Positive Natural Selection in the Human Lineage

P. C. Sabeti *et al.*

Science **312**, 1614 (2006);
DOI: 10.1126/science.1124309

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

Positive Natural Selection in the Human Lineage

P. C. Sabeti,[1,2*] S. F. Schaffner,[1*†] B. Fry,[1] J. Lohmueller,[1,3] P. Varilly,[1] O. Shamovsky,[1] A. Palma,[1] T. S. Mikkelsen,[1] D. Altshuler,[1,4,5] E. S. Lander[1,6,7,8]

Positive natural selection is the force that drives the increase in prevalence of advantageous traits, and it has played a central role in our development as a species. Until recently, the study of natural selection in humans has largely been restricted to comparing individual candidate genes to theoretical expectations. The advent of genome-wide sequence and polymorphism data brings fundamental new tools to the study of natural selection. It is now possible to identify new candidates for selection and to reevaluate previous claims by comparison with empirical distributions of DNA sequence variation across the human genome and among populations. The flood of data and analytical methods, however, raises many new challenges. Here, we review approaches to detect positive natural selection, describe results from recent analyses of genome-wide data, and discuss the prospects and challenges ahead as we expand our understanding of the role of natural selection in shaping the human genome.

Homo sapiens, like all species, has been shaped by positive natural selection. As first articulated by Darwin and Wallace in 1858, positive selection is the principle that beneficial traits—those that make it more likely that their carriers will survive and reproduce—tend to become more frequent in populations over time (*1*). In the case of humans, these beneficial traits likely included bipedalism, speech, resistance to infectious diseases, and other adaptations to new and diverse environments. Understanding the traits (and genes underlying them) that have undergone positive selection during human evolution can provide insight into the events that have shaped our species, as well as into the diseases that continue to plague us today.

Until very recently, the only practical way to identify cases of positive selection in humans was to examine individual candidate genes. Allison noted in 1954 that the geographical distribution of sickle cell disease was limited to Africa and correlated with malaria endemicity (*2*); this observation led to the identification of the sickle cell mutation in the *Hemoglobin-B* gene (*HBB*) as having been the target of selection for malaria resistance (*3*, *4*). Since then, approximately 90 different loci have been proposed as possible targets for selection (table S1 provides a review of this literature).

Some of the proposed candidates for selection, like *HBB*, have strong support in the form of a functional mutation with an identified phenotypic effect that is a likely target of selection. In the case of *HBB*, the selected mutation creates a glutamate to valine amino acid change, but the target of selection need not be in the protein-coding region of a gene. For example, the Duffy antigen (*FY*) gene encodes a membrane protein used by the *Plasmodium vivax* malaria parasite to enter red blood cells. A mutation in the promoter of *FY* that disrupts protein expression confers protection against *P. vivax* malaria and was proposed to be selected for in regions of Africa where *P. vivax* malaria has been endemic (*5*). Another example is a mutation in a regulatory region near the gene for lactase (*LCT*) that allows lactose tolerance to persist into adulthood. This particular variant was apparently selected in parts of Europe after the domestication of cattle (*6*).

Often, however, the functional target of selection is not known. In some cases, candidate genes gain support because they lie in functional pathways, such as spermatogenesis and the immune response, that are known to be frequent targets for selection in other species. One example is protamine 1 (PRM1), a sperm-specific protein that compacts sperm DNA (*7*, *8*). Such cases, however, are the exception. Most proposed candidates lack compelling biological support. Rather, the argument for selection has relied solely on comparative and population genetic evidence.

Despite its great potential to illuminate new biological mechanisms, identification of selected loci by genetic evidence alone is fraught with methodological challenges. Studies based on comparisons between species suffer from limited power to detect individual incidents of selection, whereas studies based on human genetic variation have suffered from difficulties with assessing statistical significance. The evidence for positive selection has traditionally been evaluated by comparison with expectations under standard population genetic models, but the model parameters (especially those relating to population history) have been poorly constrained by available data, leading to large uncertainties in model predictions. One solution would be to assess significance by comparing empirical results from different studies, but this has been challenging because of the varied statistical tests, sizes of genomic region, and population samples used (see table S2 for examples).

The advent of whole-genome sequencing and increasingly complete surveys of genetic variation represent a turning point in the study of positive selection in humans. With these advances, humans can now join model organisms such as *Drosophila* (*9*) at the forefront of evolutionary studies. Newly available tools allow systematic survey of the genome to find the strongest candidate loci for natural selection, as well as to reevaluate previously proposed candidate genes, in comparison with genetic variation in the genome as a whole (the genome-wide empirical distribution). Although they permit us to make progress even while working out remaining theoretical issues, they also bring analytical challenges of their own, because they represent imperfect samples of genetic variation.

Here, we review genetic methods for detecting natural selection, discuss initial results about positive selection based on recent whole-genome analyses, and outline the potential and the challenges ahead in going from candidates of selection to proven examples of adaptive evolution.

Methods for Detecting Selection

When alleles (genetic variations) under positive selection increase in prevalence in a population, they leave distinctive "signatures," or patterns of genetic variation, in DNA sequence. These signatures can be identified by comparison with the background distribution of genetic variation in humans, which is generally argued to evolve largely under neutrality (*10*). This is in accord with the neutral theory, which proposes that most observed genetic variation, both within and between species, is neutral (i.e., has no effect on an individual's fitness), so that its population prevalence changes over time by chance alone (so-called "genetic drift") (*11*). A great challenge for population genetics–based signatures (sections ii to v below) is determining whether a signature is due to selection or to the confounding effects of population demographic history, such as bottlenecks (periods of reduced population size), expansions, and subdivided populations.

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Harvard Medical School, Boston, MA, USA. [3]Brown University, Providence, RI, USA. [4]Departments of Genetics and Medicine, Harvard Medical School, Boston, MA, USA. [5]Department of Molecular Biology, Center for Human Genetic Research, and Diabetes Unit, Massachusetts General Hospital, Boston, MA, USA. [6]Department of Biology, MIT, Cambridge, MA, USA. [7]Whitehead Institute for Biomedical Research, Cambridge, MA, USA. [8]Department of Systems Biology, Harvard Medical School, Boston, MA, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: sfs@broad.mit.edu

Many specific statistical tests have been proposed to detect positive selection (table S3 provides a review), but they are all based broadly on five signatures. Below, we describe the nature of each signature, an estimate of the window of evolutionary time in which it can be used to detect moderately strong selection in humans (Fig. 1), and its strengths and weaknesses in human studies. Several excellent reviews (*12*–*18*) provide more information, as well as background on coalescent modeling and on other types of selection (e.g., purifying selection and balancing selection). It should be noted that many instances of selection are likely not detectable by any currently proposed method—for example, if the selective advantage is too small or selection acts on an allele that is already at an appreciable frequency in the population (*19*).

