PLOS GENETICS

RESEARCH ARTICLE

The estimates of effective population size based on linkage disequilibrium are virtually unaffected by natural selection

Irene Novo[1]*, **Enrique Santiago**[2], **Armando Caballero**[1]

1 Centro de Investigación Mariña, Universidade de Vigo, Facultade de Bioloxía, Vigo, Spain,
2 Departamento de Biología Funcional, Facultad de Biología, Universidad de Oviedo, Oviedo, Spain

* irene.novo.gimenez@uvigo.es

OPEN ACCESS

Citation: Novo I, Santiago E, Caballero A (2022) The estimates of effective population size based on linkage disequilibrium are virtually unaffected by natural selection. PLoS Genet 18(1): e1009764. https://doi.org/10.1371/journal.pgen.1009764

Editor: Bret Payseur, University of Wisconsin–Madison, UNITED STATES

Received: August 5, 2021

Accepted: December 21, 2021

Published: January 25, 2022

Data Availability Statement: The code generated during this study is available at github address https://github.com/irene-novo-g/Selection-and-Ne.

Funding: We acknowledge financial support from Grant PID2020-114426GB-C21 funded by MCIN/AEI/10.13039/501100011033 (AC), Xunta de Galicia (GRC, ED431C 2020-05) and Centro singular de investigación de Galicia accreditation 2019-2022 (AC), and the European Union (European Regional Development Fund - ERDF), Fondos Feder "Unha maneira de facer Europa" (AC). IN was funded by a predoctoral (FPU18/

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. Directional 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 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, which 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, or they do it very weakly, with recombination rate, nucleotide diversity, proportion of polymorphic sites, background selection statistic, minor allele frequency of SNPs, loss of function and missense variants and gene density. This suggests that N_{eLD} measures mainly reflect demographic changes in population size across generations.

Author summary

The inference of the demographic history of populations is of great relevance in evolutionary biology. This inference can be made from genomic data using coalescence methods or linkage disequilibrium methods. However, the assessment of these methods is usually made assuming neutrality (absence of selection). Here we show by computer simulations and analyses of human data that the estimates of historical effective population size obtained from linkage disequilibrium between markers are virtually unaffected by natural selection, either positive or negative. In contrast, estimates obtained by coalescence

04642) grant from Ministerio de Ciencia, Innovación y Universidades (Spain). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

mutation-recombination-based methods can be strongly affected by it, which could have important consequences for recent estimations of human demography.

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 population [1]. The effective size of a population depends on its demographic history and structure as well as the selection regime affecting the population [2–4]. 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 both the genetic distance (or recombination rate, c) between marker sites 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 species [10], 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 colleagues [11] 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 (software GONE) [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 MSMC [18] 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). In this situation, N_e can be estimated from the mean and variance of progeny numbers contributed by parents (N_{eVk}). This value depends on many factors, such as the number of breeding males and females, changes in census size across generations, system of mating, overlapping generations, etc. [1–4], and reflects the amount of neutral genetic drift affecting the whole genome. For a close population with stable demography and breeding system, N_{eVk} becomes constant over generations. Under selection, however, N_e (which always refers to neutral loci) gets reduced over generations because the cumulative effect of selection on genetic drift, to reach an asymptotic value which is lower than N_{eVk} [2,20,21]. This reduction occurs even with free recombination and it can be large under artificial selection, as shown initially by Robertson [20]. Under natural selection, the effect is not expected to be so large as for artificial selection except when there is tight linkage [2,22–27]. Thus, under selection and linkage, N_e is lower than N_{eVk}, so that the genetic drift ascribed to neutral genes is higher than that quantified by N_{eVk}.

Natural populations are predicted to encode many deleterious variants [28,29] that can affect the outcome of N_e. 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 regions [31–34] and because of the differential impact of natural selection on them [35]. For example, selective sweeps of favorable mutations are expected to hitch-hike close-by neutral SNPs producing sharp decreases in diversity in linked regions [36–38]. Thus, because the reduction in N_e depends on the recombination rate and the intensity of selection, and these are variable across the genome, the amount of genetic drift for neutral genes is not expected to be the same in different genomic regions, what is called genomic heterogeneity for N_e [2,39–41]. The distribution of genetic variability, both within and between genomes, is affected by the impact of selection on genetic drift and, in particular, nucleotide diversity can be strongly reduced by selection when linkage is tight. Ignoring the heterogeneity in N_e may lead to biased estimates of past demography [42].

It has been shown that selective sweeps of favourable mutations generate LD between close-by neutral loci [43], although this LD increase is transient [38–44] and may be small [45]. In this paper we assess the impact of selection on the estimates of historical N_e obtained by GONE [13], which is based on linkage disequilibrium between SNP markers, in comparison with other coalescence mutation-based methods, MSMC [18] 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. We also consider a model of heterozygote advantage for fitness and another with partial self-fertilization, assuming or not selection. 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 Finnish [46] and Koryaks [47] 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 are virtually unaffected by selection.

Results

Effect of selection and recombination on the estimation of N_e

Fig 1 shows the joint effect of recombination and selection on the estimates of N_e, for a population of invariable census size of N = 1,000 individuals assuming different recombination rates (RR) per Mb across the genome. Under a neutral scenario, linkage disequilibrium estimates by GONE (N_{eLD}) provide virtually unbiased estimates of the expected effective population size from the variance of progeny numbers (N_{eVk}) for all recombination rates. In the random mating scenario, $N_{eVk} = N$ = 1,000, the number of breeding individuals. For partial self-fertilization with a proportion 0.5 of selfed mating, $N_{eVk} = 3N/4$ = 750. The same results 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 accurate estimates of N_{eVk} except for the most extreme cases of recombination rate 5 or 0.01 cM per Mb.

