Reconstructing human origins in the genomic era

Daniel Garrigan and Michael F. Hammer

Abstract | Analyses of recently acquired genomic sequence data are leading to important insights into the early evolution of anatomically modern humans, as well as into the more recent demographic processes that accompanied the global radiation of *Homo sapiens*. Some of the new results contradict early, but still influential, conclusions that were based on analyses of gene trees from mitochondrial DNA and Y-chromosome sequences. In this review, we discuss the different genetic and statistical methods that are available for studying human population history, and identify the most plausible models of human evolution that can accommodate the contrasting patterns observed at different loci throughout the genome.

Hominin
All the taxa on the human lineage after the split from the common ancestor with the chimpanzee.

Neutral DNA polymorphism
Nucleotide variants that segregate in a population but have frequencies that are not influenced by natural selection.

Demographic processes
Changes in population size, distribution and structure.

Bottleneck
A transient reduction in the abundance of a population. This could occur, for example, because of environmental catastrophe or after the founding of a new population.

Division of Biotechnology, University of Arizona, Tucson AZ 85721, USA. Correspondence to M.F.H. e-mail: mfh@u.arizona.edu
doi:10.1038/nrg1941

In 1967 two hominin skulls, thought to be 130,000 years old, were unearthed in the Omo Valley of south-western Ethiopia[1]. The volcanic pumice in which the two skulls were embedded was recently dated to 195,000 years ago (195 kya) (REF. 2), making one of these morphologically distinct forms (Omo I) the earliest known evidence of the anatomically modern human (AMH) phenotype. From this equatorial African homeland, *Homo sapiens* succeeded in radiating across Africa, Eurasia, near Oceania and the Americas by 15 kya, with the remote islands of the Pacific being settled as recently as 850 years ago.

The evolution of *H. sapiens* can be conceptually divided into two distinct epochs, which can be reconstructed not only through palaeontological and archaeological records, but also from patterns of neutral DNA polymorphism in the human genome. The first epoch includes the evolutionary processes that took place in the lineage that culminated in the emergence of AMH. The aforementioned fossils from the Omo Valley in Ethiopia represent one (or possibly two) of many taxa of *Homo* that ranged throughout Africa, Europe and Asia for nearly 2 million years[3,4]. Do the genes of AMH trace back to a single ancestral population in East Africa, or was the evolution of the modern phenotype in sub-Saharan Africa a gradual process, with assimilation of genes from more widely distributed, distinct forms of archaic *Homo*? The answer to this question has important implications for understanding how the suite of traits that distinguishes us as modern humans was assembled in our genome. The second epoch focuses on the demographic processes that accompanied the global diaspora of AMH after their origin in Africa. Was there a bottleneck or bottlenecks associated with migrations of AMH out of Africa? What were the sizes of ancestral populations in the Pleistocene, and when did they first begin to grow dramatically? When did ancestral populations diverge from one another and how much gene flow occurred among their descendant populations? Did AMH completely replace archaic forms without interbreeding? The answers to these questions have implications that go well beyond the desire to understand the evolution of our species: formulating a null model of human demographic history from selectively neutral polymorphisms facilitates the discovery of genetic variants that might contribute to disease susceptibility or other components of human fitness[5].

Until recently, efforts to reconstruct the origin and dispersal of AMH primarily focused on DNA sequence data sets obtained from the two haploid compartments of the genome, namely mitochondrial DNA (mtDNA) and the non-recombining Y chromosome (NRY). Early research on mtDNA and the NRY had a decisive role in the long-standing palaeontological debate over human origins[6] by providing a relatively simple picture of human demographic history: the AMH phenotype originated in a small, isolated population in Africa during the Late Pleistocene. This population is then thought to have completely replaced archaic forms of *Homo* as it expanded its size and range throughout the Old World[7–9]. Although this widely acknowledged model is known by many names (such as Recent African Replacement, Out of Africa 2, Garden of Eden), we refer to it here as the 'single origin' model. Today there is an abundance

© 2006 Nature Publishing Group

Box 1 | **Genomic compartments**

When making inferences about human demographic history from DNA sequence data, it is important to take account of the processes that differentially affect autosomal, X-chromosomal, non-recombining Y-chromosomal (NRY) and mitochondrial DNA. There are differences in copy number and inheritance patterns, as well as in recombination and mutation rates, among these four genomic compartments (see accompanying table).

Genes carried on mtDNA or the NRY are inherited from only one parent. X-linked genes are present in two copies in females and only one copy in males, but can be transmitted by either sex. Autosomal genes are maintained on two chromosomes in both sexes. As a result, when equal numbers of males and females are breeding, the relative effective population size (N_e) of the autosomes, X chromosomes, NRY and mtDNA is 4:3:1:1, respectively. The reduced N_e of the haploid loci is expected to result in shallower times to the most recent common ancestor (TMRCA), higher levels of differentiation among human populations and possibly smaller effects of natural selection. The larger N_e of the X chromosome and autosomes means that loci in these genomic compartments are expected to have deeper ancestry, allowing inferences of evolutionary processes that took place well before the TMRCA of the haploid regions.

Another point to consider is that the haploid regions each behave evolutionarily as a single locus, and selection acting on any particular site affects patterns of polymorphism on the entire mtDNA and NRY. By contrast, the X chromosome and autosomes experience recombination, a force that tends to decouple the genealogical histories of sites along chromosomes (that is, neutral sites and those under selection). Non-coding sequences from regions of the genome that experience higher rates of recombination are less likely to be perturbed from a neutral, equilibrium state by the effects of selection at linked sites. Therefore, these regions are the best candidates for evaluating several important parameters of interest in models of human evolution, including TMRCA and N_e. Even under neutrality, each non-recombining locus provides only a single outcome of the highly stochastic evolutionary process, and so many independent loci are needed to infer human evolutionary history.

	Genomic compartment			
Feature	**Autosomes**	**X chromosomes**	**NRY**	**mtDNA**
Location	Nuclear	Nuclear	Nuclear	Cytoplasmic
Inheritance	Bi-parental	Bi-parental	Uni-parental	Uni-parental
Ploidy	Diploid	Haploid–diploid	Haploid	Haploid
Relative N_e	4	3	1	1
Recombination rate	Variable	Variable	Zero	Zero
Mutation rate	Low	Low	Low	High

of DNA polymorphism data from two other genomic compartments — the X chromosome and the autosomes. Analyses of these data, with increasingly sophisticated computational tools, are yielding new insights into both epochs of human evolutionary history.

In this review, we concentrate both on the methods used to reconstruct human evolutionary history from DNA resequencing data and on the population genetics models that are supported by all the available data. We pay particular attention to how recent findings from X-chromosomal and autosomal loci contrast with earlier inferences gleaned from mtDNA and the NRY. Differences in ploidy and other inherent properties of different genomic compartments mean that they provide different levels of resolution for the two epochs of human evolution (BOX 1). Finally, we attempt to synthesize inferences from the literature with the goal of providing possible alternatives to the single origin model.

Pleistocene
An epoch of the Quaternary period beginning 1.8 million years ago and transitioning to the Holocene epoch approximately 10,000 years ago. The Pleistocene is characterized by a cool climate and extensive glaciation of northern latitudes.

Population phylogeny
The hierarchical relationship among individual populations, typically inferred from pairwise genetic differences between populations.

Evolutionary analyses of molecular genetic data

The first efforts to use genetic data to make inferences about human history relied on allele frequencies of classical protein polymorphisms to reconstruct population phylogeny[10,11]. With the advent of DNA sequencing, population history was inferred by following one of two routes: reconstruction of gene trees or analysis of summary statistics[12]. However, these two approaches are relatively inflexible when a wide range of candidate models need to be considered. Simulation methods that are based on a third approach, the coalescent approach, allow greater flexibility for testing a variety of non-equilibrium models. The three modern approaches are described below. As we discuss, the coalescent approach is particularly useful for inferring the history of changes in human population size and structure.

Analysis of gene trees. Unlike allele frequency data, DNA sequences lend themselves to analyses that are based on the reconstruction of evolutionary trees for individual loci (gene trees). Such methods are primarily implemented in the analysis of non-recombining regions of the human genome. By incorporating homologous sequences from closely related species (such as humans and chimpanzees), the topology of a gene tree can be inferred along with the direction and rate of mutation.

The field of intraspecific phylogeography[13] combines information about the ancestral–descendant relationships in the haplotype tree with the frequency and distribution of haplotypes in a sample of sequences. This approach allows qualitative inferences to be made, such as the place of the most recent common ancestor (PMRCA) for a given genomic region and the migration routes of derived haplotypes. Because the gene trees from both mtDNA[7,14,15] and the NRY[16,17] are paraphyletic with respect to extant African populations, Africa was inferred to be the PMRCA. Analyses of autosomal and X-linked loci also typically point to Africa as the PMRCA. For example, Takahata *et al.*[18] found that 9 of the 10 loci they examined had gene trees that unambiguously rooted in Africa. One clear exception to this rule of human gene tree topology comes from a recent study of polymorphism in the X-linked ribonucleotide reductase M2 polypeptide pseudogene 4 (*RRM2P4*), which yields a gene tree that roots in East Asia[19] (BOX 2a,b).