(*i*) *High proportion of function-altering mutations* (*age, many millions of years*). Genetic variants that alter protein function are usually deleterious and are thus less likely to become common or reach fixation (i.e., 100% frequency) than are mutations that have no functional effect on the protein (i.e., silent mutations). Positive selection over a prolonged period, however, can increase the fixation rate of beneficial function-altering mutations (*20*, *21*), and such changes can be measured by comparison of DNA sequence between species. The increase can be detected by comparing the rate of nonsynonymous (amino acid–altering) changes with the rate of synonymous (silent) or other presumed neutral changes, by comparison with the rate in other lineages, or by comparison with intraspecies diversity. One extreme example of this kind of signature is found in the gene *PRM1*, mentioned earlier, which has 13 nonsynonymous and 1 synonymous differences between human and chimpanzee (*7*, *8*) (Fig. 2). Statistical tests commonly used to detect this signature include the Ka/Ks test, relative rate tests, and the McDonald-Kreitman test (*20*–*22*). Similar tests can also be applied to other functional sites, such as noncoding regulatory sequences, and their development is an area of active research (*23*, *24*).

This signature can be detected over a large range of evolutionary time scales. Moreover, it focuses on the beneficial alleles themselves, eliminating ambiguity about the target of selection. Its power is limited, however, because multiple selected changes are required before a gene will stand out against the background neutral rate of change. It is thus typically possible to detect only ongoing or recurrent selection. In practice, when the human genome is surveyed in this manner, few individual genes will give statistically significant signals, after correction for the large number of genes tested. However, the signature can readily be used to detect positive selection across sets of multiple genes (*25*). For example, genes involved in gametogenesis clearly stand out as a class having a high proportion of nonsynonymous substitutions (*25*–*27*).

(*ii*) *Reduction in genetic diversity* (*age <250,000 years*). As an allele increases in population frequency, variants at nearby locations on the same chromosome (linked variants) also rise in frequency. Such so-called "hitchhiking" leads to a "selective sweep," which alters the typical pattern of genetic variation in the region. In a complete selective sweep, the selected allele rises to fixation, bringing with it closely linked variants; this eliminates diversity in the immediate vicinity and decreases it in a larger region. New mutations eventually restore diversity, but these appear slowly (because mutation is rare) and are initially at low frequency. Positive selection thus creates a signature consisting of a region of low overall diversity, with an excess of rare alleles.

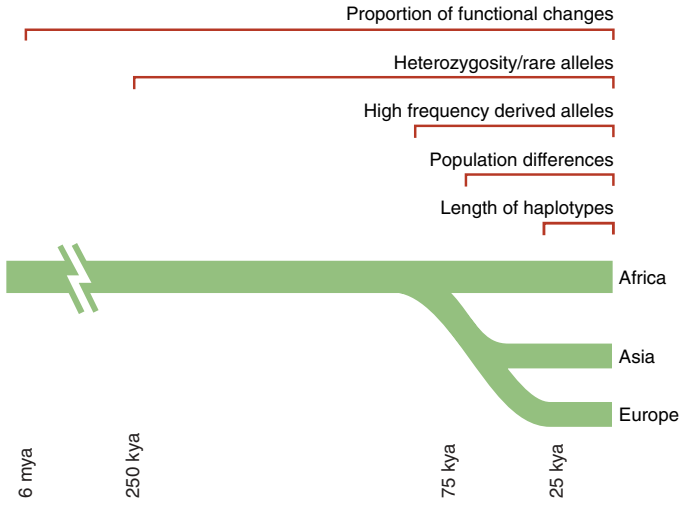

Fig. 1. Time scales for the signatures of selection. The five signatures of selection persist over varying time scales. A rough estimate is shown of how long each is useful for detecting selection in humans. (See fig. S1 for details on how the approximate time scales were estimated).

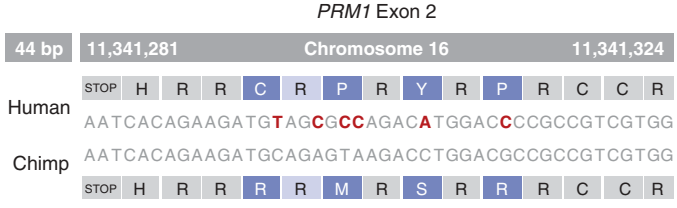

Fig. 2. Excess of function-altering mutations in *PRM1* exon 2. The *PRM1* gene exon 2 contains six differences between humans and chimpanzees, five of which alter amino acids (*7*, *8*).

Unlike excess functional changes, which involve differences between species, selective sweeps are detected in genetic variation within a species. The most common type of variant used is the single-nucleotide polymorphism (SNP). As an example, Akey *et al*. identified a 115-kb region containing four genes including the Kell blood antigen, which showed an overall reduction in diversity and more rare alleles in Europeans than expected under neutrality (Fig. 3) (*28*). Statistical tests commonly used to detect this signal include Tajima's D, the Hudson-Kreitman-Aguadé (HKA) test, and Fu and Li's D* (*29*–*32*).

Reduction in genetic diversity can be particularly useful because it persists longer than other population genetic signatures. The characteristic time for new mutations to drift to high frequency under neutral evolution in the human population is ~1 million years. This means that statistically significant signals of selection can persist for several hundred thousand years, long enough to encompass the origins of modern humans.

The size of the genome region affected by a sweep depends on the strength of positive selection and, thus, the speed at which the selected allele reached high frequency. That is, rapid sweeps affect large regions. If an allele confers a selective advantage of 1% (considered moderately strong selection), the modal size of the affected genomic region has been estimated to be roughly 600,000 base pairs (600 kb) (*27*). Such a large size facilitates detection, although it also makes the subsequent task of identifying the causal variant more difficult. Another challenge is that the signature may be difficult to distinguish from effects of demographic history, e.g., an expanding population increases the fraction of rare alleles.

(*iii*) *High-frequency derived alleles* (*age <80,000 years*). Derived (that is, nonancestral) alleles arise by new mutation, and they typically have lower allele frequencies than ancestral alleles (*33*). In a selective sweep, however, derived alleles linked to the beneficial allele can hitchhike to high frequency. Because many of these derived alleles will not reach complete fixation (as a result of an incomplete sweep or recombination of the selected allele during the sweep), positive selection creates a signature of a region containing many high-frequency derived alleles. A good example of this kind of

signature is the 10-kb region around the Duffy red cell antigen (*FY*), which has an excess of high-frequency derived alleles in Africans, thought to be the result of selection for resistance to *P. vivax* malaria (Fig. 4) (*34*, *35*). The most commonly used test for derived alleles is Fay and Wu's H (*36*).

