Theoretical Population Biology 100 (2015) 63–78

Contents lists available at ScienceDirect

Theoretical Population Biology

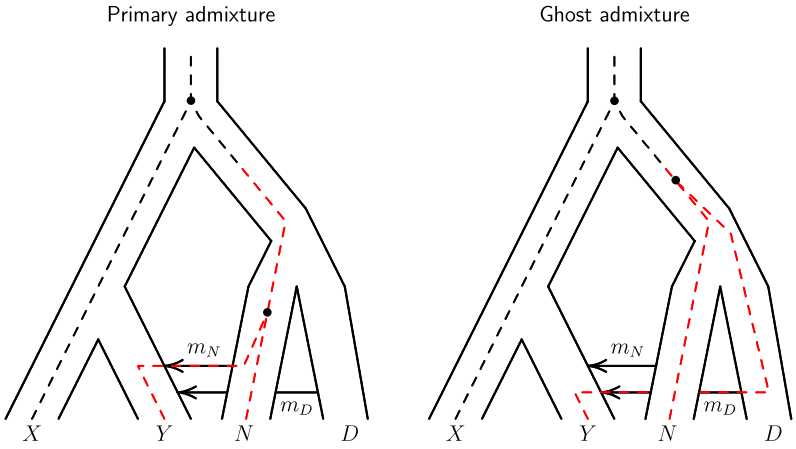
journal homepage: www.elsevier.com/locate/tpb

Bias in estimators of archaic admixture

Alan R. Rogers *, Ryan J. Bohlender

Department of Anthropology, 270 S 1400 E, University of Utah, Salt Lake City, UT 84112, United States

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 29 July 2013
Available online 6 January 2015

Keywords:
Archaic admixture
Population genetics
Gene genealogy
Human evolution
Neanderthal
Denisovan

ABSTRACT

This article evaluates bias in one class of methods used to estimate archaic admixture in modern humans. These methods study the pattern of allele sharing among modern and archaic genomes. They are sensitive to "ghost" admixture, which occurs when a population receives archaic DNA from sources not acknowledged by the statistical model. The effect of ghost admixture depends on two factors: branch-length bias and population-size bias. Branch-length bias occurs because a given amount of admixture has a larger effect if the two populations have been separated for a long time. Population-size bias occurs because differences in population size distort branch lengths in the gene genealogy. In the absence of ghost admixture, these effects are small. They become important, however, in the presence of ghost admixture. Estimators differ in the pattern of response. Increasing a given parameter may inflate one estimator but deflate another. For this reason, comparisons among estimators are informative. Using such comparisons, this article supports previous findings that the archaic population was small and that Europeans received little gene flow from archaic populations other than Neanderthals. It also identifies an inconsistency in estimates of archaic admixture into Melanesia.

1. Introduction

Forty years ago, Howells (1976) discussed the origin of modern humans, emphasizing two extreme views. One of these, which would now be called the multiregional hypothesis (Wolpoff, 1989), held that modern humans evolved across a broad front within

* Corresponding author.
E-mail addresses: rogers@anthro.utah.edu (A.R. Rogers), ryan.bohlender@gmail.com (R.J. Bohlender).

http://dx.doi.org/10.1016/j.tpb.2014.12.006
0040-5809/

Table 1
The site patterns studied in this analysis (with 0 and 1 representing the ancestral and derived alleles); the gene tree implied by each pattern; and the counts (I_{uv}) of such sites for a San sample, *x*, a French sample, *y*, a Neanderthal sample, *n*, and a chimpanzee sample, *o* (Patterson et al., 2010a, p. S138).

Site pattern					Gene tree	Count
	x	*y*	*n*	*o*		
ny	0	1	1	0	$((x,(y,n)),o)$	$I_{ny} = 103,612$
nx	1	0	1	0	$(((x,n),y),o)$	$I_{nx} = 95,347$
xy	1	1	0	0	$(((x,y),n),o)$	$I_{xy} = 303,340$

a worldwide population held together by gene flow. The other, which would now be called the replacement hypothesis (Stringer and Andrews, 1988), involved "a single origin, outward migration of separate stirps, like the sons of Noah, and an empty world to occupy, with no significant threat of adulteration by other gene pools or even evaporating gene puddles" (Howells, 1976, p. 480). But Howells also considered a third hypothesis, which also proposed expansion from a single point of origin. This expansion, however, involved "encounters between populations of modern man and of other forms, with consequent gene flow" (Howells, 1976, p. 492). This hypothesis has been endorsed by various paleoanthropologists (Bräuer, 1984, 1989; Smith et al., 1989; Trinkaus, 2005). During the past decade, it has also received support from genetics.

In the preceding decade, geneticists were less supportive. At that time, human evolutionary genetics dealt mainly with mitochondrial DNA (mtDNA), which is remarkably homogeneous in modern human samples. Stoneking (1993) argued that the mtDNA of Neanderthals ought to lie well outside the narrow range of variation seen in modern human samples. The absence of such divergent mtDNAs argued that their frequency within the human species must be low. Yet as Stoneking observed, this did not refute the hypothesis of archaic admixture. Introgressed archaic mtDNAs might simply have been lost by genetic drift. Nordborg (1998) developed a model of this process, which showed that mitochondrial data have low power to detect archaic admixture.