Another important piece of information comes from the estimated time depth of gene trees. In the context of gene-tree analysis, estimates of the time to the most recent common ancestor (TMRCA) are based on the assumptions of the molecular clock, in which the number of nucleotide mutations separating two sequences is a linear function of time. The estimated TMRCA for the mtDNA gene tree is 171 kya, which is based on complete sequences of the mitochondrial genome from a global sample of humans[15]. For a neutrally evolving locus, the expected TMRCA is a function of the effective population size (N_e) — therefore, we expect a three or fourfold older TMRCA for the X chromosome and the autosomes, respectively (BOX 1).

Interestingly, the TMRCA for the male-specific NRY is roughly half that of the female-specific mtDNA[20]. This

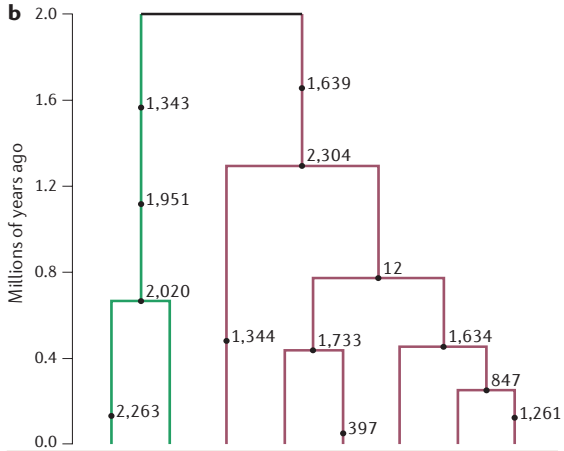
© 2006 Nature Publishing Group

Box 2 | **Different approaches to analysing human polymorphism data**

a

Nucleotide position	Chimpanzee haplotype	Human haplotypes A	B	C	D	E	F	G	H
12	T	.	.	.	C	C	C	C	C
397	G	.	.	.	.	C	.	.	.
847	C	.	.	.	.	.	.	T	T
1,261	A	.	.	.	.	.	.	.	G
1,343	T	A	A	.	.	.	.	.	.
1,344	G	.	.	T	.	.	.	.	.
1,634	T	.	.	.	.	.	C	C	C
1,639	A	.	.	G	G	G	G	G	G
1,733	A	.	.	.	G	G	.	.	.
1,951	T	G	G	.	.	.	.	.	.
2,020	G	A	A	.	.	.	.	.	.
2,263	C	A	.	.	.	.	.	.	.
2,304	T	.	.	C	C	C	C	C	C

b



	A	B	C	D	E	F	G	H
Africa	.	.	1	2	1	1	5	.
Asia	2	1	.	8	.	.	.	1
Europe	.	.	.	9	.	.	1	.
America	.	.	.	3	.	.	.	7

c

Population	*n*	*S*	θ	π	*D*
Africa	10	6	2.12	2.16	0.068
Europe	10	3	1.06	0.60	−1.562
Asia	12	11	3.64	3.74	0.114
America	10	4	1.41	1.87	1.229

Polymorphism in the X-linked ribonucleotide reductase M2 polypeptide pseudogene 4 (*RRM2P4*) suggests that divergent evolutionary histories could be sampled from the genome. Whereas data from all genomic compartments largely agree on placing the most recent common ancestor (MRCA) in Africa, analysis of the *RRM2P4* gene points to East Asia as the place of the MRCA by all three approaches described below. All three approaches begin with the collection of DNA sequence data from a sample of individuals who represent different continental populations. Panel **a** shows the haplotype obtained from a single common chimpanzee and eight haplotypes (A–H) obtained from 42 human individuals. There are a total of 13 polymorphic sites in 2,385 resequenced nucleotides; identity with the ancestral state (represented by the chimpanzee sequence) is depicted by a dot.

The phylogenetic approach, shown in panel **b**, begins with the reconstruction of a single most parsimonious gene tree, which can be constructed from the *RRM2P4* locus because there is no evidence for recombination or back mutation. *RRM2P4* nucleotide polymorphism is partitioned into two major branches. Shown below the *RRM2P4* gene tree are the frequencies of eight haplotypes in the four sampled continental populations (Africa, Asia, Europe and America). On the left side of the root, there are only three Asian sequences (haplotypes A and B; green), whereas both Asian and non-Asian sequences are to the right of the root (C–H; pink). The numbers on the branches correspond to the nucleotide position of the mutation shown in panel **a**. Under the phylogenetic branching model, the simplest explanation for this gene-tree topology is that Asia is the place of the most recent common ancestor (PMRCA). The coalescent approach is illustrated by the timescale to the left of the *RRM2P4* gene tree in panel **b**, and estimates the ages of individual mutations and the time to a most recent common ancestor (TMRCA). These quantities were estimated by coalescent simulation, conditioned on the topology of the gene tree. When one million coalescent simulations are carried out under an island model of population structure, the probability of the PMRCA being in Asia is indeed the highest at $P = 0.92$, whereas the values are $P = 0.05$ for Africa, $P = 0.01$ for Europe and $P = 0.02$ for the Americas.

Certain inferences about *RRM2P4* polymorphism can also be drawn from summary statistics (panel **c**). In the table summarizing *RRM2P4* polymorphism, the second and third columns give sample sizes (n) for each population and the number of polymorphic nucleotide sites (S). The statistics θ and π are two different estimates of $3N_e\mu$, where N_e is the effective population size and μ is the mutation rate. The statistic θ (equation 1) is based on the number of segregating sites.

$$\theta = S / \sum_{i=1}^{n-1} i^{-1} \qquad (1)$$

However, π is shown as equation 2, where d_{ij} is the number of nucleotide differences between sequences i and j.

$$\pi = \binom{n}{2}^{-1} \sum_{i \neq j} \sum_{j} d_{ij} \qquad (2)$$

The penultimate column gives Tajima's D statistic, and is a summary of the polymorphism frequency distribution (equation 3).

$$D = \pi - \theta / \sqrt{var(\pi - \theta)} \qquad (3)$$

Tajima's D is negative in Europeans, indicating an excess of rare polymorphisms, whereas it is positive in Americans, indicating a deficiency of rare polymorphisms. However, in both of these examples, D does not significantly differ from its expected value of zero. The relative values of θ and π are consistent with a larger Asian population size for this particular region of the genome. Diagrams modified with permission from REF. 19 © (2005) Oxford University Press.

Summary statistics
Statistics that describe some aspect of polymorphism data, such as the number of polymorphic sites, the distribution of mutation frequencies or the extent of association between linked polymorphisms. Summary statistics are often estimates of parameters in an evolutionary model.

Coalescent approach
A probabilistic construct that describes the hierarchical common ancestry of a sample of gene copies. The probability that two gene copies share a common ancestor (or coalesce) in the preceding generation is proportional to the reciprocal of the size of the entire population.

Haplotype
A contiguous DNA sequence of arbitrary length along a chromosome that has a primary structure that is distinct from that of other homologous regions in a given population.

Place of the most recent common ancestor
The geographical area, of arbitrary scale, where the ancestry of a current sample of gene copies can be traced back to a single, endemic ancestral population.

Derived
The state of genotypic or phenotypic character, possessed by some biological entity, which has mutated from a common ancestral state.

Paraphyletic
When the common ancestor of one natural group is shared with any other such group.

Time to the most recent common ancestor
The number of generations back in time when a single gene copy gave rise to all of the gene copies in a contemporary sample. If n gene copies are sampled from a population of size N, the time to a most recent common ancestor for an autosomal locus is expected to be $4N(1 - 1/n)$ generations.

© 2006 Nature Publishing Group

Effective population size
The number of individuals of a given generation that contribute gametes to the subsequent generations. This abstract quantity depends on the breeding sex ratio, number of offspring per individual and type of mating system.

Island model of population structure
A commonly used model to describe gene flow in a subdivided population in which each subpopulation of constant size, *N*, receives and gives migrants to each of the other subpopulations at the same rate, *m*. Under the Island model, $F_{ST} = 1/(4Nm + 1)$.

Standard neutral model
A population genetics model that assumes all individuals in a population are replaced by their offspring each generation, so that the population size remains constant, mating occurs randomly and each parent produces a Poisson-distributed number of offspring. Under these conditions, the model predicts the fate of mutations that are not affected by natural selection.

Harmonic mean
One method for calculating an average, defined as the reciprocal of the arithmetic mean of the reciprocals of a specified set of positive numbers.

Tajima's *D*
A statistic used to test the standard neutral model for a given region of DNA sequence. It is the standardized difference between the number of pairwise nucleotide differences and the total number of segregating sites.

Frequency spectrum
The distribution of polymorphism frequencies in a sample of DNA sequences. For example, 30% of polymorphisms might occur in a single gene copy, 20% in two gene copies, and so on. Under the standard neutral model, the frequency spectrum is expected to follow a geometric distribution.

difference in male and female TMRCA has been attributed to a lower male N_e resulting from a higher variance in male reproductive success[20], natural selection[21], higher variance in mtDNA mutation rates[11] and/or stochasticity in the evolutionary process[22].

