Genome-wide scans for loci under selection in humans

*James Ronald and Joshua M. Akey**

University of Washington, Seattle, Washington, USA
**Correspondence to*: Tel: +1 206 543 7254; E-mail: akeyj@u.washington.edu

Date received (in revised form): 31st January 2005

Abstract

Natural selection, which can be defined as the differential contribution of genetic variants to future generations, is the driving force of Darwinian evolution. Identifying regions of the human genome that have been targets of natural selection is an important step in clarifying human evolutionary history and understanding how genetic variation results in phenotypic diversity, it may also facilitate the search for complex disease genes. Technological advances in high-throughput DNA sequencing and single nucleotide polymorphism genotyping have enabled several genome-wide scans of natural selection to be undertaken. Here, some of the observations that are beginning to emerge from these studies will be reviewed, including evidence for geographically restricted selective pressures (ie local adaptation) and a relationship between genes subject to natural selection and human disease. In addition, the paper will highlight several important problems that need to be addressed in future genome-wide studies of natural selection.

***Keywords:** genetic variation, evolutionary genomics, natural selection, single nucleotide polymorphisms (SNPs)*

Introduction

Phenotypic diversity is a ubiquitous characteristic of natural populations. Individuals vary in almost every conceivable way, including physical appearance, behaviour, disease susceptibility, ability to detoxify drugs and perception of environmental stimuli.[1] Although environmental forces undoubtedly contribute to phenotypic variation, so too does genetic variation. Therefore, explaining the evolutionary forces that create, maintain and shape patterns of human genetic variation is of fundamental importance in understanding phenotypic variation.[2]

An important goal in studies of human genetic variation is to identify loci that have been targets of natural selection due to their variable effects on the fitness of individuals throughout a population's history. Signatures of natural selection delimit regions of the genome that are, or have been, functionally important. Therefore, identifying such regions will facilitate the identification of genetic variation that contributes to phenotypic variation and help to functionally annotate the genome. Unfortunately, inferring the action of natural selection remains a challenge. This is likely to change in the near future, as high-throughput methods for cataloguing genetic variation on a genome-wide scale and new statistical tools for detecting selection have been, and continue to be, developed.

Much important work has been done on genome scans for natural selection in model organisms such as *Drosophila*;[3–5] this review, however, will focus on studies performed in human populations. Firstly, there will be a summary of the effects of natural selection and population history on patterns of genetic variation and some of the common statistical methods used to test for deviations from neutrality will be presented. Next, a critical evaluation of several empirical genome-wide scans for selection will be presented. Finally, the paper will highlight several important problems, both practical and conceptual, that need to be addressed in future studies.

Human genetic variation: The neutral expectation

The evolutionary sojourn of a newly arisen mutation depends upon how it affects the fitness of the individual who possesses it. The neutral theory of molecular evolution posits that the vast majority of polymorphisms in a population are selectively neutral and have no appreciable effects on fitness.[6,7] Under neutrality, changes in allele frequency are governed by the stochastic effects of genetic drift in populations of finite size. Thus, the effective population size, N_e, and neutral mutation rate, μ_o, determine levels of polymorphism within species and the rate of divergence between species.[8] In addition, the effect of mutations with small fitness effects can be rendered 'nearly neutral' if the product of N_e and s (which measures the strength of selection) is <1.[9,10] For human populations, N_e is approximately

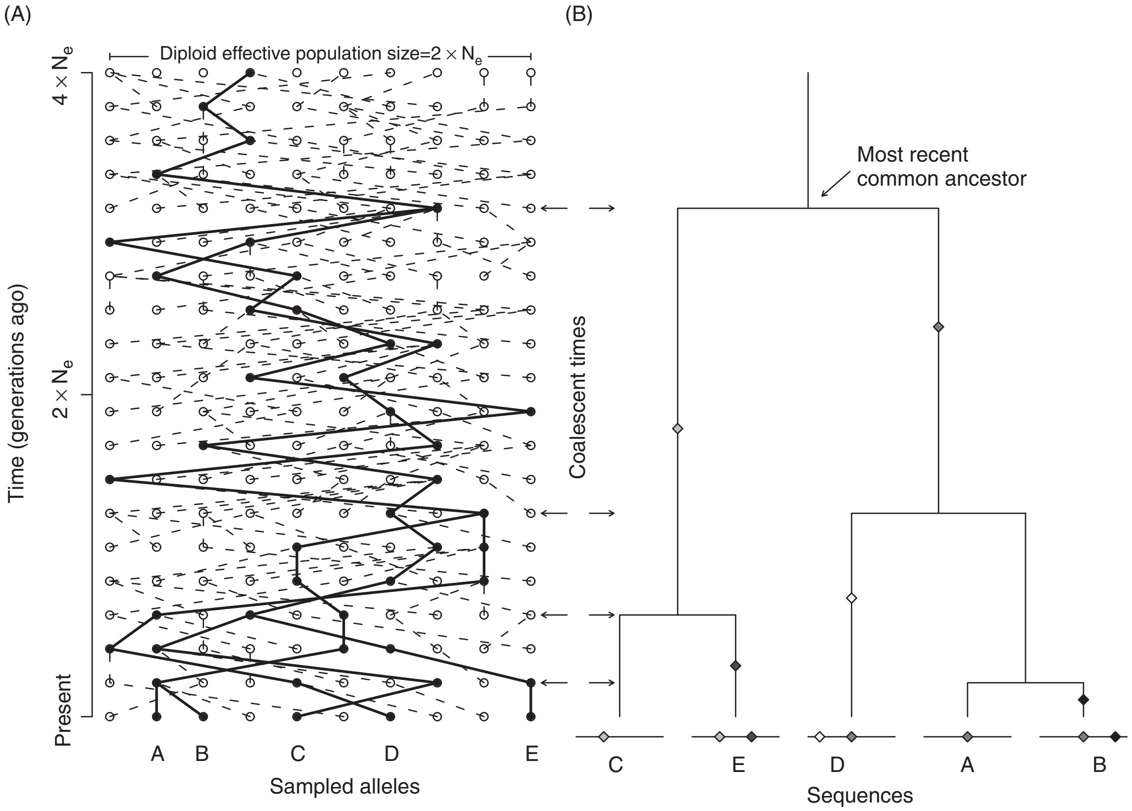

Figure 1. Coalescent representation of a neutrally evolving sequence. (A) Explicitly tracing the history of a sample of alleles from the population, each progenitor allele is derived from a randomly chosen parental allele. Occasionally, two progenitor alleles are derived from the same parent, causing the lineages of these two alleles to unite or coalesce when they are followed backward in time from the present. Note that if progeny are derived from parents at random, the probability that two lineages coalesce increases as the number of distinct lineages increases and as the effective population size decreases. Thus, for a constant-sized population, a characteristic distribution of waiting times between coalescent events is expected. (B) The untangled, coalescent representation of (A) is created by treating lineages as branches, ignoring the intermediate ancestors between coalescent events. Under neutrality, mutational events (represented by shaded diamonds) are uniformly distributed throughout time, hence the number of mutations that occur on a branch is proportional to the length of the branch. Note that mutations occurring on external branches are rare, appearing on only a single allele, whereas mutations occurring on internal branches are common.

10,000 and therefore $|s|$ must be greater than 10^{-4} to overcome the stochastic effects of genetic drift. Because the neutral theory makes explicit and quantitative predictions about expected patterns of genetic variation within and between species, it is an indispensable tool in studies of natural selection. Specifically, the neutral theory provides an essential foundation for evaluating the evidence either for or against selection in empirical data, as it serves as the null hypothesis when exploring alternative evolutionary models.[11,12]

In combination with the neutral theory, coalescent theory provides a powerful framework for conceptualising and making inferences about evolutionary forces. The coalescent is a stochastic model of gene geneaologies[13–17] and has emerged as the primary analytical tool in studies of genetic variation. In classical population genetics theory, the initial state of a population is defined and one observes the evolution of the entire population by looking forward in time. By contrast, the coalescent is a sample-driven theory that traces the history of coalescent events backwards in time (see Figure 1 for some basic properties of the coalescent). Several excellent and detailed reviews of coalescent theory can be found elsewhere.[18,19] Deviations from the standard neutral model distort the branch lengths, topology and coalescent times of gene genealogies, as described below.

