Natural selection does not affect the estimates of effective population size based on linkage disequilibrium

Irene Novo,^{1,*} Enrique Santiago² and Armando Caballero¹

¹ Centro de Investigación Mariña, Universidade de Vigo, Facultade de Bioloxía, 36310 Vigo, Spain.

² Departamento de Biología Funcional, Facultad de Biología, Universidad de Oviedo, 33006 Oviedo, Spain

* Correspondence: irene.novo.gimenez@uvigo.es (I.N.)

Abstract

The effective population size (N_e) is a key parameter to quantify the magnitude of genetic drift and inbreeding, with important implications in human evolution. The increasing availability of high-density genetic markers allows the estimation of historical changes in N_e across time using measures of genome diversity or linkage disequilibrium between markers. Selection is expected to reduce diversity and N_e , and this reduction is modulated by the heterogeneity of the genome in terms of recombination rate. Here we investigate by computer simulations the consequences of selection (both positive and negative) and of recombination rate heterogeneity in the estimation of historical N_e . We also investigate the relationship between diversity parameters and N_e across the different regions of the genome using human marker data. We show that the estimates of historical N_e obtained from linkage disequilibrium between markers (N_{eLD}) are virtually unaffected by selection. In contrast, those estimates obtained by coalescence mutation-recombination-based methods can be strongly affected by it, what could have important consequences for the estimation of human demography. The simulation results are supported by the analysis of human data. The estimates of N_{eLD} obtained for particular genomic regions do not correlate with recombination rate, nucleotide diversity, polymorphism, background selection statistic, minor allele frequency of SNPs, loss of function and missense variants and gene density. This suggests that N_{eLD} measures are merely indicative of demographic changes in population size across generations.

36 **Author summary**

37 The inference of the demographic history of populations is of great relevance in
 38 evolutionary biology. This inference can be made from genomic data using coalescence
 39 methods or linkage disequilibrium methods. However, the assessment of these methods
 40 is usually made assuming neutrality (absence of selection). Here we show by computer
 41 simulations and analyses of human data that the estimates of historical effective
 42 population size obtained from linkage disequilibrium between markers are unaffected
 43 by natural selection, either positive or negative. In contrast, estimates obtained by
 44 coalescence mutation-recombination-based methods can be strongly affected by it,
 45 which could have important consequences for recent estimations of human demography.
 46 Thus, only linkage disequilibrium methods appear to provide unbiased estimates of the
 47 population census size.

Introduction

The effective population size (N_e) is a parameter of paramount relevance in evolutionary biology, plant and animal breeding and conservation genetics, because its magnitude reflects the amount of genetic drift and inbreeding occurring in the population¹. The effective size of a population depends on its demographic history and structure as well as the selection regime affecting the population²⁻⁴. Estimates of N_e can be obtained by methods using information from genetic markers^{3,5,6}, and those based on linkage disequilibrium (LD) between them are generally acknowledged to be reliable and robust.⁷⁻⁸ The idea behind these methods is that, for neutral loci in an isolated population LD is inversely proportional to the genetic distance (or recombination rate, c) between any pair of markers and the effective size of the population.⁹

With the increasing availability of high-density marker information, such as that of single nucleotide polymorphisms (SNP) panels and whole genome sequences for more and more species¹⁰, methods based on LD that allow an estimate of the temporal changes of N_e in the recent past have been developed.¹¹⁻¹³ The basic 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. Thus, Hayes and colleagues¹¹ suggested that LD between loci with a recombination rate c approximately reflects the ancestral effective population size $1/(2c)$ generations ago. Thus, they proposed to estimate N_e at a given generation t from pairs of SNPs at a genetic distance $1/(2t)$ Morgans. This method has become increasingly popular for estimating the past and present N_e in human^{12,14} and livestock^{15,16} populations, and a number of bioinformatic tools have been developed to allow its implementation (e.g. Hollenbeck et al.¹⁷).

The original application of the above method for estimating historical N_e is, however, restricted to the assumption of constant or linear population growth or decline.¹¹ Thus, drastic population size changes such as bottlenecks or sudden severe declines in census size, which are common in natural populations or at the start of breeding programs, cannot be detected accurately with this method. A late development has been shown to accurately detect drastic changes in historical N_e .¹³ Over relatively recent timespans of about 200 generations back in time, the method has been shown to be more accurate than other alternative coalescence and mutation-recombination-based methods, such as MSMC¹⁸ and Relate,¹⁹ which are expected to be applied for long term evolutionary estimations.

