SIGNATURES OF NATURAL SELECTION IN THE HUMAN GENOME

Michael Bamshad[*‡] *and Stephen P. Wooding*[*]

During their dispersal from Africa, our ancestors were exposed to new environments and diseases. Those who were better adapted to local conditions passed on their genes, including those conferring these benefits, with greater frequency. This process of natural selection left signatures in our genome that can be used to identify genes that might underlie variation in disease resistance or drug metabolism. These signatures are, however, confounded by population history and by variation in local recombination rates. Although this complexity makes finding adaptive polymorphisms a challenge, recent discoveries are instructing us how and where to look for the signatures of selection.

FITNESS
The ability of an individual to reproduce his or her genetic makeup, which is not always equivalent to individual reproductive success.

Humans differ from each other in many ways, ranging from their physical appearance, behaviour and susceptibility to disease, to their likelihood of experiencing an adverse drug reaction. Although part of this phenotypic variability results from differences in environmental exposures or chance, it is clear that gene variants are also responsible. The identification of these variants could lead to insights into how genes predispose individuals to disease, and might, therefore, inform the development of improved therapeutic and disease-prevention strategies.

One way to find functionally important variants is to identify genes that have been acted on by natural selection. Variants that increase the FITNESS of an individual in its environment might increase in frequency as a result of positive selection (FIG. 1), whereas moderately to severely deleterious gene variants tend to be eliminated by negative (or purifying) selection — a force to which all genes are probably subject, to maintain function. Looking for evidence of positive selection is an attractive strategy for finding functional variants because the dispersal of early humans from Africa to Europe, Asia and the Americas — each with different climates, pathogens and sources of food — varied the selective pressures that challenged human populations. The effects of selection could have been accentuated by the marked changes in population size, population density and cultural conditions that accompanied the introduction of agriculture at the beginning of the Neolithic period ~10,000 years ago[1,2]. Today, the functional consequences of the genetic variants that facilitated survival in ancestral human populations might underlie the phenotypic differences between individuals and groups. So, the analysis of genetic variation in populations has become central to understanding the function of genes.

In this review, we highlight some of the types of natural selection and their effects on the patterns of DNA variation in the human genome. We explain the relative strengths and weaknesses of the strategies that can be used to detect the signatures of natural selection at individual loci. These strategies are illustrated by their application to empirical data from the gene variants that are associated with differences in disease susceptibility. We also outline the methods proposed to scan the genome for evidence of selection. Finally, we discuss the problems that are associated with identifying signatures of selection and with making inferences about the nature of the selective process. There are several philosophical issues (for example, defining the units of selection), theoretical concepts and empirical studies in other species, which are beyond the scope of this review. For further information, the interested reader is referred to several excellent resources[3–6].

[*]Department of Human Genetics, University of Utah, Salt Lake City, Utah 84112, USA.
[‡]Department of Pediatrics, University of Utah, Salt Lake City, Utah 84112, USA.
e-mails: mike@genetics.utah.edu; swooding@genetics.utah.edu
doi:10.1038/nrg999

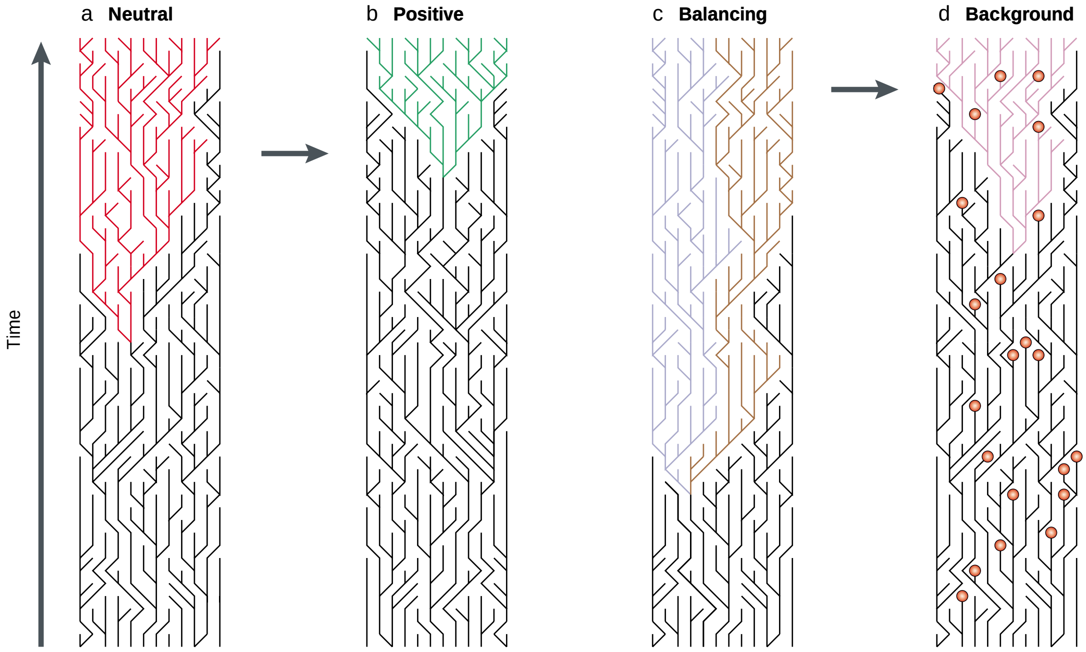


Figure 1 | **Effects of natural selection on gene genealogies and allele frequencies.** Each panel (**a**–**d**) represents the complete genealogy for a population of 12 haploid individuals. Each line traces the ancestry of a lineage, and coloured lines trace all descendants who have inherited an allele that is either neutral (**a**) or affected by natural selection (**b**–**d**) back to their common ancestor (that is, the coalescence of the genealogy). **a** | The genealogy of a neutral allele (red) as it drifts to FIXATION. **b** | The genealogy of an allele (green) that is driven to fixation more quickly after the onset of positive selection (arrow) compared with expectations under a neutral model. Note that the genealogy has a more recent coalescence. **c** | The genealogy of two alleles (blue and gold) under BALANCING SELECTION, which are driven neither to fixation nor to extinction. As a result, the genealogy of the two alleles has an older coalescence. **d** | The genealogy of an allele (purple) that drifts to fixation under the influence of BACKGROUND SELECTION. Each circle represents the elimination of a deleterious (that is, lethal) mutation by background selection. The coalescence of the lineage is more recent than expected under a neutral model because a linked deleterious mutation caused the extinction of one lineage (arrow) more quickly than would be predicted.

FIXATION
The increase in the frequency of a genetic variant in a population to 100%.

BALANCING SELECTION
A selection regime that results in the maintenance of two or more alleles at a single locus in a population.

BACKGROUND SELECTION
The elimination of neutral polymorphisms as a result of the negative selection of deleterious mutations at linked sites.

POLYMORPHISM
The contemporary definition refers to any site in the DNA sequence that is present in the population in more than one state. By contrast, the traditional definition referred to an allele with a population frequency >1% and <99%.

GENETIC DRIFT
The random fluctuations of allele frequencies over time due to chance alone.

Genetic variation: the raw material of selection

Natural selection can act in a population only if mutation has generated heritable genetic differences (or POLYMORPHISMS) among individuals in the population. Otherwise, the differences in fitness between individuals could not be transmitted from one generation to the next. In humans, it has been estimated that ~4 new amino-acid-altering mutations arise per diploid genome per generation[7]; these mutations can be broadly categorized as advantageous (that is, adaptive), deleterious or neutral (that is, exerting no effects on the fitness of an individual). The fixation of adaptive mutations in a population is an important theme underlying Darwin's theory on the origin of species by natural selection. For the past few decades, however, it has been widely believed that most genetic variation — both polymorphisms in species and divergence between species — is neutral and that polymorphisms are eliminated or fixed in populations as a consequence of the stochastic effects of GENETIC DRIFT. The logic underlying this supposition was outlined by Motoo Kimura and became known as the neutral theory of molecular evolution[8–11] (BOX 1).

The statistics that are used to summarize polymorphism data, and to test for the effects of natural selection, compare DNA or amino-acid variation in populations or species and/or the degree of divergence between them (BOX 2; TABLE 1; see REF. 12 for a review). The power of these tests is typically determined by carrying out simulations under a restricted range of demographic models and parameters to estimate the critical values that support rejection of the neutral model[13–15]. To this end, an understanding of population history is crucial for identifying the genes that are subject to selection.

The confounding effects of population history

Interest in characterizing the patterns of genetic variation within and among human populations has grown over the past few years (see REFS 16,17 for a review). However, until recently, studies have been hampered by the relatively small number of polymorphic loci that were typed in each individual and by the restricted sampling of human populations. In the past few years, many publications have reported results on the basis of extensive surveys of variation of the mitochondrial genome[18], the Y chromosome[19] and various autosomal regions, using microsatellite or single nucleotide polymorphism (SNP) markers[20]. These data have allowed broad inferences to be made

Box 1 | The neutral theory of molecular evolution

Before the late 1960s, many evolutionary biologists assumed that most of the polymorphisms in a population were maintained by balancing selection. However, because the maintenance of balanced polymorphisms was predicted to impose a large GENETIC LOAD, most genes were thought to be monomorphic. However, perspectives began to change as the proliferation of protein sequencing and electrophoresis led to the discovery of extensive amino-acid polymorphisms both within, and between, species. In a series of papers published during the 1960s and 1970s, Motoo Kimura and others suggested that the patterns of protein polymorphism seen in nature were more compatible with the hypothesis that most polymorphisms and fixed differences between species are selectively neutral[8,9]. This proposition was called the neutral theory of molecular evolution, or the neutral theory.

