Recent demographic history inferred by high-resolution analysis of linkage disequilibrium

Enrique Santiago[1], Irene Novo[2], Antonio F. Pardiñas[3], María Saura[4], Jinliang Wang[5] and Armando Caballero[2]

[1]Departamento de Biología Funcional, Facultad de Biología, Universidad de Oviedo, Oviedo, Spain
[2]Departamento de Bioquímica, Genética e Inmunología (Facultade de Bioloxía) y Centro de Investigación Mariña (CIM-UVIGO), Universidade de Vigo, Vigo, Spain
[3]MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK
[4]Departamento de Mejora Genética Animal, INIA, Madrid, Spain
[5]Institute of Zoology, Zoological Society of London, London, UK

Corresponding author: Enrique Santiago. Departamento de Biología Funcional, Facultad de Biología, Universidad de Oviedo, Oviedo, Spain. e-mail: esr@uniovi.es

Keywords: Effective population size; genetic drift; demography; SNP; ancient DNA

Abstract

Inferring changes in effective population size (N_e) in the recent past is of special interest for conservation of endangered species and for human historiography. Current methods for estimating the very recent historical N_e are unable to detect complex demographic trajectories involving multiple episodes of bottlenecks, drops and expansions. Here we develop a theoretical and computational framework to infer with high resolution the demographic history of a population within the past 100 generations from the observed spectrum of linkage disequilibrium (LD) of pairs of loci over a wide range of recombination rates in a sample of contemporary individuals. The contributions of all of the previous generations to the observed LD are individually included in our model, and a genetic algorithm is used to search for the sequence of historical N_e values that best explains the observed LD. The method can be applied to samples of fewer than 10 individuals using various types of genotyping and DNA sequencing data: haploid, diploid with phased or unphased genotypes and pseudo-haploid data from low-coverage sequencing. The method was tested by computer simulation for sensitivity to genotyping errors, temporal heterogeneity of samples, population admixture and structural division into subpopulations, showing a high tolerance to deviations from the assumptions of the model. Computer simulations also show that the proposed method outperforms other leading approaches when the inference concerns recent timeframes. Analysis of a variety of human and animal populations gave results in agreement with previous estimations by other methods or with records of historical events.

Introduction

Several models and sophisticated mathematical tools have been developed to extract demographic information from the growing amount of genomic data. These models focus on different aspects of the genetic variability generated by mutation and recombination. When recombination is not considered, the only free parameter is the mutation rate, which becomes the metronome of the coalescence process (Hudson 1990). Because mutations accumulate slowly, these models are suitable for estimating the effective population size (N_e) from very ancient times (Atkinson et al. 2008) with the limit given by the coalescence time of all the sequences in the sample. The inclusion of recombination reflects better the reality of nuclear genomes and improves the estimations of past N_e not only for more recent times but also for distant times as several genome sequences can be considered in the same analysis (Li and Durbin 2011; Palacios et al. 2015; Terhorst et al. 2017; Schiffels and Durbin 2014; Speidel et al. 2019). However, the role of mutation remains central in the estimation of the lengths of genealogy branches and the impact of recombination is restricted to a small genomic scale. With fairly accurate estimates of N_e in the ancient past of several thousands of generations, these methods are not expected to provide good estimations for very recent timeframes.

Models based exclusively on the theory of linkage disequilibrium (LD) between loci measure the time by the rate of occurrence of recombination events, which typically can take values much larger than mutation rates when loci are distant. Thus, the occurrence of mutations becomes irrelevant and the inference of population sizes from LD concerns essentially the recent demographic history, which is key to understand the current genetic composition of small populations. To some extent, the structure of LD of a population can be described by the distribution of lengths of identity by descent (IBD) segments and, from it, the recent demography can be inferred by the principle that longer segments shared by individuals correspond to more recent common ancestors (Hayes et al. 2003; Palamara et al. 2012; Browning and Browning 2015). However, only long IBD segments, which are infrequent in small samples from large populations, can be reliably identified. Thus, large samples of phased genotypes are usually needed in order to reach some resolution for a general trend.

A simplified representation of the structure of LD is given by the correlation between alleles of pairs of loci (Sved and Hill 2018). Two locus statistics provide additional power over one locus statistics in recovering past demography (Ragsdale and Gutenkunst 2017). This basic theory has proven to be useful for estimating the current N_e of small populations from LD between unlinked loci (Waples 2006; Sved et al. 2013; Waples and Do 2008; Wang et al. 2016) and has also been extended to infer changes in N_e in the recent past from LD between linked loci (Hayes et al. 2003; Tenesa et al. 2007; Qanbari et al. 2010; Corbin et al. 2012; Mörseburg et al. 2016). The fundamental idea is that LD between pairs of SNPs at different genetic distances provides differential information on N_e at different time points in the past.

Several methods assume that the expected LD between loci at a particular recombination rate is the result of genetic drift at a particular generation (Barbato et al. 2015; Mezzavilla and Ghirotto 2015; Hollenbeck et al. 2016). By assuming that the observed LD between loci pairs at a genetic distance $1/(2t)$ Morgans reflects the N_e value t generations back in time, they are able to estimate general trends with slow increases or decreases in population size, which is a remarkable achievement for a rather simplistic approach. However, although LD for closely linked loci depends more strongly on genetic drift occurred far in the past than LD for loosely linked loci, the magnitude of LD between loci at any given genetic distance is the result of the cumulative effects of genetic drift (determined by N_e^{-1}, which generates LD) and recombination (determined by genetic distance, which reduces LD) occurred over all the previous generations.

Here, we derive equations for the expected contributions of each of the past generations to the LD of pairs of loci separated by a particular genetic distance. We also develop corrections for the sampling effects (i.e. LD due to finite sample size), covering the most general types of SNP data from both genotyping and DNA sequencing: diploid unphased genotypes, diploid phased genotypes and pseudo-haploid genotypes of low-coverage genomes usually resulting from sequencing ancient DNA (Haak et al. 2015). Based on the principle that the observed LD for different genetic distances provides differential information of past N_e at different generations, we develop an iterative optimization approach (GONE; Genetic Optimization for N_e Estimation) to infer the recent demographic history of a population from SNP data of a small sample of contemporary

individuals. The method is validated by simulation under different demographic scenarios, and is compared with the previous leading methods, MSMC (Schiffels and Durbin 2014), Relate (Speidel et al. 2019) and the algorithms used by previous LD-temporal N_e methods, such as SNeP (Barbato et al. 2015), NeON (Mezzavilla and Ghirotto 2015) or LinkNe (Hollenbeck et al. 2016). We next inferred the historic population sizes from a number of real datasets from animal and human populations.

Results

Theoretical developments

We derived the expectations for the squared covariance between the alleles of a given pair of loci (D^2) and the product of their two genetic variances (W), such that the linkage disequilibrium (LD) between the loci is measured by the standardized quantity $\delta^2 = E[D^2]/E[W]$ (Ohta and Kimura 1969) (see Supplementary File).