Tests based on derived alleles require knowledge of the ancestral allele. In practice, the ancestral allele is inferred from the allele present in closely related species, with the assumption that mutation occurred only once at this position and that it occurred after the two species diverged (*36*). Determination of the ancestral allele in humans is facilitated by the availability of the chimpanzee genome sequence and by the growing data from additional primate genomes. The derived-alleles signature differs from the rare-allele signature discussed above in two important ways. First, different demographic effects are potential confounders [for example, population expansion is a major confounder for rare-alleles tests but not for derived-alleles tests (*36*), whereas population subdivision is more of a problem for the latter (*37*)]. Second, the signature persists for a shorter period (*37*) because high-frequency derived alleles rapidly drift to or near fixation.

(*iv*) *Differences between populations* (*age <50,000 to 75,000 years*). When geographically separate populations are subject to distinct environmental or cultural pressures, positive selection may change the frequency of an allele in one population but not in another. Relatively large differences in allele frequencies between populations (at the selected allele itself or in surrounding variation) may therefore signal a locus that has undergone positive selection. For example, the *FY*O* allele at the Duffy locus is at or near fixation in sub-Saharan Africa but rare in other parts of the world, an extreme case of population differentiation (Fig. 5) (*34*, *38*). Similarly, the region around the *LCT* locus demonstrates large population differentiation between Europeans and non-Europeans, reflecting strong selection for the lactase persistence allele in Europeans (*6*). Commonly used statistics for population differentiation include F_{ST} and p_{excess} (*39*–*41*).

Population differentiation can only arise when populations are at least partially isolated reproductively. For humans, it thus pertains largely to events that occurred after the major human migrations out of Africa some 50,000 to 75,000 years ago (Fig. 1). As with other population genetic signatures, distinguishing between genuine selection and the effect of demographic history, especially population bottlenecks, on genetic variation can be difficult.

(*v*) *Long haplotypes* (*age <30,000 years*). Under positive selection, a selected allele may rise in prevalence rapidly enough that recombination does not substantially break down the association with alleles at nearby loci on the ancestral chromosome. Such a collection of alleles in a chromosomal region that tend to occur together in individuals is termed a haplotype. Selective sweeps can produce a distinctive signature that would not be expected under neutral drift—namely, an allele that has both high frequency (typical of an old allele) and long-range associations with other alleles (typical of a young allele). The long-range associations are seen as a long haplotype that has not been broken down by recombination. For example, the lactase persistence allele at the *LCT* locus lies on a haplotype that is common (~77%) in Europeans but that extends largely undisrupted for more than 1 million base pairs (1 Mb) (Fig. 6) (*6*), much farther than is typical for an allele of that frequency. This signature can be detected with the long-range haplotype (LRH) test, haplotype similarity, and other haplotype-sharing methods (*42*–*45*). Developing such tests is an area of vigorous current investigation (*46*, *47*).

Long haplotypes are useful for detecting partial selective sweeps, with allele frequencies as low as ~10%. Tests for this signature are relatively robust to the choice of genetic markers used (ascertainment bias), an important issue in practical applications. Another advantage of this test is that it can identify a narrow candidate region, even a single gene. One limitation of the test is that long-range haplotypes persist for relatively short periods of time, because recombination rapidly breaks down the haplotype. After 30,000 years, a typical chromosome will have undergone more than one crossover per 100 kb, leaving fragments that are too short to detect. A critical issue with this kind of signature is accurate control for variation in recombination rate; evidence that recombination rates may vary between haplotypes is a concern (*48*).

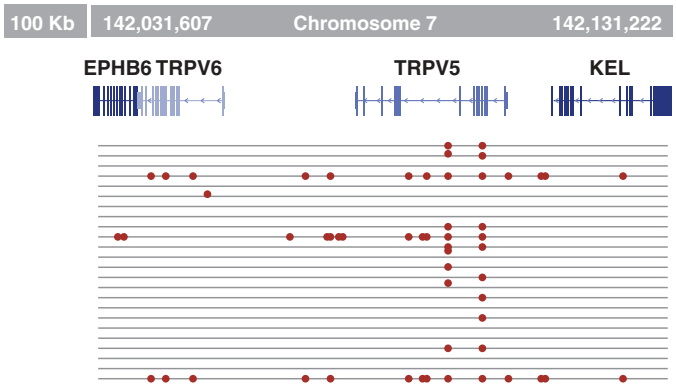

Fig. 3. Low diversity and many rare alleles at the Kell blood antigen cluster. On the basis of three different statistical tests, the 115-kb region (containing four genes) shows evidence of a selective sweep in Europeans (*28*).

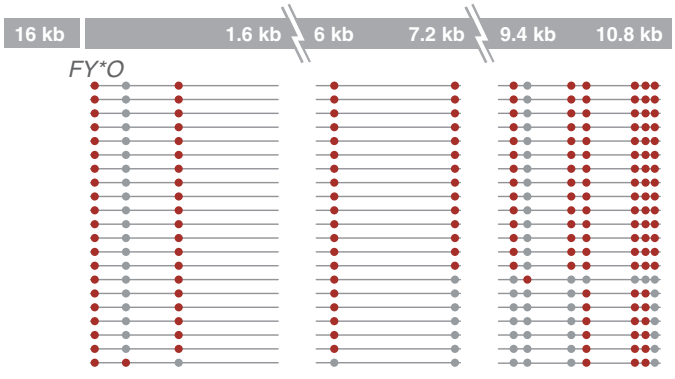

Fig. 4. Excess of high-frequency derived alleles at the Duffy red cell antigen (*FY*) gene (*34*). The 10-kb region near the gene has far greater prevalence of derived alleles (represented by red dots) than of ancestral alleles (represented by gray dots).

Genome-Wide Studies

Genome-wide studies of positive selection in humans have recently become possible because of the availability of (i) a near-complete sequence of the human genome (*49*, *50*), together with an increasing number of genome sequences for other species such as chimpanzee (*27*), mouse (*51*), and dog (*52*), and (ii) large catalogs of human genetic variation, such as those created by the SNP consortium (*53*), International Haplotype Map (HapMap) Project (*54*), and Perlegen Sciences (*55*). The current data are still limited. Additional closely related species are needed, and the polymorphism data are incomplete and not fully representative of human genetic variation. Nonetheless, the data are expanding rapidly: Sequencing of macaque and orangutan is far along, and sequencing of gorilla is beginning. The HapMap project has completed data collection for its second phase, with a SNP density higher by a factor of 4 that will include more than a third of all the estimated 10 to 12 million common human SNPs.