Throughout later sections, we revisit the concept of TMRCA and how estimates of TMRCA vary across the genome.

Summary statistics. In most cases, the analysis of gene trees does not explicitly assume an underlying population genetics model, so that it is difficult to test alternative hypotheses. However, another widely used approach uses a statistical summary of DNA sequence data and compares the observed value with that expected under a particular population genetics model. The expected values of many summary statistics are available under the standard neutral model[23], which assumes that N_e remains stationary over time, and that evolutionary forces have reached equilibrium (BOX 2c). It should be noted that it can be difficult to obtain expectations for many summary statistics under non-equilibrium models, and so inference is largely limited to accepting or rejecting the standard neutral model. Two fundamental summary statistics of polymorphism data are the number of segregating sites[23] and the average number of pairwise nucleotide differences[24] in a sample of sequences. Both summary statistics are estimates of the population mutation rate (θ). If the neutral mutation rate, μ, is known, then these summaries can be used to estimate N_e under the assumptions of the standard neutral model, as $\theta = 4N_e\mu$ for an autosomal locus.

Estimates of human N_e from summary statistics consistently find that it is on the order of 10,000 individuals, which is conspicuously smaller than that of other great ape species. Assuming an equal sex ratio, the estimates of female N_e from human mtDNA are close to the total consensus estimate of 10,000 individuals[9], whereas X-chromosome and autosomal estimates range from 7,500 to 32,300 (REF. 25) and from 10,000 to 15,000 (REFS 26–28), respectively. Human N_e values are generally half the estimates of chimpanzee N_e, which range from 12,000 to 50,000 for autosomal loci[29,30]. Likewise, gorilla and orangutan N_e are estimated to be about twice and more than three times that of humans, respectively[31]. Interpreting these relative great ape values of N_e under the standard neutral model leads to the conclusion that humans have been the least abundant species. However, many alternative interpretations that violate the assumptions of the standard neutral model are also plausible. Long-term N_e behaves as a harmonic mean and is therefore disproportionately influenced by low values. Therefore, one probable explanation for the apparent small size of the ancestral human population is that human N_e has fluctuated more intensely over its evolutionary history than that of other great apes.

Summary statistics have been designed to extract specific information from different aspects of DNA polymorphism data. For example, the summary statistic approach can be used to estimate rates of recombination, gene flow, or even to understand potential changes in N_e. With regard to changes in N_e, Tajima's *D* and related statistics summarize the frequency spectrum of polymorphisms in a sample of DNA sequences[32–34]. Rapid population growth is expected to result in an excess of low-frequency polymorphisms ($D < 0$), whereas severe reductions in N_e might result in a deficiency ($D > 0$) over that expected under the standard neutral model (where $D \approx 0$). Recombination is considered to be a neutral process that occurs at a much higher rate than mutation, and so estimates of linkage disequilibrium (LD) for X-linked and autosomal sequences have become an invaluable tool for evolutionary inference.

One challenge for interpreting summary statistics is that different evolutionary processes might result in identical values of the chosen statistic. For example, locus-specific processes such as directional selection and balancing selection could also result in an excess or a deficiency, respectively, of rare polymorphisms. Therefore, multilocus analysis of the frequency spectrum is needed to distinguish between demographic and selective processes, because the former is expected to affect all loci in the genome, whereas the latter affects only a distinct region of the genome. Moreover, multiple neutral demographic processes could result in similar summary statistics. For example, changes in N_e can be easily confounded with population structure[35–37]. Few analytical expectations for summary statistics have been obtained for models that include both population structure and changes in N_e, even though these might be the most important models for understanding patterns of human DNA polymorphism.

Coalescent-based inference. The most rapidly evolving set of tools for evolutionary sequence analysis and inference rely heavily on the *n*-coalescent approximation[38,39]. Because the coalescent process limits itself to tracing the ancestry of only those chromosomes that appear in a sample, it is extremely efficient for simulating DNA sequence polymorphism data sets (see BOX 2B for an example). Furthermore, violations of the assumptions of the standard neutral model often yield predictable results for the shape and length of ancestral coalescent genealogies. Ease of coalescent simulation has facilitated the development of likelihood-based methods for model parameter estimation. Several recent review papers focus on the latest advances that involve these estimation methods, including the development of Bayesian, Markov chain Monte Carlo, and importance sampling techniques[40,41].

Approximate likelihood analysis is one extremely flexible and convenient implementation of coalescent-based inference[42]. In this approach, coalescent simulations are used to find the evolutionary model that is most likely to produce the observed summary (or summaries) of the data. The relative ease with which this inferential procedure can be carried out has made it a popular choice in the recent literature, and many of the examples we discuss make use of it. Summaries of the frequency spectrum and LD are commonly used for approximate likelihood analyses of historical changes in human N_e. Moreover, the precision of the likelihood estimate

© 2006 Nature Publishing Group

might be greatly improved if multiple summaries, each describing some different informational content of the data, are considered simultaneously[43,44]. Although the approximate likelihood approach represents a substantial improvement over both gene-tree and summary statistical methods, inference might still be ambiguous if different evolutionary processes result in identical summaries of the data.

What have we learned from the genome?

Many of the studies from the literature consider potential changes in N_e during both epochs of human evolutionary history (that is, both before/during and after the emergence of AMH). An important question surrounding the earlier epoch is whether the origin of the AMH phenotype at ~200 kya was accompanied by a radical change in N_e (in the form of a 'speciation' bottleneck). For the more recent epoch of human history, a similar question focuses on whether a bottleneck (or bottlenecks) occurred as AMH emigrated from Africa (an 'out-of-Africa' bottleneck). A third question asks when local populations (demes) began to grow. With the world's population at over 6.5 billion people, it is clear that humans have undergone rapid population growth from a much smaller ancestral N_e. What remains uncertain is whether this expansion began with hunter-gatherer groups in the Pleistocene, perhaps as a result of the invention of new technologies or behavioural innovations, or much more recently in the Neolithic[45]. However, many of the inferences we discuss regarding changes in N_e depend on what is assumed about the structure of human populations from the Pleistocene to the present.

In the following sections we concentrate on evidence from haploid and non-haploid loci for changes in N_e in both epochs of human evolution, and then review studies that suggest that the ancestral population might have been geographically structured. We then discuss the implications of these inferences for the single origin model.

Changes in effective population size. The shallow TMRCA estimated from both the mtDNA and NRY data was initially interpreted as evidence that the AMH population experienced a speciation bottleneck. Furthermore, the reduced diversity observed in non-African mtDNA and NRY suggested a later bottleneck as AMH migrated out of Africa. In support of these models, several studies in the early 1990s concluded that the excess of low-frequency polymorphisms in human mtDNA over those expected under the standard neutral model is the signal of a rapid population expansion[46,47]. The onset of the mtDNA expansion(s) was estimated to have occurred between 30 and 130 kya, indicating that there has been one or more episodes of marked population growth during the more recent epoch of human evolution[48]. However, once polymorphism data sets from X-linked and autosomal loci began to appear, it was immediately clear that they did not agree with the mtDNA pattern of rapid expansion from a small ancestral N_e (REFS 26,34,49). X-linked and autosomal sequences showed no excess of low-frequency polymorphism; instead, there was a tendency for non-African populations to have positive Tajima's *D* values (indicating a deficiency of rare polymorphisms), and African populations to have only slightly negative values[50,51] (FIG. 1).

Some investigators still favoured the rapid growth model gleaned from mtDNA data: widespread balancing selection on the X chromosome and autosomes was invoked as an explanation for the observed discrepancy between the frequency spectra of haploid and non-haploid polymorphisms[8,52]. On the other hand, Fay and Wu[34] pointed out that, after a population bottleneck, a period of time is expected in which there is an excess of low-frequency polymorphisms in the haploid compartments and an excess of intermediate-frequency polymorphisms for autosomal loci. This discrepancy occurs because a given demographic event happens at different relative depths of haploid and autosomal genealogies. Interestingly, Tajima's *D* values at X-linked loci in non-African populations are intermediate to those of haploid and autosomal loci, consistent with the predictions of Fay and Wu's model of a recent non-African bottleneck (FIG. 1).

When and how did such a bottleneck or bottlenecks occur? Specific parameters of a bottleneck model have been estimated recently for a handful of non-African populations for which large neutral polymorphism data sets exist. Most of these studies use the approximate likelihood approach to parameter estimation. Voight *et al.*[44] simultaneously examined both Tajima's *D* and measures of linkage disequilibrium from 50 non-coding, autosomal locus pairs from a European (Italian) and an East Asian (Han Chinese) population. These authors found that Italian polymorphism summaries are compatible with a bottleneck model that includes a short period of severely reduced N_e 40 kya, or an older, less severe bottleneck. The Han Chinese polymorphism data were consistent with a similar, but even more severe, bottleneck. Interestingly, analysis of a single African population (the Hausa) failed to reject the standard neutral model.