All methods of estimation of past N_e trajectories assume neutrality (absence of selection). It is well known, however, that selection significantly reduces N_e ,^{2,20-21} particularly in regions of low recombination.^{2,22-27} Natural populations are predicted to encode many deleterious variants²⁸⁻²⁹ that can affect the outcome of N_e estimation methods. The fate of these variants, as well as of that of advantageous ones, also relies on linkage, because selection is less effective in genomic regions of low recombination.³⁰

Genomes are also heterogeneous for genetic variation due to differences in recombination rates across chromosomal regions³¹⁻³⁴ and because of the differential impact of natural selection on them.³⁵ For example, selective sweeps of favourable mutations are expected to hitch-hike close-by neutral SNPs producing sharp decreases in diversity in linked regions.³⁶⁻³⁸ Ignoring the heterogeneity in N_e may lead to biased estimates of past demography³⁹ and, in fact, a heterogeneity of N_e across the genome has been found for different eukaryotic species.⁴⁰⁻⁴²

It has been shown that selective sweeps of favourable mutations generate LD between close-by neutral loci,⁴³ although this LD increase can be negligible⁴⁴ and, in any case, it is transient and disappears quickly.^{38,45} If so, estimates of N_e based on LD would be unaffected by selection. In this paper we assess the impact of selection on the estimation of historical N_e using the method of Santiago and colleagues¹³, which is based on linkage disequilibrium between SNP markers, in comparison with other coalescence mutation-based methods, MSMC¹⁸ and Relate.¹⁹ Using individual-based forward simulations, we compare the estimates of historical N_e provided by these methods assuming selective sweeps of favourable mutations and background selection on deleterious mutations, and considering the heterogeneity in recombination rates across the genome. In addition, we investigate the relationship between the estimates of linkage disequilibrium N_e and other diversity and genomic parameters across the human genome, using SNP data obtained from genome sequencing of Finnish⁴⁶ and Koryaks⁴⁷ populations. We obtained the correlation between the estimates of local N_e and several diversity variables over small windows across the genome. Both the simulation results and the analyses of human genome data provide strong evidence that estimates of effective population size based on LD reflect the true census of reproductive individuals in populations.

Results

Effect of selection and recombination on the estimation of N_e

Figure 1 shows the joint effect of recombination and selection on the estimates of N_e , for a population of constant size of $N=1,000$ individuals assuming different recombination rates (RR) per Mb across the genome. Under a neutral scenario, linkage

disequilibrium estimates (N_{eLD}) provide virtually unbiased estimates of N for all recombination rates. The same result can be observed, as expected, for estimates of N_e based on nucleotide diversity (π) and calculated as $N_{e\pi} = \pi/(4\mu)$, where μ is the nucleotide mutation rate, assumed also to be constant across the genome. Estimates obtained from Relate ($N_{eRelate}$) and MSMC (N_{eMSMC}) give also precise estimates of N except for the most extreme cases of recombination rate 5 and 0.01 cM/Mb.

Under background selection and selective sweep models, N_{eLD} estimates give the same results as for the neutral model, indicating that these estimates are very little or not affected by selection, either negative or positive (Fig. 1). As expected, estimates of $N_{e\pi}$ diverges from the true census size (N) of the population, with decreasing values as recombination rate decreases. Relate and MSMC estimates show a pattern similar to that of $N_{e\pi}$ but $N_{eRelate}$ estimates increase sharply for tight linkage scenarios.

- Figure 1 -

Simulation results for historical N_e estimates

Figure 2 shows estimates of historical N_e assuming a constant population census size ($N = 1,000$ or $10,000$), either with a constant or a variable recombination rate. Estimates of historical N_{eLD} reflect the population census size regardless of selection and the variability in genomic recombination rates. Estimates from $N_{eRelate}$ and N_{eMSMC} can give unbiased values of N under a neutral and background selection model if the initial generations are discarded and population size is not too large ($N = 1000$). Otherwise, they can show important differences from the population size, particularly under a selective sweep model, where N is underestimated by $N_{eRelate}$, and can be either underestimated or overestimated by N_{eMSMC} . In general, variation in the recombination

rate across the genome has a limited effect on the estimates of N with respect to a constant recombination rate model.

- Figure 2 -

Regarding historical estimates with variable N_e (Fig. 3), estimates of N_{eLD} predict accurately a recent demographic change in population size (occurred 30 generations in the past) regardless of selection and recombination rate heterogeneity. Relate's estimates show a certain noise, particularly for selective sweeps and/or variable recombination rate scenarios, but give the approximate N value except for an underestimation in several selective sweep cases. MSMC's estimates are also accurate, except for the most recent generations in the exponential N growth case. Selective sweeps tend to lead to an underestimation of N_{eMSMC} , and recombination rate variability only biases estimates in the recent exponential growth case.

- Figure 3 -

When population size changes occur in more ancient times (around 300 generations ago; Fig. 4), both $N_{eRelate}$ and N_{eMSMC} are unable to detect these demographic changes, while N_{eLD} is generally more accurate. Nevertheless, N_{eLD} estimates can be noisy, and sometimes, unable to detect the change in N when selection is present. A variable recombination rate does not substantially affect the estimates.

- Figure 4 -

Correlation between regional estimates of π and N_{eLD} with other diversity parameters using human data

The correlation between the average nucleotide diversity (π) in each genomic region and the other related variables followed the expected trends (Fig. 5a). A strong positive correlation was found between π and recombination rate RR , the background selection statistic B , nucleotide polymorphism P and MAF of SNPs. Nucleotide diversity was also weakly negatively correlated with loss-of-function, missense mutations and gene density, but only significantly for the Finnish population. Finally, no correlation was found between π and N_{eLD} .

