Title

Possible Ancestral Structure in Human Populations

Permalink

https://escholarship.org/uc/item/7mm6q32q

Journal

PLOS Genetics, 2(7)

ISSN

1553-7390

Authors

Plagnol, Vincent
Wall, Jeffrey D

Publication Date

2006-07-01

DOI

10.1371/journal.pgen.0020105

Possible Ancestral Structure in Human Populations

Vincent Plagnol[*], **Jeffrey D. Wall**

Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California, United States of America

Determining the evolutionary relationships between fossil hominid groups such as Neanderthals and modern humans has been a question of enduring interest in human evolutionary genetics. Here we present a new method for addressing whether archaic human groups contributed to the modern gene pool (called ancient admixture), using the patterns of variation in contemporary human populations. Our method improves on previous work by explicitly accounting for recent population history before performing the analyses. Using sequence data from the Environmental Genome Project, we find strong evidence for ancient admixture in both a European and a West African population ($p \approx 10^{-7}$), with contributions to the modern gene pool of at least 5%. While Neanderthals form an obvious archaic source population candidate in Europe, there is not yet a clear source population candidate in West Africa.

Citation: Plagnol V, Wall JD (2006) Possible ancestral structure in human populations. PLoS Genet 2(7): e105. DOI: 110.1371/journal.pgen.0020105

Introduction

A long-standing controversy in the field of human evolution concerns the origin of modern humans [1,2]. The debate focuses on the relationship between various groups of archaic humans, such as Neanderthals or Asian *Homo erectus,* and anatomically and behaviorally modern *Homo sapiens* (i.e., modern humans). At one end of the spectrum, the multiregional model claims that modern humans evolved in concert across the Old World from various archaic groups [3]. At the other end, the Recent African Origin (RAO) model posits that modern humans evolved in a single location in Africa and from there spread and replaced all other existing hominids [4]. Currently, most but not all of the fossil evidence seems to support the RAO model [5,6]. From a genetic perspective, we can rephrase the debate in terms of what contribution archaic human populations have made to the contemporary human gene pool. The multiregional model predicts that this contribution would be substantial while the RAO model predicts that this contribution is negligible. Other models predict intermediate contributions of archaic populations to the modern gene pool [7].

The easiest way to answer this question is through a direct comparison of DNA sequences from both archaic and modern populations. Recently, researchers have managed to sequence fragments of Neanderthal mtDNA from fossil bones [8–11]. All published Neanderthal mtDNA sequences are quite different from all known modern human mtDNA sequences, and it is extremely unlikely that Neanderthals made any contribution to the modern human mtDNA gene pool. Although this observation is consistent with the RAO model, it does not prove that Neanderthals and modern humans did not interbreed—the two groups may have mixed but Neanderthal mtDNA may have been lost by the chance action of genetic drift. Subsequent studies have concluded that the data are consistent with a Neanderthal contribution of up to 25% of the modern gene pool [10,12]. A comparison of Neanderthal nuclear DNA with modern human nuclear DNA has the potential to clarify the precise genetic relationship between Neanderthals and modern humans. So far no Neanderthal nuclear DNA sequences have been determined, though recent technological advances give us the hope that such sequences may be recovered in the future [13,14].

In this paper, we take a different approach to the question. We look for signs of Neanderthal admixture by analyzing the patterns of linkage disequilibrium (LD) in contemporary human DNA sequences. Our method relies on the observation that the genetic signature of ancient admixture is so strong that even tens of thousands of years of random mating is not enough to obscure it [15]. To see this, consider the following crude approximation: at the time of (putative) admixture, extensive LD would extend across the whole genome. After 2,000 generations of random mating (40,000 years, assuming a generation time of 20 years), LD would still extend roughly 0.05 cM on average, equivalent to approximately 40 Kb, assuming 1.25 cM/Mb (cf. [16]). We look for evidence of ancient admixture in patterns of LD at intermediate distances (e.g., 5–50 Kb).

To avoid possible confounding effects, we first use extant sequence data to estimate parameters for a demographic null model that incorporates several known features of modern human history: recent population growth, a bottleneck in Europeans, and population differentiation between European and African populations [17]. Then, we introduce a new measure of LD called S^* which generalizes the work of [15].

Editor: Anna Di Rienzo, University of Chicago, United States of America

Received February 17, 2006; **Accepted** May 26, 2006; **Published** July 28, 2006

A previous version of this article appeared as an Early Online Release on May 26, 2006 (DOI: 110.1371/journal.pgen.0020105.eor).

DOI: 110.1371/journal.pgen.0020105

Abbreviations: ARG, ancestral recombination graph; CEPH, Centre d'Etude du Polymorphisme Humain 1980 database of people living in Utah with ancestry from Northern and Western Europe; EGP, Environmental Genome Project; RAO, recent African origin

[*] To whom correspondence should be addressed. E-mail: vincent.plagnol@normalesup.org

Synopsis

Determining the evolutionary relationships between modern humans and fossil hominine groups such as Neanderthals has been a question of enduring interest in human evolutionary genetics. In this paper, Plagnol and Wall present a new method for addressing whether archaic human groups contributed to the modern gene pool. Using sequence data from the Environmental Genome Project, they find strong evidence for ancient admixture in both a European and a West African population, with contributions to the modern gene pool of at least 5%. While Neanderthals form an obvious archaic source population candidate in Europe, there is not yet a clear source population candidate in West Africa. The authors' results have direct implications for the competing models of modern human origins. In particular, their estimates of non-negligible contributions of archaic populations to the modern gene pool are inconsistent with strict forms of the Recent African Origin model, which posits that modern humans evolved in a single location in Africa and from there spread and replaced all other existing hominines.

We use S^* to test whether there has been archaic admixture in the history of modern Europeans and West Africans.

