Review TRENDS in Genetics Vol.22 No.8 August 2006

Genomic insights into positive selection

Shameek Biswas and Joshua M. Akey

Department of Genome Sciences, University of Washington, 1705 NE Pacific, Seattle, WA 98195, USA

The traditional way of identifying targets of adaptive evolution has been to study a few loci that one hypothesizes *a priori* to have been under selection. This approach is complicated because of the confounding effects that population demographic history and selection have on patterns of DNA sequence variation. In principle, multilocus analyses can facilitate robust inferences of selection at individual loci. The deluge of large-scale catalogs of genetic variation has stimulated many genome-wide scans for positive selection in several species. Here, we review some of the salient observations of these studies, identify important challenges ahead, consider the limitations of genome-wide scans for selection and discuss the potential significance of a comprehensive understanding of genomic patterns of selection for disease-related research.

The benefits of studying positive selection

An enduring goal of evolutionary biology is to understand the forces that govern how populations and species evolve. In terms of molecular evolution, this problem has often been framed in elucidating the relative contributions of genetic drift and natural selection to extant patterns of genetic variation [1,2]. More specifically, the neutral theory posits that most polymorphisms are either neutral or slightly deleterious and changes in allele frequency are primarily governed by the stochastic effects of genetic drift in populations of finite size [3–5]. Thus, the effective population size, N_e, and neutral mutation rate, μ_o, determine the levels of polymorphism within species and the rate of divergence between species, respectively. An alternative view is that a significant proportion of variation does affect the ability of an organism to survive and reproduce and will therefore be subject to natural selection [1].

The continuing development of large-scale catalogs of genetic variation has stimulated renewed interest in finding targets of positive selection, which ultimately will help to clarify the roles of drift and selection in evolutionary processes. Furthermore, signatures of positive 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 diversity and help to annotate the genome functionally. In addition to these more utilitarian benefits, each target of positive selection has a story to tell about the historical forces and events that have shaped the history of a population.

Several genome-wide analyses for positive selection have been performed in a variety of species. In this review, we summarize some of the recent studies, primarily focusing on humans, critically evaluate what genome-wide scans for selection are and are not likely to find and suggest future avenues of research. A brief overview of statistical methods used to detect deviations from neutrality is summarized in Box 1. For more detailed discussions, see Refs [6,7].

Thinking genomically

Positive selection perturbs patterns of genetic variation relative to what is expected under a standard neutral model. For example, signatures of positive selection

Glossary

Ascertainment bias: a systematic bias in measuring the true frequency of a phenomenon due to the way in which the data are collected. Ascertainment bias introduced by how genetic variation was discovered is of particular interest to genome-wide analyses of polymorphism. For example, many human SNPs were discovered in chromosomes from a few individuals, which resulted in an over-representation of common alleles and a deficit of low frequency alleles.
Balancing selection: a general type of selective regime in which multiple alleles at a locus are maintained in a population. Specific types of balancing selection include heterozygote advantage and frequency dependent selection.
Purifying selection: the removal of deleterious alleles from a population.
Background selection: describes the case where purifying selection against deleterious mutations also removes variation at linked neutral sites.
Effective population size: the size of an idealized Wright–Fisher population that contains the same amount of genetic drift observed in the actual population under consideration. Note that there are different types of effective population size depending on how the deviation from an ideal population is quantified, such as inbreeding effective size, variance effective size and eigenvalue effective size.
Genetic drift: the random change in allele and haplotype frequencies in populations of finite size.
Genetic hitchhiking: the influence of a beneficial mutation on patterns of linked neutral variation.
Linkage disequilibrium: The non-random association of alleles between two or more loci. Although it is commonly used to measure correlations between linked loci, linkage disequilibrium can also form between unlinked loci for many reasons including population structure, admixture and epistatic selection.
Nucleotide polymorphism, θ_w: a measure of DNA sequence variation based on the observed number of segregating sites and number of chromosomes in a sample.
Nucleotide diversity (or heterozygosity), π: a measure of DNA sequence variation based on the average pairwise distance between all sequences in the sample.
Population bottleneck: a reduction in population size that increases the effects of drift and introduces skews the allele frequency spectrum of polymorphisms. The effect of a bottleneck on patterns of genetic variation depends on how severe the decrease in population size is and the duration of the bottleneck.
Standard neutral model: denotes an idealized constant-size, neutrally evolving and randomly mating population at mutation-drift equilibrium.

Corresponding author: Akey, J.M. (akeyj@u.washington.edu).