- Figure 5 -

The correlations among all genetic variables studied are shown in the Supplemental material (Fig. S2), and follow the expected trends. For example, recombination rate RR , the B statistic, polymorphism P and MAF of SNPs were highly positively correlated among them. The B statistic was strongly negatively correlated with gene density and deleterious variation (LoF, missense), and these latter were highly correlated among them.

The mean estimates (and standard deviation) of N_{eLD} across regions were 8,377 (5,462) for the Finnish population and 2,034 (3,660) for the Koryaks population. Thus, the standard deviation of the regional estimates of N_{eLD} relative to a mean of one, for comparison, were 0.65 for the Finnish population and 1.80 for the Koryaks population.

The correlation between the estimates of N_{eLD} and other diversity parameters for different genomic regions are shown in Figure 5b. The correlations did not follow the trends observed for nucleotide diversity. In particular, there was no significant

correlation between N_{eLD} and π , RR , B , P , MAF, loss-of-function variants, missense variants and density of genes, except for a small significant correlation between N_{eLD} and P for the Koryaks population.

Discussion

Our results show that the estimates of census population size obtained from linkage disequilibrium between pairs of SNPs (N_{eLD}) are virtually unaffected by either positive or negative selections. This has been deduced from simulation data assuming constant or variable population size and different selection models (Figs. 1-4). The simulation results are supported by those obtained from real human genomic data. The estimates of N_{eLD} in genomic windows are uncorrelated with recombination rate, the B statistic (which quantifies the strength of background selection), nucleotide diversity and polymorphism, as well as the number of deleterious variants (loss-of-function and missense variants) and density of genes (Fig. 5). Thus, the results show that the estimates of N_{eLD} are basically unaffected by selection.

Although positive selection is known to generate linkage disequilibrium between neutral loci closed to selective loci,⁴³ this effect is transient and disappears very fast.^{38,45} Stephan and colleagues⁴⁵ showed that the increase in linkage disequilibrium between two neutral loci occurs if both are at less than a genetic distance of $c = 0.1s$, where s is the selection coefficient of the advantageous allele in homozygosis. We performed simulations with $s = 0.02$, which implies that LD is generated between loci located at a genetic distance of $c = 0.002$, or 0.2 cM. In the most extreme linkage scenario simulated in Figure 1 we assumed a rate of recombination of $RR = 0.01$ cM per Mb, which implies a total genetic distance of 1 cM for the whole simulated genome sequence of 100 Mb.

Even in this extreme scenario, the estimates of N_{eLD} are unaffected by positive selection. Note that, although some transient disequilibrium may be generated during selective sweeps for close-by neutral loci, the estimates of N_{eLD} are based on all pairs of loci across the genome, so it is expected that many pairs are not affected by any effect of selection on linkage.

In contrast, and as it is well known (e.g. Charlesworth², Santiago and Caballero²⁷ Gillespie⁴⁸), the nucleotide diversity effective population size ($N_{e\pi}$) is drastically reduced in regions of low recombination under selection (Fig. 1). The observed strong correlations between the regional genomic values of π and the recombination rate, the background selection statistic, and the deleterious variants from human data (Fig. 5a), also confirms this observation. Estimates of N_e obtained by mutation-recombination-based coalescence methods (MSMC and Relate) are also affected by selection (Figs. 1-4). In fact, a Relate Selection Test has been proposed based on estimating the speed of spread of a particular lineage relative to other competing lineages.¹⁹ MSMC estimates generally follow the pattern of $N_{e\pi}$ estimates except for large recombination rates, for which it provides overestimates of the true population size (Fig. 1). Relate estimates also follow $N_{e\pi}$ estimates down to recombination rates of 0.1 cM/Mb but, for lower values, the estimates increase drastically above the true population size (Fig. 1). These coalescence methods have been used to investigate ancient demography of human populations and are not generally applicable to short-term historical changes in population size (see Figs. 2-4). In fact, it has been acknowledged that MSMC with 8 haplotypes works for estimations before about 70 generations in the past, i.e. about 2,000 years for human populations¹⁸, and Relate seems to discriminate after about 1,000 years back.¹⁹ Thus, MSMC was able to detect the out-of-Africa bottleneck in non-

African populations from 200,000 years ago until 50,000 years ago,¹⁸ while Relate detected it from 40,000 to 20,000 years ago.¹⁹ The possible impact of selection on these demographic inferences is an issue to be further investigated.

The effect of recombination rate heterogeneity on historical N_e estimates is not very noticeable in most cases (Fig. 2-4), particularly when GONE and MSMC are used. This is in accordance with the analyses made by Schiffels and Durbin (their Supplementary Figure 4)¹⁸, showing that simulated estimates from MSMC obtained using chunks of the human recombination map do not differ much from those using a constant recombination rate. For GONE estimates, varying recombination rate seems to only affect some of the background selection scenarios, while estimates obtained under a neutral and selective sweep models remain similar to those obtained with the constant recombination rate scenario. For Relate estimates, recombination rate heterogeneity seems to affect the estimates of recent demographic changes (Figure 3), generating noisier estimations. MSMC estimates are, in general, only slightly affected by variable recombination rate.