Constant effective population size: When population size is kept constant over generations, the expected values $E[D^2]$ and $E[W]$ in consecutive generations can be obtained by considering a third statistic $E[D(1-2p)(1-2q)]$, where p and q are the allele frequencies at both loci (Hill and Robertson 1968; Hill 1975). This third statistic is equivalent to the moment of order (2,2)th that we approximate in terms of D^2 and W by assuming that most of the new LD produced at any generation is built by drift acting on old variation (see Supplementary File).

At equilibrium, after many generations with constant effective population size N_e, constant mutation rate and recombination rate c, δ^2 can be predicted by N_e and c as

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

Note that δ^2 is, in fact, the squared correlation coefficient $r^2 = D^2/W$ (Hill and Robertson 1968; Rogers 2014) weighted by the product of variances, *i.e.* $\delta^2 = E[r^2W]/E[W]$. Under simplified assumptions (negligible c^2 and N_e^{-1}), equation (1) is close to the classical Sved´s (1971) approximation, $r^2 \approx 1 / [4N_ec + 2]$, for the case of unknown phase. Equation (1) is valid for the whole range of c values. For independent loci ($c = ½$), neglecting the term N_e^{-1}, equation (1) is simplified to $5/(6N_e)$. Likewise, the corresponding equation for haploid genomes (Eqn. S2 in Supplementary File) reduces to $2/(3N_e)$. The quantitative difference

between δ^2 and r^2 has been considered typically small, particularly for intermediate allele frequencies. However, important biases in the estimation of N_e could be found if r^2 instead of δ^2 is used (Supplementary Fig. S1).

In practice, sampling could also generate LD (equivalent to one extra-generation of recombination and drift) and thus its effects need to be corrected to obtain the population estimate of δ^2. Approximate corrections for several data types (haploids, phased diploids, unphased diploids and pseudo-haploid genomes) are given in the Supplementary File.

Variable effective population size: When population size changes with time, the above equation for δ^2 does not hold and the historical series of N_e cannot be inferred from a single δ^2 value. For a particular recombination rate (c), the expectation of the current D_c^2 can be expressed as

$$E[D_c^2] \approx \sum_{g=0}^{\infty}(C_g \cdot 2N_g\mu) \ ,$$

where C_g (Supplementary File) is the contribution to the current squared covariance of a single mutation occurred at generation g back in time and the term $2N_g\mu$ is the number of new mutations at that generation, N_g being the effective population size at generation g and μ the mutation frequency that is assumed to be constant across loci and generations.

In the same way, $E[W_c]$ can be expressed as (Supplementary File):

$$E[W_c] = \sum_{g=0}^{\infty}(w_g \cdot 2N_g\mu) \approx \mu \sum_{g=0}^{\infty} \left[V_x \cdot \prod_{i=0}^{g-1}\left(1-\frac{1}{N_i}\right)\right],$$

where w_g is the contribution to the current product of variances from a mutation occurred at generation g, and V_x is the background neutral variance. The product of the sequence of terms with negative upper bound equals 1. Note that the expression in the right-hand side shows the decline in genetic variation by genetic drift. The ratio of expectations $E[D_c^2]$ and $E[W_c]$ for a particular recombination value c becomes independent of μ,

$$\delta_c^2 = \frac{E[D_c^2]}{E[W_c]} = \frac{\sum_{g=0}^{\infty}(C_g \cdot 2N_g)}{\sum_{g=0}^{\infty}\left[V_x \cdot \prod_{i=0}^{g-1}\left(1-\frac{1}{N_i}\right)\right]} \ .$$

An estimate of the temporal series of N_g values can be obtained from the observed δ_c^2 values for pairs of markers with different recombination rates c. Consequently we developed a genetic algorithm implemented into a computer program (GONE) to search for the temporal N_g values that minimize the sum of squares of the difference between the

expected (calculated above) and observed δ_c^2 values (see Methods). Supplementary Fig. S2 shows the close agreement between the observed and optimized values of δ_c^2 for different demographic scenarios.

Simulation results

Over 10^8 replicates were simulated for each combination of recombination rate and population size in order to check the accuracy of the predictions of δ^2 for constant population sizes for diploids (Eqn. 1) and haploids (Eqn. S2 in Supplementary File). Predictions resulted to be very close to simulations over the whole range of recombination rates (Supplementary Table S1). They are accurate even at the two boundaries of the range of recombination rates $c = 0.5$ and $c = 0$, where the true δ^2 value used to be controversial. Moreover, δ^2 marginally increases when N decreases in both predictions and simulations at both c bounds. The table also shows predictions by other methods.

We evaluated GONE for the ability to infer the true historic series of N_e values of simulated populations. Inferences were carried out from LD data between loci with recombination frequencies from 0.001 to 0.5. Several profiles of changes in population size were simulated, and the resulting genetic data were analyzed by GONE in comparisons with three of the leading methods, MSMC (Schiffels and Durbin 2014), Relate (Speidel et al. 2019), and the algorithms used by the previous LD-temporal N_e methods (such as SNeP, NeON or LinkNe). The results are shown in Figure 1 for a representative sample of demographic scenarios. Within the range of the most recent 200 generations, GONE outperforms any of the other methods, which are, at most, able to detect a general trend for both phased and unphased data. The previous LD-temporal N_e approach, which is a simple method based on bi-locus LD, performs fairly well when compared with Relate and MSMC, particularly for unphased data. Relate is prone to large deviations in recent generations, which suggests that coalescence methods are better suited for ancient N_e estimations.

Figure 2 illustrates different characteristics of the estimations by GONE. First, the accuracy of the estimations decreases with time: Ancient demographic changes, like a bottleneck at generation 140 in the figure (panel B), are detected with lower precision than recent ones (panel A). Second, overlapping generations causes some underestimations in

the recent generations estimates and a wildly series of estimates in latter generations (Panel C). Third, the inferences from synthetic populations created by mixing of several populations in past times do not show distortions in N_e estimations from the time of mixing to present (panel D). Fourth, no distortion or bias occurs when the analysis deals with metapopulations structured according to the standard island model, and the migration rate between subpopulations is low without extinctions (panel E). The estimates correspond to the total size of the metapopulation, in agreement with the expected effective population size from the classical N_e theory. However, there are substantial biases in the estimates for recent generations when the migration rate is high (panel F). Fifth, base calling errors do not affect estimates in a significant way if they are not larger than 1%, which is a reasonable assumption for data from common commercial genotyping and sequencing platforms (panel G). Other methods need high quality sequences or the application of a threshold MAF to eliminate the distortion caused either on genealogies or on correlations between alleles at different loci. Sixth, the sampling of non-contemporary individuals causes a bias in the estimations of the most recent generations (panel H). This scenario assumes that each of the individuals are sampled in each of the last 100 generations. The distortion in these estimates seems to be significant but affecting a time of inference which is smaller (about a quarter) than the length of the sampling period. Finally, the random selection of individuals of a small sample leads to differences in the estimations from different samples, particularly for the most recent generations (panel I). These differences are mitigated if data from distant loci (say $c > 0.05$) are not included in the analysis, leading to more consistent estimations (panel J).