These inferences are concordant with the results of other approximate likelihood analyses of both the frequency spectrum[53,54] and linkage disequilibrium[55] in autosomal SNPs and of the frequency spectrum for resequencing data[53,56–58]. Among these studies (some of which were carried out on samples from the United States[53–55]), there is some discrepancy among the estimated time of the out-of-Africa bottleneck, with estimates ranging from 27–53 kya (REF. 55) to 58–112 kya (REF. 54).

So, although the occurrence of an out-of-Africa bottleneck enjoys support from both the haploid and non-haploid compartments of the genome, the existence of an earlier speciation bottleneck continues to be debated. We might expect that the signal of the putative speciation bottleneck would be most evident in extant African polymorphisms: more recent demographic upheaval in non-African populations could obscure the signal of earlier events. Most analyses of African polymorphism either fail to reject the null model of constant population size[44] or find support for mild growth from a constant-sized ancestral population[44,54]. One study of resequencing data

Linkage disequilibrium
The non-random association of polymorphisms at two linked loci. Linkage disequilibrium is created by mutation, but broken down over time primarily by crossing over between the two loci.

Directional selection
A form of positive selection in which a single mutation has a selective advantage over all other mutations, resulting in the selected mutation rapidly reaching fixation (that is, a frequency of 100%) in the population.

Balancing selection
A form of positive selection that maintains polymorphism in the population. One well-known form of balancing selection is heterozygote advantage, where an individual who is heterozygous at a selected locus has a higher fitness than either of the homozygous genotypes.

Population structure
Arises when the individual members of a population do not mate at random with respect to geography, age class, language, culture or some other defining characteristic.

Likelihood-based method
A class of statistical methods that calculate the probability of the observed data under varying hypotheses, in order to estimate model parameters that best explain the observed data and determine the relative strengths of alternative hypotheses.

Bayesian technique
An approach to inference in which probability distributions of model parameters represent both what we believe about the distributions before looking at data and the likelihood of the parameters given the observed data.

© 2006 Nature Publishing Group

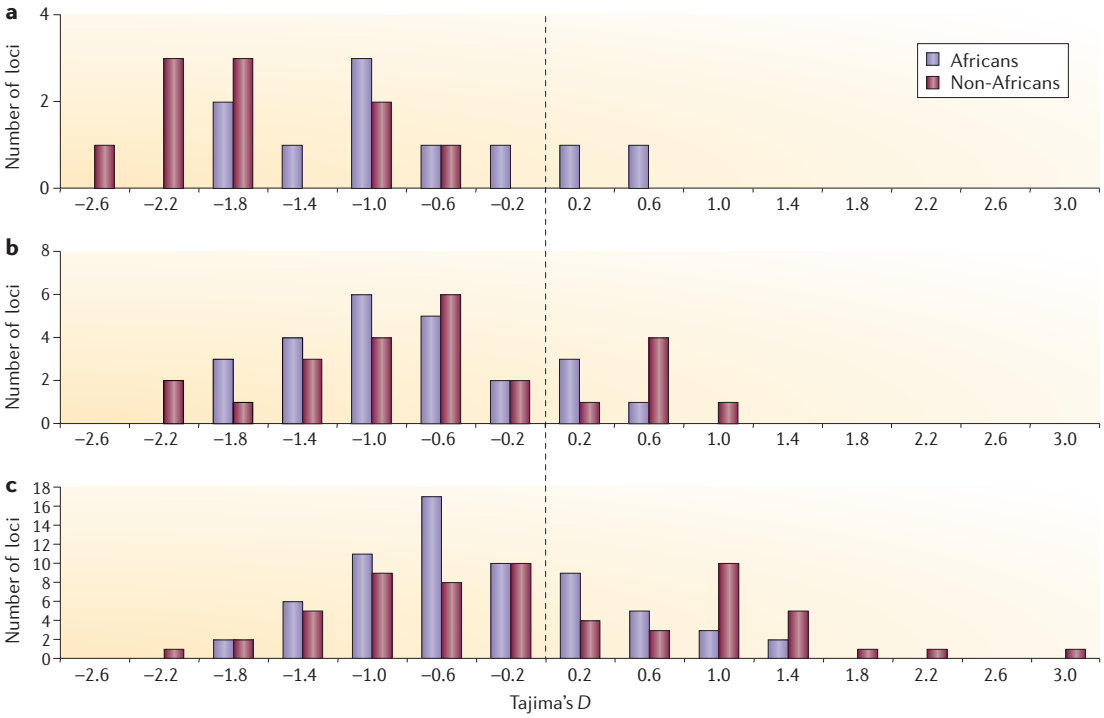


Figure 1 | **Frequency distribution of Tajima's *D*.** The frequency of Tajima's *D* statistic observed in DNA resequencing studies of human populations. Data from each locus were partitioned into African and non-African samples. **a** | mtDNA and Y-chromosome data from 5 African populations and 5 non-African populations[82]. **b** | Published data from 24 X-chromosomal loci[25,83]. **c** | Published data from 65 autosomal loci[26–28,44,69,73,84–92]. Mean Tajima *D* values for Africans are all slightly negative, ranging from −0.58 for both the haploids and the X chromosome to −0.20 for the autosomes. By contrast, mean Tajima *D* values for non-Africans show a trend from extremely negative (−1.50) to slightly negative (−0.46) to slightly positive values (0.13) for the haploid loci, X chromosome and autosomes, in corresponding order. The non-African pattern fits the expectations of a recent bottleneck, whereas the African pattern of slightly negative Tajima's *D* values for each genomic compartment is consistent with either slow growth from a small size, or a bottleneck that is much older than that of non-Africans.

Markov chain Monte Carlo technique
A simulation technique for producing samples from an unknown probability distribution. By evaluating the probability of the observed data at each step in the Markov chain, an estimate of the probability distribution of model parameters can be obtained by observing the behaviour of the chain as it proceeds through many steps.

Importance sampling
An efficient simulation method for integrating an unknown function, in which only those parameters that can actually produce the observed data are considered.

Approximate likelihood
A measure of the fit of some hypothetical model to a statistic calculated from observed data. For example, if 50% of polymorphisms occur in single individual chromosomes, a population growth model might have a higher likelihood of producing the observed number of singleton mutations than a model of population reduction.

Deme
A geographically localized population of a species that can be considered a distinct, interbreeding unit.

Neolithic
A human cultural period, beginning approximately 10,000 years ago, marked by the appearance in the archaeological record of industries such as polished stone and metal tools, pottery, animal domestication and agriculture.

Panmictic
Describes a diploid population in which each individual of a particular sex has an equal chance of producing offspring with any other member of the opposite sex in the population.

argues that an older bottleneck (~173 kya) is responsible for the summaries of LD in African populations[59]; although in this case, alternative models were not explicitly considered.

Ultimately, the genetic consequences of a speciation bottleneck might depend on the structure of the population at the time of the origin of AMH. If, for example, this ancestral African population were panmictic, we might expect a strong signal of an expansion from a small size in extant African populations. Conversely, if this ancestral population were geographically structured, the patterns of extant African polymorphism would depend greatly on the dynamics of gene flow during the transition to the AMH phenotype[60].

Ancestral population structure. One intriguing observation concerning the two hominin skulls recovered from the Omo valley (Omo I and Omo II) is that they are morphologically distinct, despite being found in the same stratigraphic layer. The Omo I skull seems to have the fully modern features associated with *H. sapiens*. The Omo II skull has many modern features; however, it also has an angular occipital and a flattened frontal bone, two morphological features that are characteristic of *Homo erectus*[61]. The spatial and temporal proximity of these two specimens offers a glimpse into the variation that existed at the time of the transition to the AMH phenotype. Did Omo I and Omo II belong to the same population? If so, was it a randomly mating population or was it genetically structured? This distinction could have enormous consequences for explaining the patterns of contemporary neutral genomic polymorphism.

The aforementioned studies of recent AMH demographic history analyse each population sample separately, thereby neglecting the effects of population structure. This approach can be misleading for many reasons; for example, gene flow and non-random mating can lead to skews in the frequency spectrum that is used to infer population history[60]. However, several studies explicitly account for the effects of population structure. For example, one study of genome-wide SNP-discovered loci made use of a coalescent-based method and inferred that ancestral population structure could have had the effect of increasing the ancestral N_e (REF. 62). Likewise, analyses of multilocus resequencing data have led some authors to argue that the ancestral population must have been genetically structured[63,64]. Because mtDNA or NRY polymorphism contain little information on this earlier epoch of human history, it is necessary to address the

