Declines in Neanderthal and Denisovan Population Sizes are Probably not an Artifact of Subdivision

Alan R. Rogers

19 May 2023

Abstract

In a subdivided population, effective population size (N_e) reflects not only the number of individuals, but also the rate and pattern of gene flow. If we sample two haploid genomes from the same deme, estimates of N_e will be small in the recent past but large in the distant past, even if there were no changes in census size or in the rate or pattern of gene flow. This article asks whether this effect can plausibly account for observed declines in effective sizes of archaic populations.

Archaic human populations (Neanderthals and Denisovans) had very low heterozygosity. Under the hypothesis of equilibrium, heterozygosity alone allows us to estimate that each of these populations had a total effective size (including all subdivisions) of about 3600 individuals.

Both archaic populations also exhibit a pronounced decline in N_e over a period of about 400 thousand years. Under the island model of population structure, these data imply an implausible level of separation among archaic demes. Each deme receives a migrant every 14 generations, and $F_{ST} = 0.69$. A different problem arises when one considers a larger number of demes distributed across a 2-dimensional landscape and exchanging mates only with nearest neighbors. Because of the constraint on global population size, the number of demes can be large only if each one is small. Small groups tend to be exogamous, so the rate of local gene flow must be high. With 144 demes, each of size 25, this model converged even faster than the island model did—too fast to generate the pattern of declining N_e described above.

It seems unlikely that any model assuming equilibrium can be consistent with the pattern observed in archaic DNA. That pattern apparently reflects change either in the number of individuals or in the rate or pattern of gene flow.

1 Introduction

A variety of statistical methods use genetic data to estimate the history of effective population size, N_e [4, 11–13, 18]. In a subdivided population, these estimates depend not only on the number of individuals but also on gene flow between subdivisions [9, 19]. Mazet, Chikhi and their colleagues [6–8, 10] have extended this idea to show that any history of change in N_e can be mimicked by a model that varies rates and patterns of gene flow without changing the number of individual. I focus here on the same phenomenon, but emphasize that estimates of N_e decline even when there has been no change either in the number of individuals or in the rate or pattern of gene flow.

This article will ask whether this phenomenom can explain the estimated histories of archaic population sizes shown in Fig. 1. These show an apparent decline in population size between roughly 700 kya and 300 kya—an interval of 400 ky. Is it possible that this decline might result from subdivision alone, without requiring any changes in deme sizes or in the rate or pattern of gene flow?

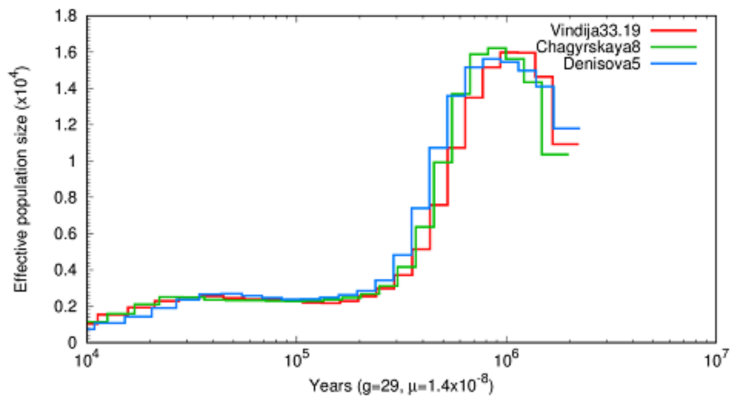


Figure 1: PSMC estimates of the history of population size based on three archaic genomes: two Neanderthals and a Denisovan [5].

2 Methods

2.1 Coalescent hazard and effective population size

Suppose that we sample two genes from a population at time 0 and trace the ancestry of each backwards in time. Eventually, the two lines of descent will *coalesce* into a common ancestor [1, 17]. The *hazard*, $h(t)$, of a coalescent event at time t is the conditional probability density that it occurs at t given that it did not occur between 0 and t [2, p. 6]. In a randomly-mating population of constant size N, this hazard is $h(t) = 1/2N$. In a population of more complex structure, this formula doesn't hold. Nonetheless, we can define an *effective population size* as half the reciprocal of the hazard.

$$N_e = \frac{1}{2h(t)} \tag{1}$$

It is this is the measure of population size that one obtains when N_e is estimated from genetic data.