Application to real data

We next apply the method to make inferences on the recent demographic changes of several human and animal populations (Figure 3) with large differences in size. In order to reduce the effect of sampling in recent generations observed in simulations, LD data for recombination frequencies larger than 0.05 were excluded from the analysis. Inferences of N_e from a herd of domestic pigs, which was founded from a population of unknown origin and then maintained under controlled mating conditions for 26 generations before sampling, are in agreement with estimates obtained from the observed genealogical

information of individuals (Saura et al. 2015) except for generations close to the setup of the population. This deviation is exactly the kind of artifact expected after mixing of different populations as shown by simulations (Fig. 2D).

The estimated N_e values in pigs contrast with the large recent N_e values inferred from a sample of 99 individuals from the Finnish population, which has experienced a rapid growth during the last 15 generations. In this case, the data refers to sequencing analysis and a large number of SNPs (more than 9 million) were available. Thus, 20 replicates of estimation were carried out for each of which 50,000 SNPs were randomly sampled per chromosome. The red thick line is the average over replicates and the shadow area gives the interval of confidence obtained from the replicates. These estimations show some differences with a previous study based on the analysis of IBD segments of a much larger sample of 5,402 individuals (Browning and Browning 2015). While the IBD inference assumed a monotonic increase of population size, we detect a reduction in the Finnish population during the middle ages, which could be in fact a result of the admixture of partially differentiated populations in iron age and medieval times (Översti et al. 2019). Our estimations for recent times are clearly under the actual numbers of Finns. This deviation can only be partially explained by the substantial differences between effective sizes (N_e) and census sizes (N) generally observed in natural populations. In general, large sample sizes (n) are needed by GONE to infer large population sizes with some precision (see Methods), particularly for very recent generations, which relates to the difference between the drift signal (proportional to $1/N$) and the magnitude of sampling error (proportional to $1/n$). Additionally, Figure 3 shows that the alternative use of a map with constant recombination rate of 1.2 cM/Mb across the genome (thin continuous line) does not make a big difference in the estimations of demography of the Finnish population.

The analyses of salmon samples composed by individuals born between 1985 and 1992 from two tributaries of River Dee in Scotland highlights the consistency of the method when applied to replicates. Both estimates are coincident with a drop in population size about 10 generations before sampling. While fine-scale recombination maps were used for pigs and humans, this salmon analysis assumes a constant rate of recombination of 1 cM/Mb for the whole genome, which is an approximated average of estimates by several authors (Philips et al. 2009; Lien et al. 2011; Tsai et al. 2016). Salmon genome underwent a

recent event of diploidization and several chromosome rearrangements (Lien et al. 2016) and is still polymorphic for some of them. Consequently, there is a lack of continuity between the assumed physical and the estimated genetic maps but, by ignoring large recombination rates (over $c = 0.05$ in this analysis), we avoid most complications due to gaps or lacks of continuity.

Analysis of samples of ancient human remains dated between 2,500 and 4,500 years BCE (Olalde et al. 2018) produces N_e estimates between 2,000 and 6,000 individuals from two Scottish samples. The "random draw" method of genotyping of these ancient-DNA samples results in pseudo-haploid genomes (Haak et al. 2015). While other N_e estimators do not perform adequately with this type of data, our method can be straightforwardly modified to accommodate it (Supplementary File). Simulation results accounting for an extended sampling period of 100 generations (Fig. 2H) showed estimation bias for about a quarter of the time of sampling. Therefore, most recent N_e estimations from these samples should be disregarded.

Inferences from two samples of Ashkenazi Jews from Eastern and Western Europe (Behar et al. 2010) show similar N_e trajectories with increased deviations for the most distant generations. The strong reduction in N_e inferred around generation 60 is approximately contemporary with the Jewish-Roman wars of the First Century, which are commonly considered to have contributed to the expansion of the Jewish diaspora across Europe, Africa and Asia (Goodman 2004). The large expansion of this ethnic group in recent times (Slatkin 2004) is not observed in our results, which only show a moderate increase. This, again, illustrates the difficulties of the method in detecting large increases of N_e in recent times from very small samples. The analysis of Mizrahim genomes does not show any decline in N_e at generation 60, which is coincident with the fact that these communities were included in the Parthian Empire by that time and were not affected by the Jewish-Roman wars (Goodman 2004). No significant effect of the later expansion of Islam on N_e is observed but a sharp drop in N_e is detected particularly in Caucasian Mizhrahims, which is coincident with the repeated invasions of the region between the 13^{th} and 16^{th} centuries (Singer et al. 1906), and a later decline is observed in Mizhahims from Iran and Iraq.

Discussion

Our method is able to infer demographic histories within a hundred generations in the past from both phased and unphased genotypes. These short-term inferences appear to be more accurate than those obtained by current coalescence methods. The mapping of mutations to estimate the length of branches of genealogical trees makes coalescence theory rather more suitable for modeling ancient demography because mutations accumulate very slowly in populations. Consequently, estimations from coalescence methods deviate from the real N_e for recent generations as can be observed for Relate estimations from simulated data (Fig. 1). On the contrary, MSMC makes use of the observed changes in heterozygosity across the genome to infer demography, which considers both mutation and recombination events. Although MSMC performs better than Relate, it lacks enough power to resolve recent demographic changes. The reason is probably because few recombination events between consecutive sites are dated in recent times even when eight haplotypes are included in the sample. The inclusion of more haplotypes could improve the recent N_e estimates but the method would probably become computationally intractable.

GONE makes use of the information from a wide range of recombination rates, including distant loci for which at least one crossover event is expected in every meiosis. Every new mutation generates a small amount of LD between the mutation site and any other polymorphic site. This LD is expected to increase by genetic drift over consecutive generations at a rate which depends on N_e. At the same time, LD is constantly removed by recombination at a rate which depends on the genetic distance between loci. Thus, the observed LD between distant loci is mainly the result of the recent drift because the effect of old drift is removed by intense recombination in a few generations, whereas LD between closely linked loci is the result of drift generated both recently and remotely in the past (Hayes et al. 2003).

Relevant aspects of GONE allow the detection of demographic changes in scenarios where previous LD methods fail. One of them is the use of δ^2 (Ohta and Kimura 1969) to measure LD instead of the generally used Pearson's r. The use of r^2 to infer temporal changes of N_e is problematic, as there are not analytic solutions for its sampling error. This makes difficult to reach accurate predictions of the cumulative effects of drift on LD over generations, particularly when the recombination rate is small. The general approximation

by Fisher (1915) for the normal distribution and some related variations (Tenesa et al. 2007) are inaccurate for a bivariate binomial distribution, for which r^2 depends on gene frequencies in an intricate way. On the contrary, δ^2 is the ratio of two statistics whose expectations in consecutive generations can be established . In addition, because δ^2 is a measure of LD weighted by the genetic variances of the involved loci (Rogers 2014), it is much less affected than r^2 by sampling of low frequency variants and by genotyping errors, which usually generate singleton variants in samples. Methods using r^2 (Tenesa et al. 2007; Saura et al. 2015; Mörseburg et al. 2016; etc.) are prone to overestimations of N_e under those circumstances, which are only partially corrected by applying an arbitrary MAF threshold to data (Supplementary Fig. S1). For our method, however, MAF should not be applied *a priori*. In fact, the application of MAF thresholds results in slightly biased estimates of N_e. However, there is one scenario in which MAF thresholds clearly results in improved estimations: when there are sequencing errors. The application of MAF results in acceptable estimates of N_e except when the rate of errors is extremely high (say 10%) (Figure 2G). We have derived accurate and computationally efficient equations to predict the change of δ^2 over consecutive generations. This accuracy is critical because the inference of N_e across time is the result of the comparison of the accumulated contributions of all previous generations to the observed δ^2 values for pairs of loci with different recombination rates. We also derived appropriate corrections for sampling, some of them similar but more accurate than previous developments, and extended them to new sampling methods.

