.

2.4- Requirements of the software *currentNe*.

The *currentNe* program is written in C++ and has been tested on computers running Linux with the distributions Arch, Debian and Ubuntu. The minimum requirements are a 64-bit CPU and 3 Gb of free RAM space, although it runs faster on multi-processor systems. Input data must follow either the *vcf* format (Danecek et al. 2011), or the *ped* or *tped* PLINK formats (Chang et al. 2015). The execution of the software also requires either the approximate size of the genome in Morgans or the estimated number of chromosomes of the species. The output is a single file with concise genetic information and the estimate of N_e using Eq. (3) with the corresponding confidence interval. If the assignments of SNPs to chromosomes are also available in the input file, an additional estimate of N_e

is made based on Eq. (1), i.e. considering only pairs of SNPs located in different chromosomes. Usage information is available with the program using the -h modifier ("./currentNe -h").

3- Results

Predictions with *currentNe* were compared with those using the method of *NeEstimator.v2* (Do et al. 2014), a widely used software for estimating N_e from LD data, and the correction for linked markers from Waples et al. (2016). The results of these comparisons are shown in Figure 4. *NeEstimator* is based on an empirical equation for linkage calculated from simulation results (Waples and Do 2008), and corrected for the linkage of SNP also from simulation results (Waples et al. 2016). In comparison, *currentNe* performs particularly well in the scenarios where sample sizes are small, for which *NeEstimator.v2* generally provides negative estimates of N_e. With some variation in specific cases, the confidence intervals provided by the ANN are more realistic than those based on parametric methods, which tend to underestimate the true intervals, or resampling which tends to provide infinite upper bounds for small sample estimates. While LD methods based on genetic maps such as GONE (Santiago et al, 2020) are prone to bias due to map errors, estimates of contemporary N_e are largely unaffected, as expected (Figure S2). However, these latter are sensitive to the highly non-uniform distribution of markers, which effectively reduces the size of the genetic map because closely linked markers are more common than would be expected if sites were randomly distributed.

The effect of deviating from the assumptions of random sampling and panmixia is shown in Fig. S3. Sampling restricted to a subset of families causes a reduction in N_e estimates that is somewhat proportional to the reduction space of the sampling, i.e., the group of families that could be sampled. The theory implemented in *currentNe* also assumes that the spectrum of frequencies of the SNPs included in the analysis is representative of the frequencies in the population, so no MAF threshold should be applied. The use of SNP arrays with markers selected for high variability is quite equivalent to the application of a MAF threshold, leading to a significant reduction in N_e estimates at high MAF values (Figure S3). The theory also assumes a single random mating population. The consequences of

subdividing the population into different demes depend on both the migration rate and the sampling location. If only one subpopulation is sampled, the estimate will reflect the size of that subpopulation unless the migration rate m is very high. However, if the entire metapopulation is sampled, the result depends on the product $N_e m$: if $N_e m > 1$, the estimate reflects the size of the entire metapopulation, but as $N_e m$ becomes smaller than 1, the LD measured on the entire population increases and the estimate decreases even below the size of the subpopulations. These results are in agreement with those found by Waples and England (2011) and Ryman et al. (2019). Deviations from the discrete-generation assumption lead to overestimation or underestimation of N_e, depending on sampling: When a single cohort is sampled, the estimate falls between the expected N_e and the census of reproducers, but when all cohorts are sampled proportionally to their abundance, the true N_e is underestimated, consistent with Waples et al. (2014).

Figure 5 shows N_e estimates from samples taken from four real populations, that are expected to be very different in size. In addition to the N_e estimate based on all pairs of loci using Eq. (3), a second estimate based on pairs of loci on different chromosomes using Eq. (1) was performed, as for these species the assignment of markers to particular chromosomes is well known. This additional estimation is performed by *currentNe* if the assignments of SNPs to chromosomes is also available in the input file. This information is used by *currentNe* to identify pairs of SNPs located on different chromosomes, but the exact locations within the chromosomes, if available, are not used. Estimates from a domestic pig herd (N_e = 26 and 32 for whole genome and unlinked marker estimates, respectively), which was maintained at a roughly constant size for several generations prior to sampling (Saura et al. 2015), are consistent with estimates derived from observed genealogical information ($N_e \approx 24$). Estimates from a salmon sample, consisting of a mixture of individuals born between 1985 and 1992 from the River Dee in Scotland, are in close agreement with the result of a previous analysis of the same sample by Santiago et al. (2020) using the GONE software ($N_e \approx 200$). Both estimates for pig and salmon populations were made using the default option for the number of siblings, i.e. allowing the programme to estimate the corresponding k (0.69 for pig and 0.10 for salmon populations) value from the sample data. The analysis showed that full sib pairs were present

in the samples from the pig and salmon populations. The analysis of the Koryak population was carried out with a relatively small sample (n = 16), which explains the large confidence intervals. However, the central estimate is around two or three thousand individuals, which is about three times lower than the current census (Minority Rights Group International, 2018), which has not changed significantly in recent times (Novo et al. 2022). Estimates for the Finnish population were made using a random subset of SNPs for 99 individuals which are available at www.internationalgenome.org (1000 Genomes Project). The Finnish population has been growing rapidly for 20 generations, especially during the last century (O'Neill 2022). The apparent discrepancy between the N_e estimates, with confidence intervals in the range of a few tens of thousands to hundreds of thousands, and the observed census sizes of over 3,000,000 for the few most recent generations deserves more attention, as past demography also affects *currentNe* estimates. The estimate based on unlinked locus pairs is larger than the estimate based on all locus pairs (Figure 5), which is consistent with the observed increase in Finnish population size because linkage makes the estimates to be more affected by demography in past generations (Figure 1). Also, census size, which aggregates a wide range of ages, and N_e, which is more related with reproducer census size, could be of very different magnitudes in human populations, where N_e to census ratios are in the range of 0.1 - 0.6 (Felsenstein 1971; Frankham 1995; Urnikyte et al. 2017). The Koryak and Finnish populations were assumed to be monogamous because, even if a non-trivial fraction of progeny arises from extra-pair matings, the effect of reducing the degree of monogamy within the higher range, say between 0.6 and 1, is small (Figure S1). Consequently, the expected number of full siblings of a random individual was assumed to be 2, i.e. *k* was set to 2 using the optional *-k* modifier when running *currentNe*. The software is rather fast. For example, the analysis of the Koryaks sample with 16 individuals and around 90,000 SNPs took 1,5 minutes of computing time in a Linux x64 machine with 8 processors running in parallel. For the Finnish data, including 99 individuals and 100,000 SNPs, the computing time was 16 minutes.

4- Discussion

The integration of the equation for LD over the whole genome leads to a method for estimating contemporary N_e from a set of unmapped markers, which requires only the total length of the genetic map and assumes that markers are randomly distributed along the genome. Although methods to infer past demography using LD (e.g. GONE; Santiago et al. 2020) also provide estimates of the contemporary N_e, a detailed genetic map of markers is required for these methods to be applied. In contrast, map information is not available for the vast majority of the species, for which tools like *NeEstimator* or *currentNe* are the only options to estimate contemporary size. In addition, contemporary N_e estimates obtained by *currentNe* may be less biased than estimates by methods designed to reconstruct the whole demography (Figures S2 and S3).

The estimates of N_e are more consistent than the method of Waples et al. (2016), presumably due to the increased accuracy of the integral equation. Analyzing the details of the theory leads to a better understanding of what it is actually estimated by LD methods and facilitates the identification of the ultimate factors influencing the estimates. It was made clear that what is called contemporary N_e is actually an average of the most recent generations, starting from the generation prior to the sampling generation. The smaller the genome, the more generations contribute to the estimate. For average genomes of about 20 Morgans the estimated contemporary N_e is approximately the average of the two most recent generations in the past.

By following the theoretical derivation, in particular the composition of the cross products of the squared covariance between markers [see derivation of Eq. (S5) of the Appendix], it was possible to determine that full siblings were the ultimate cause of the deviation of estimates with lifetime pairing, either monogamy or polygamy, when compared with random pairing predictions. It is difficult to visualize the biological concepts by approximating eigenvectors for the transition matrix of two-locus descent measures as in Weir and Hill (1980), although this has the advantage of being a thorough method. Another advantage of our derivations is that the use of generic variables (X and Y in

390 the Appendix), without any assumption about their distribution, instead of allelic frequencies of
391 binomial distributions, facilitates the abstraction and consequently the interpretation of the results.
392 Since the general equation for N_e only requires the value of k, i.e. the expected number of full siblings
393 that a random individual has, it immediately follows the connections with other reproductive methods,
394 such as in multiparous species and with factors affecting the variance of family size, such as selection.

395 An important implication of our results is that because in natural scenarios one might find full
396 siblings in samples from populations of species with unknown mating systems, LD methods that
397 ignore this, will tend to produce downwardly biased estimates of contemporary N_e. The solution of
398 excluding full siblings from the sample, with the intention of compensating for this bias, introduces a
399 further bias in the opposite direction, as already found by Waples and Anderson (2017), who
400 considered the effect of purging siblings from samples to compensate for family overrepresentation.
401 The correct procedure for obtaining unbiased estimates is to include the k value estimated from the
402 sample in the analysis, provided that the sample is truly random. Estimates of contemporary N_e
403 obtained under the assumption of no or negligible numbers of full siblings should be reconsidered,
404 especially when analyses are performed to infer whether the effective size of the population is below
405 or above the minimum size assumed for its persistence (Frankham et al. 2014; Pérez-Pereira et al.
406 2022). It is important to note, however, that the presence of siblings is not expected to affect estimates
407 of past demography based on the LD of closely linked markers. In contrast to contemporary N_e
408 inference, which relies mainly on LD between pairs of loci on different chromosomes, inference of
409 past demography (Tenesa et al. 2007; Santiago et al. 2020) typically uses recombination rates between
410 markers in the lower range of c values, e.g. below c=0.1. As the effect of k on N_e estimates is scaled
411 by c^2, as shown in Eq. (1), the effective bias is expected to be negligible.

412 Another point of interest is the use of ANN to solve the complex problem of predicting the
413 sampling variance of the N_e estimates. Jackknife resampling has been found to be a good method for
414 estimating confidence intervals of N_e estimates (Do et al. 2014), but it often produces infinite intervals
415 with small samples. Although the jackknife resampling is an efficient method for estimating the

sampling variance, there are two reasons that are likely to reduce its efficiency in this particular case. First, samples are not independent individuals from a population, but the individuals are connected by genealogies that determine the measure of the correlations between markers, i.e. there is no simple variable that is measured on individuals. Second, the intervals have to be computed for a transformation of the correlation: the correlation is squared, then a correction factor for sampling is subtracted and, finally an operation similar to the inverse is performed to obtain the N_e estimate for each resampling. This leads to a high instability of the estimates, especially when the squared correlations are small, since the subtraction of the correction for sampling can lead to negative results.