Under background selection and selective sweep models in random mating populations, N_{eLD} estimates give basically the same results as for the neutral model (with some minor deviations for intermediate recombination rates), indicating that these estimates are very little or not affected by selection, either negative or positive (Fig 1). As expected, estimates of $N_{e\pi}$ diverge from N_{eVk}, with decreasing values as recombination rate decreases. Relate and MSMC

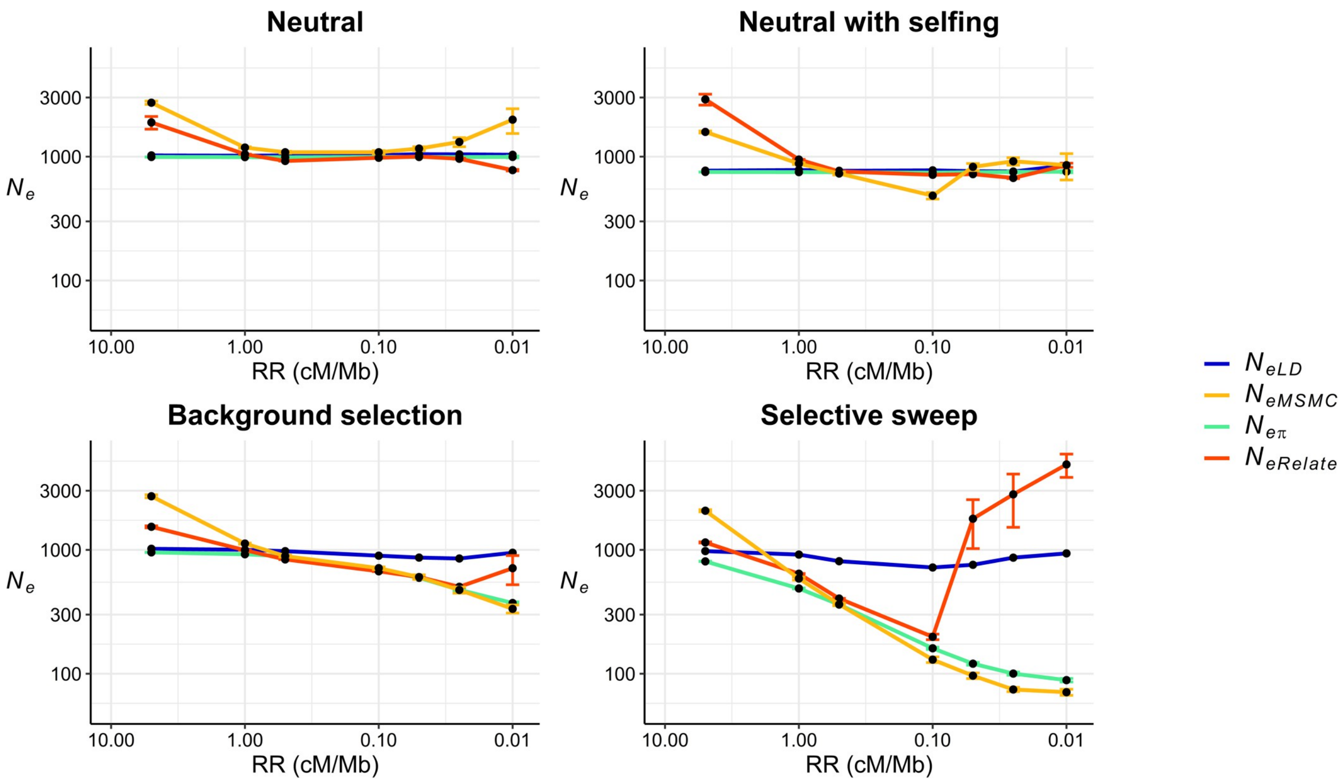

Fig 1. Estimates of effective population size from linkage disequilibrium (N_{eLD}, sample size of n = 100 individuals), Relate ($N_{eRelate}$, n = 100), MSMC (N_{eMSMC}, n = 4), and from nucleotide diversity (π), this latter calculated as $N_{e\pi} = \pi/(4\mu)$, where μ is the nucleotide mutation rate, for scenarios with different recombination rates (*RR* in cM/Mb) uniform across the genome. Simulations assume a fixed population size of N = 1,000 individuals under neutrality (random mating or partial self-fertilization with a frequency of 50% selfed progeny), background selection and selective sweeps (both for random mating populations), with constant mutation rate $\mu = 10^{-8}$ per base per generation. Estimates were obtained including windows of recombination rate between pairs of SNPs ranging from c = 0.0025 to 0.0250 for N_{eLD} and averaging historical estimates of N_e between generations 150 to 350 for $N_{eRelate}$ and N_{eMSMC}. Error bars represent one standard error above and below the mean of the simulation replicates. N_{eLD} estimates were obtained pooling all replicates to speed up computation. All simulations were run for up to 100 replicates.

https://doi.org/10.1371/journal.pgen.1009764.g001

estimates show a pattern similar to that of $N_{e\pi}$ but $N_{eRelate}$ estimates increase sharply for tight linkage scenarios. Note that the lowest recombination rate assumed (*RR* = 0.01) implies a genetic map size of only 1 cM for a genome of 100Mb, so we are evaluating very extreme scenarios of linkage. For the partial self-fertilization model, the results under background selection and selective sweeps are similar to those with random mating (S1 Fig), although N_{eLD} appears to increase with the hitch-hiking model for extreme cases of tight linkage.

Finally, a model of overdominance for fitness under random mating (S2 Fig) shows that N_{eLD} is almost unaffected by selection for recombination rates down to 0.05 cM per Mb (a genome size of 5 cM). Below this threshold, there is a sharp reduction in N_{eLD} caused by the appearance of linkage blocks of mutations in heterozygous state. Nucleotide diversity increases for low recombination rates, as expected with this model, and the same behaviour is observed for the estimates from $N_{eRelate}$ and N_{eMSMC}.

Simulation results for historical N_e estimates

Fig 2 shows estimates of historical N_e assuming an invariable population census size (N = 1,000 or 10,000), either with a fixed or a variable recombination rate. Estimates of historical N_{eLD} reflect the effective population size in the absence of selection ($N_{eVk} = N$) regardless of selection and the variability in genomic recombination rates. Estimates from $N_{eRelate}$ and

