Population Growth Inflates the Per-Individual Number of Deleterious Mutations and Reduces Their Mean Effect

Elodie Gazave,* Diana Chang,* Andrew G. Clark,*,† and Alon Keinan*,[1]
*Department of Biological Statistics and Computational Biology and †Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853

ABSTRACT This study addresses the question of how purifying selection operates during recent rapid population growth such as has been experienced by human populations. This is not a straightforward problem because the human population is not at equilibrium: population genetics predicts that, on the one hand, the efficacy of natural selection increases as population size increases, eliminating ever more weakly deleterious variants; on the other hand, a larger number of deleterious mutations will be introduced into the population and will be more likely to increase in their number of copies as the population grows. To understand how patterns of human genetic variation have been shaped by the interaction of natural selection and population growth, we examined the trajectories of mutations with varying selection coefficients, using computer simulations. We observed that while population growth dramatically increases the number of deleterious segregating sites in the population, it only mildly increases the number carried by each individual. Our simulations also show an increased efficacy of natural selection, reflected in a higher fraction of deleterious mutations eliminated at each generation and a more efficient elimination of the most deleterious ones. As a consequence, while each individual carries a larger number of deleterious alleles than expected in the absence of growth, the average selection coefficient of each segregating allele is less deleterious. Combined, our results suggest that the genetic risk of complex diseases in growing populations might be distributed across a larger number of more weakly deleterious rare variants.

THE human population size has been growing rapidly, most notably since the advent of agriculture ~10,000 years ago. The rate of population growth has increased over time to as much as 10–30% per generation during the last 500 years (Cohen 1996; Hawks *et al.* 2007; United Nations Department of Economic and Social Affairs Population Division 2011; Keinan and Clark 2012). This recent demographic growth is reflected in recent estimates from genetic data of the human current effective population size (N_e), with all estimates being much higher than the conventionally estimated historical variance effective population size of ~10,000 (Coventry *et al.* 2010; Keinan and Clark 2012; Nelson *et al.* 2012; Tennessen *et al.* 2012). The growth in effective population size has resulted in an excess of rare alleles due to the very large number of recent mutations (Coventry *et al.* 2010; Fu *et al.* 2012; Keinan and Clark 2012; Nelson *et al.* 2012; Tennessen *et al.* 2012). Not only are the vast majority of protein-coding variants rare and recent, but also both the group of rarer and the group of more recent variants are enriched for deleterious mutations (Fu *et al.* 2012; Nelson *et al.* 2012; Tennessen *et al.* 2012). The extent to which recent population expansion has contributed to these empirical observations depends on how purifying selection has operated during the epoch of growth. Growth can potentially increase the number of deleterious alleles carried by each individual and the burden of disease-causing alleles in the human population.

It is crucial to understand the interaction of demographic expansion and natural selection, and in particular purifying (negative) selection, to assess whether the recent explosive growth has affected the genetic basis and architecture of complex disease. Several theoretical predictions of natural selection in a population of varying size have been formulated, but there is no adequate coverage of the situation

doi: 10.1534/genetics.113.153973
Manuscript received June 3, 2013; accepted for publication August 9, 2013
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.153973/-/DC1.
[1]Corresponding author: Cornell University, 102C Weill Hall, Ithaca, NY 14853. E-mail: ak735@cornell.edu

where a population started growing only recently and is far from reaching a new mutation–selection–drift equilibrium. On one hand, it has been suggested that rapid growth of human populations may have led natural selection to be inefficient at removing deleterious mutations (Casals and Bertranpetit 2012), which could lead to a rapid accumulation of deleterious alleles (Kondrashov 1988; Lynch 2010). On the other hand, since an increase in N_e increases the efficacy of natural selection to raise the frequency of favorable mutations and reduce the frequency of deleterious mutations (Kimura 1955, 1957), it has also been suggested that with the current human N_e, even slightly deleterious mutations can be purged (Reed and Aquadro 2006). Indeed, for a population that has been exponentially growing throughout its history, it has been established that deleterious mutations have a lower probability of fixation compared to those in a nongrowing population, and most deleterious mutations will eventually be lost (Waxman 2011). However, for mutations that reach fixation, the expected time until fixation (conditional upon fixation) increases with population size (Kimura and Ohta 1969; Muruyama and Kimura 1974; Waxman 2012). Thus, weakly deleterious mutations can segregate longer, and increase in their number of copies in the population.

In this study, we examine the characteristics of deleterious mutations in a scenario of recent explosive growth of approximately the magnitude experienced by human populations, with the aim of answering three questions. First, is the population growth expected to lead to an increase in the number of deleterious mutations carried by each individual in the extant, large population? Second, how does the growth affect the purging of rare deleterious derived alleles? And third, has growth affected the average selection coefficient of segregating variants and, as a consequence, the average individual fitness?

Due to the limitations of theoretical results that assume equilibrium, we addressed these questions, using forward-in-time simulations that track the evolution of newly arisen mutations with known selective effects over time. Importantly, using simulations allowed the comparison of models with and without recent growth, as well as with and without purifying selection, thereby revealing how population growth alone has altered the efficacy of purifying selection.

The main results of the simulations are that population growth largely increases the number of segregating sites in the population and slightly increases the average number of mutations carried by each individual, including that of deleterious mutations. Studying the relationship between the number of copies of a derived allele and its selection coefficient, our results indicate that the increased efficacy of natural selection in a growing population results in a faster elimination of newly introduced mutations that are strongly deleterious. As a consequence, while each individual in a growing population carries a larger number of deleterious variants, the average selection coefficient of each variant copy is less deleterious than in a population that has remained at constant size in recent history. Finally, we discuss the implications of these results on the overall individual fitness in extant human populations and on the genetic architecture of complex diseases.

Materials and Methods

Simulated loci

We performed forward-in-time population genetic simulations using the program SFS_CODE (Hernandez 2008), including a few minor revisions to its source code. We simulated two types of loci, one evolving neutrally and one evolving under purifying selection. For each type of locus, we considered two demographic scenarios, one that includes recent population growth and one without growth. For each of these four locus-by-demography combinations, we simulated genetic sequences of 5 kb with a mutation rate of 2×10^{-8} per nucleotide. The value chosen for the mutation rate is higher than recent estimates (Campbell *et al.* 2012; Keightley 2012), but does not affect our conclusions since these are based on comparing the different simulated combinations. For computational efficiency, each individual 5-kb locus is in complete linkage, with no recombination. We simulated 10,000 independent replicates of each of the four models, and considered summary statistics in each replicate and then averaged them across the 10,000 replicates. Relevant SFS_CODE command lines are provided in Supporting Information, File S1.

Models of population history

The two demographic scenarios (with growth and without growth) occur in two distinct populations that split from a common ancestral population. At the split time, SFS_CODE makes a copy of the ancestral population such that each new population conserves the full ancestral effective population size of $N_e = 10{,}000$ (throughout, quantities denote *effective* population sizes). This ensures that both populations start in identical states. Prior to the population split, the ancestral population follows a demographic model of European history with two population bottlenecks, as described in Keinan *et al.* (2007). Specifically, after the burn-in phase, the population undergoes a first bottleneck of intensity $F = 0.264$ at 4720 generations ago followed by a quick recovery to the ancestral population size of $N_e = 10{,}000$. The second bottleneck occurs 720 generations ago with an intensity of $F = 0.09$, also followed by a quick recovery to the ancestral population size (Keinan *et al.* 2007). Then, 420 generations before present the ancestral population splits into two populations. One population (referred to as the "with growth" population) starts growing exponentially 400 generations before present (~10,000 years ago) at a rate of 1.74% per generation thereby reaching a final size of $N_e = 10{,}000{,}000$ at the end of the simulation (Keinan and Clark 2012). The other population (the "no growth" population) maintains a constant $N_e = 10{,}000$ from the split until the present.

The with growth and no growth models are therefore identical over their entire history except for the last 400 generations (Figure S1). We followed mutations during the last 440 generations of the simulation, covering the entire exponential growth phase and the 40 preceding generations when the two models are still identical.

We considered additional models of population history, including varying recent growth and ancient history, as well as a baseline of a population that has been of constant size throughout history (File S1). These additional models show that all results presented throughout the study are robust to the details of ancestral model of European history (Keinan *et al.* 2007) and the recent explosive growth model (File S1). In all models we considered a generation time of 25 years.

Mutations and natural selection

For each model of population history, we simulated loci in which either all mutations are neutral (selection coefficient $s = 0$) or all mutations are deleterious ($s < 0$). For each deleterious mutation, s was obtained from the population-scaled selection coefficient, $\gamma = 2Ns$ with γ following the opposite of a gamma distribution, (α, β). We chose the shape parameter $\alpha = 0.206$ and rate parameter $\beta = 1/2740$ from Boyko *et al.* (2008), after rescaling to an ancestral N_e of 10,000. With these parameter values, the average s was -0.028. We simulated all loci as "noncoding" in SFS_CODE nomenclature since "coding" regions assume no fitness effect ($s = 0$) at approximately one-third of sites. Mutations with $s < -1$were set to $s = -1$.

Following the implementation in SFS_CODE (Hernandez 2008), the fitness of each individual is the product of the fitness effect of the mutations it carries, which for a mutation with selection coefficient s ($s < 0$ for deleterious mutations) is $1 + s$ in heterozygotes and $(1 + s)^2$ in homozygotes. The selection model implemented in SFS_CODE is a model of shift in fitness, such that the selection coefficient of alleles that reach fixation is reset to 1. Throughout, we therefore ignored mutations that reach fixation. The use of a shift in fitness model has a minimal effect on the results since fixations are extremely rare (<0.3% of the mutations) during the 400 generations followed here. In addition, this model has recently been shown to be a realistic fit to human mutation load (Keightley *et al.* 2011; Lesecque *et al.* 2012). We refer to derived alleles (*i.e.*, the new alleles introduced by the mutation process) as *lost* from the population when they reach 0 copies.

Simulation scaling

For computational efficiency, and since the extant effective population size following growth is very large (10,000,000), we scaled down both the effective population size and the time by a factor of 10. This scaling approach has been shown to lead to little change in resulting patterns of variation (Hernandez 2008). While scaling does not broadly alter allele frequencies, it does affect the nominal number of copies of each allele. For example, the simulated ancestral population size is 1000 individuals, in which a singleton (allele appearing in a single copy) is of frequency 0.05%, while a singleton in a population of 10,000 individuals is of frequency 0.005%. For this reason, we do not make direct quantitative inference on the distribution of rare variants in the real-sized human population, but rather focus all analyses and conclusions on a comparison of the models with and without growth.

Analyses and summary statistics

At each generation, we recorded the number of segregating sites, the number of copies of the derived allele at each segregating site, and their selection coefficients. We calculated and averaged the following quantities for each simulated scenario. S is the number of segregating sites, *i.e.*, genomic positions that carry one or more copies of a derived allele (Figure 1A). The derived allele count (DAC) is the number of copies of the derived allele of a segregating site, which we denote by δ_i for site i. The proportion of segregating sites lost, $\%S_{\text{lost}}$, is given by R/S, where R is the number of segregating sites lost (Figure 1B). The proportion of segregating sites lost is also computed within different categories of DAC, *i.e.*, R_k/S_k for different values of DAC $k = 1, \ldots, 6$ (Figure 2). The fraction of derived alleles at lost sites, $\%\text{DA}_{\text{lost}}$, is given by $\sum_{j=1}^{R}\delta_j/\sum_{i=1}^{S}\delta_i$. We partitioned deleterious segregating sites into three categories based on their selection coefficient s: "very deleterious" ($s \in [-1, -0.01]$), "mildly deleterious" ($s \in (-0.01, -0.0001]$), and "nearly neutral" ($s \in (-0.0001, 0]$). The percentage of derived alleles in each of these categories is given by $\%\text{DA}_k = \sum_{j=1}^{S_k}\delta_j/\sum_{i=1}^{S}\delta_i \times 100$, where S_k is the total number of segregating sites in category k (Figure 3). The average fitness effect across copies of derived alleles, w_{DA}, is given by $\sum_{i=1}^{S}\delta_i s_i/\sum_{i=1}^{S}\delta_i$, where s_i is the selection coefficient of the derived allele at site i (Figure 4). The average number of mutations per individual chromosome is calculated as $L = \sum_{i=1}^{S}\delta_i/2N_e$, where N_e is the simulated (scaled) number of individuals (Figure 5).

Results

Accumulation and loss of segregating sites

We first measured the accumulation of mutations in the population by tabulating S, the number of segregating sites. Since our simulations assume an infinite-sites model, each new mutation introduces a new segregating site. As expected (Watterson 1975; Tajima 1989), S increases rapidly over time as the population grows, culminating in over two orders of magnitude increase following 400 generations of growth, both with and without selection (Figure 1A). For both demographic scenarios, loci with deleterious mutations have on average fewer segregating sites than loci with solely neutral mutations (Figure 1A), as expected by purifying selection (*e.g.*, Przeworski *et al.* 1999), but the relative difference between the two becomes smaller as the population grows (Figure 1A).

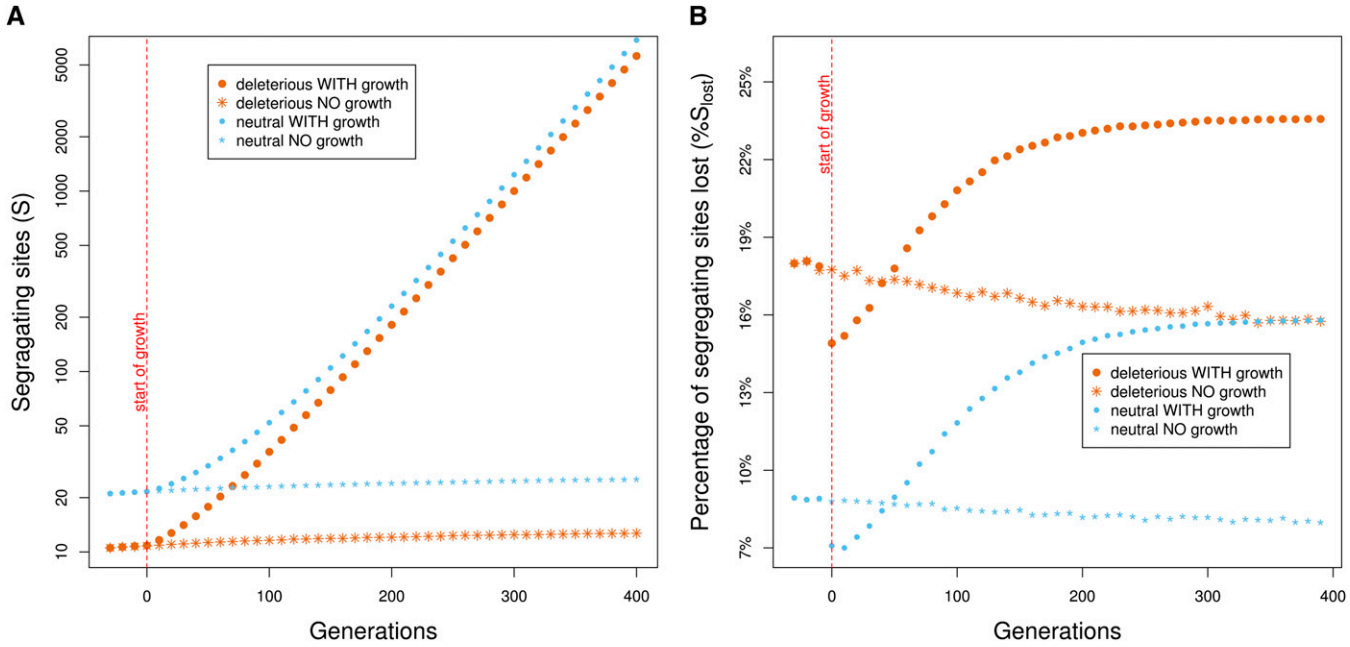


Figure 1 Population growth increases the number of segregating sites, but also the fraction of sites that are lost. (A) S, the number of segregating sites of the whole population (on a log scale); (B) $\%S_{lost}$, the percentage of segregating sites lost from the population in a single simulated generation (*Materials and Methods*). Both panels present the two simulated demographic scenarios (with growth and with no growth) for each selection model (neutral or deleterious). Results are presented every 10 generations (corresponding to a single simulated generation) during the last 440 generations. Population growth increases both S and $\%S_{lost}$. S is smaller for deleterious than for neutral mutations, while $\%S_{lost}$ is higher. Trends with time in the models without growth are due to the preceding population bottlenecks (Figure S3).