Results

Estimation of Demographic Parameters

Our analysis is based on 135 finished genes from the NIEHS Environmental Genome Project (EGP, as of February 2006, see [18,19] and Materials and Methods for details). We first model the recent history of European (CEPH [Centre d'Etude du Polymorphisme Humain 1980 database of people living in Utah with ancestry from Northern and Western Europe]) and Yoruba populations. We restrict our study to these two samples in order to limit the number of parameters that need to be estimated. The model must be simple enough to allow a precise estimation of its parameters yet sophisticated enough to capture the characteristics of both populations. Differences between the samples are illustrated by commonly used summary statistics (Table 1): Watterson's estimator of θ [20], Hudson's estimator of ρ [21], Tajima's D [22], and F_{ST} [23]. The bottom part of the table gives the values of these statistics for our best-fitting model.

We use a simple two-island model with islands representing European and African populations. Initially, we considered models where there was no migration between the two populations after they split. These models did not fit the data well (unpublished data) so we use a model that incorporates a low level of migration between the populations. We include population growth in each population as well as a bottleneck in the European branch. We estimate a total of six parameters and the likelihood is estimated over a grid of values to find the maximum (see Materials and Methods for details). The scaling in years has been done assuming an ancestral population size of 10,000 diploid individuals (as estimated in [17]) and a generation time of 20 years. We fitted a gamma distribution to the variability of the recombination rate ρ to reproduce the variability observed in the Yoruba sample (see Materials and Methods).

To estimate parameters we use a composite-likelihood approach based on various summary statistics (cf. [17]). We used two sets of summary statistics. The first set consists of four statistics: Tajima's D in each sample, Fu and Li's D^* in the CEPH sample, and F_{ST}. Tajima and Fu and Li's D^* measures the frequency spectrum, and F_{ST} the level of divergence between the populations. For a given value of the parameter, the joint likelihood of these statistics is estimated by fitting to the data a multivariate Gaussian distribution. Parameters of this distribution are estimated using Monte Carlo simulations.

The second set of summary statistics divides the SNPs at a locus in three categories: private in the CEPH sample, private in the Yoruba sample, and segregating in both samples. Sites segregating in both samples are subsequently divided between low and high frequency (we set the threshold at 10%). SNPs segregating in both samples and at low frequency are characteristic of recent migrations and help to estimate this rate. This set of statistics has the useful property that the joint distribution can be computed exactly for a given realization of the genealogical process (Ancestral Recombination Graph [ARG] [24]). The final likelihood is then averaged over a large number of ARG using Monte Carlo simulations.

Even though these two sets of summary statistics are correlated, we could not estimate their joint distribution. To estimate the overall likelihood we use a composite-likelihood approximation: precisely, we assume that both sets of summary statistics are independent.

Table 1. Distribution of Summary Statistics for the NIEHS-EGP Dataset and Our Best-Fitting Model

Dataset/Model	Summary Statistics	Africa	Europe	Overall
NIEHS dataset, 135 genes, 12 Yoruba, and 22 CEPH individuals	$\hat{\theta}$per kb (SD)	0.87 (0.38)	0.58 (0.3)	0.88 (0.35)
	$\hat{\rho}$per kb (SD)	0.38 (0.38)	0.18 (0.26)	–
	Tajima's D (SD)	−0.54 (0.68)	−0.09 (1.1)	−0.75 (0.75)
	F_{ST} (SD)	–	–	0.15 (0.11)
Best-fitting model, 12 Yoruba and 22 CEPH individuals	$\hat{\theta}$per kb (SD)	0.87 (0.23)	0.57 (0.2)	0.87 (0.18)
	$\hat{\rho}$per kb (SD)	0.37 (0.44)	0.17 (0.24)	–
	Tajima's D (SD)	−0.43 (0.76)	−0.07 (1.1)	−0.67 (0.81)
	F_{ST} (SD)	–	–	0.15 (0.11)

We use Watterson's estimator of θ, Hudson's estimator of ρ, Tajima's D, and F_{ST}.
SD, standard deviation.
DOI: 110.1371/journal.pgen.0020105.t001

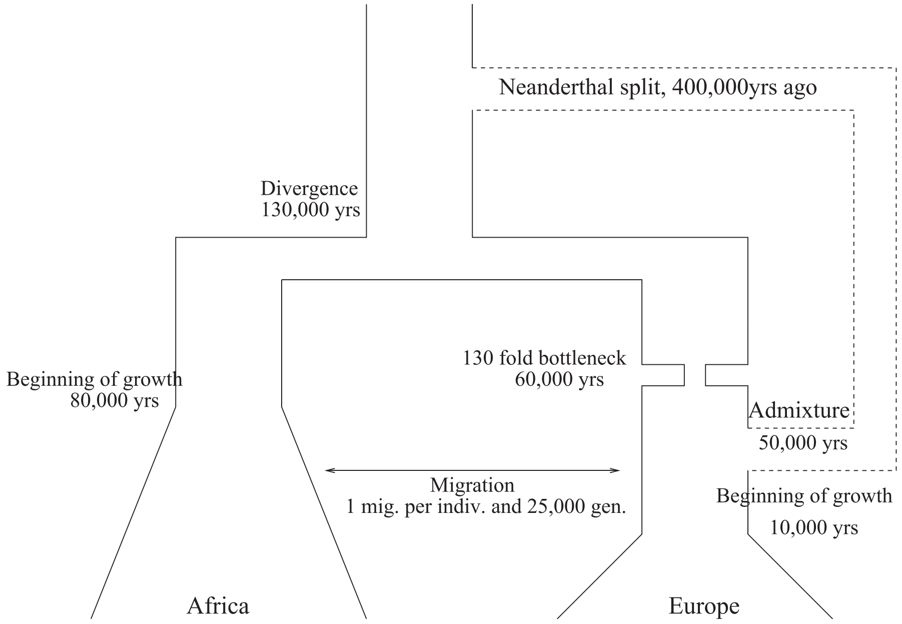

Figure 1. Demographic Model for European and African Populations with the Value of Our Best-Fitting Parameters
DOI: 110.1371/journal.pgen.0020105.g001

Our approach does not provide an accurate estimate of the date of divergence between the populations. Interpreting confidence intervals is difficult because we use a composite-likelihood approach. Nevertheless, a χ^2 approximation for the composite-log-likelihood ratio provides our best estimate of the confidence interval. Using this approximation we find a lower bound at 120,000 years and no upper bound. Precisely, the goodness-of-fit for an equilibrium island model with a low rate of migration between both populations is only slightly worse than for our best-fitting model. We set the divergence date to the lower bound of the confidence interval (more consistent with our knowledge of human history) and verified that this choice does not affect qualitatively the results presented in this paper (see Discussion).