Gossmann and colleagues⁴⁰ quantified the heterogeneity of N_e across the genome of ten eukaryotic species (including humans) through the nucleotide diversity of genome sites and accounting for the differences in mutation rate between loci by considering the divergence between species. Thus, they obtained estimates of $N_{e\pi}$ finding a modest but statistically significant variability of N_e for most species. Gossmann and colleagues⁴⁰ found that N_e was only positively correlated with recombination rate for *Drosophila* ($r = 0.45$), and negatively correlated with gene density for *Arabidopsis* ($r = -0.11$) and humans ($r = -0.19$). These correlations generally agree with those found for nucleotide diversity in our analysis (Fig. 5a), i.e., r

= 0.51 and 0.60 between π and RR , and $r = -0.10$ and 0.001 between π and gene density, for Finnish and Koryaks, respectively. The failure of Gossmann and colleagues⁴⁰ to detect further correlations was attributed to the small variation of N_e observed or to only having considered neutral diversity (synonymous changes). In another analysis of genome N_e heterogeneity, Jiménez-Mena and colleagues⁴¹ also found significant variation in N_e across the genome of cattle populations using the temporal N_e estimation method, but this variation did not correlate with the recombination rate, the density of genes, or the presence of loci under artificial selection. This negative result was attributed to the assumption of large genomic windows in order to have a large enough number of markers in each of them, or to the fact that the temporal method of estimation of N_e was only based on a single-generation interval.⁴¹⁻⁴²

The correlations found between the different parameters analyzed with genomic data follow the expected trends (Fig. S2) and agree with previous estimations. For example, analyzing 100 kb windows of a Danish population, Lohmueller and colleagues⁴⁹ found significant correlations between the recombination rate and the number of SNPs in the windows ($r = 0.20$), the SNP diversity ($r = 0.11$) and SNP MAF ($r = 0.062$). These correlations are comparable with those found in our study between RR and genome diversity parameters (Fig. S2).

The large reduction in $N_{e\pi}$ with respect to N_{eLD} in Figure 1 for low recombination rates confirms that linked selection can reduce very substantially neutral genetic diversity. This result may have an implication on Lewontin's⁵⁰ paradox, i.e. the observation that differences in genetic diversities at neutral sites for most eukaryotes and prokaryotes species fall within a range of an order of magnitude,⁵¹⁻⁵³ whereas

differences in the average census sizes between species differ in more than three orders of magnitude. This has been explained by alternative hypotheses. One of them suggests that species with large population sizes would generally have a high variation in reproductive success and would suffer from a higher impact from demographic changes.⁵⁴ Another explanation is the possible constrain in genomic variation imposed by natural selection via fixation of beneficial mutations and elimination of deleterious ones in linked neutral variation.⁵⁵ Theoretical developments showing that there is a nonlinear relationship between effective population size and census size, may contribute to explain this paradox.²⁷ In this sense, because N_{eLD} is not affected by selection, it would be expected to be closer to population size N than to N_e from nucleotide diversity.

In summary, our results show that the estimates of historical effective size obtained from linkage disequilibrium between pairs of SNPs are not altered by selection, either positive or negative, and are not either strongly affected by the heterogeneity in recombination rate across the genome. Therefore, linkage disequilibrium N_e reflects the true demographic changes in population size over generations. In contrast, other methods based on mutation and recombination, from which recent estimates of human demography have been obtained, can be strongly affected by selection.

Methods

Computer simulations

Genomic data of populations under different demographic and evolutionary scenarios were simulated with the software SLiM 3.⁵⁶ Random mating populations of constant or changing size (ranging between $N = 100$ and 10,000) were run for up to 10,000

generations. Different demographic scenarios (constant population size, bottlenecking, exponential growth, etc.) were assumed. Genome sequences with a length of 100 Mb were considered where mutations occur at a rate between 10^{-7} and 10^{-9} mutations per nucleotide and generation, depending on the demographic scenario and simulation, with different recombination rates (ranging from 0.01 to 5 cM per Mb). For each locus, values of fitness of 1, $1 + s/2$, and $1 + s$ were considered for the wild-type homozygote, the heterozygote, and the mutant homozygote, respectively. Fitness of an individual was assumed to be multiplicative across loci. Three mutation models were assumed. (1) A neutral model for all mutations ($s = 0$). (2) Background selection (BS), where 90% of mutations are neutral and 10% are deleterious with selection coefficient $s = -0.02$. (3) Selective Sweeps (SS), where 99% of mutations are neutral and 1% are assumed to be advantageous with effect $s = 0.02$. To investigate the heterogeneity of the genome in recombination rates, the simulated sequence was divided in 70 regions of equal length and the particular rate of recombination for each region was randomly chosen from the distribution of recombination rates observed in analogous genome windows of the human genome (Suppl. Figure S1). Each simulation was run for up to 100 replicates.