The development of the neutral theory was motivated by two principal observations. First, between-species comparisons of amino-acid substitution rates showed that they were regular, or clock-like: although clock-like rates would be expected if amino-acid substitutions occurred stochastically, they would not be expected if natural selection were pervasive. Under the pressure of natural selection, amino-acid substitutions would be expected to occur irregularly, reflecting the unpredictability of environmental change. Second, when the amino-acid substitution rates inferred from comparisons between species were taken into account, the levels of diversity within species were found to be roughly proportional to the EFFECTIVE POPULATION SIZE. Such a pattern would not be expected if natural selection were acting to balance new variants. Taken together, these patterns were interpreted as evidence that genetic drift, rather than natural selection, was responsible for maintaining most polymorphisms.

Kimura emphasized that the neutral theory is not incompatible with the idea of an important role for natural selection in shaping human genetic variation. Strong negative selection, for example, which removes variants from a population, could affect most new mutations but still have little effect on the levels of polymorphism seen. Some positive selection, which would sweep mutations to fixation, could also exist without jeopardizing the conclusion that most of the fixed differences between genes are neutral. The remaining polymorphisms, Kimura argued, represent a mixture of selectively neutral alleles and mildly deleterious alleles that have not yet been removed by natural selection.

The emphasis of the neutral theory on the importance of genetic drift changed the focus of population genetic analysis. It transformed commonly held views on the role of natural selection in maintaining variation, introducing a more sophisticated outlook on the balance of selection and drift, which persists in present evolutionary theory.

about the population history of humans, such as the degree of population subdivision and changes in population size[21].

Most of these studies indicate that the human population has not maintained a constant size, having increased from tens of thousands of individuals to more than 6 billion during the past 100,000 years[22,23]. Furthermore, the human population also shows substantial POPULATION STRUCTURE[24,25]. These findings are relevant because population growth and subdivision can both cause departures from the neutral model that are indistinguishable from those caused by natural selection. For example, population genetics theory predicts that the proportion of variants with a low frequency in a population will increase in an expanding population, because in such a population new mutations are lost at a lower rate[26]. The excess of low-frequency variants seen in humans for many types of genetic marker that are presumed to be neutral has been interpreted as evidence of the rapid expansion in human population size[23,27]. However, positive selection can produce a similar excess of low-frequency variants[28].

It is also possible for a departure from the neutral model at any specific locus to be caused by a combination of both population history and selection. An important distinction is that demographic processes similarly affect all loci, whereas the effects of selection are restricted to specific loci. Therefore, one way that the confounding effects of population history can be treated empirically is by comparing the pattern of variation at a candidate locus with the genome-wide pattern estimated from a set of neutral markers that have been typed in the same individuals or populations[29,30]. Empirical distributions of the summary statistics that are sensitive to selection and population history (for example, Tajima's *D*, see BOX 2), estimated from hundreds of coding and non-coding regions, are also becoming available[31]. These distributions can be used to compare a candidate locus to other regions of the genome to determine whether it has a pattern of variation that is significantly different. However, the comparison of results across studies is often difficult because the distributions are sensitive to the populations studied and the sampling strategies used, and these often vary[32].

The impact of positive selection

In contrast to demographic processes, which affect the entire genome, natural selection affects specific functionally important sites in the genome. In addition, depending on the local rate of recombination, whenever selection acts on a mutation it will affect linked sites as well, leaving its signature in the adjacent chromosomal region. This signature is manifest as a reduction in variation at linked sites for two types of selection. The first — background selection — removes deleterious mutations and eliminates variation at linked sites[33]. The strength of this effect will vary with the recombination rate, the magnitude of selection and the mutation rate[34]. The second — genetic hitchhiking — predicts that if a mutation increases in frequency in a population as a result of positive selection, linked neutral variation will be dragged along with it[35]. As a consequence, variation that is not

GENETIC LOAD
The proportion of a population's maximum fitness that is lost as a result of selection against the deleterious genotypes it contains.

EFFECTIVE POPULATION SIZE
The size of the ideal population in which the effects of random drift would be the same as those seen in the actual population.

POPULATION STRUCTURE
A departure from random mating as a consequence of factors such as inbreeding, overlapping generations, finite population size and geographical subdivision.

linked to the adaptive mutation is eliminated, resulting in a SELECTIVE SWEEP (FIG. 2). Therefore, models predict that genetic hitchhiking will cause a greater overall reduction in genetic diversity, and that the effect will be more pronounced in regions of lower recombination. Both types of selection will result in an overall positive correlation between genetic diversity and recombination rate if the strength and frequency of positive and/or background selection are sufficiently high throughout the genome[36,37].

Empirical data from several plant and animal species, including mice[38] and fruitflies[37] — the best-studied species so far — are consistent with the predictions of recurrent selective sweeps. In humans, it seems that about half of the VARIANCE in nucleotide diversity might be explained by the local recombination rate[39], although more comprehensive studies are needed. These results indicate that selection might have been more important in shaping patterns of variation in the genome than was previously anticipated, although the relative importance of background selection and genetic hitchhiking remains unknown. Indeed, it is possible that both processes contribute to the patterns of variation[40,41].

Most of what is known about the impact of adaptive evolution on the human genome comes from studies of the patterns of genetic differences between humans and other species. In general, the genetic variation in the coding region of a gene is compared between humans and one or more other species to determine whether the rate of amino-acid substitution is higher or lower than expected (BOX 2; see REFS 42,43 for review). The current wealth of DNA sequence data is making this approach increasingly popular, and an emphasis has been placed on comparing loci between humans and chimpanzees for evidence of positive selection[44,45]. The identification of such loci might provide insights into the nature of changes, such as the origin of speech, which led to the evolution of modern humans[46] (see REF. 47 for a review).

In contrast to the investigation of selection among species, relatively few studies of local positive selection in humans have been carried out, even though the loci that show differential selection across populations might be determinants of disease resistance. Such studies might be useful if common diseases, such as diabetes, obesity and atherosclerotic heart disease, are caused by genetic variants that were positively selected in ancestral environments but are detrimental at present (for example, the so-called 'thrifty' genotypes)[48]. Studies of local positive selection can be designed to screen the entire genome for regions that are affected by selection (see below), or they can focus on testing specific candidate genes. The testing of specific candidates is limited by our lack of knowledge about the most suitable genes for

Box 2 | Measures of genetic variation

Several descriptive statistics are commonly used to summarize polymorphism data in a sample of DNA sequences. For example, θ_W describes the proportion of segregating sites in a sample (corrected for the size of the sample)[112], and nucleotide sequence diversity (π) describes the mean number of differences per site between two sequences chosen at random from a sample of sequences[113]. The average π in humans (~7.5×10^{-4}) is relatively low[114,115], although π can vary by more than an order of magnitude among genomic regions[116]. Each of these statistics is an estimator of the population mutation rate, $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the neutral mutation rate per generation. Therefore, π can also be used to estimate the N_e of humans, which, on the basis of diverse genetic data, is ~10,000 (REFS 27,117). This is smaller than our census size and indicates that probably only those polymorphisms that had substantial effects on fitness were likely to have overcome the effects of genetic drift.

Departure from a STANDARD NEUTRAL MODEL can be assessed using several test statistics that use comparisons of estimators of θ (TABLE 1). One frequently used test statistic, Tajima's D, compares the estimates of θ_W and π from a single, non-recombining region of the genome[118]. The difference between θ_W and π is expected to be zero under the neutral model, and so a non-zero D is a sign of a departure from the neutral model due to a relative excess or deficiency of polymorphisms of various frequencies. For example, background selection will eliminate polymorphisms linked to deleterious alleles, allowing them to reach only low frequencies, whereas positive selection will tend to eliminate older, high-frequency alleles, and newer, low-frequency alleles will hitchhike with the target of selection (FIG. 1). As a consequence, positive or background selection can to lead to an excess of polymorphisms at low frequencies.

Other estimators of θ vary in their sensitivity to polymorphisms of different frequencies. For example, alleles that have been targets of recent positive selection might exist at a relatively high frequency. A recently developed estimator, θ_H, is more sensitive to such alleles, and therefore the test statistic based on it, H, might be more powerful for detecting recent positive selection[119,120]. It is worth noting that all of these test statistics use the allelic distribution and/or the level of allele variability, both of which are dependent on the genealogy of a locus. Therefore, they rely on assumptions about the demographic histories of the populations in which the samples were ascertained. As a consequence, the interpretation of the results of these test statistics can be challenging, and they rarely provide unambiguous evidence of selection.

By contrast, tests for selection that are independent of the genealogy of a locus have provided clear evidence for selection. These tests use comparisons of the variability and divergence of different types of mutation at a locus, such as the rate of non-synonymous mutations (d_N) versus the rate of synonymous mutations (d_S) (TABLE 1). In this example, a significantly increased rate ratio (d_N/d_S) has indicated that human olfactory receptor genes[121], human leukocyte antigen (HLA) loci[101] and breast cancer 1, early onset (*BRCA1*)[122] have all been subject to positive selection. These tests are generally conservative, because the substitution rates are averaged across all the amino-acid sites tested. Alternative strategies include testing functionally important domains of a protein individually, testing single amino-acid sites[123] or separately testing the lineages in a phylogenetic tree.

SELECTIVE SWEEP
The process by which positive selection for a mutation eliminates neutral variation at linked sites.

STANDARD NEUTRAL MODEL
A hypothetical panmictic (randomly mating) population of constant size in which genetic variation is neutral and follows a model (the 'infinite sites model') in which each new mutation occurs at a site that has not previously mutated.

VARIANCE
A statistic that quantifies the dispersion of data about the mean.