Several authors reached solutions for the expected value of δ^2 (Ohta and Kimura 1971; Hill 1975; McVean 2002; Weir and Hill 1980). Recently Ragsdale and Gravel (2020) developed a combinatorial method to find estimators of several statistics related with δ^2, which were combined with the predictive theory by Hill and Robertson (1968) in order to consider sampling-without-replacement in the genetic transition of a population from one generation to the next one. The resulting predictions of LD at equilibrium when $c = 0.5$ and population size is constant over time, were $\delta^2 = 1/(6N)$ and $\delta^2 = 1/(3N)$ for haploid and diploid populations, respectively. Simulations show that our predictions of δ^2 with constant population size are generally more accurate for the whole range of recombination rates than

356 those predicted by previous theory (Supplementary Table S1). Particularly for $c = 0.5$, our
357 result is $\delta^2 \approx 2/(3N)$ and $5/(6N)$ for diploids and haploids, respectively.
358 As we have explained above, the expected LD for a particular recombination rate is
359 not only a consequence of the N_e at a particular generation. Previous two-loci LD-based
360 methods (Hayes et al. 2003; Tenesa et al. 2007; Barbato et al 2015; Mezzavilla and
361 Ghirotto 2015; Hollenbeck et al. 2016) assume a univocal correspondence between N_e at a
362 particular generation g in the past and the observed LD between pairs of loci with a
363 particular recombination rate $c = 1/(2g)$. This relationship was deduced by Hayes et al.
364 (2003) in the context of the probability that two chromosome segments, which are flanked
365 by two markers with recombination rate c, come from a common ancestor without
366 intervening recombination. As stated by Hayes et al. (2003), this approach would be only
367 valid for constant N_e or a linear increment or decrement of N_e across generations (Hayes et
368 al. 2003). Our method, however, provides a solution for the inference of the historical N_e
369 without any previous assumption on the magnitude or the trend of changes. In addition, the
370 method is quite robust for base-calling errors, deviations for the genetic map and deviations
371 from the assumption of a single unstructured population. Overlapping generations tend to
372 produce underestimations of the recent N_e, as has been reported for the estimations of the
373 current N_e (Waples et al. 2014). Also, while the admixture of differentiated populations
374 distorts the structure of LD, inferences are valid for the derived population up to nearly the
375 generation of admixture.
376 Although all bins for pairs of SNPs at different distances can be used in the
377 estimation procedure, it is advised in practice to ignore those corresponding to the largest
378 recombination frequencies. In fact, the default largest value of c used in our application is
379 0.05. The reason for this is tripled. First, random sampling of few individuals can lead to
380 deviations from the average coancestry of the population (Fig. 2I). The consequences of
381 these deviations on the inference of temporal N_e are larger for large c values than for small
382 ones because genealogies of a finite sample of individuals mix progressively with the
383 population backwards in time. That is, inferences of recent N_e are more affected by
384 sampling than inferences of ancient N_e. These biases are partially corrected by disregarding
385 large values of c (cf. Fig. 2I and 2J). Second, the observed LD for any particular c value
386 does not depend exclusively on the N_e at a particular generation back in time. However,

while LD of SNP pairs with $c = 0.5$ depends on the N_e of a few recent generations (say a couple generations back in time), LD of bins with smaller c values depends on the historical N_e values of a wider span of time from past to present, including the recent generations. As the inferences of N_e at different generations are interconnected in this way, biases in the measure of LD of bins with large c values affect more the inference of the whole series of temporal N_e than biases of LD of small c values do. Finally, when populations are strongly geographically structured, the distortion in LD can be very large (Fig. 2F). This effect is relatively similar to the random sampling of a few individuals in a panmictic population. By ignoring bins of large c values, the distortion in the inference of past N_e is mitigated (see Fig. 2F). Nevertheless, our recommendation of considering the largest value c as 0.05 is a compromise solution which can be changed by the user by setting the switch of this option to any other value between 0 and 0.5. For example, for simulation results, where the sampling of individuals is a random sample of the population, the use of the largest c values is justified unless the sample size is very small.

Inferences by GONE are restricted to recent changes in N_e, with the highest resolution within a hundred generations before sampling. Drastic demographic changes partially erase the linkage disequilibrium footprint of older events. Therefore, if older changes are relatively small or there are many demographic changes involved in the time period considered, the method will fail to detect them accurately or will only detect the most recent ones. The lack of precision of N_e estimates of ancient events (Fig. 2A vs. 2B) could be a consequence of the fact that ancient N_e estimates rely on a large number of measures of LD of different recombination-rate bins. Thus, cumulative errors are expected to be larger for ancient estimates than for recent ones.

To a good approximation, the accuracy of the estimations is proportional to the sample size, to the squared root of the number of pairs of SNPs included in the analysis and to the inverse of the effective population size (see Methods and Supplementary File). That is, halving the sample size can be approximately compensated by doubling the number of SNPs included in the analysis. This is consistent with previous findings related to N_e estimation by the temporal method (Waples 1989). Note, however, that this approximation relies on the assumption that the individuals analysed are a truly random sample from the population. Even so, if the sample size is very small, the accuracy of population parameter

estimates cannot be compensated by a larger number of SNPs. As noted by King et al. (2018), with more and more loci the estimates converge on the true parameter values for the pedigree of the sampled individuals, but not necessarily on the pedigree of the population as a whole. For deep coalescent evaluations this is not such a big problem, as all recent pedigrees coalesce to the same ancestral lineages as one moves back in time. However, this is an important issue for recent generations.

Here we have introduced a method to infer very recent changes in effective population size from the distribution of LD between pairs of SNPs from chip genotyping or sequencing data. Its temporal space of inference is of particular interest in the survey and assessment of perspectives of endangered populations and could also be a useful historiographic tool to study human demography. It is computationally efficient, accurate and fairly stable against deviations from the assumptions of the model such as genotyping errors, non-random mating, admixture of populations, overlapping generations, and alterations of the genetic map. It is applicable to populations with a wide range of demographic changes and different types of genomic data. In summary, this method facilitates the immediate use of a large amount of genomic information to study the recent demography of populations.

Methods

Estimation of the historical N_e

In a first step, SNP data files with *map* and *ped* formats are processed by a custom program to calculate linkage disequilibrium (sample d_c^2) for bins of pairs of SNPs with different genetic distances (c). The analysis is made for individual chromosomes, which can be run in parallel on several processors. It has a number of options: (a) the number and length of bins assumed; (b) the use of the observed genetic distances between SNPs, if available in the *map* file, or the use of genetic distances calculated under the assumption of a given number of cM per Mb of sequence; (c) the use of Haldane´s or Kosambi´s corrections for genetic distances, or none of them; (d) the exclusion or inclusion of SNPs with missing data; (e) the use of phased diploid data, unphased diploid data, or pseudo-haploid data; (f) a predefined maximum number of SNPs to be analyzed per chromosome, taken at random among all available SNPs, and excluding loci with more than two alleles; and (g) the

application of a threshold MAF if desired. Values of d_c^2 from all chromosomes are then combined in a single file for estimation of historical series of N_e, although estimates from individual chromosomes can also be performed.