Analysis of human genomes

Data comes from the genomic sequencing of samples from two human populations: 99 individuals from a Finnish population,⁴⁶ with a total of about 9.4 million SNPs, and 16 individuals from a Koryaks' population,⁴⁷ with about 4.6 million SNPs. The Koryaks population data coordinates, in genome version hg18, were converted to hg19 using liftOver UCSC tool.⁵⁷ In the process, 65.088 variants were not found and were excluded from the analyses. However, there is a large number of SNPs available in both populations, allowing the study of relatively small regions of the genome. Only

autosomal chromosomes were taken into account. Genomic data was divided in 2 cM regions in which local N_e and other genetic variables were estimated to investigate the correlations between one another. For an accurate estimation of linkage disequilibrium N_e , only regions with more than 250,000 pairs of SNPs were considered. In addition, telomeric regions shorter than 2 cM and regions with very noisy N_e outlier estimates ($> 30,000$) or negative ones were also excluded. Thus, following these criteria, 655 regions were excluded from the analyses of Finnish and Koryaks data, respectively, and the final number of regions analysed was 1,502 and 1,154, respectively.

Estimation of N_e

The software GONE was used to obtain historical estimates of linkage disequilibrium N_e (N_{eLD}) using all pairs of SNPs available from simulation data at distances between $c = 0.5$ and 0.001 Morgans (M) in samples of 100 individuals. The software MSMC¹⁸ and Relate¹⁹ were applied to the same data, except that only samples of four randomly sampled individuals were analysed with MSMC because of computation time restrictions with this software. MSMC version 2 (downloaded in December 2019) was used with the “fixedRecombination” flag, as recommended by the user’s guide. Since the software needs several chromosomes to be run, sets of 10 replicates were run and considered as chromosomes. Relate (downloaded in December 2019) was run without monomorphic SNPs, providing the mutation rate of the simulations, the number of haplotypes of the sample, a seed, 300 bins and a threshold value from 50 to 30 depending on the simulation scenario. It was run for each simulation replicate and the results averaged over replicates.

For the analysis of specific genomic regions with real data, N_{eLD} estimates were directly obtained with equations S4b and S5 of the Supplemental Material of Santiago et al.¹³, which applies to the scenario of constant population size of diploid populations when the genetic phase of genotypic data is unknown. In this case, because SNPs in the regions are necessarily at relatively low genetic distances, pairs of SNPs at distances ranging between 1/50 and 1/100 M were considered in order to obtain at least 250,000 pairs per genomic region. The software Relate and MSMC were not used in these analyses, as they are assumed to apply only to historical estimates of N_e .

Estimation of other genomic variables

Recombination rates (RR ; in cM/Mb) between all pairs of consecutive SNPs for each of the genomic regions were obtained from the human genetic map³⁴ and averaged for each region. Estimates of the background selection statistic (B)⁵⁸ were obtained for each site and averaged for each genomic region. A reduction in neutral diversity at a given genomic region is a function of the intensity of purifying selection and the rate of recombination, as the impact of selection on reducing diversity is higher in tight linkage regions.^{22,26} The B statistic measures the impact of background selection on nucleotide diversity. Thus, it fluctuates between one (no background selection affecting diversity) and zero (almost complete exhaustion of diversity as a result of background selection), with an average for the human autosomal genome of about 0.74 - 0.81.⁵⁸

Average nucleotide diversity (π), polymorphism (P) and minor allele frequency (MAF) of SNPs were calculated for each genomic region. The number of Loss of Function (LoF) and missense variants in each genomic region were also obtained using data from the 0.3.1 version of the ExAC browser,⁵⁹ downloaded on 14th October 2019.

Only high confidence variants were taken into account. The gene density of each genomic region was obtained using the RefSeq database.⁶⁰ When a gene spanned over different regions, we considered it to be in the region where its middle point was located. Only genes with a straightforward chromosome code were used (e.g. NC_000001.10 corresponding with chromosome 1).

Acknowledgements

We thank Humberto Quesada for useful comments on the topic. This work was funded by Agencia Estatal de Investigación (AEI) (PID2020-114426GB-C21), Xunta de Galicia (GRC, ED431C 2020-05) and Centro singular de investigación de Galicia accreditation 2019-2022, and the European Union (European Regional Development Fund - ERDF), Fondos Feder “Unha maneira de facer Europa”. I.N. was funded by a predoctoral (FPU18/04642) grant from Ministerio de Ciencia, Innovación y Universidades (Spain).

Declaration of interests

The authors declare no competing interests.

Web resources

387 LiftOver UCSC tool, <http://genome.ucsc.edu>

388 RefSeq database.

389 https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh37_latest/refseq_identifiers/GRCh37_latest_genomic.gff.gz

390 ifiers/GRCh37_latest_genomic.gff.gz

391 **Data availability**

392 The authors affirm that all data necessary for confirming the conclusions of the article
393 are present within the article, figures, and tables. The code generated during this study is
394 available at github address [...].

395 **Author Contributions**

396 **Conceptualization:** Enrique Santiago, Armando Caballero.

397 **Formal analysis:** Irene Novo.

398 **Supervision:** Enrique Santiago, Armando Caballero.

