Spread of an Inactive Form of Caspase-12 in Humans Is Due to Recent Positive Selection

Yali Xue,[1] Allan Daly,[1] Bryndis Yngvadottir,[1] Mengning Liu,[1] Graham Coop,[3] Yuseob Kim,[4] Pardis Sabeti,[5] Yuan Chen,[2] Jim Stalker,[1] Elizabeth Huckle,[1] John Burton,[1] Steven Leonard,[1] Jane Rogers,[1] and Chris Tyler-Smith[1]

[1]The Wellcome Trust Sanger Institute and [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom; [3]Human Genetics Department, University of Chicago, Chicago; [4]Department of Biology, University of Rochester, Rochester, NY; and [5]Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA

The human caspase-12 gene is polymorphic for the presence or absence of a stop codon, which results in the occurrence of both active (ancestral) and inactive (derived) forms of the gene in the population. It has been shown elsewhere that carriers of the inactive gene are more resistant to severe sepsis. We have now investigated whether the inactive form has spread because of neutral drift or positive selection. We determined its distribution in a worldwide sample of 52 populations and resequenced the gene in 77 individuals from the HapMap Yoruba, Han Chinese, and European populations. There is strong evidence of positive selection from low diversity, skewed allele-frequency spectra, and the predominance of a single haplotype. We suggest that the inactive form of the gene arose in Africa ~100–500 thousand years ago (KYA) and was initially neutral or almost neutral but that positive selection beginning ~60–100 KYA drove it to near fixation. We further propose that its selective advantage was sepsis resistance in populations that experienced more infectious diseases as population sizes and densities increased.

Our evolutionary past profoundly influences our current genetic makeup, including our predispositions to health and disease, and is also of great intrinsic interest. We know from fossil and archaeological records that both the human genus *Homo* and our species *sapiens* evolved in Africa. Anatomically modern *Homo sapiens* was present in Ethiopia ~195 thousand years ago (KYA) (McDougall et al. 2005), but modern human behavior developed much later (~50–100 KYA) (Henshilwood et al. 2002), and populations outside Africa derive their genes almost entirely from migrations of humans who were both anatomically and behaviorally modern, beginning ~50–60 KYA, followed by further local adaptations (Jobling et al. 2004). Many of the changes in phenotype must have had a genetic component, and we would like to understand these, but few of the relevant genes have been identified.

Two approaches have been used to search for these evolutionarily important genes. One is to start from a phenotype of interest, for which biological information or a Mendelian disorder may point toward a particular gene. The identification of genes involved in resistance to malaria (e.g., Saunders et al. 2002) or dietary adaptation (e.g., Bersaglieri et al. 2004) and the *ASPM* gene (in which mutation produces microcephaly) (Zhang 2003; Evans et al. 2004; Mekel-Bobrov et al. 2005) and the *FOXP2* gene (in which mutation produces speech/language disorder) (Enard et al. 2002; Zhang et al. 2002) testifies to the utility of this approach, but, in many cases, we do not know which genes influence the phenotype of interest. A complementary approach is, therefore, to identify changes in DNA sequence, gene copy number, or expression level without prior information about their phenotypic relevance. The chimpanzee genome sequence allows for genomewide comparisons (Chimpanzee Sequencing and Analysis Consortium 2005), whereas targeted studies have examined coding regions (Clark et al. 2003), gene-copy-number changes (Fortna et al. 2004), and expression-level changes, particularly in brain (reviewed by Preuss et al. [2004]). These analyses provide information mainly about fixed differences, whereas studies of variation within humans provide information about more-recent genetic changes that still differ between populations (e.g., Kayser et al. [2003]). These genomewide studies often identify large numbers of differences, most of which are likely to be neutral, but provide candidates for further investigation.

Once candidate genes have been identified, their relevance needs to be evaluated. Evolutionarily important genes will have undergone positive selection, and this leaves an imprint on the gene and its surrounding region. There is no single test for past positive selection, but patterns of amino acid change, nucleotide diversity, allele-frequency spectra, differentiation between populations, and haplotype structure can all provide some information (Ronald and Akey 2005). Genes or alleles that

Received November 17, 2005; accepted for publication February 1, 2006; electronically published February 21, 2006.

Address for correspondence and reprints: Dr. Chris Tyler-Smith, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom. E-mail: cts@sanger.ac.uk

Am. J. Hum. Genet. 2006;78:659–670. © 2006 by The American Society of Human Genetics. All rights reserved. 0002-9297/2006/7804-0012$15.00

have been positively selected can show such properties as rapid amino acid change, low diversity, high frequencies of rare or derived alleles, large differences between populations, and/or extended haplotypes, depending on the time when selection began and the frequency of the selected allele. Genes that do show evidence of positive selection fall into two categories: those that are selected in a wide variety of species, because they are involved in processes like host-pathogen interactions or reproduction, and those involved in more human-specific traits (Vallender and Lahn 2004).

Changes in protein sequence and expression pattern have been considered general molecular mechanisms underlying human evolution (King and Wilson 1975). One variant of these, gene loss, may be a common way in which populations and species adapt, because the multiplicity of ways in which a gene can be inactivated means that loss-of-function mutations are readily available for selection to act on (Olson 1999). Examples relevant to human evolution include (1) the inactivation, several million years ago, of a myosin heavy chain gene (*MYH16*), expressed predominantly in masticatory muscles, that may have influenced the anatomy of the head and removed a constraint to the development of the modern brain (Stedman et al. 2004), and (2) the *CCR5* Δ32 deletion that inactivates the CCR5 protein, with the result that Δ32/Δ32 homozygotes are strongly protected against HIV infection and AIDS and heterozygotes receive some protection (Dean et al. 1996). Although the Δ32 mutation must confer a selective advantage now and its current abundance has been suggested to result from selection by prehistoric infections, the variation of this gene is compatible with a neutral evolutionary past (Sabeti et al. 2005).

The caspase-12 gene (*CASP12*) provides another example of gene loss. It exists in two forms: full length (ancestral, active) or truncated in the middle by a stop codon at aa 125 (derived, inactive) (Saleh et al. 2004). This polymorphism has been shown to have significant phenotypic consequences: individuals with the full-length form produce lower levels of cytokines after stimulation by bacterial lipopolysaccharides, which leads to a lower initial immune response. If bacteria enter the bloodstream, however, they are at greater danger of immune overreaction and sepsis (Saleh et al. 2004). The active form was reported at a frequency of ~20% in Africa but was rare elsewhere. Because of the biological interest of this gene and the limited amount of information available about its evolutionary history, we investigated whether the predominant inactive form spread by neutral genetic drift or because of a selective advantage associated with gene loss. We conclude that it has spread through most of the human population within the past 100 KY because of positive selection.

Material and Methods

Population Samples

The examined samples consisted of 1,064 individuals from the CEPH Human Genome Diversity Panel (HGDP-CEPH) (Cann et al. 2002) and 77 individuals—26 Yoruba from Ibadan, Nigeria (YRI); 26 Han Chinese from Beijing (CHB); and 25 CEPH Utah residents with ancestry from northern and western Europe (CEU)—from the HapMap panel (International HapMap Consortium 2003).