2.2 The island model

I will consider two models, both of which assume that population sizes and the rates and pattern of gene flow are constant. The first of these is Wright's [20] island model of population structure, which assumes that demes of equal size exchange migrants at equal rates. For any given level of gene flow, this model converges rapidly toward its equilibrium. This rapid convergence will exacerbate the difficulty of fitting it to the slow decline in Fig. 1. Nonetheless, the simplicity of this model makes it useful for illustrating the principles involved. After studying this model, I will then consider a more realistic model involving isolation by distance on a two-dimensional surface.

The island model assumes a population comprising d demes, each with N diploid individuals. Migration occurs at rate m per gene per generation. When a migration occurs (as we look backwards into the past), a gene in one deme moves to one of the $d-1$ other demes, which is chosen at random. We study a pair of homologous genes sampled from a single deme.

To study this process, let us embed it in one with a slightly higher rate, involving what I will call "perturbations." These occur at rate M. When one occurs, the affected gene moves to a deme chosen at random from all d demes. With probability $1/d$, the two demes are the same, so nothing happens. The rest of the time, a perturbation is a migration. If at least one perturbation has occurred, the gene is equally likely to be in all d demes, and its current location is independent of that at which its descendant was sampled. The rates of perturbation and migration are related by $M = md/(d-1)$.

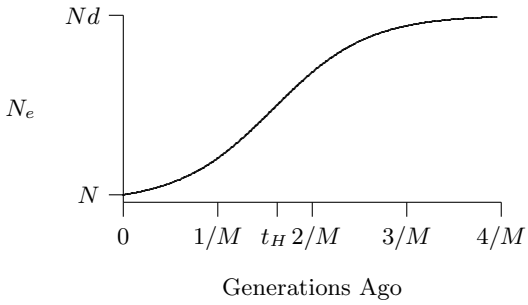


Figure 2: Time path of effective population size when two genes are sampled from the same deme of a subdivided population. Assumptions: $d = 25$ demes of size $N = 500$ with perturbation rate M per generation. t_H is the time at which N_e is half way from N to Nd.

As we trace the history of the two sampled genes backwards in time, the hazard of a perturbation is M per gene copy per generation. The coalescent hazard is $1/2N$ for two genes in the same deme and 0 for genes in different demes. At time t, the probability that neither gene has been perturbed is e^{-2Mt}. In this case, the two genes are still in the original deme, and the hazard of a coalescent event is $1/2N$. With probability $1 - e^{-2Mt}$, at least one perturbation has occurred. The coalescent hazard is the product of $1/d$ (the probability that the two genes are in the same deme) and $1/2N$ (the hazard of coalescence for two genes in the same deme). Putting this together, the coalescent hazard in generation t is $h(t) = e^{-2Mt}/2N + (1 - e^{-2Mt})/2Nd$. The effective population size is half the reciprocal of the hazard,

$$N_e(t) = \frac{Nd}{1 + e^{-2Mt}(d-1)}. \tag{2}$$

$N_e(t)$ is graphed in Fig. 2. It increases in value from N in the recent past to Nd in the distant past. Although it does not converge exponentially, it gets half way from its initial to its final value in

$$t_H = \frac{\log(d+1)}{2M} = \frac{\log(d+1)}{2md/(d-1)} \tag{3}$$

generations.

Under what circumstances should a curve like that in Fig. 2 be visible in data? Intuition suggests that this will depend on the speed of the coalescent process relative to that of migration. If at nearly all loci, the migration process converges to equilibrium before the two genes coalesce, then the population should behave like one with constant effective size Nd. On the other hand, if the process of coalescence is much faster than that of migration, the population should behave like one of constant size N.

To constrain guesses about the value of $4Nm$, I make use of the fact that under the island model, F_{ST} (a measure of differentiation among demes) is

$$F_{ST} \approx \frac{1}{1 + 4Nmd^2/(d-1)^2} \tag{4}$$

where the approximation is good if u is small [15].

2.3 More complex models of population structure

We are interested in the ancestry of genes sampled at time 0 from a single deme. In the model described below, all demes are the same size (N), and they all connected within a symmetric pattern of gene flow. No deme is more central or more peripheral than any other. Consequently, it doesn't matter which deme we sample, and I will assume for convenience that samples are from deme 1.

The pattern of migration is described by the migration matrix, $\mathbf{M}$, whose ijth entry is the expected fraction of deme i that derives from j in the previous generation. Let $\mathbf{x}(t)$ represent a

vector whose jth entry, $x_j(t)$, is the probability that t generations ago the ancestor of a random gene, sampled from deme 1 at time 0, was in deme j. At time 0, $\mathbf{x}(0) = (1, 0, 0, \ldots, 0)^T$, where the superscript T indicates the matrix transpose. At time t, $\mathbf{x}(t) = \mathbf{M}^t\mathbf{x}(0)$.