399 **Writing – review & editing:** Armando Caballero, Irene Novo, Enrique Santiago.

400

401 **References**

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

403 2. Charlesworth, B. (2009). Effective population size and patterns of molecular
404 evolution and variation. *Nat. Rev. Genet.* *10*, 195–205.

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

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

- 408 5. Luikart, G., Ryman, N., Tallmon, D.A., Schwartz, M.K., and Allendorf, F.W. (2010).
409 Estimation of census and effective population sizes: The increasing usefulness of
410 DNA-based approaches. *Conserv. Genet.* *11*, 355–373.
- 411 6. Gilbert, K.J., and Whitlock, M.C. (2015). Evaluating methods for estimating local
412 effective population size with and without migration. *Evolution* *69*, 2154–2166.
- 413 7. Waples, R.S., and Do, C. (2010). Linkage disequilibrium estimates of contemporary
414 N_e using highly variable genetic markers: A largely untapped resource for applied
415 conservation and evolution. *Evol. Appl.* *3*, 244–262.
- 416 8. Waples, R.S., and England, P.R. (2011). Estimating contemporary effective
417 population size on the basis of linkage disequilibrium in the face of migration.
418 *Genetics* *189*, 633–644.
- 419 9. Hill, W. G. (1981). Estimation of effective population size from data on linkage
420 disequilibrium. *Genet. Res.* *38*, 209–216.
- 421 10. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A.,
422 and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and
423 Translation. *Am. J. Hum. Genet.* *101*, 5–22.
- 424 11. Hayes, B.J., Visscher, P.M., McPartlan, H.C., and Goddard, M.E. (2003). Novel
425 multilocus measure of linkage disequilibrium to estimate past effective population
426 size. *Genome Res.* *13*, 635–643.
- 427 12. Tenesa, A., Navarro, P., Hayes, B.J., Duffy, D.L., Clarke, G.M., Goddard, M.E., and
428 Visscher, P.M. (2007). Recent human effective population size estimated from
429 linkage disequilibrium. *Genome Res.* *17*, 520–526.

- 430 13. Santiago, E., Novo, I., Pardiñas, A.F., Saura, M., Wang, J., and Caballero, A.
431 (2020). Recent Demographic History Inferred by High-Resolution Analysis of
432 Linkage Disequilibrium. *Mol. Biol. Evol.* 37, 3642–3653.
- 433 14. Mörseburg, A., Pagani, L., Ricaut, F.X., Yngvadottir, B., Harney, E., Castillo, C.,
434 Hoogervorst, T., Antao, T., Kusuma, P., Brucato, N., et al. (2016). Multi-layered
435 population structure in Island Southeast Asians. *Eur. J. Hum. Genet.* 24, 1605–
436 1611.
- 437 15. Qanbari, S., Pimentel, E.C.G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A.R.,
438 Simianer, H. (2010). The pattern of linkage disequilibrium in German Holstein
439 cattle. *Anim. Genet.* 41, 346–356.
- 440 16. Saura, M., Tenesa, A., Woolliams, J.A., Fernández, A., and Villanueva, B. (2015).
441 Evaluation of the linkage-disequilibrium method for the estimation of effective
442 population size when generations overlap: An empirical case. *BMC Genomics* 16,
443 922.
- 444 17. Hollenbeck, C.M., Portnoy, D.S., and Gold, J.R. (2016). A method for detecting
445 recent changes in contemporary effective population size from linkage
446 disequilibrium at linked and unlinked loci. *Heredity* 117, 207–216.
- 447 18. Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation
448 history from multiple genome sequences. *Nat. Genet.* 46, 919–925.
- 449 19. Speidel, L., Forest, M., Shi, S., and Myers, S.R. (2019). A method for genome-wide
450 genealogy estimation for thousands of samples. *Nat. Genet.* 51, 1321–1329.
- 451 20. Robertson, A. (1961). Inbreeding in artificial selection programmes. *Genet. Res.* 89,
452 275–280.
- 453 21. Santiago, E., and Caballero, A. (1995). Effective Size of Populations Under
454 Selection. *Genetics* 139, 1013–1030.