The quality of the ANN solution depends on the quality of the training data. Here we tried to use a wide combination of parameters with a uniform data density distribution over the training space. To simplify the complexity of the network, cross connections were also pruned. This leads to a reduction in noise, especially the turbulence of the estimates at the boundaries of the training space. The resulting network, which like any other simple network, consists of an equation with input variables and an output, predicts confidence intervals of N_e as a function of the number of individuals in the sample, the number of markers, the genetic size of the genome and the estimated effective population size. This equation, included in the code of the *currentNe* program, is similar in accuracy to jackknife with large samples but with the advantage of working with small samples. ANN also makes it possible to visualize the main effects and interactions between factors that affect the sampling variance of N_e estimates, as shown in Figure 3.

Acknowledgements

The authors thank José Antonio Godoy, Josephine Pemberton, Robin Waples and two anonymous reviewers for useful comments. This study forms part of the Marine Science Programme (ThinkInAzul) supported by the Ministerio de Ciencia e Innovación and Xunta de Galicia with funding from the European Union NextGenerationEU (PRTR-C17.I1) and European Maritime and Fisheries Fund. We also acknowledge financial support from 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”. IN was funded by a predoctoral (FPU18 04642) grant from the Ministerio de Ciencia, Innovación y Universidades (Spain).

References

Chang C.C., Chow C.C., Tellier L.C.A.M., Vattikuti S., Purcell S.M. & Lee J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R. & 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156-2158.

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 Resour*ces, 14, 209-214.

Felsenstein J. (1971). Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics*, 68, 581– 597.

Frankham R., Bradshaw C.J. & Brook B.W. (2014). Genetics in conservation management: revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biological Conservation,* 170, 56e63.

Frankham R. (1995). Effective population size / adult population size ratios in wildlife: a review. *Genetical Research*, 66, 95-107.

Haller B.C. & Messer P.W. (2019). SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution*, 36, 632-637.

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, 635-43.

Hill W.G. & Robertson A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38, 226–231.

Minority Rights Group International. (2018). *World Directory of Minorities and Indigenous Peoples - Russian Federation : Koryaks*, available at: https://www.refworld.org/docid/49749cbcc.html [accessed 26 April 2023].

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.

Otto S. & Payseur B.A. (2019). Crossover interference: shedding light on the evolution of recombination. *Annual Review of Genetics,* 53, 19-44.

O'Neill A. (2022). Population of Finland 1750-2020. https://www.statista.com/statistics/1009145/total-population-finland-1750-2020/ [accessed 26 April 2023]

Pérez-Pereira N., Wang J., Quesada H. & Caballero A. (2022). Prediction of the minimum effective size of a population viable in the long term. *Biodiversity and Conservation,* 31, 2763–2780.

Rogers A. (2014). How population growth affects linkage disequilibrium. *Genetics,* 197, 1329–1341.

Ryman N., Laikre L. & Hössjer O. (2019). Do estimates of contemporary effective population size tell us what we want to know? *Molecular Ecology*, 28, 1904–1918.

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, 3642-3653.

Saura M., Tenesa A., Woolliams J.A., Fernández A. & Villanueva B. (2015). Evaluation of the linkage-disequilibrium method for the estimation of effective population size when generations overlap: an empirical case. *BMC Genomics* 16, 922.

Tenesa A., Navarro P., Hayes B.J., Duffy D.L., Clarke G.M., Goddard M.E. & Visscher P.M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research,* 17, 520-526.

Urnikyte A., Molyte A. & Kucinskas V. (2017). Recent effective population size estimated from segments of identity by descent in the Lithuanian population. *Anthropological Science*, 125(2), 53-58.

Waples R.S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics*, 7, 167–184.

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

Waples R.S. & England P.R. (2011). Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. *Genetics*, 189(2), 633-644.

Waples R.S. (2014). Effects of overlapping generations on linkage disequilibrium estimates of effective population size. *Genetics* 197(2), 769-780.

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, 233–240.

Waples R.S. & Anderson E.C. (2017). Purging putative siblings from population genetic data sets: a cautionary view. *Molecular Ecology* 26, 1211-1224.

Waples R.S., Waples R.K. & Ward E.J. (2022). Pseudo-replication in genomic-scale data sets. *Molecular Ecology Resour*ces, 22,503-518.

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

Wright S. (1933). Inbreeding and homozygosis. *Proceedings of the National Academy of Sciences* USA, 19, 411–420

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

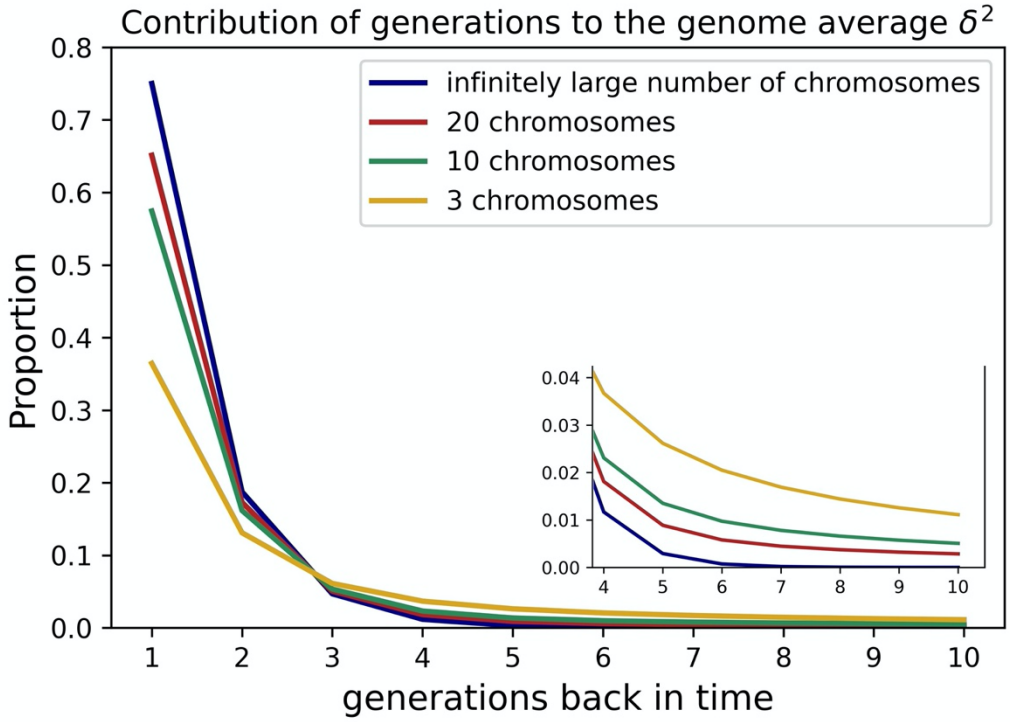
Data Accessibility

The software to apply this method is available on GitHub at https://github.com/esrud/currentNe.

Author Contributions

E.S. and A.C. conceived the work. E.S. developed the theory and the computational solution. E.S., A.C. and C.K. wrote the article. E.S., A.C. and. I.N. performed simulations. E.S. and C.K. developed the ANN solution.

527 Figure 1.



528

529 **Figure 1.** Contributions of past generations of populations with 1000 individuals to the average LD
530 (δ^2) over all site pairs for genomes with 3, 10, 20 and an infinite number of chromosomes of one
531 Morgan length. The latter corresponds to an effective recombination rate $c = 0.5$ between all marker
532 pairs. Generation 1 is the generation immediately preceding the generation of the sample. The small
533 section is an enlargement of the tail of the figure. Contributions were calculated using the cumulative
534 equations in the Supplementary File, Section 10 "Prediction of δ^2 when N changes", in Santiago et al.
535 (2020).

536

Figure 2.

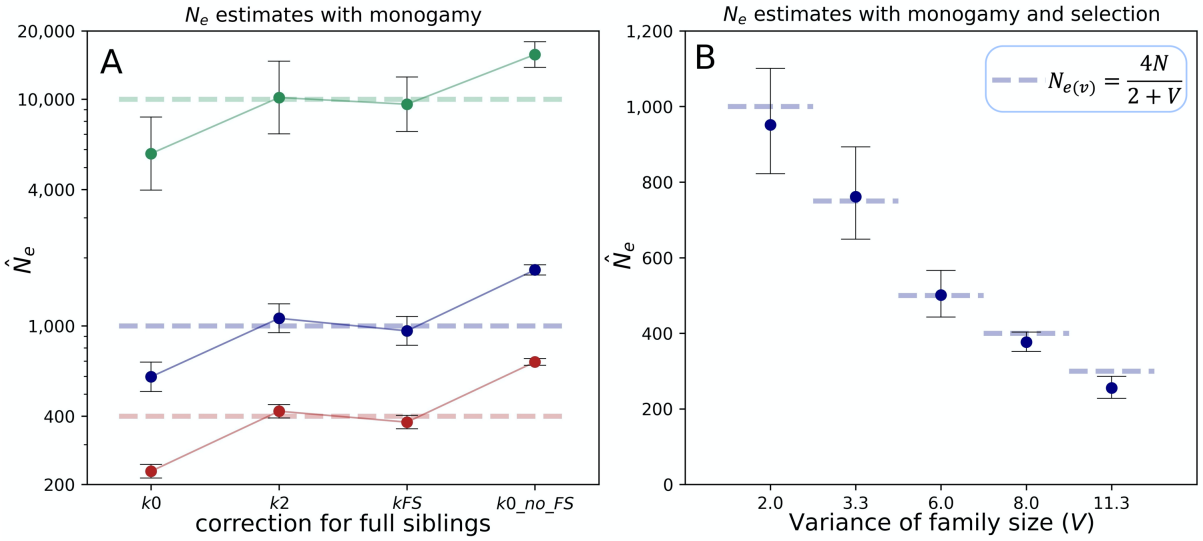


Figure 2. N_e estimates and 95% confidence intervals with complete lifetime monogamy and unphased genotypes. **A)** Each point is the average of 10 estimates, each with 100 samples, from simulated populations with census sizes 10,000, 1000 or 400 individuals (in green, blue and red from top to bottom). Analyses were performed using the *currentNe* software with four different options. With option *k0*, monogamy was erroneously ignored, i.e. the number of full siblings, *k* ,of a random individual was set to 0 ($k = 0$). With option *k2*, the correct value $k = 2$, corresponding to full lifetime monogamy, was used in the analysis. With the *kFS* option, *k* values were estimated using the observed incidence of full-sibs in each sample. With the *k0_no_FS* option, one random sibling from each pair of full-sibs was discarded from the samples and the analyses were performed with the incorrect assumption of no monogamy, i.e. $k = 0$. **B)** Each dot is the average of 10 estimates with 100 samples each from simulated populations with census size $N = 1{,}000$. Analyses were performed with *currentNe* using the *k* values estimated from the frequency of full siblings in the sample. Selection for a non-inherited trait was applied to families at different intensities in such a way that the variance V of family size increased above its expected value in the absence of selection (i.e. 2.0). Consequently, the predicted "variance effective number" $N_{e(v)}$ (dashed lines and equation from Wright (1938) in the legend) decreases.