Because the coalescent hazard at time t is conditioned on absence of any coalescent event between 0 and t, the locations of the ancestors of two genes, both sampled from deme 1 at time 0, are statistically independent. These two ancestors are both in deme j with probability $x_j^2(t)$. They are both in a single unspecified deme with probability $\sum x_j^2(t)$. Thus, the hazard of a coalescent event at time t is $h(t) = \sum x_j^2(t)/2N$. The effective population size at time t is half the reciprocal of the hazard

$$N_e(t) = \frac{N}{\sum x_j^2(t)} \tag{5}$$

2.4 Heterozygosity

In evaluating models of history, we can exclude those that imply values of within-deme heterozygosity, H_W, that are far from those observed in archaic genomes. Expected heterozygosity at equilibrium is approximately

$$E[H_W] \approx 4Ndu \tag{6}$$

when u, the mutation rate per site per generation, is small [14, 16]. This is the same as the heterozygosity of a randomly-mating population whose size equals that of the metapopulation. This result holds with remarkable generality, being affected neither by variation in the sizes of demes nor by the rate or pattern of migration among them, provided that the migration matrix is irreducible and aperiodic. (This means that a gene in one deme may eventually reach any other deme and that the pattern of migration isn't cyclical.) In the real world, some demes may have suffered recent bottlenecks, which reduced their heterozygosity. But we are interested in whether the pattern in Fig. 1 can be explained without recourse to changes in population size and migration rates. Under this hypothesis all demes—no matter how large or small, how central or peripheral—have the same heterozygosity.

3 Results

Archaic genomes consistently exhibit lower heterozygosities than those of modern humans. For example, Mafessoni et al. [5, table S8.2] list values between 1.5×10^{-4} and 2×10^{-4} for four archaic genomes: three Neanderthal and one Denisovan. To be conservative, I will take the upper end of this range as representative of archaic heterozygosity. I also assume a mutation rate of 1.4×10^{-8} per nucleotide site per generation. If heterozygosity were at equilibrium, then Eqn. 6 would imply that the entire population had effective size 3571, irrespective of the number of demes, their sizes, or the pattern of migration among them. Small as this number may seem, it overstates the number of individuals, because effective size exceeds census size in subdivided populations [9]. This small size seems implausible in view of the wide geographic extent of the Neanderthal population and the richness of its archaeological record. We should therefore be skeptical of the hypothesis of equilibrium. Nonetheless, let us ask what it implies in the context of two specific models of geographic population structure.

Under the island model, Eqn. 2 and Fig. 2 show that N_e converges monotonically from N to Nd in backwards time. In the curves in Fig. 1, the early values of N_e are about 5 times the later ones. Thus, d must equal 5 if we are to fit these data to the island model. Dividing metapopulation size (3571) by 5 gives a deme size of $N = 714$. The observed decline in N_e occupies an interval of

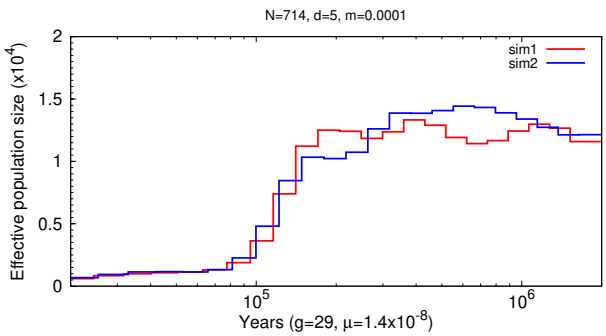


Figure 3: PSMC estimates based on simulations of island-model migration, with $d = 5$, $N = 714$, $m = 0.0001$. Simulations used msprime [3] and involved 100 chromosomes, each 2 Mb in size.

about 400 ky, so t_H is 200 ky or 6897 generations. These values, together with Eqn. 3, imply that $m \approx 0.0001$. Fig. 3 shows PSMC estimates [4] based on msprime simulations [3] of this model.

As expected, N_e increases backwards in time over an interval of about 400,000 years. Furthermore, the heterozygosities observed in the two simulation runs—1.8×10^{-4} and 2.2×10^{-4}—were in good agreement with observed archaic values. In these respects, the island model provides a good fit to archaic data. However, this model implies an improbable degree of isolation among archaic demes: demes receive an immigrant every 14 generations on average, and $F_{ST} = 0.69$—a much higher value than is typically seen in populations of large mammals.