Our procedure estimates that the bottleneck event is more ancient than the putative admixture event. We find that precise dating of this bottleneck is difficult because beyond 50,000 years, a change in the date of the bottleneck has very little effect on the pattern of polymorphism. The parameters of this model are presented in Figure 1 and average values of commonly used summary statistics are presented in Table 1. The dashed line in Figure 1 represents the potential admixture with an archaic population. We provide the likelihood profiles in Figure S1 and the associated *ms* [25] command line that generates this model in Protocol S1.

Goodness of Fit

To be able to assess the significance of the pattern of LD, one needs to measure the goodness of fit of the model to confirm that it captures the main features of European and West African demography. We provide quantile–quantile plots between data and simulated distributions in Figures S2, S3, and S4 for the summary statistics used in the inference procedure as well as comparison of the simulated and observed frequency spectrum. These figures show that the model is mostly consistent with the data and explains well the summary statistics used in the fitting procedure. Though not directly comparable, it appears that our null model provides at least as good a fit as the demographic model proposed by [26]. However, we find some limitations in the goodness of fit. Precisely, the frequency spectrum (see Figure S4) does not fit very well: our model tends to simulate more singletons and fewer low-frequency SNPs (excluding singletons) than are observed in the Yoruba sample.

An important feature of our demographic model is that it reproduces well the ratio of the estimated recombination rate ρ between the CEPH and Yoruba populations. This is remarkable because no aspect of LD information was used in our fitting procedure.

Measure of LD: S^*

We now show that we can detect a specific aspect of the level of LD that is directly affected by the level of admixture and that is not captured by the estimator of the recombination rate ρ. Our statistic S^* is designed to identify which SNPs are the most likely to have mutated in a putative archaic population. Typically these mutations accumulate on the same branch of the genealogical tree, generating an identical pattern of mutations called congruent sites [15]. S^* generalizes this concept and can extract from the sequence the largest subset of SNPs which are almost congruent, a concept that we define formally in the Methods. S^* is highly sensitive to ancient admixture, with higher levels leading to larger values if S^*. We compute three different versions of S^*. The first version uses all the available polymorphism data. However, if the admixture occurs within the European population and in the absence of migration (or at least at a very low level), the SNPs that originated in the archaic population must be private to the European sample. Hence, to test for a recent admixture in the European sample we need to restrict the computation of S^* to SNPs private to the CEPH sample. Alternatively, we only use SNPs private to the Yoruba sample when we test for admixture within the African branch. We denote these values as S^*_{All},S^*_{Yor}, and S^*_{CEPH}.

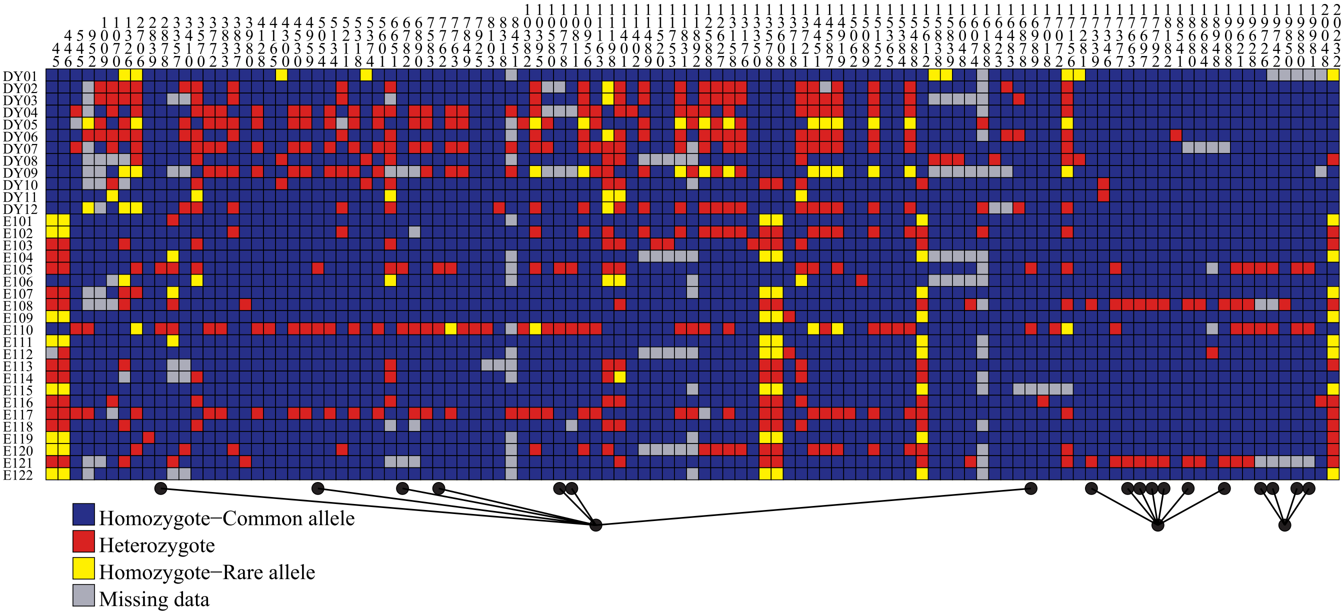

Figure 2. Polymorphism Data Using the Visual Genotype Display Format

See [35,36] for the gene *chrna4*. The first 12 rows are Yoruba genotypes and the last 22 rows are CEPH genotypes.
For the gene *chrna4*, S^*_{CEPH} picks 18 SNPs divided into three congruent sets. All selected SNPs, denoted by a black dot, are segregating in the European sample and fixed in the Yoruba sample. The associated p-value is 0.039.
DOI: 110.1371/journal.pgen.0020105.g002

While S^*_{All} typically captures information about the oldest and deepest branches of the genealogical tree, S^*_{CEPH} provides information about branches internal to the European tree. These branches are expected to be the signature of an ancient admixture in Europe. An illustration of what S^*_{CEPH} does is provided in Figure 2.