To further investigate the effect of genetic drift and natural selection on the number of segregating sites under population growth, we estimated at each generation the percentage of segregating sites that are not observed in the next generation ($\%S_{lost}$). After a few generations of mutation accumulation, $\%S_{lost}$ becomes higher for the model with population growth, both for neutral and for deleterious loci (Figure 1B), implying that population growth increases not only the number of segregating sites, but also the rate at which they are lost. This phenomenon is explained by the larger fraction of singletons (Figure S2) and very rare variants in the growing population, which have a higher probability of loss (Figure 2).

Derived allele count of segregating sites

Each segregating site in the population can be categorized by the number of sequences that carry the derived allele. The average DAC per segregating site (see *Materials and Methods*) is a measure of the prevalence of those sites in the population. This measure is pertinent when comparing populations of different sizes since allele frequencies are difficult to interpret as the sample size (which is here the population size) increases in the population with growth. Furthermore, it is the allele count, rather than the allele frequency, that affects its probability of loss or transmission (File S1).

Our simulations reproduce a well-established effect of population growth (Slatkin and Hudson 1991; Wakeley 2008; Coventry *et al.* 2010; Keinan and Clark 2012; Nelson *et al.* 2012; Tennessen *et al.* 2012) by showing an increase in the proportion of singletons (sites with DAC = 1) (Figure S2). The proportion is further elevated at deleterious loci for both population models (Figure S2). To investigate the efficacy of purifying selection in a growing population free of the expected skew in the site frequency spectrum, we consider instead the DAC of lost segregating sites.

We computed the percentage of lost sites within each category of DAC, with $\%S_{lost}$ for DAC = k being the percentage of segregating sites with k derived alleles that are lost within one generation (*Materials and Methods*). In contrast to the increase of $\%S_{lost}$ when considered across all DACs (Figure 1B), we observe that within each DAC category, population growth decreases $\%S_{lost}$ (Figure 2A), both for neutral (36.7% to ~30.4%) and for deleterious loci (~44.1% to ~39.7% for singletons). This differential direction (Figure 1B vs. Figure 2A) is due to the greater percentage of variants with low DAC in the growing population. For example, singletons represent 46.8% of all segregating sites under growth with neutral mutations and only 17.9% in the same scenario without growth (Figure S2). As such, the percentage of variants that are singleton and lost (without conditioning on being a singleton) from all segregating sites is higher in the growing population than in the scenario without growth (30.4% × 46.8% = 14.2% *vs.* 36.7% × 17.9% = 6.6%, respectively).

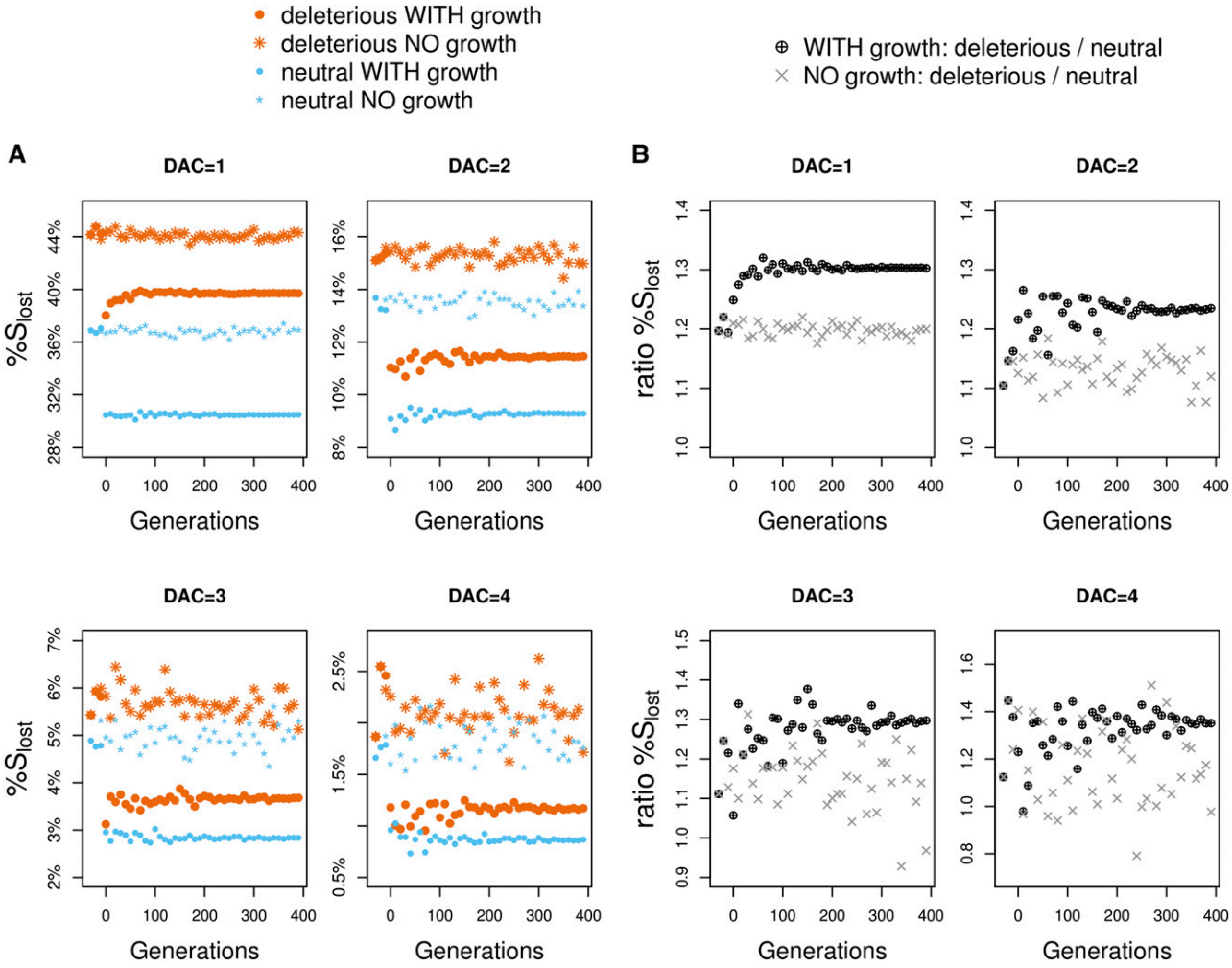


Figure 2 Rare variants are less likely to be lost during population growth, but deleterious ones are purged more efficiently. (A) Percentage of sites of a given derived allele count (DAC) that are lost (%S_{lost}) in a single simulated generation. For example, at a neutral locus in the scenario without growth, just over 36% of all the singletons in the population are not observed in the next generation, with the other 64% being transmitted (with any number of copies) to the next generation. Note the different scaling of the *y*-axis in the different panels, which also explains the noisier trends as DAC increases. (B) For each demographic scenario, the same data as in A are presented as the ratio of %S_{lost} at the deleterious over %S_{lost} at the neutral loci. Population growth increases this ratio, which reflects the higher efficacy of natural selection.

In addition to the decrease in %S_{lost} in each DAC, we also observe that %S_{lost} is always higher at the loci under selection than at the neutral loci. Both these results follow expectations of the action of population growth and selection, respectively, but interestingly the increase in %S_{lost} due to selection is proportionally higher in the growing population (Figure 2B). For singletons for instance, the proportion of segregating sites lost is 1.3 times higher under selection with growth, while it is only 1.2 times higher under selection without growth (Figure 2B). These results show that purifying selection further facilitates the purging of deleterious sites in a growing population.

Segregating sites that are lost are overwhelmingly sites with low DACs. Specifically, among all segregating sites that are lost in a given generation, singletons and doubletons make up >95% in both population models (Figure S4). Sites with DAC $>$ 10 represent between 0% and 3.3×10^{-5}% of the lost sites at deleterious loci for the model without and with growth, respectively. Hence, lost segregating sites represent only a small fraction of all the copies of derived alleles present in the population. While ~16% of segregating sites are lost at the neutral locus with growth (Figure 1B), these sites compose only 0.1% of the total number of copies of derived alleles in the population (%DA_{lost}, *Materials and Methods*; Figure S5). At a locus under purifying selection, %DA_{lost} is an order of magnitude higher in both demographic models (Figure S5), thus showing the contribution of natural selection in the process of allele loss and removal.

In summary, by carefully considering the DAC at lost sites, we have shown that the action of natural selection is not invalidated under population growth. In addition, although it decreases the proportion of segregating sites lost in each DAC category, the larger size of the growing

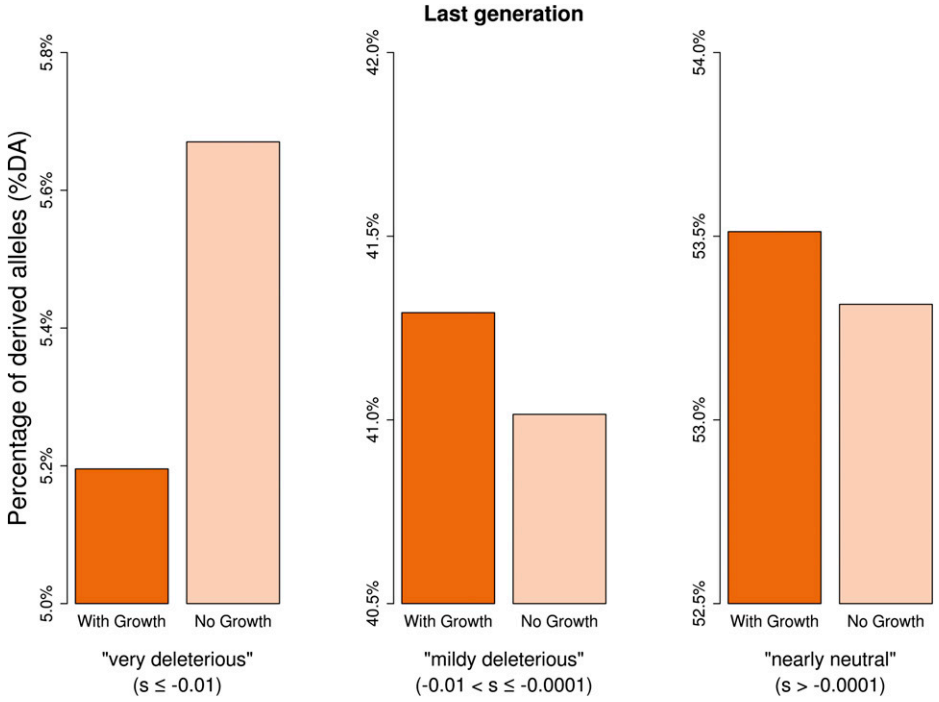


Figure 3 Higher efficiency of natural selection in a growing population decreases the percentage of the most deleterious allele copies. Segregating sites are classified into three discrete categories of fitness effect. For each category, the percentage of derived alleles (%DA) is the sum of the number of copies of derived alleles observed across all the segregating site in the category divided by the total number of derived alleles across all segregating sites in the population × 100 (%DA of the three categories sums up to 100). Data are shown for the last generation of the simulation, both with and without recent growth. Vertical bars denote ±SE based on 10,000 replicates. Population growth leads to a lower percentage of derived alleles in the most deleterious category.

population improves the efficacy of natural selection, and deleterious sites are more readily eliminated.

Fitness effect of deleterious alleles in a growing population

Average fitness effect of a deleterious mutation: To go beyond the burden in the number of deleterious mutations and consider their effects, we compared the distribution of selection coefficients in the population models with and without growth. We computed the average fitness effect (selection coefficient) for derived alleles that are lost and derived alleles that are transmitted to the next generation by averaging the fitness effect of each allele weighed by its number of copies (*Materials and Methods*). As expected, in both demographic scenarios, lost sites are much more deleterious than sites that are transmitted (Figure S6). Interestingly, this phenomenon is more pronounced in a growing population, again pointing to the higher efficacy of selection in a larger population (Figure S6).

To obtain a snapshot of the fitness effect of all segregating variation (*i.e.*, independently of whether sites are lost or transmitted to the following generation), we partitioned segregating sites into three categories corresponding to very deleterious, mildly deleterious, and nearly neutral (*Materials and Methods*). Considering the number of copies of each site, we measured the percentage of copies of derived alleles (%DA) that fall into each category. In the very deleterious category, %DA decreases progressively over time as the population grows (Figure S7). At the last generation of the simulation, %DA is significantly lower (by 8.4%) than in the model without growth (Figure 3). The effect of the population model on the other two categories is much smaller and nonsignificant (Figure S7 and Figure 3). The stronger effect of population growth on the most deleterious alleles is also visible in the site frequency spectrum (Figure S8). More generally, the selection coefficient s averaged across all derived allele copies (w_{DA}) becomes less deleterious as the population grows (Figure 4). At the end of the simulation, an allele chosen randomly is 15.8% less deleterious in the population that has undergone growth (Figure 4). We note that the average selection coefficient also increases—although to a smaller extent—in the absence of growth (Figure 4). This is because the population model without growth is also not at equilibrium due to the preceding population bottlenecks. The role of the bottlenecks becomes evident in comparison to a model of a population that has been of constant size throughout history (File S1; Figure S9).

Despite the accumulation of deleterious segregating sites in the growing population, we show a stronger increase in the average fitness effect of derived alleles in the growing population (Figure 4). This effect is particularly evident when considering the relative amount of very deleterious alleles. In the scenario with growth, for one copy of a very deleterious allele there are 130 copies of nearly neutral ones; the respective is only 1 to 48 (Table S1A). The new mutations accumulated due to growth tend to be more deleterious due to their recency, but while less deleterious alleles increase in number of copies faster as the population grows, the very deleterious alleles are purged more effectively in this scenario.