Table 1 | Commonly used tests of neutrality

Test	Compares	References
Tests based on allelic distribution and/or level of variability		
Tajima's *D*	The number of nucleotide polymorphisms with the mean pairwise difference between sequences	118
Fu and Li's *D*, *D**	The number of derived nucleotide variants observed only once in a sample with the total number of derived nucleotide variants	129
Fu and Li's *F*, *F**	The number of derived nucleotide variants observed only once in a sample with the mean pairwise difference between sequences	129
Fay and Wu's *H*	The number of derived nucleotide variants at low and high frequencies with the number of variants at intermediate frequencies	119
Tests based on comparisons of divergence and/or variability between different classes of mutation		
d_N/d_S, K_a/K_s	The ratios of non-synonymous and synonmyous nucleotide substitutions in protein coding regions	130,131
HKA	The degree of polymorphism within and between species at two or more loci	132
MK	The ratios of synonymous and non-synonymous nucleotide substitutions in and between species	128

HKA, Hudson–Kreitman–Aguade; MK, McDonald–Kreitman.

analysis. However, our increasing knowledge of the biological mechanisms that underlie phenotypic traits has led to the accumulation of circumstantial evidence, indicating that certain loci might have been the targets of selection. This has led to notable recent successes in finding signatures of local positive selection in human populations (for examples, see REFS 49–66). In turn, a review of the properties of these loci might refine our strategies for finding the signatures of selection.

Selection and the site frequency spectrum

Among the most promising candidates to test for a signature of local positive selection are genes that encode the proteins that are involved in drug transport and metabolism. Many of these genes show marked differences in allele frequencies between populations, and gene variants have been associated with variable responses to foods and drugs[67]. An example is cytochrome P450 1A2 (*CYP1A2*), which encodes an enzyme that oxidizes carcinogenic arylamines, acetaminophen and several widely prescribed anti-psychotic drugs[68]. The hepatic mRNA expression of *CYP1A2* varies by as much as 15-fold among individuals and, accordingly, variants of *CYP1A2* might underlie interindividual variation in cancer susceptibility, as well as variation in response to toxins or medications[69]. The regulatory region of this gene was a logical candidate to screen for a signature of selection.

Analysis of 3.7 kb of a non-coding DNA sequence 5′ of *CYP1A2* in Africans, East Asians and Europeans showed a pattern of SNP frequencies that indicated recent positive selection[49]. To illustrate this, we focus on the site frequency spectrum. In most spectra, each MINOR ALLELE is inferred to be the derived (that is, the younger or more recent) allele, because derived alleles typically exist at lower frequencies than ancestral ones. Moreover, if positive selection was recent, linked derived variants might not be fixed, and so exist at higher frequencies than expected. This can be inferred by comparing the human sequence to an OUTGROUP. Chimpanzees and gorillas — the two species with which humans most recently shared a common ancestor — are commonly used as outgroups when examining human frequency spectra. Under a neutral model, this distribution has a characteristic shape, which can be skewed by natural selection (FIG. 2). If we know which allele is ancestral and which allele is derived at each site, we can make inferences about the type of selection that has affected a locus. Although positive selection is expected to skew the spectrum towards an excess of low-frequency alleles, such a skew is not generally expected in regions that are subject to background selection[28].

In the case of four of the SNPs in *CYP1A2*, the minor allele in humans was the fixed allele in chimpanzees and gorillas, so the common SNPs in humans were inferred to be the derived state (even though they each had a frequency >90%). By comparison with the expected fraction of variants at each frequency estimated under the neutral model, the site frequency spectrum for *CYP1A2* showed an excess of both low- and high-frequency alleles. This indicates that *CYP1A2* might have been influenced by both positive selection and recent population growth, although the relative strengths of each are unclear.

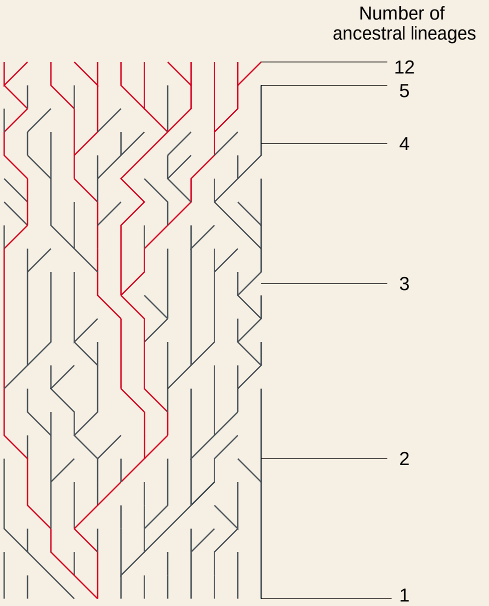
Gene genealogies and selection

Natural selection can also affect the genealogy of alleles, the relationships of which can be depicted in a tree or network (FIG. 2a). The parameters of the genealogical process that have given rise to a tree can be estimated using coalescence theory (BOX 3). Departures from the neutral model are thought to reflect the effect that population history and natural selection have had on the shape of the genealogy (see REF. 70 for a review). For example, positive selection that sweeps an adaptive variant to fixation can distort the genealogy to create a star-like pattern[71], which is a sign of an excess of low-frequency variants that are connected to a common ancestor by branches with similar, often short, lengths. Coalescence theory can be used to test whether natural selection has produced such a genealogy.

MINOR ALLELE
The less frequent of two alleles at a locus.

OUTGROUP
A closely related species that is used for comparison, for example, to infer the ancestral versus the derived state of a polymorphism.

Box 3 | The coalescent process



If a sample of genes is collected, and the ancestry of those genes is traced backwards in time, common ancestors will be encountered. As these common ancestors are encountered, the number of distinct ancestral lineages decreases from *n* (the original sample size) to *n* – 1, *n* – 2 and so on, eventually ending in one lineage, which is the common ancestor of the entire sample; this is called the coalescent process. The diagram shows the genealogy of a population of 12 haploid individuals. The red lines trace the ancestry of their 12 lineages back to the most recent common ancestor.

The coalescent process is a point of key interest in efforts to detect natural selection in human genes. Because natural selection can affect the rate at which common ancestors are encountered (for example, compare the time at which the common ancestor is encountered in each panel of FIG. 1), the shape of gene genealogies can inform us about the selective processes that have been involved. Selective sweeps, for example, result in 'shallow' gene genealogies with a few, rare mutations. By contrast, balancing natural selection results in 'deep' gene genealogies in which many mutations are found at intermediate frequencies.

LIKELIHOOD ANALYSIS
A statistical method that calculates the probability of the observed data under varying hypotheses, in order to estimate model parameters that best explain the observed data and determine the relative strengths of alternative hypotheses.

PHYLOGENETICS
Reconstruction of the evolutionary relationships (that is, the phylogeny) of a group of taxa, such as species.

LINKAGE DISEQUILIBRIUM
(LD). The non-random association of alleles in haplotypes.

The use of LIKELIHOOD ANALYSIS, based on the coalescent, to test hypotheses of selection in humans is becoming more common[54,55]. In the meantime, most human polymorphism data are still analysed using summary statistics based on coalescence theory, as well as methods adopted from PHYLOGENETICS. In the latter case, once polymorphism data are available, the genealogical tree of the locus is estimated and compared with inferences made using other methods. This strategy has practical value, particularly if questions about the haplotype structure of a locus are being explored. In a tree of *CYP1A2* haplotypes, for example, the common haplotype that shares the four high-frequency-derived variants (asterisk in FIG. 3a) is connected to other haplotypes by short branches, in a star-like pattern (FIG. 3a). This pattern is reminiscent of a tree distorted by positive selection (FIG. 2). This tree can also be used to facilitate the cloning of functional variants, because haplotypes sharing functional variants that are associated with the same phenotypic trait generally share a common internal branch in the network. The sequences of these haplotypes can be compared with one another to identify all of the mutations shared exclusively among them. Each of these mutations can then be tested, alone or in combination, to determine whether they are of functional significance[72], thereby reducing the number of mutations that need to be screened for functional effects. Although the topology of these networks can help generate hypotheses about the history of a locus, it must be noted that they should not be used, alone, to make inferences of selection.

Linkage disequilibrium and selection

Most of the strategies that are used to detect whether natural selection has affected a specific allele measure the departure of the frequency of the allele from expectations under a neutral model, with the assumption that there has been no local recombination. Indeed, recombination has often been considered a nuisance in this context. It is, therefore, ironic that the block-like nature of LINKAGE DISEQUILIBRIUM (LD) across the human genome offers a new way to detect a signature of recent positive selection[50,73]. The logic underlying this strategy is straightforward. When a mutation arises, it does so on an existing background haplotype characterized by complete LD between the new mutation and the linked polymorphisms (FIG. 4). Over time, new mutations and recombination reduce the size of this haplotype block such that, on average, older and relatively common mutations will be found on smaller haplotype blocks (that is, there is only short-range LD between the mutation and linked polymorphisms). Younger, low-frequency mutations might be associated with either small or large haplotype blocks. A signature of positive selection is indicated by an allele with unusually long-range LD and high population frequency. The formal implementation of this strategy has recently been introduced as the long-range haplotype (LRH) test[50].

Glucose-6-phosphate dehydrogenase (*G6PD*) is one of several examples of genes in which alleles that are common in sub-Saharan Africans have been associated with resistance to infection[74–76]. G6PD is the only enzyme in red blood cells that can recycle nicotinamide adenine dinucleotide phosphate (NADP), which is needed to prevent oxidative damage to the cell. Hundreds of *G6PD* variants have been identified, although most of them are relatively uncommon[77]. *G6PD-202A* reduces enzyme activity to ~10% of its normal value, and is found at frequencies as high as 25% in sub-Saharan African populations[78]. This variant is advantageous in certain environments, because it reduces the risk of malarial disease by 40–60% in heterozygous females and hemizygous males[79].