Evolutionary forces perturb patterns of genetic variation

Natural selection and population demographic history perturb patterns of genetic variation relative to what is expected under a standard neutral model (constant sized, randomly mating, panmictic population at mutation drift equilibrium). Below, the way in which selection and demographic history affect patterns of genetic variation will be considered, from a coalescent point of view.

Theoretical studies have investigated the evolutionary dynamics of genetic variation subject to a variety of selective pressures, including, purifying[20–22] positive[23–26] and balancing selection.[27–30] This paper will focus on positive and balancing selection, as these have been the primary types of selection that current genome-wide scans have studied. Positive selection acts to increase the frequency of advantageous alleles in a population. Strongly advantageous mutations are rapidly swept to fixation, hence the term 'selective sweep'. Importantly, through a process referred to as 'genetic hitchhiking', positive selection also affects patterns of neutral polymorphisms linked to an advantageous mutation.[23] Positive selection leads to a shallow star-like genealogy (Figure 2) with a decreased time to the most recent common ancestor. Tracing the history of alleles backwards in time, these effects are a direct consequence of the rapid coalescence of lineages in the small but expanding progenitor population.[31] The signature of positive selection includes reduced levels of genetic variation compared with neutral expectations,[24–26] a skew in the allele frequency spectrum towards low-frequency alleles[32] (including an excess of high-frequency derived alleles[33]) and elevated levels of linkage disequilibrium.[34]

Balancing selection occurs when polymorphisms are selectively maintained in a population. By contrast with positive selection, the genealogy of a locus subject to balancing selection is characterised by an increased time to the most recent common ancestor and long internal branches (Figure 2). The effect of balancing selection on gene genealogies can be understood by considering balanced alleles as distinct subpopulations, such that coalescence events can occur rapidly within a subpopulation but slowly between subpopulations.[27] The signature of balancing selection includes elevated levels of polymorphism relative to neutral expectations and a skew of the allele frequency distribution towards an excess of intermediate frequency alleles.[29,30,35]

In addition to natural selection, population demographic history can also have strong influences on patterns of genetic variation, which often mimic the effect of natural selection.[36,37] In other words, inferences of natural selection are confounded by population demographic history. For example, both positive selection and increases in population size have similar effects on gene genealogies (Figure 2); both processes therefore lead to an excess of low-frequency alleles in a population. In fact, strong positive selection can be thought of as a rapid population expansion of an advantageous allele as it sweeps through a population. Similarly, population structure and balancing selection both result in subdivided genealogies and therefore both processes are expected to result in an excess of intermediate-frequency alleles in a population (Figure 2). Population bottlenecks can lead to an excess of either low- or intermediate-frequency alleles relative to neutral expectations, depending on the age and severity of the bottleneck. Figure 2 demonstrates the effect of a severe and recent bottleneck, which forces all lineages to coalesce at the time of the size reduction and results in a genealogy that is similar to positive selection. Human populations clearly do not meet all of the assumptions of the standard neutral model; hence, rejecting the standard neutral model for a particular locus cannot be interpreted as unambiguous evidence for selection.

Detecting the signature of natural selection

Before presenting the results from genome-wide scans for natural selection, there now follows a brief description of some commonly used statistical methods designed to detect departures from neutrality, highlighting some of their strengths and limitations. The following is not meant to be an exhaustive discussion of such tests, and descriptions of many interesting and useful methods will not be included here. For further study, the reader is encouraged to see an excellent review by Kreitman.[38]

Statistical tests of neutrality can broadly be classified into three categories, based upon the type of data that they use: 1) within species tests; 2) within and between species tests; and 3) between species tests (Table 1). The most common class of within-species tests compares summary statistics of the observed allele frequency distribution at a locus with the values expected under neutral evolution; it includes Tajima's D,[39] Fu and Li's D and F,[40] Fay and Wu's H[33] and Fu's F_s.[41] An attractive feature of these tests is that they do not require any *a priori* classification of functional versus non-functional sites, thus making them equally suitable for protein and non-protein coding regions. In a thorough examination of the power of Tajima's D and Fu and Li's D and F, Simonsen *et al.*[42] found that these tests can only detect selective sweeps in a narrow time interval in the recent past, and that they can only detect balancing selection if it has acted for a very long time period. Interestingly, Simonsen *et al.* also found that the power of these tests could drop below the nominal false-positive rate, α, if non-neutral evolution did not occur within these critical time windows, thus creating the undesirable scenario in which rejection of the null is more likely if the null is true than if it is false. Fu observed similar results, although he found that F_s was most powerful and performed better at detecting more ancient positive selection.[41]

The site-frequency spectrum tests discussed above are confounded by demographic events such as population growth, bottlenecks and subdivision (Figure 2) and are rendered conservative by intra-locus recombination. The desire to estimate population demographic parameters, recombination rates and evolutionary parameters has