Figure 3.

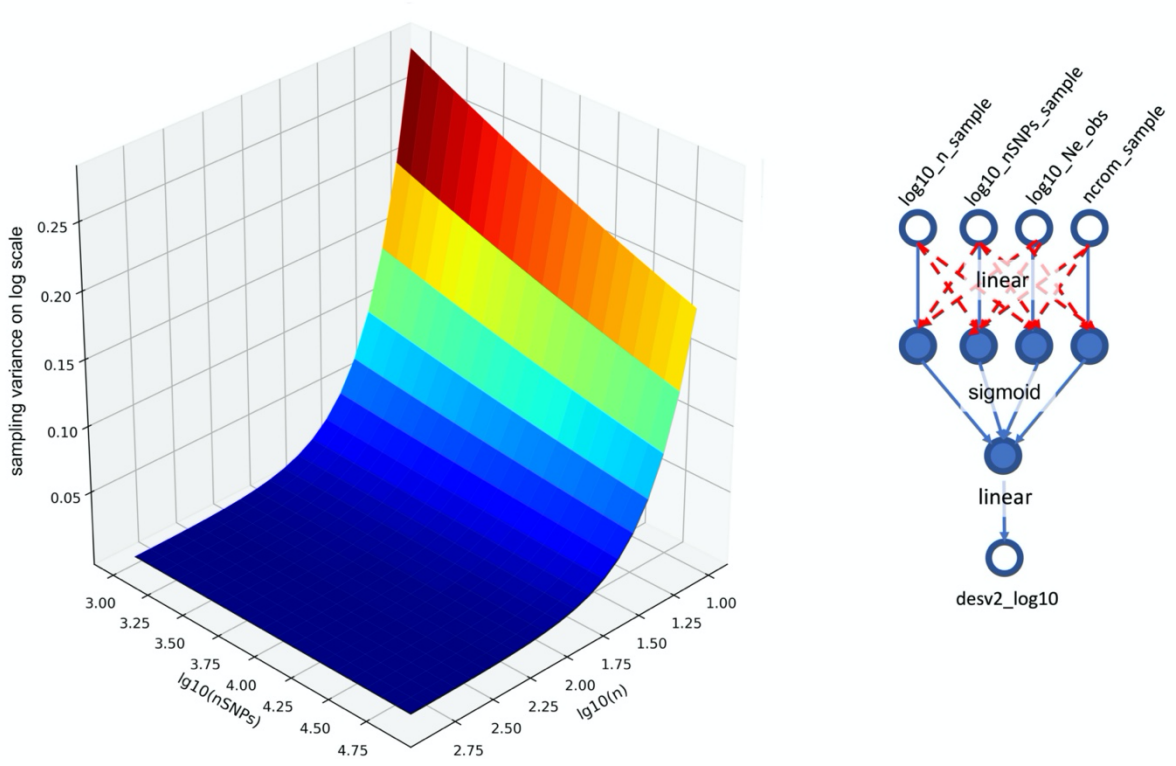
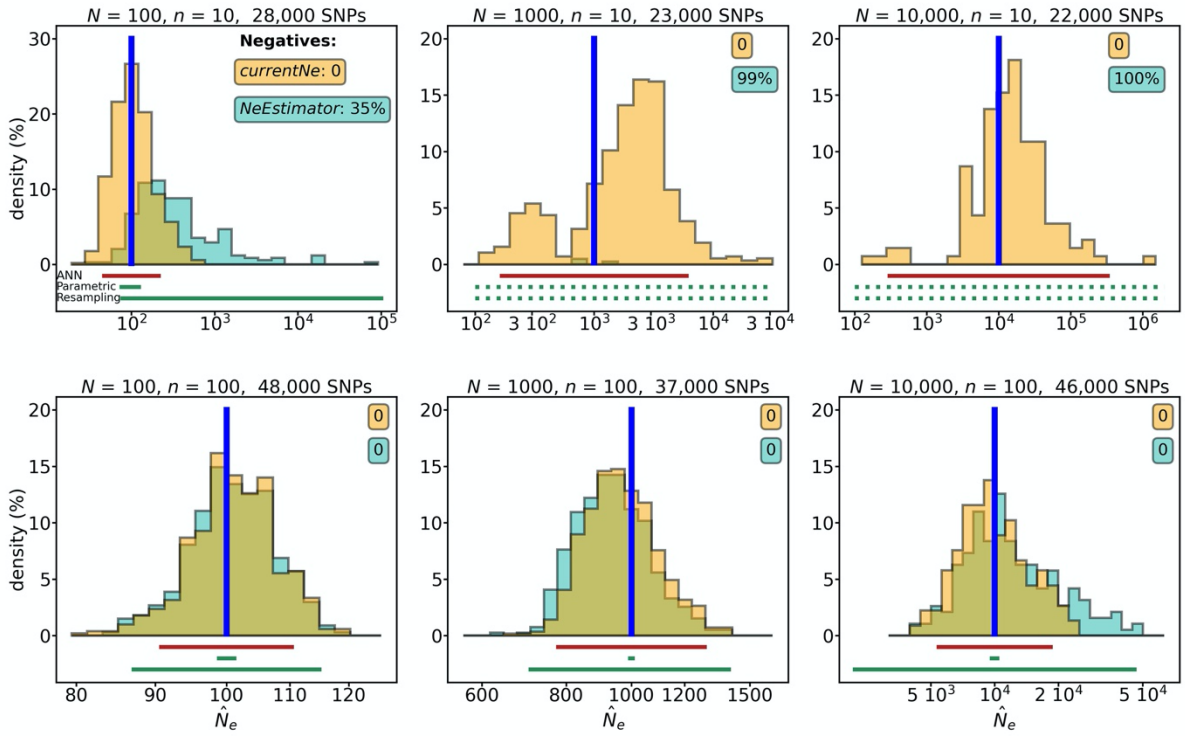


Figure 3. Sampling variance of N_e estimates on a base 10 logarithmic scale (vertical axis) generated by the ANN. This 3D plot corresponds to populations with N = 1,000 individuals and 20 chromosomes of 1 Morgan length. The sampling variance is given as a function of the number of individuals in the sample (n) and the number of markers (nSNPs), both in logarithmic scale. The small graph at the side of the plot corresponds to the design of the ANN. The four nodes of the input layer are the number of individuals in the sample (log10_n_sample), the number of SNPs used in the sample data (log10_nSNPs_sample), the genome size in Morgans (ncrom_sample) and the estimate of N_e from the data (log10_Ne_obs). The output layer is the squared difference between the N_e estimate and the true N_e on a logarithmic scale. The labels on the connections refer to the activation functions used between layers. Red connections were pruned during the training period.

570 Figure 4.

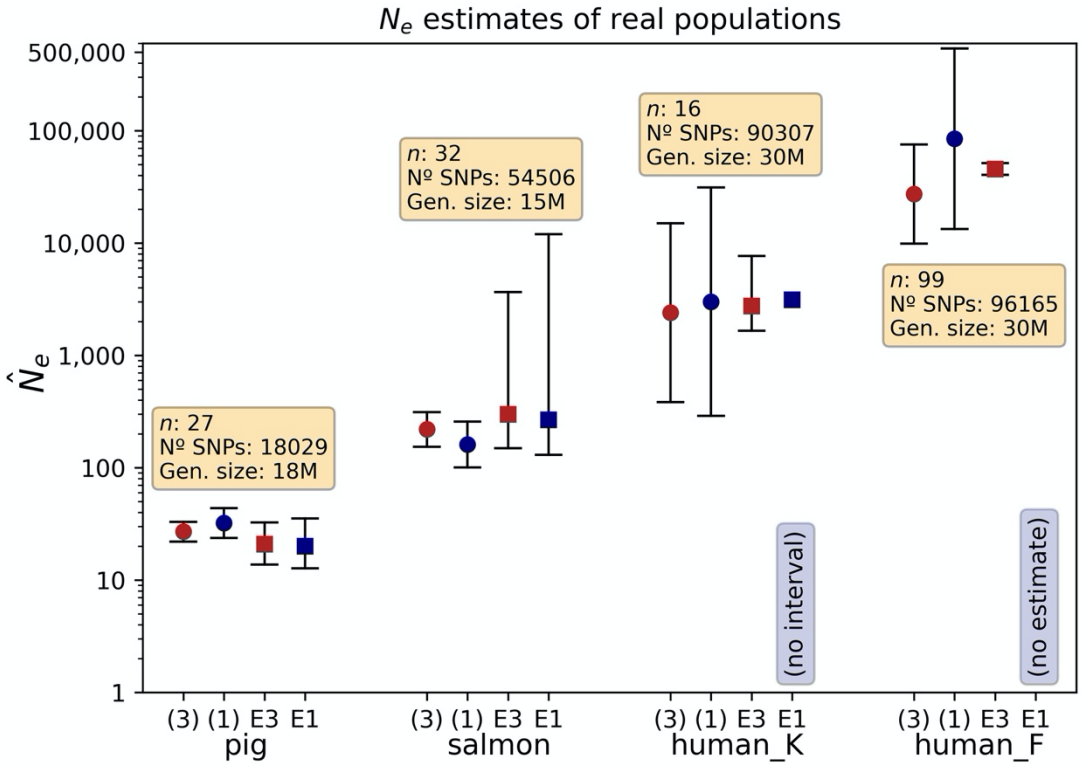


571

572 **Figure 4.** Distribution of estimates of contemporary N_e of simulated populations with constant sizes N
573 = 100, 1000 or 10,000 individuals, sample size n = 10 or 100 individuals and approximately 20,000 to
574 50,000 markers randomly distributed in genomes with 20 chromosomes of 1 Morgan length each.
575 Distributions of estimates made using *currentNe* are shown in salmon colour. Those obtained with the
576 method of Waples and Do (2008) for independent markers, assuming a minor allele frequency of
577 0.05, and further corrected for tight linkage with the corrections given by Waples et al. (2016)
578 accounting for the number of chromosomes, are overlapped in green colour. The estimations assumed
579 random mating for pig and salmon samples and monogamy for human samples. Boxes show the
580 percentage of estimates with negative N_e estimates, infinite or "not available" results. Simulations are
581 based on 300 to 600 replicates. The 95% confidence intervals are shown below the x axis by lines
582 centred on the true N_e value, which equals the real census size N of the population. Intervals for
583 *currentNe* estimates were calculated using the ANN (salmon coloured lines). The two intervals for
584 Waples et al. (2016) estimates (green lines) were calculated using the *NeEstimator.v2* software (Do et
585 al. 2014), using the parametric and resampling options only on five replicates per interval. Dotted
586 lines indicate that no intervals were generated by the method.