Average number of mutations per chromosome: We next considered the burden of deleterious mutations as the number of mutations present in each of the $2N_e$ chromosomes in

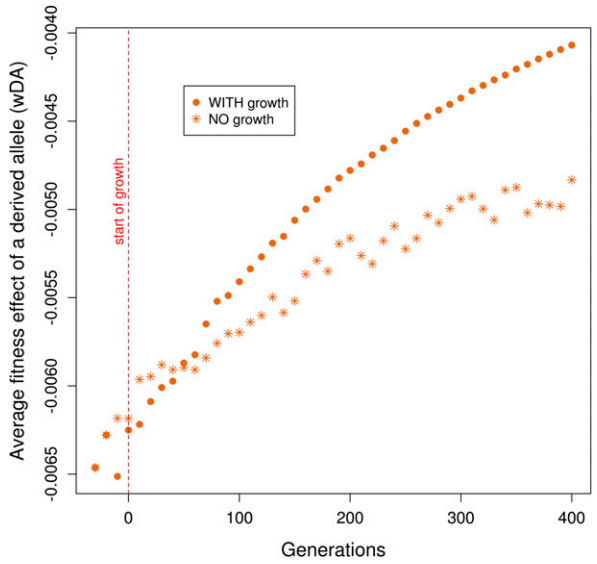


Figure 4 The average selection coefficient across alleles present in the population is increased by population growth. The average selection coefficient of a derived allele (*w*DA) is obtained by weighting the selection coefficient of each segregating site by its number of copies (*Materials and Methods*). The increase in average selection coefficient of derived alleles shows that alleles are on average less deleterious over time. The increase in the population model without growth is due to the preceding population bottlenecks (File S1). The increase is faster for the model with population growth.

the population. As expected, the average number of mutations per chromosome, L, is much larger at the neutral loci than at the deleterious loci (Figure 5). L is also larger—both with and without selection—in the growing population (Figure 5). This increase in L is steady over the generations of population growth, but in stark contrast to the several orders of magnitude increase of S (Figure 1A), L increases by only 0.9% relative to the model without growth at the neutral locus, and by only 6% at the deleterious locus (Figure 5; Figure S10). This can be understood by considering that the time to the most recent common ancestor (t_{MRCA}) of a neutral locus, which underlies L, can increase only by as many generations as growth lasted, no matter how extreme that growth has been. For the demographic models considered here, t_{MRCA} for a pair of chromosomes is on the order of 15,000 generations, and thus the very recent growth starting only 400 generations ago can lead only to a relatively small increase (File S1).

Average fitness of individuals: We established that on the one hand, population growth slightly increases the average number of mutations carried by an individual and, on the other hand, each of these alleles is slightly less deleterious. The overall fitness of an individual is a function of the combined effect of all deleterious alleles it carries. In our simulations, the above two effects counteract each other such that individual fitness is similar between the two population models (Figure S11). We note, however, that this specific result might vary as a function of growth model parameters and dominance level.

Discussion

In this study, we address whether an extremely fast and recent population growth can hinder the action of purifying selection. We studied this independent of any effects that the transition to agriculture itself might have had on purifying selection (Eshed *et al.* 2004; Gage and DeWitte 2009). The joint effect of natural selection and population growth on deleterious alleles is not trivial to understand since these two forces have opposite effects: purifying selection purges deleterious alleles while population growth introduces an excess of deleterious mutations into the population. To disentangle the effects of population growth and purifying selection, we compared simulated loci with neutral mutations to those with deleterious mutations, each in population models with and without a recent epoch of exponential growth. The particular choice of the population growth model does not affect the results, as we showed by repeating the analyses using other published models of recent European history (File S1, Figure S12, Figure S13, and Figure S14).

Our results show that population expansion is accompanied by an accumulation of segregating sites. At both locus types and in both demographic scenarios, mutations that are not transmitted to the next generation are typically singletons or doubletons, which constitute the majority of segregating sites but only a small fraction of all copies of derived alleles present in the population. Beyond known differences in the site frequency spectrum, we showed that mutations lost during population growth have on average more deleterious fitness effects than in a population that does not experience growth. This effect is attributable to the increased efficacy of purifying selection as the population size increases. As a result, derived alleles present in the growing population have on average a less deleterious effect when averaged per allelic copy. This result may seem at odds with recent sequencing studies that have shown that human populations carry a burden of recent mutations that tend to be more deleterious due to their recency (Fu *et al.* 2012; Nelson *et al.* 2012; Tennessen *et al.* 2012). However, recent mutations are expected to be more deleterious also in the absence of population growth. Thus, to understand how population growth has affected the selection coefficient of segregating mutations, the empirical comparison should be made to a human population that has not experienced recent growth. Here, we show that averaging all segregating sites, across frequency or age, the selection coefficient of an allele copy picked at random in the extant human population is less deleterious than it would have been had the population not gone through an epoch of extreme growth. We conclude that natural selection purges the most deleterious alleles more efficiently in the scenario with growth.

While the selection coefficient of derived alleles is on average less deleterious, our results also show that each

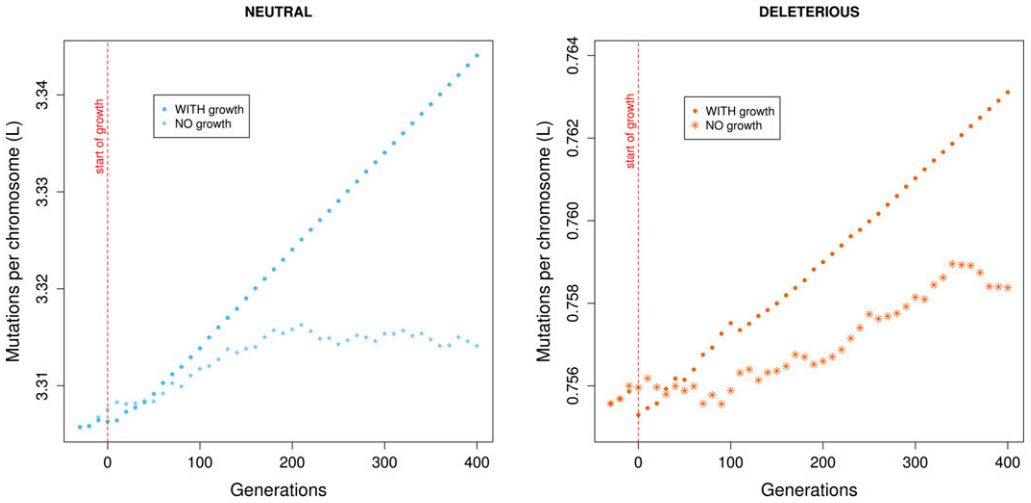


Figure 5 The number of mutations carried by each individual chromosome is higher in a growing population. The number of mutations per chromosome L (*Materials and Methods*) is presented for neutral and deleterious loci. L increases slowly as the population grows: L increases by only 1.14% and 1.03% at the neutral and the deleterious locus, respectively, between the beginning and the end of the growth.

individual carries a larger number of deleterious alleles in a growing population though only modestly so due to the recency of the growth in the human genealogical scale. Overall, the two effects balance out and the average individual fitness is not different in the growing population from that in the population without growth. Importantly, the simulations presented in this study considered only the case of mutations that are partially dominant, with fitness effect being $1 + s$ for heterozygotes and $(1 + s)^2$ for homozygotes. For the vast majority (>94%) of mutations, the selection coefficient s is distributed between 0 and -0.01, for which s^2 is negligible and this model is approximately additive ($1 + s$ and $1 + 2s$), which is commonly assumed in genome-wide association studies (GWAS). The actual distribution of dominance degree (h) is difficult to obtain in humans, except to note that most Mendelian disorders are largely recessive. In model organisms, estimates of dominance vary considerably and are biased in various ways (Fernandez *et al.* 2004; Agrawal and Whitlock 2011). In addition, it has been established that h covaries with s (Simmons and Crow 1977; Caballero and Keightley 1994; Phadnis and Fry 2005). While it remains to be tested how our results translate to alleles of varying dominance models, the effect of recent population growth is characterized by very rare alleles, which are seldom observed as homozygous.

Our results also show that deleterious derived alleles are expected to have fewer copies in the population than neutral ones and especially so in the growing population. This result is in agreement with empirical data showing that the odds ratio of rare variants being functional compared to variants with minor allele frequency >0.5% is 4.2 (Tennessen *et al.* 2012). Importantly, low frequency is not the sole predictor of functionality in a demographic expansion scenario, since we observed that growth also leads to accumulation of extremely rare neutral variants. Applying Murayama's theory (Muruyama 1974) to population growth, Maher *et al.* (2012) and Kiezun *et al.* (2013) showed that—conditioned on allele frequency—allele age can be powerful in predicting selective effect.

The impact of changes in population size on deleterious variants has received considerable attention in recent years. Lohmueller *et al.* (2008), comparing 15 African American and 20 European American individuals, showed that human populations that went through a rapid and recent population bottleneck present a higher proportion of deleterious variation. Comparing the same two populations, Tennessen *et al.* (2012) found that this result was dependent on the criteria used to classify variants as putatively deleterious. In non-European samples that experienced population bottlenecks, Szpiech *et al.* (2013) showed that recent inbreeding increased the proportion of mildly deleterious homozygous mutations. The impact of ancient bottlenecks on the average selection coefficient of an allele is also visible in our simulations. While our simulations show that population growth does not have a downward impact on the average selection coefficient of a derived allele copy, population bottlenecks had more notably affected deleterious variation.

Population growth generated a strong increase in the number of segregating sites in the population that can potentially play a role in complex disease risk. Since the vast majority of these variants are extremely rare, recent growth leads only to a moderate increase in the number of derived alleles carried by each individual. This supports the claim that common diseases may frequently be subject to strong genetic heterogeneity (Lango Allen *et al.* 2010), with different patients that have a similar diagnosis carrying rare or private mutations at the same or different loci (Galvan *et al.* 2010; McClellan and King 2010; Ravanbod *et al.* 2012). At the same time, our results have no implications on the heterogeneity of *de novo* mutations since these are not affected by demographic history. Importantly, a larger population size by itself (and the consequent higher efficiency of selection), without population growth, would also result in increased genetic heterogeneity. In summary, we showed that the recent rapid growth experienced by many human populations can be partially responsible for increased levels of genetic heterogeneity in the architecture of complex disease and traits, via two effects: (1) the introduction of a much

larger number of variants and (2) improved purging of the most deleterious alleles and maintenance of more mildly deleterious alleles, with the latter having smaller effect sizes on average for diseases that are under purifying selection.

Our results indicate that the recent rapid expansion of human populations has perturbed different population genetic attributes of our species. These can be used to suggest directions of exploration, define strategies in medical genetics, refine association methods, and tests of positive natural selection optimized for the genetic diversity segregating in human populations (Yu *et al.* 2009). Our results have clear relevance to the issue of missing heritability in genome-wide association studies (Maher 2008; Manolio *et al.* 2009; Eichler *et al.* 2010). The impact of population growth on individual mutation load and the genetic architecture of complex diseases deserve further study, both theoretically and empirically. For example, a careful study of recombination and linkage patterns would benefit the association methods based on identity-by-descent (Gusev *et al.* 2011; Browning and Thompson 2012; Zhuang *et al.* 2012). Comparing large studies of deep sequencing between functional and nonfunctional regions will provide further empirical insight into ways that rapid growth affects the balance of mutation, drift, and selection. Finally, theoretical models need to be extended to accommodate additional factors, including variation in the degree of dominance, variation in family size (Wakeley *et al.* 2012), and changes in variance in reproductive success over time.

Acknowledgments

The authors thank Leonardo Arbiza for help with optimization of analysis tools, and to Joshua Akey, Leonardo Arbiza and Kevin Mitchell for comments on earlier versions of this paper. A.G.C. and A.K. were supported in part by the National Institutes of Health (U01-HG005715). E.G. was supported in part by a Cornell Center for Comparative and Population Genomics fellowship. A.K. was also supported by The Ellison Medical Foundation, an Alfred P. Sloan Research Fellowship, and the Edward Mallinckrodt, Jr. Foundation.

Literature Cited

Agrawal, A. F., and M. C. Whitlock, 2011 Inferences about the distribution of dominance drawn from yeast gene knockout data. Genetics 187: 553–566.

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4: e1000083.

Browning, S. R., and E. A. Thompson, 2012 Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics 190: 1521–1531.

Caballero, A., and P. D. Keightley, 1994 A pleiotropic nonadditive model of variation in quantitative traits. Genetics 138: 883–900.

Campbell, C. D., J. X. Chong, M. Malig, A. Ko, B. L. Dumont *et al.*, 2012 Estimating the human mutation rate using autozygosity in a founder population. Nat. Genet. 44: 1277–1281.

Casals, F., and J. Bertranpetit, 2012 Genetics. Human genetic variation, shared and private. Science 337: 39–40.

Cohen, J. E., 1996 *How Many People Can the Earth Support?* Ed. 1. W. W. Norton, New York.

Coventry, A., L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell *et al.*, 2010 Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat. Commun. 1: 131.

Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal *et al.*, 2010 Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11: 446–450.

Eshed, V., A. Gopher, T. B. Gage, and I. Hershkovitz, 2004 Has the transition to agriculture reshaped the demographic structure of prehistoric populations? New evidence from the Levant. Am. J. Phys. Anthropol. 124: 315–329.

Fernandez, B., A. Garcia-Dorado, and A. Caballero, 2004 Analysis of the estimators of the average coefficient of dominance of deleterious mutations. Genetics 168: 1053–1069.

Fu, W., T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis *et al.*, 2012 Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493: 216–220.

Gage, T. B., and S. DeWitte, 2009 What do we know about the agricultural demographic transition? Curr. Anthropol. 50: 649–655.

Galvan, A., J. P. Ioannidis, and T. A. Dragani, 2010 Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. Trends Genet. 26: 132–141.

Gusev, A., E. E. Kenny, J. K. Lowe, J. Salit, R. Saxena *et al.*, 2011 DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. Am. J. Hum. Genet. 88: 706–717.

Hawks, J., E. T. Wang, G. M. Cochran, H. C. Harpending, and R. K. Moyzis, 2007 Recent acceleration of human adaptive evolution. Proc. Natl. Acad. Sci. USA 104: 20753–20758.

Hernandez, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24: 2786–2787.

Keightley, P. D., 2012 Rates and fitness consequences of new mutations in humans. Genetics 190: 295–304.

Keightley, P. D., L. Eory, D. L. Halligan, and M. Kirkpatrick, 2011 Inference of mutation parameters and selective constraint in mammalian coding sequences by approximate Bayesian computation. Genetics 187: 1153–1161.

Keinan, A., and A. G. Clark, 2012 Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740–743.

Keinan, A., J. C. Mullikin, N. Patterson, and D. Reich, 2007 Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat. Genet. 39: 1251–1255.

Kiezun, A., S. L. Pulit, L. C. Francioli, F. van Dijk, M. Swertz *et al.*, 2013 Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. PLoS Genet. 9: e1003301.

Kimura, M., 1955 Stochastic process and distribution of genes frequencies under natural selection. Cold Spring Harb. Symp. Quant. Biol. 20: 33–53.

Kimura, M., 1957 Some problems of stochastic processes in genetics. Ann. Math. Stat. 28: 882–901.

Kimura, M., and T. Ohta, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics 61: 763–771.

Kondrashov, A. S., 1988 Deleterious mutations and the evolution of sexual reproduction. Nature 336: 435–440.

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon *et al.*, 2010 Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838.