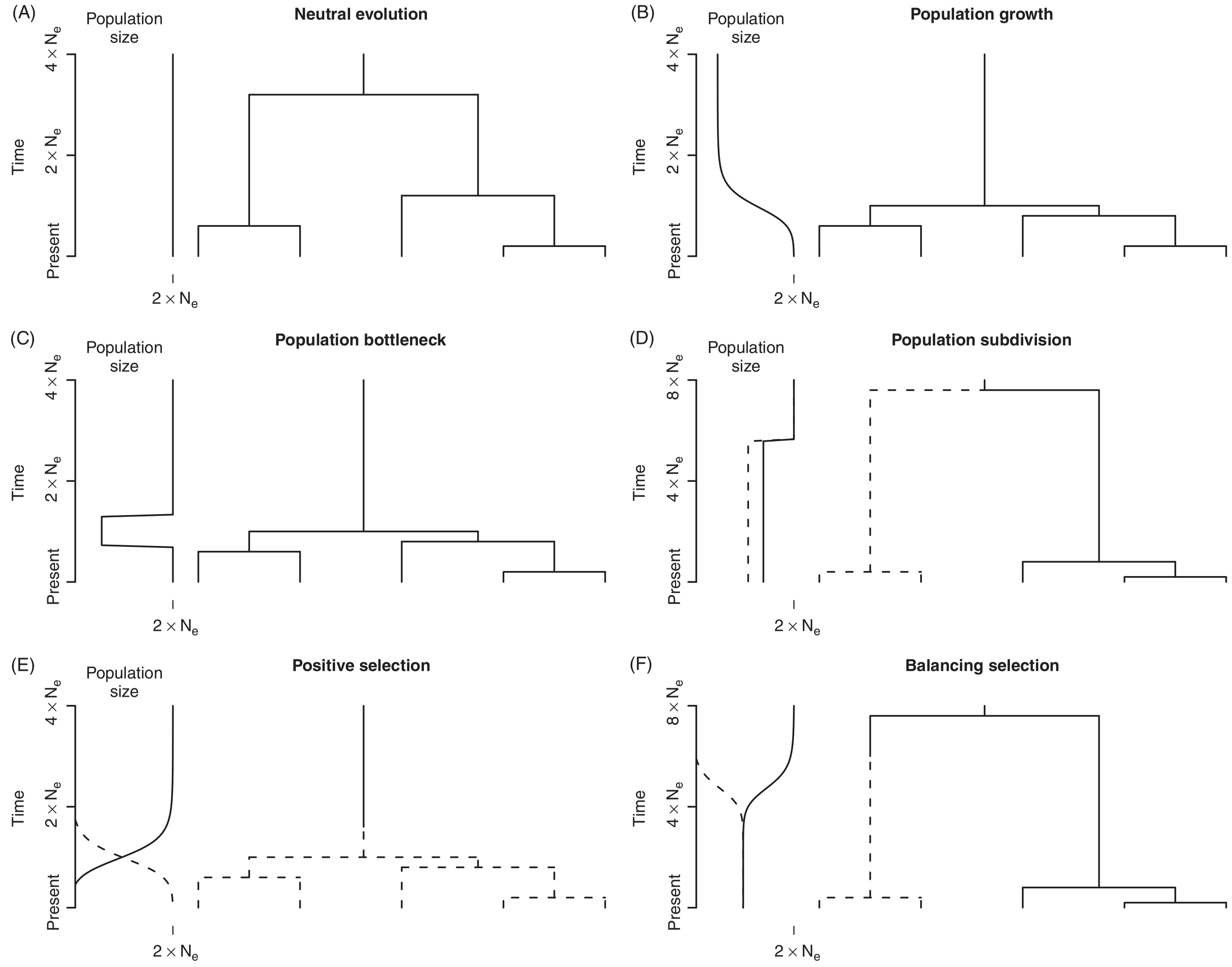

Figure 2. Effects of deviations from neutrality on gene genealogies. (A) Neutral evolution. (B) Population growth. (C) Population bottleneck. Here, only one ancestral lineage passes through the bottleneck, leading to a short tree with relatively long external branches. (D) Population subdivision. An initial population (represented by solid lines) separates into two subpopulations, which are denoted by dashed and solid lines. (E) Positive selection. An advantageous allele (represented by the dashed lines) sweeps through the population to fixation. Note that the genealogy of a selective sweep is similar to that produced by a population growth or bottleneck. (F) Balancing selection. An allele that is advantageous only in the heterozygous state (dashed line) appears in the population and is maintained at an intermediate frequency. Note that the genealogy under balancing selection is similar to population subdivision. N_e, effective population size.

Table 1. Statistical tests of neutrality.

Class	Strengths	Limitations
Within-species tests		
Site-frequency spectrum	Most thoroughly studied	Powerful only if non-neutral evolution occurred within a critical time period
Coalescent likelihood methods	In principle, more powerful than summary statistic methods, possible to estimate multiple parameters simultaneously	Computationally intensive
F_{ST}	Does not require sequence data, can be performed with marker genotypes only	Low power to detect balancing selection
Long-range haplotype test (LRH)	Does not require sequence data, can be performed with marker genotypes only	Sensitivity of LRH test to population demographics and haplotype construction not well studied; not applicable to detecting balancing selection
Within- and between-species tests		
Hudson–Kreitman–Aguade (HKA)	May be useful for detecting balancing selection	Requires sequences from multiple individuals in two species; may be difficult to interpret significant HKA test
McDonald–Kreitman	More sensitive than raw measure of the ratio of one number of non-synonymous amino acid substitution in a gene to the number of synonymous substitutions (d_n/d_s); may be fairly robust to population demographics and recombination	Requires sequences from multiple individuals in two species; selection to change codons may adversely affect test; applicable to protein-coding regions only
Between-species tests		
Ratio of non-synonymous to synonymous substitutions	Requires only a single sequence from each species; raw d_n/d_s measure is unlikely to be confounded by demographics or recombination	Raw d_n/d_s measure is extremely stringent; applicable to protein-coding regions only

prompted the development of maximum likelihood-based methods which use the complete data, rather than summary statistics.[31,43,44] These methods are computationally intensive and are not currently feasible for large datasets, but they potentially allow for substantial gains in statistical power relative to summary statistics methods and are likely to become increasingly important tools in the future (for a general discussion, see Felsenstein[45]).

Another within-species test that has been used to detect selection is to compare the variation in allele frequencies between populations, which can be quantified by the statistic F_{ST}. Under selective neutrality, F_{ST} is determined by genetic drift, whereas natural selection is a locus-specific force that can cause systematic deviations in F_{ST} values for a selected gene and nearby genetic markers. For example, geographically restricted directional selection may lead to an increase in F_{ST} of a selected locus, whereas balancing or species-wide directional selection may lead to a decrease in F_{ST} compared with neutrally evolving loci.[46–50] In a series of simulation experiments analysing two different F_{ST} test implementations, Beaumont and Balding found that this approach yielded sufficient power to detect positive selection provided that the selective coefficient was approximately five times larger than the migration rate, but that F_{ST} had little power to detect balancing selection.[50]

Positive selection is also expected to increase levels of linkage disequilibrium (LD) relative to neutral expectations. Recently, a new statistical test was developed, the long-range haplotype (LRH) test,[33] which takes advantage of ancestral recombination events and the associated decay in LD to identify genes subject to positive selection. The rationale for this test is that a common allele with long-range LD potentially represents a site that has appeared recently and was driven to high frequency before recombination could erode LD. The LRH approach does not detect balancing selection, however, and the robustness of the test to non-neutral population demographics, the choice of haplotype defining markers and phase misspecification have not been well studied.

The second major class of neutrality tests compares levels of within-species polymorphism and between-species divergence and includes the Hudson–Kreitman–Aguade (HKA)[51] and McDonald–Kreitman (MK)[52] tests. The HKA method tests the goodness of fit of the observed levels of polymorphism within species and the observed divergence between species to those predicted under neutral theory. In order to determine polymorphism and divergence expectations under neutrality, data are required from at least two loci in each species, so that a simultaneous estimate can be made of a time-since-speciation parameter and a relative population size parameter. Under the HKA test, rejection of the null is formally interpreted as elevated polymorphism at one locus or reduced polymorphism at the other, or excess divergence at one locus or limited divergence at the other. Thus, it may not be obvious which locus or which process is responsible for producing a statistically significant test. McDonald[53] has described improvements to the HKA test which may ameliorate this problem.

In the MK test, a 2 × 2 contingency table is formed to compare the number of non-synonymous and synonymous sites that are polymorphic within a species (P_N and P_S) and fixed between species (D_N and D_S). Under neutrality, the ratio of non-synonymous to synonymous sites that are polymorphic equals the ratio of non-synonymous to synonymous sites that are fixed (ie $P_N/P_S = D_N/D_S$). Under positive selection, however, these two ratios are no longer equal and $D_N/D_S > P_N/P_S$.[54] Among the strengths of the MK test are that it does not require assumptions about population demographic history (although under some circumstances the test can be adversely affected by increases in effective population size[54]) and is relatively insensitive to intra-locus recombination. Positive or purifying selection for codon usage may, however, bias the MK test.[38]

The final class of neutrality tests uses between-species data to test for adaptive protein evolution. The classic test of positive selection compares the number of non-synonymous amino acid substitutions in a gene (d_n) with the number of synonymous amino acid substitutions (d_s). Under neutrality, the mutation rate at both categories of sites is the same, and d_n/d_s is expected to equal one; however, $d_n/d_s < 1$ for proteins subject to purifying selection and $d_n/d_s > 1$ for proteins under adaptive evolution. Although $d_n/d_s > 1$ provides strong evidence for adaptive protein evolution, it is a very conservative test, particularly if only a small number of codons have been selected for. The basic test has also been extended by Nielsen and Yang[55] and others to include models of codon and transition/transversion bias, to detect variation in d_n/d_s ratios among lineages and to identify specific codons under selection.[56,57]

Key advantages of genome-wide analyses

As alluded to above, distinguishing between the confounding effects of natural selection and population demographic history is difficult when studying a single locus. When many unlinked genes are considered, however, a clear strategy emerges. Population demographic history affects patterns of variation at all loci in a genome in a similar manner, whereas natural selection acts upon specific loci.[12,37,46,58] Therefore, by sampling a large number of unlinked loci throughout the genome, empirical distributions of test statistics can be constructed and genes subject to locus-specific forces, such as natural selection, can be identified as outlier loci.