587

588 Figure 5.



589

590 **Figure 5**. N_e estimates and 95% confidence intervals resulting from the analysis of four populations
591 using the *currentNe* software: pig, salmon, Koryak (human_K) and Finnish (human_F). Red dots
592 correspond to estimates based on all possible SNP pairs using Eq. (3) and blue dots correspond to
593 estimates based only on unlinked SNP pairs using Eq. (1), i.e. between SNPs on different
594 chromosomes. This method calculates the confidence intervals using an ANN. Analogously, estimates
595 using the *NeEstimator*.v2 software (Do et al. 2014), assuming MAF = 0.05, random mating for pigs
596 and salmon, and monogamy for human samples, are represented by squares: red squares (E3) are
597 estimates using all locus pairs corrected for linkage with the equations of Waples et al. (2016)
598 accounting for number of chromosomes, and blue squares (E1) are estimates based on unlinked pairs.
599 For the pig and salmon estimates with *NeEstimator*.v2 (E1, E3), missing data was excluded from the
600 data files, as faulty estimates were obtained when they were included. The number of individuals in
601 the sample, the number of SNPs and the approximate autosomal genome size in Morgans are given in
602 the boxes. This method calculates confidence intervals using jackknife resampling. The labels "no
603 interval" and "no estimate" mean that the software has generated an infinite result.

APPENDIX

Estimation of the contemporary effective population size from SNP data while accounting for mating structure.

Enrique Santiago

1- Derivation of the equation of LD at equilibrium when each offspring is generated by a new random mating.

In order to facilitate the exposition, the derivation of the expected linkage disequilibrium (LD) in a randomly mating population given by Santiago et al. (2020) is repeated here:

The population consists of 2*N* haploid genomes randomly arranged in *N* diploid individuals. That is, each offspring is generated by a new random mating. Let *c* be the recombination rate between two polymorphic sites *X* and *Y*, *t* the current generation and D_t^2 the squared covariance of allele states between sites in haploid genomes at this generation (point 1 in Figure SF1). The expected value of the squared covariance in the next generation D_{t+1}^2 can be approximated in a two-step derivation: first the effect of recombination and then the sampling process.

The particular way in which the genomes are paired in parents at generation t determines the expectation of gametes (point 2 in figure SF1), since recombination is restricted to genomes within individuals. If *c*>0, new combinations of alleles are expected from recombination events within individuals and, consequently, the squared covariance in the infinite pool of gametes changes to $D_t'^2$ as shown in the figure. The expectation is given by the following equation, where x_i and y_i represent the allele values at sites *X* and *Y* respectively in genome *i* at generation *t*.

$$D_t'^2 = \left[\frac{x_1y_1\frac{(1-c)}{2} + x_2y_2\frac{(1-c)}{2} + x_1y_2\frac{c}{2} + x_2y_1\frac{c}{2} + \cdots + x_{2N-1}y_{2N-1}\frac{(1-c)}{2} + x_{2N}y_{2N}\frac{(1-c)}{2} + x_{2N-1}y_{2N}\frac{c}{2} + x_{2N}y_{2N-1}\frac{c}{2}}{N}\right]^2$$

$$= \left[\frac{(1-c)}{2}\frac{\sum_{i=1}^{2N} x_iy_i}{N} + \frac{c}{2}\frac{\sum_{i=1}^{2N} x_iy_j}{N}\right]^2 \quad \text{(S1)}$$

Here, x_i and y_i are deviations from the population mean values. Note that the coincidence of subscripts in both terms of a product of allele values (e.g. x_iy_i) refers to the original alleles in a given haploid genome of a diploid individual from generation *t* (i.e., alleles in a non-recombinant gamete of that individual). Otherwise (e.g. x_iy_j), the product refers to values of alleles in a recombinant gamete produced by that individual.

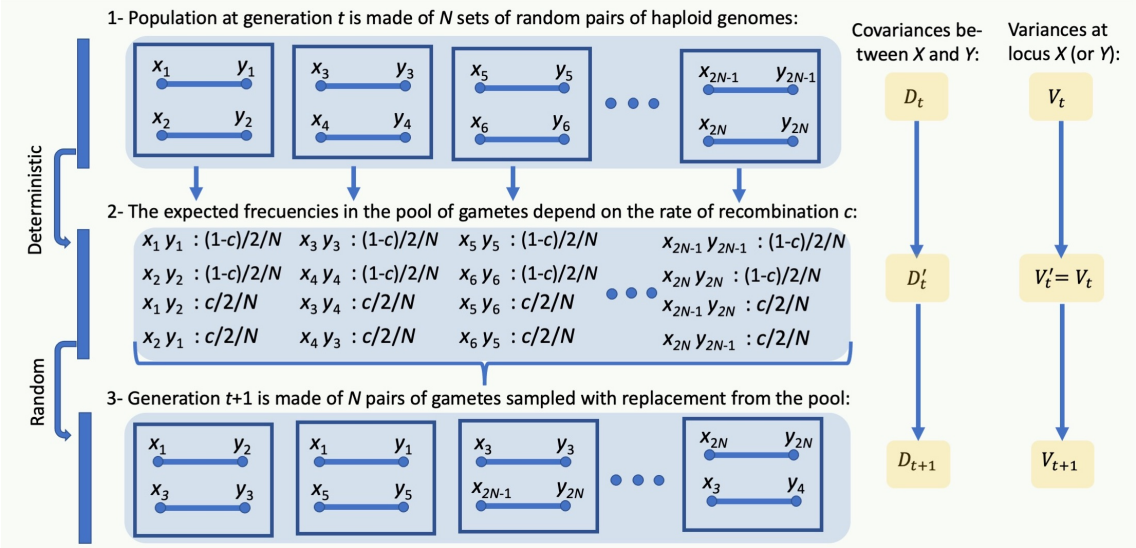


Figure SF1. Transition from one generation to the next one in a random mating population. The terms x_i and y_i refer to the allelic values at loci *X* and *Y* respectively in genome *i*.

Expanding the equation, we get:

$$D_t'^2 = (1-c)^2 \left[\frac{\sum_{i=1}^{2N} x_i y_i}{2N}\right]^2 + c^2 \left[\frac{\sum_{i=1}^{2N} x_i y_j}{2N}\right]^2 + 2c(1-c)\frac{\sum_{i=1}^{2N} x_i y_i}{2N}\frac{\sum_{l=1}^{2N} x_l y_k}{2N}$$

where the subscripts *k,l* correspond to alleles in a recombinant gamete of one individual and *i,j* correspond to alleles in a recombinant gamete of another individual, which may be the same.

The first term on the right-hand side is the remaining part of the original squared covariance D_t^2 after recombination. The third term is the sum of the cross-products of the allelic values of the gametes of different individuals, which has a marginal value under random mating. Then we get the simplification:

$$D_t'^2 \approx (1-c)^2 D_t^2 + c^2 \left[\frac{\sum_{i=1}^{2N} x_i y_j}{2N}\right]^2 \quad \text{(S2)}$$

The equation shows that $D_t'^2$ is the sum of two terms, the first one is due to contributions from parental gametes and the second one is due to recombinant gametes. The latter does not exist in haploid models with random mating. However, when genomes are paired within diploid individuals, i.e. when recombination is restricted to fixed pairs of genomes, it becomes significant. Its expansion is:

$$c^2 \left[\frac{\sum_{i=1}^{2N} x_i y_j}{2N}\right]^2 = c^2 \frac{\sum_{i=1}^{2N} x_i^2 y_j^2}{4N^2} + c^2 \frac{\sum_{i\neq j}^{2N} x_i y_j x_j y_i}{4N^2} + c^2 \frac{\sum_{i\neq j}^{2N} \sum_{i\neq j\neq k\neq l}^{2N} x_i y_j x_k y_l}{4N^2}$$

The expectation of $x_i^2 y_j^2$ in the first term on the right-hand side is equal to the product $V_x V_y$ of the variances at both locations, hereafter *W*, which is assumed to be constant throughout the process. The expectation of $x_i y_j x_j y_i$ in the second term is effectively D_t^2 (only a term due to autocorrelation makes a small difference). The summands in the

third term cancel each other under random mating and non-overlapping generations. Therefore, the equation for $D_t'^2$ reduces to:

$$D_t'^2 \approx (1-c)^2\, D_t^2 + c^2 \frac{W}{2N} + c^2 \frac{D_t^2}{2N}$$

D_t^2 is several orders of magnitude smaller than W at equilibrium and, consequently, the third term becomes irrelevant, especially when c is large. The equation reduces to:

$$D_t'^2 \approx (1-c)^2\, D_t^2 + c^2 \frac{W}{2N} \tag{S3}$$

The expected D_{t+1}^2 after random sampling of $2N$ gametes (point 3 in Figure SF1) is the sum of the remaining $D_t'^2$ (reduced by sampling) and the increase in covariance due to drift acting on standing variation (see Appendix in Santiago et al. 2020):

$$D_{t+1}^2 \approx D_t^{2\prime}\left(1 - \frac{2.2}{2N}\right) + W\frac{1}{2N}$$

This is a good approximation for systems where most of the LD at equilibrium has been generated in old generations, i.e. N is large and c is small. Otherwise the factor 2.2 increases slightly but becomes irrelevant because $W \gg D_t'^2$, especially when c is large.