A second program (GONE) implements a genetic algorithm (Mitchell 1998) to search for the global optimal solution of the historical N_e series that best fits the observed δ_c^2 values, which are obtained from the d_c^2 values previously calculated by the first program, after correction for sample size. The function to be minimized is the sum of the squared differences between observed and predicted δ_c^2 values for the whole range of recombination rates c considered in the analysis. An output of the program is the series of observed and predicted d_c^2 values over the range of recombination rates and the sum of squares of their differences. In this genetic algorithm, an "individual" is a particular sequence of temporal N_e values for all the previous generations. In order to reduce the complexity of the optimization procedure, the entire time space from 0 (i.e. at the sampling point) to an infinite number of generations in the past is split into consecutive blocks, with the same N_e value for all the generations within each block. In order to generate each initial "individual", the time space is randomly split into four blocks with a boundary set at generation $1/c_{min}$, where c_{min} is the minimum c value among all pairs of SNPs included in the analysis, and random N_e values are assigned to each block. Thus, 1,000 "individuals" are randomly generated and fitness values are assigned as the inverse of the sum of the squared differences between observed and predicted δ_c^2 values calculated from the set of N_e values of the "individual". Then, the fittest 100 "individuals" are selected to be parents of the next generation. In order to produce each "individual" of the next generation, two "parents" are randomly selected, "crossovers" (interchange of sections of temporal N_e series) between both "parents" are carried out and "mutations" (changes in the boundaries of blocks and the N_e values of blocks) occur randomly. Each "crossover" introduces a new boundary, but the number of blocks can also be reduced by random "mutations" that merge two consecutive blocks. In this way, a new set of 1,000 "individuals" is generated and selection of parents starts again to produce the next generation. The block from generation $1/c_{min}$ up to infinity will remain without further divisions during the whole optimization. The selective process is repeated for 750 generations and the average N_e series of the best 10 "individuals" is considered to be the solution of the optimization process. As this

480 solution could be an “adaptive peak”, that is a local optimal solution, the selective process
481 is repeated a desired number of times (say 40) and the final solution is calculated as the
482 average value of the available solutions, e.g. 40 × 10 = 400 “individuals”. The replicated
483 estimations can also be run in parallel using several processors. Thus, GONE provides a
484 solution of consensus or general trend for the demographic history of a population. We
485 have found that this solution is more consistent and repeatable than any particular optimal
486 solution. An example of the fit between optimized values of δ_c^2 and the observed simulated
487 values is given in Supplementary Figure S2.
488 The method does not generate parametric confidence intervals for the estimate.
489 However, if the number of SNPs per chromosome is large, such as occurs with sequencing
490 data or with some large chips, it is possible to make estimation replicates by choosing
491 different sets of SNPs per chromosome with a functionality implemented in the scripts, as
492 mentioned above. This would allow empirical confidence limits to be obtained. An example
493 of this application is shown in Figure 3 for the Finnish population.
494
495 **Simulation programs**
496 To check the accuracy and statistical properties of the new LD based N_e estimation method,
497 simulations were performed with the software SLiM (Messer 2013; Haller et al. 2019), a
498 forward simulator of SNPs, as well as with in-house programs. For most cases, sequences
499 of 250Mb of length were run for 10,000 generations assuming absence of selection under
500 different demographic scenarios (changes in N over generations), such as bottlenecks, drops
501 or expansions of the population within the last 200 generations. Mutation and
502 recombination rates per nucleotide were assumed to be $m = c = 10^{-8}$, which implies 1 Mb =
503 1 cM. At the last generation, a sample of n diploid individuals (20 or 100) without
504 replacement was taken for analysis. We also considered sampling with replacement in some
505 cases to check the corresponding estimations under this sampling scenario. In general, no
506 pruning was made regarding MAF, but some simulations were run by applying MAF <
507 0.05 and 0.1 to check the effects of rare alleles. Simulation results were based on 10-100
508 replicates for each scenario. A custom program was used to obtain the *map* and *ped* files
509 needed to start the estimation procedure.
510

Estimation of temporal N_e with other methods

The *map* and *ped* files of a number of simulated scenarios were transformed into the necessary file formats for MSMC (Schiffels and Durbin 2014) and Relate (Speidel et al. 2019) and parameters were set to the default options. Analyses of unphased genotypes were implemented by indicating all the possible phasing modes in MSMC and by randomization of pairs of allele copies of the same individual in Relate. Likewise, the d_c^2 values obtained in the simulations were analysed by assuming the approach of previous estimator of temporal N_e with LD (Tenesa et al. 2007; Barbato et al 2015; Mezzavilla and Ghirotto 2015; Hollenbeck et al. 2016) with the corresponding corrections for phased and unphased genotypes.

Sample size estimation

By assuming some simplifications (Supplementary File), it can be shown that the power of detecting fluctuations in N_e is roughly proportional to:

$$G = \frac{n \cdot \sqrt{\vartheta}}{N_e},$$

where n is the sample size and ϑ is the number of loci pairs included in the analysis. As a general rule for experiments in which the range of c values varies from 0.5 to 0.001, good estimations of effective population sizes are obtained when $G > 100$ and very poor estimations are obtained when $G < 10$.

Generation time

In order to compare inferences of N_e with references to historical events, generation time was set to 30 years for humans (Fenner 2005).

Relationship between physical and recombination maps

A genetic map in centi-Morgans (cM) and a map function are needed to estimate the recombination frequency c between any pair of loci from their physical positions in the genome. A fine-scale recombination map was used for humans (Myers et al. 2005) and an inferred map from data by Tortereau et al. (2012) was used for pigs.

There is not a consensus on physical and genetic maps to date for salmon, probably due to the complexity of the chromosome rearrangements in this species. We used the salmon reference genome assembly ICSASG_v2 (Lien et al. 2011) to assign locations to SNPs and considered a constant ratio of 1 cM/Mb between genetic and physical maps, which is an approximate average over several studies (Philips et al. 2009; Lien et al. 2011; Tsai et al. 2016). Tsai et al. (2016) showed the lack of continuity between the assumed physical and the estimated genetic maps, particularly for some chromosomes, with gaps of up to 150 cM. However, by ignoring recombination rates over 0.05 (with the option *-hc* 0.05) we avoided most complications due to gaps or lacks of continuity in the genome. Note that, at 1cM/Mb, a recombination rate of 0.05 corresponds to 5.3Mb assuming Haldane´s function. Using SNPs closer than this distance makes improbable to have a significant representation of SNP pairs at different sides of a gap.