But since the observed decline in archaic N_e is very slow, the island model is not ideal. For any given level of immigration, the island model converges toward equilibrium more rapidly than alternative models do. This rapid convergence happens because each deme exchanges genes with every other deme. Convergence is slower if demes exchange genes only with their immediate neighbors. In that case, it can take many generations for a gene to travel from one deme to another that is far away. Let us therefore consider a model in which each deme exchanges migrants with its immediate neighbors on a square lattice. Each deme receives (on average) m immigrants per generation, $m/4$ from each of its 4 nearest neighbors to the north, south, east, and west. To eliminate edge effects, I wrap the lattice around a torus (a geometrical object like a doughnut).

In constructing this model we encounter a problem. The low archaic heterozygosity implies that the total population contains only about 3600 individuals, as explained above. If demes are numerous, they must be very small. Small groups of animals tend to be exagomous, presumably because selection favors mating patterns that minimize close inbreeding. The rate of gene flow between neighboring demes will therefore be high. Let us assume that demes are of size 25, as is typical of bands of modern human foragers. To stay within our budget of 3600 individuals, we can afford 144 demes, arrayed on a 12×12 lattice. I assume an immigration rate of 0.1 per generation. This seems implausibly low for demes of this size, but I'm trying to bias things in favor of slow convergence.

Figure 4 is based on Eqn. 5 and shows how N_e declines with time under our model of 2-dimensional isolation by distance on a torus. N_e declines across an interval of about 300 generations, or 8700 years. This decline is much faster than that seen in data from archaic humans.

Msprime simulations of this model generate heterozygosities similar to those observed in archaic genomes. But estimates of N_e show a very different trajectory. This model converges so rapidly that the pattern of declining N_e seen in Fig. 4 is not visible in Fig. 5. Instead, the PSMC estimates show a population size that is large and constant. These estimates are telling us about the size of the metapopulation, not those of the demes. Because it fails to generate the observed decline in archaic N_e, this model is not consistent with the archaic data.

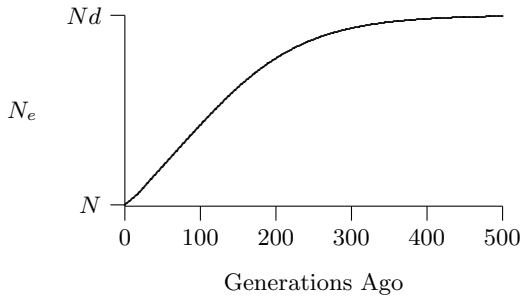


Figure 4: Time path of effective population size under isolation by distance on a torus. Assumptions: demes of size $N = 25$ are on a 12×12 lattice, which is wrapped around a torus. The migration rate $m = 0.1$ per generation.

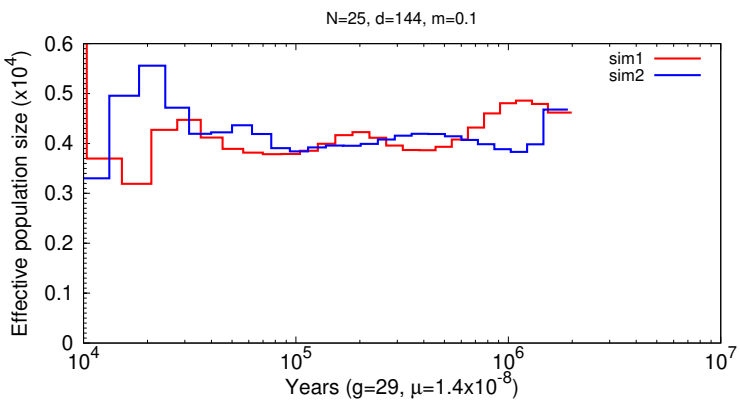


Figure 5: PSMC estimates based on simulations of 2-dimensional stepping-stone migration. Demes are on a 12×12 lattice, wrapped around a torus. Each deme is of size $N = 25$, and the rate of immigration is $m = 0.1$ per generation.

4 Discussion and Conclusions

This article has asked when it is reasonable to interpret a decline in effective population size as the consequence of geographic population structure, without positing change in census population size or in the rate or pattern of gene flow among demes. In other words, what sorts of decline in N_e are consistent with equilibrium, and what sorts are inconsistent?