Haplotypes that bear *G6PD-202A* have significantly less microsatellite variability than predicted by a coalescence model. This low level of variability, in conjunction with the high frequency of *G6PD-202A*, indicates that *G6PD-202A* might have risen in frequency so rapidly that there was no time to accumulate new variation in nearby polymorphisms[51,52]. Long-range LD around haplotypes that bear *G6PD-202A* extends for hundreds of kilobases (REF. 50), which is significantly longer than the LD of other *G6PD* variants of comparable frequency. Both of these patterns differ from the pattern of haplotype variation and LD seen at other loci in the same populations, providing molecular evidence that *G6PD-202A* has been a target of recent positive selection. The date of origin of *G6PD-202A* estimated from these data ranges

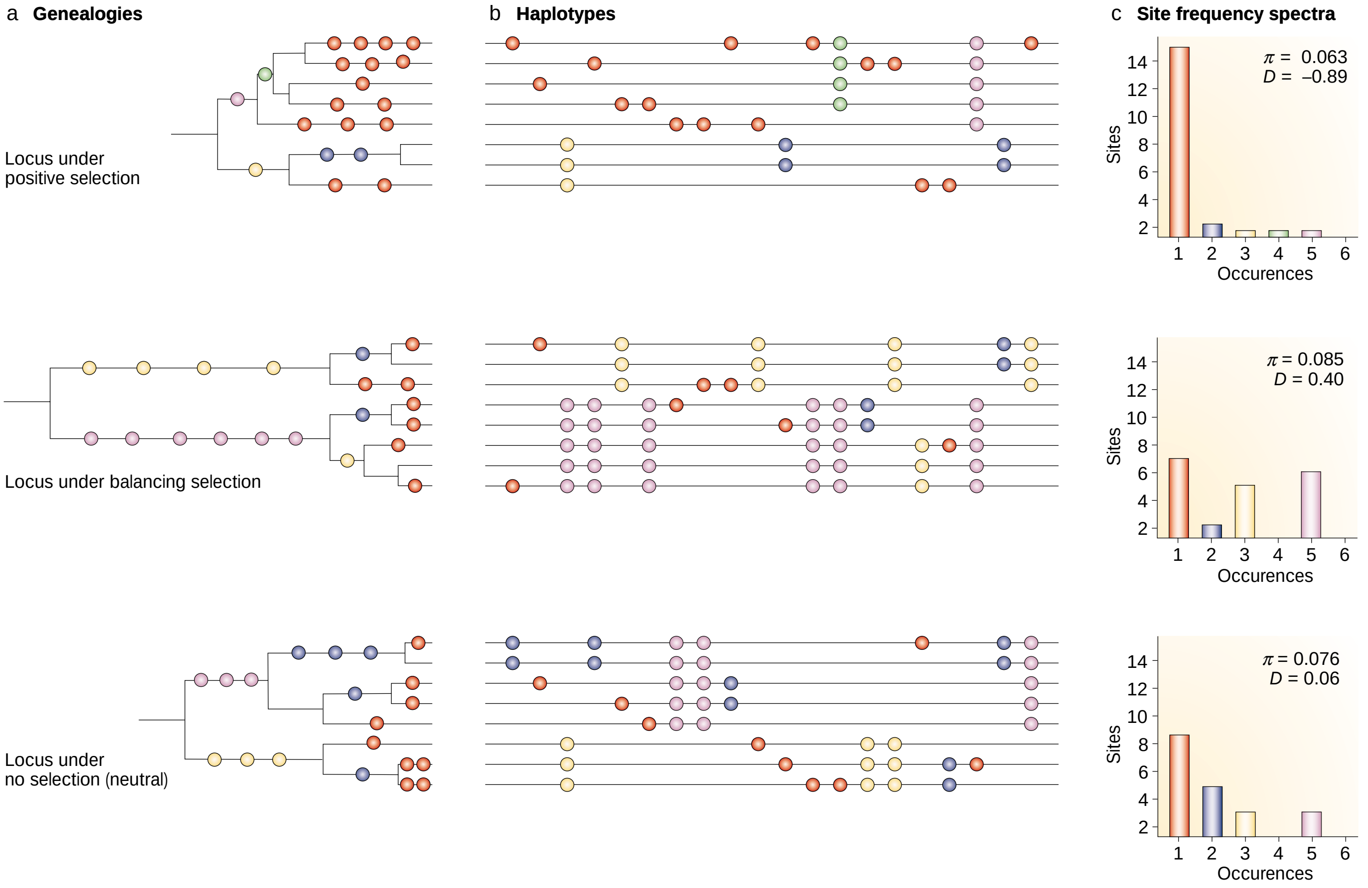


Figure 2 | **The effects of selection on the distribution of genetic variation. a** | The genealogies of three genes that are typical of loci under positive selection (top), balancing selection (middle) and no selection (neutral) (bottom) are depicted. Each circle represents a mutation, and the colour shows the final frequency of each mutation in the sampled HAPLOTYPES (**b**) and the SITE FREQUENCY SPECTRA (**c**). For each gene, the number of segregating sites is 20. **b** | Each haplotype contains mutations that have accumulated on each lineage in the gene genealogy, assuming no recombination. **c** | The site frequency spectrum of each gene. Positive selection (top) can result in a lower level of sequence diversity (*π*), an excess of low-frequency variants (red) and, consequently, a negative value of Tajima's *D* (BOX 2). Balancing selection (middle) can result in a higher level of sequence diversity (*π*), an excess of intermediate-frequency variants (purple) and, consequently, a positive value of Tajima's *D*. The diversity estimate and site frequency spectrum of a neutral locus (bottom) can be used for comparison.

from ~2,500 to 6,500 years ago[50,51]. Interestingly, these dates are in agreement with archaeological data that indicate that malaria might have had a substantial impact on sub-Saharan Africans only in the past 10,000 years, concordant with a recent expansion in an already large effective population of *Plasmodium falciparum*[80] and with the diversification of *Anopheles gambiae*, a mosquito vector of malaria[81].

Evidence of local positive selection has also been found outside Africa. Idiopathic haemochromatosis is an autosomal-recessive disorder caused by mutations in the *HFE* gene, which is characterized by excessive intestinal iron uptake. Iron accumulates in the heart, liver, pancreas, joints and skin; this can lead to hepatic cirrhosis, diabetes, arthropathy and heart failure. Because iron accumulates slowly in affected individuals, disease symptoms are not usually seen until the fifth or sixth decade in males, if at all. Among females, the age of onset is even later because of iron loss due to menstruation, pregnancy and lactation.

A mutation that produces a Cys–Tyr substitution at position 282 of the mature polypeptide (C282Y) accounts for ~85% of all haemochromatosis mutations[82]. This mutation is almost exclusive to individuals of European descent[83], in whom it has a frequency of 5–10%. An assessment of LD between the C282Y mutation and linked polymorphisms showed a substantial degree of LD, indicating that the mutation probably arose only ~60 generations ago[84,85]. It is intriguing that C282Y has reached a relatively high frequency in such a short period of time. A coalescence-based analysis of the frequency of C282Y and its age (based on the level of the LD between the mutation and the linked polymorphisms) showed that it is improbable that the observed frequency of C282Y could have been achieved by genetic drift alone[53]. Instead, recent positive selection is probably responsible for the high frequency of the mutation in Europeans.

Support for the hypothesis of a selective sweep by C282Y comes from data on iron deficiency among

HAPLOTYPE
The combination of alleles or genetic markers found on a single chromosome of a given individual.

SITE FREQUENCY SPECTRUM
The fraction of polymorphic sites at which a minor or derived allele is present in one copy, two copies and so on.

SINGLE-LINKAGE JOINING ALGORITHM
A simple clustering algorithm that begins with all data points (for example, haplotypes) in separate clusters, and then iteratively joins pairs of similar clusters.

DUFFY BLOOD GROUP
This group is defined by variants in a chemokine receptor that is present on the surface of several types of cell, including red blood cells. This receptor must be present for *Plasmodium vivax* to invade cells and cause malaria.

individuals who are heterozygous for a haemochromatosis mutation. A study of >1,000 American heterozygotes showed that 32% of normal homozygous females of reproductive age had iron deficiency (defined as a serum ferritin level $<12\ \mu g\ l^{-1}$), compared with only 21% of haemochromatosis heterozygotes[86]. Among males over the age of 18 years, the corresponding percentages were 4% and 2%. Iron deficiency, which might have been more common in earlier populations, is associated with an increased risk of preterm delivery and with low birth weight[87]. Therefore, the C282Y allele might have improved the fitness of both male and female heterozygous carriers. Because of the late age of onset of symptoms, an adverse effect in homozygotes might have had less of an effect on fitness. It is worth noting, however, that even mutations having an adverse effect later in life might have weakly deleterious effects early in life, that could alter the frequency distribution of an allele[88]. Alternatively, the 'Grandmother hypothesis' suggests that mutations with late-onset deleterious effects could have strong effects if they impair the ability of their host to care for its descendants, indirectly diminishing the fitness of the host[89].

Evidence of positive selection acting on other genes is slowly beginning to accumulate. Signatures of positive selection have been found in genes that have been used as classical models of selection, including those at the β-globin[54] and the DUFFY BLOOD GROUP loci[56,57]. At least one gene that contributes to the diversification of morphological traits among humans — the melanocortin-1 receptor (*MC1R*) — seems to have been under various selective pressures[58,59]. Evidence for positive selection has also been found in genes that encode the dopamine receptor D4 (DRD4)[60], calpain-10 (REF. 29), factor IX (REF. 61), CD40 ligand[50], dystrophin[62], monoamine oxidase A (MAOA)[63], lactase-phlorizin hydrolase[64] and chemokine (C-C motif) receptor 5 (CCR5)[65,66]. Most of these genes were candidates for study because of our knowledge of their biology. Relatively little is known about most of the 30,000 or so genes in the human genome. Although it is easy to begin developing *ad hoc* stories about how selection might have influenced a candidate, it is prudent to consider each candidate with a degree of scepticism. Many differences between the genes of humans and other species have been affected by selection, but it is less clear how many genes have been affected by local adaptive processes, and much less clear whether these genes are important for understanding human phenotypic variation.