Samples

The different sample sizes of individuals analyzed (n) and the number of SNPs (N_{SNP}) analyzed in the estimations are as follows. Guadyerbas population of Iberian pig (Saura et al. 2015) (n = 219; N_{SNP} = 19,144), Finnish population (1000 Genomes Project Consortium) (n = 99; N_{SNP} = 1,100,000), Salmon from River Dee (n = 16 for each population; N_{SNP} = 104,354), Neolithic West Scotland (Olalde et al. 2018) (n = 17 [10.8], where the number in brackets refers to the actual sample size disregarding missing genotyping data; N_{SNP} = 552,191), Neolithic North Scotland (Olalde et al. 2018) (n = 21 [14.8]; N_{SNP} = 594,385), Ashkenazi East (Behar et al. 2010) (n = 9; N_{SNP} = 478,394), Ashkenazi West (Behar et al. 2010) (n = 9; N_{SNP} = 477,884), Mizhrahi caucasus (Behar et al. 2010) (n = 12; N_{SNP} = 486,075), Mizhrahi Iran & Iraq (Behar et al. 2010) (n = 15; N_{SNP} = 485,199).

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Program codes, binaries for Linux and Mac, and the scripts necessary to apply the method are available at github address XXXXXXXXX.

(Only for reviewing purposes temporarily at Dropbox address: https://www.dropbox.com/sh/pyvhfjxkia06qz2/AADUH2nwNFk3RtavjWzI4QVRa?dl=0).

Acknowledgments

We thank Humberto Quesada and two anonymous referees for helpful comments, Beatriz Villanueva for providing pig data and fruitful discussion, John Taggart for providing salmon samples and Paloma Morán for collaborating in salmon genotyping and for helpful comments. This work was funded by Agencia Estatal de Investigación (AEI) (CGL2016-75904-C2-1-P), Xunta de Galicia (ED431C 2016-037) and Fondos Feder: "Unha maneira de facer Europa". UVigo Marine Research Centre (CIM-UVIGO) is funded by the "Excellence in Research (INUGA)" Program from the Regional Council of Culture, Education and Universities, with co-funding from the European Union through the ERDF Operational Program Galicia 2014-2020. Pig and salmon genotyping were funded by the Ministerio de Economía y Competitividad of Spain (grants RZ2010-00009-00-00 and RZ2012-00011-C02-00). Irene Novo acknowledges support from a FPU grant from Ministerio de Ciencia, Innovación y Universidades. Antonio F. Pardiñas acknowledges support from a Medical Research Council Project Grant (MC_PC_17212).

Author Contributions

E.S. and A.C. conceived the work and wrote the article. E.S. developed the theory and the computational solution. A.C. designed the structure of data and the analysis. I.N. compared methods. A.F.P. contributed human data and investigations. M.S. contributed animal data and analysis. J.W. provided intellectual input.

References

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56-65.

Atkinson QD, Gray RD, Drummond AJ. 2008. mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Mol Biol Evol*. 25: 468–474.

Barbato M, Orozco-terWengel P, Tapio M, Bruford MW. 2015. SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Front Genet*. 6: 109.

Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, Rootsi S, Chaubey G, Kutuev I, Yudkovsky G, et al. 2010. The genome-wide structure of the Jewish people. *Nature* 466: 238-42.

Browning SR, Browning BL. 2015. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am J Hum Genet*. 97: 404-418.

Corbin LJ, Liu AYH, Bishop SC, Woolliams JA. 2012. Estimation of historical effective population size using linkage disequilibria with marker data. *J Anim Breed Genet*. 129: 257-70.

Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Antropol.* 128: 415-423.

Fisher RA. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 4: 507-521.

Goodman M. 2004. Trajan and the origins of Roman hostility to the Jews. *Past & Present*. 182: 3-29.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, *et al.* 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522: 207-211.

Haller BC, Messer PW. 2019. SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Mol Biol Evol.* 36: 632-637.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13: 635-43.

Hill WG. 1975. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite populations. *Theor Pop Biol.* 8: 117-126.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. *Theor Appl Genet.* 38: 226-231.

629 Hollenbeck CM, Portnoy DS, Gold JR. 2016. A method for detecting recent changes in
630 contemporary effective population size from linkage disequilibrium at linked and
631 unlinked loci. *Heredity* 117: 207-16.
632 Hudson RR. 1990. Gene genealogies and the coalescent process. *Oxf Surv Evol Biol*. 7: 1-
633 44.
634 King L, Wakeley J, Carmi S. 2018. A non-zero variance of Tajima's estimator for two
635 sequences even for infinitely many unlinked loci. *Theor Popul Biol.* 122: 22-29.
636 Li H, Durbin R. 2011. Inference of human population history from individual whole-
637 genome sequences. *Nature* 475: 496-493.
638 Lien S, Gidskehaug L, Moen T, Hayes BJ, Berg PR, Davidson WS, Omholt SW, Kent MP.
639 2011. A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals
640 extended chromosome homeologies and striking differences in sex-specific
641 recombination patterns. *BMC Genom.* 12: 615.
642 Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong
643 JS, Minkley DR, Zimin A, *et al*. 2016. The atlantic salmon genome provides unsights
644 into rediploidization. *Nature* 533: 200-205.
645 McVean GAT. 2002. A genealogical interpretation of linkage disequilibrium. *Genetics* 162:
646 987-991.
647 Messer P. 2013 SLiM: simulating evolution with selection and linkage. *Genetics* 194:
648 1037-1039.
649 Mezzavilla M, Ghiroto S. 2015. Neon: An R package to estimate human effective
650 population size and divergence time from patterns of linkage disequilibrium between
651 SNPS. *J Comput Sci Syst Biol.* 8: 1.
652 Mitchell M. 1998. *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press.
653 Mörseburg A, Pagani L, Ricaut F-X, Yngvadottir B, Harney E, Castillo C, Hoogervorst
654 T, Antao T, Kusuma P, Brucato N, et al. 2016. Multi-layered population structure in
655 island Southeast Asians. *Eur J Hum Genet*. 24, 1605-1611.
656 Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A Fine-Scale Map of
657 Recombination Rates and Hotspots Across the Human Genome. *Science* 310: 321-324.
658 Ohta T, Kimura M. 1969. Linkage disequilibrium due to random genetic drift. *Genet Res.*
659 13: 47-55.