These data sets, although still limited, have already enabled initial genome-wide empirical studies of natural selection in the human genome. Seven large-scale studies of positive selection have recently been published, including

four surveys of amino acid–altering mutations in the human lineage and three surveys of human genetic diversity; more studies will likely be in print by the time of this publication (*46*, *47*). They provide a first look into the genome-wide distribution of diversity, identify high-priority candidates for natural selection without regard to previous biological hypotheses, and allow us to begin to re-evaluate earlier reports. These new studies also reveal the challenges ahead in extracting a coherent picture of adaptive selection in humans from the flood of new information.

Function-altering mutations in the human lineage. Four studies have examined natural selection in the ancient human lineage by considering amino acid–altering mutations. Each of the studies used one of two basic approaches: two-way comparison of human and chimpanzee orthologs (genes in multiple species evolved from a common ancestor) (*25*–*27*) or three-way comparison of human, chimpanzee, and murid (mouse or rat) orthologs (*27*, *56*). The advantage of the latter strategy is that it can distinguish between events that occurred in the human or chimpanzee lineage, but the use of a distant species for comparison limits the number of genes that can be studied (~7000 versus 11,000 to 14,000 genes for the two-way comparison). One of these studies (*26*) also used human polymorphism data to provide additional information about expected rates of change in different parts of the genome.

These analyses offer preliminary insights into the evolution of various functional classes of genes. In particular, they suggest that ongoing positive selection on humans has been strongest for genes related to immune response, reproduction (especially spermatogenesis), and sensory perception (especially olfaction). The studies are not completely consistent with each other (e.g., of two studies, using largely the same data, one found strong evidence for selection in spermatogenesis-related genes and the other did not) (*25*, *56*), but the overall picture is consistent with studies in other mammals, and the results seem plausible in terms of evolutionary predictions.

The studies also found that X-linked genes are significantly overrepresented among rapidly evolving genes (*25*, *27*). Much of the increased selection seen on the X chromosome likely arises from the larger number of sperm- and testis-associated genes (*25*), which are frequent targets of selection, on the chromosome. In addition, the hemizygosity of the X chromosome in males exposes recessive alleles to selective pressure, which may promote rapid evolution (*57*).

The great majority of genes identified in these studies as candidates for positive selection are novel, with the potential to illuminate previously unsuspected biological mechanisms. They include several genes with testis-specific expression (*USP26*, *C15orf2*, and *HYAL3*), several involved in immune regulation (e.g., *CD58*, *APOBEC3F*, and *CD72*), several tumor antigens (e.g., *SAGE1* and *MAGEC2*), and many more with as-yet-unknown functions (e.g., *FLJ46156I*, *ABHD1*, and *LOC389458*) (*25*, *26*). Because power to detect selection at any single gene (as compared to categories of genes) is limited using this approach, false positives are a concern, but one study was able to demonstrate that the candidate genes as a group had significant evidence for selection (*25*).

Interestingly, of 39 previously reported candidates based on function-altering mutations, only 4 were in the top 1% of candidates for selection in these genome-wide empirical studies (table S4). These four genes encode two sperm-related proteins [*PRM1* and *PRM2*, with the former being the strongest candidate for selection in one of the studies (*25*)], one antiviral enzyme (*APOBEC3G*), and an Rh blood antigen (*RHCE*). Of the remaining 35 genes, some had missing or incomplete data and a few had weaker but suggestive evidence for selection (e.g., *SEMG1*, *VIRL1*, and *SRY*). However, some may well be false positives due to previously insufficient knowledge of gene variation across the genome under neutrality.

The increasing number of genome sequences of closely related species will greatly expand the set of genes that can be studied by multiple-species comparison. These new data should also improve estimates of the neutral substitution rate (the rate at which fixed differences between species accumulate under neutral evolution). Humans and chimpanzees show an average of only 4.5 synonymous differences per gene (a number often used to estimate the neutral substitution rate); 10 times that number can be expected between human and macaque. The power of these studies will also be increased by better ways to recognize likely functional changes. These could come from both a greater understanding of the effect of specific mutations on protein structure and function and from a clearer understanding of the noncoding regulatory regions of the genome.

Genetic variation within human populations. Three published genome-wide surveys have used human genetic variation data to study recent selective sweeps (up to ~250,000 years ago). One of these (*27*) identified regions of the human genome with unusually low diversity, with subsequent confirmation by testing for an excess of high-frequency derived alleles. A second study (*54*) used the HapMap data to examine three signatures of selection: population differentiation, allele frequency spectrum, and long haplotypes. The

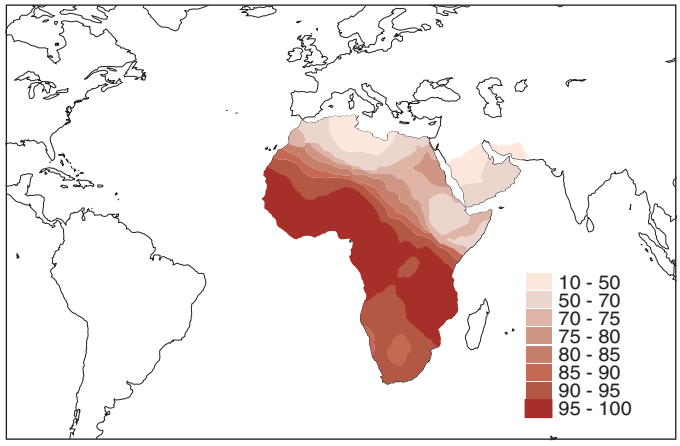

Fig. 5. Extreme population differences in *FY*O* allele frequency. The *FY*O* allele, which confers resistance to *P. vivax* malaria, is prevalent and even fixed in many African populations, but virtually absent outside Africa (*38*).