© 2006 Nature Publishing Group

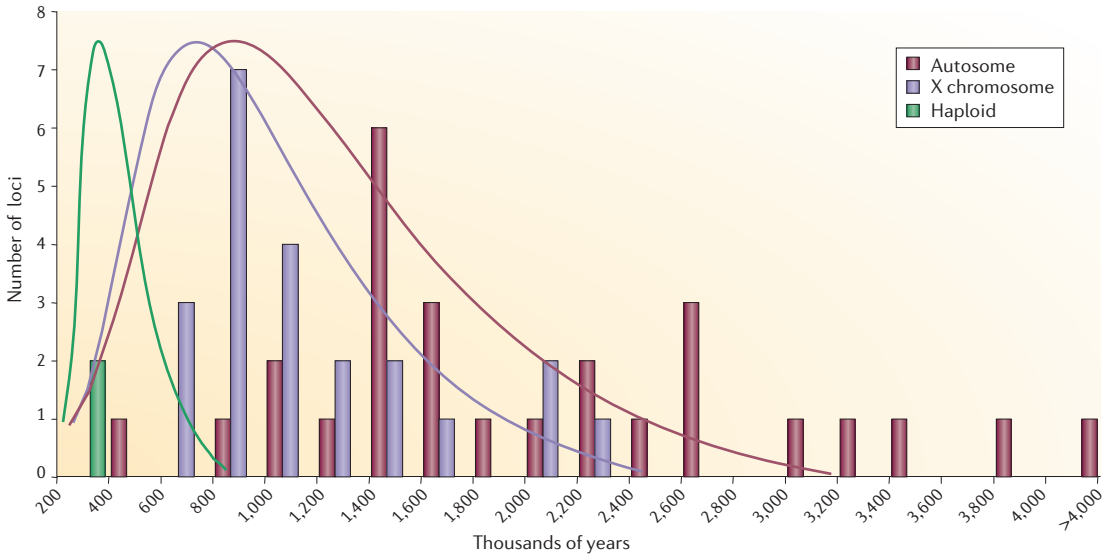


Figure 2 | **Frequency distribution of times to the most recent common ancestor (TMRCA).** Bars show the frequencies of TMRCA estimated from published resequencing studies of the two haploid loci, 22 X-linked loci [18,25,63,93–95], and 27 autosomal loci[18,63,69,71–73,85,88,93,95–97]. The three colour-coded curves are the corresponding expected probability distributions of TMRCA in a panmictic population that has an effective size of 10,000. The observed distribution of estimated TMRCA is in general agreement with the expected distribution, with the exception that there seems to be an excess of loci in the upper tail (that is, loci that have large estimated TMRCA values). One explanation for this apparent excess of ancient TMRCA values is that the anatomically modern human population is descended from a structured ancestral population (see main text).

question of ancestral population structure with data from the X chromosome and the autosomes.

One prediction of the single origin model is that all shared AMH polymorphisms trace their origins back to a single deme in Africa (that is, other demes of archaic *Homo* contributed no genetic material to the AMH genome). One way to assess the validity of this prediction is to examine the distribution of TMRCA throughout the genome. If the ancestral population was both small and isolated, we expect a unimodal distribution of TMRCA, with autosomal and X-linked loci having a (respectively) four and threefold deeper TMRCA than the haploid loci. FIGURE 2 juxtaposes the frequency distribution of estimated TMRCA from several resequencing studies with the theoretical distributions that are expected under the single origin model. Several regions of the nuclear genome have older TMRCA than expected, both on the X chromosome[19,25,65–67] and on the autosomes[68–72]. The observed genomic distribution of TMRCA suggests that there might be too many loci with deep genealogical histories to be compatible with a simple, single origin model.

There are also instances of individual X-linked or autosomal loci that reject some basic predictions of the single origin model. One such locus is Xp21.1, a non-coding region of the X chromosome[67]. The Xp21.1 locus has two lineages of haplotypes that are estimated to have diverged from one another nearly 2 million years ago (mya). However, all the nucleotide mutations that separate the two lineages are in perfect LD, whereas mutations within one lineage of haplotypes show extensive evidence for historical recombination. It has been shown that this high level of nucleotide divergence and LD at the Xp21.1 locus statistically rejects the predictions of the single origin model[67]. There are a number of other loci that show a similar pattern of polymorphism, some of which also reject the single origin model[69,73]. One illustrative case is the CMP-*N*-acetylneuraminic acid hydroxylase pseudogene genealogy[72], in which two highly divergent haplotypes coalesce ~2.9 mya (FIG. 3).

The increasing rate of discovery of loci that are incompatible with a single origin model, in conjunction with the consistent inference that African populations do not show evidence of radical changes in historical population size, suggest that it is time to consider new working models of human origins that can accommodate the observed variance in the genome.

Revised models of human evolution

One viable explanation for the unexpected antiquity of haplotypes at several loci is that the AMH population is descended from multiple archaic subpopulations. An expected consequence of ancient population subdivision is to increase the historical N_e and therefore push back the TMRCA. This increased N_e might not be visible at all loci, depending on the proportion of the genome descending from multiple subpopulations. In this section we discuss a set of hierarchically nested models that incorporate ancestral population structure. We further show that this simple class of models can account for the variance in TMRCA and frequency spectra observed among loci in the genome. Our models assume that there were d demes of archaic *Homo* before the emergence of AMH approximately 200 kya.

© 2006 Nature Publishing Group

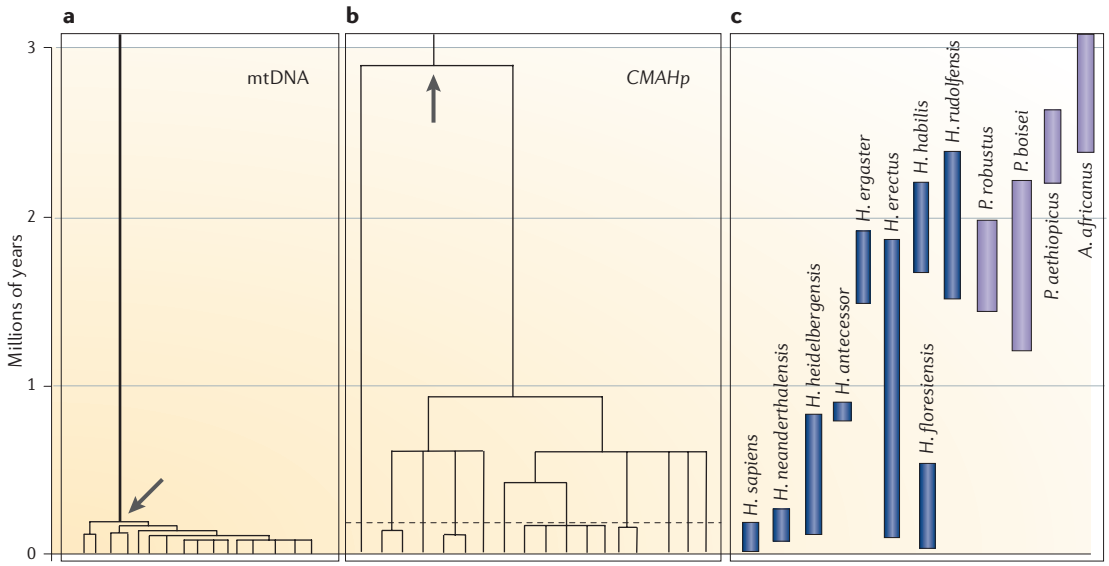


Figure 3 | **Time to most recent common ancestor and hominin fossil dates.** A comparison of the time depths of genealogies from different compartments of the genome. Panel **a** shows a timescaled mitochondrial DNA (mtDNA) gene tree[15]. Panel **b** shows a timescaled gene tree for the cytidine monophosphate-*N*-acetylneuraminic acid hydroxylase pseudogene (*CMAHp*) on human chromosome 6 (REF. 72). These trees can be compared with the reconstruction of hominin evolution as inferred from the fossil record (for example, see REF. 3), shown in panel **c**. Taxa in the genus *Homo* are shown in dark blue; *Paranthropus* and *Australopithecus* are shown in violet. Although deep lineages of mtDNA (**a**) are found in some African hunter-gatherer populations, such as the Khoisan or the Mbuti, these lineages trace back only to the time of the global most recent mtDNA ancestor (that is, <170,000 years ago; arrow in panel **a**). By contrast, polymorphisms on the X chromosome and autosomes (**b**) contain information about human population structure well before the emergence of anatomically modern humans (dotted line in panel **b**, corresponding to top of the *H. sapiens* bar in panel **c**), with a subset of loci characterized by the segregation of two highly divergent sequence haplotypes that coalesce before the Pliocene–Pleistocene boundary (~1.8 million years ago, when early forms of *Homo* and *Paranthropus* were extant; arrow in **b**). Data in panel **b** are from REF. 72.

After this time, one deme expanded its size and range and the other demes became extinct. Assuming that the d ancestral demes are all of the same size N_A, we would expect the total size of the ancestral population to be $N_e = N_A d(1 + 1/4N_e m)$, where m is the symmetrical rate of gene flow. FIGURE 4 shows a set of four models with an arbitrary number of demes ($d = 4$) that vary only with respect to the rate and timing of gene flow among demes. In the single origin model $m = 0$ (FIG. 4a), so only a single deme contributes genetically to the AMH population. FIGURE 4b,c depicts two models with different rates of gene flow among archaic demes. In the high-migration model (FIG. 4b), multiple demes contribute equally to the AMH population. In the low-migration model (FIG. 4c), a single archaic deme is a major contributor to AMH genetic diversity, whereas two others make only minor contributions. As the rate of ancestral gene flow increases among archaic demes, the high-migration model becomes equivalent to the hourglass model of Harpending *et al.*[9], which posits that the size of the archaic population that was ancestral to AMH had been large throughout much of its history. On the other hand, the low-migration model is novel in that it recognizes Africa as the probable source of AMH diversity but allows for low levels of gene flow among archaic populations[74,75]. Finally, the 'isolation-and-admixture' model (FIG. 4d) assumes that archaic demes that were isolated before the emergence of AMH exchanged genes with the AMH population as it expanded. This model is similar to a range of other models that allow for the hybridization between AMH and archaic populations in Eurasia[64,76–79].