Genotyping the Stop-Codon Polymorphism

The stop-codon polymorphism was genotyped in the HGDP-CEPH samples by SNaPshot primer extension as part of a tetraplex reaction. A fragment containing the stop-codon polymorphism was amplified using the forward primer 5′-CTCAACATCCGCAACAAAGA-3′ and the reverse primer 5′-TTGCTCTTTCAGCTGCCAAT-3′, followed by PCR extension with the primer 5′-GTATCCAAGGTTTTCAAGTAGATCTC-3′, through use of the ABI Prism SNaPshot Multiplex Kit (Applied Biosystems) according to the manufacturer's guidelines, with minor modifications.

Resequencing and Detection of Variants

PCR-amplified fragments of ~9–11 kb were generated from genomic DNA, then fragments of 500–700 bp overlapping by 200–400 bp were reamplified and sequenced. Primer and PCR details are given in table 1. For each individual, each nucleotide position was determined from both strands by at least two reads each. The *CASP12* genomic DNA sequence (GenBank accession number NC_000011) was used as the reference sequence, and the chimpanzee *Casp12* sequence (GenBank accession number NW_113990) was used in some analyses. The seven exons in the standard transcript (AceView) and an eighth exon (exon 3) present in some splice variants (Fischer et al. 2002) were considered.

Sequence traces were processed by the program ExoTrace, developed at The Wellcome Trust Sanger Institute (S. Leonard, unpublished material). Potentially polymorphic positions were flagged by the program and then were checked manually. Variants were accepted if they lay in a region with a Phred score ⩾30 and were detectable in other relevant high-quality reads. In a blind test, 1,328 (99.4%) of the 1,336 SNPs identified in this way corresponded to the genotype of the same sample in the HapMap database, equivalent to the accuracy of the HapMap data themselves (International HapMap Consortium 2005).

Data Analysis

For the stop-codon polymorphism, allele frequencies were determined by direct counting. A χ^2 test was performed for each

Table 1

Primers and PCR Conditions

The table is available in its entirety in the online edition of *The American Journal of Human Genetics*.

population as well as on the pooled world population, to evaluate Hardy-Weinberg equilibrium (HWE). F statistics were calculated according to the methods of Weir and Cockerham (1984) with the program FSTAT (Goudet 1995) and resulted in a value between 0 and 1, where 0 would imply no differentiation between populations and 1, complete differentiation.

Linkage disequilibrium (LD) blocks were inferred from genotype data through use of the Haploview program (Barrett et al. 2005). Haplotypes were reconstructed using PHASE 2.1 (Stephens et al. 2001; Stephens and Donnelly 2003; see the Stephens Web site). A median-joining network was constructed using NETW4.1.0.9 (Bandelt et al. 1999; Fluxus). The ratio of the number of nonsynonymous substitutions per nonsynonymous site (K_a) to the number of synonymous substitutions per synonymous site (K_s), K_a/K_s, and summary statistics (Tajima 1989; Fu and Li 1993; Fu 1997; Fay and Wu 2000) were calculated using DnaSP 4.00 (Rozas et al. 2003). Several of these tests compare different estimators of the population mutation parameter, θ. Tajima's D compares θ estimated from the number of polymorphic sites with θ estimated from the nucleotide diversity; negative values indicate an excess of rare variants, whereas positive values indicate an excess of intermediate-frequency variants. Fu and Li's tests compare θ estimated from the number of mutations in external branches of a gene tree rooted using the chimpanzee as an outgroup, with θ estimated from the number polymorphic sites (giving D) or the nucleotide diversity (giving F), and negative values indicate an excess of singleton mutations. Fay and Wu's H compares θ estimated from the nucleotide diversity with a θ estimator based on the frequency of derived variants, and negative values indicate an excess of high-frequency derived alleles. The other tests compare the observed haplotype distribution with that expected under a chosen population model. Fu's F_s is based on the probability of obtaining a sample with an equal or larger number of haplotypes than that observed, whereas the common-haplotype frequency test measures whether the most common single haplotype is expected to reach the frequency observed. Coalescent simulations were performed using the program ms (Hudson 2002) via a custom Perl script to process the output. Extended haplotype homozygosity (EHH) or relative EHH (REHH), which measures the decay of the ancestral extended haplotype with distance due to recombination (Sabeti et al. 2002), was analyzed using the program Sweep 1.0.

Frequency-based ages of the stop-codon mutation were estimated as described elsewhere (Griffiths 2003). Phylogeny-based time to the most recent common ancestor (TMRCA) estimates were obtained from NETW4.1.0.9 with use of a mutation rate based on 82 fixed differences between chimpanzee and human in the 8.6-kb LD block, under the assumption that 41 mutations occurred on each lineage and that the lineages split 7 million years ago. The hypotheses of a partial and complete selective sweep were compared using a composite-likelihood-ratio test (Meiklejohn et al. 2004). Likelihoods for both partial and complete sweeps were calculated from the entire data under the assumption that the selective target is located at the site of the stop-codon mutation. To estimate the strength of putative selection on the stop-codon mutation, we applied the composite likelihood analysis (Kim and Stephan 2002) to the subsample of haplotypes carrying the inactive gene, again assuming that the selective target is the stop-codon mutation. We assumed that $N_e = 10{,}000$ and $r = 10^{-8}$, where N_e is the effective population size and r is the recombination rate per base per generation across the ~13-kb region. In addition, we applied a full likelihood method (Coop and Griffiths 2004) to estimate the selection coefficient of the stop-codon polymorphism under a model of genic selection. The method assumes no recombination, so we restricted the analysis to an ~2-kb region, surrounding the polymorphism, that showed no evidence of recombination under the four-gamete test (Hudson and Kaplan 1985). The maximum-likelihood estimate of the selection coefficient was then used to estimate the age of the polymorphism by use of the same method (Coop and Griffiths 2004). In performing this analysis, we used the per base mutation rate obtained above, an N_e of 10,000 and a generation time of 25 years. Both methods used to estimate the strength of selection make the assumption of a single panmictic, constant-sized population.

Results

Worldwide Distribution of the Stop-Codon Polymorphism

We first investigated the distribution of the active and inactive forms of the caspase-12 gene in the HGDP-CEPH panel of 1,064 individuals from 52 worldwide populations. The results (fig. 1 and table 2) show that the active form of the gene predominates in some sub-Saharan African populations but is very rare outside Africa. Mbuti Pygmies and San have the highest frequencies of the active form—60% and 57%, respectively; the average for the sub-Saharan African populations is 28%. Outside Africa, the active allele was detected at low frequency in Israel, Pakistan, China, and Mexico (fig. 1), but the average was <1%, and 65% of the population samples were fixed for the inactive form. Although recent admixture may account for some of the active copies outside Africa (e.g., Mexico), other populations carrying active genes have no known history of recent African admixture.

No disagreement with HWE was observed in individual populations, but the pooled sample departed significantly from HWE ($P < .01$), reflecting subdivision. The large interpopulation variation, mainly caused by differences between the African and non-African populations, led to an F_{ST} value of 0.274, calculated using the frequency in each individual population. To assess whether this was unusually high, we compared it with empirically derived F_{ST} values. These are not available on a large scale for the HGDP-CEPH panel but are available for American populations of African, Han Chinese, and European origin (Hinds et al. 2005). We therefore recalculated the caspase-12 F_{ST} for sub-Saharan Africans versus Han Chinese or Europeans (table 3). Since F_{ST} is dependent on minor-allele frequency, we used SNPs matching the caspase-12 minor-allele frequency averaged across the pair of populations in each comparison and noted the 95% empirical ranges of these control F_{ST} values

(table 3). The African-Chinese caspase-12 F_{ST} value is not unusually high; the African-European value is the maximum possible for its allele frequency but, again, is not unusual and falls within the 95% CI of the control SNPs.