 0168-9525/$ - see front matter doi:10.1016/j.tig.2006.06.005

Box 1. Methods and approaches to detect positive selection

Tests based on polymorphisms within species

Tajima's D: this statistic measures the difference between two estimators of the population mutation rate, θ_w and π [53]. Under neutrality, the means of θ_w and π should be approximately equal to one another. Therefore, the expected value of Tajima's D for populations conforming to a standard neutral model is zero. Significant deviations from zero indicate a skew in the allele frequency distribution relative to neutral expectations. Positive values of Tajima's D arise from an excess of intermediate frequency alleles and can result from population bottlenecks, structure and/or balancing selection. Negative values of Tajima's D indicate an excess of low frequency alleles and can result from population expansions or positive selection.

Fu and Li's D and F: this set of tests is similar to Tajima's D in that it tests for a skew in the allele frequency spectrum, but makes the distinction between old and recent mutations as determined by where they occur on the branches of genealogies. The D and F statistics compare an estimate of the population mutation rate based on the number of derived variants seen only once in a sample (referred to as singletons) with θ_w or π, respectively [54]. Similar to Tajima's D, the expected value of D and F is zero, and both positive and negative deviations are informative about distinct demographic and/or selective events.

Fay and Wu's H test: a statistic that detects the presence of an excess of high frequency derived alleles in a sample, which is a hallmark of positive selection [55].

Long range haplotype (LRH) test: this test examines the relationship between allele frequency and the extent of LD [23]. Positive selection is expected to accelerate the frequency of an advantageous allele faster than recombination can break down LD at the selected haplotype. Thus, a hallmark of recent positive selection is an allele that has greater long-range LD given its frequency in the population relative to neutral expectations. To capture this signature, the LRH test begins by selecting a 'core' haplotype (note this could also be applied to a single SNP). Next, the decay in LD is assessed for flanking markers by calculating EHH, which is defined as the probability that two randomly chosen chromosomes carrying the core SNP or haplotype are identical by descent. For each core, haplotype homozygosity is initially 1 and decays to 0 at increasing distances. Positive selection is formally tested by finding core haplotypes that have elevated EHH relative to other core haplotypes at the locus conditional on haplotype frequency. By focusing on relative EHH, the various core haplotypes control for local rates of recombination.

iHS: this statistic is applied to individual SNPs and begins by calculating the integrated EHH (iHH), which is defined as the integral of the observed decay of EHH (i.e. the area under the curve of EHH versus distance) away from a specified core allele until EHH reaches 0.05 [22]. The log ratio of iHH for the ancestral and derived alleles is then standardized such that it has a mean of 0 and variance of 1 irrespective of allele frequency at the core SNP. Large positive and negative values of iHS indicate unusually long haplotypes carrying the ancestral and derived allele, respectively.

LD decay (LDD): The goal of this test, similar in spirit to the iHH and EHH statistics, is to detect large differences in the extent of LD between two alleles at a particular locus [21]. The test begins by identifying individuals who are homozygous for the SNP being considered, which eliminates the need to infer haplotypes. Individuals are then sorted according to whether they are homozygous for the major or minor allele. The fraction of recombinant chromosomes (FRC) is computed for all adjacent SNPs within an *a priori* defined window and the FRC and distance from the target SNP are then used to calculate an ALnLH statistic. SNPs with high ALnLH values imply that the decay in LD for one allele is unusual compared with that of the alternative allele, in which the pattern of LD decay is within an *a priori* defined bound of the genome-wide average.

F_{ST}: a statistic that quantifies levels of differentiation between subpopulations [56]. Many estimators of F_{ST} have been proposed, but a conceptually simple one is $(H_T\text{-}H_S)/H_T$. Here H_T is an estimate of total heterozygosity and H_S is a measure of the average heterozygosity across subpopulations. Thus, one way to interpret F_{ST} is the reduction in heterozygosity among subpopulations relative to what is expected under random mating. Under neutrality, levels of F_{ST} are largely determined by genetic drift and migration, but local adaptation can accentuate levels of population differentiation at particular loci thus resulting in large F_{ST} values.

Tests based on polymorphisms within species and the divergence between species

Hudson–Kreitman–aguade (HKA) test: the neutral theory predicts a positive correlation between levels of polymorphism within species and divergence between species [57]. The HKA test is used to determine if levels of nucleotide variation within and between species at two or more loci conform to this expectation. A significant HKA test can thus be caused by increased levels of polymorphism at one locus or reduced levels of polymorphism at the other, or by excess divergence at one locus or limited divergence at the other.

McDonald Kreitman (MK) test: this test also makes use of polymorphism and divergence data, but compares different types of mutations, such as synonymous versus non-synonymous sites at a specific locus [58]. 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 $P_N/P_S = D_N/D_S$, whereas positive selection leads to an increase in non-synonymous divergence ($D_N/D_S > P_N/P_S$).