The expected decline is illustrated for two models of population structure in Figs. 2 and 4. These declines reflect the fact that we have sampled two genes from a single deme. In the recent past, it is likely that the two ancestors of these genes are still neighbors. Coalescent hazard is therefore high and N_e is small. As we move farther into the past, the ancestors are less likely to be neighbors, so the coalescent hazard declines and N_e increases. This will have an appreciable effect on genetic data provided that the time scales of the processes of coalescence and migration are similar. This similarity of time scales holds in Fig. 3, where the predicted pattern of declining N_e is visible in simulated data. It does not hold in Fig. 5, where N_e converges too rapidly for this process to be visible.

The island model provided a poor fit to archaic human data. Under that model, the low archaic heterozygosity together with the slow decline in archaic N_e (Fig. 1) imply an implausible degree of isolation among archaic demes. This suggests that a model of isolation by distance, which should converge more slowly than the island model, might provide a better fit. However, at equilibrium, the low heterozygosity of Neanderthals implies that if there are many demes, each one must be small. Because small demes tend to be exagomous, it was not possible to find a plausible model of isolation by distance that converged slowly.

It therefore seems unlikely that any model assuming equilibrium can account for the declines in N_e observed in Neanderthals and Denisovans. These declines must result either from changes in census size or from changes in the rate or pattern of gene flow.

References

[1] Richard R. Hudson. "Gene genealogies and the coalescent process". In: *Oxford Surveys in Evolutionary Biology*. Ed. by Douglas Futuyma and Janis Antonovics. Vol. 7. Oxford: Oxford University Press, 1990, pp. 1–44.

[2] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. New York: Wiley, 1980. DOI: 10.2307/3315078.

[3] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. "Efficient coalescent simulation and genealogical analysis for large sample sizes". *PLoS Computational Biology* 12.5 (May 2016), pp. 1–22. DOI: 10.1371/journal.pcbi.1004842.

[4] Heng Li and Richard Durbin. "Inference of human population history from individual whole-genome sequences". *Nature* 475.7357 (2011), pp. 493–496. DOI: 10.1038/nature10231.

[5] Fabrizio Mafessoni et al. "A high-coverage Neandertal genome from Chagyrskaya Cave". *Proceedings of the National Academy of Sciences, USA* 117.26 (2020), pp. 15132–15136. DOI: 10.1073/pnas.2004944117.

[6] O. Mazet et al. "On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference?" *Heredity* 116.4 (2016), pp. 362–371. DOI: 10.1038/hdy.2015.104.

[7] Olivier Mazet and Camille Noûs. "Population genetics: coalescence rate and demographic parameters inference". *arXiv* (2023). DOI: 10.48550/ARXIV.2207.02111.

[8] Olivier Mazet, Willy Rodríguez, and Lounès Chikhi. "Demographic inference using genetic data from a single individual: separating population size variation from population structure". *Theoretical Population Biology* 104 (2015), pp. 46–58. DOI: j.tpb.2015.06.003.

[9] Masatoshi Nei and Naoyuki Takahata. "Effective population size, genetic diversity, and coalescence time in subdivided populations". *Journal of Molecular Evolution* 37 (1993), pp. 240–244.

[10] Willy Roberto Rodriguez et al. "The IICR and the non-stationary structured coalescent: demographic inference with arbitrary changes in population structure". *bioRxiv* (2018). DOI: 10.1101/341750.

[11] Alan R. Rogers. "Legofit: estimating population history from genetic data". *BMC Bioinformatics* 20 (2019), p. 526. DOI: 10.1186/s12859-019-3154-1.

[12] Alan R. Rogers. "An efficient algorithm for estimating population history from genetic data". *Peer Community Journal* 2 (2022), e32. DOI: 10.24072/pcjournal.132.

[13] Stephan Schiffels and Richard Durbin. "Inferring human population size and separation history from multiple genome sequences". *Nature Genetics* 46.8 (Aug. 2014), pp. 919–925. DOI: https://doi.org/10.1038/ng.3015.

[14] Montgomery Slatkin. "The average number of sites separating DNA sequences drawn from a subdivided population". *Theoretical Population Biology* 32.1 (1987), pp. 42–49. DOI: 10.1016/0040-5809(87)90038-4.

[15] Montgomery Slatkin. "Inbreeding coefficients and coalescence times". *Genetical Research, Cambridge* 58 (1991), pp. 167–175.