Substituting $D_t'^2$ by its value given in Eq. (S3):

$$\begin{aligned} D_{t+1}^2 &= \left[(1-c)^2\, D_t^2 + c^2\frac{W}{2N}\right]\left(1 - \frac{2.2}{2N}\right) + W\frac{1}{2N} \\ &\approx (1-c)^2\, D_t^2\left(1 - \frac{2.2}{2N}\right) + W\frac{c^2}{2N} + W\frac{1}{2N} \end{aligned}$$

The next section shows that the contribution of mutations in one generation time to the standing LD is very small compared to the effects of drift and recombination. Ignoring for the moment the effect of mutations, D^2 at equilibrium is equal to the sum of the remaining D^2 after recombination and the increase in D^2 due to drift:

$$D^2 = (1-c)^2 D^2\left(1 - \frac{2.2}{2N}\right) + W\left(\frac{1}{2N} + \frac{c^2}{2N}\right) \tag{S4}$$

Dividing both sides of the equation by W and rearranging:

$$\delta_c^2 = \frac{1 + c^2}{2N(1-(1-c)^2) + 2.2(1-c)^2}$$

where δ^2 is a measure of population LD given by the ratio of expectations D^2/W.

2- The contribution of new mutations.

Consider a population at mutation drift equilibrium where *Y* is a monomorphic site in a very large genome, such that mutation events per site occur at a very low rate $\mu \ll 1/2N$. When a new mutation occurs at site *Y*, the genetic variance induced at that site and the expected squared covariance with other polymorphic site *X* are:

$$V_y = \frac{1}{2N} \cdot \frac{2N-1}{2N} \approx \frac{1}{2N}$$

$$E[D_{XY}^2] = E\left[p_x \left(\frac{1-p_x}{2N}\right)^2 + (1-p_x)\left(-\frac{p_x}{2N}\right)^2\right] = \frac{V_x}{4N^2} = \frac{W_{XY}}{2N}$$

where p_x is the frequency of the reference allele of site *X*, V_x is the genetic variance of site *X* and W_{XY} is the product of the genetic variances at both sites.

As site *Y* mutates with probability $2N\mu$ for the whole population, the expected increment of D^2 due to new mutations is:

$$E[\Delta D_{XY}^2] = 2N\mu \cdot 2 \cdot \frac{E[V_x]}{4N^2}$$

The factor "2" is included because the creation of a new polymorphic site in a genetic system with a number of previously polymorphic sites generates new covariances at twice that number. The increase in D^2 is very small compared to the product of the genetic variances *W* of pairs of sites at equilibrium:

$$E[W] = E\left[{V_x}^2\right] = 2N\mu \cdot E[V_x]$$

Therefore,

$$\frac{E[\Delta D_{XY}^2]}{E[W]} = \frac{1}{2N^2}$$

Now, including the increment of the squared covariance due to mutation in Eq. (S4) at equilibrium:

$$D^2 = (1-c)^2 D^2 \left(1 - \frac{2.2}{2N}\right) + W\frac{1+c^2}{2N} + E[\Delta D_{XY}^2]$$

That is, D^2 at equilibrium is equal to the sum of the fraction of D^2 remaining after recombination and the increases in D^2 due to drift and mutation. Divide both sides by *W* and rearrange:

$$\delta_c^2 = \frac{1 + c^2 + 1/N}{2N(1-(1-c)^2) + 2.2(1-c)^2}$$

Therefore, the small factor $1/N$ in the numerator represents the contribution of new mutations to LD. This contribution will be ignored in the following.

3- The number of full siblings that a random individual has.

With random mating and random contribution of parents, that is when each offspring is generated by a new random mating, each individual is expected to have $4/N$ full siblings among the set of N offspring, assuming an equal number of sexes (demonstration not shown). This number is too small to have a significant effect on the above predictions. In contrast, with lifetime monogamy, an individual sampled from a monogamous family is expected to have two full siblings, assuming constant population size and a random distribution of family size.

Let m=1 be the proportion of monogamous pairings in a population with a random distribution of family sizes, i.e., each monogamous family is expected to contribute with two offspring to the next generation. If the population size is not too small, the family size follows a Poisson distribution and the frequency of families with z offspring is:

$$\frac{e^{-2}2^z}{z!}$$

And the probability that a random individual comes from a family with z full siblings is

$$P(z) = \frac{e^{-2}2^z}{z!} \cdot \frac{z}{2}$$

where the factor "2" in the denominator is the average family size. Therefore, the expected size (number of siblings) of the family of a random individual is

$$\frac{\sum_{z=0}^{\infty} P(z) \cdot z}{\sum_{Z}^{\infty} P(z)} = 3$$

In summary, the number of full siblings of a random individual is 2 when m=1. Therefore, for any given proportion m of monogamous matings in a population, the expected number of full siblings that a random individual will have is $k = 2m$.

An alternative demonstration for monogamy can be made in a more direct way. Given a population of $N_e/2$ full-sib families, if we choose a single reference offspring to be the first one for the next generation, each of the subsequent N_e-1 offspring will be a full sibling of the reference offspring with probability $2/N_e$; hence the expected number of full siblings that the reference offspring (or any other individual) has in the whole population is the product of the population size (N_e-1≈ N_e) and the probability $2/N_e$, i.e. $k = (N_e\text{-}1)\cdot 2/N_e \approx 2$. Note that this derivation assumes that each of the subsequent N_e-1 individual samples is independent of the others.

Applying the same logic to polygyny in a population with N_m males and N_f females, the probability of an offspring being a full sibling of the reference offspring is $1/N_f$, regardless of the sex of the offspring, because there are N_f full-sib families. For populations with different numbers of males and females (Wright 1933),

$$N_e = \frac{4N_mN_f}{N_m + N_f}$$

This means that for the variance of gene frequencies and for the inbreeding, the population behaves as if, in each generation, N_e monoecious individuals were randomly selected from families to be the parents of the next generation. This does not change the probability of two individuals coming from the same family ($1/N_f$). Consequently in an idealized population of size N_e, which is the size referred to in Eq. (1), k is the product of the size N_e and the probability:

$$k = N_e \cdot \frac{1}{N_f} = \frac{4N_m N_f}{N_m + N_f} \cdot \frac{1}{N_f} = \frac{4N_m}{N_m + N_f}$$

Following a similar but more heuristic argument, in multiparous species with several (but not too few) offspring per litter and L litters per female sired by different males, the expected number of full siblings that a random individual has among the N reproducers has is $k = 2/L$. The expected value of k increases when a single male sires more than one litter from the same female. Let L_i be the expected number of litters sired by a given male such that:

$\sum_{i=1}^{s} L_i = L$, where s is the number of sires per female.

This sum represents the average number of litters per male. Consequently the probability of sampling with replacement two offspring (one after the other) of the same sire among all the offspring of a female is:

$$\sum_{i=1}^{s} \left(\frac{L_i}{L}\right)^2 = \frac{1}{L_e}$$

If we call the inverse of this probability as the "effective number of litters" L_e, then the expected number of full siblings of a random individual (among reproducers) is $k = \frac{2}{L_e}$.

Selection and other factors also affect the variance of the full-sib family size and hence the k-value. Let $f(x)$ be the frequency of full-sib families of size x. The mean M and the variance V of this distribution are

$$M = \sum_{x=0}^{\infty}[f(x) \cdot x] \,, \quad V = \sum_{x=0}^{\infty}[f(x) \cdot x^2] - M^2$$

The probability that a random individual comes from a family of size x full-sibs is

$$\frac{f(x) \cdot x}{M}$$

And the average number of full siblings that a random individual has is

$$k = \frac{\sum_0^{\infty}[f(x) \cdot x \cdot (x-1)]}{\sum_0^{\infty}[f(x) \cdot x]} = \frac{\sum_0^{\infty}[f(x) \cdot x^2]}{\sum_0^{\infty}[f(x) \cdot x]} - 1 = \frac{V + M^2}{M} - 1 = \frac{V}{M} + M - 1$$

With random contribution of families to the next generation (i.e. Poisson distribution of full-sib family size), V is equal to M and the equation reduces to $k=M$. Under selection, however, V rises above M and the value of k could even be greater than 2, which is the expected value under full lifetime monogamy.

4- Considering the effect of full siblings on LD.

The key difference between a lifetime mating model and a random mating model (each offspring from a new random mating) is that in the former some individuals (represented in point 1 of figure SF2) can be full siblings and the expectations of some sums of products, which are neglected in random mating, become significant.

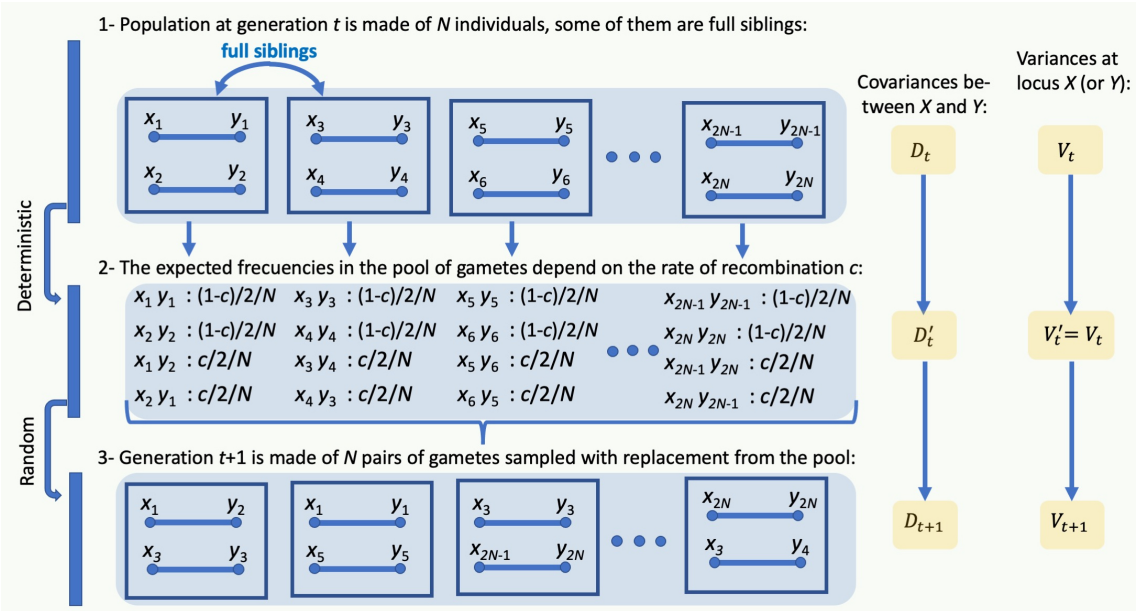


Figure SF2. Transition from one generation to the next one in a population with full siblings. The terms x_i and y_i refer to the allelic values at loci X and Y respectively in genome i.