Lesecque, Y., P. D. Keightley, and A. Eyre-Walker, 2012 A resolution of the mutation load paradox in humans. Genetics 191: 1321–1330.
Lohmueller, K. E., A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez *et al.*, 2008 Proportionally more deleterious genetic variation in European than in African populations. Nature 451: 994–997.
Lynch, M., 2010 Rate, molecular spectrum, and consequences of human mutation. Proc. Natl. Acad. Sci. USA 107: 961–968.
Maher, B., 2008 Personal genomes: the case of the missing heritability. Nature 456: 18–21.
Maher, M. C., L. H. Uricchio, D. G. Torgerson, and R. D. Hernandez, 2012 Population genetics of rare variants and complex diseases. Hum. Hered. 74: 118–128.
Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff *et al.*, 2009 Finding the missing heritability of complex diseases. Nature 461: 747–753.
McClellan, J., and M. C. King, 2010 Genetic heterogeneity in human disease. Cell 141: 210–217.
Muruyama, T., 1974 The age of a rare mutant in a large population. Am. J. Hum. Genet. 26: 669–673.
Muruyama, T., and M. Kimura, 1974 A note on the speed of gene frequency changes in reverse directions on a finite population. Evolution 28: 161–163.
Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean *et al.*, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337: 100–104.
Phadnis, N., and J. D. Fry, 2005 Widespread correlations between dominance and homozygous effects of mutations: implications for theories of dominance. Genetics 171: 385–392.
Przeworski, M., B. Charlesworth, and J. D. Wall, 1999 Genealogies and weak purifying selection. Mol. Biol. Evol. 16: 246–252.
Ravanbod, S., M. Rassoulzadegan, G. Rastegar-Lari, M. Jazebi, S. Enayat *et al.*, 2012 Identification of 123 previously unreported mutations in the F8 gene of Iranian patients with haemophilia A. Haemophilia 18: e340–e346.
Reed, F. A., and C. F. Aquadro, 2006 Mutation, selection and the future of human evolution. Trends Genet. 22: 479–484.
Simmons, M. J., and J. F. Crow, 1977 Mutations affecting fitness in Drosophila populations. Annu. Rev. Genet. 11: 49–78.
Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129: 555–562.
Szpiech, Z. A., J. Xu, T. J. Pemberton, W. Peng, S. Zollner *et al.*, 2013 Long runs of homozygosity are enriched for deleterious variation. Am. J. Hum. Genet. 93: 90–102.
Tajima, F., 1989 The effect of change in population size on DNA polymorphism. Genetics 123: 597–601.
Tennessen, J. A., A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337: 64–69.
United Nations, Department of Economic and Social Affairs, Population Division, 2011 World Population Prospects: The 2010 Revision, Volume I: Comprehensive Tables. ST/ESA/SER.A/313. United Nations, New York.
Wakeley, J., L. King, B. S. Low, and S. Ramachandran, 2012 Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. Genetics 190: 1433–1445.
Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.
Waxman, D., 2011 A unified treatment of the probability of fixation when population size and the strength of selection change over time. Genetics 188: 907–913.
Waxman, D., 2012 Population growth enhances the mean fixation time of neutral mutations and the persistence of neutral variation. Genetics 191: 561–577.
Yu, F., A. Keinan, H. Chen, R. J. Ferland, R. S. Hill *et al.*, 2009 Detecting natural selection by empirical comparison to random regions of the genome. Hum Mol Genet 18: 4853–4867.
Zhuang, Z., A. Gusev, J. Cho, and I. Pe'er, 2012 Detecting identity by descent and homozygosity mapping in whole-exome sequencing data. PLoS ONE 7: e47618.

Communicating editor: N. A. Rosenberg

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.153973/-/DC1

Population Growth Inflates the Per-Individual Number of Deleterious Mutations and Reduces Their Mean Effect

Elodie Gazave, Diana Chang, Andrew G. Clark, and Alon Keinan

DOI: 10.1534/genetics.113.153973

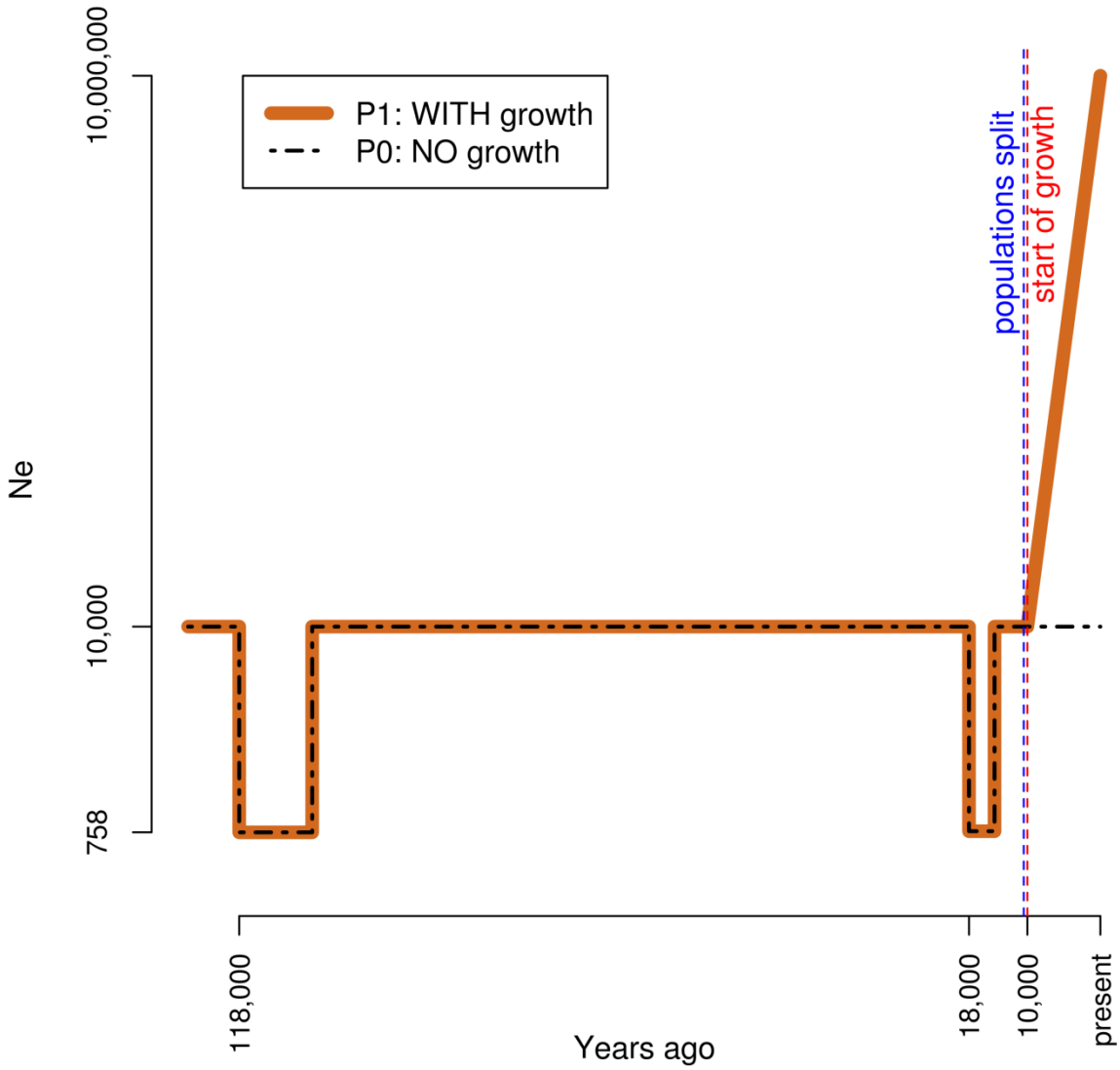


Figure S1 Main demographic models used. The y-axis represents Ne at different epochs (x-axis) on a log scale. Both populations (with growth, no growth) have an ancestral population size of N_e = 10,000. They undergo a first bottleneck of intensity F = 0.264 at 4,720 generations ago followed by a quick recovery to the ancestral population size. The second bottleneck occurs 720 generations ago with an intensity of F = 0.09, also followed by a quick recovery to the ancestral population size. Both populations share a common ancient history until 420 generations ago where a copy (P1) of the ancestral population (P0) is created. Then, 400 generations before present, P1 starts growing at a rate of 1.73% per generation, reaching a final size of N_e = 10,000,000 at the end of the simulation. In contrast, P0 continues to evolve maintaining its size constant at N_e = 10,000 until present.

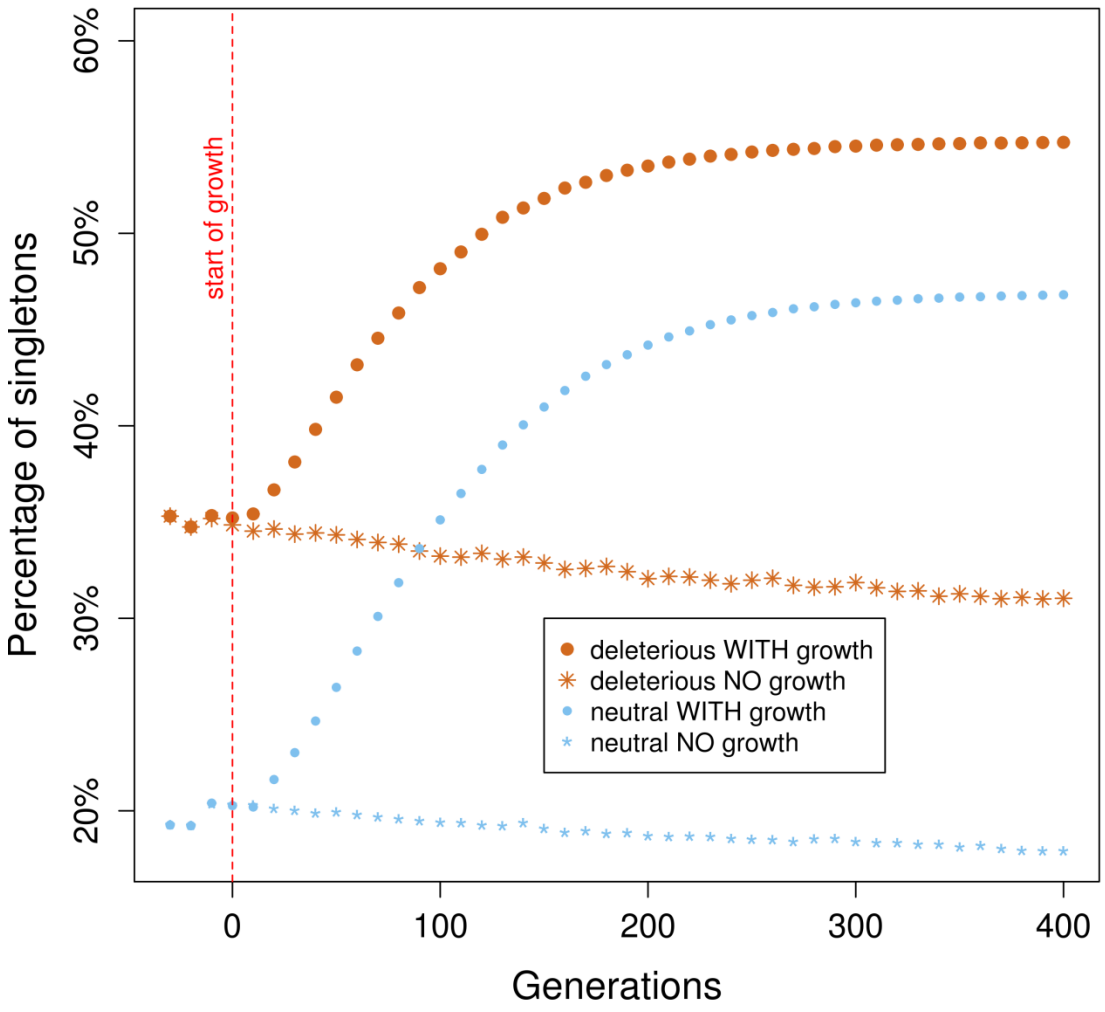


Figure S2 The percentage of singletons increases in a growing population. The percentage of singletons (sites with derived allele count, DAC = 1) out of all the segregating sites is shown for the last 440 generations of the simulation, for loci with either deleterious or neutral mutations, for both population models (NO growth, WITH growth).

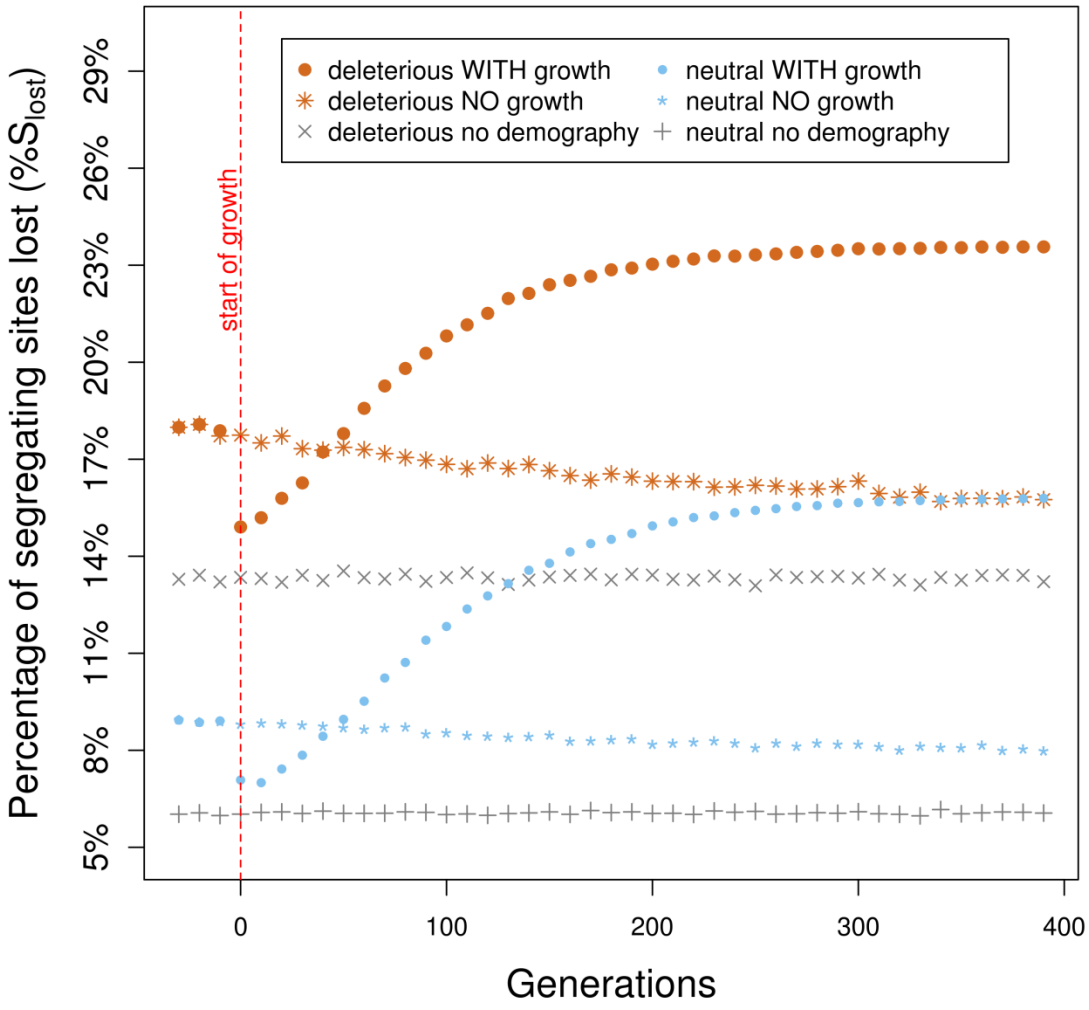


Figure S3 The percentage of segregating sites lost at each generation decreases in population without growth while it is constant in a population without any demography. The growing and constant populations data are the same as presented in Figure 1b. The population with no demography has a constant effective population size of 10,000 throughout history without any demographic events, thus differing from the constant size model by the absence of ancestral bottlenecks. The comparison of the constant size population and the population with no demography (gray) allows assessing the effect of past demographic events (bottlenecks). The percentage of segregating sites lost $\%S_{lost}$ at each generation is slightly decreasing in the population without growth, but not in the scenario without demographic event. This comparison reveals that the population without growth, although evolving at constant size for the last 580 generations has not recovered its mutation-drift balance that was altered during the ancient bottlenecks.