Tests between species

d_n/d_s test: in the simplest forms, the ratio of non-synonymous (d_n) to synonymous (d_s) substitutions is compared in protein coding loci [59–61]. The d_n/d_s ratio provides information about the evolutionary forces operating on a particular gene. For example, under neutrality $d_n/d_s = 1$. For genes that are subject to functional constraint such that non-synonymous amino acid substitutions are deleterious and purged from the population, $d_n/d_s < 1$. For positively selected genes, $d_n/d_s > 1$. Although the observation of $d_n/d_s > 1$ provides strong evidence for positive selection, it is conservative if only a few sites have been targets of adaptive evolution. The basic d_n/d_s test has been extended to include models of codon and transition and/or transversion bias, to detect variation in d_n/d_s ratios among lineages and to identify specific sites that might be under selection.

might include a skew in the allele frequency distribution (i.e. an excess of low and/or high frequency derived alleles), reduced levels of genetic variation and elevated levels of linkage disequilibrium (LD) relative to neutral expectations (Figure 1) [8,9]. However, population demographic history can impart similar patterns on DNA sequence variation, making inferences of selection difficult. For example, population expansions can also lead to an excess of low frequency alleles compared with the number expected under the standard neutral model. Thus, when a single locus is studied, distinguishing between the confounding effects of natural selection and population demographic history is challenging. One potential solution to this conundrum is to study numerous loci spanning the genome to identify loci that possess patterns of genetic variation that are unusual relative to the rest of the genome. Specifically, population demographic history affects patterns of variation at all loci in a genome, whereas natural selection acts on specific loci [10]. Thus, identifying genes that seem exceptionally unusual (referred to as outliers) compared with numerous loci is an intuitively appealing approach for detecting genes subject to selection.