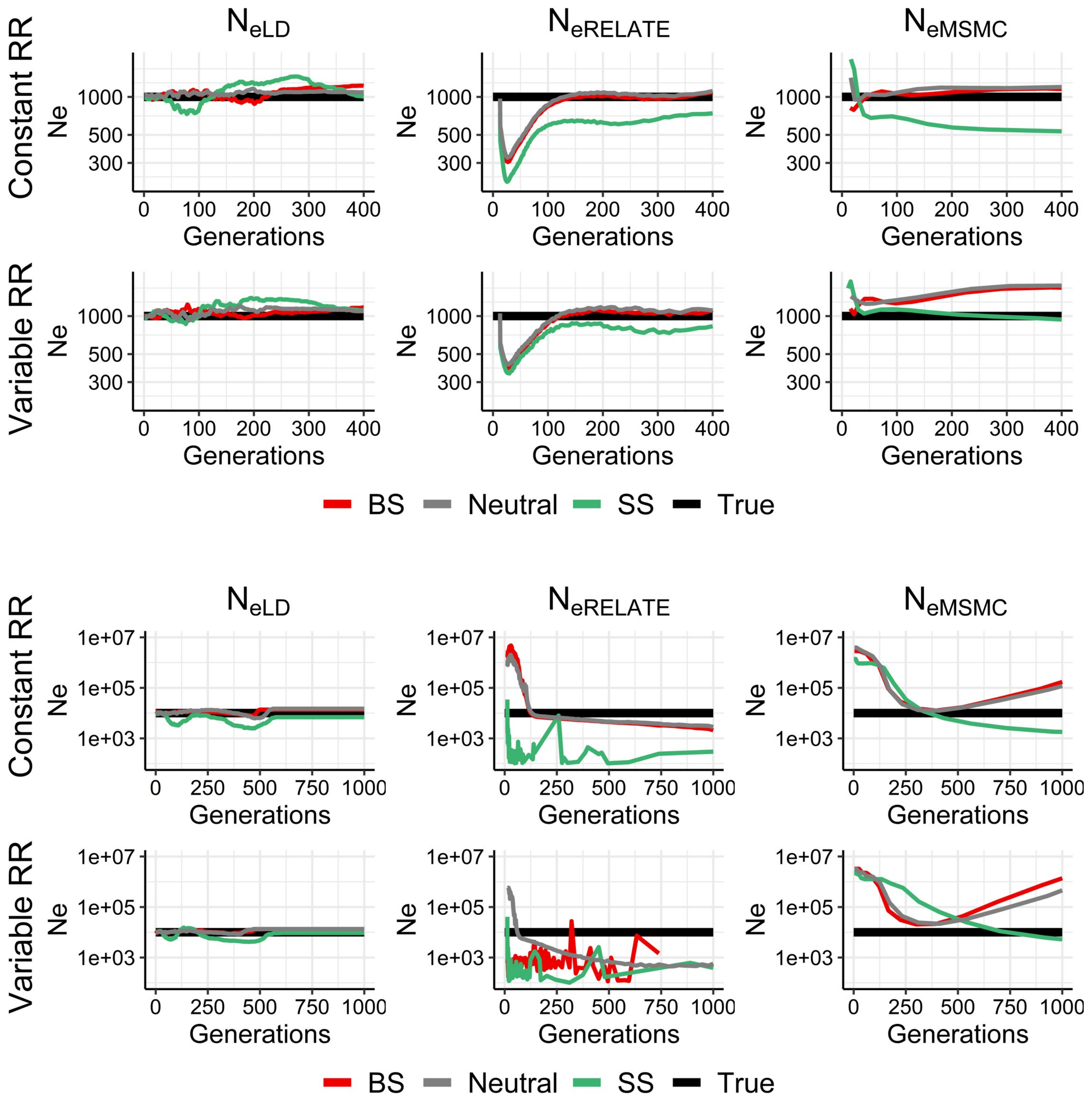

Fig 2. Estimates of historical effective population size from linkage disequilibrium (N_{eLD}, sample size n = 100 individuals), Relate ($N_{eRelate}$, n = 100) and MSMC (N_{eMSMC}, n = 4) from the present generation (generation 0) back to 400 or 1,000 generations in the past. 100 replicates were run of simulations with constant population size (N) under random mating and neutrality (grey), background selection (BS, red) or selective sweeps (SS, green), with constant (1 cM/Mb) or variable recombination rates (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 effective size from variance of family size ($N_{eVk} = N$) is shown in black.

https://doi.org/10.1371/journal.pgen.1009764.g002

N_{eMSMC} can give unbiased values of N_{eVk} 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 N_{eVk}, particularly under a selective sweep model, where N_{eVk} 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 impact on the estimates of N_{eVk} with respect to a fixed recombination rate model.

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_{eVk} value except for an overestimation in some cases. N_{eMSMC} estimates are also generally accurate, although they may show some over or underestimations.

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 and appear to be affected by selective sweeps, while N_{eLD} is generally more accurate and it is not affected by selection. The reason why N_{eLD} performs better to detect recent rather than ancient changes (cf. Figs 3 and 4) is that the linkage signals induced by drift are lost by recombination at a rate c per generation, so that changes occurred in the ancient times are less likely to persist. A variable recombination rate does not substantially affect the estimates.

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, the proportion of polymorphic sites 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}.

The correlations among all genetic variables studied are shown in the Supplemental material (S3 Fig), and follow the expected trends. For example, recombination rate RR, the B statistic, the proportion of polymorphic sites 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, median and standard deviation of the estimates of N_{eLD} across regions were 11,600, 7,202, and 13,638 for the Finnish population, respectively, and 5,866, 985, and 16,063 for the Koryaks population, respectively. Thus, the standard deviation of the regional estimates of N_{eLD} relative to a mean of one, for comparison, were 1.18 for the Finnish population and 2.74 for the Koryaks population. The distribution of N_{eLD} values across genomic regions for both populations is shown in S4 Fig.

The correlation between the estimates of N_{eLD} and other diversity parameters for different genomic regions are shown in Fig 5B. The correlations did not follow the trends observed for nucleotide diversity. There was no significant correlation between N_{eLD} and the rest of variables except for a small significant correlation between N_{eLD} and RR for the Finnish population and another between N_{eLD} and P for the Koryaks population.

Discussion

Our results show that the estimates of the effective population size obtained from linkage disequilibrium between pairs of SNPs (N_{eLD}) are virtually unaffected by either positive or negative