To illustrate the efficiency of the method, we compare a scenario where there is no archaic population to another where the admixture level is set to 5% in the European population. We assume the admixture occurred 50,000 years ago and use our best-fitting model, with and without admixture. We simulate 40-kb regions and use the recombination and mutation rates that were estimated from the EGP data.

We show a quantile–quantile plot comparing both simulated distributions of S^*_{CEPH} in Figure 3. Distributions of S^*_{All} and S^*_{Yor} are not significantly affected by this admixture (unpublished data). We find that for S^*_{CEPH} the difference of both simulated means is 60% of the standard deviation (computed under the no-admixture hypothesis). With such values a power study shows that a t-test with a sample size of 110 loci would provide a power of 95% for a type I error of 5%. Hence, at least in theory, the 135 genes in the dataset are sufficient to distinguish both hypotheses.

Distribution of S^* in the Data

To test the null hypothesis of no ancient admixture, we calculate S^*_{All},S^*_{Yor} and S^*_{CEPH} for each of the 135 loci. We estimate a p-value for each locus and each statistic by running simulations under the null model described in Figure 1 and comparing the actual S^* values to the distribution of simulated S^* values.

These 135 p-values are then combined to test if the data are consistent with our null model (see Materials and Methods for details). We then obtain an overall statistic that measures the consistency of the data with our expectations. Under the null this statistic is distributed as χ^2 with $2n = 270$ degrees of freedom (n is the number of loci in the dataset). We used various models of recombination rate heterogeneity (see Materials and Methods) to assess the robustness of our findings. Results are reported in Table 2.

We first find that a constant recombination rate (within and between loci) cannot account for the distribution of S^*_{All}. This observation is consistent with the fact that the observed variability of $\hat{\rho}$ exceeds the variance of the simulated distribution under the assumption of a uniform ρ. This discrepancy is associated with the strong heterogeneity of the recombination rate in the human genome: a large fraction of the loci has a lower recombination rate than the genome-wide average, generating elevated values of S^*_{All}.

To account for this variability, we consider a random distribution for ρ fitted to reproduce the variability observed in the data (precisely by fitting the mean and the standard deviation, see Materials and Methods). This model still assumes that the recombination rate is homogeneous within a locus. We find that this model reproduces well the observed values of S^*_{All}. This consistent fit (see Table 2 and Figure 4) provides additional evidence that our null model explains the data well and that we calibrated the distribution of ρ reasonably.

We investigated a third model of recombination rate variability. In this model, we assume a uniform background rate and a random number of hotspots (parameters were estimated based on [27], see Materials and Methods). Using this model yields comparable results (see Table 2).

However, based on the values of S^*_{CEPH} we find that within the CEPH sample the level of LD is higher than predicted by our model. This discrepancy is very strong and observed for all models of recombination. Even setting the recombination rate to zero does not account for the large values of S^*_{CEPH} in the CEPH sample ($p \approx 0.04$). This observation is also true for the Yoruba sample: values of S^*_{Yor} found in the data are

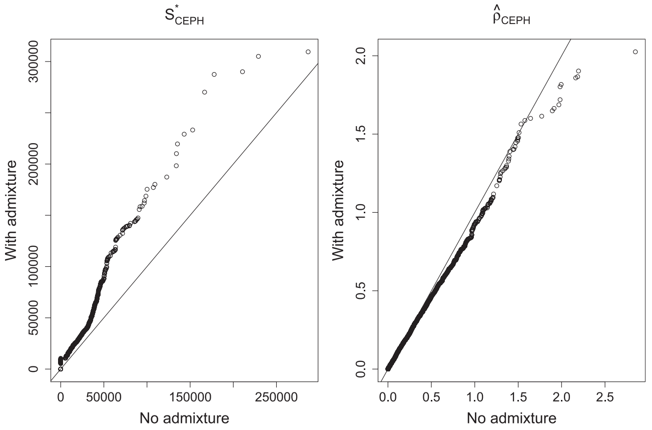

Figure 3. Quantile–Quantile Plot Comparing the Distribution of S^* (Left Graph) and the Recombination Rate $\hat{\rho}_{CEPH}$ (Right Graph), When There Is No Admixture (*x*-Axis) and When the Level of Admixture is 5% (*y*-Axis)

Since for S^*_{CEPH} many points are far away from the diagonal, we can conclude that the two models are easily distinguishable from each other. This would not be possible based on the distribution of $\hat{\rho}_{CEPH}$.

DOI: 110.1371/journal.pgen.0020105.g003

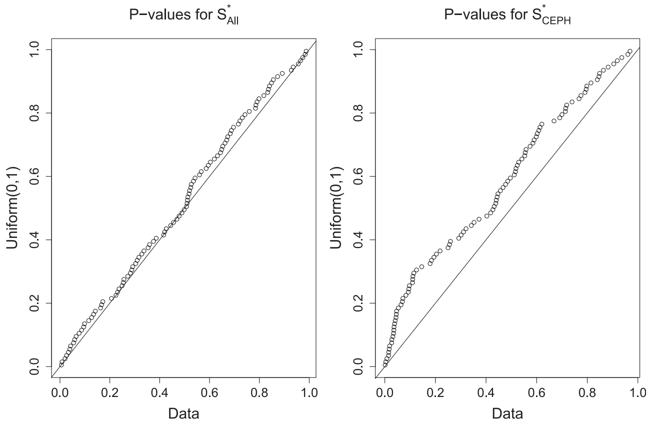

Figure 4. Quantile–Quantile Plot Comparing the Distribution of *p*-Values Associated with S^*_{All} (Left) and S^*_{CEPH} (Right) with the Expected Uniform Distribution between 0 and 1

The deviation from the diagonal line shows a discrepancy between the data and the null model for S^*_{CEPH} but not for S^*_{All}.