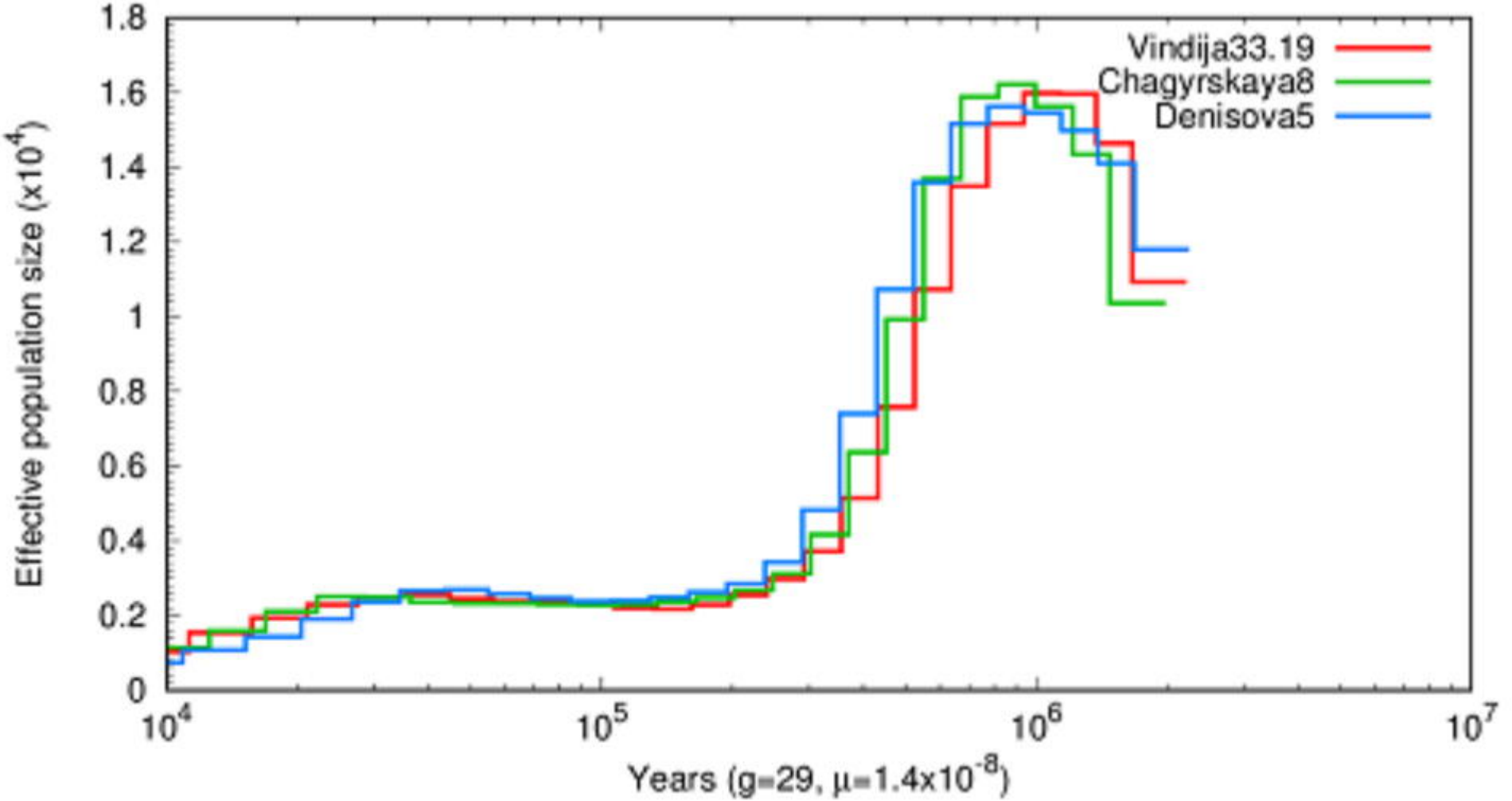
[16] Curtis Strobeck. "Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision". *Genetics* 117.1 (1987), pp. 149–153. DOI: 10.1093/genetics/117.1.149.