To provide some examples of how genome-wide analyses can facilitate inferences of natural selection, Figure 3 shows empirical distributions of Tajima's D, F_{ST} and d_n/d_s, along with their theoretical distributions, simulated under both standard neutral models and alternative demographic histories. Figure 3 highlights two important points. First, empirical distributions provide important information that can be used to infer population demographic history. For example, the demographic models used in simulating Tajima's D (Figure 3A and B) recapitulate the empirical distributions much more closely than data simulated under a standard neutral model. Secondly, outlier loci can be identified with greater precision and accuracy with more realistic models of human demographic history. Specifically, the best-fitting non-neutral distributions dramatically reduce the number of test statistics that are apparent outliers under neutrality. Conversely, some test statistics that do not appear to deviate from neutrality are outliers under the best non-neutral distributions. Thus, in principle, empirical distributions of test statistics can be used both to reduce the false-positive rate and to improve power. Although this general strategy has recently been dubbed 'population genomics', the theoretical foundation of searching for outlier loci to find targets of natural selection was outlined decades ago.[46,47]

In addition to providing empirical distributions, genome-wide scans for natural selection offer several additional

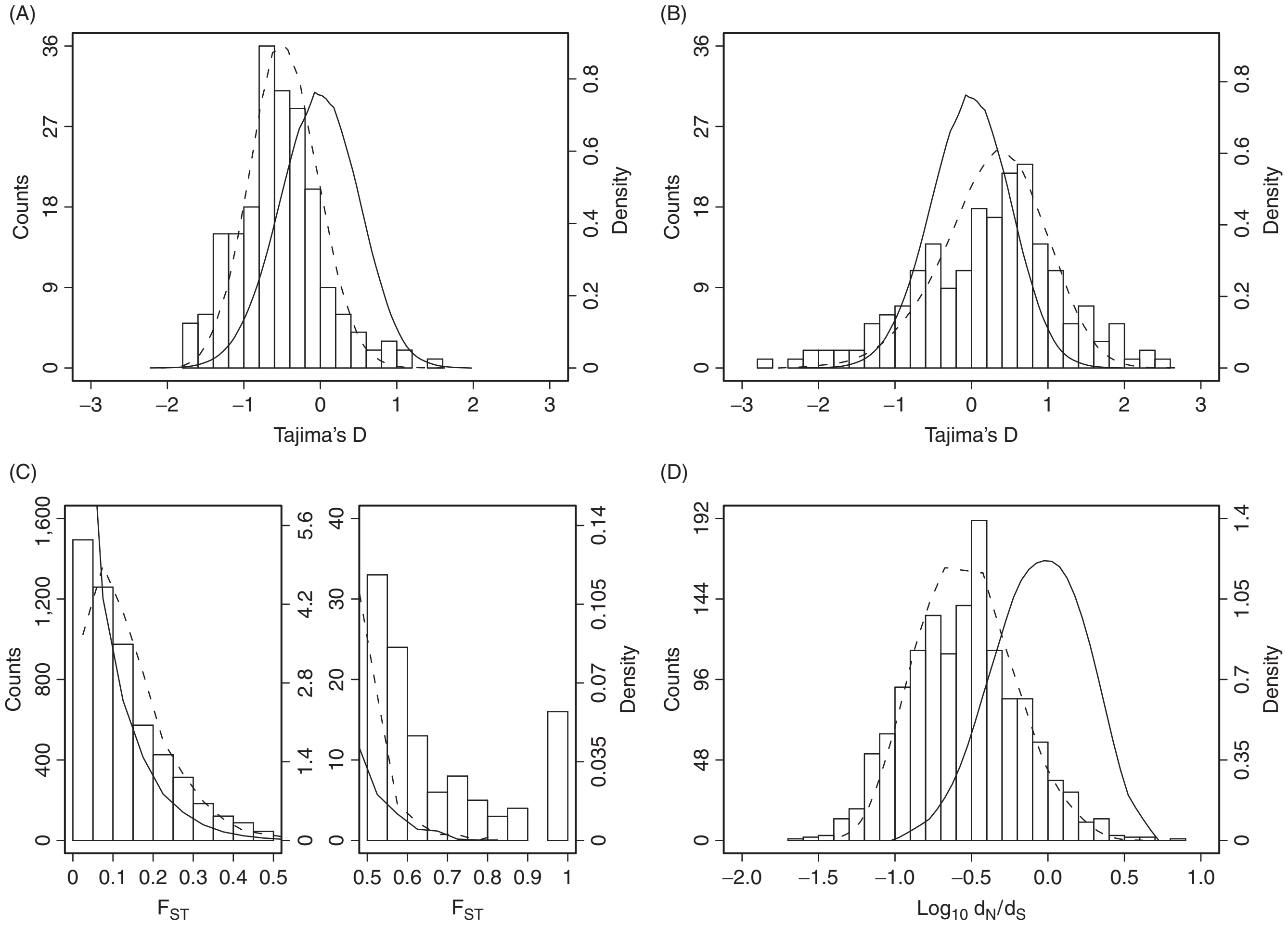

Figure 3. Empirical distributions of some commonly used test statistics and their theoretical distributions. (A) Distribution of Tajima's D in 201 genes in an African-American sample. The empirical distribution is denoted with bars. The solid line indicates the distribution of Tajima's D simulated under a standard neutral model with recombination using the ms program.[59] The dashed line indicates the distribution of Tajima's D simulated under the best fitting population demographic model from Akey *et al.*[60] (exponential expansion starting 50,000 years ago with a growth rate of 10^{-3} per generation). (B) Distribution of Tajima's D from the same 201 genes in a European-American sample. The best-fitting population demographic model is a bottleneck beginning 40,000 years ago with an inbreeding coefficient of 0.175.[60] Data used to calculate Tajima's D in panels (A) and (B) was obtained from the SeattleSNPs project (http://pga.gs.washington.edu/). (C) The empirical distribution of F_{ST} for 5,590 chromosome 7 single nucleotide polymorphisms (SNPs) obtained from the HapMap project.[61] The theoretical distributions of F_{ST} were simulated using ms.[59] An island migration model was assumed, with a constant migration parameter between each pair of populations. The solid line shows the expected distribution of F_{ST} under neutrality, whereas the dashed line shows the expected distribution under neutrality with an ascertainment bias favouring common SNPs. To approximate the ascertainment bias in the HapMap data, a 'double-hit' SNP discovery strategy was modelled[61] by randomly selecting four simulated chromosomes and only analysing SNPs where each allele was observed twice. The migration parameter was the same for both distributions and was chosen such that the mean of the biased F_{ST} distribution (0.138) closely matched the mean observed F_{ST}. (D) The empirical distribution of d_n/d_s from Clark *et al.*[62] Only those genes with $d_n > 0.001$ and $d_s > 0.001$ are shown. The solid line shows the distribution of d_n/d_s estimated using the method of Yang and Nielsen[63] for neutrally evolving coding sequences simulated with the PAML program.[64] The dashed line shows the distribution of d_n/d_s for sequences under negative selection, with the magnitude of the selective force chosen such that the mean $\log_{10} d_n/d_s$ (−1.25) matched the mean of the Clark *et al.*[62] distribution. The length of the simulated sequences (450 codons) was chosen to match the mean length of sequences from Clark *et al.*[62]

advantages compared with single-locus studies. Genome-wide scans can suggest general principles about the types of variation that natural selection acts most forcefully upon. Datasets derived from an unbiased sampling of loci throughout the genome allow for the discovery of novel functional elements whose presence is revealed by evidence for selection. Whole-genome scans also have the potential to reveal networks of genes whose evolutionary histories are correlated due to their collaboration in executing cellular functions. Finally, it is important to stress that genome-wide analyses do not preclude single-locus analyses, and that achieving a detailed and thorough understanding of the selective and demographic forces acting upon a locus will necessitate focused single-locus analyses drawing from multiple scientific disciplines.

Genome scans for natural selection

Several genome-wide scans for natural selection have recently been performed and are summarised in Table 2. These studies have used a variety of different statistical approaches, data and populations, but are united by the common theme of sampling a large number of loci and making inferences of natural selection. Below, some of these studies will be considered in more detail, to highlight the salient results emerging from genome-wide scans for selection.