DOI: 110.1371/journal.pgen.0020105.g004

significantly higher than expected. The distribution of the *p*-values associated with S^*_{CEPH} and S^*_{All} are shown in Figure 4 (a list of the most significant genes in each sample is provided in Tables S1 and S2). Overall *p*-values are approximately equal to 10^{-7} for both samples, depending on the recombination model one considers (see Table 2). A complete list of the SNPs selected by S^*_{CEPH} and S^*_{Yor} is provided in Figures S5 and S6 (in the same format as Figure 2).

Evidence for Ancestral Admixture

We now consider the effect of ancestral admixture on our inference procedure and show that it significantly improves the fit of our model, indicating strong evidence in favor of some form of ancestral admixture in the history of European and West African populations. We use the following approach: for different levels of admixture we reestimate the demographic parameters, and investigate if the newly estimated demography is consistent with the observed distribution of S^*. One should remark that date of split and admixture with the archaic population are not estimated but chosen to correspond to plausible values for putative Neanderthal admixture.

We first observe that the level of admixture in Europe has little effect on the average $\hat{\rho}$ in the European sample. Table 1 shows a slight decrease of the average $\hat{\rho}$ in the CEPH sample, which is the expected trend in the presence of admixture (higher level of LD is associated with a lower estimated ρ).

Second, in the presence of admixture, the estimated demographic parameters are only slightly modified (see Protocol S1 for the associated *ms* command line). Moreover, adding a 5% level of admixture significantly improves the value of the composite likelihood: the $\log_{10}$-ratio between the maximum value estimated with a 5% admixture and no admixture equals three.

Third, this putative admixture in the European sample has a limited effect on the distribution of S^*_{All} and S^*_{Yor}. However, it increases strongly the values of S^*_{CEPH}, as shown in Table 2. A level of admixture set to 3% is still not sufficient to explain the high values of S^*_{CEPH} observed in the data ($p \approx 0.03$). We find that approximately 5% is required to account for the distribution of S^*.

Table 2. *p*-Values Associated with S^*_{All}, S^*_{Yor}, S^*_{CEPH}, and Average Values of $\hat{\rho}$ in Each Sample for Three Different Levels of Neanderthal Admixture and Three Scenarios for Recombination Rate Heterogeneity

Model for ρ	Neanderthal Admixture	Overall	Yoruba		CEPH	
		S^*_{All}	S^*_{Yor}	$\hat{\rho}$ (SD)	S^*_{CEPH}	$\hat{\rho}$ (SD)
Hotspot	0%	0.163	3.88e-08	0.36 (0.43)	5.42e-07	0.2 (0.27)
	3%	0.34	3.79e-08	0.37 (0.44)	0.0261	0.18 (0.24)
	5%	0.404	1.95e-09	0.37 (0.44)	0.275	0.18 (0.25)
Variable between loci	0%	0.0914	8.13e-09	0.38 (0.45)	1.12e-06	0.17 (0.22)
	3%	0.267	8.78e-09	0.39 (0.45)	0.0396	0.15 (0.19)
	5%	0.338	6.58e-10	0.39 (0.45)	0.358	0.15 (0.19)
Uniform	0%	8.66e-09	9.55e-13	0.38 (0.24)	1.54e-08	0.16 (0.11)
	3%	1.1e-05	7.84e-13	0.39 (0.24)	0.0195	0.14 (0.1)
	5%	5.04e-05	2.95e-14	0.39 (0.24)	0.232	0.14 (0.1)

SD, standard deviation of this estimate, computed across loci.
See Materials and Methods for details about the different recombination rate models.
DOI: 110.1371/journal.pgen.0020105.t002

Discussion

We use a model-based approach to describe the history of European and West African populations. This model predicts what pattern of polymorphism to expect in the absence of ancient admixture, and how such an event would affect the data. In the absence of admixture, the comparison of the data with our model shows a clear discrepancy that can be explained by an admixture rate of 5% in the European population. Even though we cannot exclude the possibility that an alternative demographic scenario is the cause of this pattern, this aspect of the data was chosen to be the most sensitive to an admixture event.

If the signal we observe is indeed the result of an admixture event, then these results would change our understanding of the origins of modern humans. It would indicate that archaic populations such as Neanderthals must have made a substantial contribution to the modern gene pool in Europe. We observe a similar pattern for West African populations even though a clear source population has not yet been found.

While the putative source population may not be as obvious as in Europe (Neanderthals), the fossil record shows that transitional forms of Homo were widespread in Africa, even after the time of emergence of modern humans. Other genetic studies have also found evidence for ancient structure in African populations [28–30]. In two of the three studies [28,30], the divergent lineage was found only in Pygmies, which suggests that the African population source differs from the European one.

Our model was designed to be as simple as possible while still capturing the main features of human polymorphism. We assessed qualitatively its goodness of fit and we found that it fits the data well: both the statistics we fitted and also the estimates of the recombination rate in both populations are consistent with the data. Our model makes simpler and fewer assumptions about human demography than a previous study with different findings [31], and we believe this makes our estimates of significance more reliable.

Our inference procedure based on summary statistics has similarities to two previous studies [17,26]. However, our statistical approach differs as both of these studies try to minimize an ad-hoc distance between simulations and data using the mean and variance of the summary statistics computed across loci. Instead, we estimate the composite likelihood independently for each locus, which should be more informative. Moreover, [26] use ascertained markers so their results are sensitive to the particular method used to correct for the ascertainment. In addition, the scale of the NIEHS-EGP dataset is much larger than the datasets used in both of these studies: Voight et al. [17] resequenced a total of 118 kb whereas the 135 loci we analyze add up to 3,305 kb of sequences for a total of 13,460 markers (compared with 3,738 markers studied in [26]).