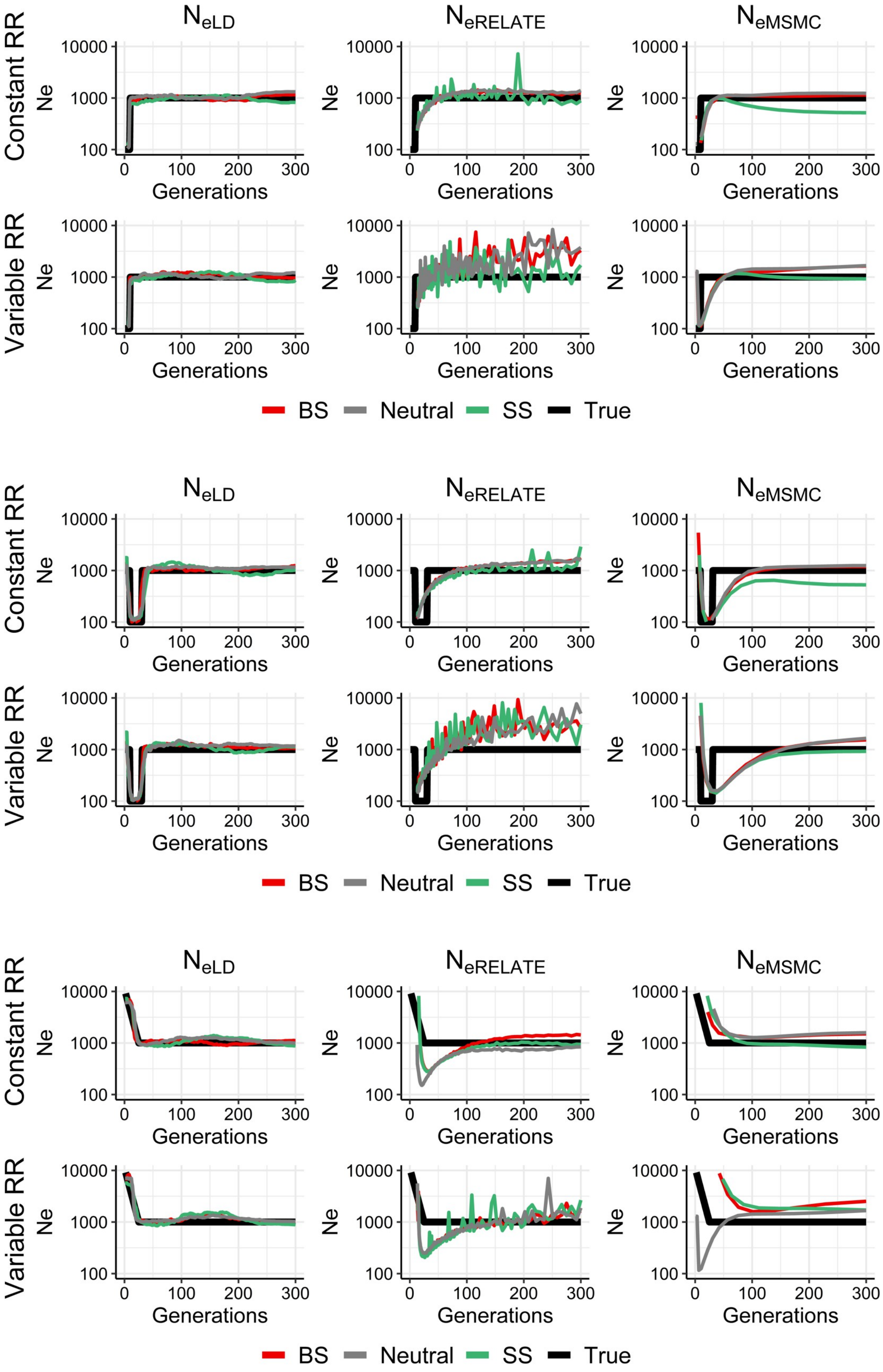

Fig 3. Estimates of historical effective population size from linkage disequilibrium (N_{eLD}, sample size n = 100 individuals), Relate ($N_{eRelate}$, n = 100) and MSMC (N_{eMSMC}, n = 4) from the present generation (generation 0) back to

300 generations in the past. 100 replicates were run of simulations with constant population size (N) under random mating and neutrality (grey), background selection (BS, red) or 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 effective size from variance of family size ($N_{eVk} = N$) is shown in black.

https://doi.org/10.1371/journal.pgen.1009764.g003

selection, thus providing estimates of the effective size from the variance of family sizes (N_{eVk}). This has been deduced from simulation data assuming fixed 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 generally uncorrelated or weakly correlated 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.

Our interest here was to quantify the impact of selection on the estimates of historical $N_{e.}$ Thus, we assumed a relatively large sample size for the analyses (100 individuals) in order to obtain reasonably accurate estimates. Estimates with lower sample sizes generally would produce noisier estimates (as seen before for GONE estimations [13]) and less accurate inferences about the evolutionary history of the population as a whole [48,49]. The MSMC method could not be applied with more than eight haplotypes for practical reasons, so in that sense it has some disadvantage with respect to the other methods. However, the method worked well in many situations even with this low sample size.

We considered the most characteristic models of natural selection (background selection on deleterious mutations and selective sweeps for advantageous mutations). The observed lack of an impact of selection on the estimates of N_{eLD} occurs for both models, and this was shown both for random mating and partially self-fertilising populations. We also assumed a model of overdominance for fitness. For this model, the nucleotide diversity is increased with tight linkage (S2 Fig), as expected, which contrasts with the empirical evidence showing that nucleotide diversity is generally reduced in regions of low recombination [35]. This reduction can be clearly seen from the human data analysed here (see S5 Fig). The heterozygote advantage for fitness assumed does not affect either the estimates of N_{eLD} unless the genetic length of the genome is so small (less than 5 cM) that linkage blocks of balanced mutations in heterozygote state are created. This artefact drastically increases linkage disequilibrium and reduces N_{eLD} (S2 Fig).

Although positive selection is known to generate linkage disequilibrium between neutral loci close to selective loci [43], this effect is transient and may disappear quickly [38,44]. Stephan and colleagues [44] showed that the increase in linkage disequilibrium between two neutral loci occurs if the recombination rate between the selected locus and the neutral loci is less than $c = 0.1s$, where s is the selection coefficient of the advantageous allele in homozygosis. We performed simulations with an average $s = 0.02$, 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 Fig 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. Thus, in this extreme scenario it is likely that the linkage disequilibrium between many pairs of close-by SNPs can be transitorily affected by selection. However, we found that the estimates of N_{eLD} appear to be basically unaffected by positive selection even with tight linkage. To explain this result, we should take into account that the estimation of N_{eLD} is not based only on the linkage disequilibrium between consecutive or close-by SNPs. It is rather based on the linkage disequilibrium between all pairs of loci across the genome. Recent estimates of N_{eLD} rely more