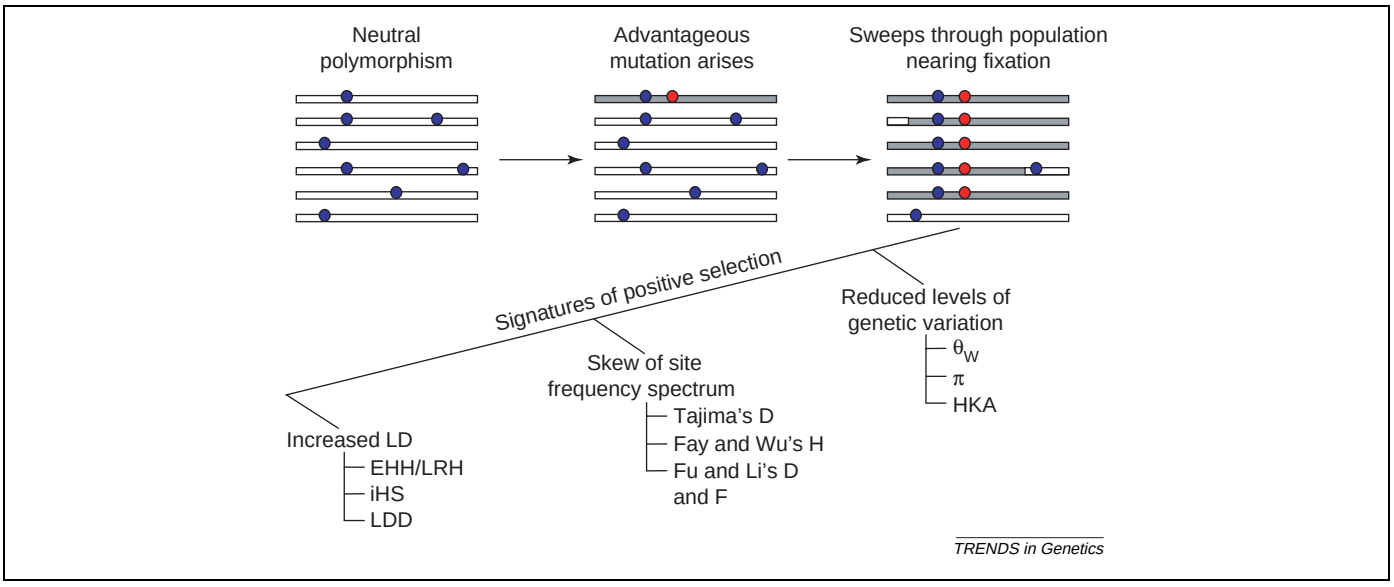

Figure 1. Signatures of positive selection. On the left, patterns of neutral polymorphism (denoted as blue circles) are shown for a sample of six haplotypes. A new advantageous mutation (indicated by the red circle) arises on a specific haplotype (middle panel highlighted in gray). As the advantageous allele increases in frequency it drags along linked neutral polymorphisms. On the right, an incomplete selective sweep is shown such that the advantageous allele has not yet reached fixation. This process perturbs patterns of genetic variation relative to neutral expectations and imparts signatures such as reduced levels of genetic variation, a skew in the site frequency spectrum (also referred to as allele frequency distribution), and increased levels of LD. Recombination between haplotypes carrying and not carrying the advantageous allele delimit the region over which the signature of selection extends. Commonly used summary statistics that have been proposed to test for these signatures are also indicated and described in more detail in Box 1. Note that the relative magnitude of these signatures of positive selection depend on many parameters such as when the advantageous allele arose, the strength of selection, whether the sweep is ongoing or has reached fixation, the amount of time that has elapsed since fixation, and local rates of recombination and mutation.

As a motivating example, Figure 2 shows the distribution of Tajima's D versus nucleotide diversity for 259 genes distributed across 22 autosomes and the X chromosome that were resequenced in 24 African–American and 23 European–American individuals by the SeattleSNPs program (http://pga.gs.washington.edu). A cursory glance of Figure 2 reveals two important points. First, the distribution of Tajima's D in both the African–American and European–American sample is extremely variable. Coalescent theory provides a ready explanation

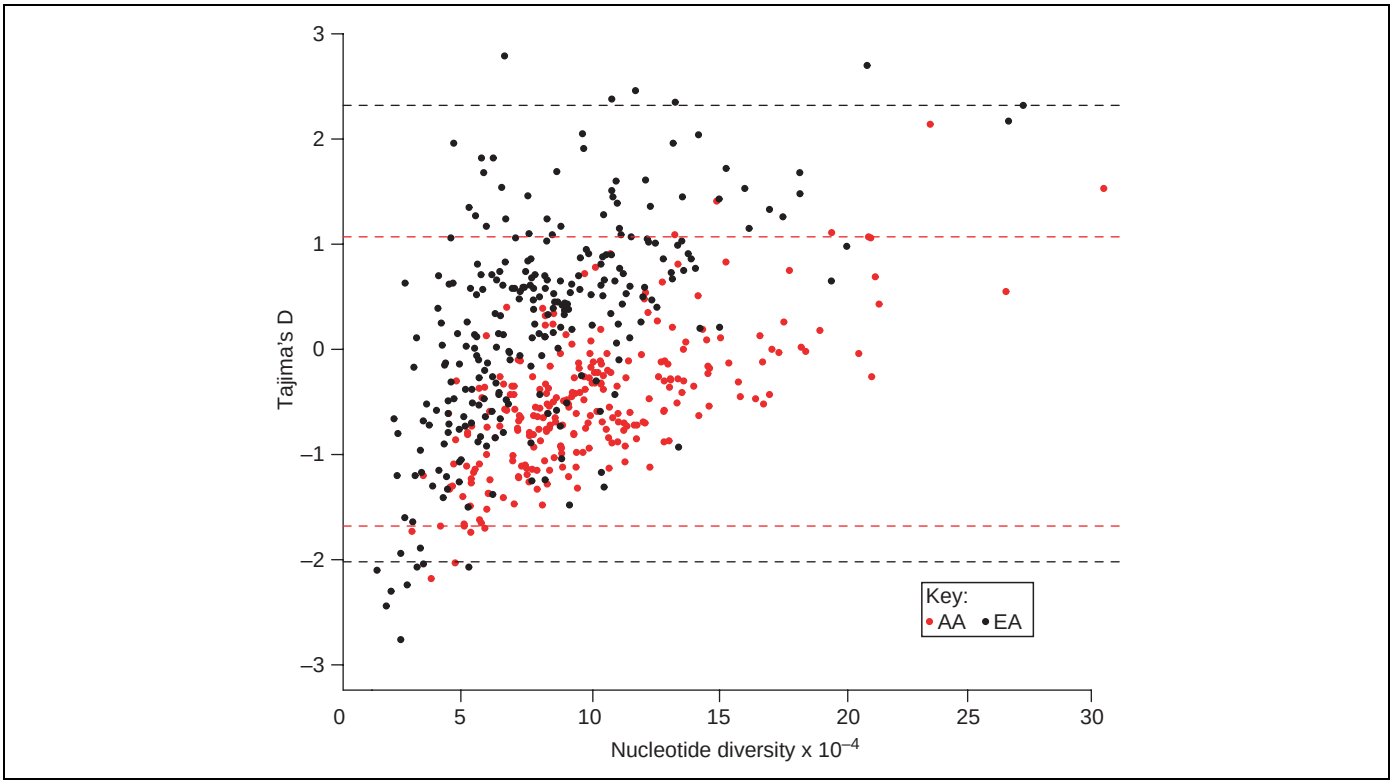

Figure 2. The distribution of Tajima's D versus nucleotide diversity for 259 genes. Each circle denotes a gene and black and red circles correspond to European–American (EA) and African–American (AA) samples, respectively. The dashed lines correspond to the upper and lower 2.5th percentiles of the empirical distribution of Tajima's D for each sample (black, European-American; red, European-American). Tajima's D varies extensively both within each population sample and between EA and AA populations.