One of the first genome-wide screens for selection to be performed analysed 26,530 single nucleotide polymorphisms (SNPs), which were genotyped in three human populations: African-Americans, East Asians and European-Americans.[65] An empirical distribution of F_{ST} was constructed and outlier SNPs in gene regions were identified. As discussed above, geographically restricted selection (local adaptation) can accentuate levels of population structure by creating large differences in allele frequencies between populations. Conversely, balancing selection can lead to lower than expected levels of population structure. In total, 174 candidate selection genes were identified whose levels of population structure were significantly different compared with neutral expectations (156 genes had exceptionally high values of F_{ST} and 18 had exceptionally low values of F_{ST}). In addition, the average F_{ST} was significantly different between SNPs located in exons, introns and non-genic regions, which is consistent with the action of purifying selection. One limitation of this study was that it relied upon markers that were discovered in a small number of chromosomes, which can lead to significant ascertainment bias (ie in this case, an over-representation of intermediate-frequency alleles). Such ascertainment bias complicates inferences of natural selection, and, as the authors note, additional analyses are needed to confirm the signature of selection in these genes.

Three genome-wide scans for natural selection have also been performed with microsatellite markers,[66–68] the largest of which analysed 5,257 microsatellite markers in 28 individuals of European descent.[66] A sliding window analysis across the genome revealed 43 bins that contained a significant reduction in heterozygosity relative to neutral expectations. Interestingly, the recombination rate in these 43 bins was significantly reduced compared with the genome-wide average, which is consistent with theoretical predictions

Table 2. Summary of genome-wide scans for selection.

Data	Number of loci	Statistical method	Comments	Reference
SNP	26,530 markers	Population structure	174 candidate selection genes were identified whose levels of population structure were inconsistent with neutrality	[65]
Microsatellite	5,257 markers	Intraspecific polymorphism	43 sliding windows were identified that contained significant deficits in heterozygosity relative to neutral expectations	[66]
Microsatellite	332 markers	Population structure	15 loci were identified with levels of population structure inconsistent with neutrality	[67]
Microsatellite	624 markers	Population structure	13 loci were identified with levels of population structure inconsistent with neutrality	[68]
DNA sequence	7,645 genes	Interspecific divergence	1,547 genes with $d_n/d_s > 1$ in the human lineage	[62]

SNP = single nucleotide polymorphism; d_n = number of non-synonymous amino acid substitutions in a gene; d_s = number of synonymous amino acid substitutions in a gene.

that positive selection will be easier to detect in regions of the genome with low recombination rates.[23]

The other two microsatellite based genome-wide scans for selection included multiple populations and searched for evidence of local adaptation by identifying outlier loci that exhibited large levels of population structure relative to the empirical distribution of all loci. Specifically, Kayser *et al.*[67] studied 332 microsatellite markers in 47 Europeans and 47 Africans (23 Ethiopians and 24 South Africans). The test statistics R_{ST}, a multiallelic analogue of F_{ST}, and ln RV, which is the natural log of the variance in allele sizes between populations,[69] were calculated for all loci. Numerous outlier loci were detected and 11 were studied further by genotyping additional microsatellite markers in these regions. The additional microsatellite analyses confirmed the large differences in genetic differentiation, which strengthens the hypothesis that outlier loci have been targets of geographically restricted selective pressures. Similarly, Storz *et al.*[68] analysed a total of 624 microsatellite loci that were previously genotyped in multiple populations from Africa, Europe and Asia. Again, measures of population structure were calculated for all markers (F_{ST} and an analogue to ln RV) and outlier loci were identified. In total, 13 outlier loci were found and all but one had significant reductions in heterozygosity in non-African populations; this was interpreted as evidence that local adaptation was more common outside of Africa. An important limitation of the microsatellite analyses is that the high mutation rate of microsatellites may obscure signatures of selection, except in low-recombining regions of the genome.[70,71]

In one of the largest gene-based genome-wide screens performed to date, Clark *et al.*[62] analysed 7,645 orthologous genes from humans, chimpanzees and mice (see also Figure 3D). Maximum-likelihood models were fitted to protein-coding DNA sequences to estimate rates of synonymous (d_s) and non-synonymous (d_n) substitutions. In total, 1,547 genes had d_n/d_s ratios >1 in humans, which is commonly interpreted as evidence for positive selection, but the neutral model could be formally rejected at $p < 0.05$ for only six of these genes. Using an alternative statistical method with greater sensitivity, branch site models were fitted to the data in order to detect accelerated rates of d_n/d_s in the human lineage for a subset of nucleotide sites (ie d_n/d_s does not have to be >1 for the entire gene). A total of 667 genes were identified as significant at $p < 0.05$ in this analysis; subsequent bioinformatics analyses revealed two interesting observations. First, accelerated rates of evolution were found for several functional classes of genes, including olfactory, nuclear transport and sensory perception. Secondly, genes with evidence for positive selection were enriched for genes that are associated with human diseases, as defined by the Online Mendelian Inheritance of Man (OMIM) database. OMIM primarily contains monogenic disease genes with large phenotypic effects, and it will therefore be interesting to see if these results also extend to complex disease genes. Indeed, signatures of natural selection have been described for several genes associated with various complex diseases.[34,72–78] If complex disease genes are enriched for signatures of natural selection, finding targets of adaptive evolution may be a useful strategy for prioritising candidate genes in disease-mapping studies.

It is important to note that a recent theoretical study has suggested that maximum-likelihood branch site models may have a high false-positive rate[79] and, therefore, the 667 significant (at $p < 0.05$) genes in the study by Clark *et al.*[62] may contain a higher than anticipated fraction of false positives. In addition, increased rates of d_n/d_s along a lineage do not always indicate the action of positive selection and can also occur due to relaxation of purifying selection.[79,80] As the authors point out, obtaining polymorphism data from human populations would provide further insight into the evolutionary history of these genes and help to clarify some of the issues raised above.

Local adaptation

An interesting observation that has consistently emerged from large-scale studies of selection is that local adaptation may be a more common feature of recent human evolutionary history than previously thought.[52–59,60,63–68] Human populations have clearly had dramatic range expansions during the past 100,000 years that, at least theoretically, may have led to geographically restricted selective pressures, such as unique dietary, pathogenic and climatic challenges. Several genes that possess patterns of genetic variation consistent with local adaptation have previously been reported (Table 3).[60,72–74,76,81–85] As an illustrative example, Figure 4 shows patterns of genetic variation for a 115 kilobase region on chromosome 7q33 that possesses a striking signature of local adaptation in European-American populations.[60] Two of the genes in this region, *TRPV5* and *TRPV6*, mediate the rate-limiting step of dietary calcium absorption;[86,87] given the fact that lactase persistence and related metabolic pathways were selected for in northern European populations,[83] they are particularly strong candidates for the gene or genes driving this pattern of local adaptation.

In addition, several studies have found that non-African populations possess more evidence for selection relative to African populations.[60,67,68] As most studies have considered only a single African population, however, it is difficult to determine whether the observed differences in the frequency of selective events between African and non-African populations is a general phenomenon or simply reflects the need to sample African populations more comprehensively. Furthermore, theoretical studies have demonstrated that the power to detect a recent selective sweep is greater compared with an older sweep.[41,42,88,89] Therefore, the frequency of

Table 3. Genes with evidence of local adaptation.

Gene	Potential selective pressure	Reference
MAO-A	Behavioural	[81]
MC1R	Climate	[82]
Lactase	Dietary	[83]
CAPN10	Dietary	[76]
TRPV5/TRPV6	Dietary	[60]
CCR5	Infectious disease	[75,84]
G6PD	Infectious disease	[73]
FY	Infectious disease	[74]
IL4	Infectious disease	[85]
IL1A	Infectious disease	[60]
DCN	Infectious disease	[60]
ACE2	Infectious disease	[60]

selective events may be similar in African and non-African populations, but may be easier to detect in non-African populations if they occurred more recently.