Since the late 1990s, the field has relied increasingly on nuclear DNA. Because unlinked loci provide essentially independent replicates of the evolutionary process, the nuclear genome provides far greater power to detect admixture. A variety of statistical methods has been developed. Some rely on information in the site frequency spectrum (Eswaran et al., 2005; Yang et al., 2012). Others are based on linkage disequilibrium (Wall, 2000; Wall et al., 2009; Wall and Hammer, 2006; Plagnol and Wall, 2006; Moorjani et al., 2011; Hammer et al., 2011; Abi-Rached et al., 2011; Evans et al., 2005; Mendez et al., 2012; Cox et al., 2008).

Our focus here is on a different class of methods, which capitalizes on the availability of archaic DNA sequences. These methods infer admixture from the frequency with which derived alleles are shared by pairs of samples. In the most common pattern, the derived allele is shared by genes drawn from closely related populations. Two samples uniquely share a derived allele only if a mutation occurs in a uniquely shared ancestor. For example, at many of the loci in Table 1, the derived allele is present only in the French and African samples. These derived alleles arose in genes that were ancestral to the French and African samples but not to the Neanderthal or the Chimpanzee. Such sites are common in the data, because the French and African populations are conspecific and thus share a portion of their evolutionary history.

What then of the other two patterns, in which the derived allele is shared by a Neanderthal and one of the two modern human samples? In the absence of admixture, these site patterns can arise only through incomplete lineage sorting. If random mating prevailed within the population ancestral to humans and Neanderthals, these two patterns ought to occur in equal frequencies (Pamilo and Nei, 1988). Yet in Table 1 the *ny* pattern occurs more often than the *nx* pattern. This excess supports the hypothesis of admixture between Neanderthals and the ancestors of Europeans. Several published methods use this principle to estimate the fraction of archaic genes in modern populations (Green et al., 2010; Reich et al., 2010; Durand et al., 2011; Reich et al., 2011; Meyer et al., 2012; Patterson et al., 2012b).

These methods rely on the assumption of random mating in the ancestral population. If instead that population were geographically structured, with limited gene flow between geographic subdivisions, this could result in biased frequencies such as those seen in Table 1 (Slatkin and Pollack, 2008). This hypothesis of "ancestral subdivision" has been seen as an alternative to that of archaic admixture (Durand et al., 2011; Eriksson and Manica, 2012; Blum and Jakobsson, 2011). This issue is still contentious, with some authors arguing that it has been refuted (Yang et al., 2012; Sankararaman et al., 2012; Wall et al., 2013) and others that it has not been properly tested (Eriksson and Manica, 2014).

Whatever the outcome of this dispute, there are also other potential biases. Several published estimators allow for gene flow from only one archaic population. Estimates may be biased if the modern population also received genes from other archaic populations, a phenomenon known as "ghost admixture" (Beerli, 2004; Slatkin, 2005; Durand et al., 2011, p. 2240; Harris and Nielsen, 2013). Some estimators also assume that population size has been constant throughout the human gene tree. These estimators may be biased if populations have varied in size. In what follows, we explore the magnitudes of these biases.

2. Methods

Because this article is about bias, we focus on expected values and ignore statistical uncertainties. Following Durand et al. (2011, p. 2241), we assume that admixture occurs at discrete points in time. Between these events, the isolation of populations is complete. Within populations, we assume mating is at random.

2.1. Population sizes and coalescent time scale

We use single upper-case letters, such as X and Y, to label individual populations. The notation XY refers to the population ancestral to X and Y but not ancestral to other sampled populations. The diploid sizes of X, Y, and XY are written as N_X, N_Y, and N_{XY}. The symbol N_0 represents the diploid size of the ancestral human population—the ancestors of modern humans, Neanderthals, and Denisovans, but not of chimpanzees.

In this population, the hazard of a coalescent event between a single pair of lineages is $1/2N_0$ per generation, and their mean coalescence time is $2N_0$ generations (Hudson, 1990). However, let us adopt a coalescent time scale, with time units of $2N_0$ generations. On this scale, the mean and hazard are both unity, and the mutation rate is $U \equiv 2N_0u$, where u is the mutation rate per generation.

We allow for changes in population size at branch points in the population tree. Between branching points, we assume the population is constant. For example, Fig. 1 implies that population XY existed within the time interval (ζ, λ). Within this interval, we assume that it had constant size N_{XY}. Let $K_{XY} = N_{XY}/N_0$. In words, K_{XY} is the size of population XY relative to that of the ancestral human population. For the duration of population XY, the coalescent hazard for a single pair of lineages is $1/K_{XY}$ per unit of coalescent time. Other ratios, such as K_X and K_Y, are defined similarly.

The "survival function", $S_{XY}^{(\zeta,\lambda)} \equiv e^{-(\lambda-\zeta)/K_{XY}}$, is the probability that a pair of lineages within XY remain distinct throughout interval (ζ, λ). The "cumulative distribution function", $F_{XY}^{(\zeta,\lambda)} = 1 - S_{XY}^{(\zeta,\lambda)}$, is the probability that the pair coalesces within this interval.