Is the observed predominance of the inactive form of caspase-12 due to positive selection, or does it result from factors such as a bottleneck associated with human migration out of Africa that acted on a neutral variant?

Sequence Variation of the Caspase-12 Gene

To address this question, we resequenced a 13.3-kb stretch of DNA that covers the whole caspase-12 gene and ~0.7 kb on each side in 77 individuals from the HapMap collection (26 YRI, 26 CHB, and 25 CEU), and we investigated the evolutionary history of the region. Of our sample of 155 chromosomes (including the reference sequence), 8 carried the active form of the gene: 6 YRI, 1 CHB, and the reference sequence of unknown origin, roughly reflecting the worldwide geographical distribution. All the rest of the chromosomes carried the inactive form. A total of 123 SNPs were detected (table 4 and online-only tab-delimited SNP table.txt, which can be downloaded and opened into a spreadsheet), but these were distributed very unevenly among the forms of the gene and populations. In the inferred haplotypes, the active genes were much more diverse: the eight chromosomes carried 61 SNPs and showed a nucleotide diversity of 19.7×10^{-4}, whereas the 147 inactive chromosomes carried 76 SNPs and had a nucleotide diversity almost 10 times lower, 2.0×10^{-4}. This led to higher diversity in the YRI (9.1×10^{-4}) than in the other populations (1.9×10^{-4} and 0.5×10^{-4} in the CHB and CEU, respectively—a ratio more extreme than any encountered in a study of 132 genes in African American and European American populations [Akey et al. 2004]), although it did not entirely account for the high YRI diversity. The inactive genes were also more diverse in Africa than outside ($\pi = 4.4 \times 10^{-4}$ and $\pi = 0.7 \times 10^{-4}$, respectively; table 4). The low diversity of the inactive genes, particularly outside Africa, provided the first indication that their spread might have been rapid and thus due to positive selection.

Many analyses can be performed most simply on regions that have experienced little or no recombination.

Table 2

Worldwide Distribution of the Active and Inactive Forms of the Caspase-12 Gene in the HGDP-CEPH Diversity Panel

The table is available in its entirety in the online edition of *The American Journal of Human Genetics*.

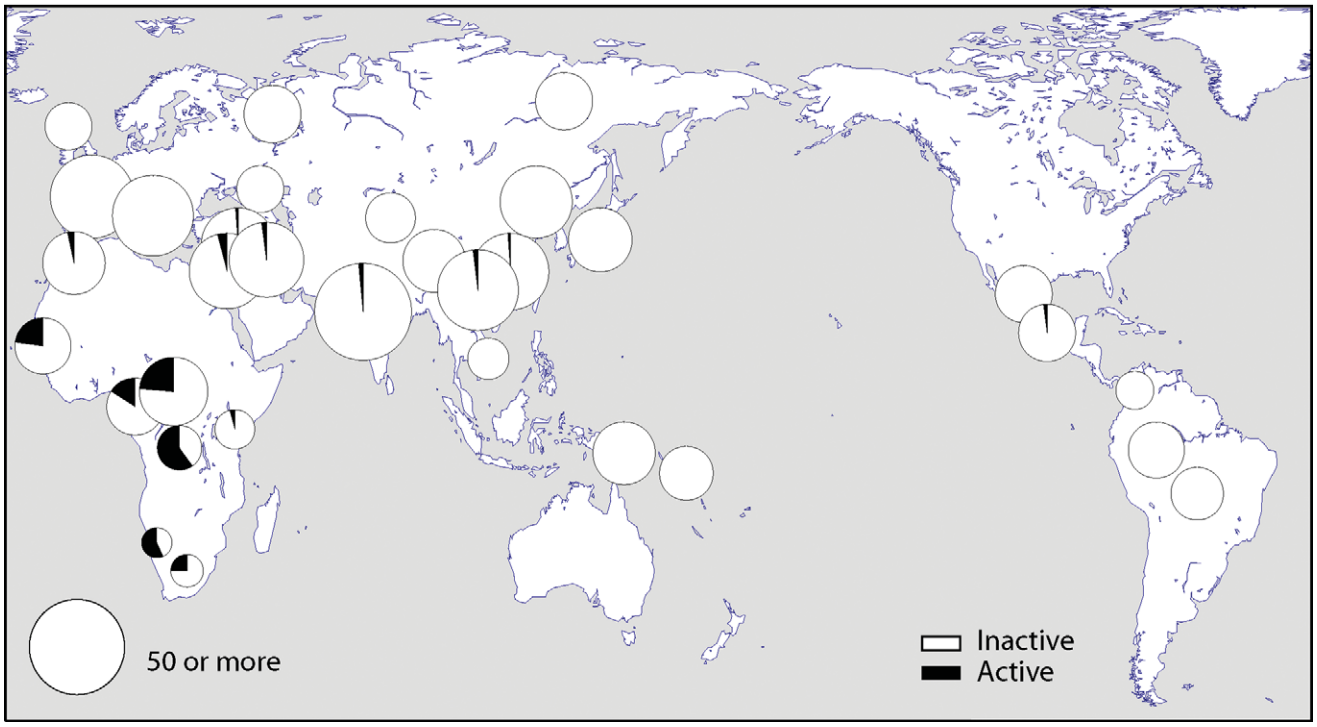

Figure 1 Worldwide distribution of the active and inactive forms of the caspase-12 gene in the HGDP-CEPH diversity panel. Circle area is proportional to sample size, up to a maximum of 50 individuals.

Table 3

Caspase-12 and Control SNP F_{ST} Values

Comparison	Caspase-12		Control SNPs		
	Frequency[a]	F_{ST}	Frequency Range	No.[b]	95% F_{ST} Range
Sub-Saharan Africa and Chinese Han	.138	.172	.132–.143	22,943	.016–.271
Sub-Saharan Africa and Europe	.132	.253	.127–.138	25,552	.011–.266

[a] Frequency of the minor allele.
[b] Number of control SNPs lying in this frequency range.

We therefore investigated the LD structure of the region and identified an ~8.6-kb LD block containing SNPs 10–99, with the stop codon in its center, and used it, together with the complete region, in further analyses (fig. 2). Haplotypes were again inferred for the LD block, a task facilitated by the observation that 57 (74%) of the 77 individuals carried zero or one SNP in this section. We then investigated whether the inferred pattern of variation was compatible with neutral evolution.

Neutrality Tests

We first examined the evolution of the coding region, expressed as the K_a/K_s ratio based on the human and chimpanzee sequences. This was 0.55, indicative of purifying selection over most of the evolutionary period but providing little insight into the most recent phase. Tests based on the variation within humans are better able to do this.

Neutral models of evolution provide predictions of expected allele-frequency characteristics, and observed patterns can be compared with these. We have calculated Tajima's *D* (Tajima 1989), Fu and Li's *D* and *F* (Fu and Li 1993), and Fay and Wu's *H* (Fay and Wu 2000); results are summarized in table 4. Neutrality is rejected for both the entire region and the 8.6-kb LD block by all tests with use of the whole data set. In individual populations, neutrality is similarly rejected by all tests for the CHB, but only by Tajima's *D* and Fay and Wu's *H* for the YRI and by Tajima's *D* for the CEU. These results can readily be understood in terms of a selective sweep that has proceeded to different stages in the different populations (see the "Discussion" section).