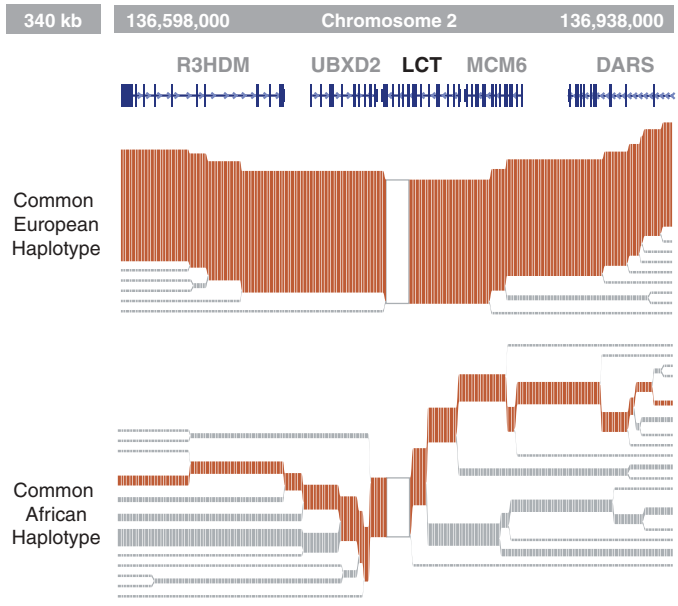

Fig. 6. Long haplotype surrounding the lactase persistence allele. The lactase persistence allele is prevalent (~77%) in European populations but lies on a long haplotype, suggesting that it is of recent origin (*6*).

third investigation (*58*) searched for regions with rare alleles in the Perlegen data and examined high-frequency derived alleles for confirmation.

The three surveys used data that were not developed primarily for the purpose of studying selection. Thus, potential biases in the choice of genetic markers studied (ascertainment biases) had to be taken into account in their analyses (*59*, *60*). For example, the procedure for selecting SNPs in the HapMap data was biased toward high-frequency derived alleles; the analysis of the HapMap data therefore avoided tests based on frequencies of derived alleles and focused instead on tests of overall diversity. Only subsequent work with truly unbiased data sets will reveal how successfully these studies succeeded in avoiding ascertainment bias.

The statistical power of these studies to detect selective events is limited by the still-incomplete nature of current SNP catalogs and by the limited number of individuals genotyped. Greater SNP density will permit more complete dissection of haplotypes, with finer granularity and increased power, particularly for frequency-based measures (as much as 50% greater for older sweeps, simulations suggest). Genotyping of more individuals will be essential for detecting partial sweeps at low-frequency alleles.

Despite these limitations, the initial analyses are striking and will fuel much additional research. The majority of strong candidates for selection found in these surveys were novel. These included the *LARGE*, *ALMS1*, and *SLC24A5* genes and several gene clusters that include the genes *CD36*, *ITGAE*, *FAF1*, *SYT1*, and *GRAP2* (*54*). Interestingly, some high-scoring regions contained no known genes at all, which may reflect selection on noncoding sequences (*27*, *54*).

Also notable is that fewer than half of loci previously identified as targets of positive selection ranked among the top candidates in the empirical genome-wide analyses. When 81 previously proposed candidate genes were reexamined in the genome-wide data, using seven different tests for selection, only 25 were in the top 1% of the genome on at least one test (Fig. 7 and table S4). For the remaining genes, which include well-known cases like *FOXP2*, *G6PD*, and *MC1R*, the genome-wide evidence is weaker or absent.

Why have many earlier results fared poorly in genome-wide studies? In some cases the explanation is insufficient power, either because of the tests employed in the new studies or because of inadequate coverage of some genomic regions. An example of the latter effect may be *G6PD*, which scores in the top 3% of loci but is not an extreme outlier, despite having an established association with malaria resistance (*43*, *61*). This may be because the locus lies in a genomic region (Xq28) with low SNP density in HapMap data, although it may also indicate that the locus has a relatively modest effect. The possibility that these tests could miss some signals is illustrated by the Duffy locus. The role of selection there is well established, but the genetic signal is completely invisible under most of the tests used in the genome-wide surveys [the exception being a single-marker population differentiation test, in which the signal is clearly observed (table S4)]. For other loci, a signal might be missed by all of the tests (*62*).

Based on these genome-wide empirical comparisons, however, some previous claims of association may well represent false positives. That is, it is now clear that some signals that stood out in comparison to simple population genetic models do not stand out relative to the genome-wide distribution of diversity. For example, haplotypes that span relatively short distances (e.g., tens of kilobases) and long haplotypes around rare alleles (e.g., the *CCR5*-Δ*32* mutation) can now be seen to be common features in the genome (*47*, *63*, *64*) and not particularly suggestive of selection. As understanding of the genome-wide landscape improves, the precision of these tests will undoubtedly improve as well.

On the other hand, many of the most well-studied and convincing cases for selection [e.g., *LCT*, *HBB*, *FY*, and the major histocompatibility complex (MHC)] are clear outliers in the empirical distributions. Many of these are genes already associated with adaptive evolu-