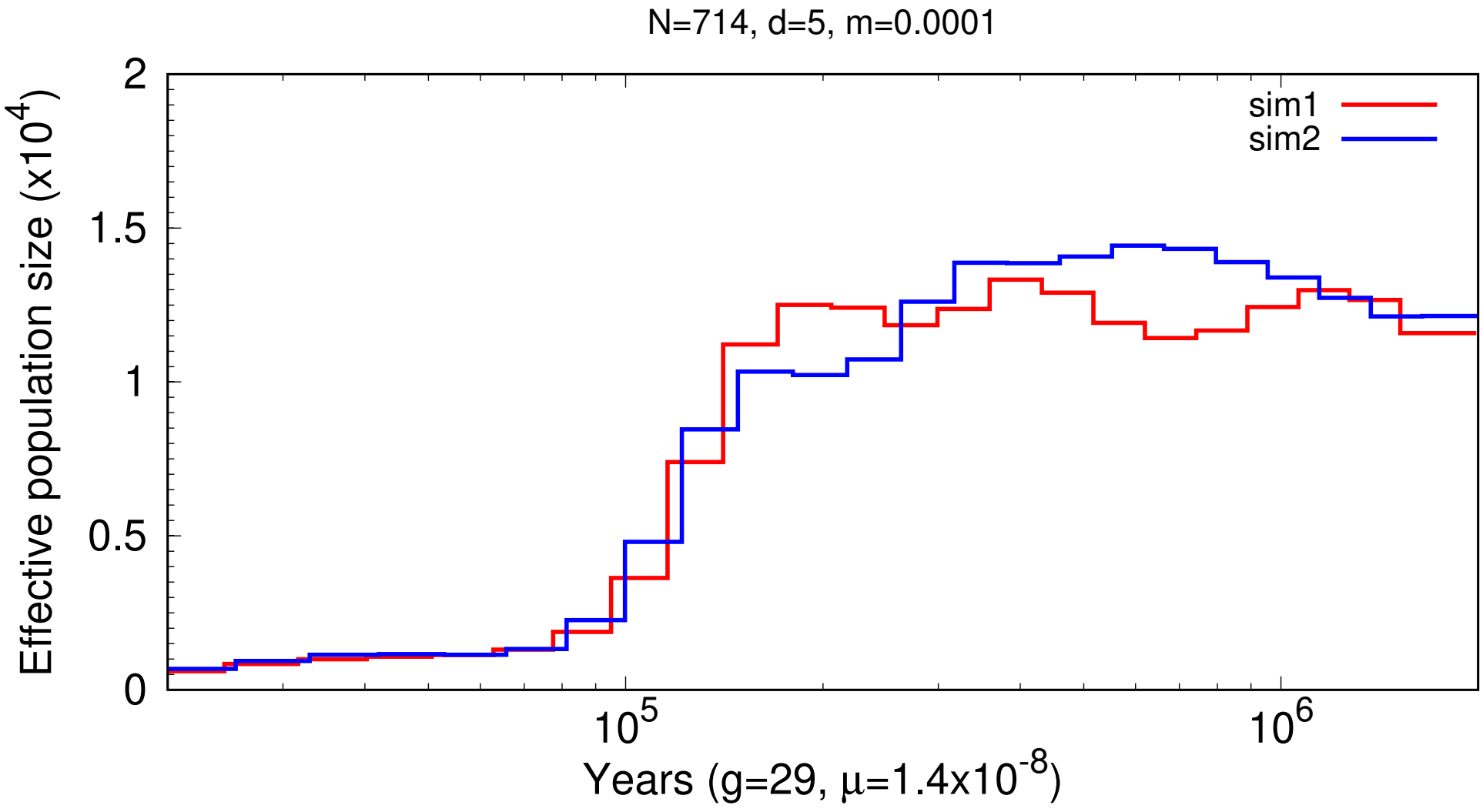
[17] Simon Tavaré. "Line-of-descent and genealogical processes, and their applications in population genetics models". *Theoretical Population Biology* 26 (1984), pp. 119–164. DOI: 10.1016/0040-5809(84)90027-3.