Looking ahead: The HapMap project

The HapMap project (http://www.hapmap.org/) is a large international collaboration to describe patterns of common haplotype variation throughout the human genome.[61] The initial goal of the HapMap project is to genotype 600,000 SNPs in 270 individuals: 90 individuals of northern and western European ancestry (30 trios consisting of two parents and an adult child), 90 Yoruban individuals from Ibadan, Nigeria (30 trios), 45 unrelated Japanese individuals from Tokyo, Japan, and 45 unrelated Han Chinese individuals from Beijing, China. Although the HapMap project was initially developed to facilitate the search for complex disease genes, it will provide a powerful resource for population genetics and evolutionary studies. Specifically, it will provide a unifying publicly-available resource of genome-wide variation data to interrogate systematically for signatures of natural selection. As numerous evolutionary analyses will undoubtedly be

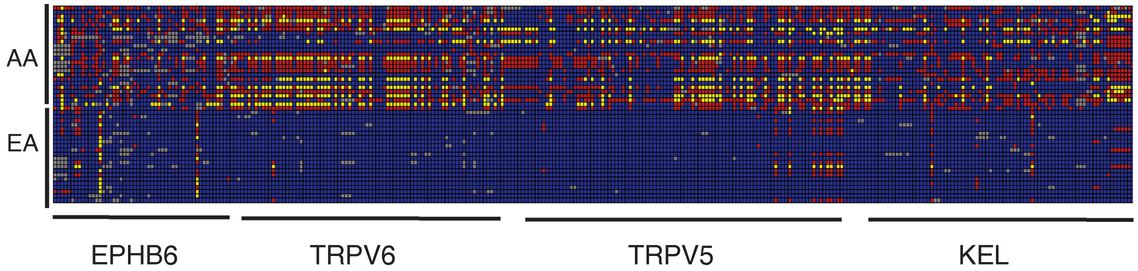

Figure 4. Signature of local adaptation on chromosome 7q33. A graphical representation of genotypes is shown for 23 European-American (EA) and 24 African-Americans (AA) across a 115 kilobase region on chromosome 7q33, which encompasses four genes. Rows correspond to individuals and columns denote a particular single nucleotide polymorphism (SNP). For each SNP, blue, red and yellow boxes indicate whether the individual is homozygous for the common allele, heterozygous or homozygous for the rare allele, respectively. Grey boxes indicate missing data. Notice the significant reduction in polymorphism in the European-American sample, which is consistent with the hypothesis that variation in one or more of these four genes conferred a selective advantage to European-Americans but not African-Americans. See Akey *et al.*[60] for more details. This figure was produced using genotype data from SeattleSNPs project (http://pga.gs.washington.edu/).

conducted on the HapMap data, results can be verified across studies, which will allow prioritising candidate selection genes for subsequent studies.

Future challenges

It is important to temper our enthusiasm for genome-wide scans of natural selection because several analytical and conceptual challenges remain. For example, as indicated above, thousands of hypothesis tests will be performed in a typical study and it is necessary to correct for multiple tests to avoid an unacceptably high false-positive rate. One particularly appealing approach is to control the false discovery rate,[90,91] which is more powerful than traditional methods such as Bonferroni corrections and has been used in a wide variety of genomics analyses. Furthermore, as numerous genome-wide scans for selection will be applied to common datasets, such as the HapMap, methods for combining results across studies would be invaluable.

A critical issue that has already arisen in current genome-wide scans for selection is the need to verify the signature of selection through replication studies and by alternative experimental approaches. The importance of follow-up studies cannot be overstated because in their absence we will simply be left with a list of interesting 'candidate selection genes'. The problem of follow-up replication in genome-wide studies is a general one that has been considered in linkage analysis[92] and genetic association studies.[93] Clearly, replication in independent samples from the same population is an important criterion that can be used to discard false positives that accumulate from the multiple testing inherent in genome scans. Genome-wide study designs are known to suffer from the 'winner's curse' phenomenon, however, whereby the effect sizes of statistically significant loci are systematically over-estimated.[93,94] If such concerns are ignored, the statistical power of subsequent replication attempts is likely to be over-estimated, leading the community to place undue faith in the veracity of failed replication attempts. Even if signatures of selection are confirmed, it remains difficult to identify the specific variants that have been subject to selection. Ideally, suspected targets of selection will be functionally characterised, which will facilitate inferences on genotype – phenotype correlations and ultimately on how the putative selected alleles affect fitness. Finally, more powerful methods to estimate evolutionary parameters, such as the timing of selective events and the strength of selection, need to be developed.

In addition to the issues described above, it is important to note that all of the statistical methods and studies considered in this review are predicated upon simple theoretical models of natural selection. For example, tests such as Tajima's D search for signatures of selection that act on a single locus. Genes do not exist in isolation, however, and it is possible—perhaps even likely—that selection acts on combinations of alleles, a process that is referred to as epistatic selection.[95] Recently, two studies in *Drosophila melanogaster* demonstrated strong empirical evidence for epistatic selection.[96,97] It seems likely that that progress in reconstructing gene and protein networks will serve as a valuable guide in beginning to explore epistatic selection in humans.

Conclusions

The intersection of high-throughput methods to access human genetic variation on a genome-wide scale and statistical tools to identify signatures of natural selection will undoubtedly provide a deeper understanding of how adaptive processes helped to shape our genomes. Furthermore, the same resources used to scan the genome for signatures of selection will also provide a more comprehensive understanding of human demographic history, which will be necessary to understand how neutral and non-neutral evolutionary forces have interacted to shape extant patterns of human genetic and phenotypic diversity. Although many hurdles are likely to be encountered, the evolutionary insights obtained from genome-wide analyses will have implications for many contemporary issues, such as the functional annotation of the human genome and the discovery of complex disease genes.

Acknowledgments

We thank Jennifer Madeoy, Dayna Akey and an anonymous reviewer for critical reading of the manuscript and providing valuable comments. J.R. is supported by the University of Washington Medical Scientist Training Program. J.M.A. is supported by a Pilot and Feasibility Award from the Clinical Nutrition Research Unit at the University of Washington.

References

1. Valle, D. (2004), 'Genetics, individuality, and medicine in the 21st century', *Am. J. Hum. Genet.* Vol. 74, pp. 374–381.
2. Bamshad, M. and Wooding, S.P. (2003), 'Signatures of natural selection in the human genome', *Nat. Rev. Genet.* Vol. 4, pp. 99–111.
3. Harr, B., Kauer, M. and Schlotterer, C. (2002), 'Hitchhiking mapping: A population-based fine mapping strategy for adaptive mutations in *Drosophila melanogaster*', *Proc. Natl. Acad. Sci. USA* Vol. 99, pp. 12949–12954.
4. Kauer, M.O., Dieringer, D. and Schlotterer, C. (2003), 'A microsatellite variability screen for positive selection associated with the "Out of Africa" habitat expansion of *Drosophila melanogaster*', *Genetics* Vol. 165, pp. 1137–1148.
5. Schofl, G. and Schlotterer, C. (2004), 'Patterns of microsatellite variability among X chromosomes and autosomes indicate a high frequency of beneficial mutations in non-African *D. simulans*', *Mol. Biol. Evol.* Vol. 21, pp. 1384–1390.
6. Kimura, M. (1968), 'Evolutionary rate at the molecular level', *Nature* Vol. 217, pp. 624–626.
7. King, J.L. and Jukes, T.H. (1969), 'Non-Darwinian evolution', *Science* Vol. 164, pp. 788–798.
8. Kimura, M. (1983), *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, UK.
9. Ohta, T. (1973), 'Slightly deleterious mutant substitutions in evolution', *Nature* Vol. 246, pp. 96–98.
10. Ohta, T. and Gillespie, J.H. (1996), 'Development of neutral and nearly neutral theories', *Theor. Popul. Biol.* Vol. 49, pp. 128–142.