660 Ohta T, Kimura M. 1971. Linkage disequilibrium between two segregating nucleotide sites
661 under the steady flux of mutations in a finite population. *Genetics* 68: 571-580.
662 Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, Rohland N, Mallick
663 S, Szécsényi-Nagy A, Mittnik A, et al. 2018. The beaker phenomenon and the genomic
664 transformation of Northwest Europe. *Nature* 555: 190-196.
665 Översti S, Majander K, Salmela E, Salao K, Arppe L, Belskiy S, Etu-Sihvola H, Laakso
666 V, Mikkola E, Pfrengle S, et al. 2019. Human mitochondrial DNA lineages in Iron-Age
667 Fennoscandia suggest incipient admixture and eastern introduction of farming-related
668 maternal ancestry. *Sci Rep.* 9: 16883.
669 Palacios JA, Wakeley J, Ramashandran S. 2015. Bayesian nonparametric inference of
670 population size changes from sequential genealogies. *Genetics* 201: 281-304.
671 Palamara PF, Lencz T, Darvasi A, Pe'er I. 2012. Length distributions of identity by descent
672 reveal fine-scale demographic history. *Am J Hum Genet.* 91: 809-822.
673 Phillips RB, Keatley KA, Morasch MR, Ventura AB, Lubieniecki KP, Koop BF,
674 Danzmann RG, Davidson WS. 2009. Assignment of Atlantic salmon (*Salmo salar*)
675 linkage groups to specific chromosomes: conservation of large syntenic blocks
676 corresponding to whole chromosome arms in rainbow trout (*Oncorhynchus mykiss*).
677 *BMC Genet*. 10: 46.
678 Qanbari S, E. C. G. Pimentel, J. Tetens, G. Thaller, P. Lichtner, A. R. Sharifi, et al. The
679 pattern of linkage disequilibrium in German Holstein cattle. *Anim Genet*. 41: 346-56.
680 Rasgdale A, Gutenkunst N. 2017. Inferring demographic history using two-locus statistics.
681 *Genetics* 206: 1037-1048.
682 Ragsdale A, Gravel S. 2020. Unbiased estimation of linkage disequilibrium from unphased
683 data. *Mol Biol Evol.* 37: 923-932.
684 Rogers A. 2014. How population growth affects linkage disequilibrium. *Genetics* 197:
685 1329-1341.
686 Saura M, Tenesa A, Woolliams JA, Fernández A, Villanueva B. 2015. Evaluation of the
687 linkage-disequilibrium method for the estimation of effective population size when
688 generations overlap: an empirical case. *BMC Genom.* 16: 922.
689 Schiffels S, Durbin R. 2014. Inferring human population size and separation history from
690 multiple genome sequences. *Nat Genet*. 46: 919-925.

691 Singer I (eds.). 1906. *The Jewish Encyclopedia*. Funk & Wagnalls, New York.
692 Slatkin M. 2004. A population-genetic test of founder effects and implications for
693 Ashkenazi Jewish diseases. *Am J Hum Genet*. 75: 282-293.
694 Speidel L, Forest M, Shi S, Myers SR. 2019. A method for genome-wide genealogy
695 estimation for thousands of samples. *Nat Genet*. 51: 1321-1329.
696 Sved JA. 1971. Linkage disequilibrium and homozygosity of chromosome segments in
697 finite population. *Theor Popul Biol*. 2: 125-141.
698 Sved JA, Cameron EC, Gilchrist AS. 2013. Estimating effective population size from
699 linkage disequilibrium between unlinked loci: Theory and application to fruit fly
700 outbreak populations. *PLOS One* 8: e69078.
701 Sved JA, Hill WG. 2018. One hundred years of linkage disequilibrium. *Genetics* 209: 629-
702 636.
703 Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007.
704 Recent human effective population size estimated from linkage disequilibrium.
705 *Genome Res*. 17: 520-526.
706 Terhorst J, Kamm JA, Song YS. 2017. Robust and scalable inference of population history
707 from hundreds of unphased whole-genomes. *Nat Genet.* 49: 303-309.
708 Tsai HY, Robledo D, Lowe NR, Bekaert M, Taggart JB, Bron JE, Houston RD. 2016.
709 Construction and annotation of a high density SNP linkage map of the Atlantic salmon
710 (*Salmo salar*) genome. *G3.* 6: 2173-2179.
711 Tortereau F, Servin B, Frantz L, Megens HJ, Milan D, Rohrer G, Wiedmann R, Beever
712 J, Archibald AL, Schook LB, Groenen MAM. 2012. A high density recombination
713 map of the pig reveals a correlation between sex-specific recombination and GC
714 content. *BMC Genom.* 13: 586.
715 Wang J, Santiago E, Caballero A. 2016. Prediction and estimation of effective population
716 size. *Heredity* 117: 93-206.
717 Waples RS. 1989. A generalized approach for estimating effective population size from
718 temporal changes in allele frequency. *Genetics* 121: 379-391.
719 Waples RS. 2006. A bias correction for estimates of effective population size based on
720 linkage disequilibrium at unlinked gene loci. *Conserv Genet*. 7: 167-184.

721 Waples RS, Antao T, Luikart G. 2014. Effects of overlapping generations on linkage
722 disequilibrium estimates o effective population size. *Genetics* 197: 769-780.
723 Waples RS, Do C. 2008. LDNE: a program for estimating effective population size from
724 data on linkage disequilibrium. *Mol Ecol Resour*. 8: 753-756.

Figure 1.

Estimates of temporal N_e of simulated populations from phased (left) and unphased (right) data under different demographic scenarios from present (generation 0) to 220 generations in the past. The green area is the true (simulated) population size. The black, red, blue and purple lines are respectively estimations by GONE, MSMC, Relate and LinkNe software. Samples were composed of 4 diploid individuals (8 haplotypes) for MSMC and 20 diploid individuals for the other methods. The total number of SNPs involved in the estimations ranged between 255,000 and 450,000 depending on the scenarios. No MAF threshold was applied to the data.