Balancing selection

Natural selection does not always increase or decrease the frequency of a single allele at a locus. Sometimes, selection tends to maintain two or more alleles at a locus

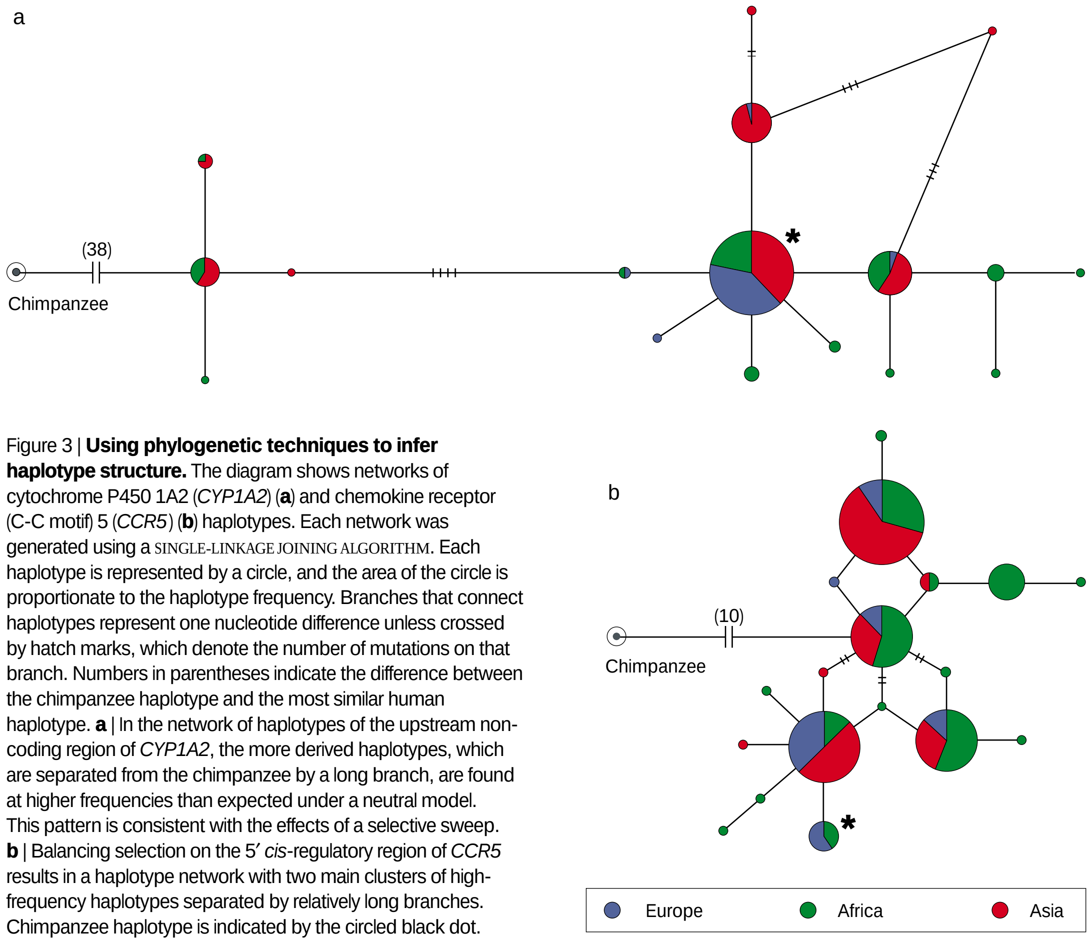


Figure 3 | **Using phylogenetic techniques to infer haplotype structure.** The diagram shows networks of cytochrome P450 1A2 (*CYP1A2*) (**a**) and chemokine receptor (C-C motif) 5 (*CCR5*) (**b**) haplotypes. Each network was generated using a SINGLE-LINKAGE JOINING ALGORITHM. Each haplotype is represented by a circle, and the area of the circle is proportionate to the haplotype frequency. Branches that connect haplotypes represent one nucleotide difference unless crossed by hatch marks, which denote the number of mutations on that branch. Numbers in parentheses indicate the difference between the chimpanzee haplotype and the most similar human haplotype. **a** | In the network of haplotypes of the upstream non-coding region of *CYP1A2*, the more derived haplotypes, which are separated from the chimpanzee by a long branch, are found at higher frequencies than expected under a neutral model. This pattern is consistent with the effects of a selective sweep. **b** | Balancing selection on the 5′ *cis*-regulatory region of *CCR5* results in a haplotype network with two main clusters of high-frequency haplotypes separated by relatively long branches. Chimpanzee haplotype is indicated by the circled black dot.

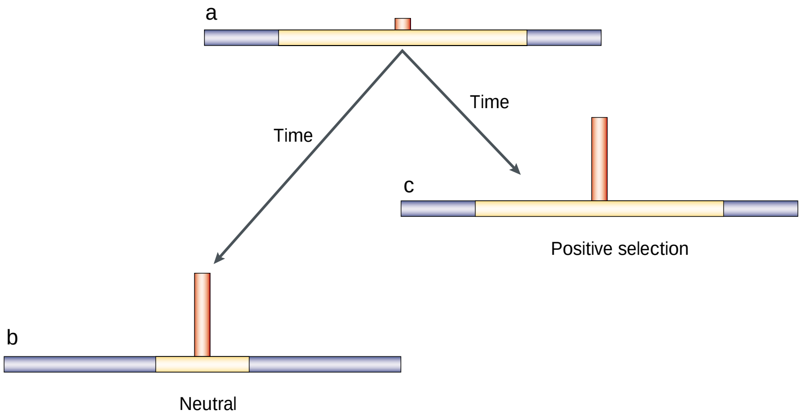


Figure 4 | **Detecting recent positive selection using linkage disequilibrium analysis.** **a** | A new allele (red) exists at a relatively low frequency (indicated by the height of the red bar) on a background haplotype (blue) that is characterized by long-range linkage disequilibrium (LD) (yellow) between the allele and the linked markers. **b** | Over time, the frequency of the allele increases as a result of genetic drift, and local recombination reduces the range of the LD between the allele and the linked markers (that is, it creates short-range LD). **c** | An allele influenced by recent positive selection might increase in frequency faster than local recombination can reduce the range of LD between the allele and the linked markers.

in a population (FIG. 2). This is known as balancing selection because the frequencies of alleles are maintained in a balance, often as a result of a rare allele advantage. Balancing selection can, therefore, maintain an excess of alleles at intermediate frequencies, and variation at linked loci can also accumulate because of genetic hitchhiking[90,91]. In many plants and animals, balancing selection seems to be involved in maintaining diversity at the loci that coordinate recognition between self and non-self[92]. In humans, this has been best studied at loci that are involved with host–pathogen responses, including human leukocyte antigen (HLA) class I and class II genes[93], β-globin[54], *G6PD*[94], glycophorin A[95], interleukin 4 receptor-α[96] and *CCR5* (REF. 30).

CCR5 is a seven-transmembrane G-protein-coupled chemokine receptor that, along with CD4, is required on the surface of a cell for the entry of the human immunodeficiency virus type 1 (HIV-1). The role of CCR5 in the pathogenesis of AIDS was highlighted by the observation that a small fraction of individuals that are resistant to infection by HIV-1 are homozygous for a 32-bp deletion (*CCR5-Δ32*) in its open reading frame (ORF), which eliminates the cell-surface expression of CCR5 (REF. 97). This allele is found at unusually high frequencies only in populations from North-eastern Europe, where it seems to have been a target of local positive selection[65,66]. Nevertheless, most polymorphisms in *CCR5* that are associated with HIV-1 disease progression are in the 5′ *cis*-regulatory region that flanks the ORF. As in the *HLA* genes, genetic diversity in this region is higher than expected, with a site frequency spectrum characterized by an excess of intermediate frequency alleles[30] (FIG. 2).

Loci that are subjected to balancing selection, which favours intermediate-frequency alleles, are expected to show a different pattern of sequence diversity compared with neutral loci[98,99]. Balancing selection increases within-population diversity relative to total diversity, because alleles are kept in more equal frequencies compared with a neutral locus, and an advantageous allele that is introduced to a population by migration will be positively selected. This increases the chances of survival of an advantageous allele compared with a neutral allele. Both of these processes are expected to decrease population differentiation, commonly measured using WRIGHT'S FIXATION INDEX (F_{ST}). The F_{ST} estimate for the 5′ *cis*-regulatory region of *CCR5* is 1.6%, which is nearly tenfold lower than the typical F_{ST} estimates of 10–15% that are found for other regions of the genome among various populations throughout the world[16].

WRIGHT'S FIXATION INDEX (FST). The fraction of the total genetic variation that is distributed among subpopulations in a subdivided population.

Because balancing selection maintains two or more lineages over a longer period of time than expected, the genealogy of the locus is expected to differ from that of a neutral locus[100]. Genealogies are generally characterized by two or more classes of lineages that are separated by relatively long branches (FIGS 1c, 2), a pattern recapitulated in the tree of *CCR5* haplotypes (FIG. 3b). Similar genealogies can be caused by population subdivision, in which haplotypes are restricted to specific populations.

The length of each branch is, however, ultimately dependent on the age of the mutation and the length of time that it has been under balancing selection. If a mutation has arisen relatively recently, or the onset of balancing selection is recent, both of which might be the case for *G6PD*, the branch lengths might be short. Branch lengths for mutations that are subjected to recent positive selection, such as *CCR5-Δ32* in Europeans (see asterisk in FIG. 3b), will also be short. The latter shows that more than one type of selection can affect the pattern of genetic variation at a locus. Whereas local positive selection has recently increased the frequency of *CCR5-Δ32* in Europeans, polymorphisms in the 5′ *cis*-regulatory region of *CCR5* that are associated with disease progression have been maintained by balancing selection. This pattern is similar to that seen at the major histocompatibility complex (MHC) locus, in which allele frequency variation has been affected by both positive and balancing selection[101].

Balancing selection involves rare-allele advantage. Two types of selection that feature rare-allele advantage are negative frequency-dependent selection and generalized overdominance. In negative frequency-dependent selection, the fitness of an allele decreases as it becomes more common. In generalized overdominance, heterozygotes maintain a selective advantage over homozygotes, and, therefore, the rare alleles benefit from their representation in the heterozygotes. This latter type of selection is thought to maintain the high levels of allelic variation seen at the MHC locus[102], an insight derived from functional data showing that MHC heterozygotes can present an expanded spectrum of antigens to T cells compared with that of MHC homozygotes.