The four human origins models depicted in FIG. 4 make different predictions about the expected distribution of TMRCA and frequency spectra from the genomes of extant African populations (BOX 3). The single origin model is consistent with observed African frequency spectra; however, as already discussed, loci with deep TMRCA seem to be too frequently sampled from the genome to be compatible with this model (FIG. 2). On the other hand, the high-migration model predicts a similar probability of observing loci with recent and ancient TMRCA, as well as many loci with a deficiency of rare polymorphisms. This is inconsistent with current observations[44,53,54,57], and so this model might not be appropriate for describing the early epoch of human history. The low-migration model predicts a slight excess of low-frequency polymorphism and a long upper tail on the genomic distribution of TMRCA. Additionally, under this model, many loci will have a high probability of single-deme ancestry, especially haploid loci that have a low N_e.

Therefore, the ancestral population structure model with low migration seems to be more strongly

© 2006 Nature Publishing Group

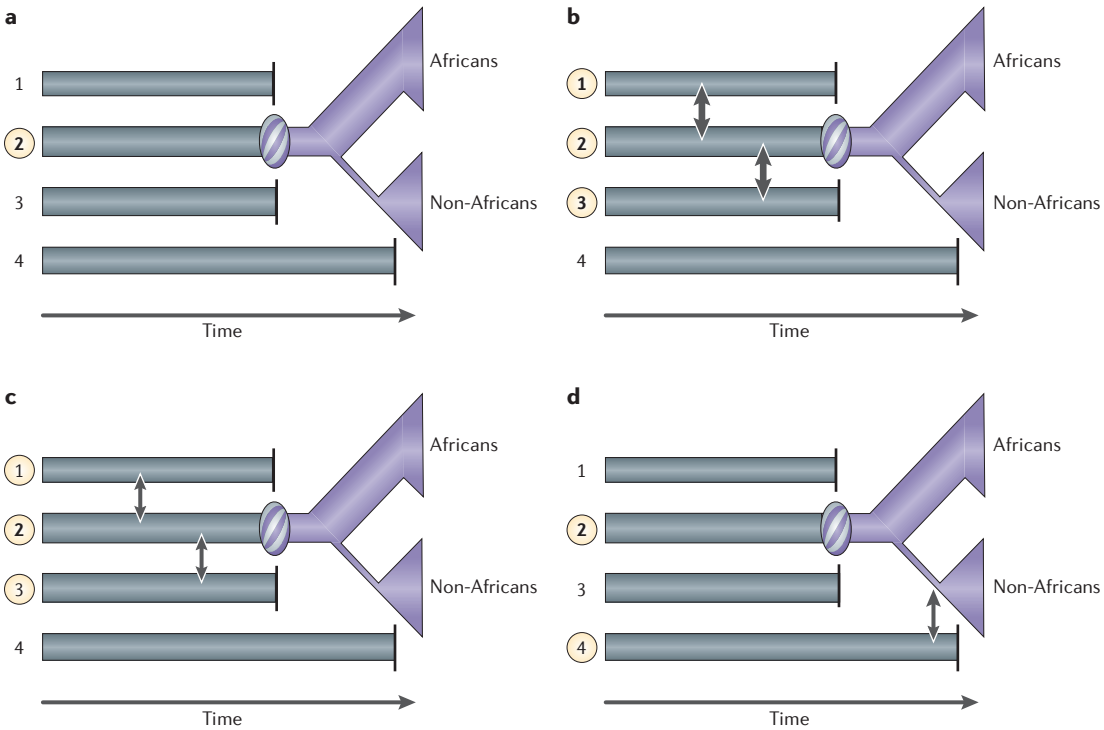


Figure 4 | **Models of human origins. a** | Single origin model. **b** | High-migration model. **c** | Low-migration model. **d** | Isolation and admixture model. Numbers to the left of the diagrams indicate demes (1–4): those that are circled contribute genetic material to the anatomically modern human (AMH) genome (numbers in bold type make a major contribution whereas numbers not in bold type indicate a minor contribution). Grey horizontal bars represent archaic demes; thick and thin vertical double arrows refer to high and low rates, respectively, of (symmetrical) gene flow among demes; violet cross-hatched circles represent transition to AMH; short vertical black bars depict extinction events; the violet horizontal bar represents the incipient AMH deme; increases in population size and divergence between African and non-African demes are indicated by angled violet lines and a bifurcation, respectively; bottlenecks in the non-African lineage are indicated by a thin violet line. Models that include ancestry from multiple archaic demes (**b**–**d**) might be favoured over the single origin model (**a**) because they predict a much higher variance in evolutionary patterns at different loci throughout the genome.

supported by the current data than either the single origin or high-migration models. To account for the discovery of a subset of loci with highly divergent haplotypes and long-range LD, the possibility of more recent admixture between AMH and previously isolated archaic demes also remains viable. Indeed, a recent analysis of multilocus sequence data found evidence for ancient admixture in both a European and West African population, with contributions to the modern gene pool of ~5% (REF. 80).

If the AMH genome contains any degree of dual ancestry (that is, archaic and modern) the single origin model must be rejected. Although most of the AMH genome might descend from a single African population, if further studies confirm a non-negligible contribution of archaic genetic material to the AMH genome, it would imply that the evolutionary lineage leading to AMH did not evolve reproductive isolation from other archaic hominin subpopulations and, therefore, cannot be considered a distinct biological species. Full resolution of this question awaits systematic resequencing surveys of many independent regions of the genome that are far from functional regions (that is, to avoid the possible confounding effects of natural selection). Future analyses of such data could help to resolve whether admixture, if it occurred, was before (low-migration model) or after (isolation and admixture model) the emergence of the AMH phenotype. For this purpose, additional summaries of the DNA resequencing data should be evaluated for their power to distinguish among the predictions of these models.

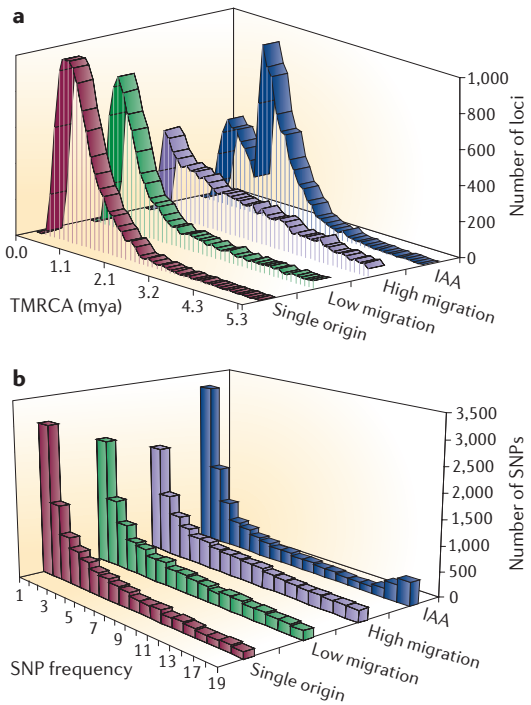
It should be noted that alternative models of sustained population subdivision (such as metapopulation models) also remain feasible[22]; however, further theoretical predictions that are based on these models must be obtained before the specific nature of the putative historical subdivision can be fully addressed.

Conclusions

The genomic era has ushered in vast amounts of new polymorphism data that are helping to unravel the demographic history of human populations. Although many questions remain unanswered, clear patterns are beginning to emerge for both early and recent epochs of human evolution. Although the genomic signatures of recent human population expansion are readily observable at the haploid loci for many populations, patterns of non-African autosomal polymorphism indicate a period of severe reductions in N_e before the