Box 2. Coalescent interpretations of nucleotide variation

The coalescent is a stochastic model of gene genealogies that has become the central theoretical tool in population genetics for understanding, interpreting and simulating genetic variation [62–64]. The coalescent process starts with *n* lineages. As the process is followed backwards in time, lineages merge, or coalesce at distinct time intervals. This process continues until a single lineage exists, referred to as the most recent common ancestor (MRCA). Patterns of genetic variation are shaped by two stochastic events: (i) the history of coalescent events, which can be modeled as exponential random variables; and (ii) the history of mutational events, which can be modeled as poisson random variables. Under neutrality, mutations are uniformly distributed along the branches of a genealogy, and therefore the number of mutations occurring on a branch is proportional to its length (i.e. time to coalescence). Because of recombination, different regions of the genome can have distinct gene genealogies. To illustrate these points, Figure I shows two genealogies for a sample of six sequences simulated from a standard neutral model assuming $\theta = 4N_e\mu = 2$. Mutations are indicated by solid circles. Notice that the history of coalescent events and number of mutations (S) differs for each genealogy, which leads to distinct patterns of genetic variation in sampled sequences that are present at the tips of the branches, as indicated by the statistic Tajima's D.

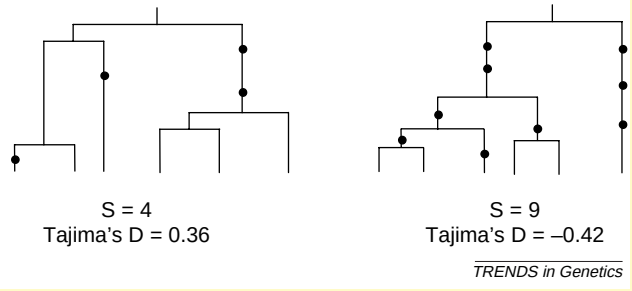

Figure I. Using the coalescent model to understand genetic variation.

for such variability. Specifically, patterns of polymorphism at a locus depend on the stochastic nature of both coalescent and mutational events (Box 2) and therefore locus-to-locus variability is expected. Second, the distribution of Tajima's D is considerably different in African–Americans compared with that in European–Americans, with an average of 0.22 and −0.50, respectively. These differences potentially reflect distinct demographic histories, distinct selective histories or a combination of both [11]. In Figure 2, there are clearly genes that have patterns of genetic variation that are unusual compared with those of other loci, but are they targets of selection or simply the extreme cases of a neutral process? The enthusiasm for genome-wide analyses of genetic variation is predicated on the hope that if variation resulting from demographic perturbations (e.g. genetic drift) can be properly accounted for then genes with robust signatures of natural selection will be revealed.

Genome-wide scans for recent positive selection within humans

With the recent publication of two large-scale studies of human genetic variation [12,13], humans are arguably becoming the best 'model organism' for genome-wide scans of positive selection. Perlegen Biosciences (http://www.perlegen.com) genotyped >1.58 million single nucleotide polymorphisms (SNPs) in 71 individuals of African, Asian or European descent [12]. In addition, the HapMap project (http://www.hapmap.org) recently completed Phase I, which entailed genotyping 210 individuals from four populations (Northern and Western Europeans, Japanese, Han Chinese and Yorubans) for ~600 000 SNPs [13]. This catalog has expanded to >3 million SNPs (as of May 2006). These resources have already been used to interrogate the genome systematically for signatures of selection using various approaches that attempt to capture distinct features of positive selection.