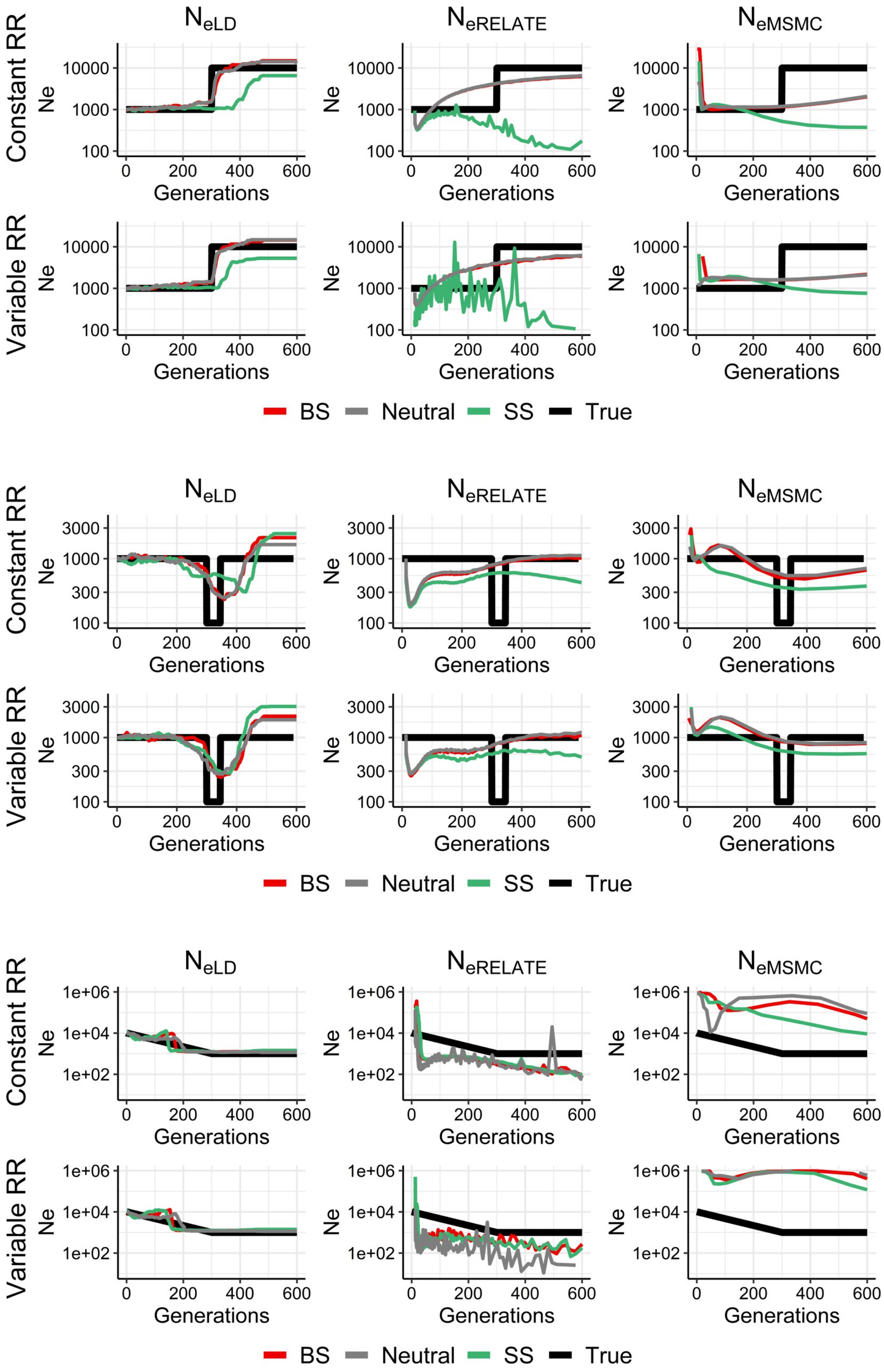

Fig 4. Estimates of historical effective population size from linkage disequilibrium (N_{eLD}, sample size n = 100 individuals), Relate ($N_{eRelate}$, n = 100) and MSMC (N_{eMSMC}, n = 4) from the present generation (generation 0) back to

600 generations in the past. 100 replicates were run of simulations with constant population size (N) under random mating and neutrality (grey), background selection (BS, red) or 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 effective size from variance of family size (N_{eVk}) is shown in black.

https://doi.org/10.1371/journal.pgen.1009764.g004

strongly on pairs of SNPs at large genetic distances, whereas old estimates rely more strongly on SNPs at close genetic distances, and the latter are more affected by selection. However, even so, most pairs of SNPs used in the estimation of N_{eLD} are likely to be far away from selective loci even in the background selection model, where we assumed that only 5% of mutations are deleterious (0.1% in the selective sweep model of advantageous mutations). Therefore, even though a selective locus may have some impact on the linkage disequilibrium of close-by neutral SNPs, the average linkage disequilibrium of all pairs of SNPs is expected to be weakly affected.

In contrast with the above result of a near independence of N_{eLD} from selection, the effective population size obtained from nucleotide diversity ($N_{e\pi}$) is expected to be drastically reduced in regions of low recombination under selection [2,27,50] (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 confirm 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 based on estimating the speed of spread of a particular lineage relative to other competing lineages has been proposed [19]. MSMC estimates generally follow the pattern of $N_{e\pi}$ values except for large recombination rates, for which it provides overestimates of the effective population size in the absence of selection (N_{eVk}; Fig 1). Relate estimates also follow $N_{e\pi}$ values down to recombination rates of 0.1 cM/Mb but, for lower rates, 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 populations [18], and Relate seems to discriminate before 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 (Figs 2–4), particularly when GONE and MSMC are used. This is in accordance with the analyses made by Schiffels and Durbin (their S4 Fig) [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 Relate estimates, recombination rate heterogeneity seems to affect the estimates of recent demographic changes (Fig 3), generating noisier estimations.

Gossmann and colleagues [39] 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 this parameter for most species. Gossmann and colleagues [39] found that $N_{e\pi}$ 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