A second class of neutrality test examines haplotypes rather than single variable positions. A total of 36 haplotypes were identified (fig. 3), but one haplotype carrying the stop codon occurred 99 times and accounted for 64% of the sample (and 76% of non-African chromosomes). Thirty-six individuals (47%) were homozygous for this haplotype, so its high frequency cannot be an artifact of the haplotype-inference procedure. Fu's F_s test (Fu 1997), performed on the entire region, shows that significantly fewer haplotypes are found in the whole sample and in CEU than expected under neutrality (table 4). In the 8.6-kb block, fewer haplotypes than expected are found in these populations also. We also used coalescent simulations (Hudson 2002) to evaluate how often a single haplotype would be expected to occur in ⩾99 of 154 chromosomes under neutrality and how often a single haplotype would be expected at the observed frequencies in the individual populations. Except among the CEU, the observed frequencies were highly significant (table 4; last column, headed "Common Haplotype Frequency").

Therefore, according to all the tests used, sequence variation in the caspase-12 gene is significantly different from that expected under neutrality, and the properties of the LD block resemble those of the complete region. Departures from neutral expectation at a single locus can arise in many ways, including stochastic variation and demographic change, but, as discussed below, the simplest explanation for all these deviations is positive selection.

Haplotype Structure and Phylogeny

A median-joining network was constructed to show the relationships between the inferred haplotypes of the 8.6-kb LD block (fig. 4). The network had a simple structure, with little evidence of recombination or recurrent mutation, as expected from the way the region had been chosen. The eight haplotypes carrying active genes are all different from one another and from the inactive genes. All of the inactive haplotypes clustered together, with 99 chromosomes at the center of the cluster, 29 one step away, 6 two steps away, and a few more distant. Outside Africa, the most distant inactive haplotype lay only three steps from the center, whereas there was more diversity among the inactive haplotypes in Africa, and not all radiated directly from the central haplotype.

EHH (Sabeti et al. 2002) is a feature of regions that have recently experienced positive selection. We have therefore explored the extended haplotype structure surrounding the caspase-12 gene. Fortunately, the stop SNP was included in the HapMap set (International HapMap

Table 4

Caspase-12 Summary Statistics

The table is available in its entirety in the online edition of *The American Journal of Human Genetics*.

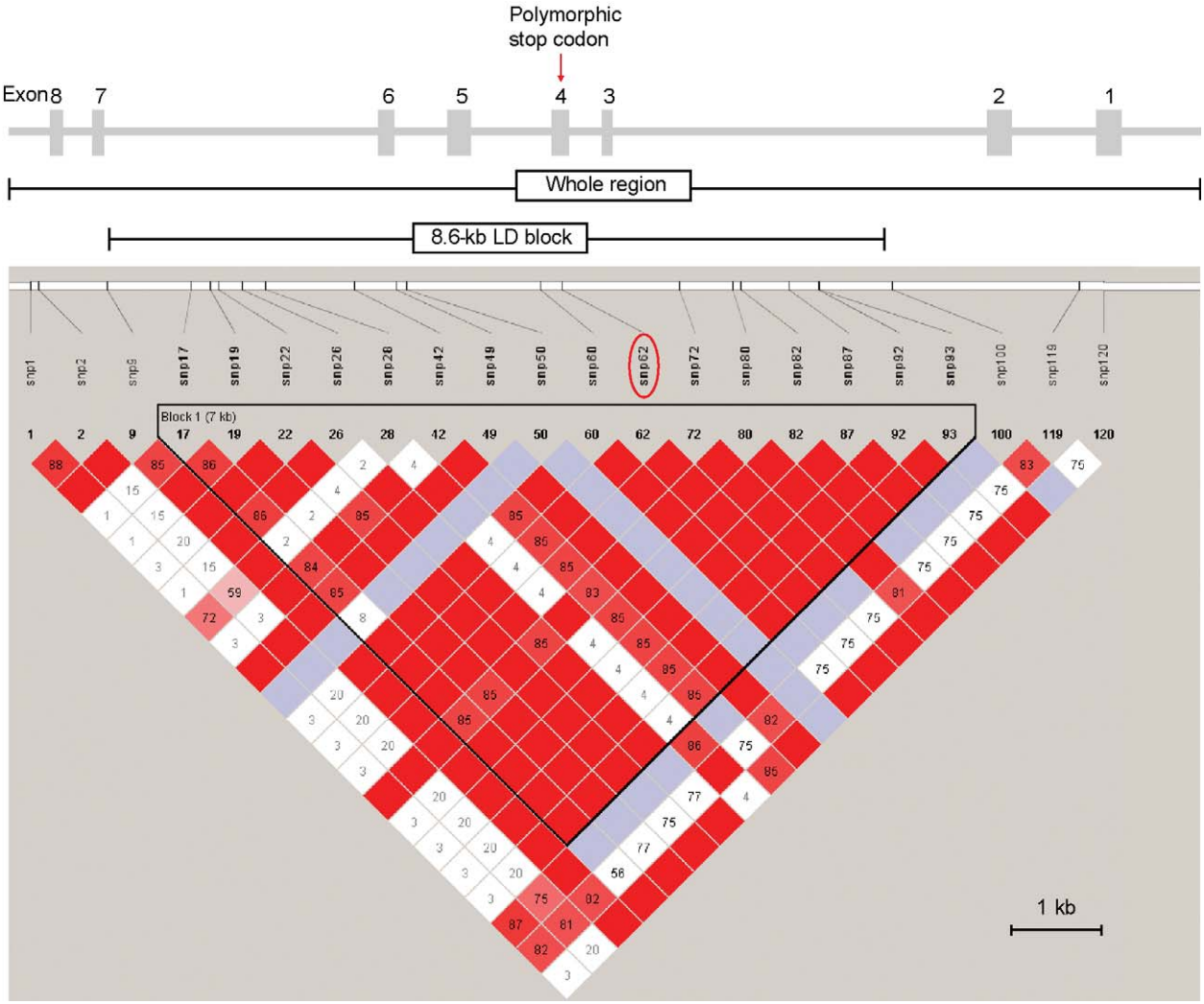

Figure 2 Structure of the caspase-12 gene. The exon-intron structure—including exon 3, which is present only in some transcripts—is shown at the top, as is the whole sequenced region and the location of the 8.6-kb LD block. The lower part of the figure shows the LD block identified using Haploview. Each square represents a pairwise value of D', with the standard color coding (red indicates LOD $\geqslant 2$ and $D' = 1$; pink indicates LOD $\geqslant 2$ and $D' < 1$; blue indicates LOD < 2 and $D' = 1$; white indicates LOD < 2 and $D' < 1$).

Consortium 2003); therefore, we could perform this analysis entirely *in silico*. We first selected cores containing this SNP and tested regions of 10–100 kb on either side, but we found that neither EHH nor REHH was significantly different from the genome average. We then measured the genetic distance over which EHH remains above a threshold value (0.5 or 0.2) and compared this with the corresponding distances for all alleles on chromosome 11. These distances were 0.013 cM and 0.079 cM and fell in the 58th and 41st percentile, respectively, so, again, were not unusual (see fig. 5). A related analysis by Pritchard and coworkers revealed a similarly nonsignificant value of the measure iHH (integrated EHH) (B. Voight, S. Kudaravalli, and J. Pritchard; personal communication). One explanation could be that sufficient time has elapsed for the long-range structure of the selected haplotype to decay; therefore, we wished to understand the timing of selection more fully.