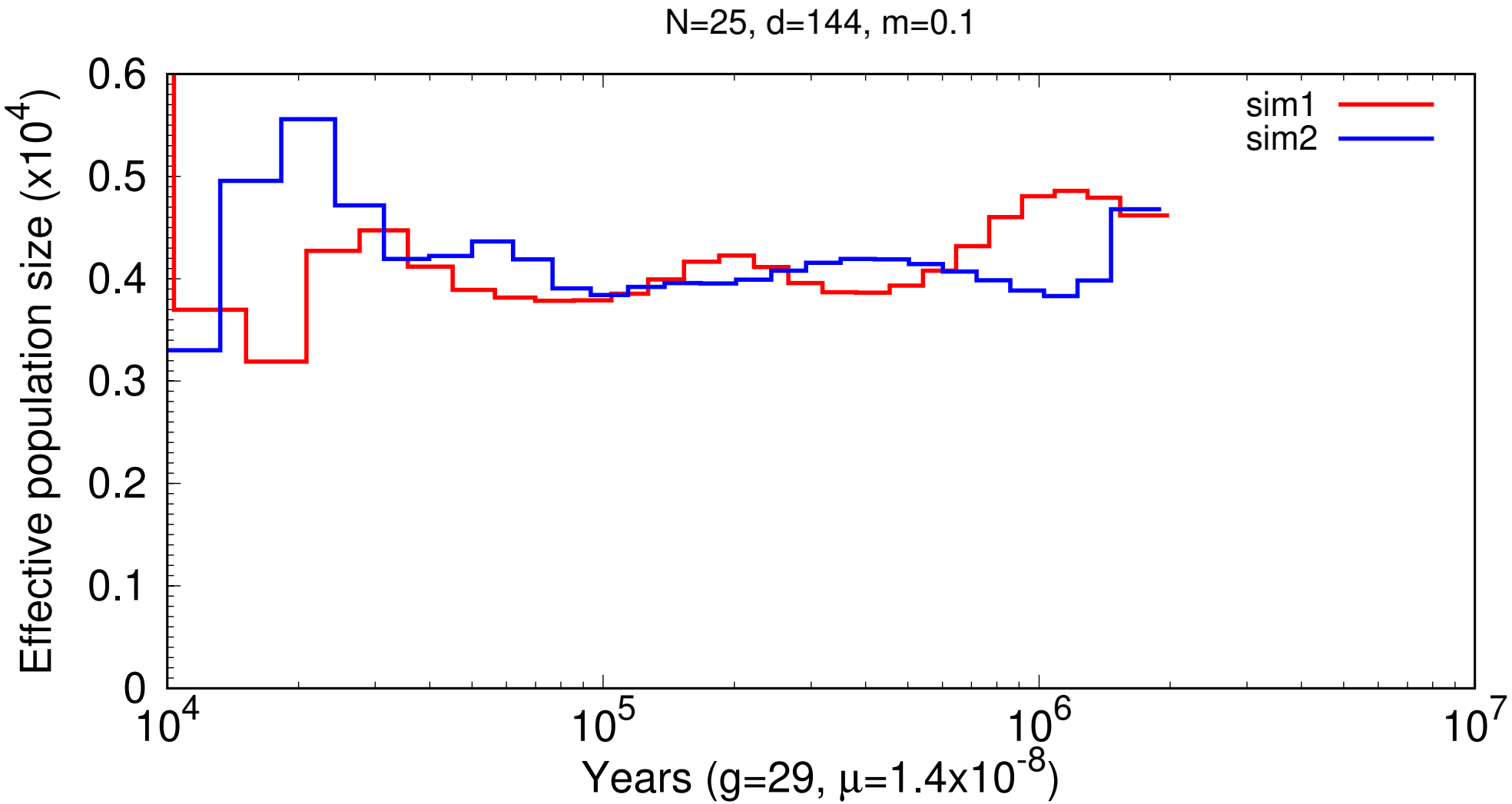
[18] Jonathan Terhorst, John A Kamm, and Yun S Song. "Robust and scalable inference of population history from hundreds of unphased whole genomes". *Nature Genetics* 49.2 (2017), pp. 303–309. DOI: 10.1038/ng.3748.

[19] M. C. Whitlock and Nicholas H. Barton. "The effective size of a subdivided population". *Genetics* 146 (1997), pp. 427–441.

[20] Sewall Wright. "The genetical structure of populations". *Annals of Eugenics* 15 (1951), pp. 323–354. DOI: 10.1111/j.1469-1809.1949.tb02451.x.


Vindija33.19
Chagyrskaya8
Denisova5
Effective population size (x10^4)
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
10^4
10^5
10^6
10^7
Years (g=29, μ=1.4x10^-8)


N=714, d=5, m=0.0001
sim1
sim2
Effective population size (x10^4)
2
1.5
1
0.5
0
10^5
10^6
Years (g=29, μ=1.4x10^-8)


N=25, d=144, m=0.1
sim1
sim2
Effective population size (x10^4)
0.6
0.5
0.4
0.3
0.2
0.1
0
10^4
10^5
10^6
10^7
Years (g=29, μ=1.4x10^-8)