Many coding regions in the human genome do not have an excess of low-frequency alleles. This indicates that balancing selection might be more common than is generally perceived[55]. Although this point of view was previously common, it lost support for several reasons. One objection was that a large number of loci maintained by overdominance would

Box 4 | Approaches to scanning the genome for selection

Many analytical approaches have been developed to screen a genome for evidence of selection[124]. In general, for each of several loci, these tests compare the degree of differentiation among sample populations with the overall level of diversity. Under a neutral model, the variation in allele frequencies among populations is determined by genetic drift alone, the impact of which depends only on the demographic history of a population. Therefore, all loci are expected to show the same degree of differentiation. If positive selection has increased the frequency of an allele in one population, but not the other, a higher fraction of variation will be distributed between populations. Positive selection will also tend to reduce the total variability in a population. Both effects of positive selection are expected to increase the level of differentiation and, consequently, the F_{ST} between populations.

The first approach that was developed to screen multiple loci was the Lewontin–Krakauer test, which used the variance of F_{ST} values among loci to identify those with an F_{ST} that deviated more than was expected[125]. This test was criticized for being unreliable because, given certain population histories, it inflated the expected variance. This led to the development of more refined tests that are robust over a wide range of demographic models[126,127]. These tests use comparisons of the relationship of F_{ST} to heterozygosity, estimates of genetic distance between populations or reduced variability at a locus compared with expectations under a neutral model or various demographic models (FIG. 5). For each of these tests, the ability to detect signatures of selection depends on the marker density, the distance between markers and the site under selection, the local recombination rate, the strength of selection and the assumptions made about population history.

An alternative strategy to screen the genome for selection is to use tests that do not depend on assumptions about population history. One example is the McDonald–Kreitman test, in which the d_N/d_S of polymorphisms in species is compared with the d_N/d_S of fixed differences between species in a 2×2 CONTINGENCY TABLE[128]. If polymorphism and divergence are the consequence of only mutation and drift, the ratio of the number of fixed differences to polymorphisms is the same for both non-synonymous and synonymous mutations.

impose an excessive genetic load on a population[103]. However, recent models show that the number of loci that could be maintained by overdominance without a substantial genetic load far exceeds the number of loci in the human genome[104].

Scanning the genome for natural selection

The approaches that are used to detect selection at individual candidate loci are, in principle, adaptable to scanning the entire human genome for signatures of selection — albeit with the same limitations imposed by the effects of population history (BOX 4). In addition, these scans might improve our understanding of the impact of natural selection across the entire genome. Such scans would rely on information from genetic markers that were assayed in a representative set of individuals from selected populations. Although neutral regions would be expected to have a similar pattern of variability and allele frequency distribution among populations, the patterns of variation in regions that are affected by selection might differ (FIG. 5). For example, an excess of rare alleles in a region, or more than expected differentiation among populations at a marker (that is, a high F_{ST}), might be a signal of recent positive selection. A region that is characterized by an excess of intermediate frequency alleles, or by less than the expected differentiation among populations at a marker (that is, a low F_{ST}), might be under balancing selection.

The wealth of nucleotide polymorphism data that has become available during the past few years has provided an exciting opportunity to carry out genome scans for selection. Several scans of the human genome have been undertaken to search for regions under natural selection, and more are underway[31,105–108]. These scans vary in strategy by, for example, using either a battery of microsatellites versus SNPs, or screening anonymous regions of the genome versus coding regions. A recent survey of >5,000 microsatellites typed in a sample of Europeans identified more than 40 regions with extreme skews in the site frequency spectrum[105]. Some of these regions are likely to have been affected by selection. The question of whether these regions contain genes that have been a target of positive selection is, however, complicated by our limited understanding of the impact of selection on microsatellite variability. Additionally, the distance over which hitchhiking is detectable might be too small, depending on the local rate of recombination, compared with the average physical distance between the microsatellites tested and the genes that might have been influenced by selection[109]. This limitation indicates that a more efficient strategy might be to assay markers in or near genes.

A recent example of an approach in which markers in or near genes were analysed, examined ~9,000 SNPs and identified the outliers in the extreme tails of the empirical distributions F_{ST} (FIG. 5), an approach that does not depend on assumptions about population history. SNPs with F_{ST} patterns that indicated that they had been subject to natural selection identified 174 candidate genes[108], including peroxisome proliferative activated receptor-γ (*PPARG*), which has been associated with type II diabetes. This screen could be considered conservative because the SNP data set that was analysed might have been over-represented with common SNPs, which are expected to be shared across populations. Therefore, these SNPs would be expected to have smaller differences in allele frequencies between populations.

In general, the results of screens for signatures of selection have been similar, with a small percentage of loci seeming to deviate substantially from expectations. It remains unclear, however, what proportion of these genes have been targets of positive selection. The next step will be to test the predictions of these screens, perhaps by more direct tests of selection, such as the

2×2 CONTINGENCY TABLE
A 2×2 table that describes the cross-classification of data that are divided into two groups with two categories in each.

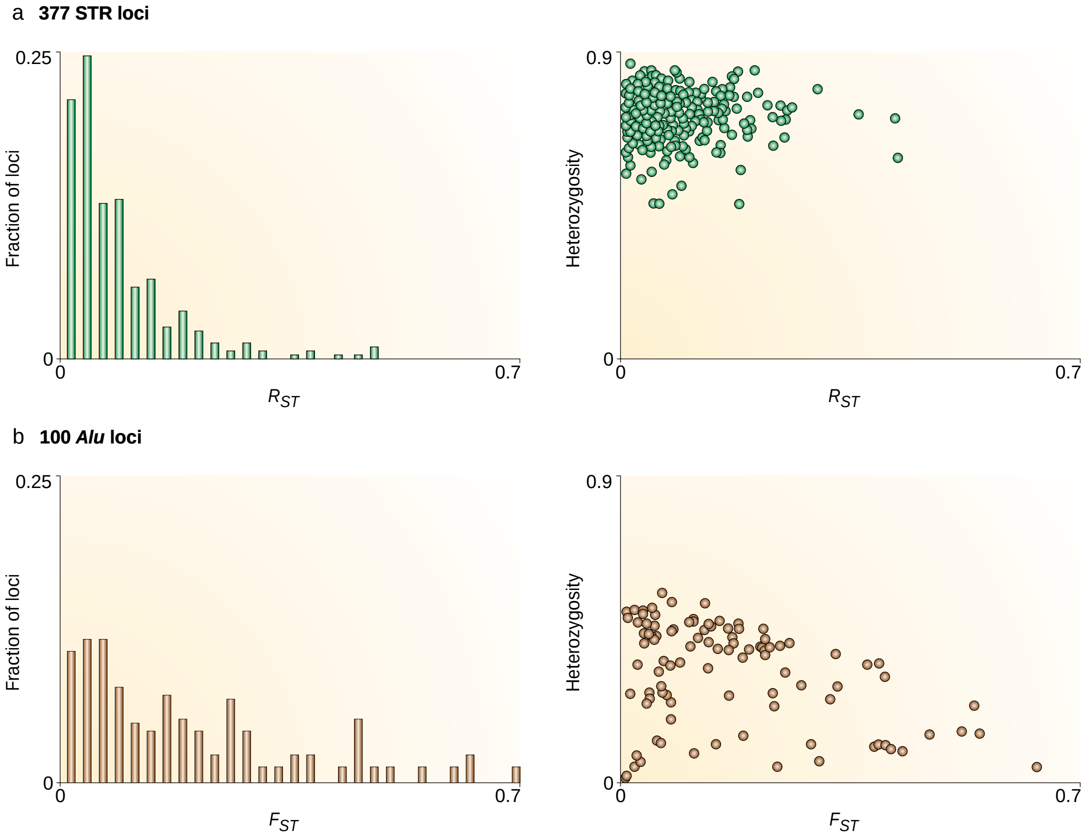


Figure 5 | **Screening the human genome for signatures of natural selection**. Heterozygosity and F_{ST} or R_{ST} values (R_{ST} is an analogue of F_{ST} designed for microsatellites) for two large data sets of markers that are distributed throughout the human genome. **a** | A set of 377 microsatellites (short tandem repeats) typed in 958 individuals from Africa, Asia and Europe[24]. **b** | A set of 100 *Alu* insertion polymorphisms typed in 207 individuals from Africa, Asia and Europe[25]. Left-hand panels: one strategy to find loci that have been subject to natural selection is to identify outliers in the empirical frequency distribution of F_{ST} or R_{ST}. A high F_{ST} indicates that the region might have been subject to local positive selection, whereas a low F_{ST} can be seen in regions under balancing selection (see text for details). Right-hand panels: another strategy to identify regions that are affected by selection is to identify outliers in a plot of F_{ST} or R_{ST} versus heterozygosity. Loci in regions that are far from the origin in each plot, including markers with exceptionally high heterozygosity, exceptionally high F_{ST} (or R_{ST}) or both, are the most obvious candidates for regions that are affected by selection. However, a robust inference depends on comparisons to a null distribution generated under a demographic model that must make assumptions about human population history. The success of both strategies depends on the proximity of each marker to the target of selection and the local recombination rate.

comparison of non-synonymous to synonymous substitutions. Some of the genes that deviate from expectations are members of the same gene family. In some cases, these within-family trends might be due to the clustering of genes with similar function to the same region of the genome, or to the co-evolution of genes that interact with one another. Genes that encode mediators in the same metabolic pathway or developmental programme might show a similar signature of selection. Even after a candidate locus has been shown to be subject to selection, a substantial amount of work will be required to identify the causal variants and to understand their relationship to a human phenotype.