© 2006 Nature Publishing Group

Box 3 | **Genetic predictions of human origins models**



Do the four models presented in FIG. 4 make different predictions about patterns of polymorphism throughout the genome? Because there is strong support for a more recent non-African bottleneck, these predictions apply only to extant African populations. The expected genomic distribution of the time to the most recent common ancestor (TMRCA; panel **a**) and the frequency spectrum of SNPs from a sample of 20 chromosomes (panel **b**) was obtained by coalescent simulation under the four models. In these simulations, the present-day effective population size (N_e) of the hypothetical African anatomically modern human (AMH) population is 20,000 and the size each of the four ancestral demes is 2,000. Under the single origin model (see also FIG. 4a), there is a unimodal distribution of TMRCA, with a mean of approximately 800,000 years ago. The frequency spectrum expected in this model shows a slight excess of low-frequency polymorphisms because of the assumption that the AMH population has experienced recent growth. By contrast, the high-migration model (see also FIG. 4b) assumes that $4N_em = 10$ (m is the rate of migration per generation), so that there is a low probability that the sampled chromosomes will have a line of descent that is restricted to a single ancestral deme. The high rate of ancestral gene flow results in a relatively flat distribution of TMRCA, suggesting that loci randomly sampled from the genome will have a wide range of equally probable TMRCA. The high-migration model also predicts a deficit of low-frequency polymorphisms, compared with the standard neutral model. The low-migration model (see also FIG. 4c) assumes that $4N_em = 0.5$ and therefore some loci will have a line of descent that traces back to a single ancestral deme, whereas others will have experienced migration to another ancestral deme. This partitioning of loci in the genome results in a unimodal distribution of TMRCA, as in the single origin model, but also includes a long upper tail. This long upper tail means that loci with ancient TMRCA can be sampled from the genome, but with less frequency than under the high-migration model. The expected frequency spectrum under the low-migration model is also intermediate to the previous two models, showing only a slight excess of low-frequency polymorphisms. In the isolation-and-admixture (IAA) model (see also FIG. 4d), the admixture proportion is set to 5% and the two admixing populations are assumed to have split from one another 2 million years ago, roughly the time when *Homo erectus* emigrated from Africa. This model predicts a bimodal distribution of TMRCA, where loci with more ancient TMRCA are the byproducts of the admixture event. Adjusting the admixture proportion will adjust the relative heights of each mode. The IAA model predicts a frequency spectrum with an excess of both low-frequency and high-frequency polymorphisms, reflecting the fact that only a few sequences from the AMH sample trace the archaic ancestry and that these sequences will differ by many mutations from the remainder of the AMH sequences. In addition to TMRCA and the frequency spectrum, a more comprehensive consideration of the predictions of models of human population structure might also examine other summaries of data, such as levels of linkage disequilibrium.

recent period of growth. Whether the reduction in N_e outside Africa was a bottleneck resulting from a single migrant pool leaving Africa, a series of bottlenecks, or fluctuating N_e resulting from frequent extinction and recolonization of demes[35], remain outstanding questions. The single origin model of human evolution was originally formulated mainly on the basis of patterns of haploid polymorphism; however, mtDNA and the NRY are but two of potentially millions of independent realizations of evolutionary history that constitute our genome. Although many of the published X-linked and autosomal resequencing data are concordant with the single origin model, there are a growing number that are not. Indeed, the genome seems to be made up of loci with widely varied evolutionary histories[81].

This mosaic nature of our genome might be most easily explained by models that incorporate gene flow among a group of ancestral demes. The dynamics of gene flow in a structured ancestral population could have mitigated the effects of a 'speciation' bottleneck during the transition to the AMH phenotype. Finally, the persistence of highly divergent haplotypes with elevated LD, both inside and outside Africa, suggests that replacement of archaic *Homo* by the AMH population might have been accompanied by some degree of genetic assimilation. Future work directed towards obtaining reliable estimates of the genomic frequency spectrum for African populations, more resequencing studies that target non-coding regions of the genome to arrive at accurate TMRCA distributions, further theoretical work on the predictions of the various models, and improved methods of data analysis will undoubtedly bring us closer to understanding the neutral history of our genome.

© 2006 Nature Publishing Group

1. Day, M. H. Omo human skeletal remains. *Nature* **222**, 1135–1138 (1969).
2. McDougall, I., Brown, F. H. & Fleagle, J. G. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **433**, 733–736 (2005).
3. Wood, B. Hominid revelations from Chad. *Nature* **418**, 133–135 (2002).
4. Anton, S. C. & Swisher, C. C. Early dispersal of *Homo* from Africa. *Annu. Rev. Anthropol.* **33**, 271–296 (2004).
5. Teshima, K. M., Coop, G. & Przeworski, M. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* **16**, 702–712 (2006).
6. Stringer, C. B. & Andrews, P. Genetic and fossil evidence for the origin of modern humans. *Science* **239**, 1263–1268 (1988).
   An early synthesis of genetic and fossil evidence that supports the recent African origin of anatomically modern humans.
7. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
   An influential paper that discusses one of the first attempts to use mtDNA data to infer the origin of anatomically modern humans.
8. Excoffier, L. Human demographic history: refining the recent African origin model. *Curr. Opin. Genet. Dev.* **12**, 675–682 (2002).
9. Harpending, H. C. *et al.* Genetic traces of ancient demography. *Proc. Natl Acad. Sci. USA* **95**, 1961–1967 (1998).
   One of the first comparisons of haploid and non-haploid sequence data in the context of explicit models that incorporate a history of bottlenecks and different ancestral effective population sizes for modern humans.
10. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, New Jersey, 1994).
11. Cavalli-Sforza, L. L. & Feldman, M. W. The application of molecular genetic approaches to the study of human evolution. *Nature Genet.* **33**, S266–S275 (2003).
12. Hey, J. & Machado, C. A. The study of structured populations — new hope for a difficult and divided science. *Nature Rev. Genet.* **4**, 535–543 (2003).
13. Avise, J. C. *et al.* Intraspecific phylogeography: the mitochondrial bridge between population genetics and systematics. *Annu. Rev. Ecol. Syst.* **18**, 489–522 (1987).
14. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507 (1991).
15. Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000).
16. Hammer, M. F. *et al.* Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol. Biol. Evol.* **15**, 427–441 (1998).
17. Underhill, P. A. *et al.* Y chromosome sequence variation and the history of human populations. *Nature Genet.* **26**, 358–361 (2000).
18. Takahata, N., Lee, S. H. & Satta, Y. Testing multiregionality of modern human origins. *Mol. Biol. Evol.* **18**, 172–183 (2001).
19. Garrigan, D., Mobasher, Z., Severson, T., Wilder, J. A. & Hammer, M. F. Evidence for archaic Asian ancestry on the human X chromosome. *Mol. Biol. Evol.* **22**, 189–192 (2005).
20. Wilder, J. A., Mobasher, Z. & Hammer, M. F. Genetic evidence for unequal effective population sizes of human females and males. *Mol. Biol. Evol.* **21**, 2047–2057 (2004).
21. Jaruzelska, J., Zietkiewicz, E. & Labuda, D. Is selection responsible for the low level of variation in the last intron of the *ZFY* locus? *Mol. Biol. Evol.* **16**, 1633–1640 (1999).
22. Harding, R. M. & McVean, G. A structured ancestral population for the evolution of modern humans. *Curr. Opin. Genet. Dev.* **14**, 667–674 (2004).
23. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
24. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
25. Hammer, M. F. *et al.* Heterogeneous patterns of variation among multiple human X-linked loci: the possible role of diversity-reducing selection in non-Africans. *Genetics* **167**, 1841–1853 (2004).
26. Harding, R. M. *et al.* Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60**, 772–789 (1997).
    A seminal paper on the use of autosomal resequencing data to infer the history of modern human populations.
27. Zhao, Z. *et al.* Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proc. Natl Acad. Sci. USA* **97**, 11354–11358. (2000).
28. Yu, N. *et al.* Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. *Mol. Biol. Evol.* **18**, 214–222 (2001).
29. Fischer, A., Wiebe, V., Pääbo, S. & Przeworski, M. Evidence for a complex demographic history of chimpanzees. *Mol. Biol. Evol.* **21**, 799–808 (2004).
30. Yu, N. *et al.* Low nucleotide diversity in chimpanzees and bonobos. *Genetics* **164**, 1511–1518 (2003).
31. Kaessmann, H., Wiebe, V., Weiss, G. & Pääbo, S. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature Genet.* **27**, 155–156 (2001).
32. Tajima, F. The effect of change in population size on DNA polymorphism. *Genetics* **123**, 597–601 (1989).
33. Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709 (1993).
34. Fay, J. C. & Wu, C. I. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Mol. Biol. Evol.* **16**, 1003–1005 (1999).
    The first paper to recognize that differences in the effective population sizes of the haploid and autosomal compartments of the genome result in a different frequency spectrum after a population bottleneck.
35. Charlesworth, B., Charlesworth, D. & Barton, D. E. The effects of genetic and geographic structure on neutral variation. *Annu. Rev. Ecol. Syst.* **34**, 99–125 (2003).
36. Hammer, M., Blackmer, F., Garrigan, D., Nachman, M. & Wilder, J. Human population structure and its effects on sampling Y chromosome variation. *Genetics* **164**, 1495–1509 (2003).
37. Ptak, S. E. & Przeworski, M. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genet.* **18**, 559–563 (2002).
38. Kingman, J. F. C. On the genealogy of a large population. *J. Appl. Probab.* **19A**, 27–43 (1982).
39. Kingman, J. F. C. The coalescent. *Stochastic Process Appl.* **13**, 235–248 (1982).
40. Stephens, M. & Donnelly, P. Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B* **62**, 605–655 (2000).
41. Beaumont, M. A. Recent developments in genetic data analysis: what can they tell us about human demographic history? *Heredity* **92**, 365–379 (2004).
42. Tavare, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
    Presents for the first time the argument that the likelihood of an evolutionary model can be calculated from a statistical summary of genetic data, rather than directly from the data.
43. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16**, 1791–1798 (1999).
44. Voight, B. F. *et al.* Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl Acad. Sci. USA* **102**, 18508–18513 (2005).
    A recent study of 50 non-coding resequenced loci that supports a strong genetic bottleneck as humans migrated from Africa.
45. Wall, J. D. & Przeworski, M. When did the human population size start increasing? *Genetics* **155**, 1865–1874 (2000).
    Among the first analyses to acknowledge that there is too much variance in the genomic frequency spectrum to be compatible with any simple neutral model of population history.
46. Slatkin, M. & Hudson, R. R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562 (1991).
47. Di Rienzo, A. & Wilson, A. C. Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc. Natl Acad. Sci. USA* **88**, 1597–1601 (1991).
48. Rogers, A. R. Genetic evidence for a Pleistocene population explosion. *Evolution* **49**, 608–615 (1995).
49. Hey, J. Mitochondrial and nuclear genes present conflicting portraits of human origins. *Mol. Biol. Evol.* **14**, 166–172 (1997).
50. Przeworski, M., Hudson, R. R. & Di Rienzo, A. Adjusting the focus on human variation. *Trends Genet.* **16**, 296–302 (2000).
51. Frisse, L. *et al.* Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* **69**, 831–843 (2001).
52. Harpending, H. & Rogers, A. Genetic perspectives on human origins and differentiation. *Annu. Rev. Genomics Hum. Genet.* **1**, 361–385 (2000).
53. Adams, A. M. & Hudson, R. R. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* **168**, 1699–1712 (2004).
54. Marth, G. T., Czabarka, E., Murvai, J. & Sherry, S. T. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372 (2004).
55. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
56. Akey, J. M. *et al.* Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* **2**, e286 (2004).
57. Pluzhnikov, A., Di Rienzo, A. & Hudson, R. R. Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* **161**, 1209–1218 (2002).
58. InternationalHapMapConsortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
59. Lonjou, C. *et al.* Linkage disequilibrium in human populations. *Proc. Natl Acad. Sci. USA* **100**, 6069–6074 (2003).
60. Wakeley, J. & Aliacar, N. Gene genealogies in a metapopulation. *Genetics* **159**, 893–905 (2001).
61. Klein, R. G. *The Human Career: Human Biological and Cultural Origins* (Univ. Chicago Press, Chicago, 1999).
62. Wakeley, J., Nielsen, R., Liu-Cordero, S. N. & Ardlie, K. The discovery of single-nucleotide polymorphisms and inferences about human demographic history. *Am. J. Hum. Genet.* **69**, 1332–1347 (2001).
63. Satta, Y. & Takahata, N. The distribution of the ancestral haplotype in finite stepping-stone models with population expansion. *Mol. Ecol.* **13**, 877–886 (2004).
64. Templeton, A. R. Out of Africa again and again. *Nature* **416**, 45–51 (2002).
65. Zietkiewicz, E. *et al.* Haplotypes in the dystrophin DNA segment point to a mosaic origin of modern human diversity. *Am. J. Hum. Genet.* **73**, 994–1015 (2003).
66. Harris, E. E. & Hey, J. X chromosome evidence for ancient human histories. *Proc. Natl Acad. Sci. USA* **96**, 3320–3324 (1999).
    One of the first papers to show the effects of ancient population structure on patterns of polymorphism at a locus in humans.
67. Garrigan, D., Mobasher, Z., Kingan, S. B., Wilder, J. A. & Hammer, M. F. Deep haplotype divergence and long-range linkage disequilibrium at Xp21.1 provide evidence that humans descend from a structured ancestral population. *Genetics* **170**, 1849–1856 (2005).
    The first resequencing data set to reject the hypothesis that humans are descended from a single, randomly mating ancestral population.
68. Baird, D. M., Coleman, J., Rosser, Z. H. & Royle, N. J. High levels of sequence polymorphism and linkage disequilibrium at the telomere of 12q: implications for telomere biology and human evolution. *Am. J. Hum. Genet.* **66**, 235–250 (2000).
69. Barreiro, L. B. *et al.* The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region. *Am. J. Hum. Genet.* **77**, 869–886 (2005).
70. Hardy, J. *et al.* Evidence suggesting that *Homo neanderthalensis* contributed the H2 MAPT haplotype to *Homo sapiens*. *Biochem. Soc. Trans* **33**, 582–585 (2005).
71. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet.* **37**, 129–137 (2005).
72. Hayakawa, T., Aki, I., Varki, A., Satta, Y. & Takahata, N. Fixation of the human-specific CMP-*N*-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. *Genetics* **172**, 1139–1146 (2006).