The expansion of Eq. (S2) is:

$$D_t'^2 = (1-c)^2\, D_t^2 + c^2 \frac{\sum_{i=1}^{2N} x_i^2 y_j^2}{4N^2} + c^2 \frac{\sum_{i\neq j}^{2N} x_i y_i x_j y_j}{4N^2} + c^2 \frac{\sum_{i\neq j}^{2N} \sum_{i\neq j\neq k\neq l}^{2N} x_i y_j x_k y_l}{4N^2}$$

The second term is $c^2 \frac{W}{2N}$ and the third term is the irrelevant contribution $c^2 \frac{D_t^2}{2N}$ as explained above, therefore:

$$D_t'^2 = (1-c)^2\, D_t^2 + c^2 \frac{W}{2N} + c^2 \frac{\sum_{i\neq j}^{2N} \sum_{k\neq l}^{2N} x_i y_j x_k y_l}{4N^2}$$

The last term was assumed to be 0 for random mating because alleles from different individuals are uncorrelated (the subscripts i,j vs. k,l, refer to alleles in recombinant gametes from different individuals). However, this double sum has a small but significant contribution in lifetime pairings because some pairs of individuals are full sibs and could produce recombinant gametes with the same allele combinations in the two sites X and Y. The expectation of $x_i y_j x_k y_l$ is equal to $x_i^2 y_j^2$, that is W, when i and k in site X, and j and l in site Y come from the same allele copies in the parents of the siblings (Figure SF3). However, there is no possibility of coincidence of allele copies for recombinant gametes coming from half siblings because these share only one parent, the other two parents being unrelated.

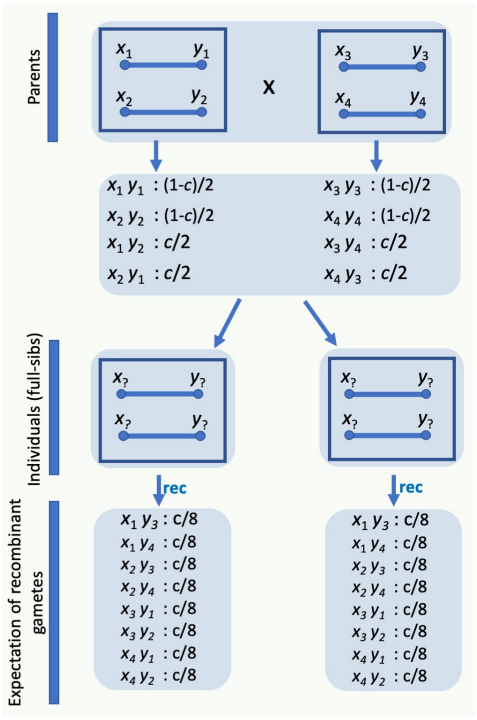


Figure SF3. With reference to the identities in parents, eight different types of recombinant gametes are expected from individuals. The expectation of each type is $c/8$.

Since there are eight different types of recombinant gametes, the total probability that any two recombinant gametes from two full siblings will be of the same type is 1/8. In other words, two recombinant gametes from two full siblings have 1/8 probability of matching each other in allele copies. This double sum is the sum of all the non-diagonal elements of a square matrix of order $2N$. Each of the N individuals is represented twice in each of the two marginals (row and column) of the matrix, once for each of its two possible gametes, and also has k full siblings, who are also represented twice, once for each of their two gametes. Therefore, the double sum of the recombinants $\sum_{i\neq j}^{2N}\sum_{k\neq l}^{2N} x_i y_j x_k y_l$ has $(2k \cdot 2N \,/\, 8)$ summands with the same type (i.e., $\mathrm{E}[x_i y_j x_k y_l] = \mathrm{E}[x_i y_j x_i y_j] = E[x_i^2 y_j^2] = W$ for each of these summands). Therefore, the expectation of the double sum is $(WkN \,/\, 2)$ and the equation reduces to

$$D_t'^2 = (1-c)^2\, D_t^2 + c^2 \frac{W}{2N} + c^2 \frac{W}{2N}\frac{k}{4} \tag{S5}$$

Now, similar to the derivation of Eq. (S4), $2N$ gametes are sampled to produce the next generation:

$$D_{t+1}^2 = D_t'^2\left(1-\frac{2.2}{2N}\right) + W\frac{1}{2N} \approx (1-c)^2\, D_t^2\left(1-\frac{2.2}{2N}\right) + c^2\frac{W}{2N} + c^2\frac{W}{2N}\frac{k}{4} + W\frac{1}{2N}$$

Dividing both sides of the equation by W and rearranging:

$$\delta_c^2 = \frac{1 + c^2 + c^2\frac{k}{4}}{2N(1-(1-c)^2) + 2.2(1-c)^2}$$

5- Correction for sampling of diploids when the phase is unknown.

If the phase is unknown, we use the covariance of the bivariate distribution of the mean of the two values at each locus *X* and *Y* in each of the *n* diploids of the sample. That is, for each individual:

$$\chi_{i,j} = \frac{x_i + x_j}{2} \quad \text{and} \quad \psi_{i,j} = \frac{y_i + y_j}{2} \tag{S6}$$

where the subscripts *i* and *j* represents the two homologous chromosomes (see figure SF4). In the case of random union of gametes, the covariance between χ and ψ can be expressed as:

$$cov_{\chi,\omega} = \frac{1}{2}D$$

where D is the covariance between the values *x* and *y* in haploid genomes. Burrows' Δ composite measure of LD for pairs of diallelic loci in a sample (Cockerham and Weir, 1977) is given by:

$$\Delta = 2\,P_{AABB} + P_{AABb} + P_{AaBB} + \frac{P_{AaBb}}{2} - 2\,p_A\,p_B$$

where capital *P* refers to the frequencies of the genotypes in the sample and small *p* refers to the allele frequencies at two loci with alleles A/a and B/b respectively. It can be shown that Δ is twice $cov_{\chi,\omega}$ (with values 1 vs. 0 for presence vs. absence of the reference allele), that is twice the component of covariance between individuals in the sample ($D_{s_{betw}}$). In this case, Eq. (A10) of the Appendix in Santiago et al. (2020) for sampling of diploids results:

$$E[\Delta^2] = 4 \cdot E\left[D^2_{s_{betw}}\right] \approx 4 \cdot D'^2_{betw} \cdot \frac{(n-1)^3 + \frac{4}{5}(n-1)^2 + (n-1)}{n^3} + 4 \cdot W'_{betw}\frac{n-1}{n^2}$$

where D'_{betw} is the covariance component between zygotes in the group from which the sample is taken and $W'_{betw} = W/4$. Looking in detail at the offspring of the first pair of parents in point 3 of figure SF4,

$$\begin{aligned} D'^2_{betw} = \Big[& \left(\chi_{1,3}\psi_{1,3} + \chi_{1,4}\psi_{1,4} + \chi_{2,3}\psi_{2,3} + \chi_{2,4}\psi_{2,4}\right)\frac{(1-c)^2/4}{N/2} \\ & + \left(\chi_{1,3}\psi_{1,4} + \chi_{1,4}\psi_{1,3} + \chi_{2,3}\psi_{1,4} + \chi_{2,4}\psi_{1,3} + \chi_{1,3}\psi_{2,3} + \chi_{2,3}\psi_{1,3} + \chi_{1,4}\psi_{2,4} + \chi_{2,4}\psi_{1,4}\right)\frac{c\,(1-c)/4}{N/2} \\ & + \left(\chi_{1,3}\psi_{2,4} + \chi_{1,4}\psi_{2,3} + \chi_{2,3}\psi_{1,4} + \chi_{2,4}\psi_{1,3}\right)\frac{c^2/4}{N/2} + \cdots (terms\ of\ the\ other\ pairs\ of\ parents) \Big]^2 \end{aligned}$$

After replacing of $\chi_{i,j}$ and $\psi_{i,j}$ by their values in Eq. (S6), expanding the equation and rearranging the terms into two classes, those that are affected by the recombination rate *c* and those that are independent of *c*, the equation simplifies to:

$$\begin{aligned} & D'^2_{betw} \\ & = \Big[\frac{x_1y_1(1-c)/2 + x_2y_2(1-c)/2 + x_1y_2c/2 + x_2y_1c/2 + \cdots (terms\ of\ the\ other\ pairs\ of\ parents)}{2N} \\ & + \frac{x_1y_3 + x_1y_4 + x_2y_3 + x_2y_4 + x_3y_1 + x_4y_1 + x_3y_2 + x_4y_2 + \cdots (terms\ of\ the\ other\ pairs\ of\ parents)}{8N}\Big]^2 \end{aligned}$$

The expectation of the product of the two terms in brackets is zero, as the factors are uncorrelated. So the equation can be simplified to a sum of two squared terms:

$$D'^2_{betw} = \left[\frac{x_1y_1(1-c)/2 + x_2y_2(1-c)/2 + x_1y_2c/2 + x_2y_1c/2 + \cdots (terms\ of\ the\ other\ pairs\ of\ parents)}{2N}\right]^2 + \left[\frac{x_1y_3 + x_1y_4 + x_2y_3 + x_2y_4 + x_3y_1 + x_4y_1 + x_3y_2 + x_4y_2 + \cdots (terms\ of\ the\ other\ pairs\ of\ parents)}{8N}\right]^2$$

Where the first term is equal to D'^2_t given in Eq. (S1). The second term is the sum of the products of the values of the alleles at different loci and at different parents of the same couple and, since the cross products cancel each other, is simplified as:

$$\frac{(x_1y_3)^2 + (x_1y_4)^2 + (x_2y_3)^2 + (x_2y_4)^2 + (x_3y_1)^2 + (x_4y_1)^2 + (x_3y_2)^2 + (x_4y_2)^2 + \cdots (terms\ of\ the\ other\ pairs\ of\ parents)}{(8N)^2}$$

$$= \frac{8N/2 \cdot W}{(8N)^2} = \frac{W}{16N}$$

Note that this term only exists if two conditions are met: there are lifetime pairings and the genotypes are unphased. It is independent of the recombination rate. As the derivation assumed full lifetime monogamy (i.e., monogamy rate $m = k/2 = 1$), the term must be multiplied by the appropriate $k/2$ value in other circumstances. Finally,

$$D'^2_{betw} = \frac{D'^2}{4} + \frac{W}{16N}\frac{k}{2}$$

Replacing the first term by its value (S5) gives the general equation:

$$D'^2_{betw} = \frac{(1-c)^2\, D_t^2 + c^2\frac{W}{2N} + c^2\frac{W}{2N}\frac{k}{4}}{4} + \frac{W}{16N}\frac{k}{2} = (1-c)^2\frac{D_t^2}{4} + \frac{W}{16N}\left[2c^2 + 2c^2\frac{k}{4} + \frac{k}{2}\right]$$