Conclusions

Our inferences of signatures of selection are constrained by an insufficient understanding of population demography and of local rates of recombination. Patterns of nucleotide variability caused by population growth and subdivision can be mixed up with selection. Moreover, the expectations of population genetic models are dependent on assumptions about demographic parameters for which estimates remain ambiguous. Therefore, continued progress towards identifying the genes that are subject to selection will depend on understanding more about the demographic structure of human populations.

Our interpretation of how selection has influenced the human genome is relatively simple, whereas, in fact, the effects of selection will probably be complex. Selection intensity can fluctuate over time, and genetic drift might dominate in some populations, whereas selection might dominate in others. Most population genetic models of positive selection assume that a neutral site is linked to only one functional site on which selection is acting. Whether this is correct depends on the rate at which positive selection sweeps mutations to fixation. If this is frequent, predictions made on the basis of these models might not be accurate.

EPISTATIC
An interaction between non-allelic genes, such that one gene masks, interferes with or enhances the expression of the other gene.

EPISTATIC effects such as synergism or interference between gene variants that lie close to one another might also affect the patterns of selection in the genome[110]. New multi-locus models of selection that are designed to explore these effects seem promising[111].

Our ability to interpret these patterns will improve as we learn more about the signatures of selection at the molecular level and as we improve our ability to link causal variants to phenotypes. The clearest interpretations of the impact of selection are for loci about which we already know a great deal. The increasing enthusiasm for characterizing signatures of natural selection is dependent, in part, on how well these signatures will be able to predict the location of gene variants of biomedical importance with little, if any, *a priori* knowledge of their functional significance. To this end, the study of population variation will continue to be of great interest and relevance to researchers and clinicians. From this process, we will learn more about the evolutionary history of our species, both the shared biology that makes individuals so similar and the small fraction of differences that explain, in part, why one individual dies of malaria, another is allergic to penicillin and another is resistant to AIDS.

1. Klein, R. G. *The Human Career: Human Biological and Cultural Origins* (Univ. of Chicago Press, Chicago, 1999).
2. Klein, J. & Takahata, N. *Where Do We Come From? The Molecular Evidence for Human Descent* (Springer, New York, 2002).
3. Sober, E. *The Nature of Selection: Evolutionary Theory in Philosophical Focus* (MIT Press, Cambridge, Massachusetts, 1993).
4. Li, W. *Molecular Evolution* (Sinauer Associates, Sunderland, Massachusetts, 1997).
An excellent introductory text that outlines the theoretical basis of molecular evolutionary analyses and provides insightful empirical examples.
5. Nei, M. *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
6. Endler, J. A. *Natural Selection in the Wild* (Princeton Univ. Press, New Jersey, 1986).
7. Eyre-Walker, A. & Keightley, P. D. High genomic deleterious mutation rates in hominids. *Nature* **397**, 344–347 (1999).
8. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
9. Kimura, M. *Neutral Theory of Molecular Evolution* (Cambridge Univ. Press,Cambridge, UK, 1985).
10. Fay, J. C. & Wu, C. I. The neutral theory in the genomic era. *Curr. Opin. Genet. Dev.* **11**, 642–646 (2001).
11. Fay, J. C., Wyckoff, G. J. & Wu, C. I. Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234 (2001).
12. Kreitman, M. Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.* **1**, 539–559 (2000).
A detailed review of analytical methods to detect the effects of natural selection on patterns of polymorphism.
13. Fu, Y. X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925 (1997).
14. Simonsen, K. L., Churchill, G. A. & Aquadro, C. F. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429 (1995).
15. Wall, J. D. Recombination and the power of statistical tests of neutrality. *Genet. Res.* **74**, 65–79 (1999).
16. Jorde, L. B., Watkins, W. S. & Bamshad, M. J. Human population genomics: a bridge from evolutionary history to genetic medicine. *Mol. Genet.* **10**, 2199–2207 (2001).
17. Przeworski, M., Hudson, R. R. & Di Rienzo, A. Adjusting the focus on human variation. *Trends. Genet.* **16**, 296–302 (2000).
18. Ingman, M., Kaessmann, H., Paabo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000).
19. Ke, Y. *et al.* African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. *Science* **292**, 1151–1153 (2001).
20. Jorde, L. B. *et al.* The distribution of human genetic diversity: a comparison of mitochondrial, autosomal and Y-chromosome data. *Am. J. Hum. Genet.* **66**, 979–988 (2000).
21. Kimmel, M. *et al*. Signatures of population expansion in microsatellite repeat data. *Genetics* **148**, 1921–1930 (1998).
22. Reich, D. E. & Goldstein, D. B. Genetic evidence for a Paleolithic human population expansion in Africa. *Proc. Natl Acad. Sci. USA* **95**, 8119–8123 (1998).
23. Wooding, S. & Rogers, A. R. A Pleistocene population X-plosion? *Hum. Biol.* **72**, 693–695 (2000).
24. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
The most comprehensive analysis of global patterns of human population structure completed so far. It shows that there is substantial geographical structure among populations, although the proportion of an individual's ancestry from one or more of these populations is highly variable.
25. Bamshad, M. *et al.* Human population genetic structure and inference of group membership. *Am. J. Hum. Genet.* (in the press).
26. Harpending, H. C. Genetic traces of ancient demography. *Proc. Natl Acad. Sci. USA* **95**, 1961–1967 (1995).
27. Takahata, N. Allelic genealogy and human evolution. *Mol. Biol. Evol.* **10**, 2–22 (1993).
28. Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**, 783–796 (1995).
29. Fullerton, S. M. *et al.* Geographic and haplotype structure of candidate type 2 diabetes susceptibility variants at the calpain-10 locus. *Am. J. Hum. Genet.* **70**, 1096–1106 (2002).
30. Bamshad, M. J. *et al.* A strong signature of balancing selection in the 5′ *cis*-regulatory region of *CCR5*. *Proc. Natl Acad. Sci. USA* **99**, 10539–10544 (2002).
31. Stephens, J. C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
An extensive survey of the level of polymorphism found in and near more than 300 genes, which includes a preliminary test of whether the patterns are consistent with neutrality.
32. Prezworski, M. The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179–1189 (2002).
33. Charlesworth, B. *et al.* The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
34. Hudson, R. R. & Kaplan, N. L. Deleterious background selection with recombination. *Genetics* **141**, 1605–1617 (1995).
35. Maynard-Smith, J. & Haigh, J. The hitch-hiking effect of a favorable gene. *Genet. Res.* **23**, 23–35 (1974).
A noteworthy exposition of the impact of positive selection on linked neutral polymorphisms.
36. Kaplan, N. L., Hudson, R. R. & Langley, C. H. The 'hitchhiking effect' revisited. *Genetics* **123**, 887–899 (1989).
37. Begun, J. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520 (1992).
38. Nachman, M. W. Patterns of DNA variability at X-linked loci in *Mus domesticus*. *Genetics* **147**, 1303–1316 (1997).
39. Nachman, M. W. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.* **17**, 481–485 (2001).
40. Kim, Y. & Stephan, W. Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* **155**, 1415–1427 (2000).
41. Fay, J. C., Wyckoff, G. J. & Wu, C. I. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024–1026 (2002).
42. Yang, Z. Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.* **12**, 1–7 (2002).
43. Bush, R. M. Predicting adaptive evolution. *Nature Rev. Genet.* **2**, 387–392 (2001).
44. Wyckoff, G. J., Wang, W. & Wu, C. I. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403**, 304–309 (2000).
45. Johnson, M. E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
46. Enard, W. *et al.* Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* **418**, 869–872 (2002).
47. Olsen, M. V. & Varki, A. Sequencing the chimpanzee genome: insights into human evolution and disease. *Genetics* **4**, 20–28 (2003).
48. Neel, J. V. Diabetes mellitus: a 'thrifty' genotype rendered detrimental by 'progress'? *Am. J. Hum. Genet.* **14**, 353–362 (1962).
49. Wooding, S. P. *et al.* DNA sequence variation in a 3.7-kb noncoding sequence 5′ of the *CYP1A2* gene: implications for human population history and natural selection. *Am. J. Hum. Genet.* **71**, 528–542 (2002).
50. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
51. Tishkoff, S. A. *et al.* Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293**, 455–462 (2001).
52. Saunders, M. A., Hammer, M. F. & Nachman M. W. Nucleotide variability at G6PD and the signature of malarial selection in humans. *Genetics* (in the press).
53. Toomajian, C. & Kreitman, M. Sequence variation and haplotype structure at the human *HFE* locus. *Genetics* **161**, 1609–1623 (2002).
54. Harding, R. M. Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **70**, 369–383 (1997).
55. Wooding, S. & Rogers, A. The matrix coalescent and an application to human single-nucleotide polymorphisms. *Genetics* **161**, 1641–1650 (2002).
56. Hamblin, M. T. & Di Rienzo, A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* **66**, 1669–1679 (2000).
57. Hamblin, M. T., Thompson, E. E. & Di Rienzo, A. Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* **70**, 369–383 (2002).
A meticulous analysis of the molecular signature of selection on a classical human trait, which illustrates the potential confounding effects of population history and the interaction of several selective forces.
58. Harding, R. M. *et al.* Evidence for variable selective pressures at *MC1R*. *Am. J. Hum. Genet.* **66**, 1351–1361 (2000).
59. Makova, K. D., Ramsay, M., Jenkins, T. & Li, W. H. Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter. *Genetics* **158**, 1253–1268 (2001).
60. Ding, Y. C. *et al.* Evidence of positive selection acting at the human dopamine receptor *D4* gene locus. *Proc. Natl Acad. Sci. USA* **99**, 309–314 (2002).
61. Harris, E. E. & Hey, J. Human populations show reduced DNA sequence variation at the factor IX locus. *Curr. Biol.* **11**, 774–778 (2001).
62. Nachman, M. W. & Crowell, S. L. Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, *Dmd*, in humans. *Genetics* **155**, 1855–1864 (2000).
63. Gilad, Y., Rosenberg, S., Przeworski, M., Lancet, D. & Skorecki, K. Evidence for positive selection and population structure at the human *MAO-A* gene. *Proc. Natl Acad. Sci. USA* **99**, 862–867 (2002).
64. Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nature Genet.* **30**, 233–237 (2002).
A good example of the difficulties of finding the functional variants under selection at a locus with a signature of positive selection.