- 455 22. Charlesworth, B., Morgan, M.T., and Charlesworth, D. (1993). The effect of
456 deleterious mutations on neutral molecular variation. *Genetics* 134, 1289–1303.
- 457 23. Hudson, R.R., and Kaplan, N.L. (1995). Deleterious background selection with
458 recombination. *Genetics* 141, 1605–1617.
- 459 24. Nordborg, M., Charlesworth, B. and Charlesworth, D. (1996). The effect of
460 recombination on background selection. *Genet. Res.* 67, 159–174.
- 461 25. Nicolaisen, L.E., and Desai, M. M. (2013). Distortions in genealogies due to
462 purifying selection and recombination. *Genetics* 195, 221–230.
- 463 26. Santiago, E., and Caballero, A. (1998). Effective Size and Polymorphism of Linked
464 Neutral Loci in Populations Under Directional Selection. *Genetics* 149, 2105–2117.
- 465 27. Santiago, E., and Caballero, A. (2016) Joint prediction of the effective population
466 size and the rate of fixation of deleterious mutations. *Genetics* 204, 1267–1279.
- 467 28. Campos, J.L., Halligan, D.L., Haddrill, P.R., and Charlesworth, B. (2014). The
468 relation between recombination rate and patterns of molecular evolution and
469 variation in *drosophila melanogaster*. *Mol. Biol. Evol.* 31, 1010–1028.
- 470 29. Lohmueller, K.E. (2014). The distribution of deleterious genetic variation in human
471 populations. *Curr. Opin. Genet. Dev.* 29, 139–146.
- 472 30. Hill, W.G., and Robertson, A. (1966). The effect of linkage on limits to artificial
473 selection. *Genet. Res.* 8, 269–294.
- 474 31. Begun, D.J., and Aquadro, C.F. (1992). Levels of naturally occurring DNA
475 polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356,
476 519–520.
- 477 32. Nachman, M.W. (1997) Patterns of DNA variability at X-linked loci in *Mus*
478 *domesticus*. *Genetics* 147, 1303–1316.

479 33. Hellmann I., Prüfer, K., Ji, H., Zody, M.C., Pääbo, S., and Ptak, S.E. (2005). Why
480 do human diversity levels vary at a megabase scale? *Genome Res.* 15, 1222–1231.

481 34. Myers, S., Bottolo, L., Freeman, C., McVean, G., Donnelly, P. (2005). Genetics: A
482 fine-scale map of recombination rates and hotspots across the human genome.
483 *Science* 310, 321–324.

484 35. Charlesworth, B., and Charlesworth, D. (2010). *Elements of Evolutionary Genetics*
485 (Roberts and Company Publishers).

486 36. Maynard-Smith, J., and Haigh J. (1974). The hitchhiking effect of a favourable
487 gene. *Genet. Res.* 23, 23–35.

488 37. Kaplan, N.L., Hudson, R.R., and Langley, C.H. (1989) The “hitchhiking effect”
489 revisited. *Genetics* 123, 887–899.

490 38. Przeworski, M. (2002). The signature of positive selection at randomly chosen loci.
491 *Genetics* 160, 1179–1189.

492 39. Zeng K., Jackson, B.C., and Barton, H.J. (2018). Methods for Estimating
493 Demography and Detecting Between-Locus Differences in the Effective Population
494 Size and Mutation Rate. *Mol. Biol. Evol.* 36, 423–433.

495 40. Gossman, T.I., Woolfit, M., and Eyre-Walker, A. (2011). Quantifying the variation
496 in the effective population size within a genome. *Genetics* 189, 1389–1402.

497 41. Jiménez-Mena, B., Hospital, F., and Bataillon, T. (2016a). Heterogeneity in
498 effective population size and its implications in conservation genetics and animal
499 breeding. *Conserv. Genet. Resour.* 8, 35–41.

500 42. Jiménez-Mena, B., Tataru, P., Brøndum, R.F., Sahana, G., Guldbrandtsen, B., and
501 Bataillon, T. (2016b). One size fits all? Direct evidence for the heterogeneity of
502 genetic drift throughout the genome. *Biol. Letters* 12.

503 43. Thomson, G. (1977). The effect of a selected locus on linked neutral loci. *Genetics*
504 85, 753–788.

505 44. Gillespie, J.H. (1997). Junk ain't what junk does: neutral alleles in a selected
506 context. *Gene* 205, 291–299.

507 45. Stephan, W., Song, Y.S., and Langley, C.H. (2006). The hitchhiking effect on
508 linkage disequilibrium between linked neutral loci. *Genetics* 172, 2647–2663.

509 46. The 1000 Genomes Project Consortium. (2010). An integrated map of genetic
510 variation from 1,092 human genomes. *Nature* 491, 56–65.

511 47. Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J.S., Albrechtsen, A., Moltke, I.,
512 Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., et al. (2010). Ancient human
513 genome sequence of an extinct Palaeo-Eskimo. *Nature* 463, 757–762.

514 48. Gillespie, J.H. (2000). Genetic drift in an infinite population: The pseudohitchhiking
515 Model. *Genetics* 155, 909–919.

516 49. Lohmueller, K.E., Albrechtsen, A., Li, Y., Kim, S.Y., Korneliussen, T.,
517 Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Feder, A.F., Grarup, N., et al.
518 (2011). Natural Selection Affects Multiple Aspects of Genetic Variation at
519 Putatively Neutral Sites across the Human Genome. *PLoS Genet.* 7, e1002326.

520 50. Lewontin, R.C. (1974). *The Genetic Basis of Evolutionary Change* (Columbia
521 University Press).

522 51. Lynch, M., and Conery, J. (2003). The origins of genome complexity. *Science* 302,
523 1402–1404.

524 52. Leffler, E.M., Bullaughey, K., Matute, D.R., Meyer, W.K., Ségurel, L., Venkat, A.,
525 Andolfatto, P., Przeworski, M. (2012). Revisiting an old riddle: What determines
526 genetic diversity levels within species? *PLoS Biol.* 10, e1001388.