Age of the Mutation: Timing and Strength of Selection

The frequency of an allele provides one guide to its age: it begins as a single copy, and the time required to rise to an observed frequency under neutrality or different selective regimes can be estimated (Griffiths 2003). According to this model, the stop codon would require almost 1 million years to reach 96% under neutrality, but this time would be greatly reduced by positive selec-

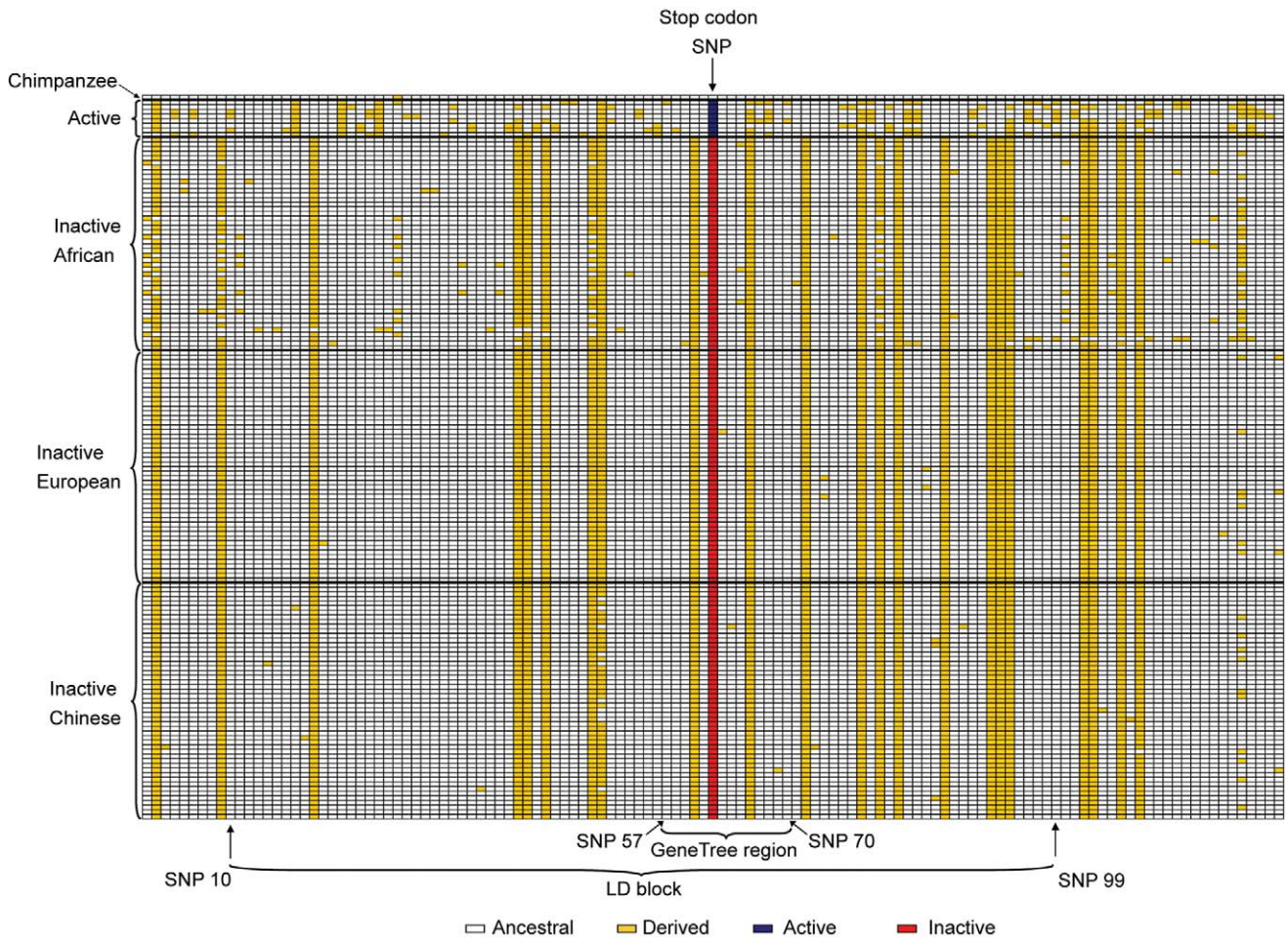

Figure 3 Inferred caspase-12 haplotypes. Only positions that are variable in humans are shown, coded according to whether they carry the same allele as chimpanzee (white) or not (yellow), except that the stop-codon polymorphism is shown in blue or red, respectively. The low diversity and high frequency of derived alleles can be seen in the inactive genes.

tion—for example, to 27 KY if it conferred a selective advantage of 1% (table 5). However, unless the selection coefficient can be estimated from other sources, this method does not provide an absolute age. We next estimated the TMRCA of the inactive alleles, using a phylogeny-based method (Bandelt et al. 1999), by means of the measure ρ, the average number of mutations from the root. This requires that a root be specified, and three different roots were investigated. Through use of root 1 (fig. 4), a mean ($\pm$SD) of 552 $\pm$ 276 KYA was obtained for the entire set of inactive haplotypes. The sensitivity of this estimate to the specification of the root is illustrated by the use of root 2, one step away, which led to a time of 397 $\pm$ 223 KYA. A TMRCA for the star-shaped cluster, through use of root 3 and without the haplotypes that lie between this root and the active genes, gave 61 $\pm$ 16 KYA. The first two times provide information about when the inactivation mutation occurred, whereas the third provides information about when a subset of inactive chromosomes started to expand, so they are expected to differ.

We then applied methods aimed at inferring the time of selection from the estimated selective advantage conferred by the stop-codon mutation. First, we attempted to estimate the strength of selection ($4N_e s$) by using parametric models that predict the spatial pattern of nucleotide diversity and allele-frequency spectrum around the putative target of selection (composite likelihood analyses) (Kim and Stephan 2002; Meiklejohn et al. 2004). We found that the data did not provide significant support for an incomplete sweep compared with a complete one: $\log(L[\text{incomplete sweep}]) - \log(L[\text{complete sweep}]) = 2.38$ (under the assumption that θ [scaled mutation rate per site] = 0.002 = observed level of mean diversity in active genes; this is likely to be the before-sweep level of variation in the entire region). This like-