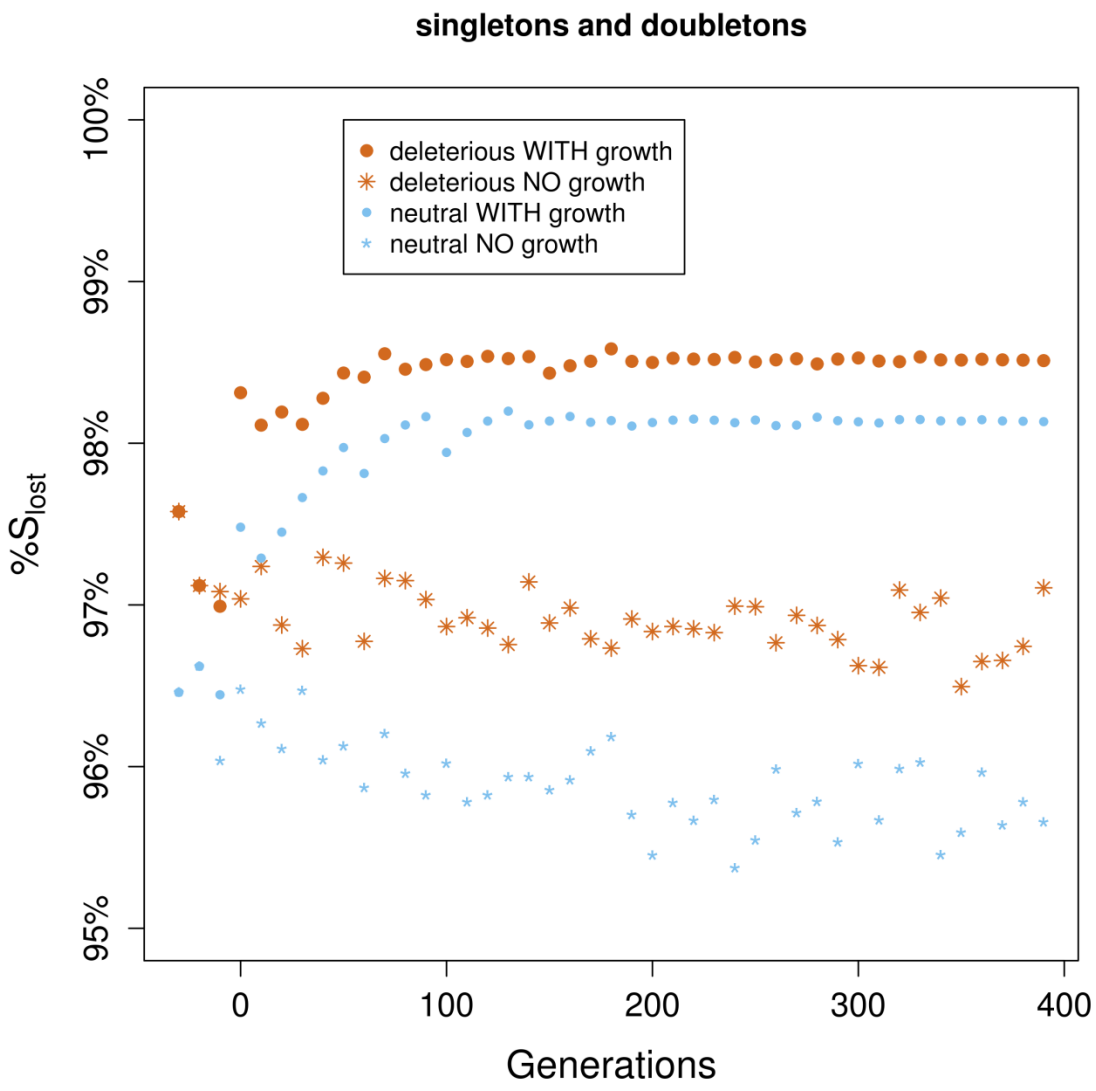


Figure S4 The contribution of singletons and doubletons to lost sites increases in a growing population. Considering all lost segregating sites, we counted the percentage that have a certain count of derived alleles (DAC = 1, DAC = 2, DAC = 3 etc). These percentages for all possible DAC (from 1 to 2 N_e - 1) sum up to 1. In the growing scenario, singletons and doubletons represent over 98% of the derived alleles lost. For higher DAC (>= 3), the contribution of segregating sites with higher DAC decreases quickly as DAC increases. For example, all sites with DAC>10 combined only represent between 0 and 0.00016% of the sites that are lost at a deleterious locus in a growing population, depending on the generation considered (data not shown).

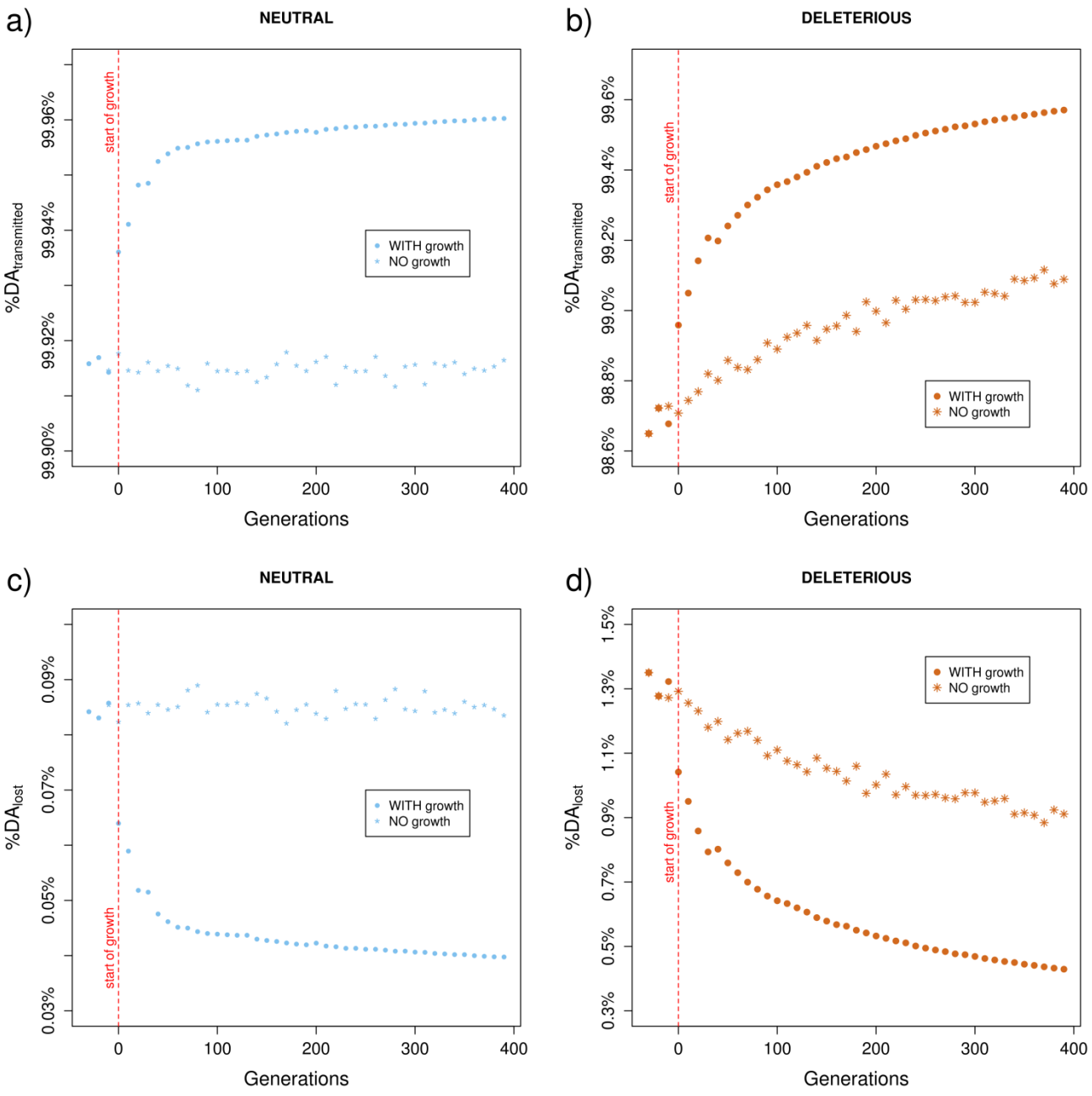


Figure S5 The fraction of derived alleles lost under population expansion is higher at a deleterious locus than at a neutral locus. The fraction of derived alleles lost, $\%DA_{lost}$, is defined as the sum of the number of copies of derived alleles at the segregating sites lost, over the sum of the number of copies of derived alleles present at all segregating sites present in the population. The fraction of derived alleles transmitted, $\%DA_{transmitted}$, is 1-$\%DA_{lost}$. The $\%DA_{transmitted}$ (a, b) and the $\%DA_{lost}$ (c, d) are shown for each locus type, neutral (a, c) or deleterious (b, d). The growing populations lose a higher percentage of derived alleles at the deleterious locus than at the neutral locus.

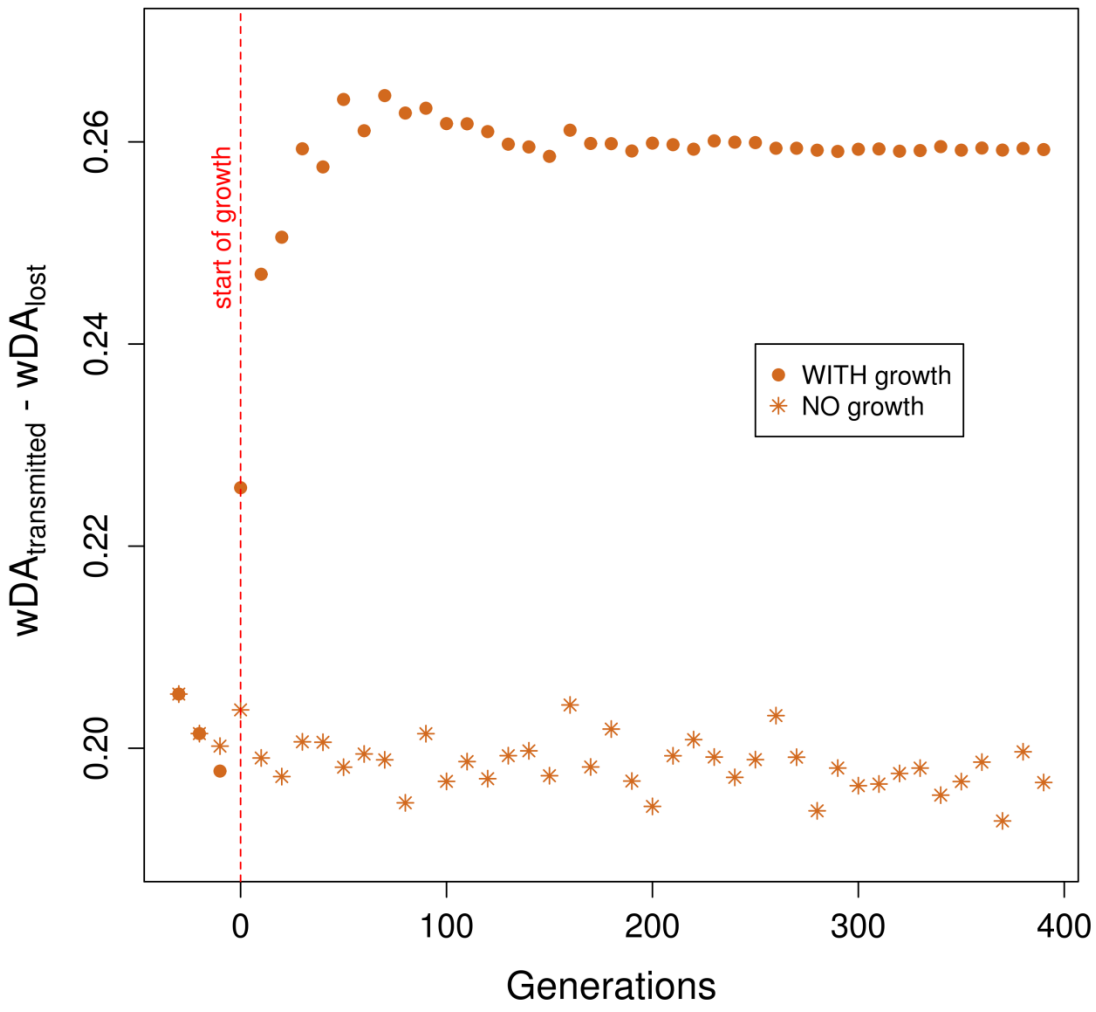


Figure S6 The difference in fitness effect between lost and transmitted derived alleles increases in a growing population. The average fitness effect (*wDA*) is defined as in Figure 4. The difference in fitness effect is the difference between *wDA* at transmitted sites and *wDA* at lost sites ($wDA_{transmitted}$ - wDA_{lost}). All the mutations are deleterious and have a negative fitness effect. Positive values on the y-axis show that the average fitness at transmitted sites is less deleterious than at lost sites. The difference $wDA_{transmitted}$ - wDA_{lost} is higher in the growing population, showing that the fraction of alleles that are transmitted have on average a better fitness effect than the ones that are lost in this demographic scenario compared to a scenario without growth.