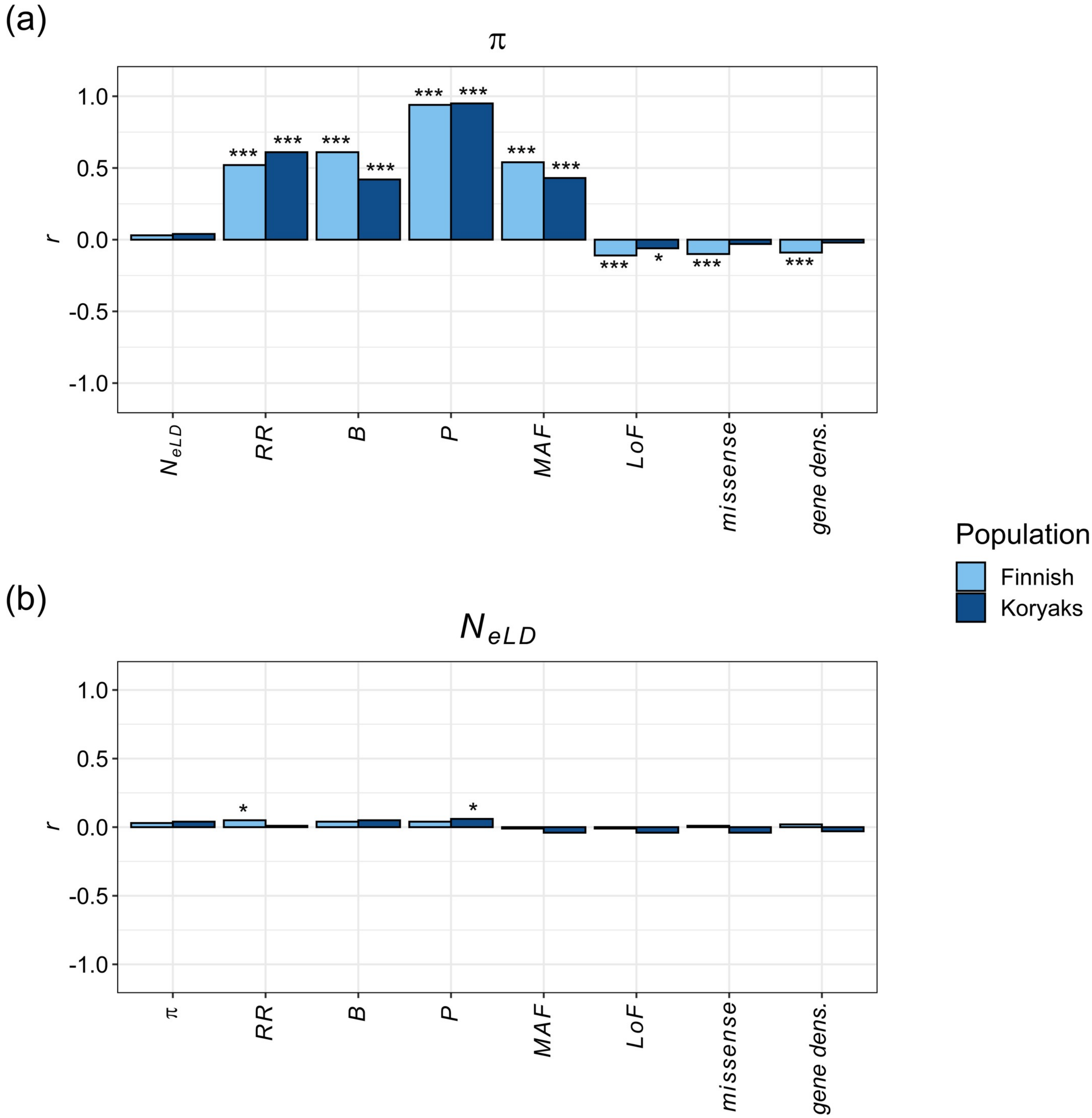

Fig 5. Spearman's correlation coefficient (*r*) between nucleotide diversity (*π*; panel a) or estimates of linkage disequilibrium effective population size (N_{eLD}; panel b) and different diversity variables, estimated within genomic regions. *RR*: recombination rate; *B*: B statistic; *P*: proportion of polymorphic nucleotides; 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.

https://doi.org/10.1371/journal.pgen.1009764.g005

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 colleagues [39] to detect further correlations was attributed to the small variation of $N_{e\pi}$ observed or to only having considered neutral diversity (synonymous changes). In another analysis of genome N_e heterogeneity, Jiménez-Mena and colleagues [40] 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 [40,41]. It can also be argued that the estimate of N_e obtained by the temporal method is closer to N_{eVk} than to $N_{e\pi}$, what may also contribute to explain the lack of significant correlations.

The correlations found between the different parameters analyzed with genomic data followed the expected trends (S3 Fig) and agree with previous estimations. For example, analyzing 100 kb windows of a Danish population, Lohmueller and colleagues [51] 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 compatible with those found in our study between *RR* and genome diversity parameters (S3 Fig).

In summary, our results show that the estimates of historical effective size obtained from linkage disequilibrium between pairs of SNPs are not substantially altered by selection, either positive or negative, nor are they 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 sometimes affected by selection.

Methods

Computer simulations

Genomic data of populations under different demographic and evolutionary scenarios were simulated with the software SLiM 3 [52]. This software simulates a Wright-Fisher model of reproduction with the possibility of adding different types of selection and non-random mating. Random mating populations of constant or changing size (ranging between $N = 100$ and 10,000) with discrete generations were run for up to 10,000 generations. Different demographic scenarios (constant population size, bottlenecking, exponential growth, etc.) were assumed. A model of partial self-fertilization (50% of selfed mating) was also simulated. 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 5 to 0.01 cM per Mb). For each locus, values of fitness of 1, $1 + sh$, 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. Four mutation models were assumed. (1) A neutral model for all mutations ($s = 0$). (2) Background selection (BS), where 95% of mutations are neutral and 5% are deleterious with selection coefficient obtained from a gamma distribution with shape parameter 0.2 and mean value $s = -0.02$ and additive gene action. (3) Selective Sweeps (SS), where 99.9% of mutations are neutral and 0.1% are assumed to be advantageous with effect obtained also from a gamma distribution with shape parameter 0.2 and mean value $s = 0.02$ and additive gene action. (4) Heterozygote advantage (overdominance) for fitness, where 99.99% of mutations are neutral and 0.01% are assumed to be advantageous with effect $s = 0.02$ and dominance coefficient $h = 1.5$. In the absence of selection and for random mating, the expected effective population size from the variance of family sizes is

$N_{eVk} = N$, the number of breeding individuals. With self-fertilization with a proportion $\beta = 0.5$ of selfed matings, N_{eVk} is expected to be $N_{eVk} = N/(1 + \alpha)$, where $\alpha = \beta /(2 - \beta)$ (see, e.g., Caballero [4], p. 101), i.e., $N_{eVk} = 3N/4$.

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 (S6 Fig). 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 [53]. In the process, 65,088 variants were not found and were excluded from the analyses. However, a large number of SNPs is 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. Telomeric regions shorter than 2 cM were also removed from the analysis. In addition, regions with extremely large N_e estimates (> 100,000) or negative ones were also excluded (see the distribution of N_{eLD} values for genomic regions in S4 Fig). Thus, following these criteria, 120 and 140 regions were excluded from the analyses of Finnish and Koryaks data, respectively, and the final number of regions analysed was 1,621 and 1,180, 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 MSMC [18] and Relate [19] 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 of minimum mutations per tree from 50 to 30 depending on the simulation scenario. It was run for each simulation replicate and the results were 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 map [34] and averaged for each region. Estimates of the background selection statistic (*B*) [54] 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 [54].

Average nucleotide diversity (π), proportion of polymorphic sites (*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 [55], 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 [56]. When a gene spanned over different regions, we considered it to be in the region were its middle point was located. Only genes with a straightforward chromosome code were used (e.g. NC_000001.10 corresponding with chromosome 1).