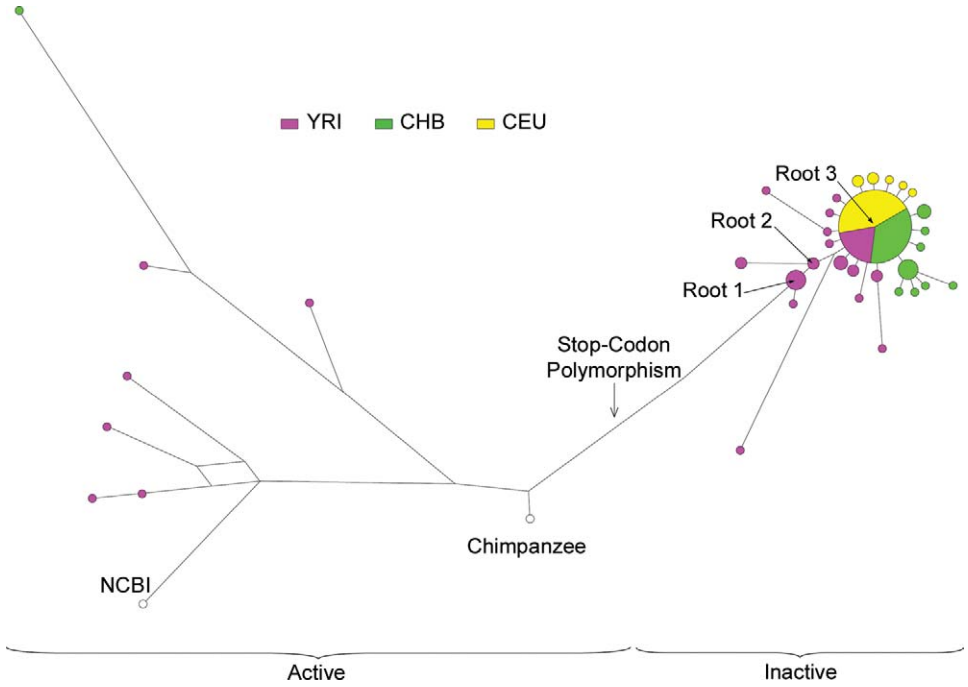

Figure 4 Median-joining network of inferred caspase-12 haplotypes from the 8.6-kb LD block. Roots 1, 2, and 3 are discussed in the text. Circle area is proportional to haplotype frequency, and circles are coded according to population.

lihood ratio is not large enough to reject a complete sweep, when assessed using data sets simulated under a complete-sweep model (Meiklejohn et al. 2004). However, this test of incomplete versus complete sweep has rather low power and assumes sampling from a single, randomly mating population, which is clearly violated by our data. Under the assumption that an incomplete sweep of the stop-codon mutation indeed shapes the haplotype structure of the data, the strength of selection ($4N_e s$) acting on the stop-codon mutation might be obtained by treating the 147 inactive haplotypes as if they represent a sample from a population in which a complete sweep had occurred (Meiklejohn et al. 2004). The model of a complete sweep (Kim and Stephan 2002) applied to the 147 sequences yields the estimate of $4N_e s = 677$. If $N_e = 10{,}000$, this corresponds to a selective advantage of ~1.7%. This suggests a time for the mutation of ~19 KYA. We also performed a full likelihood analysis of the data (Coop and Griffiths 2004), which required a data set free of recombination; we therefore restricted this analysis to a region of ~2 kb around the stop-codon polymorphism (fig. 3). The likelihood surface for the selection parameter $4N_e s$ peaked at ~315 (fig. 6), a selective advantage of ~0.8%. With use of this estimate of $4N_e s$, the time of the mutation was estimated from the ~2-kb region, by use of the method of Coop and Griffiths (2004), to be 0.058 in units of $2N_e$ generations, or ~29 KYA.

Finally, the geographical distribution of the alleles, combined with our understanding of modern human spread, provides indirect information about their age. Only one inactive haplotype appears to have left Africa, so this approach suggests that selection is likely to predate the exodus ~50–60 KYA.

Discussion

The inactive form of the caspase-12 gene has spread recently throughout most of the human population. We discuss here the evidence that this occurred as a consequence of positive selection rather than of drift, the likely time scale of events, and the significance of the inactivation of this gene for human evolution.

Positive Selection for Loss of Caspase-12

Positive selection leads to the rapid increase of a particular allele and its surrounding sequences. The available tests for neutrality/selection capture different con-

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Figure 5 EHH as a function of genetic distance in the CEU; corresponding haplotype bifurcation diagram.

Table 5

Estimates of the Age of the Mutation That Inactivated Caspase-12 or the TMRCA of Subsets of Inactive Alleles

Basis, Reference, and Conditions or Comments	KYA
Frequency (Griffiths 2003):	
Neutrality assumed	980
1% selective advantage assumed	27
5% selective advantage assumed	4.8
Phylogeny (Bandelt et al. 1999):	
Root 1 used (see fig. 3)	552 ± 276
Root 2 used (see fig. 3)	397 ± 223
Root 3 used (see fig. 3)	61 ± 16
Composite likelihood (Kim and Stephan 2002):	
1.7% selective advantage estimated	19
Full likelihood (Coop and Griffiths 2004):	
.8% selective advantage estimated	29

sequences of the process, and the population samples from Africa, China, and Europe illustrate different stages of the selective sweep, including complete fixation in the CEU sample. We would therefore expect the results of the tests to differ between the populations. Diversity becomes substantially reduced only as a sweep nears completion. The value for the caspase-12 gene worldwide was 4.5×10^{-4} (table 4), not very different from the genomewide and chromosome 11 averages of 7.5×10^{-4} and 8.4×10^{-4}, respectively (Sachidanandam et al. 2001), but the value in the CHB sample was reduced to 1.9×10^{-4} (SD = 0.9×10^{-4}), and that in the CEU was even lower, 0.5×10^{-4} (SD = 0.1×10^{-4}), both significantly lower than the YRI value (9.1×10^{-4}; SD = 1.7×10^{-4}).

Similarly, allele-frequency spectra become greatly skewed only as a sweep nears completion. They therefore show highly significant departures from neutral expectation in the worldwide data set and in the CHB sample, where there was 1 active gene and 51 inactive genes, but not in the YRI, where there were more active genes and greater diversity among the inactive ones. However, on fixation, the variants that previously contributed the low-frequency SNPs, singletons, and high-frequency–derived SNPs to the tests are no longer variable in the population; thus, these tests show slightly significant or nonsignificant results, as for the CEU. The significance of the values obtained was tested against a model of neutral evolution, but departures from neutrality can arise from causes other than selection, such as changes in population size. We have used simulations of a population bottleneck to explore the effect of one set of nonneutral demographies on some of these statistics, assuming a population size of 10,000 before 2,000 generations ago (~50 KYA at 25 years per generation, approximating the out-of-Africa migration), an instantaneous drop to a reduced size that remained until 1,000 generations ago (~25 KYA, corresponding to a commonly estimated start of growth outside Africa) and then expanded exponentially back to 10,000. The reduced size ranged from 100 to 1,000 in different runs. We found (table 6) that the value of Tajima's *D* in the CEU was not unusual under an extreme bottleneck but that the values observed in the CHB were never reached, even under the most extreme reduction in population size. It therefore seems that the summary statistic test results are not readily explained by a population bottleneck. Another way to assess the significance of the caspase-12 statistics is to compare them with empirical data, although even these comparisons must be interpreted with caution because they are based on different sample sets, which may have experienced different demographic histories. The caspase-12 values from the worldwide or CHB samples lie outside the 95% empirical range of 132 genes examined in two populations (Akey et al. 2004) and are more negative than almost all those published as examples of positive selection (table 7). Only *TRPV6* in Europeans, which was detected as the most extreme outlier from 264 analyses but for which the selective agent remains unknown (Akey et al. 2004), shows lower values.