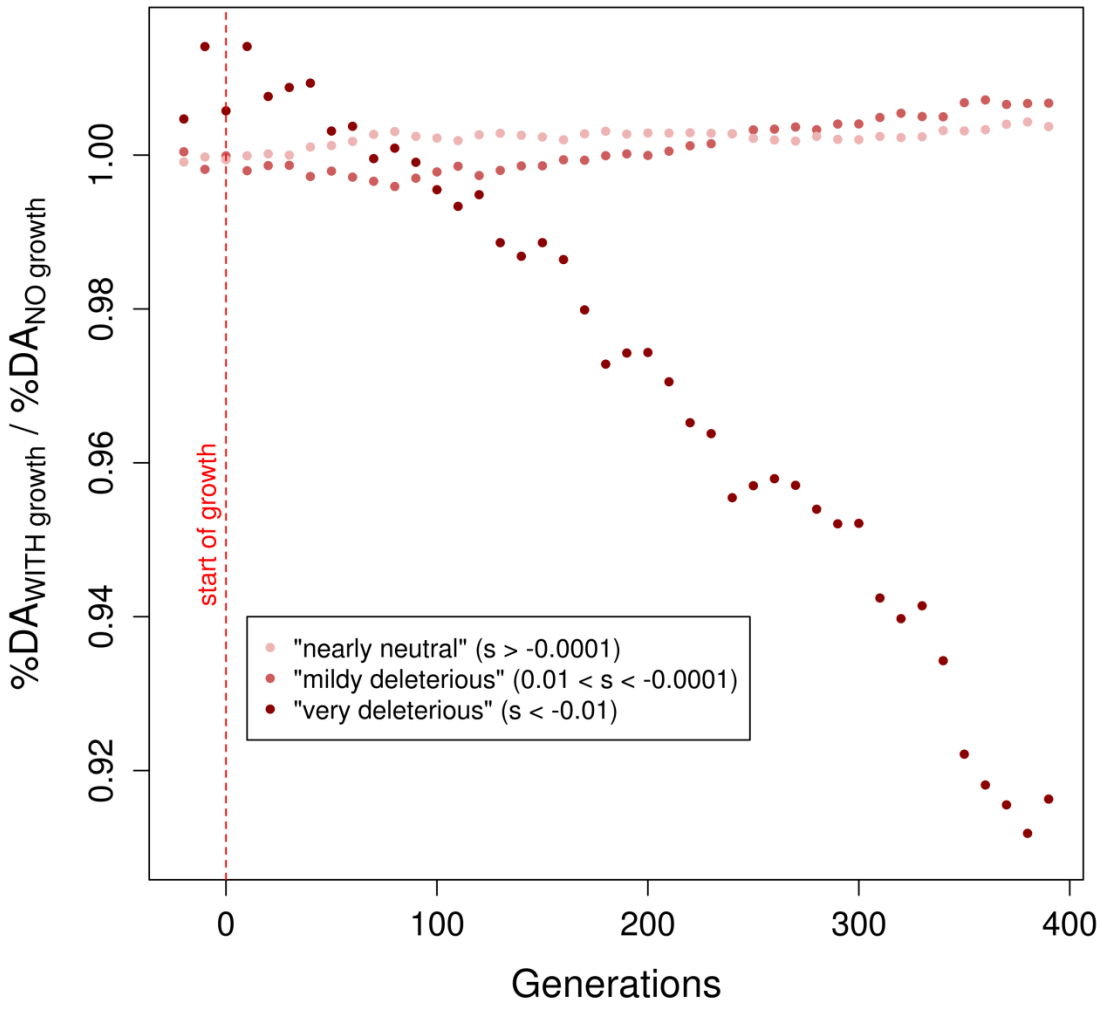


Figure S7 The percentage of derived alleles in the most deleterious category of fitness effect decreases in the growing population. The percentage of derived alleles (*%DA*) and the category of fitness effect are defined as in Figure 3. The data are presented as the ratio of *%DA* in the growing population over the *%DA* of the population of constant size. The ratio below 1 in the most deleterious category indicates that a lower percentage of derived alleles have a fitness effect < -0.01 in the scenario with growth.

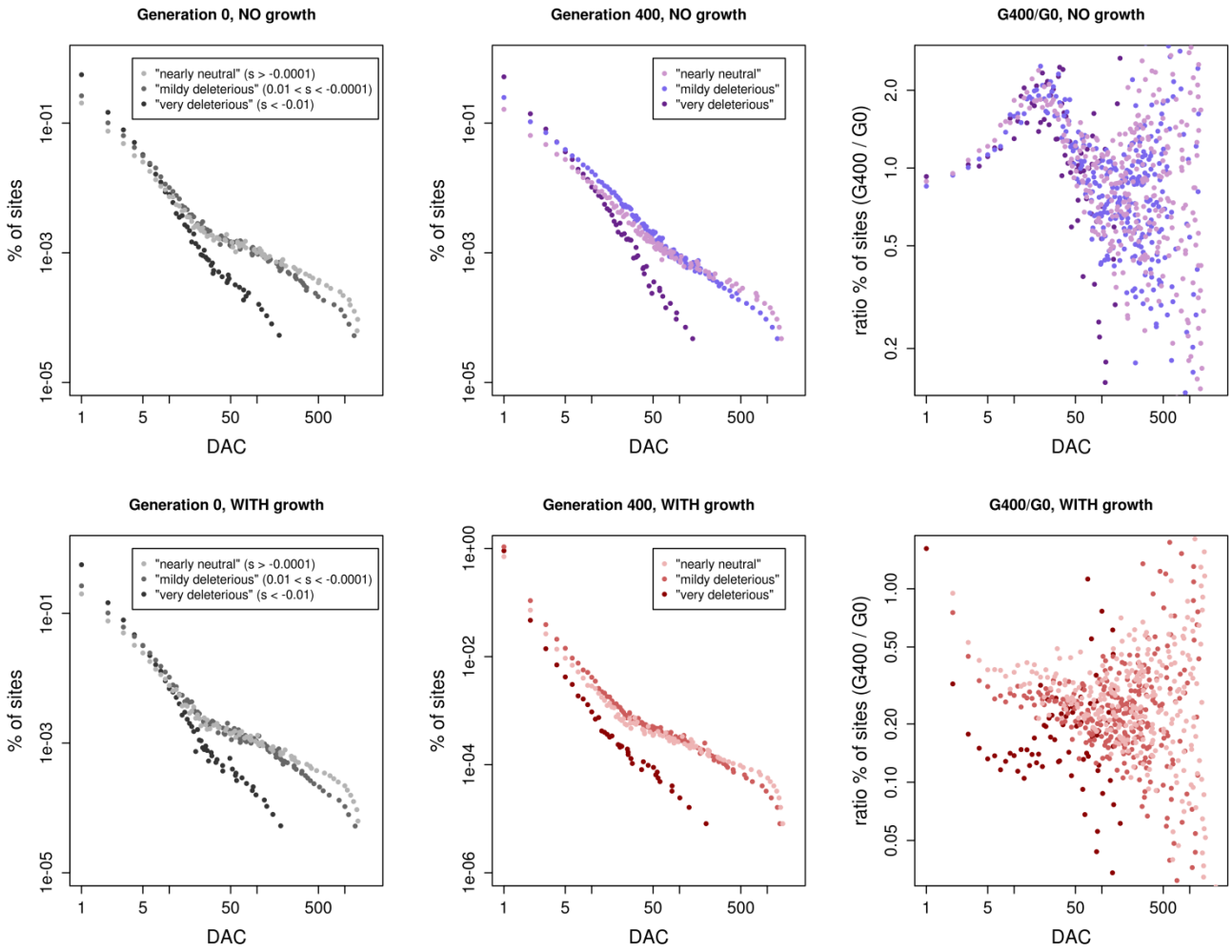


Figure S8 Population growth mainly affects the most deleterious allele category. The site frequency spectrum (SFS) of the population is shown before (a) and after 400 generations growth (b), for three different categories of fitness effect. During the same interval of time, we also show the three same SFS for a population that has not experienced the growth (d,e). Data are presented on a log-log scale. Generation 400 represents present time and generation 0 represents the time at which one population starts growing. Panels (c) and (e) show the ratio of the proportion of sites for each derived allele count (DAC) at time 400 over time 0. For DAC greater than 20, data are noisy because counts are low in each DAC. We observe that population growth has induced a skew of the SFS toward rare variants (ratio below 1). In addition, this skew is more accentuated for the alleles in the most deleterious category. In contrast, the SFS for the three categories of fitness have changed almost identically after 400 generation in the population without growth.

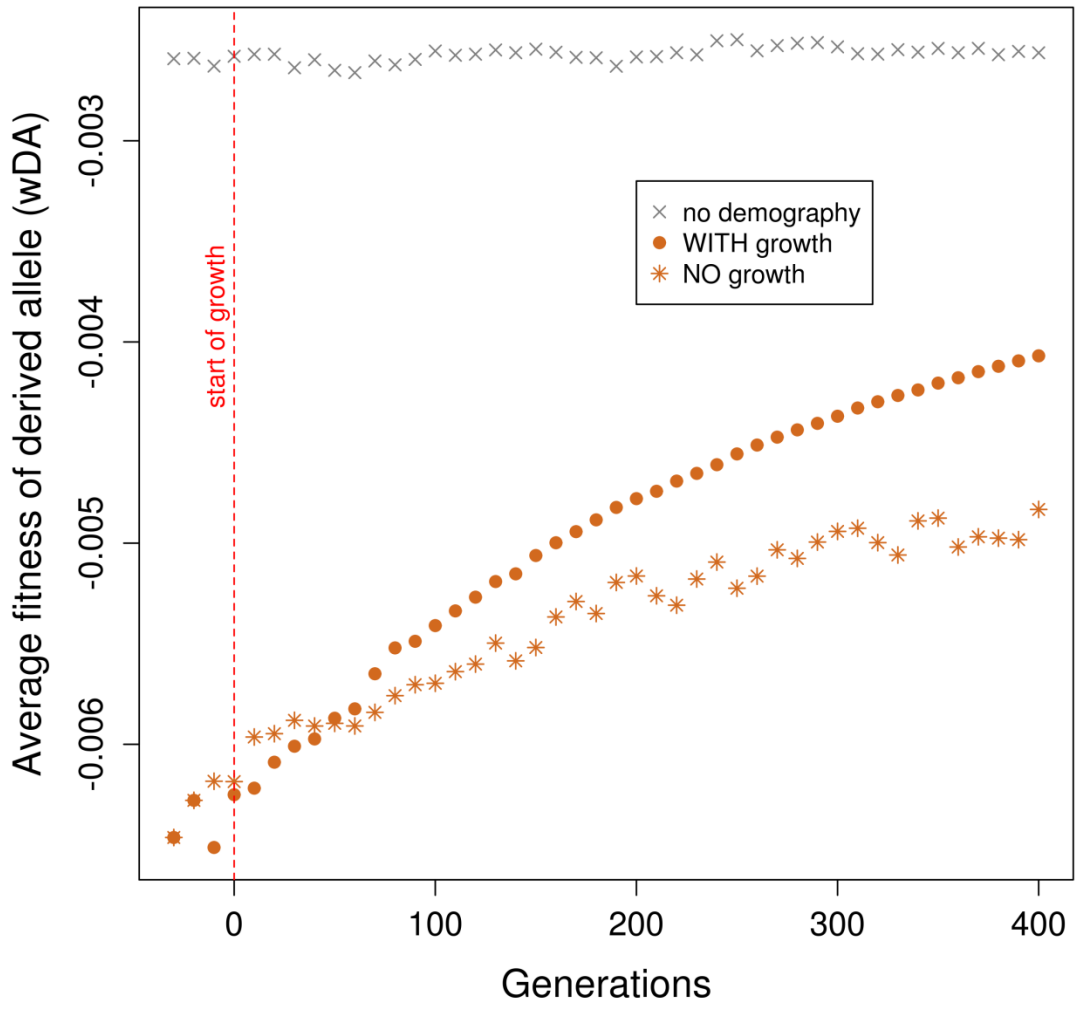


Figure S9 The mean fitness effect of the derived alleles in both demographic scenarios is lower than in a population without any demography. The growing and constant populations data are the same as presented in Figure 4. The population with no demography has a constant effective population size of 10,000 throughout history without any demographic events, thus differing from the constant size model by the absence of ancestral bottlenecks. The comparison of the constant size population and the population with no demography (gray) allows assessing the effect of past demographic events (bottlenecks). The average fitness of derived alleles in both populations with demographic events is lower than in the population with no demography. The increase in average fitness effect of the constant population size, although slower than the increase in the growing population, is explained by the progressive elimination of copies of derived alleles with low fitness effect that accumulated during the ancient bottlenecks. This effect is shown in empirical data in Lohmueller *et al.* (2008).

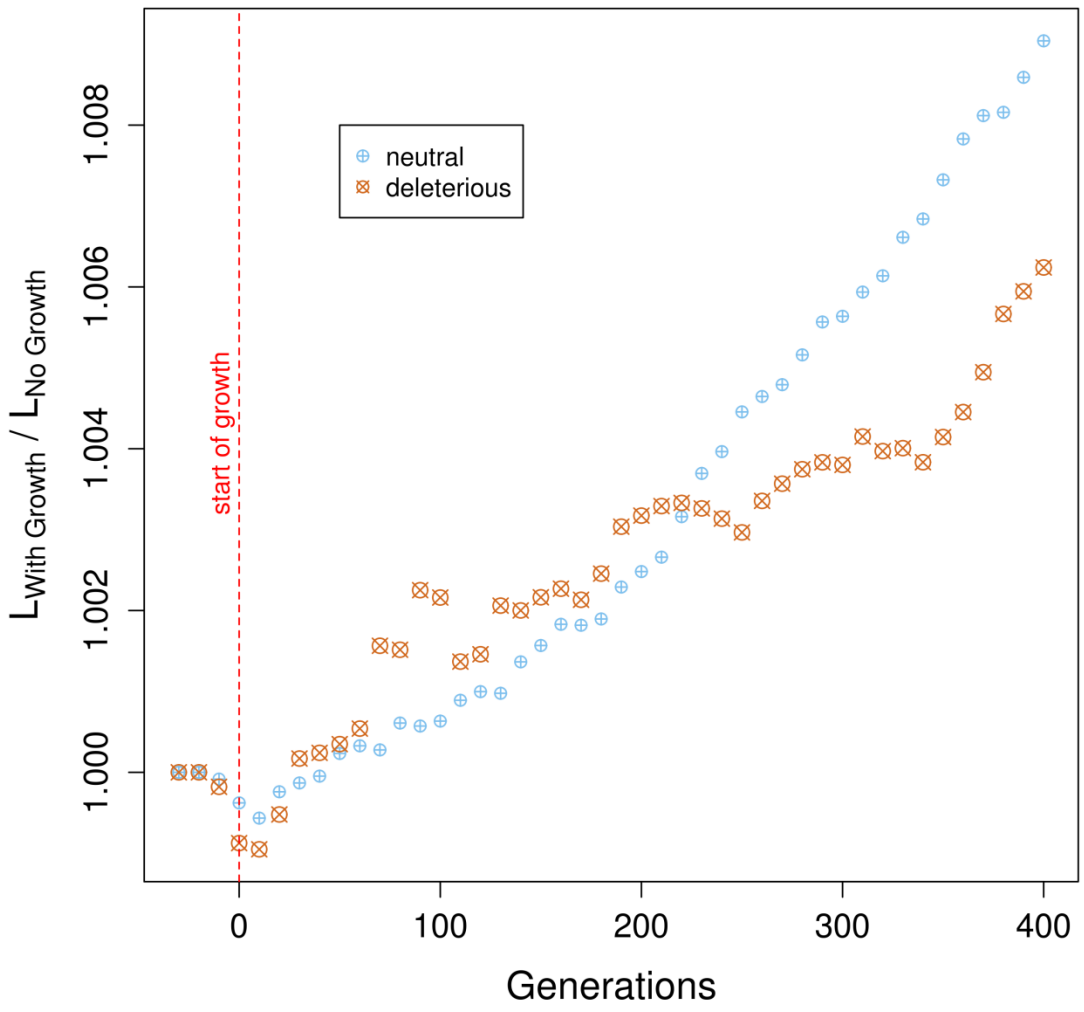


Figure S10 The number of mutations per chromosome is higher in a growing population. The number of mutations per chromosome (*L*) is defined as in Figure 5. Data shown is the ratio of *L* in a population with growth over *L* in a population without growth. Ratios above 1 indicate that *L* is higher under population expansion.