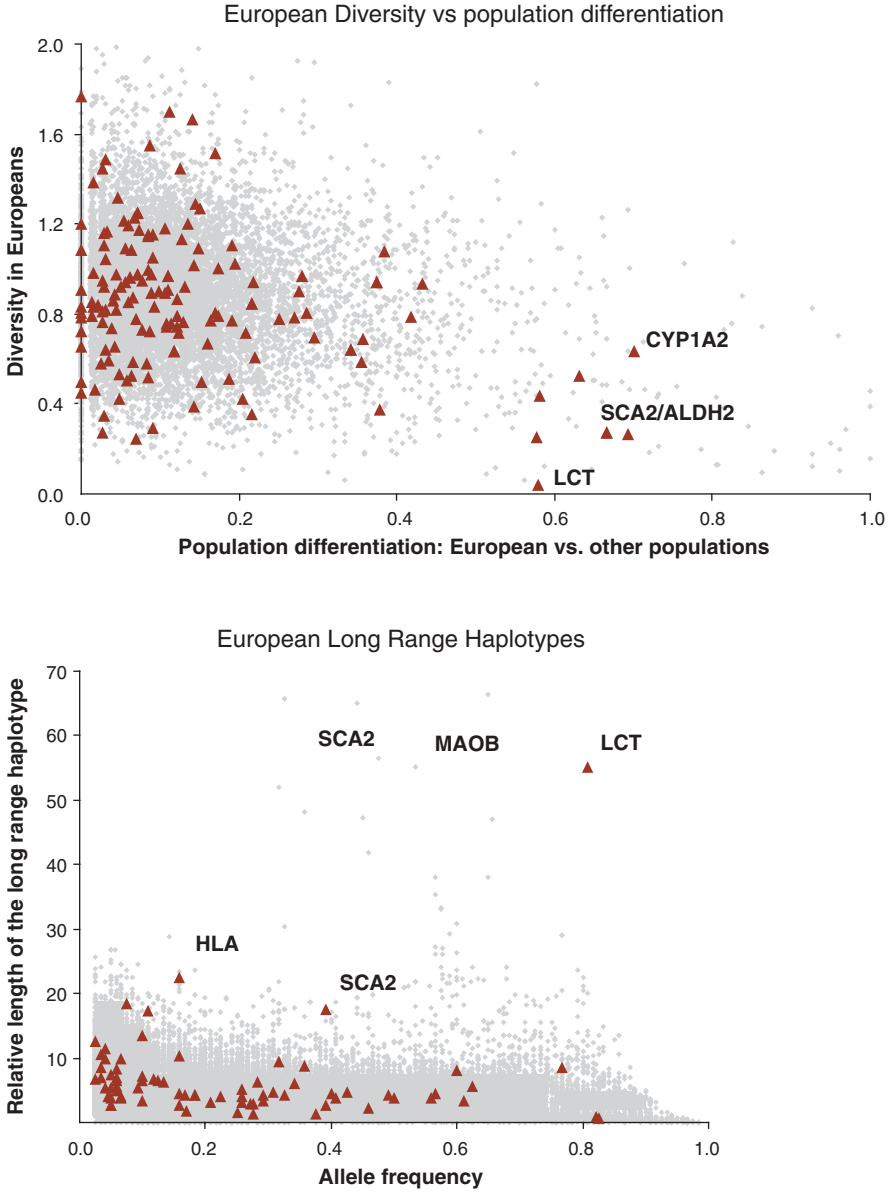

Fig. 7. The previous candidates for selection (red triangles), identified by limited empirical data, in comparison with the newly available genome-wide empirical data sets (gray diamonds). The results are presented here for a European sample for three signatures: diversity, population differentiation, and long-range haplotypes. More detailed results are presented in table S4 and fig. S2.

tion, such as those involved in resistance to malaria (*HBB*, *CD40L*, *FY*, and newly identified *CD36*) or other infectious diseases (MHC). This reinforces the notion that infectious diseases, and specifically malaria, have been among the strongest selective pressures in recent human history. Other previously identified candidates found in the survey data included LCT in Europeans (found by long haplotype and diversity/frequency tests), *DMD* and the *SCA2*/*ALDH2* cluster (long haplotypes), and the *CYP3A4*/*CYP3A5* cluster, *ALDH2*, and immunoglobulin A (diversity/frequency tests).

As with the surveys of amino acid–altering mutations, the studies based on human genetic variation found an excess of candidates on the X chromosome: 10 of 33 candidates in the HapMap-based study lie on the X chromosome, which comprises only 5% of the genome. Although it is prudent to withhold judgement (a higher rate of false positives on the X chromosome could arise from stronger effects of bottlenecks, given the smaller number of X chromosomes than autosomes in the population), it is plausible that this reflects a different impact of selection on the X chromosome. Reassuringly, a similar excess is seen in a population that has experienced major recent bottleneck effects (Europeans) and in a population that has not (West Africans).

The transition to using empirical, rather than purely theoretical, distributions as the basis for selecting candidates represents real progress and lays the foundation for fruitful work to come. It should be remembered, however, that the demonstration that a gene is a clear outlier does not definitively prove that it is the target of selection. Because we do not know the underlying proportion of loci that have experienced positive selection, we cannot calculate a precise posterior probability of selection. In the end, convincing proof will require an understanding of biological function.

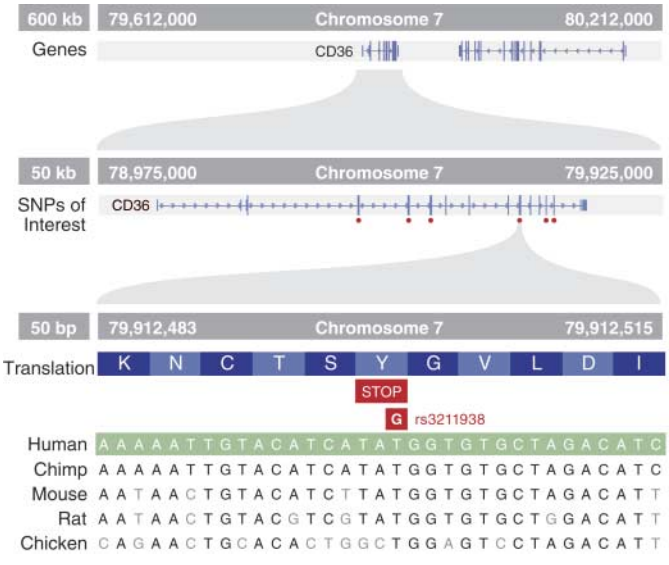

Fig. 8. Identification of functional polymorphism associated with a signature of selection at *CD36*. An allele at *CD36* identified to be under selection by the LRH test (*42*, *53*) has been associated with differential susceptibility to *P. falciparum* cerebral malaria. An animated version of the browser by B. Fry to scan the selected region for functional variations is available at www.broad.mit.edu/mpg/pubs/sabeti-science2006.

From Candidate to Function

True understanding of the role of adaptive evolution will require both better constrained models of neutral evolution (which can be derived in part from the same new data sets discussed here) and detailed, case-by-case analysis of candidate loci to identify those with biological evidence for selection. The latter, which will also help inform estimates of how common positive selection has been in humans, is the real goal, because it is the selected traits themselves that are of the most interest.

Identifying and understanding the traits that have been targets of selection will be a major challenge. Consideration of *HBB* sickle cell, one of the earliest successful dissections of adaptive evolution, demonstrates the depth of work required for this pursuit. Forty years of constant effort, by a succession of researchers including Pauling, Ingram, Allison, and Perutz, were required to unravel the association with malaria and the biochemical properties of the sickle cell mutation (*2*, *65*, *66*). Even now, there is still work to be done to understand exactly how the sickle state inhibits malaria infection.

Dissecting selection at a specific locus can be approached from two directions: finding a DNA change with functional molecular consequence or finding an association to a phenotypic difference in the human population. The first approach begins with good genetic annotation of the region, including both coding and regulatory regions, and will be enhanced by ongoing advances in comparative genomics. Depending on the kind of event under study, the functional changes might be found through comparisons between species, between populations, or between haplotypes. Such clues can be the basis of diverse means of biological experimentation. The second approach, which is only possible if the selected variant is still polymorphic in humans, usually depends on knowledge of the underlying biology of the region. The associated phenotype might be measured in human populations (for example, malaria resistance) or in cell lines (for example, protein function or expression).