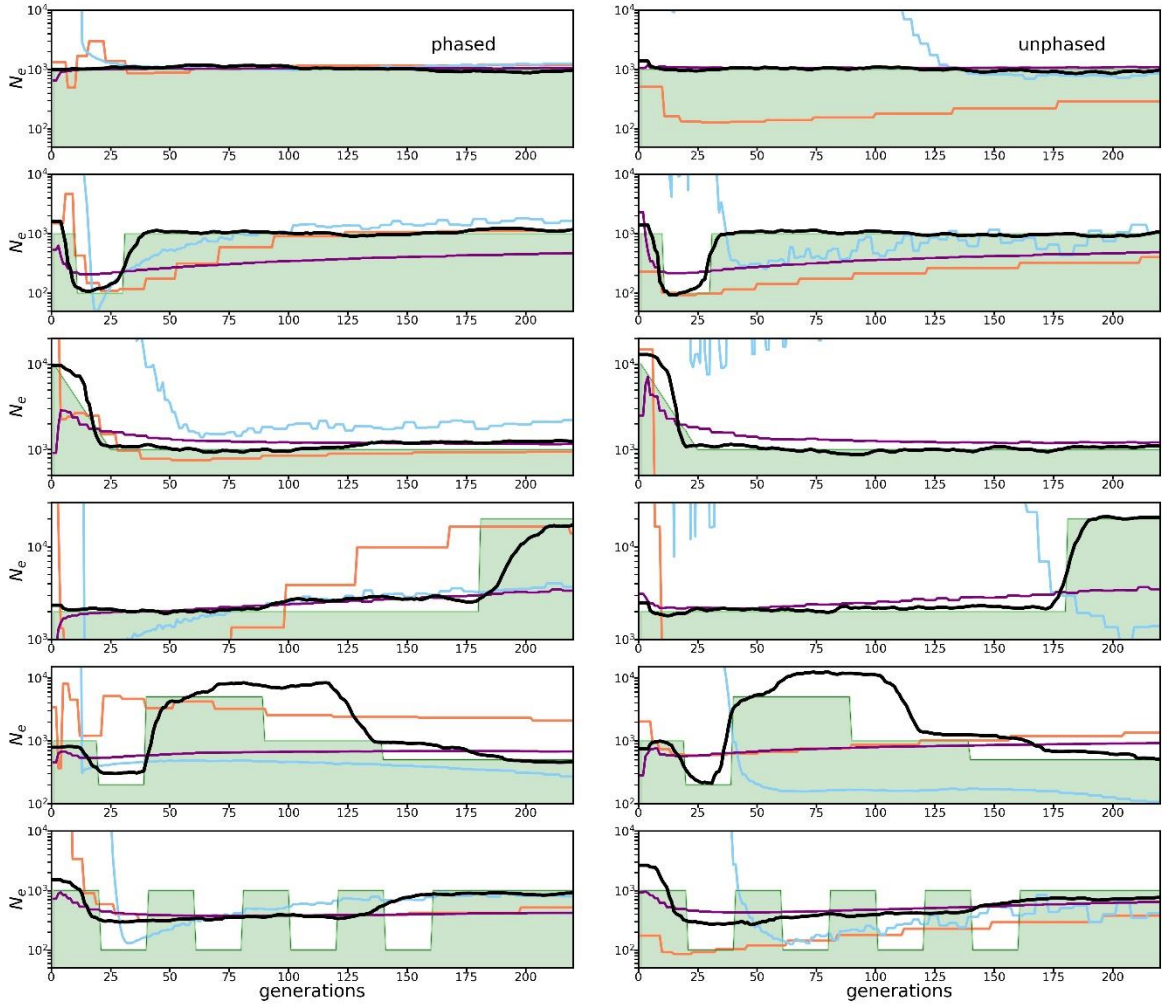

Figure 2.

Estimates of temporal N_e by GONE (red line) under different simulated demographic scenarios from present (generation 0) to 220 generations in the past. The true population size is the green shadowed area and n is the sample size of individuals for analysis. For all panels, the black lines refer to an analysis where all recombination bins from $c = 0.001$ up to $c = 0.5$ are considered (option $hc = 0.5$), whereas the red lines refer to analyses with rate bins from $c = 0.001$ up to only 0.05 ($hc = 0.05$). **(A)** and **(B)**: Detection of bottlenecks occurring at different times. **(C)**: Scenario with overlapping generations with three cohorts per generation and mixed-cohort sampling. **(D)**: A population $N_e = 1000$ was divided into two populations $N_e = 1000$ each, which were isolated for 100 generations and then mixed 50 generations ago into a single population with $N_e = 1000$. **(E)** and **(F)**: Metapopulation composed of two subpopulations $N_e = 1000$ each with 2% and 0.2% of migration, respectively, between them. **(G):** Estimations under different base-calling error rates. From top to bottom, 10%, 1%, 0.1% and 0%, the latter two being indistinguishable. **(H)**: A hundred individuals were sampled from the population over a period of 100 consecutive generations at a rate of one sampled individual per generation. **(I)** and **(J)**: Eight small samples ($n = 10$ each) were taken from the same population at the same time.

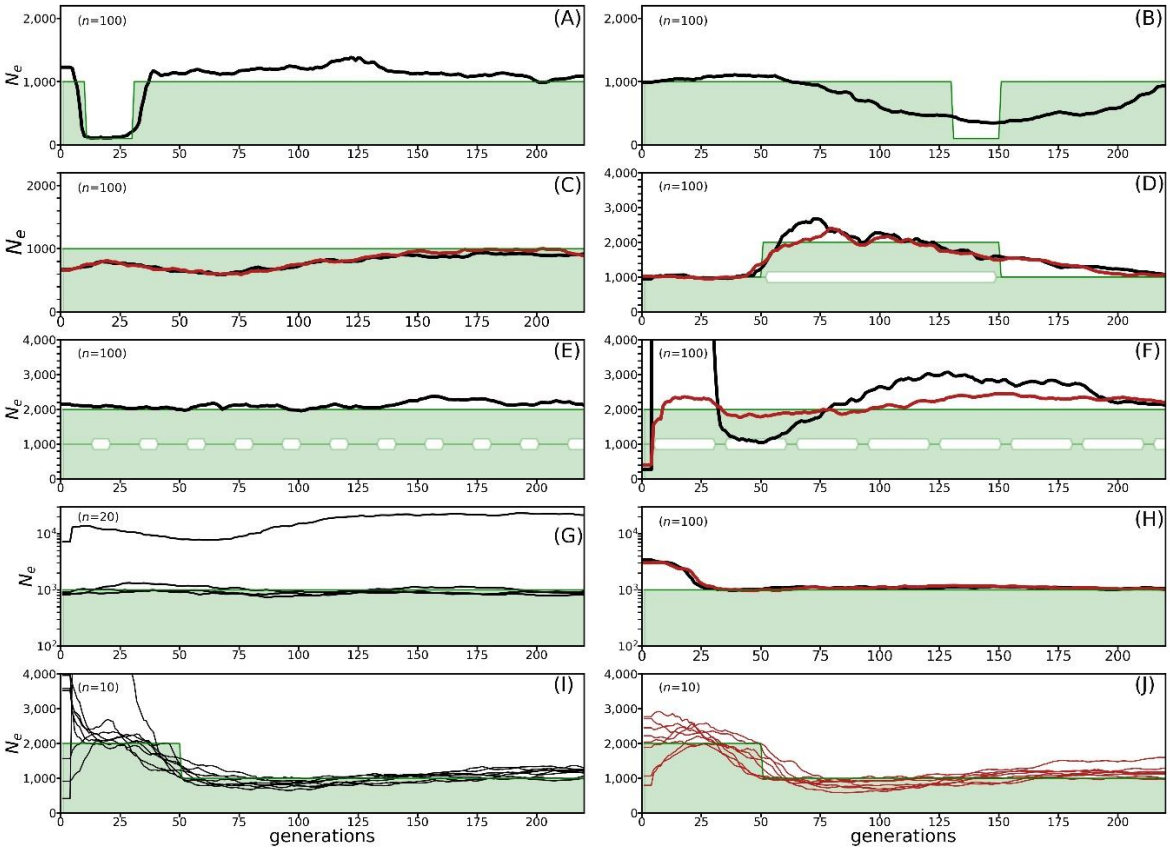

Figure 3.

Estimates of temporal N_e of real populations with different sample sizes (n). **PIGS:** Guadyerbas population of Iberian pigs. The thin blue line is the estimate of N_e using the individual contributions from genealogical data (Saura et al. 2015). **FINNISH:** Estimates of Finnish human population. The shadow area gives the confidence interval of the estimates obtained by running 20 replicates, each one corresponding to a random sample of 50,000 SNPs for each chromosome. The thin broken blue line is the estimation obtained by Browning and Browning (2015) for a Northern Finnish NFBC sample of 5,402 individuals. The thin green line is the estimate of N_e assuming a constant recombination rate of 1.2 cM per Mb. **SALMON DEE:** Atlantic salmons of two tributaries of River Dee in Scotland. **NEOLITHIC:** Two neolithic samples from West and North Scotland, where the sampling period accounts for about 60 generations. **ASHKENAZI JEWS:** Samples of eastern and western European populations. **MIZHRAHI JEWS:** Samples from a Caucasus population and from Iran and Iraq. All estimations assume no MAF threshold and unphased genomes except for the NEOLITHIC, which involves pseudo-haploid genomes.