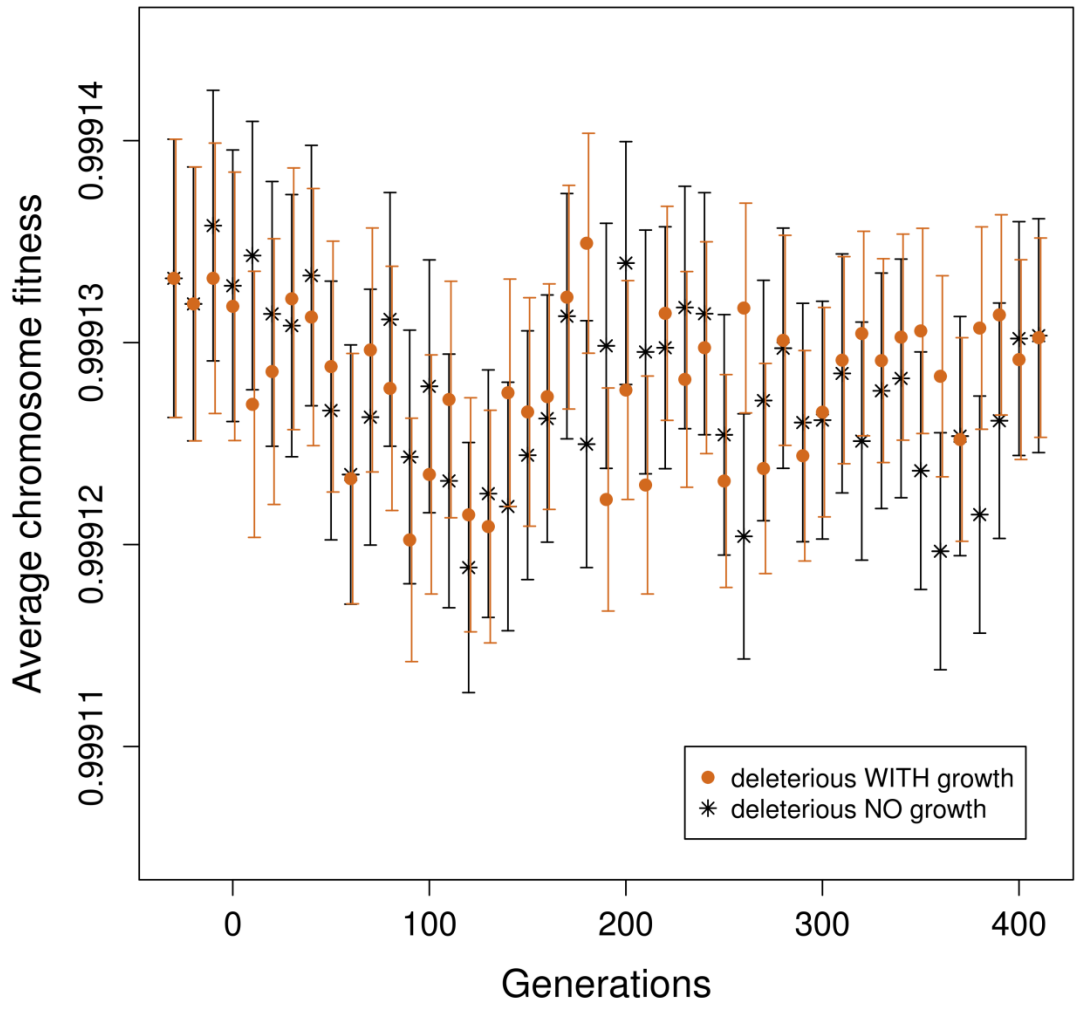


Figure S11 The average fitness of individuals is not different in the two demographic scenarios. The fitness of an individual chromosome relative to a fitness of 1 is defined as the product of selective coefficient of all mutations it carries. The average chromosome fitness, w_{CHR}, is defined as $\sum_{k=1}^{2Ne}\prod_{i=1}^{S}(1+s_{ik})\ /2Ne$, where s_{ik} is the selection coefficient of the derived allele at site i on chromosome *k*. In a growing population, individual chromosomes carry a larger number of mutations, but each one of them has on average a less deleterious effect, while in the population without growth an individual chromosome carries fewer mutations of larger fitness effect. Vertical bars represent the standard errors of the mean over 10,000 replicates.

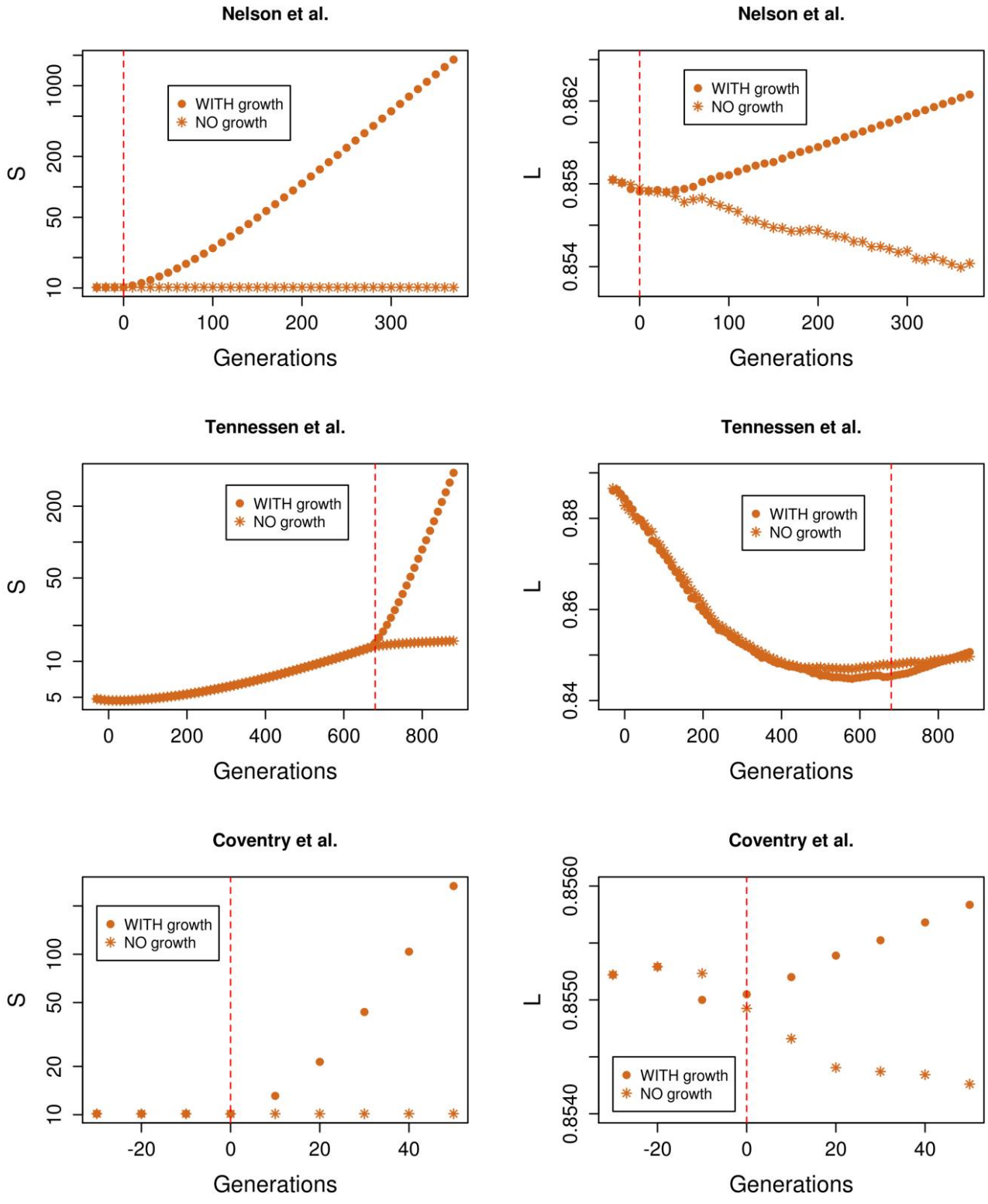


Figure S12 The number of segregating sites and the number of mutations per chromosome increase in growing populations assuming different models of European history. *S* and *L* are defined as in Figure 1 and Figure 5, respectively. Data are shown for a locus with deleterious mutations.

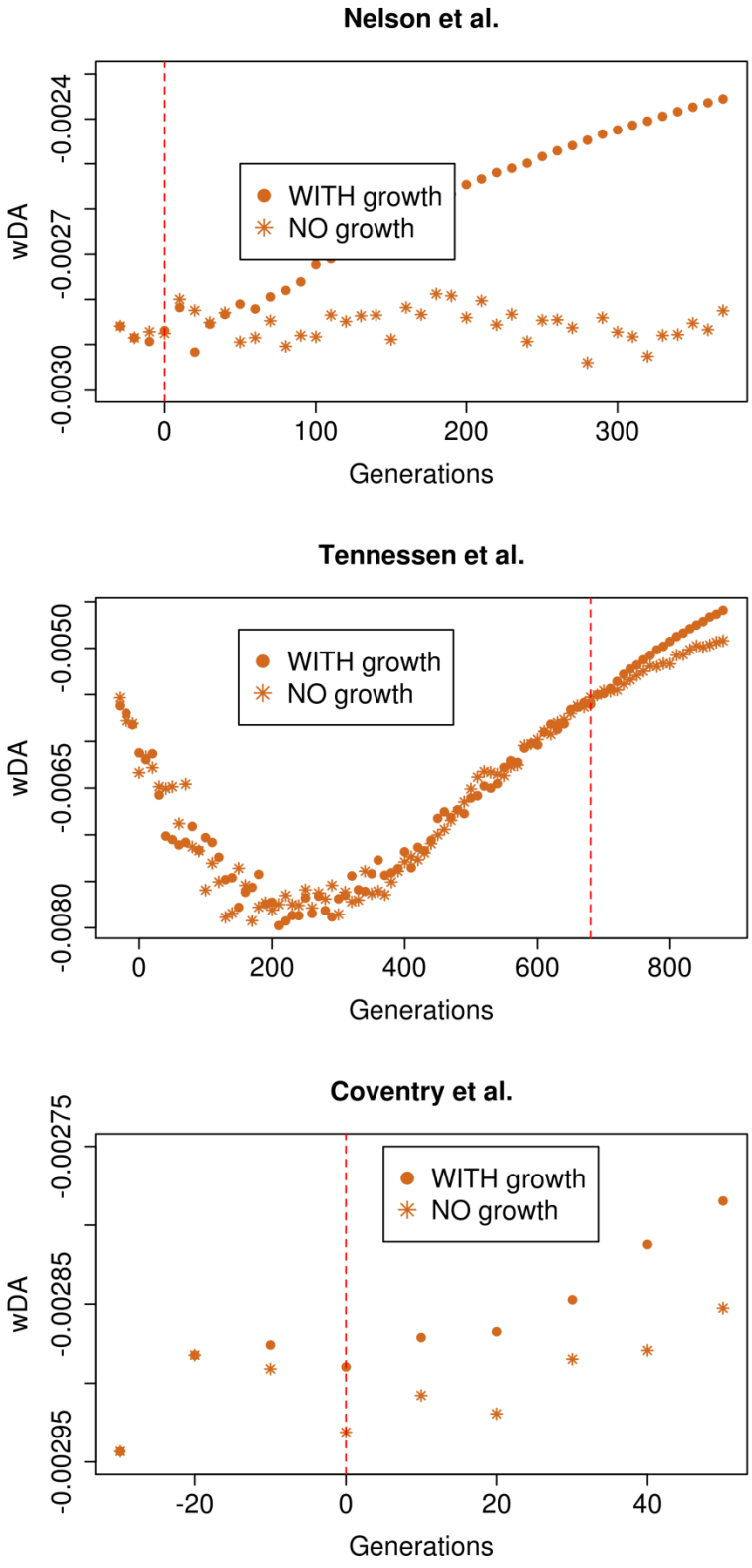


Figure S13 The average fitness effect of the alleles present in the population increases in growing populations assuming different models of European history. *wDA* is defined as in Figure 4.

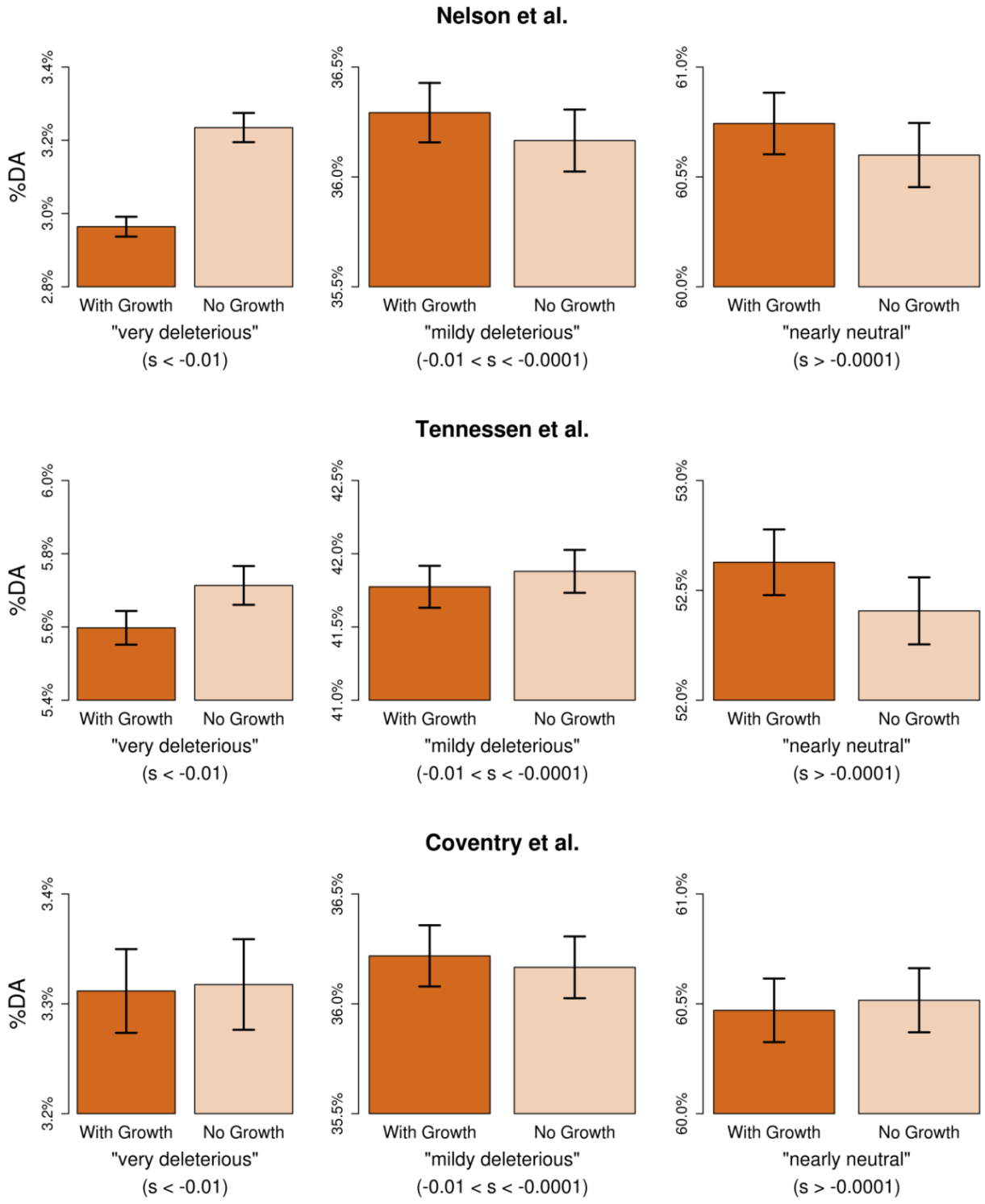


Figure S14 The percentage of sites in the most deleterious category of fitness decreases in growing populations assuming different models of European history. *%DA* and categories of fitness effect are defined as in Figure 3 and Figure S7. The growing population has a lower percentage of very deleterious mutations in all models but the one of Coventry *et al.*, where growth is extremely recent (56 generations) and its effect on very deleterious alleles is not significant.