The expectation of Δ^2 after sampling is:

$$E[\Delta^2] = 4 \cdot D'^2_{betw} \cdot \frac{(n-1)^3 + \frac{4}{5}(n-1)^2 + (n-1)}{n^3} + W'\frac{n-1}{n^2}$$

Eq. (A8) in the Appendix of Santiago et al (2020) allows the prediction of the expectation $E[W_s]$ in the sample:

$$E[W_s] \approx W'\frac{(2n-1)^3 + (2n-1)^2}{8n^3}$$

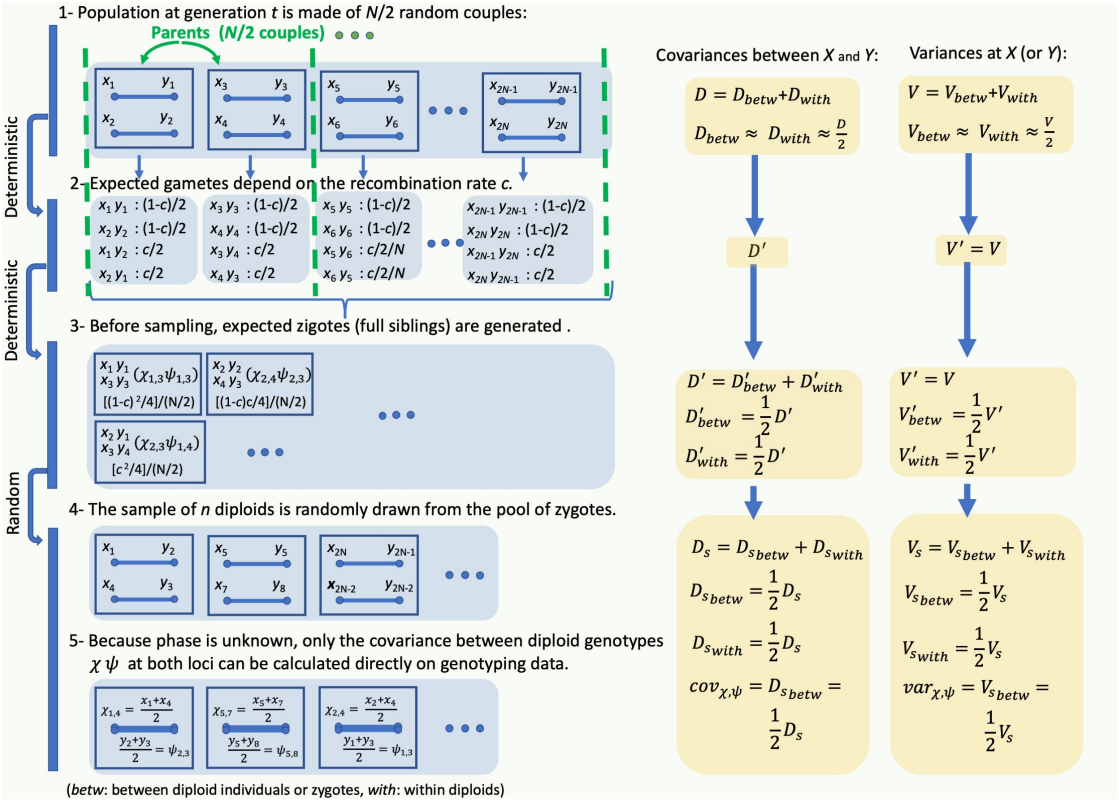


Figure SF4. Diagram of the process of sampling unphased genotypes with monogamy.

The expectation of $d^2_{s(c)}$ in the sample for a particular recombination rate c is:

$$d^2_{s(c)} = \frac{E[\Delta^2]}{E[W_s]} \approx \left[\delta_c^2(1-c)^2 Z + \frac{\frac{k}{2} + 2c^2 + 2c^2\frac{k}{4}}{4N} Z + \frac{4n-4}{(2n-1)^2}\right] \tag{S7}$$

where

$$Z = \frac{(2n-2)^3 + \frac{8}{5}(2n-2)^2 + 4(2n-2)}{(2n-1)^3 + (2n-1)^2}$$

Therefore, the estimation for δ_c^2 in the population is:

$$\widehat{\delta_c^2} \approx \frac{d^2 - \frac{4n-4}{(2n-1)^2} - \frac{\frac{k}{2} + 2c^2 + 2c^2\frac{k}{4}}{4N} Z}{Z(1-c)^2}$$

6- The integral for LD over the genome including sampling.

The prediction of the genome-wide average LD can be obtained by integrating all possible site pairs. If the species has v chromosomes, each of length l Morgans, then two random sites in this genome are in different chromosomes with probability $(v-1)/v$ and for them c is 0.5. If two random sites are in the same chromosome

(probability = $1/v$), the density of their genetic distance follows a triangular distribution with the highest frequency corresponding to full linkage and the lowest frequency corresponding to the maximum distance l in the chromosome. Assuming a random distribution of recombination events (i.e., Haldane's map function $c = (1 - e^{-2x})/2$, where x is the genetic distance in Morgans), the expected LD in a finite sample d_s^2 can be calculated as the average of the δ_c^2 values in the population over all site pairs with Eq. (S7), summing the contributions of pairs in different chromosomes (first term on the right) and pairs in the same chromosome (second term):

$$d_s^2 = \frac{v-1}{v}\left[\frac{\delta_{0.5}^2 Z}{4} + \frac{k+1+\frac{k}{4}}{8N}Z + \frac{4n-4}{(2n-1)^2}\right] + \frac{2}{vl}\int_0^l \frac{l-x}{l}\left[\delta_c^2(1-c)^2 Z + \frac{\frac{k}{2}+2c^2+2c^2\frac{k}{4}}{4N}Z + \frac{4n-4}{(2n-1)^2}\right]dx \quad \text{(S8)}$$

where,

$$c = (1 - e^{-2x})/2, \quad Z = \frac{(2n-2)^3+\frac{8}{5}(2n-2)^2+4(2n-2)}{(2n-1)^3+(2n-1)^2}, \quad \delta_c^2 = \frac{1+c^2+c^2\frac{k}{4}}{2N_e(1-(1-c)^2)+2.2(1-c)^2}$$

Eq. (S8) can be solved numerically for N_e, given the observed d_s^2 in the sample.

7- The relation to the equations of Weir and Hill (1980).

Numerical predictions of LD using Weir and Hill's (1980) equations under random pairing (each progeny from a new random pairing) and lifetime mating are compared with predictions using our equations in Table S1 in Santiago et al. (2020) and in Supplementary Tables S1 and S2 of this paper. Here, an algebraic comparison is performed to show the differences between the two theories.

After rearranging Eq. (3) for random pairing in Weir and Hill (1980), we get:

$$\frac{E[D^2]}{E[W]} = \frac{c^2+(1-c)^2}{2N_e c(2-c)} = \frac{1+c^2}{2N_e(1-(1-c)^2)} - \frac{1}{2N_e}$$

The term $-1/2N_e$ is a consequence of the particular derivation method used to obtain "the variance of disequilibria for the infinite progeny array from a set of parents" (Weir and Hill, 1980). To obtain the prediction for a finite diploid population of size N_e, the sampling term $1/2N_e$ must be added to the equation:

$$\delta_{WH}^2 = \frac{1+c^2}{2N_e(1-(1-c)^2)}$$

The corresponding equation for random pairing in Santiago et al. (2020) is:

$$\delta^2 = \frac{1 + c^2}{2N_e(1 - (1 - c)^2) + 2.2(1 - c)^2}$$

with the additional term, $2.2(1 - c)^2$ in the denominator, representing the reduction by sampling of the original squared covariance from the previous generation (see derivation of Eq. S4). Covariances, like variances, are reduced by sampling in a fraction proportional to the inverse of the sample size (here, $1/2N_e$ per generation), in addition to the reduction of the covariance by recombination by a fraction *c*. This reduction by sampling is ignored in the Weir and Hill's equation, probably because their original derivation is for an infinite progeny. This leads to deviations in the predictions of δ^2, especially when the product $N_e c$ is small.

Eq. (5) in Weir and Hill (1980) for lifetime pairing can be rearranged in a similar way,

$$\frac{E[D^2]}{E[W]} = \frac{(1 - c)^2 + 2c^2 + f[(1 - c)^2 + c^2]}{2N_e c(2 - c)(f + 1)} = \frac{1 + c^2 + \frac{c^2}{1 + f}}{2N_e(1 - (1 - c)^2)} - \frac{1}{2N_e}$$

Here, *f* is the number of females per male under lifetime polygyny (with lifetime monogamy, *f* = 1). As before, the prediction for a finite population with N_e diploid individuals is,

$$\delta^2_{WH} = \frac{1 + c^2 + \frac{c^2}{1 + f}}{2N_e(1 - (1 - c)^2)}$$

while our equation derived in Section 4 is:

$$\delta^2 = \frac{1 + c^2 + c^2\frac{k}{4}}{2N_e(1 - (1 - c)^2) + 2.2(1 - c)^2}$$

For lifetime pairing, $k = \frac{4N_m}{N_m + N_f}$, where N_m is the number of males and N_f is the number of females (Section 3 of the Appendix), assuming $N_f \geq N_m$. Expressed in terms of the number *f* of females per male, *k* reduces to $k = \frac{4}{1+f}$, and both equations show to be identical for lifetime pairing except for the term $2.2(1 - c)^2$, which represents the reduction by sampling of the original covariance coming from the previous generation, as indicated above.

SUPPLEMENTARY TABLES AND FIGURES

Estimation of the contemporary effective population size from SNP data while accounting for mating structure

Enrique Santiago, Armando Caballero, Carlos Köpke and Irene Novo

Table S1. Simulated (observed) and predicted LD values $\left(\frac{E[D^2]}{E[W]}\right)$ of phased genotypes in fully monogamous populations (i.e. $k = 2m$, where $m = 1$ is the proportion of monogamous mating) of constant size N individuals. Under the conditions of these simulations, the census size N and the expected effective size N_e are equal. Observed values were calculated by averaging the true values in the population (i.e. those for the whole population with phased genotypes) for 10^8 consecutive generations in a two-locus system for each combination of N and recombination rate c, with reintroduction of mutations after fixation or loss of alleles at either locus.

Predictions were obtained by:

- eq.1: $\delta_c^2 = \frac{E[D^2]}{E[W]} = \frac{1+c^2+c^2\frac{k}{4}}{2N_e(1-(1-c)^2)+2.2(1-c)^2}$, Eq. (1) in the main text with full lifetime monogamy (i.e., $k = 2$).
- W&H: $\frac{E[D^2]}{E[W]} = \frac{(1-c)^2+2c^2+f[(1-c)^2+c^2]}{2N_e c(2-c)(f+1)} + \frac{1}{2N_e}$, Eq. (5) for phased genotypes in Weir & Hill (1980) with a ratio females/males $f = 1$ and substituting sample size by N_e.