We found that the choice of summary statistics is very important in our inference procedure. In particular, not incorporating a statistic which measures the level of migration between European and African branches leads to a different maximum of the composite likelihood where the divergence date is much lower and the estimates of LD are biased. If we had not added this summary statistic we could not have observed this poor goodness of fit. Hence, we tried to fit in our inference procedure all components of the data which seemed relevant. However, we cannot exclude that an important feature has been missed because our summary statistics cannot measure it.

Our inference procedure cannot clearly reject an island model where both populations remain separated indefinitely with a low level of migration. Using a χ^2 approximation, the p-value associated with this null hypothesis is 10^{-3}. This is relatively low but our composite-likelihood approach is likely to narrow confidence intervals, and given the uncertainty regarding our model (constant migration rate in particular) we cannot clearly exclude this model. However, the discrepancy between observations and expectations remain unchanged when looking at S^*_{Yor} and S^*_{CEPH} ($p < 10^{-7}$) and this choice does not affect significantly our main findings. We note that an equilibrium island model, along with all models incorporating a substantial element of ancient admixture, is not compatible with simple forms of the RAO model.

We investigated the pattern of polymorphism of the SNPs selected by our method in both other samples available in the NIEHS dataset: Hispanics and Han Chinese. We found that 75% of the SNPs selected by S^*_{Yor} in the Yoruba sample are fixed in the Hispanic and Chinese samples. However this number is not significantly different from other SNPs segregating in the Yoruba sample and fixed in the CEPH sample. Most SNPs selected by S^*_{CEPH} in the CEPH sample are variable in the Hispanic (90%) and Chinese (50%) samples. These numbers also do not differ significantly from other SNPs fixed in the Yoruba sample and segregating in the CEPH sample. However, there is no clear expectation regarding those proportions because we do not precisely know where and when the admixture occurred. In addition, the pattern of polymorphism has been affected by recent migrations, in particular between non-African populations.

Some alternate explanations can potentially explain the elevated values of S^*, including selective effects. Because natural selection tends to affect single loci, while demography affects the whole genome, we believe that considering 135 different loci allows us to capture the underlying demographic signal. Nevertheless, a strong selective sweep can generate a similar pattern of elevated level of LD. Investigating the function of the genes selected based on S^* did not show any significant pattern (see Tables S1 and S2 for a complete list). We also compared these genes with two genome-wide scans for selection [32,33]. We looked for our most significant genes among those analyzed in [33] and did not find significant correlations. Likewise, the ten most significant genes selected by S^*_{CEPH} are not identified as positively selected by Voight et al. [32] in Europeans. One gene *(xrcc4)* is identified in the Yoruba sample by both studies. However, we note that we find four significant genes ($p < 0.05$ for S^*_{CEPH}) in the ADH cluster, where Voight et al. [32] find a strong signal of selection in the East Asian sample.

An advantage of our method is that in addition to showing some evidence in favor of a significant rate of admixture, we also specifically pick a subset of candidate "archaic" SNPs. The only way to be certain of the answer would be to verify that these SNPs are indeed mutated in the DNA of Neanderthal fossils. Estimating the significance of the observed pattern is difficult because modeling human demography is a complex task. However, it is likely that if there has been a significant level of admixture, the SNPs selected by S^* are the best available trace of this event. Such data is not available yet but technologies are currently being developed to sequence fossil nuclear DNA (cf. [13,14]). We should note, though, that fossil nuclear DNA sequencing studies may have a difficult time distinguishing between archaic human DNA sequences and modern human contaminants.

Materials and Methods

Dataset. Our approach requires resequencing polymorphism data free of ascertainment bias and we based our analyses on data from the EGP. The data were generated at the University of Washington (Seattle, Washington, United States) (see [18,19]).

Our analysis requires that the ethnicity of the samples is known. 135 out of the 505 genes in the EGP dataset fit this criterion. The average length of the genes included in the study is 50 kb. The average sequenced length is 25 kb. We restricted our study to the 12 Yoruba and 22 Caucasian (CEPH) individuals. We excluded genes on the sex chromosomes. Indels as well as SNPs with more than two alleles are also excluded.

Demographic inference. We make several assumptions about the demographic scenario in order to limit the number of parameters to estimate. First, we assume that the population growth is 100-fold in each population. Second, we assume that the bottleneck lasts 1,000 years, and we only estimate the reduction of the population size during this period. The six parameters left to be estimated are: the date of the beginning of growth (one parameter for each population), the date of the bottleneck, and its intensity (reduction of the population size during the bottleneck), the date of divergence between European and African populations, and the migration rate after this divergence.

We use a grid for the parameter values on which we estimate the composite likelihood. The grid consists of ten values per parameter (10^6 total) and we refined this grid several times to locate the maximum of the composite-likelihood surface. For each value of the parameters, the same set of simulations is used across different loci in the data (using the mean sequenced length). The log-likelihood is then summed across loci to obtain the final value.

The first set of summary statistics consists of Tajima's D in each sample, Fu and Li's D^* in the CEPH sample, and F_{ST}. We assume that the distribution of these parameters is multivariate Gaussian and for each value of the parameter on the grid, we estimate the vector of means and the covariance matrix using Monte Carlo simulations.