For example, using the Perlegen data set Carlson *et al.* [14] divided the genome into 100-Kb bins and performed a sliding window analysis to identify large regions that had a skew in the allele frequency distribution towards low frequency alleles using the statistic Tajima's D. In total, 7, 23 and 29 regions of the genome, encompassing 176 genes, were found to be extreme outliers in the empirical distribution in the African–American, European–American and Chinese–American samples, respectively (outliers were heuristically defined as 'regions of 20 contiguous windows where >75% of the windows were in the lower 1% of the empirical Tajima's distribution' [14]). Interestingly, twelve of these regions did not contain any known or predicted genes. Only four of the outlier regions overlapped between populations, suggesting that these loci were subject to varying selective histories in each population. Extensive resequencing of eight genes from six of the outlier regions further substantiated the evidence for positive selection and all regions possessed a significant skew in the allele frequency spectrum as assessed by Tajima's D and Fay and Wu's H. Although this analysis provides intriguing insights into regions of the genome that might have experienced strong and recent selection, no attempts were made to determine if the number of observed outliers was greater than that expected under neutrality. Furthermore, is unclear how sensitive the results are to the ascertainment bias present in the Perlegen data and to different definitions of outlier regions (e.g. window size and length of overlap between windows).

Several analyses have also been performed to detect selection based on allele frequency differences between populations. For example, in the HapMap publication [13], 926 SNPs were identified with allele frequency differences between populations even greater than that observed for the *FY* gene (also known as *DARC*, a chemokine receptor expressed on the surface of erythrocytes), which is a well-accepted example of selection where the *FY*O* allele is at or near fixation in sub-Saharan African populations but is rare in non-African populations [15]. Interestingly, 32 of the 926 extremely differentiated loci were non-synonymous SNPs, including rs1426654 in *SLC24A5* (encoding a member of the potassium-dependent sodium and calcium exchanger family), which has recently been shown to be a major determinant of skin pigmentation variation between European and African individuals [16]. Reduced levels of genetic diversity and allele frequency differences between populations were also observed at SNPs adjacent to *SLC24A5*, as expected for positive

selection. Another gene identified in this analysis was *ALMS1*, which has six non-synonymous SNPs with large allele frequency differences between populations. Mutations in *ALMS1* cause Alstrom syndrome, a rare recessive disease with a phenotype that includes obesity, insulin resistance and type 2 diabetes [17]. Thus, follow-up studies of these highly differentiated regions might provide significant insight into phenotypic diversity within and between human populations.

Weir *et al.* [18] also analyzed the genomic distribution of population structure in the HapMap and Perlegen data using a statistic that is closely related to the conventional F_{ST} (which measures population structure) [19]. They showed that levels of structure vary considerably throughout the genome, potentially indicating the action of positive selection. Importantly, however, the authors show that careful attention must be made to the inherent stochastic variation in single locus F_{ST} values if levels of population structure are to be used for identifying putative targets of selection, and argue that sliding window analyses can reduce locus-to-locus variation and facilitate inferences of selection. To this end, the average F_{ST} in 5-Mb sliding windows was calculated and regions containing putative targets of positive selection were defined as those that were three standard deviations above the chromosomal mean. In total, ~300 candidate selection regions were identified including the region containing the lactase gene, which is a well-documented example of positive selection in individuals of European descent [20]. Although sliding window analyses are appealing, as noted earlier, the size of the window is often subjectively determined, which can influence the final results and interpretations. Weir *et al.* [18] suggest that one potential refinement would be to adjust window sizes to local levels of LD, although how to account for varying levels of LD between populations remains unclear.

Finally, two complimentary studies of the HapMap and Perlegen data have been performed to detect signatures of selection based on regions of extended LD [21,22]. In each study, new statistics were developed that are similar in spirit to the Long Range Haplotype test developed by Sabeti *et al.* [23]. In the first study, Wang *et al.* [21] proposed the LD decay (LDD) test, which measures the extent of LD surrounding the alleles at a marker locus. Specifically, for each SNP an average log likelihood (ALnLH) statistic is calculated (Box 1) with high test statistic values indicating loci where one allele exhibits unusually long-range LD relative to the alternative SNP allele. The LDD test was initially applied to the Perlegen data and SNPs with ALnLH values >2.6 standard deviations from the genome-wide average were considered significant. In total, 25 386 SNPs met this criterion. The selected SNPs clustered in or around 1799 genes, and this set of genes was significantly enriched for Gene Ontology terms such as pathogen–host interaction, reproduction, DNA metabolism and/or cell cycle, protein metabolism and neuronal function (http://www.geneontology.org). Furthermore, 35% of SNPs with evidence of positive selection were located at least 100 Kb from known genes. An important observation of this study was that ~29% of the significant SNP clusters showed evidence for selection in all three Perlegen populations, whereas 78% of them showed evidence for selection in at least two populations. Interestingly, the authors argue that their data can be explained by ongoing balancing selection following the 'Out of Africa' dispersal, because many of the signatures of selection are shared across populations and the alleles have not reached fixation. Although this is an interesting proposition, the current data do not provide sufficient information to distinguish among competing hypotheses regarding the tempo and mode of the inferred selective events.