11. Otto, S.P. (2000), 'Detecting the form of selection from DNA sequence data', *Trends Genet.* Vol. 16, pp. 526–529.
12. Nielsen, R. (2001), 'Statistical tests of selective neutrality in the age of genomics', *Heredity* Vol. 86, pp. 641–647.
13. Kingman, J.F.C. (1982), 'The coalescent', *Stochastic Process Appl.* Vol. 13, pp. 235–248.
14. Kingman, J.F.C. (1982), 'On the genealogy of large populations', *J. Appl. Prob.* Vol. 19A, pp. 27–43.
15. Hudson, R.R. (1983), 'Properties of a neutral allele model with intragenic recombination', *Theor. Popul. Biol.* Vol. 23, pp. 183–201.
16. Hudson, R.R. (1983), 'Testing the constant-rate neutral allele model with protein sequence data', *Evolution* Vol. 37, pp. 203–217.
17. Tajima, F. (1983), 'Evolutionary relationship of DNA sequences in finite populations', *Genetics* Vol. 105, pp. 437–460.
18. Fu, Y.X. and Li, W.H. (1999), 'Coalescing into the 21st century: An overview and prospects of coalescent theory', *Theor. Popul. Biol.* Vol. 56, pp. 1–10.
19. Rosenberg, N.A. and Nordborg, M. (2002), 'Genealogical trees, coalescent theory and the analysis of genetic polymorphisms', *Nat. Rev. Genet.* Vol. 3, pp. 380–390.
20. Charlesworth, B., Morgan, M.T. and Charlesworth, D. (1993), 'The effect of deleterious mutations on neutral molecular variation', *Genetics* Vol. 134, pp. 1289–1303.
21. Hudson, R.R. and Kaplan, N.L. (1995), 'Deleterious background selection with recombination', *Genetics* Vol. 141, pp. 1605–1617.
22. Neuhauser, C. and Krone, S.K. (1997), 'The genealogy of samples in models with selection', *Genetics* Vol. 145, pp. 519–534.
23. Maynard Smith, J. and Haigh, J. (1974), 'The hitch-hiking effect of a favorable gene', *Genet. Res.* Vol. 231, pp. 1114–1116.
24. Thomson, G. (1977), 'The effect of a selected locus on a linked neutral locus', *Genetics* Vol. 85, pp. 752–788.
25. Kaplan, N., Hudson, R.R. and Langley, C.H. (1989), 'The "hitchhiking effect" revisited', *Genetics* Vol. 123, pp. 887–899.
26. Stephan, W., Wiehe, T.H.E. and Lenz, M.W. (1992), 'The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory', *Theor. Popul. Biol.* Vol. 41, pp. 237–254.
27. Nordborg, M. (1997), 'Structured coalescent processes on different time scales', *Genetics* Vol. 146, pp. 1501–1514.
28. Schierup, M.H., Vekemans, X. and Charlesworth, D. (2000), 'The effect of subdivision on variation at multi-allelic loci under balancing selection', *Genet. Res.* Vol. 76, pp. 51–62.
29. Kelly, J.K. and Wade, M.J. (2000), 'Molecular evolution near a two-locus balanced polymorphism', *J. Theor. Biol.* Vol. 204, pp. 83–101.
30. Nordborg, M. and Innan, H. (2003), 'The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population', *Genetics* Vol. 163, pp. 1201–1213.
31. Neilsen, R. (2000), 'Estimation of population parameters and recombination rates from single nucleotide polymorphisms', *Genetics* Vol. 154, pp. 931–942.
32. Braverman, J.M., Hudson, R.R., Kaplan, N.L. *et al.* (1995), 'The hitchhiking effect on the site frequency spectrum of DNA polymorphism', *Genetics* Vol. 140, pp. 783–796.
33. Fay, J.C. and Wu, C.I. (2000), 'Hitchhiking under positive Darwinian selection', *Genetics* Vol. 155, pp. 1405–1413.
34. Sabeti, P.C., Reich, D.E., Higgins, J.M. *et al.* (2002), 'Detecting recent positive selection in the human genome from haplotype structure', *Nature* Vol. 419, pp. 832–837.
35. Takahata, N. and Nei, M. (1990), 'Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci', *Genetics* Vol. 124, pp. 967–978.
36. Tajima, F. (1989), 'The effect of change in population size on DNA polymorphism', *Genetics* Vol. 123, pp. 597–601.
37. Przeworski, M., Hudson, R.R. and Di Rienzo, A. (2000), 'Adjusting the focus on human variation', *Trends Genet.* Vol. 16, pp. 296–302.
38. Kreitman, M. (2000), 'Methods to detect selection in populations with applications to the human', *Annu. Rev. Genomics Hum. Genet.* Vol. 1, pp. 539–559.
39. Tajima, F. (1989), 'Statistical method for testing the neutral mutation hypothesis by DNA polymorphism', *Genetics* Vol. 123, pp. 585–595.
40. Fu, Y.X. and Li, W.H. (1993), 'Statistical test of neutrality of mutations', *Genetics* Vol. 133, pp. 693–709.
41. Fu, Y.X. (1997), 'Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection', *Genetics* Vol. 186, pp. 1997–2004.
42. Simonsen, K.L., Churchill, G.A. and Aquadro, C.F. (1995), 'Properties of statistical tests of neutrality for DNA polymorphism data', *Genetics* Vol. 141, pp. 413–429.
43. Kuhner, M.K., Yamato, J. and Felsenstein, J. (1995), 'Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling', *Genetics* Vol. 140, pp. 1421–1430.
44. Kuhner, M.K., Yamato, J. and Felsenstein, J. (1998), 'Maximum likelihood estimation of population growth rates based on the coalescent', *Genetics* Vol. 149, pp. 429–434.
45. Felsenstein, J. (2004), 'Likelihood calculations on coalescents', in: Felsenstein, J. (ed.), *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA, pp. 470–487.
46. Cavalli-Sforza, L.L. (1966), 'Population structure and human evolution', *Proc. R. Soc. Lond. B Biol. Sci.* Vol. 164, pp. 362–379.
47. Lewontin, R.C. and Krakauer, J. (1973), 'Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms', *Genetics* Vol. 74, pp. 175–195.
48. Weir, B.S. and Cockerham, C.C. (1984), 'Estimating F-statistics for the analysis of population structure', *Evolution* Vol. 38, pp. 1358–1370.
49. Vitalis, R., Dawson, K. and Boursot, P. (2001), 'Interpretation of variation across marker loci as evidence of selection', *Genetics* Vol. 158, pp. 1811–1823.
50. Beaumont, M.A. and Balding, D.J. (2004), 'Identifying adaptive genetic divergence among populations from genome scans', *Mol. Ecol.* Vol. 13, pp. 969–980.
51. Hudson, R.R., Kreitman, M. and Aguade, M. (1987), 'A test of neutral molecular evolution based on nucleotide data', *Genetics* Vol. 116, pp. 153–159.
52. McDonald, J.H. and Kreitman, M. (1991), 'Adaptive protein evolution at the Adh locus in Drosophila', *Nature* Vol. 351, pp. 652–654.
53. McDonald, J.H. (1998), 'Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence', *Mol. Biol. Evol.* Vol. 15, pp. 377–384.
54. Eyre-Walker, A. (2002), 'Changing effective population size and the McDonald-Kreitman test', *Genetics* Vol. 162, pp. 2017–2024.
55. Nielsen, R. and Yang, Z. (1998), 'Likelihood models for detecting positive selected amino acid sites and applications to the HIV-1 envelope gene', *Genetics* Vol. 148, pp. 929–936.
56. Nei, M. and Gojobori, T. (1986), 'Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions', *Mol. Biol. Evol.* Vol. 3, pp. 418–426.
57. Suzuki, Y. and Gojobori, T. (1999), 'A method for detecting positive selection at single amino acid sites', *Mol. Biol. Evol.* Vol. 16, pp. 1315–1328.
58. Andolfatto, P. (2001), 'Adaptive hitchhiking effects on genome variability', *Curr. Opin. Genet. Dev.* Vol. 11, pp. 635–641.
59. Hudson, R.R. (2002), 'Generating samples under a Wright-Fisher neutral model of genetic variation', *Bioinformatics* Vol. 18, pp. 337–338.
60. Akey, J.M., Eberle, M.A., Rieder, M.J. *et al.* (2004), 'Population history and natural selection shape patterns of genetic variation in 132 genes', *PLoS Biol.* Vol. 2, pp. 1591–1599.
61. International HapMap Consortium(2003), 'The international HapMap project', *Nature* Vol. 426, pp. 789–794.
62. Clark, A.G., Glanowski, S., Nielsen, R. *et al.* (2003), 'Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios', *Science* Vol. 302, pp. 1960–1963.
63. Yang, Z. and Nielsen, R. (2000), 'Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models', *Mol. Biol. Evol.* Vol. 17, pp. 32–43.