Supporting information

S1 Fig. Estimates of effective population size from linkage disequilibrium (N_{eLD}, sample size of n = 100 individuals), Relate ($N_{eRelate}$, n = 100), MSMC (N_{eMSMC}, n = 4), and from nucleotide diversity (π), this latter calculated as $N_{e\pi} = \pi/(4\mu)$, where μ is the nucleotide mutation rate, for scenarios with different recombination rates (*RR* in cM/Mb) uniform across the genome. Simulations assume a fixed population size of N = 1,000 individuals with partial self-fertilization (50% of selfed progeny) under background selection and selective sweeps (see main text for mutational parameters), with constant mutation rate $\mu = 10^{-8}$ per base per generation. Estimates were obtained including windows of recombination rate between pairs of SNPs ranging from c = 0.0025 to 0.0250 for N_{eLD} and averaging historical estimates of N_e between generations 150 to 350 for $N_{eRelate}$ and N_{eMSMC}. Error bars represent one standard error above and below the mean of the simulation replicates. N_{eLD} estimates were obtained pooling all replicates to speed up computation.
(TIFF)

S2 Fig. Estimates of effective population size from linkage disequilibrium (N_{eLD}, sample size of n = 100 individuals), Relate ($N_{eRelate}$, n = 100), MSMC (N_{eMSMC}, n = 4), and from nucleotide diversity (π), this latter calculated as $N_{e\pi} = \pi/(4\mu)$, where μ is the nucleotide mutation rate, for scenarios with different recombination rates (*RR* in cM/Mb) uniform across the genome. Simulations assume a fixed population size of N = 1,000 individuals with random mating assuming a model of heterozygote advantage (overdominance) for fitness (see main text for mutational parameters), with constant mutation rate $\mu = 10^{-8}$ per base per generation. Estimates were obtained including windows of recombination rate between pairs of SNPs ranging from c = 0.0025 to 0.0250 for N_{eLD} and averaging historical estimates of N_e between generations 150 to 350 for $N_{eRelate}$ and N_{eMSMC}. Error bars represent one standard error above and below the mean of the simulation replicates. N_{eLD} estimates were obtained pooling all replicates to speed up computation.
(TIFF)

S3 Fig. Spearman's correlation coefficient (r) of the estimated genomic variables for the Finnish (over the diagonal) and the Koryaks (below the diagonal) populations. *RR*: recombination rate; *B*: B statistic; *P*: proportion of polymorphic nucleotides; MAF: Minor Allele Frequency; *LoF*: number of Loss of Function variants; *missense*: number of missense variants; *gene dens*: gene density. P-values: $^{*} < 0.05$, $^{**} < 0.01$, $^{***} < 0.001$.
(TIFF)

S4 Fig. Distribution of estimates of linkage disequilibrium effective size (N_{eLD}) for different genomic regions using data from Finnish and Koryaks populations.
(TIFF)

S5 Fig. Values of nucleotide diversity (π) for genomic regions with different average recombination rate (*RR*) using data from the Finnish and Koryaks populations. The lines indicate linear regressions.
(TIFF)

S6 Fig. Distribution of mean recombination rate values for 2 cM genomic windows obtained from Myers *et al.* (2005) genetic map. Recombination rates for genomic Windows used in the simulations were randomly obtained from this distribution.
(TIFF)

Acknowledgments

We thank Humberto Quesada for useful comments on the topic. Part of the analyses of this project were run using computing resources and technical support provided by CESGA (Centro de Supercomputación de Galicia).

Web resources

LiftOver UCSC tool, http://genome.ucsc.edu

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

Author Contributions

Conceptualization: Enrique Santiago, Armando Caballero.

Formal analysis: Irene Novo.

Supervision: Enrique Santiago, Armando Caballero.

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

References

1. Wright S. Evolution in mendelian populations. Genetics. 1931; 16: 97–159. https://doi.org/10.1093/genetics/16.2.97 PMID: 17246615
2. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009; 10: 195–205. https://doi.org/10.1038/nrg2526 PMID: 19204717
3. Wang J, Santiago E, Caballero A. Prediction and estimation of effective population size. Heredity. 2016; 117: 193–206. https://doi.org/10.1038/hdy.2016.43 PMID: 27353047
4. Caballero A. Quantitative Genetics. Cambridge University Press; 2020.
5. Luikart G, Ryman N, Tallmon DA, Schwartz MK, Allendorf FW. Estimation of census and effective population sizes: The increasing usefulness of DNA-based approaches. Conserv Genet. 2010; 11: 355–373.

6. Gilbert KJ, Whitlock MC. Evaluating methods for estimating local effective population size with and without migration. Evolution. 2015; 69(8): 2154–2166. https://doi.org/10.1111/evo.12713 PMID: 26118738

7. Waples RS, Do C. Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: A largely untapped resource for applied conservation and evolution. Evol Appl. 2010; 3(3): 244–262. https://doi.org/10.1111/j.1752-4571.2009.00104.x PMID: 25567922

8. Waples RS, England PR. Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. Genetics. 2011; 189(2): 633–644. https://doi.org/10.1534/genetics.111.132233 PMID: 21840864

9. Hill WG. Estimation of effective population size from data on linkage disequilibrium. Genet Res. 1981; 38(3): 209–216.

10. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. 2017; 101(1): 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005 PMID: 28686856

11. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 2003; 13(4): 635–643. https://doi.org/10.1101/gr.387103 PMID: 12654718

12. Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, et al. Recent human effective population size estimated from linkage disequilibrium. Genome Res. 2007; 17(4): 520–526. https://doi.org/10.1101/gr.6023607 PMID: 17351134

13. Santiago E, Novo I, Pardiñas AF, Saura M, Wang J, Caballero A. Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. Mol Biol Evol. 2020; 37(12): 3642–3653. https://doi.org/10.1093/molbev/msaa169 PMID: 32642779

14. Mörseburg A, Pagani L, Ricaut FX, Yngvadottir B, Harney E, Castillo C, et al. Multi-layered population structure in Island Southeast Asians. Eur J Hum Genet. 2016; 24: 1605–1611. https://doi.org/10.1038/ejhg.2016.60 PMID: 27302840

15. Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, et al. The pattern of linkage disequilibrium in German Holstein cattle. Anim Genet. 2010; 41(4); 346–356. https://doi.org/10.1111/j.1365-2052.2009.02011.x PMID: 20055813

16. Saura M, Tenesa A, Woolliams JA, Fernández A, Villanueva B. Evaluation of the linkage-disequilibrium method for the estimation of effective population size when generations overlap: An empirical case. BMC Genomics. 2015; 16: 922. https://doi.org/10.1186/s12864-015-2167-z PMID: 26559809

17. Hollenbeck CM, Portnoy DS, Gold JR. A method for detecting recent changes in contemporary effective population size from linkage disequilibrium at linked and unlinked loci. Heredity. 2016; 117: 207–216. https://doi.org/10.1038/hdy.2016.30 PMID: 27165767

18. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014; 46: 919–925. https://doi.org/10.1038/ng.3015 PMID: 24952747

19. Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy estimation for thousands of samples. Nat. Genet. 2019; 51: 1321–1329. https://doi.org/10.1038/s41588-019-0484-x PMID: 31477933