527 53. Galtier, N., and Rouselle, M. (2020). How much does N_e vary among species?
528 *Genetics* 216, 559–572.

529 54. Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A.,
530 Chiari, Y., Dernat, R., Duret, L., Faivre, N., et al. (2014). Comparative population
531 genomics in animals uncovers the determinants of genetic diversity. *Nature* 515,
532 261–263.

533 55. Corbett-Detig, R.B., Hartl, D.L., and Sackton, T.B. (2015). Natural selection
534 constrains neutral diversity across a wide range of species. *PLoS Biol.* 13,
535 e1002112.

536 56. Haller, B.C., and Messer, P.W. (2019). SLiM 3: Forward Genetic Simulations
537 Beyond the Wright-Fisher Model. *Mol. Biol. Evol.* 36, 632–637.

538 57. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H.,
539 Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome
540 Browser Database: update 2006. *Nucleic Acids Res.* 34, D590–D598.

541 58. McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic
542 signatures of natural selection in hominid evolution. *PLoS Genet.* 5, e1000471.

543 59. Karczewski, K.J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D.M.,
544 Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K.E., Cummings, B.B., et al.
545 (2017). The ExAC browser: Displaying reference data information from over 60
546 000 exomes. *Nucleic Acids Res.* 45, D840–D845.

547 60. O'Leary N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R.,
548 Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference
549 sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and
550 functional annotation. *Nucleic Acids Res.* 44, D733–D745.

552 **Figure titles**

553 **Figure 1.** Estimates of effective population size from linkage disequilibrium (N_{eLD} ,
554 sample size of $n = 100$ individuals), Relate ($N_{eRelate}$, $n = 100$), MSMC (N_{eMSMC} , $n = 4$),
555 and from nucleotide diversity (π), this latter calculated as $N_{e\pi} = \pi/(4\mu)$, where μ is the
556 nucleotide mutation rate, for scenarios with different recombination rates (RR in
557 cM/Mb) uniform across the genome. Simulations assume a constant population size of
558 $N = 1,000$ individuals under neutrality, background selection and selective sweeps, with
559 constant mutation rate $\mu = 10^{-8}$ per base per generation. Estimates were obtained
560 including windows of recombination rate between pairs of SNPs ranging from $c =$
561 0.0025 to 0.0250 for N_{eLD} and averaging historical estimates of N_e between generations
562 150 to 350 for $N_{eRelate}$ and N_{eMSMC} . Error bars represent one standard error above and
563 below the mean of the simulation replicates. N_{eLD} does not have errors as single
564 estimates were obtained using all pairs of SNPs from all replicates. All simulations were
565 run for up to 100 replicates.

566 **Figure 2.** Estimates of historical N_e with GONE (sample size $n = 100$ individuals),
567 Relate ($n = 100$) and MSMC ($n = 4$) from the present generation (generation 0) back to
568 400 or 1,000 generations in the past. 100 replicates were run of simulations with
569 constant population size (N) under neutrality (grey), background selection (BS, red) and
570 selective sweeps (SS, green), with constant (1 cM/Mb) or variable recombination rates
571 (RR), and constant mutation rate $\mu = 10^{-8}$ or 10^{-9} mutations per base per generation for N
572 $= 1,000$ or $N = 10,000$, respectively. The true simulated N is shown in black.

573 **Figure 3.** Estimates of historical N_e with GONE (sample size $n = 100$ individuals),
574 Relate ($n = 100$) and MSMC ($n = 4$) from the present generation (generation 0) back to

300 generations in the past. 100 replicates were run of simulations with constant population size (N) under neutrality (grey), background selection (BS, red) and selective sweeps (SS, green), with constant (1 cM/Mb) or variable recombination rate (RR), and constant mutation rate $\mu = 10^{-8}$ or 10^{-9} mutations per base per generation for $N = 1,000$ or $N = 10,000$, respectively. The true simulated N is shown in black.

Figure 4. Estimates of historical N_e with GONE (sample size $n = 100$ individuals), Relate ($n = 100$) and MSMC ($n = 4$) from the present generation (generation 0) back to 600 generations in the past. 100 replicates were run of simulations with constant population size (N) under neutrality (grey), background selection (BS, red) and selective sweeps (SS, green), with constant (1 cM/Mb) or variable recombination rate (RR), and constant mutation rate $\mu = 10^{-8}$ or 10^{-9} mutations per base per generation for $N = 1,000$ or $N = 10,000$, respectively. The true simulated N is shown in black.

Figure 5. Spearman's correlation coefficient (r) between nucleotide diversity (π , panel a) or estimates of linkage disequilibrium effective population size (N_{eLD} ; panel b) and different diversity variables, estimated within genomic regions. RR : recombination rate; B : B statistic; P : nucleotide polymorphism; MAF: Minor Allele Frequency; LoF : number of Loss of Function variants; *missense*: number of missense variants; *gene dens*: gene density. Estimates are based on samples of $n = 99$ for Finnish and $n = 16$ for Koryaks populations. P-values: * < 0.05 , *** < 0.001 .

Neutral

Background selection

Selective sweep

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5