The easier end of the spectrum of individual cases is illustrated by a candidate locus discovered among the HapMap data, the gene *CD36* (*54*). A *CD36* haplotype present in Africa and absent elsewhere showed evidence for recent selection by the LRH test. Closer inspection showed that this haplotype contains a nonsense mutation (amino acid changed to a stop codon), T188G, that has been associated with differential susceptibility to *P. falciparum* cerebral malaria (Fig. 8) (*67*, *68*).

A somewhat harder case is the *LARGE* gene. A haplotype residing entirely within this gene shows evidence for selection in West Africans, simplifying the issue of identifying the causal gene (*54*). However, the associated phenotype is a mystery. The function of *LARGE* (it encodes a member of the N-acetylglucosaminyltransferase gene family) is not well understood, although a mutation in it is known to cause muscular dystrophy. (Curiously, *DMD*, another gene with mutations causing muscular dystrophy, also shows evidence for selection in the same population.)

More difficult still are cases in which a causal gene has not been identified. For example, a 0.5-Mb region in chromosome 2q11.1 shows low diversity and an excess of rare alleles in the West African data (*54*). The region contains four known and two putative genes, with no indication as to which was responsible for selection. A very strong candidate region on chromosome 4 suffers from the opposite problem (*27*, *54*): It contains no known genes, although the region has been associated with severe obesity (*69*, *70*). In these cases, the best prospect is to narrow the candidate region and identify all the functional changes (both coding and regulatory) contained therein.

In many cases, comparative genomics and population-based association studies can be extremely helpful. For example, one of the genes showing strong population differentiation in HapMap data was one of unknown function, *SLC24A5* (*54*). Independently, Lamason *et al*. identified a mutation in the Zebrafish homolog of this gene that is responsible for a pigmentation phenotype (*71*). Guided by the two findings, the investigators demonstrated that a human variant in the gene explains roughly one-third of the variation in pigmentation between Europeans and West Africans and that the European variant had likely been a target of selection. In this case, the combination of biological data from a model system and genome-wide polymorphism data rapidly established a plausible link between natural selection and a human trait.

The most difficult case arises when selected alleles have risen to fixation in the modern

human population, so that no phenotypic variation remains. Such cases include some candidates identified by virtue of low diversity in population genetic data, although most are revealed by interspecies studies. Dissection of these events, some of them crucial steps in the development of modern *H. sapiens*, will require better understanding of the biological role played by the genes. For example, there is suggestive genetic evidence for positive selection at the gene *FOXP2* but no relevant variation in the modern human population; the only clue as to what the selected trait might have been is the observation that rare mutations in this gene lead to speech defects (*72*, *73*).

Conclusions

The advent of genome-wide sequence and variation data has dramatically expanded approaches to identifying possible sites of natural selection. Much work still needs to be done to create unbiased data sets of genetic variants and to refine analytical techniques. Still, we have caught a first glimpse of a vast new landscape. We now see that only a small fraction of loci with evidence for positive selection were found by previous approaches, suggesting that many more examples are likely to be found in the coming years. With an even deeper inventory of human variation, it should soon be realistic to generate a catalog of the human loci with signals for selection above a given threshold.

The field is expanding rapidly, as evidenced by the continual flood of papers claiming new regions as candidates for selection and reporting new methods for detecting selection. It will be a challenge to interpret this new information, working toward a coherent picture of human evolution. A set of community standards for reporting and interpreting data will help advance the field and are beginning to emerge. Three key components will be (i) clear demonstration of the utility of new statistical tests, (ii) more rigorous demonstration of evidence for natural selection, and (iii) the inclusion of functional evidence for candidate loci, where possible.

First, new statistical tests will continue to be introduced, both to improve existing methods and to address characteristics of particular data sets (e.g., genotype rather than sequence data). Such methods should be evaluated by direct comparison with published methods—an obvious step, but one too seldom taken. The power, robustness to demographic history, and utility for varied data sets should all be assessed by simulation studies done under a range of demographic scenarios. Application to empirical data, justification for statistical thresholds, and control for multiple testing should be clearly described.

Second, evidence for selection at new candidate loci should be evaluated both relative to theoretical model distributions (ideally, tailored to empirical data) and by comparison to empirical, genome-wide distributions. Good theoretical models are needed to interpret the significance of genome-wide outliers. Whatever a theoretical model might suggest, however, it is also crucial to report where a locus falls in the empirical distribution. In cases where the exact genome-wide distribution is not yet available for a particular test (as in the case of resequencing data), attempts should be made to provide sufficient data for empirical comparison.

Third, genetic evidence for selection is considerably enhanced by functional evidence. This is important because the actual extent of positive selection in the human lineage is unknown, making it hard to define thresholds for genetic evidence of selection. The functional evidence might take many forms, e.g., correlation of the selected allele with human phenotypic variation, model system, or in vitro laboratory studies of the selected allele. The strongest evidence would include both identification of a functional variant in humans and evidence for the advantage that the trait provides.

The quest to identify selected traits is driven not just by curiosity about the past but also by concern for human health. Positive selection, in many cases, represents a response to pathogens or other causes of illness, or to new diet and environmental conditions. Many of these forces are still present today. Moreover, positive selection has wrought changes to human biology, to which the rest of the genome may not yet have had time to adapt. As a result, polymorphic alleles at loci that have undergone recent selection may also be good candidates for risk factors for modern disease.