20. Robertson A. Inbreeding in artificial selection programmes. Genet Res. 1961; 2(2): 189–194.

21. Santiago E, Caballero A. Effective Size of Populations Under Selection. Genetics. 1995; 139(2): 1013–1030. https://doi.org/10.1093/genetics/139.2.1013 PMID: 7713405

22. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics, 1993; 134(4): 1289–1303. https://doi.org/10.1093/genetics/134.4.1289 PMID: 8375663

23. Hudson RR, Kaplan NL Deleterious background selection with recombination. Genetics. 1995; 141(4): 1605–1617. https://doi.org/10.1093/genetics/141.4.1605 PMID: 8601498

24. Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background selection. Genet Res. 1996; 67(2): 159–174. https://doi.org/10.1017/s0016672300033619 PMID: 8801188

25. Nicolaisen LE, Desai MM. Distortions in genealogies due to purifying selection and recombination. Genetics. 2013; 195(1): 221–230. https://doi.org/10.1534/genetics.113.152983 PMID: 23821597

26. Santiago E, Caballero A. Effective size and polymorphism of linked neutral loci in populations under directional selection. Genetics. 1998; 149(4): 2105–2117. https://doi.org/10.1093/genetics/149.4.2105 PMID: 9691062

27. Santiago E, Caballero A. Joint prediction of the effective population size and the rate of fixation of deleterious mutations. Genetics. 2016; 204(3): 1267–1279. https://doi.org/10.1534/genetics.116.188250 PMID: 27672094

28. Campos JL, Halligan DL, Haddrill PR, Charlesworth B. The relation between recombination rate and patterns of molecular evolution and variation in drosophila melanogaster. Mol Biol Evol. 2014; 31(4): 1010–1028. https://doi.org/10.1093/molbev/msu056 PMID: 24489114

29. Lohmueller KE. The distribution of deleterious genetic variation in human populations. Curr Opin Genet Dev. 2014; 29: 139–146. https://doi.org/10.1016/j.gde.2014.09.005 PMID: 25461617

30. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res. 1966; 8(3): 269–294. PMID: 5980116

31. Begun DJ., Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. 1992; 356: 519–520. https://doi.org/10.1038/356519a0 PMID: 1560824

32. Nachman MW. Patterns of DNA variability at X-linked loci in Mus domesticus. Genetics. 1997; 147(3): 1303–1316. https://doi.org/10.1093/genetics/147.3.1303 PMID: 9383072

33. Hellmann I, Prüfer K, Ji H, Zody MC, Pääbo S, Ptak SE. Why do human diversity levels vary at a megabase scale? Genome Res. 2005; 15: 1222–1231. https://doi.org/10.1101/gr.3461105 PMID: 16140990

34. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science. 2005; 310(5746): 321–324. https://doi.org/10.1126/science.1117196 PMID: 16224025

35. Charlesworth B, Charlesworth D. Elements of Evolutionary Genetics. Roberts and Company Publishers; 2010.

36. Maynard-Smith J, Haigh J. The hitchhiking effect of a favourable gene. Genet Res. 1974; 23: 23–35. PMID: 4407212

37. Kaplan N.L., Hudson R.R., and Langley C.H. (1989) The "hitchhiking effect" revisited. Genetics. 23(1), 887–899. https://doi.org/10.1093/genetics/123.4.887 PMID: 2612899

38. Przeworski M. The signature of positive selection at randomly chosen loci. Genetics. 2002; 160(3): 1179–1189. https://doi.org/10.1093/genetics/160.3.1179 PMID: 11901132

39. Gossmann TI, Woolfit M, Eyre-Walker A. Quantifying the variation in the effective population size within a genome. Genetics. 2011; 189(4): 1389–1402. https://doi.org/10.1534/genetics.111.132654 PMID: 21954163

40. Jiménez-Mena B, Hospital F, Bataillon T. Heterogeneity in effective population size and its implications in conservation genetics and animal breeding. Conserv Genet Resour. 2016a; 8: 35–41.

41. Jiménez-Mena B, Tataru P, Brøndum RF, Sahana G, Guldbrandtsen B, Bataillon T. One size fits all? Direct evidence for the heterogeneity of genetic drift throughout the genome. Biol Lett. 2016b; 12(7).

42. Zeng K, Jackson BC, Barton HJ. Methods for estimating demography and detecting between-locus differences in the effective population size and mutation tate. Mol Biol Evol. 2018; 36(2): 423–433.

43. Thomson G. The effect of a selected locus on linked neutral loci. Genetics. 1977; 85(4): 753–788. https://doi.org/10.1093/genetics/85.4.753 PMID: 863244

44. Stephan W, Song YS, Langley CH. The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics. 2006; 172(4), 2647–2663. https://doi.org/10.1534/genetics.105.050179 PMID: 16452153

45. Gillespie JH. Junk ain't what junk does: neutral alleles in a selected context. Gene. 1997; 205(1–2): 291–299. https://doi.org/10.1016/s0378-1119(97)00470-8 PMID: 9461403

46. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2010; 491: 56–65.

47. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010; 463: 757–762. https://doi.org/10.1038/nature08835 PMID: 20148029

48. Ralph PL. An empirical approach to demographic inference with genomic data. Theor Popul Biol. 2019; 127: 91–101. https://doi.org/10.1016/j.tpb.2019.03.005 PMID: 30978307

49. King JP, Kimmel M, Chakraborty R. A power analysis of microsatellite-based statistics for inferring past population growth. Mol Biol Evol. 2000; 17(12): 1859–1868. https://doi.org/10.1093/oxfordjournals.molbev.a026287 PMID: 11110902

50. Gillespie JH. Genetic drift in an infinite population: The pseudohitchhiking Model. Genetics. 2000; 155 (2): 909–919. https://doi.org/10.1093/genetics/155.2.909 PMID: 10835409

51. Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, et al. Natural selection affects multiple aspects of gnetic variation at putatively neutral sites across the human genome. PLoS Genet. 2011; 7(10): e1002326. https://doi.org/10.1371/journal.pgen.1002326 PMID: 22022285

52. Haller BC, Messer PW. SLiM 3: Forward genetic simulations beyond the wright-fisher model. Mol Biol Evol. 2019; 36(3): 632–637. https://doi.org/10.1093/molbev/msy228 PMID: 30517680

53. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 2006; 34: D590–D598. https://doi.org/10.1093/nar/gkj144 PMID: 16381938

54. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009; 5: e1000471. https://doi.org/10.1371/journal.pgen.1000471 PMID: 19424416

55. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: Displaying reference data information from over 60000 exomes. Nucleic Acids Res. 2017; 45 (D1): D840–D845. https://doi.org/10.1093/nar/gkw971 PMID: 27899611

56. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016; 44(D1): D733–D745. https://doi.org/10.1093/nar/gkv1189 PMID: 26553804