The second set of summary statistics divides the SNPs at a locus in four categories: private in the CEPH sample, private in the Yoruba sample, and segregating in both samples at low or high frequency (we set the threshold at 10%). For each branch of the ARG [24], all mutations on this branch will belong to a single category. Hence, given one realization of the ARG, one can parse this graph to estimate the probability (f_1,f_2,f_3,f_4) for a random SNP to belong to each of the four classes. Conditional on the ARG and the total number of SNPs n, the distribution of (n_1,n_2,n_3,n_4) is multinomial and can be obtained explicitly. By simulating a large number of ARGs and averaging the computed probabilities for each simulated ARG τ_i, we obtain an estimate of the likelihood at a given locus:

$$\mathbb{P}(n_1, n_2, n_3, n_4 \mid n) = \sum_i \mathbb{P}(n_1, n_2, n_3, n_4 \mid n, \tau_i) \tag{1}$$

For each point on the grid, we found that 80,000 simulated ARGs are needed to obtain a precise value of the likelihood. The computation of the likelihood of the data at one point of the grid requires approximately five minutes on a 1.8-GHz Opteron processor. A significant amount of time is saved by stopping the computation after 20,000 simulations if the estimated likelihood at this point of the grid is clearly lower than the current maximum. The total computing time over the 10^6 points of the grid takes approximately three days on 100 processors. All simulations in this paper use a modified version of *ms* [25].

Recombination rate. We consider various scenarios for the recombination rate. To describe the simulated distribution of ρ, we first estimate the average mutation rate θ, using all available loci, and parametrize the recombination rate ρ in terms of $f = \rho/\theta$. We consider first a model where ρ is uniform within and between loci with $f = 0.375$.

A second model includes variability between loci but not within. Specifically, we set the distribution of f to be a gamma distribution with mean $\mu = 0.375$ and standard deviation $\sigma/\mu = 0.29 \times \sqrt{500{,}000/l}$ where l is the length of the locus we simulate (to account for a larger variability of the overall ρ in shorter loci). Within each simulated locus, the rate is uniform. With these parameters, the mean and the standard deviation of $\hat{\rho}$ are consistent with observations (see Table 1).

Finally, we consider a model consisting of a background rate and a random number of hotspots. The background rate is variable with mean $\mu = \bar{f} = 0.21$ and standard deviation $\sigma/\mu = 0.2 \times \sqrt{500{,}000/l}$. The distribution of the number of hotspots is Poisson, whose intensity is chosen to obtain on average one hotspot per 57 kb (as estimated in [27]). Hotspots have a 2-kb width and an intensity 60 times higher than the background rate. Recombination rates for each locus are scaled so that the average overall recombination rate is proportional to the DECODE estimates [16]. With these parameters 80% of recombination events occur on average in 15% of the sequence (as estimated in [27]). Note that with this hotspot model the overall expected recombination rate must be significantly higher than under a uniform model to account for the distribution of $\hat{\rho}$ (including hotspots $\bar{f} = 0.65$ under this model).

Definition of S^*. We define S^* as follows:

$$S^* = S(I) = \max_{J \subset \{1,\ldots,n\}} S(J) \tag{2}$$

where

$$S(i_1, \ldots, i_k) = \sum_j S(i_j, i_{j+1}) \tag{3}$$

$\{1,\ldots,n\}$ designates the set of SNPs at this locus and I is the subset of SNPs that maximizes the score.

The score is computed as a sum over successive pairs of SNPs in the optimum subset I. Note that SNPs in I are not necessarily adjacent. We tried various definitions for $S(i,j)$, and we chose the one that maximized the difference between our null model and the same model where the level of admixture is set to 10%.

We call the distance between two SNPs the number of chromosomes in the data at which the genotypes differ. Note that when—for a given pair of SNPs—an individual is a double heterozygote, we assume that the distance between both SNPs is zero (in other words we assume that both genotypes are in phase). If the total distance (summed over all individuals) between two SNPs is zero, both sites are congruent and the score is equal to the distance in bp between them plus 5,000. If this distance is greater than five, the score is set to $-\infty$. If the distance is between one and five, the score is equal to $-10{,}000$. We also impose a minimum distance between sites within the optimum set I of at least 10 bp, to avoid contiguous and congruent SNPs (overrepresented in the human genome) to bias our estimates.

We also need to account for missing data. For a pair of SNPs to be congruent, we allow no more than two chromosomes with the property that a missing call at one SNP is associated with the minor allele at the other SNP. When one of the two SNPs has a minor allele frequency of two, we make this criteria more stringent and allow only one such chromosome.

The computation of S^* can be done efficiently using a forward–backward algorithm, sometimes called dynamic programming and typically used in the Smith-Waterman algorithm [34]. Specifically, if we define:

$$S_j^* = \underset{J \subset \{1,\ldots,j-1\}}{max} S(J \cup \{j\}), \tag{4}$$

then we have the recursion:

$$S_{j+1}^* = \underset{k=1,\ldots,j}{max} [S_k^* + S(k, j+1)] \tag{5}$$

Computation of *p*-values. Because each locus has a different length, and different regions were not scanned for polymorphism, different sets of simulations (which reproduce these precise characteristics) are used for each locus to estimate the distribution of S^*_{All},S^*_{Yor}, and S^*_{CEPH}. On each simulated ARG we place a number of mutations equal to the number of variable sites at this locus in the data. This is done to avoid biases due to variability in the mutation rate. Also, a random fraction of the genotyping calls is labeled missing. The probability of being missing is equal to the fraction of missing calls in the data at this locus.

For each simulated genealogical tree, we obtain a value for S^*_{All},S^*_{Yor}, and S^*_{CEPH} defined as $\mathbb{P}(S^* \geq S^*_{\mathrm{data}})$ where S^*_{data} is the value computed from the data. We can obtain an overall *p*-value by using the fact that if $(X_i)_{i=1}^n$ is uniformly distributed between 0 and 1 then $\sum_{i=1}^n -2\log(X_i)$ is distributed as χ^2 with $2n$ degrees of freedom.

Supporting Information

Figure S1. Profile Likelihood for the Demographic Inference

Found at DOI: 110.1371/journal.pgen.0020105.sg001 (100 KB PDF).

Figure S2. QQ-Plot between Simulated and Observed Values for the First Set of Summary Statistics

Found at DOI: 110.1371/journal.pgen.0020105.sg002 (100 KB PDF).

Figure S3. QQ-Plot between Simulated and Observed Values for the Second Set of Summary Statistics

Found at DOI: 110.1371/journal.pgen.0020105.sg003 (120 KB PDF).