Similar to Wang *et al.*, Voight *et al.* [22] developed a novel summary statistic to capture the expected increase in levels of LD surrounding a locus that is subject to positive selection. This test uses the concept of extended haplotype homozygosity (EHH) [23], which measures the decay in homozygosity as a function of distance from a 'core' SNP. The new statistic, denoted as iHS, is computed by integrating the area under the EHH curve in both directions, taking into account whether the allele is ancestral or derived. Large iHS values (both positive and negative) indicate unusually long haplotypes of high LD. The iHS values were calculated for each SNP in the HapMap samples (European, Yoruban and Asian samples) with a minor allele frequency of $>5\%$, and 100-Kb windows were examined for the presence of multiple unusually large iHS scores. Regions of the genome that contained targets of putative positive selection were defined as windows in the greatest 1% of the empirical distribution for the proportion of SNPs where $|iHS|>2$. This criterion results in ~288 significant windows for each population encompassing a total of 455 genes and 237 regions with no known genes. Thus, these results in combination with data from Carlson *et al.* [14] and Wang *et al.* [21] suggest that non-coding regions have been an important substrate for adaptive evolution. Among the set of genes with signatures of positive selection, Panther Gene Ontology terms such as metabolism of carbohydrates, lipids, phosphates, vitamin transport, gametogenesis, spermatogenesis and fertilization were significantly enriched (http://www.pantherdb.org). These results are generally consistent with those of Wang *et al.* [21], and begin to suggest general themes about the types of genes that have been targets of positive selection in humans.

Although there was an excess of shared selective events across populations, many significant regions were confined to a single population. This observation contrasts with Wang *et al.* [21], who estimated that 78% of selective events are shared by two or more populations, whereas Voight *et al.* [22] found that only 19% are shared between two or more populations. One potential explanation for this discrepancy is that Voight *et al.* [22] included all SNPs with a minor allele frequency $>5\%$, whereas Wang *et al.* [21] required minor allele frequencies to be $>22\%$. Because the frequency of an allele is (approximately) proportional to its age, Voight *et al.* [22] are probably detecting more recent selective sweeps that postdate population divergence, whereas Wang *et al.* [21] are preferentially detecting older sweeps. Further theoretical

work on the relative power of these tests, and others, to detect young versus old sweeps and strong versus weak sweeps would be particularly useful to the field and help clarify some of these issues (and see below).

Genome-wide scans for positive selection in domesticated species

The power of genome-wide analyses is of course not limited to humans, and several insightful studies have been performed in other species, particularly in domesticated species. In contrast to natural populations, domesticated species provide an exciting opportunity to understand how artificial selection promotes rapid phenotypic evolution. Given that the strength of artificial selection is expected to be more than the selection operating in natural populations, it is reasonable to hypothesize that targets of artificial selection will be easier to find in domesticates than in non-domesticates (but see Innan and Kim [24]). Genome-wide analyses can potentially enable one to reconstruct, at least in part, the series of DNA sequence changes that occurred during the domestication process, underlying the phenotypic change. Beyond satisfying human curiosity, the genetic dissection of domestication is poised to address fundamental questions about the mechanisms of evolutionary change.

One of the most intensely studied domesticated species is maize (*Zea mays ssp. maysand*), which is a direct descendent of its wild progenitor teosinte (*Z. mays ssp. parviglumis*). Although the domestication of maize occurred recently, within the past ~7500 years, the two species show striking morphological variation that presumably reflects the action of human intervention. Recently, Wright *et al.* [25] analyzed patterns of polymorphism in 774 genes in maize and teosinte. An approximate likelihood method was developed to estimate parameters of the maize domestication bottleneck that best fit the observed data. This analysis suggests that <10% of the teosinte population contributed to extant levels of maize genetic diversity. In a clever analysis, these demographic parameters were then used to test the hypothesis whether the entire data set could be explained by a single domestication bottleneck or whether the data was better explained by two classes of sites consisting of neutral and selected loci. The authors estimate that 2–4% of maize genes (i.e. 1200–2400 genes) fall into the selected category and thus have been affected by artificial selection. Among the set of selected loci, there was an enrichment of genes involved in amino acid biosynthesis and protein catabolism, and a correlation was observed between highly significant selected loci and known positions of QTL in maize that drive key morphological traits. It will be interesting to follow-up these results with detailed population genetic analyses on individual genes to localize the signature of selection. If previous results from the *tb1* gene [26], which controls plant architecture, and the *tga1* gene [27], which is responsible for releasing the grain from its hard casing in maize, serve as a guide then regulatory evolution might make an important contribution to the molecular evolution of genes identified by Wright *et al.*