© 2006 Nature Publishing Group

73. Koda, Y. *et al.* Contrasting patterns of polymorphisms at the ABO-secretor gene (*FUT2*) and plasma α(1,3)fucosyltransferase gene (*FUT6*) in human populations. *Genetics* **158**, 747–756 (2001).
74. Wolpoff, M. H. *Paleoanthropology* (McGraw-Hil, Boston, 1999).
75. Stringer, C. Human evolution: out of Ethiopia. *Nature* **423**, 692–693 (2003).
76. Brauer, G. in *The Human Revolution: Behavioural and Biological Perspectives on the Origins of Modern Humans* (eds Mellars, P. & Stringer, C.) 123–154 (Edinburgh Univ. Press, Edinburgh, 1989).
77. Smith, F. H., Jankovic, I. & Karavanic, I. The assimilation model, human origins in Europe, and the extinction of Neanderthals. *Quaternary Int.* **137**, 7–19 (2005).
78. Eswaran, V. A diffusion wave out of Africa: the mechanism of the modern human revolution? *Curr. Anthropol.* **43**, 749–774 (2002).
79. Relethford, J. *Genetics and the Search for Modern Human Origins* (Wiley-Liss, New York, 2001).
80. Plagnol, V. & Wall, J. D. Possible ancestral structure in human populations. *PLoS Genet.* **2**, e105 (2006).
81. Pääbo, S. The mosaic that is our genome. *Nature* **421**, 409–412 (2003).
82. Wilder, J. A., Kingan, S. B., Mobasher, Z., Pilkington, M. M. & Hammer, M. F. Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. *Nature Genet.* **36**, 1122–1125 (2004).
83. Kitano, T., Schwarz, C., Nickel, B. & Pääbo, S. Gene diversity patterns at 10 X-chromosomal loci in humans and chimpanzees. *Mol. Biol. Evol.* **20**, 1281–1289 (2003).
84. Alonso, S. & Armour, J. A. A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outstide Africa. *Proc. Natl Acad. Sci. USA* **98**, 864–869 (2001).
85. Martinez-Arias, R. *et al.* Sequence variability of a human pseudogene. *Genome Res.* **11**, 1071–1085 (2001).
86. Rieder, M. J., Taylor, S. L., Clark, A. G. & Nickerson, D. A. Sequence variation in the human angiotensin converting enzyme. *Nature Genet.* **22**, 59–62 (1999).
87. Clark, A. G. *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**, 595–612 (1998).
88. Fullerton, S. M. *et al.* Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am. J. Hum. Genet.* **67**, 881–900 (2000).
89. Bamshad, M. J. *et al.* A strong signature of balancing selection in the 5′ *cis*-regulatory region of *CCR5*. *Proc. Natl Acad. Sci. USA* **99**, 10539–10544 (2002).
90. Wooding, S. P. *et al.* DNA sequence variation in a 3.7-kb noncoding sequence 5′ of the *CYP1A2* gene: implications for human population history and natural selection. *Am. J. Hum. Genet.* **71**, 528–542 (2002).
91. Toomajian, C. & Kreitman, M. Sequence variation and haplotype structure at the human *HFE* locus. *Genetics* **161**, 1609–1623 (2002).
92. Harding, R. M. *et al.* Evidence for variable selective pressures at MC1R. *Am. J. Hum. Genet.* **66**, 1351–1361 (2000).
93. Harris, E. E. & Hey, J. Human populations show reduced DNA sequence variation at the factor IX locus. *Curr. Biol.* **11**, 774–778 (2001).
94. Alonso, S. & Armour, J. A. Compound haplotypes at Xp11.23 and human population growth in Eurasia. *Ann. Hum. Genet.* **68**, 428–437 (2004).
95. Templeton, A. R. Haplotype trees and modern human origins. *Yearb. Phys. Anthropol.* **48**, 33–59 (2005).
96. Patin, E. *et al.* Deciphering the ancient and complex evolutionary history of human arylamine *N*-acetyltransferase genes. *Am. J. Hum. Genet.* **78**, 423–436 (2006).
97. Vander Molen, J. *et al.* Population genetics of *CAPN10* and *GPR35*: implications for the evolution of type 2 diabetes variants. *Am. J. Hum. Genet.* **76**, 548–560 (2005).

Acknowledgements
We thank A. Di Rienzo for providing Tajima's *D* values and the following people for providing feedback on the manuscript: M. Cox, L. Excoffier, F. Mendez, C. Stringer, J. Wilder and E. Wood. Some of the work presented here was made possible by a US National Science Foundation HOMINID grant to M.F.H.

Competing interests statement
The authors declare no competing financial interests.

DATABASES
The following terms in this article are linked online to:
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
RRM2P4

FURTHER INFORMATION
Michael Hammer's laboratory homepage:
http://hammerlab.biosci.arizona.edu
Access to this links box is available online.

© 2006 **Nature Publishing Group**