File S1

Additional Results

Derived Allele Count (DAC) rather than allele frequency determines the probability of loss or transmission

In absence of selection, the probability for a segregating site of DAC = n to be lost or transmitted to the next generation follows a binomial process, where the probability of success (transmission) is $\frac{n}{2N_e}$ and the number of trials is $2N_e$ (where $2N_e$ is the number of chromosomes in the population), assuming the same number of individuals in the next generation.

To simplify, let us consider segregating sites with DAC = 1 (singletons). Each derived allele with DAC = 1 in a population of size N_e =10,000,000 has a frequency of only 1/10,000,000, which is much smaller than their frequency in a population of size N_e = 10,000. For each singleton taken individually, the probability of transmission to the next generation (probability of success in the binomial process) is lower in the larger population. However, the number of opportunities to be transmitted to the next generation (number of trials in the binomial process) increases, as more chromosomes are drawn from the current generation to participate to the next one. Consequently, the probability for one singleton to be lost is the same in the smaller population (N_e = 10,000) and in the larger population (N_e = 10,000,000). To see why this is so, consider that the probability of drawing zero copies from the binomial is Pr(X=0) = $[(N\text{-}1)/N]^N$. Using a normal approximation to the binomial distribution, we see that this value converges to 0.368 as N grows. This shows that for a segregating site, the probability of loss or transmission does not depend on the frequency of the derived allele but rather on its DAC. The same reasoning can be applied equally to other values of DAC. It also applies to a scenario of population growth, where the probability of losing a site still depends on its DAC rather than its frequency, and the number of trials is the population size at the next generation.

Effect of bottlenecks on the average fitness effect of a copy of derived allele

One additional population model was simulated (10,000 replicates) in order to test the effect of the bottlenecks in the two demographic scenarios studied here (*i.e.* the model of European history with growth and the model without growth). This additional population model has a constant effective size of 10,000 throughout history, with no demographic events at all. We refer to this population as the "no demography" population.

Both in the "no growth" and "with growth" models, the average selection coefficient of a copy of derived allele is much lower than in the model with no demography (Figure S9). This shows that the ancestral bottlenecks have reduced the average fitness of the

SNPs present in the population. Strong population contraction episodes can induce a drop on fitness through the random increase in frequency of deleterious mutations.

Command lines used to simulate the main demographic scenario

- Main model, with deleterious mutations. P0 and P1 split 420 generations ago, 20 generations before the beginning of the growth. During the last 400 generations, P1 grows to 10,000,000 and P0 remains at the constant size of 10,000.

./sfs_code 2 1 -N 1000 -t 0.0008 -n 1000 -L 1 5000 -a N -W 2 0 1 0.2 0.206 0.000365 -Td 0 P 0 0.0758 -Td 0.02 P 0 13.2 -Td 0.2 P 0 0.0769 -Td 0.207 P 0 13 -TS 0.215 0 1 -Tg 0.216 P 1 345.37501 -TE 0.236 --noSeq --trackTrajectory T 0.214 F output -o out

- Main model, with neutral mutations. P0 and P1 split 420 generations ago, 20 generations before the beginning of the growth. During the last 400 generations, P1 grows to 10,000,000 and P0 remains at the constant size of 10,000.

./sfs_code 2 1 -N 1000 -t 0.0008 -n 1000 -L 1 5000 -a N -W 0 -Td 0 P 0 0.0758 -Td 0.02 P 0 13.2 -Td 0.2 P 0 0.0769 -Td 0.207 P 0 13 -TS 0.215 0 1 -Tg 0.216 P 1 345.37501 -TE 0.236 --noSeq --trackTrajectory T 0.214 F output -o out

Additional demographic scenario

To test the robustness of our results to the demographic scenario assumed in the main text (main model), we repeated the analysis using 3 other published demographic scenarios involving a European model of ancient history and a final epoch of rapid exponential growth (Coventry et al., 2010; Nelson et al., 2012; Tennessen et al., 2012). Both the model of Coventry et al. (2010) and the model of Nelson et al. (2012) use the ancient demographic history described in Schaffner et al. (2005). The model of Tennessen et al. (2012) is based on the ancient demography of European history described in Gravel et al. (2011). The migration parameter of Schaffner et al. (2005) and Gravel et al. (2011) was ignored here, as done in the models by Tennessen et al. (2012), Nelson et al. (2012) and Coventry et al. (2010). We repeated the analysis applied to the main model, but only for a locus with deleterious mutations. The distribution of selective coefficients in the additional models was set to have the same mean (-0.28) as in the main model. Also identical to the main model, the populations with or without growth emerge from the split of a common ancient population to insure identical states when the comparison between growth and no growth starts (see main text). The populations with and without growth are followed for the whole duration of the growth, which varied according to the model. In all three additional models the final N_e is smaller

than in the main model, and therefore fewer mutations are outputted. To match the main model's level of information, we increased the number of replicates in the additional models to 50,000 replicates. The arguments used in SFS_code are the following:

- Nelson et al. (2012) model, with deleterious mutations:

./sfs_code 2 1 -N 1250 -t 0.0008 -n 1000 -L 1 5000 -a N -W 2 0 1 0.2 0.206 0.000292 -TN 0 2400 -TN 0.54 770 -Td 0.54004 0.07639 -Td 0.54404 13.1 -Td 0.6 0.3247 -Td 0.604 3.08 -TS 0.6648 0 1 -Tg 0.6652 P 1 422.441 -TE 0.68 --noSeq --trackTrajectory T 0.6636 F output -o out'

In this model, the final population sizes are N_e = 4,000,020 for the growing population and N_e = 7,700 for the population without growth.

- Tennessen et al. (2012) model, with deleterious mutations:

./sfs_code 2 1 -N 740 -t 0.0008 -n 1000 -L 1 5000 -a N -W 2 0 1 0.2 0.206 0.0004931 -TN 0 1447 -TN 0.2622 186 -TN 0.3378 104 -TS 0.3379 0 1 -Tg 0.3379 45.47 -Tg 0.3861 P 0 0 -Tg 0.3861 P 1 283.16 -TE 0.40 --noSeq --trackTrajectory T 0.3379 F output -o out

The model of Tennessen et al. (2012) uses two epochs of growth, a first one that is slow and a second one that is fast. The split occurs before the first growth epoch but the two populations grow at the same rate and at the end of the first epoch of growth, both populations have N_e = 9,210. Only one population undergoes the second (fast) epoch of growth and grows until N_e = 512,010, while the other one stops growing and remains at 9,210 individuals until the end of the simulation.

- Coventry et al. (2010) model, with deleterious mutations:

./sfs_code 2 1 -N 1250 -t 0.0008 -n 1000 -L 1 5000 -a N -W 2 0 1 0.2 0.206 0.000292 -TN 0 2400 -TN 0.54 770 -Td 0.54004 0.07639 -Td 0.54404 13.1 -Td 0.6 0.3247 -Td 0.604 3.08 -TS 0.6774 0 1 -Tg 0.678 P 1 2480.58 -TE 0.68 --noSeq --trackTrajectory T 0.6764 F output -o out

In this model, the final Ne of the population with growth is 1,100,020, while the population without growth remains as N_e = 7,700.

Time to the most recent common ancestor (t_{MRCA}) for a sample of two chromosome (n = 2)

- For a model without growth:

$$\Pr(0 < t < t_1) = \int_0^{t_1} \frac{1}{2N_a} e^{\frac{-t}{2N_a}} dt$$

$$\Pr(t_1 < t < t_2) = [1 - \Pr(0 < t < t_1)] \int_{t_1}^{t_2} \frac{1}{2N_a} e^{\frac{-(t-t_1)}{2N_a}} dt$$

$$\Pr(t_2 < t < t_3) = [1 - \Pr(0 < t < t_2)] \int_{t_2}^{t_3} \frac{1}{2N_{b_2}} e^{\frac{-(t-t_2)}{2N_{b_2}}} dt$$

$$\Pr(t_3 < t < t_4) = [1 - \Pr(0 < t < t_3)] \int_{t_3}^{t_4} \frac{1}{2N_a} e^{\frac{-(t-t_3)}{2N_a}} dt$$

$$\Pr(t_4 < t < t_5) = [1 - \Pr(0 < t < t_4)] \int_{t_4}^{t_5} \frac{1}{2N_{b_1}} e^{\frac{-(t-t_4)}{2N_{b_1}}} dt$$

$$E[t_{MRCA}(2)] = \int_0^{t_1} \frac{1}{2N_a} e^{\frac{-t}{2N_a}} t dt + [1 - \Pr(0 < t < t_1)] \int_{t_1}^{t_2} \frac{1}{2N_a} e^{\frac{-(t-t_1)}{2N_a}} t dt + [1 - \Pr(0 < t < t_2)] \int_{t_2}^{t_3} \frac{1}{2N_{b_2}} e^{\frac{-(t-t_2)}{2N_{b_2}}} t dt$$
$$+ [1 - \Pr(0 < t < t_3)] \int_{t_3}^{t_4} \frac{1}{2N_a} e^{\frac{-(t-t_3)}{2N_a}} t dt + [1 - \Pr(0 < t < t_4)] \int_{t_4}^{t_5} \frac{1}{2N_{b_1}} e^{\frac{-(t-t_4)}{2N_{b_1}}} t dt$$
$$+ [1 - \Pr(0 < t < t_5)] \int_{t_5}^{\infty} \frac{1}{2N_a} e^{\frac{-(t-t_5)}{2N_a}} t dt$$

- For a model with growth:

$$P(0 < t < t_1) = \int_0^{t_1} L(t) e^{-G(t)} dt$$

Where: $L(t) = \frac{1}{2N_a e^{g(t_1 - t)}}$ $\quad G(t) = \int_0^t \frac{1}{L(s)} ds$ g = rate of growth (*)

$$E[t_{MRCA}(2)] = \int_0^{t_1} L(t) e^{-G(t)} t dt + [1 - \Pr(0 < t < t_1)] \int_{t_1}^{t_2} \frac{1}{2N_a} e^{\frac{-(t-t_1)}{2N_a}} t dt + [1 - \Pr(0 < t < t_2)] \int_{t_2}^{t_3} \frac{1}{2N_{b_2}} e^{\frac{-(t-t_2)}{2N_{b_2}}} t dt$$
$$+ [1 - \Pr(0 < t < t_3)] \int_{t_3}^{t_4} \frac{1}{2N_a} e^{\frac{-(t-t_3)}{2N_a}} t dt + [1 - \Pr(0 < t < t_4)] \int_{t_4}^{t_5} \frac{1}{2N_{b_1}} e^{\frac{-(t-t_4)}{2N_{b_1}}} t dt$$
$$+ [1 - \Pr(0 < t < t_5)] \int_{t_5}^{\infty} \frac{1}{2N_a} e^{\frac{-(t-t_5)}{2N_a}} t dt$$

* From Griffiths and Tavaré (1994)

With the following parameters (time measures in generations):

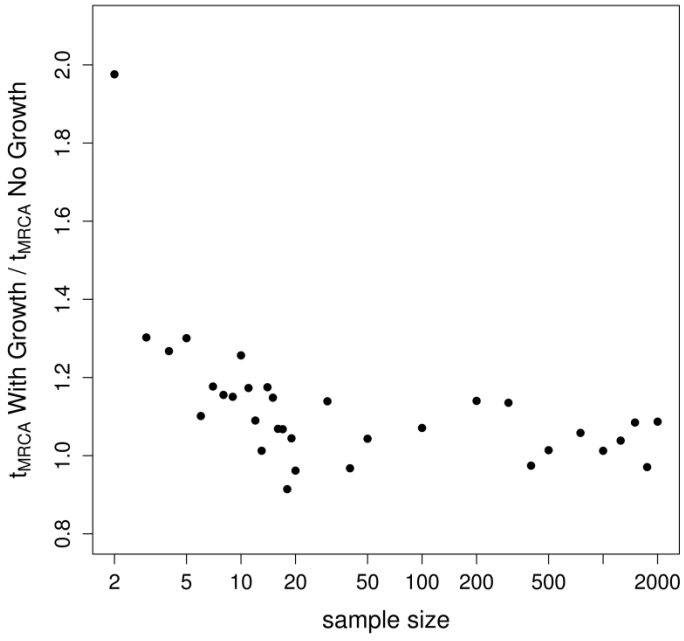
Parameter	Value
t_1	400
t_2	580
t_3	720
t_4	4320
t_5	4720
N_a	10,000
N_{b1}	757.57
N_{b2}	769.23
g	0.01726939

We obtain the following estimates for t_{MRCA} (**):

Model	Time (Generations)
E[noGrowth]	15264.76
E[Growth]	15526.00

** integrals numerically evaluated using: http://www.solvemymath.com/online_math_calculator/calculus/definite_integral/index.php

Using the two estimates calculated above, the difference in t_{MRCA} for a pair of chromosomes between the scenario with growth and without growth is 1.7%. This value is higher than the difference in the number of mutations per chromosomes (*L*) computed for the whole sample at neutral loci (Figure 5). This is because the difference in t_{MRCA} is maximal for a pair of samples and decreases with sample size. This effect is shown on the figure below: while the t_{MRCA} is expected to increase with sample size, the difference in t_{MRCA} decreases as sample size increases. The t_{MRCA} are computed over 1,000,000 replicates with the coalescent simulator ms (Hudson, 2002).



References

Coventry, A., Bull-Otterson, L.M., Liu, X., Clark, A.G., Maxwell, T.J., Crosby, J., Hixson, J.E., Rea, T.J., Muzny, D.M., Lewis, L.R., et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun 1, 131.

Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108, 11983-11988.

Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337-338.

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337, 100-104.

Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15, 1576-1583.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64-69.

Table S1 Characteristics of the mutations segregating in the population after 400 generations of growth or without growth.

	a)	Average number of copies per segregating site			b)	Average age of segregating sites (in generations)		
Bin of fitness		With Growth	No Growth	With Growth/No Growth		With Growth	No Growth	With Growth/No Growth
s<=-0.01		6.93	4.82	1.45		2.46	9.76	0.252
-0.01<s<=-0.0001		334.97	113.95	2.94		4.34	254.29	0.017
s>-0.0001		904.29	232.84	3.88		5.83	534.54	0.011

a) The number of copies at each segregating site is higher in the growing population for each category of fitness. However, the relative number of very deleterious variants compared to nearly neutral variants is much lower in the population with growth. For every copy of a very deleterious allele, there are almost 130 copies of a nearly neutral variant in the scenario with growth, while in the scenario with no growth there are 48 copies of a nearly neutral variant for each copy of a very deleterious variant.

b) In each category of fitness, the average age of the variants in the population with growth is more recent in all bins of fitness because the majority of the mutations segregating in the population have been introduced recently. Importantly, although the mutation rate is the same in both scenarios (meaning that the average fitness of the mutations introduced at each generation is the same in both scenarios), the average age of variants in the most deleterious category is younger in the model with growth because very deleterious mutations are eliminated more rapidly and more efficiently, as shown on figure S7.