64. Yang, Z. (1997), 'PAML: A program package for phylogenetic analysis by maximum likelihood', *Comput. Appl. BioSci.* Vol. 13, pp. 555–556.
65. Akey, J.M., Zhang, G., Zhang, K. *et al.* (2002), 'Interrogating a high-density SNP map for signatures of natural selection', *Genome Res.* Vol. 12, pp. 1805–1814.
66. Payseur, B.A., Cutter, A.D. and Nachman, M.W. (2002), 'Searching for evidence of positive selection in the human genome using patterns of microsatellite variability', *Mol. Biol. Evol.* Vol. 19, pp. 1143–1153.
67. Kayser, M., Brauer, S. and Stoneking, M. (2003), 'A genome scan to detect candidate regions influenced by local natural selection in human populations', *Mol. Biol. Evol.* Vol. 20, pp. 893–900.
68. Storz, J.F., Payseur, B.A. and Nachman, M.W. (2004), 'Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa', *Mol. Biol. Evol.* Vol. 21, pp. 1800–1811.
69. Schlötterer, C. (2002), 'A microsatellite-based multilocus screen for the identification of local selective sweeps', *Genetics* Vol. 160, pp. 753–763.
70. Schlötterer, C. and Wiehe, T. (1999), 'Microsatellites, a neutral marker to infer selective sweeps', in: Goldstein, D., Schlötterer, C. (eds.), *Microsatellites — Evolution and Applications*, Oxford University Press, Oxford, UK, pp. 238–248.
71. Wiehe, T. (1998), 'The effect of selective sweeps on the variance of the allele distribution of a linked multi-allele locus-hitchhiking of microsatellites', *Theor. Popul. Biol.* Vol. 53, pp. 272–283.
72. Hamblin, M.T. and Di Rienzo, A. (2000), 'Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus', *Am. J. Hum. Genet.* Vol. 66, pp. 1669–1679.
73. Tishkoff, S.A., Varkonyi, R., Cahinhinan, N. *et al.* (2001), 'Haplotype diversity and linkage disequilibrium at human G6PD: Recent origin of alleles that confer malarial resistance', *Science* Vol. 293, pp. 455–462.
74. Hamblin, M.T., Thompson, E.E. and Di Rienzo, A. (2002), 'Complex signatures of natural selection at the Duffy blood group locus', *Am. J. Hum. Genet.* Vol. 70, pp. 369–383.
75. Bamshad, M.J., Mummidi, S., Gonzalez, E. *et al.* (2002), 'A strong signature of balancing selection in the 5′ *cis*-regulatory region of CCR5', *Proc. Natl. Acad. Sci. USA* Vol. 99, pp. 10539–10544.
76. Fullerton, S.M., Bartoszewicz, A., Ybazeta, G. *et al.* (2002), 'Geographic and haplotype structure of candidate type 2 diabetes susceptibility variants at the calpain-10 locus', *Am. J. Hum. Genet.* Vol. 70, pp. 1096–1106.
77. Rockman, M.V., Hahn, M.W., Soranzo, N. *et al.* (2004), 'Positive selection on MMP3 regulation has shaped heart disease risk', *Curr. Biol.* Vol. 14, pp. 1531–1539.
78. Nakajima, T., Wodding, S., Sakagami, T. *et al.* (2004), 'Natural selection and population history in the human angiotensinogen gene (AGT): 736 complete ATG sequences in chromosomes from around the world', *Am. J. Hum. Genet.* Vol. 74, pp. 898–916.
79. Zhang, J. (2004), 'Frequent false detection of positive selection by the likelihood method with branch-site models', *Mol. Biol. Evol.* Vol. 21, pp. 1332–1339.
80. Rooney, A.P. and Zhang, J. (1999), 'Rapid evolution of a primate sperm protein: Relaxation of functional constraint or positive Darwinian selection?', *Mol. Biol. Evol.* Vol. 16, pp. 706–710.
81. Gilad, Y., Rosenberg, S., Przeworski, M. *et al.* (2002), 'Evidence for positive selection and population structure at the human MAO-A gene', *Proc. Natl. Acad. Sci. USA* Vol. 99, pp. 862–867.
82. Rana, B.K., Hewett-Emmett, D., Jin, L. *et al.* (1999), 'High polymorphism at the human melanocortin 1 receptor locus', *Genetics* Vol. 151, pp. 1547–1557.
83. Bersaglieri, T., Sabeti, P.C., Patterson, N. *et al.* (2004), 'Genetic signatures of strong recent positive selection at the lactase gene', *Am. J. Hum. Genet.* Vol. 74, pp. 1111–1120.
84. Stephens, J.C., Reich, D.E., Goldstein, D.B. *et al.* (1998), 'Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes', *Am. J. Hum. Genet.* Vol. 62, pp. 1507–1515.
85. Rockman, M.V., Hahn, M.W., Soranzo, N. *et al.* (2003), 'Positive selection on a human-specific transcription factor binding site regulating IL4 expression', *Curr. Biol.* Vol. 13, pp. 2118–2123.
86. Nijenhuis, T., Hoenderop, J.G.J., Nilius, B. and Bindels, R.J.M. (2003), '(Patho)physiological implications of the novel epithelial $Ca2^+$ channels TRPV5 and TRPV6', *Pflugers Arch.* Vol. 446, pp. 401–409.
87. van de Graaf, S.F., Hoenderop, J.G., Gkika, D. *et al.* (2003), 'Functional expression of the epithelial $Ca2^+$ channels (TRPV5 and TRPV6) requires association of the S100A10-annexin 2 complex', *EMBO J.* Vol. 22, pp. 1478–1487.
88. Kim, Y. and Stephan, W. (2000), 'Joint effects of genetic hitchhiking and background selection on neutral variation', *Genetics* Vol. 155, pp. 1415–1427.
89. Przeworski, M. (2002), 'The signature of positive selection at randomly chosen loci', *Genetics* Vol. 160, pp. 1179–1189.
90. Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *J.R. Stat. Soc.* Vol. 57, pp. 289–300.
91. Storey, J.D. and Tibshirani, R. (2003), 'Statistical significance for genome-wide experiments', *Proc. Nat. Acad. Sci. USA* Vol. 100, pp. 9440–9445.
92. Lander, E. and Kruglyak, L. (1995), 'Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results', *Nat. Genet.* Vol. 11, pp. 241–247.
93. Lohmueller, K.E., Pearce, C.L., Pike, M. *et al.* (2003), 'Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease', *Nat. Genet.* Vol. 33, pp. 177–182.
94. Goring, H.H., Terwilliger, J.D. and Blangero, J. (2001), 'Large upward bias in estimation of locus-specific effects from genomewide scans', *Am. J. Hum. Genet.* Vol. 69, pp. 1357–1369.
95. Lewontin, R.C. and Kojima, K. (1960), 'The evolutionary dynamics of complex polymorphisms', *Evolution* Vol. 14, pp. 458–472.
96. Takano-Shimizu, T., Kawabe, A., Inomata, N. *et al.* (2004), 'Interlocus nonrandom association of polymorphisms in *Drosophila* chemoreceptor genes', *Proc. Natl. Acad. Sci. USA* Vol. 101, pp. 14156–14161.
97. Zapata, C., Nunez, C. and Velasco, T. (2002), 'Distribution of nonrandom associations between pairs of protein loci along the third chromosome of *Drosophila melanogaster*', *Genetics* Vol. 161, pp. 1539–1550.