The unusually high frequency of a single haplotype—21 (40%) of 52 chromosomes, even in the YRI sample, and higher in the other populations—provided a robust signal of departure from neutrality. It has been shown elsewhere that a single haplotype from a 62-kb region carrying 166 SNPs is unlikely to reach a frequency of even 21% under a wide range of demographic models (Mekel-Bobrov et al. 2005), so such a signal is also robust to the demographic specification. We found no evidence of an unusually extended haplotype associated with

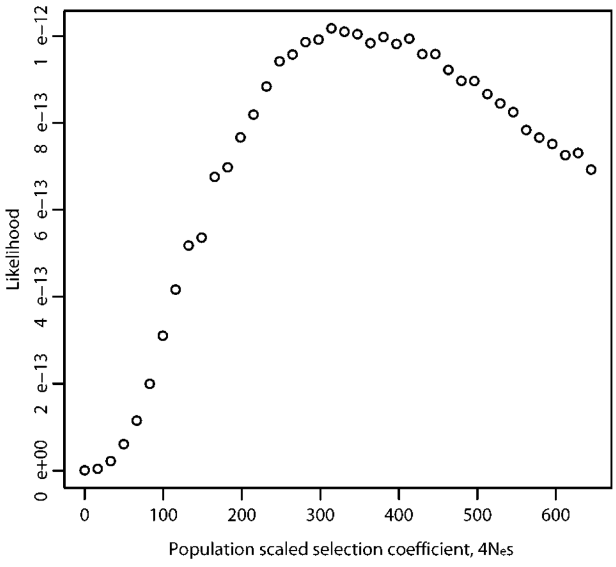

Figure 6 Likelihood surface for the selection parameter $4N_e s$

Table 6

Summary Statistic Tests with Use of a Bottleneck Model in the CEU and CHB Populations

The table is available in its entirety in the online edition of *The American Journal of Human Genetics*.

the caspase-12 gene. This can be explained by two factors. The first is the near fixation of the inactive gene, which drives other haplotypes down to low frequencies and consequently leads to low power to detect differences between haplotypes. The second is the time since the sweep began: the most significant EHH/REHH values have been reported for sweeps beginning <10 KYA (Sabeti et al. 2002; Bersaglieri et al. 2004). In conclusion, no plausible combination of demographic and stochastic factors can account for sequence variation surrounding the caspase-12 gene, but it shows exactly the signatures expected for a selective sweep that began early enough to have reached fixation in some populations but not in others. Indeed, it shows the clearest evidence of any locus documented thus far for a worldwide selective sweep in humans.

Target, Timing, and Strength of Selection

The rapid decay of LD within and surrounding the caspase-12 gene (fig. 2 and results not shown) indicates that selection is likely to be acting on the central region of the gene itself rather than on another gene in LD. Since the stop-codon polymorphism affects the phenotype and is the only variant in this region that is known to do so, we conclude that it is very likely to be the target of the selection.

Estimates of the age of the mutation or timing of selection depend on the method used, and all have wide CIs; nevertheless, all suggest that selection began in the Paleolithic period, a conclusion that is also consistent with the lack of EHH/REHH signal. The most recent—~19 KYA—is likely to be an underestimate, since it assumed that the inactive genes represented a complete sweep, whereas the sweep is evidently incomplete, and additional time is required for fixation. Furthermore, several of the methods required assumptions about demography (a panmictic population of a constant size of 10,000) that are commonly made but are obviously oversimplifications. Interactions with other advantageous genes—a kind of assortative mating for "survivorship"—could lead to additional departures from these simple models. The date based on geography—before 50–60 KYA—thus seems to provide the firmest lower date for the time of origin of the mutation, but the upper limit remains poorly defined. Despite the considerable uncertainty about the strength and timing of selection, a selective advantage of ~0.5%–1% beginning 60–100 KYA would explain most of our observations.

Selective Pressure

"Sepsis is the most common cause of death in infants and children in the world," according to a recent review (Watson and Carcillo 2005, p. S3); deaths ascribed to the four major killers pneumonia, diarrhea, malaria, and measles often occur via a common pathway leading to fatal sepsis. Its incidence is likely to have been even higher before the availability of modern sanitation and medicines, and its action early in life would have made it a potent selective force. In modern hospitals, individuals with two copies of the inactive caspase-12 gene are both ~7.8-fold more likely to escape severe sepsis and more likely to survive if they do develop it, whereas heterozygotes show an intermediate level of protection (Saleh et al. 2004). We therefore suggest that the avoidance and survival of severe sepsis was the selective force that led to the spread of the inactive form of the caspase-12 gene.

This hypothesis leads to the question of why, if the inactive caspase-12 gene is so advantageous, it has not been fixed in humans and, indeed, in other species. Many infectious diseases require large host population sizes to maintain themselves and thus would have been rare or absent in archaic humans, when population sizes were small (Dobson 1992). Consequently, in small populations, there would have been no advantage associated with the inactive gene, and the evolutionary conservation of the gene (illustrated by a low human/chimpanzee K_a/K_s ratio) suggests that there may even have been a disadvantage, although the nature of this remains to be identified. Thus, selection for the inactive gene would have occurred only when the human population size became large.

When did the population start to grow? The Neolithic transition beginning ~10 KYA was associated with population growth and close contact with domestic animals, both of which would have increased the number of infections, but genetic studies suggest that the population started to grow long before the Neolithic period (Wall and Przeworski 2000). For example, one analysis suggested a start of expansion in sub-Saharan Africa 49–640 KYA (Reich and Goldstein 1998). According to our model, there would therefore have been an intermediate stage in which the active/inactive status of the gene was neutral or fluctuated between somewhat advantageous and disadvantageous in time or space. This could account for the accumulation of relatively diverse inactive haplotypes in Africa before the enormous expansion of a single inactive haplotype (fig. 4). But why did only a single haplotype expand? We cannot find any plausible biological difference between the most frequent haplotype and the more ancestral inactive ones—the SNPs that distinguish them lie in introns—so suggest that it could reflect either drift or some other advantage arising in a single population; if the latter, further studies of the caspase-12 gene may help to pinpoint the population and

Table 7

Published Summary Statistics for Human Genes

Genes	Population	Tajima's *D*	Fu and Li's *D*	Fay and Wu's *H* (*P*)[a]	Reference
Control genes:					
132 genes (95% range)	African Americans and European Americans	−1.66 to 1.56		−26.9 to 5.5 (.006–.940)	Akey et al. 2004
Genes showing positive selection:					
TRPV6[b]	European Americans	−2.74		−45.4 (.0001)	Akey et al. 2004
FOXP2	World	−2.20		−12.24 (<.05)	Enard et al. 2002
G6PD	World	−1.43	−1.13	NS[c]	Saunders et al. 2002
Duffy (FY)	Mandinka[b]	−1.40	−1.81		Hamblin and Di Rienzo 2000
TAS2R16	Brahui[b]	−1.69	−.49	−5.4 (.002)	Soranzo et al. 2005
MATP (AIM1)	Europeans[b]	−2.23	−2.90	−8.0 (<.025)	Soejima et al. 2006
CYP3A4	Europeans[b]	−1.76		(.006)[c]	Thompson et al. 2004

[a] These values should not be compared directly.
[b] NS = not significant. Gene or population showing the lowest values in studies involving many genes or populations.
[c] Numerical values were not given.

possibly the time in which this hypothetical key advance arose. More generally, selection on the caspase-12 gene appears to have started during a key period in human evolution, when modern behavior was developing. It therefore provides an example of the signature of selection that we may expect from this time period when unknown genes that may have contributed to modern human behavior may have experienced selection, although the pattern at any particular gene will depend on many factors, including stochastic variation, local mutation and recombination rate, and the strength of selection.