65. Stephens, J. C. *et al.* Dating the origin of the *CCR5-Δ32* AIDS-resistance allele by the coalescence of haplotypes. *Am. J. Hum. Genet.* **62**, 1507–1515 (1998).
66. Leber, F. *et al.* The Δ32-ccr5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in Northeastern Europe. *Hum. Mol. Genet.* **7**, 399–406 (1998).
67. Roses, A. D. Pharmacogenetics and the practice of medicine. *Nature* **405**, 857–865 (2001).
68. Scordo, M. G. & Spina M. Cytochrome P450 polymorphisms and response to antipsychotic therapy. *Pharmacogenomics* **31**, 1–18 (2002).
69. Ikeya, K. *et al.* Human *CYP1A2*: sequence, gene structure, comparison with the mouse and rat orthologous gene, and differences in liver 1A2 mRNA expression. *Mol. Endocrinol.* **3**, 1399–1408 (1989).
70. Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Rev. Genet*. **3**, 380–390 (2002).
71. Hudson, R. R. & Kaplan, N. L. The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840 (1988).
72. Shi, Y., Radlwimmer, F. B. & Yokoyama, S. Molecular genetics and the evolution of ultraviolet vision in vertebrates. *Proc. Natl Acad. Sci. USA* **98**, 11731–11736 (2001).
73. Nordborg, M. & Tavare, S. Linkage disequilibrium: what history has to tell us. *Trends Genet.* **18**, 83–90 (2002).
74. Livingstone, F. B. Malaria and human polymorphisms. *Annu. Rev. Genet.* **5**, 33–64 (1974).
75. Cooke, G. S. & Hill, A. V. S. Genetics of susceptibility to human infectious disease. *Nature Rev. Genet.* **2**, 967–977 (2001).
76. Miller, L. H. Impact of malaria on genetic polymorphism and genetic diseases in Africans and African Americans. *Proc. Natl Acad. Sci. USA* **91**, 2415–2419 (1974).
77. Vulliamy, T. J., Mason, P. & Luzzatto, L. The molecular basis of glucose-6-phosphate dehydrogenase deficiency. *Trends Genet.* **8**, 138–142 (1992).
78. Beutler, E. G6PD deficiency. *Blood* **84**, 3613–3636 (1994).
79. Ruwende, C. *et al.* Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature* **376**, 246–249 (1995).
80. Austin, L. H. & Federica, V. Very large long-term effective population size in the virulent human malaria parasite *Plasmodium falciparum*. *Proc. R. Soc. Lond. B* **268**, 1855–1860 (2001).
81. Coluzzi, M. The clay feet of the malaria giant and its African roots: hypotheses and inferences about origin, spread and control of *Plasmodium falciparum*. *Parassitologia* **41**, 277–283 (1999).
82. Feder, J. N. *et al.* A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature Genet.* **13**, 399–408 (1996).
83. Merryweather-Clarke, A. T., Pointon, J. J., Shearman, J. D. & Robson, K. J. Global prevalence of putative haemochromatosis mutations. *J. Med. Genet.* **34**, 275–278 (1997).
84. Ajioka, R. S. *et al.* Haplotype analysis of hemochromatosis: evaluation of different linkage-disequilibrium approaches and evolution of disease chromosomes. *Am. J. Hum. Genet.* **60**, 1439–1447 (1997).
85. Thomas, W. *et al.* Haplotype and linkage disequilibrium analysis of the hereditary hemochromatosis gene region. *Hum. Genet.* **102**, 517–525 (1998).
86. Bulaj, Z. J., Griffen, L. M., Jorde, L. B., Edwards, C. Q. & Kushner, J. P. Clinical and biochemical abnormalities in people heterozygous for hemochromatosis. *N. Engl. J. Med.* **335**, 1799–1805 (1996).
87. Scholl, T. O., Hediger, M. L., Fischer, R. L. & Shearer, J. W. Anemia vs. iron deficiency: increased risk of preterm delivery in a prospective study. *Am. J. Clin. Nutr*. **55**, 985–988 (1992).
88. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
89. Hawkes, K., O'Connell, J. F., Blurton Jones, N. G., Alvarez, H. & Charnov, E. L. Grandmothering, menopause, and the evolution of human life histories. *Proc. Natl Acad. Sci. USA* **95**, 1336–1339 (1998).
90. Lewontin, R. C. & Hubby, J. L. A molecular approach to the study of genetic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54**, 595–609 (1966).
91. Kaplan, N. L., Darden, T. & Hudson, R. R. The coalescent process in models with selection. *Genetics* **120**, 819–829 (1988).
92. Richman, A. D. & Kohn, J. R. Self-incompatibility alleles from *Physalis*: implications for historical inference from balanced genetic polymorphisms. *Proc. Natl Acad. Sci. USA* **96**, 168–172 (1999).
93. Hughes, A. L. & Yeager, M. Natural selection at major histocompatibility complex loci of vertebrates. *Annu. Rev. Genet*. **32**, 415–435 (1998).
94. Verrelli, B. C. *et al.* Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. *Am. J. Hum. Genet.* **71**, 1112–1128 (2002).
95. Baum, J., Ward, R. H. & Conway, D. H. Natural selection on the erythrocyte surface. *Mol. Biol. Evol.* **19**, 223–229 (2002).
96. Wu, X., Di Rienzo, A. & Ober, C. A population genetics study of single nucleotide polymorphisms in the interleukin 4 receptor a (*IL4RA*) gene. *Genes Immun.* **2**, 128–134 (2001).
97. Liu, R. *et al.* Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply exposed individuals to HIV-1 infection. *Cell* **86**, 367–377 (1996).
98. Schierup, M. H., Vekemans, X. & Charlesworth, D. The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genet. Res.* **76**, 51–62 (2000).
99. Charlesworth, B., Nordborg, M. & Charlesworth, D. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res.* **70**, 155–174 (1997).
100. Takahata, N. & Nei, M. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* **124**, 967–978 (1990).
101. Salamon, H. *et al.* Evolution of HLA class II molecules: allelic and amino acid site variability across populations. *Genetics* **152**, 393–400 (1999).
102. Grimsley, C., Mather, K. A. & Ober, C. *HLA-H*: a pseudogene with increased variation due to balancing selection at neighboring loci. *Mol. Biol. Evol.* **15**, 1581–1588 (1998).
103. Muller, H. J. Our load of mutation. *Am. J. Hum. Genet.* **2**, 111–176 (1950).
104. Gillespie, J. H. *The Causes of Molecular Evolution* (Oxford Univ. Press, New York, 1991).
105. Payseur, B. A., Cutter, A. D. & Nachman, M. W. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol.* **19**, 1143–1153 (2002).
106. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
107. Sunyaev, S. R., Lathe, W. C., Ramensky, V. E. & Bork, P. SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet.* **16**, 335–337 (2000).
108. Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805–1814 (2002).
109. Wiehe, T. The effect of selective sweeps on the variance of the allele distribution of a linked multiallele locus: hitchhiking of microsatellites. *Theor. Popul. Biol.* **53**, 272–283 (1998).
110. Comeron, J. M. & Kreitman, M. Population, evolutionary and genomic consequences of interference selection. *Genetics* **161**, 389–410 (2002).
111. Navarro, A. & Barton, N. H. The effects of multilocus balancing selection on neutral variability. *Genetics* **161**, 849–863 (2002).
112. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
113. Tajima, F. Evolutionary relationships of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
114. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
115. Li, W. H. & Sadler, L. A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
116. Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135–142 (2002).
117. Zietkiewicz, E. *et al.* Genetic structure of the ancestral population of modern humans. *J. Mol. Evol.* **47**, 146–155 (1998).
118. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
119. Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413 (2000).
Introduces a new statistical test of neutrality on the basis of the prediction that, immediately after a selective sweep, an excess of high-frequency-derived polymorphisms is expected at linked sites.
120. Przeworski, M. The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179–1189 (2002).
121. Gilad, Y. *et al.* Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. *Nature Genet.* **26**, 221–224 (2000).
122. Huttley, G. A. *et al.* Adaptive evolution of the tumor suppressor *BRCA1* in humans and chimpanzees. *Nature Genet.* **25**, 410–413 (2000).
123. Suzuki, Y. & Gojobori, T. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **16**, 1315–1328 (1999).
124. Schlotterer, C. Towards a molecular characterization of adaptation in local populations. *Curr. Opin. Genet. Dev.* **12**, 1–4 (2002).
125. Lewontin, R. C. & Krakauer, J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175–195 (1973).
126. Bowcock, A. M. *et al.* Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proc. Natl Acad. Sci. USA* **88**, 839–843 (1991).
127. Beaumont, M. A. & Nichols, R. A. Evaluating loci for use in genetic analysis of population structure. *Proc. R. Soc. Lond. B* **263**, 1619–1626 (1996).
128. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *ADH* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
129. Fu, X. Y. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709 (1993).
130. Li, W. H., Wu, C. I. & Luo, C. C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**, 150–174 (1985).
131. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1985).
132. Hudson, R. R., Kreitman, M. & Aguade, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159 (1987).

Acknowledgements

We thank L. B. Jorde, A. R. Rogers and three anonymous reviewers for comments and criticisms. The authors are supported by funds from the US National Institutes of Health and the National Science Foundation.

Online links

DATABASES

The following terms in this article are linked online to:

LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
BRCA1 | calpain-10 | CCR5 | *CYP1A2* | DRD4| *G6PD* | MAOA | *MC1R* | *PPARG*
OMIM: http://www.ncbi.nlm.nih.gov/Omim
idiopathic haemochromatosis | type II diabetes
Access to this interactive links box is free online.