References and Notes

1. C. Darwin, A. R. Wallace, *Proceedings of Linnean Society of London* **3**, 45 (1858).
2. A. C. Allison, *Br. Med. J.* **4857**, 290 (1954).
3. M. Currat *et al.*, *Am. J. Hum. Genet.* **70**, 207 (2002).
4. J. Ohashi *et al.*, *Am. J. Hum. Genet.* **74**, 1198 (2004).
5. M. T. Hamblin, A. Di Rienzo, *Am. J. Hum. Genet.* **66**, 1669 (2000).
6. T. Bersaglieri *et al.*, *Am. J. Hum. Genet.* **74**, 1111 (2004).
7. A. P. Rooney, J. Zhang, *Mol. Biol. Evol.* **16**, 706 (1999).
8. G. J. Wyckoff, W. Wang, C. I. Wu, *Nature* **403**, 304 (2000).
9. R. R. Hudson, K. Bailey, D. Skarecky, J. Kwiatowski, F. J. Ayala, *Genetics* **136**, 1329 (1994).
10. I. Hellmann, I. Ebersberger, S. E. Ptak, S. Paabo, M. Przeworski, *Am. J. Hum. Genet.* **72**, 1527 (2003).
11. M. Kimura, *Nature* **217**, 624 (1968).
12. D. L. Hartl, A. G. Clark, *Principles of Population Genetics* (Sinauer Associates, Sunderland, Mass., 2nd ed., 1989).
13. E. J. Vallender, B. T. Lahn, *Hum. Mol. Genet.* **13 Spec No 2**, R245 (October 1, 2004).
14. J. Ronald, J. M. Akey, *Human Genomics* **2**, 113 (2005).
15. R. Nielsen, *Annu. Rev. Genet.* (August 31, 2005).
16. M. Kreitman, *Annu. Rev. Genomics Hum. Genet.* **1**, 539 (2000).
17. A. M. Bowcock *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 839 (1991).
18. M. Bamshad, S. P. Wooding, *Nat. Rev. Genet.* **4**, 99 (2003).
19. M. Przeworski, G. Coop, J. D. Wall, *Evolution Int. J. Org. Evolution* **59**, 2312 (2005).
20. W. H. Li, C. I. Wu, C. C. Luo, *Mol. Biol. Evol.* **2**, 150 (1985).
21. A. L. Hughes, M. Nei, *Nature* **335**, 167 (1988).
22. J. H. McDonald, M. Kreitman, *Nature* **351**, 652 (1991).
23. P. Andolfatto, *Nature* **437**, 1149 (2005).
24. M. V. Rockman *et al.*, *PLoS Biol.* **3**, e387 (2005).
25. R. Nielsen *et al.*, *PLoS Biol.* **3**, e170 (2005).
26. C. D. Bustamante *et al.*, *Nature* **437**, 1153 (2005).
27. Chimpanzee Sequencing and Analysis Consortium, *Nature* **437**, 69 (2005).
28. J. M. Akey *et al.*, *PLoS Biol.* **2**, e286 (2004).
29. F. Tajima, *Genetics* **123**, 585 (1989).
30. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
31. R. R. Hudson, M. Kreitman, M. Aguade, *Genetics* **116**, 153 (1987).
32. Y. X. Fu, W. H. Li, *Genetics* **133**, 693 (1993).
33. G. A. Watterson, H. A. Guess, *Theor. Popul. Biol.* **11**, 141 (1977).
34. M. T. Hamblin, E. E. Thompson, A. Di Rienzo, *Am. J. Hum. Genet.* **70**, 369 (2002).
35. A. A. Escalante *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1980 (2005).
36. J. C. Fay, C. I. Wu, *Genetics* **155**, 1405 (2000).
37. M. Przeworski, *Genetics* **160**, 1179 (2002).
38. L. L. Cavalli-Sforza, P. Menozzi, A. Piazza, *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, NJ, 1994).
39. J. M. Akey, G. Zhang, K. Zhang, L. Jin, M. D. Shriver, *Genome Res.* **12**, 1805 (2002).
40. J. Hastbacka *et al.*, *Cell* **78**, 1073 (1994).
41. R. C. Lewontin, J. Krakauer, *Genetics* **74**, 175 (1973).
42. C. Toomajian, R. S. Ajioka, L. B. Jorde, J. P. Kushner, M. Kreitman, *Genetics* **165**, 287 (2003).
43. P. C. Sabeti *et al.*, *Nature* **419**, 832 (2002).
44. Y. Kim, R. Nielsen, *Genetics* **167**, 1513 (2004).
45. N. A. Hanchard *et al.*, *Am. J. Hum. Genet.* **78**, 153 (2006).
46. E. T. Wang, G. Kodama, P. Baldi, R. K. Moyzis, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 135 (2006).
47. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, *PLoS Biol.* **4**, e72 (2006).
48. C. L. Yauk, P. R. Bois, A. J. Jeffreys, *EMBO J.* **22**, 1389 (2003).
49. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
50. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
51. R. H. Waterston *et al.*, *Nature* **420**, 520 (2002).
52. K. Lindblad-Toh *et al.*, *Nature* **438**, 803 (2005).
53. D. Altshuler *et al.*, *Nature* **407**, 513 (2000).
54. International HapMap Consortium, *Nature* **437**, 1299 (2005).
55. D. A. Hinds *et al.*, *Science* **307**, 1072 (2005).
56. A. G. Clark *et al.*, *Science* **302**, 1960 (2003).
57. S. F. Schaffner, *Nat. Rev. Genet.* **5**, 43 (2004).
58. C. S. Carlson *et al.*, *Genome Res.* **15**, 1553 (2005).
59. A. G. Clark, M. J. Hubisz, C. D. Bustamante, S. H. Williamson, R. Nielsen, *Genome Res.* **15**, 1496 (2005).
60. G. McVean, C. C. Spencer, R. Chaix, *PLoS Genet.* **1**, e54 (2005).
61. S. A. Tishkoff *et al.*, *Science* **293**, 455 (2001).
62. K. M. Teshima, G. Coop, M. Przeworski, *Genome Res.,* 10 May 2006, in advance of print.
63. N. Mekel-Bobrov *et al.*, *Science* **309**, 1720 (2005).
64. P. C. Sabeti *et al.*, *PLoS Biol.* **3**, e378 (2005).
65. V. M. Ingram, *Biochim. Biophys. Acta* **36**, 402 (1959).
66. L. Pauling *et al.*, *Science* **110**, 543 (1949).
67. T. J. Aitman *et al.*, *Nature* **405**, 1015 (2000).
68. A. Pain *et al.*, *Lancet* **357**, 1502 (2001).
69. R. Arya *et al.*, *Am. J. Hum. Genet.* **74**, 272 (2004).
70. S. Stone *et al.*, *Am. J. Hum. Genet.* **70**, 1459 (2002).
71. R. L. Lamason *et al.*, *Science* **310**, 1782 (2005).
72. J. Zhang, D. M. Webb, O. Podlaha, *Genetics* **162**, 1825 (2002).
73. W. Enard *et al.*, *Nature* **418**, 869 (2002).

74. P.C.S. is funded by the Damon Runyon Cancer Fellowship and the L'Oreal for Women in Science Award. We thank C. Bustamante, R. Nielsen, J. Akey, Y. Gilad, and C. Carlson for providing information from their previous publications. We also thank F. Steele, D. Reich, D. Hartl, R. Neilsen, D. Richter, C. Langley, M. Przeworski, A. Clark, J. Fay, S. Myers, T. Farhadian, T. Herrington, A. Foster, P. Sabeti, and four anonymous referees for careful review of our manuscript.

Supporting Online Material

www.sciencemag.org/cgi/content/full/312/5780/1614/DC1
Figs. S1 and S2
Tables S1 to S4

10.1126/science.1124309