Figure S4. Comparison of Frequency Spectrum between Data and Best-Fitting Model

Found at DOI: 110.1371/journal.pgen.0020105.sg004 (64 KB PDF).

Figure S5. List of Selected SNPs in the CEPH Sample

Found at DOI: 110.1371/journal.pgen.0020105.sg005 (2.8 MB PDF).

Figure S6. List of Selected SNPs in the Yoruba Sample

Found at DOI: 110.1371/journal.pgen.0020105.sg006 (2.8 MB PDF).

Protocol S1. *ms* Command Lines Associated with the Best-Fitting Models

Found at DOI: 110.1371/journal.pgen.0020105.sd001 (28 KB PDF).

Table S1. Most Significant Genes in the CEPH Sample

Found at DOI: 110.1371/journal.pgen.0020105.st001 (21 KB PDF).

Table S2. Most Significant Genes in the Yoruba Sample

Found at DOI: 110.1371/journal.pgen.0020105.st002 (29 KB PDF).

Acknowledgments

We thank M. Hammer, M. Przeworski, and three anonymous referees for helpful comments on a previous draft of this paper.

Author contributions. VP and JDW analyzed the data and wrote the paper.

Funding. This work was supported by NSF grant BCS-0423123 to JDW.

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

References

1. McBrearty S, Brooks AS (2000) The revolution that wasn't: A new interpretation of the origin of modern human behavior. J Hum Evolution 39: 453–563.
2. Wolpoff MH (1999) Paleoanthropology. New York: Mc Graw-Hill. 878 p.
3. Wolpoff MH, Wu X, Thorne AG (1984) Modern Homo sapiens origins: A general theory of hominid evolution involving the fossil evidence from East Asia. In: In Smith FH, Spencer F, editors. The origins of modern humans: A world survey of the fossil evidence. New York: Liss. pp. 411–483.
4. Stringer CB, Andrews P (1988) Genetic and fossil evidence for the origin of modern Humans. Science 239: 1263–1268.
5. Duarte C, Mauricio J, Pettitt PB, Souto P, Trinkaus E, et al. (1999) The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia. PNAS 96: 7604–7609.
6. Foley R, Lahr MM (1997) Mode 3 technologies and the evolution of modern humans. Camb Arch 7: 3–36.
7. Brauer G (1984) The Afro–European sapiens hypothesis and hominid evolution in East Asia during the late Middle and Upper Pleistocene. Cour Forschungsinst Senckenb 69: 145–165.
8. Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, et al. (1997) Neandertal DNA sequences and the origin of modern Humans. Cell 90: 19–30.
9. Krings M, Capelli C, Tschentscher F, Geisert H, Meyer S, et al. (2000) A view of Neandertal genetic diversity. Nat Genet 26: 144–146.
10. Serre D, Langaney A, Chech M, Teschler-Nicola M, Paunovic M, et al. (2004) No evidence of Neandertal mtDNA contribution to early modern humans. PLoS Biol 2: e57. DOI: 10.1371/journal.pbio.0020057
11. Ovchinnikov IV, Gotherstrom A, Romanova GP, Kharitonov VM, Liden K, et al. (2000) Molecular analysis of Neanderthal DNA from the Northern Caucasus. Nature 404: 490–493.
12. Nordborg M (1998) On the probability of Neanderthal ancestry. Am J Hum Genet 63: 1237–1240.
13. Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, et al. (2005) Genomic sequencing of Pleistocene cave bears. Science 309: 597–599.
14. Poinar HN, Schwarz C, Qi J, Shapiro B, MacPhee RDE, et al. (2006) Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. Science 311: 392–394.
15. Wall JD (2000) Detecting ancient admixture in humans using sequence polymorphism data. Genetics 154: 1271–1279.
16. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, et al. (2002) A high-resolution recombination map of the human genome. Nat Genet 31: 241–247.
17. Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, et al. (2005) Interrogating multiple aspects of variation in a full resequencing dataset to infer human population size changes. PNAS 102: 18508–18513.
18. Livingston RJ, Niederhausern A, Jegga AG, Crawford DC, Carlson CS, et al. (2004) Pattern of sequence variation across 213 environmental response genes. Genome Res 14: 1821–1831.
19. Nickerson DA, Rieder MJ, Crawford DC (2003) An overview of the environmental genome project, essays on the future of environmental health research. EHP Online 113.
20. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Pop Biol 7: 256–276.
21. Hudson RR (2001) Two-locus sampling distributions and their applications. Genetics 159: 1805–1817.
22. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
23. Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583–589.
24. Griffiths RC, Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. J Comp Biol 3: 479–502.
25. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics 18: 337–338.
26. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15: 1576–1583.
27. HapMap (2005) A haplotype map of the human genome. *Nature* 437: 1299–1320.
28. Garrigan D, Mobasher Z, Severson T, Wilder JA, Hammer MF (2005) Evidence for archaic Asian ancestry on the human X chromosome. Mol Biol Evol 22: 189–192.
29. Barreiro LB, Patin E, Neyrolles O, Cann HM, Gicquel B, et al. (2005) The heritage of pathogen pressures and ancient demography in the human innate-immunity cd209/cd209l region. Am J Hum Genet 77: 869–886.
30. Hayakawa T, Aki I, Varki A, Satta Y, Takahata N (2006) Fixation of the human-specific CMP-N-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. Genetics 172: 1139–1146.
31. Currat M, Excoffier L (2004) Modern humans did not admix with Neanderthals during their range expansion into Europe. PLoS Biol 2 (12): e421.
32. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e42. DOI: 10.1371/journal.pbio.0040042
33. Clark AG, Glanowsk S, Nielsen R, Thomas PD, Kejariwal A, et al. (2003) Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. Science 302: 1960–1963.
34. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195–197.
35. Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, et al. (1995) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet 19: 233–240.
36. Rieder MJ, Taylor SL, Clark AG, Nickerson DA (1999) Sequence variation in the human angiotensin converting enzyme. Nat Genet 22: 59–62.