The "less is more" hypothesis of the importance of gene loss for human evolution (Olson 1999) remains to be systematically evaluated, but capsase-12 provides one striking example of the advantage that gene inactivation can confer and its role in human evolution.

Acknowledgments

We thank Joe Greenhill and Jonathan Bailey, for their contributions to generating sequence data; Chris Gillson, for assistance; Kate Rice and Bob Griffiths, for useful discussions; Peter Donnelly, for advice; Benjamin Voight, Sridhar Kudaravalli, and Jonathan Pritchard, for permission to refer to their unpublished work; and two referees, for suggesting improvements to the manuscript. We particularly thank Molly Przeworski for her advice and suggestions during the course of this study and for comments on the manuscript. This work was supported by The Wellcome Trust.

Web Resources

Accession numbers and URLs for data presented herein are as follows:

AceView, http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cgi?db=35g&c=Gene&l=CASP12P1
DnaSP, http://www.ub.es/dnasp/
Fluxus, http://www.fluxus-technology.com/ (for Network4.1.0.9)
FSTAT, http://www2.unil.ch/popgen/softwares/fstat.htm
GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for the *CASP12* genomic DNA sequence [accession number NC_000011] and the chimpanzee *CASP12* sequence [accession number NW_113990])
Haploview, http://www.broad.mit.edu/mpg/haploview/
ms, http://home.uchicago.edu/~rhudson1/source/mksamples.html
Stephens Web site, http://www.stat.washington.edu/stephens/software.html (for PHASE)
Sweep, http://www.broad.mit.edu/mpg/sweep/download.html (for version 1.0)

References

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2:e286
Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48
Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265
Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74:1111–1120
Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, et al (2002) A human genome diversity cell line panel. Science 296:261–262
Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87
Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963
Coop G, Griffiths RC (2004) Ancestral inference on gene trees under selection. Theor Popul Biol 66:219–232
Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, Goedert JJ, Buchbinder SP, Vittinghoff E, Gomperts E, Donfield S, Vlahov D, Kaslow R, Saah A, Rinaldo C, Detels R, O'Brien SJ (1996) Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the *CKR5* structural gene. Science 273: 1856–1862
Dobson A (1992) People and disease. In: Jones S, Martin R, Pilbeam D (eds) The Cambridge encyclopedia of human evolution. Cambridge University Press, Cambridge, United Kingdom, pp 411–420
Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Pääbo S (2002) Molecular evolution of *FOXP2*, a gene involved in speech and language. Nature 418:869–872

Evans PD, Anderson JR, Vallender EJ, Gilbert SL, Malcom CM, Dorus S, Lahn BT (2004) Adaptive evolution of *ASPM,* a major determinant of cerebral cortical size in humans. Hum Mol Genet 13: 489–494

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Fischer H, Koenig U, Eckhart L, Tschachler E (2002) Human caspase 12 has acquired deleterious mutations. Biochem Biophys Res Commun 293:722–726

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, Karimpour-Fard A, Glueck D, McGavran L, Berry R, Pollack J, Sikela JM (2004) Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol 2:937–954

Fu Y-X (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147: 915–925

Fu Y-X, Li W-H (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709

Goudet J (1995) FSTAT (vers 1.2): a computer program to calculate F-statistics. J Hered 86:485–486

Griffiths RC (2003) The frequency spectrum of a mutation, and its age, in a general diffusion model. Theor Popul Biol 64:241–251

Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66:1669–1679

Henshilwood CS, d'Errico F, Yates R, Jacobs Z, Tribolo C, Duller GA, Mercier N, Sealy JC, Valladas H, Watts I, Wintle AG (2002) Emergence of modern human behavior: Middle Stone Age engravings from South Africa. Science 295:1278–1280

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079

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

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164

International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796

——— (2005) A haplotype map of the human genome. Nature 437: 1299–1320

Jobling MA, Hurles ME, Tyler-Smith C (2004) Human evolutionary genetics. Garland Science, New York and Abingdon

Kayser M, Brauer S, Stoneking M (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol Biol Evol 20:893–900

Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765–777

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188:107–116

McDougall I, Brown FH, Fleagle JG (2005) Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433:733–736

Meiklejohn CD, Kim Y, Hartl DL, Parsch J (2004) Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. Genetics 168:265–279

Mekel-Bobrov N, Gilbert SL, Evans PD, Vallender EJ, Anderson JR, Hudson RR, Tishkoff SA, Lahn BT (2005) Ongoing adaptive evolution of *ASPM,* a brain size determinant in *Homo sapiens*. Science 309:1720–1722

Olson MV (1999) When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet 64:18–23

Preuss TM, Caceres M, Oldham MC, Geschwind DH (2004) Human brain evolution: insights from microarrays. Nat Rev Genet 5:850–860

Reich DE, Goldstein DB (1998) Genetic evidence for a Paleolithic human population expansion in Africa. Proc Natl Acad Sci USA 95: 8119–8123

Ronald J, Akey JM (2005) Genome-wide scans for loci under selection in humans. Hum Genomics 2:113–125

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837

Sabeti PC, Walsh E, Schaffner SF, Varilly P, Fry B, Hutcheson HB, Cullen M, Mikkelsen TS, Roy J, Patterson N, Cooper R, Reich D, Altshuler D, O'Brien S, Lander ES (2005) The case for selection at *CCR5-Δ32*. PLoS Biol 3:e378

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Saleh M, Vaillancourt JP, Graham RK, Huyck M, Srinivasula SM, Alnemri ES, Steinberg MH, Nolan V, Baldwin CT, Hotchkiss RS, Buchman TG, Zehnbauer BA, Hayden MR, Farrer LA, Roy S, Nicholson DW (2004) Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. Nature 429:75–79

Saunders MA, Hammer MF, Nachman MW (2002) Nucleotide variability at *G6pd* and the signature of malarial selection in humans. Genetics 162:1849–1861

Soejima M, Tachida H, Ishida T, Sano A, Koda Y (2006) Evidence for recent positive selection at the human *AIM1* locus in a European population. Mol Biol Evol 23:179–188

Soranzo N, Bufe B, Sabeti PC, Wilson JF, Weale ME, Marguerie R, Meyerhof W, Goldstein DB (2005) Positive selection on a high-sensitivity allele of the human bitter-taste receptor TAS2R16. Curr Biol 15:1257–1265

Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, Bridges CR, Shrager JB, Minugh-Purvis N, Mitchell MA (2004) Myosin gene mutation correlates with anatomical changes in the human lineage. Nature 428:415–418

Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Thompson EE, Kuttab-Boulos H, Witonsky D, Yang L, Roe BA, Di Rienzo A (2004) *CYP3A* variation and the evolution of salt-sensitivity variants. Am J Hum Genet 75:1059–1069

Vallender EJ, Lahn BT (2004) Positive selection on the human genome. Hum Mol Genet Spec No 2 13:R245–R254

Wall JD, Przeworski M (2000) When did the human population size start increasing? Genetics 155:1865–1874

Watson RS, Carcillo JA (2005) Scope and epidemiology of pediatric sepsis. Pediatr Crit Care Med 6:S3–S5

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370

Zhang J (2003) Evolution of the human *ASPM* gene, a major determinant of brain size. Genetics 165:2063–2070

Zhang J, Webb DM, Podlaha O (2002) Accelerated protein evolution and origins of human-specific features: FOXP2 as an example. Genetics 162:1825–1835