			Predicted $\frac{E[D^2]}{E[W]}$	
Rec. rate	*N* diploids	observed $\frac{E[D^2]}{E[W]}$	eq.1	W&H
$c = 0.5$	10	**0.08488**	0.08842	0.09167
	100	**0.00909**	0.00913	0.00917
	1000	**0.00092**	0.00092	0.00092
	10000	**0.00009**	0.00009	0.00009
$c = 0.1$	10	**0.18814**	0.18183	0.26710
	100	**0.02527**	0.02551	0.02671
	1000	**0.00265**	0.00266	0.00267
	10000	**0.00027**	0.00027	0.00027
$c = 0.001$	10	**0.48850**	0.44731	25.01254
	100	**0.38127**	0.38530	2.50125
	1000	**0.15504**	0.16146	0.25012
	10000	**0.02353**	0.02371	0.02501
$c = 0$	10	**0.49712**	0.45454	Indeterminable
	100	**0.45570**	0.45454	
	1000	**0.45410**	0.45454	
	10000	**0.45408**	0.45454	

Table S2. Simulated (observed) and predicted LD values $\left(\frac{E[\Delta^2]}{E[W]}\right)$ of unphased genotypes in fully monogamous populations (i.e. $k = 2m$ where $m = 1$ is the proportion of monogamous mating) of constant size N individuals. Under the conditions of these simulations, the census size N and the expected effective size N_e are equal. Observed values were calculated by averaging the true values (i.e. those for the whole population with phased genotypes) for 10^8 consecutive generations in a two-locus system for each combination of N and recombination rate c, with reintroduction of mutations after fixation or loss of alleles at either locus.

Predictions were obtained by:

- eq.1: $\delta_c^2 = \frac{E[D^2]}{E[W]} = \frac{1+c^2+c^2\frac{k}{4}}{2N_e(1-(1-c)^2)+2.2(1-c)^2}$, Eq. (1) in the main text with full monogamy (i.e., $k = 2$) and, then, substituting the result into the Eq. (S7) for sampling (Appendix) to obtain the prediction of $\frac{E[\Delta^2]}{E[W]}$.
- W&H: $\frac{E[\Delta^2]}{E[W]} = \frac{1+2c^2+f[(1-c)^2+c^2]}{2N_ec(2-c)(f+1)} + \frac{1}{N_e}$, Eq. (5) in Weir & Hill (1980) for unphased genotypes with a ratio females/males $f = 1$ and substituting sample size by N_e.

			Predicted $\frac{E[\Delta^2]}{E[W]}$	
Rec. rate	*N* diploids	observed $\frac{E[\Delta^2]}{E[W]}$	eq.1	W&H
c = 0.5	10	**0.13182**	0.15830	0.16667
	100	**0.01629**	0.01658	0.01667
	1000	**0.00166**	0.00167	0.00167
	10000	**0.00017**	0.00017	0.00017
c = 0.1	10	**0.22326**	0.25364	0.34211
	100	**0.03228**	0.03296	0.03421
	1000	**0.00340**	0.00341	0.00342
	10000	**0.00034**	0.00034	0.00034
c = 0.001	10	**0.48720**	0.51906	25.08754
	100	**0.38384**	0.39242	2.50875
	1000	**0.15560**	0.16219	0.25087
	10000	**0.02361**	0.02379	0.02509
c = 0	10	**0.49479**	0.52629	Indeterminable
	100	**0.45736**	0.46160	
	1000	**0.45429**	0.45525	
	10000	**0.44147**	0.45462	

Figure S1. Effects of varying the recombination rate c with full monogamy **(A)** and varying the monogamy rate with $c = 0.5$ **(B)** on N_e estimates. Two-locus systems with 1000 diploid individuals were simulated for each c and k value during 10^8 consecutive generations with reintroduction of mutations when an allele was fixed or lost at either locus. Genotypes were unphased. Estimates were made solving Eq. (S7) (Appendix) for δ_c^2 and substituting this value in Eq. (1) in the main text to estimate N_e. Red circles are estimates of N_e using the true value of k. Blue triangles are estimates N_e ignoring monogamy ($k = 0$), i.e. with the incorrect assumption that each offspring was generated with a new random mating.

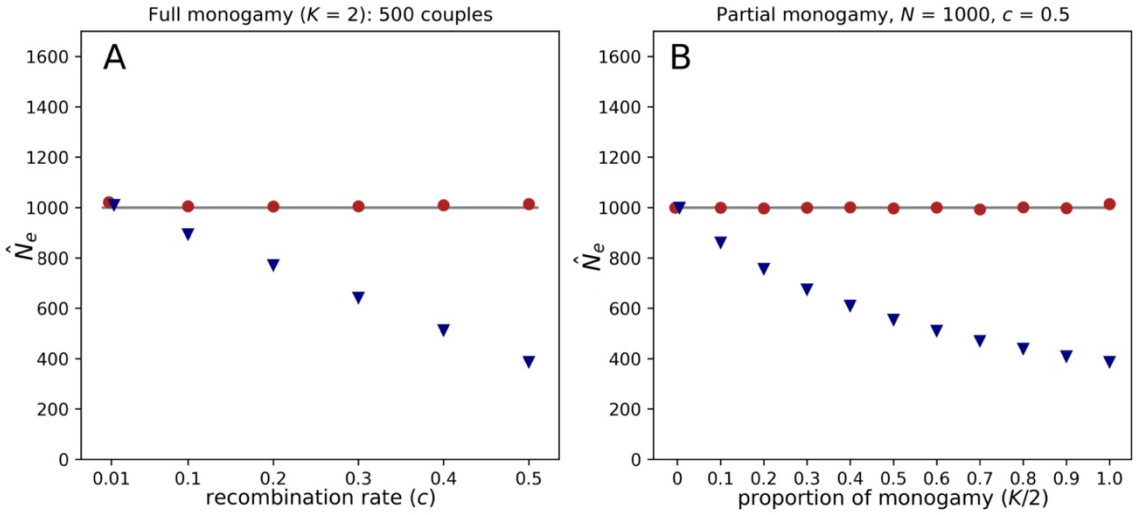


Figure S2. Effects of different map alterations on the contemporary N_e estimates (red dots with 95% confidence intervals in blue). Forty simulations were performed for each of the genetic maps: ordering errors within chromosomes (each chromosome map was split into four segments that were randomly rearranged within chromosomes to reconstruct a false map), ordering errors between chromosomes (the four segments were randomly changed between chromosomes), splitting the chromosome maps into scaffolds (four scaffolds per chromosome) and uneven distribution of markers on the true genetic map. For the latter, 90% of the markers were evenly distributed in the first half of each chromosome and the remaining 10% in the other half. In all simulations, the population size was kept constant with 1000 individuals per generation and the analyses were performed on samples with n = 100 individuals and 10,000 markers randomly distributed across 20 chromosomes of one Morgan length. Two different estimates of N_e were made for each sample: the estimate for the most recent generation using the GONE software (G), which requires a genetic map, and the estimate using the *currentNe* software based on Eq. (3) in the main text (cNe).

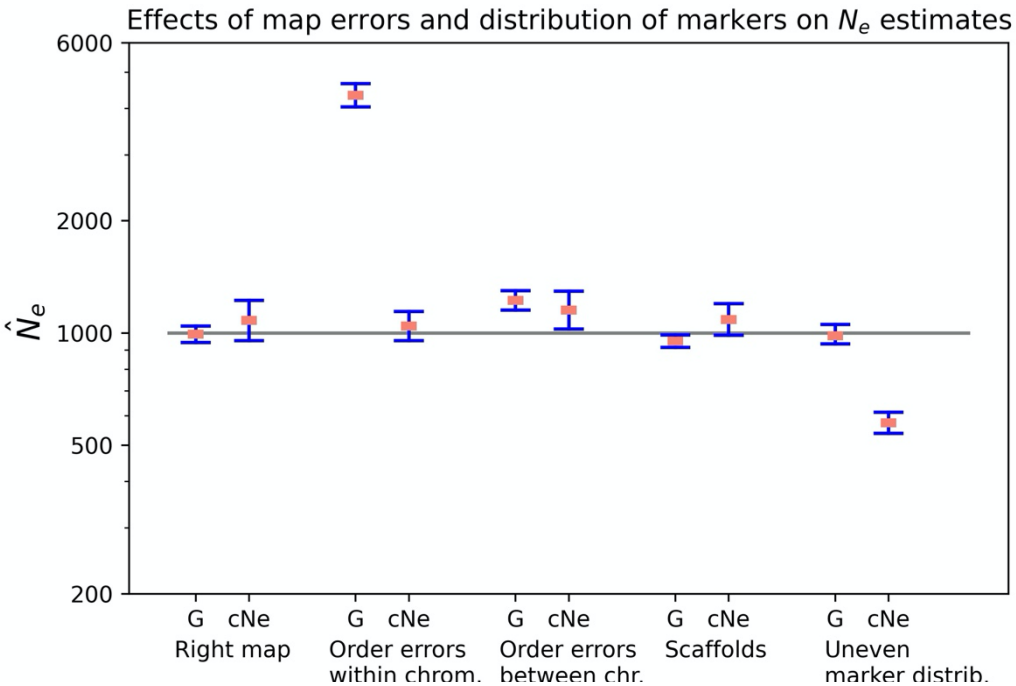


Figure S3. Effects on contemporary N_e estimates (red dots with 95% confidence intervals in blue) of deviations from assumptions of the model about sampling and population structure. From left to right in the abscissa: the analysis of a panmictic population of 1000 individuals using DNA arrays excluding SNPs with MAF below 0.2; the analysis of metapopulations composed of two subpopulations of 1000 individuals each (sampling only one subpopulation or both subpopulations and considering two migration rates: 0.02 and 0.002); the analysis of populations with overlapping generations (three cohorts of 888, 222, and 222 individuals, resulting in N_e=1000), with sampling of either one cohort or all cohorts; finally, sampling limited to the offspring of a random subset of families (50% and 20%) in populations of 1000 individuals. Forty replicates were simulated in each scenario, and each analysis was performed on a sample of 100 individuals and approximately 10,000 SNPs. The genome consisted of 20 chromosomes of one Morgan length each. Two different estimates of N_e were made for each sample: the estimate for the most recent generation using the GONE software (G), which requires a genetic map, and the estimate using the currentNe software based on Eq. (3) in the main text (cNe).