Another domesticated species that is likely to become a popular model for evolutionary studies is the domesticated dog (*Canis familiaris*). Dogs are among the most phenotypically diverse animals and the ~400 breeds come in a spectacular assortment of sizes, shapes, colors, temperaments and are afflicted by many of the same diseases as humans [28]. Pollinger *et al.* [29] performed extensive coalescent simulations to investigate the utility of an approach referred to as 'selective sweep mapping' [30] to identify genes responsible for breed-specific characteristics. The rationale of selective sweep mapping is that during breed creation artificial selection should impart a distinct signature on genomic regions harboring loci that influence the specific phenotype that is selected. A number of statistics for selective sweep mapping were investigated, including F_{ST} and statistics based on quantifying decreased heterozygosity. The authors conclude that low-resolution genome scans using microsatellite markers have high power to identify outlier regions that potentially contain genes contributing to inter-breed phenotypic variation. The theoretical predictions from the simulations were empirically tested by selective sweep mapping of the locus responsible for achondroplasia (shortened limbs) in Dachshunds. Specifically, levels of heterozygosity for 302 microsatellite markers were surveyed in Dachshunds and a panel of control dogs from breeds not affected by achondroplasia. Three linked markers spanning a 15-Mb region on canine chromosome 3 were found to be monomorphic in Dachshunds and polymorphic in the control dogs, which is an unusual observation based on the authors simulation of neutral patterns of genetic variation. Interestingly, this locus contains the gene encoding fibroblast growth factor receptor 3 (FGFR3), a mutation in which causes achondroplasia in humans [31], although limited resequencing of the transmembrane domain failed to reveal achondroplasia-causing mutations in dogs [32]. Thus, the putative mutation for foreshortened limbs might occur in protein-coding regions of *FGFR3* instead of in the transmembrane domains, in a regulatory region or in another gene located in this region. Although these results are encouraging, the distinct demographic history of dogs, including severe bottlenecks, inbreeding and non-random mating (such as sire-effects), can produce considerable heterogeneity in genome-wide patterns of variation within and between breeds, confounding inferences of selection. However, we expect that the recent publication of the dog genome [33] will facilitate additional evolutionary analyses, and ideally will stimulate comparative genome-wide surveys of polymorphism across many breeds [34].

Genome-wide scans for positive selection between species

In contrast to analyses based on polymorphisms within species, which are most suited to studying selective events within populations, analyses of inter-specific divergence facilitate inferences of adaptation occurring between species. A classic test of adaptive protein evolution between species is to calculate the ratio of non-synonymous (d_N) to synonymous (d_S) substitutions. Although this has traditionally been performed for single genes, the

expanding repertoire of genome sequences from multiple species enables global analyses of protein evolution. For example, Nielsen *et al.* [35] performed a large-scale study on 13 731 human–chimpanzee orthologs. In total, 35 genes were found to have significantly ($P<0.05$) elevated ratios of d_N/d_S, implying the action of positive selection (although the data do not enable one to determine whether selection occurred in the human lineage, the chimp lineage or both). Classification based on functional annotations suggests that genes involved in immunity-defense, sensory perception and spermatogenesis are enriched for positive selection, consistent with a similar analysis of human–chimp–mouse orthologs using a smaller data set [36]. The 50 genes with the strongest evidence for selection were resequenced in 20 European–Americans and 19 African–Americans. The polymorphism data revealed an excess of high-frequency-derived alleles for non-synonymous single nucleotide polymorphisms (SNPs), strengthening the hypothesis of positive selection.

Bustamante *et al.* [37] performed a complimentary study to Nielsen *et al.* [35] by analyzing 11 624 protein coding genes that were sequenced in 20 European–Americans, 19 African–Americans, and one chimpanzee, one of the largest surveys of human genetic variation described to date. An extension of the classic McDonald–Kreitman (MK) test was used to compare levels of polymorphism within humans with the levels of divergence between chimpanzee and human for all informative loci. Consistent with previous studies, the authors found a highly significant excess of amino acid polymorphism within humans relative to the divergence between chimpanzee and human indicating that a significant portion of amino acid variation in humans is slightly to moderately deleterious. In addition, the authors used the polymorphism and divergence data to estimate the population selection parameter ($\gamma=2N_es$, where N_e is the effective population size and s is the selection coefficient) for each gene. Genes evolving under positive selection would have a $\gamma>0$ and genes subject to purifying selection are expected to have a $\gamma<0$. Overall, 304 out of 3377 informative loci (9.0%) showed significant ($P<0.05$) evidence for positive selection, whereas 813 out of 6033 informative loci (13.5%) showed significant evidence for purifying selection.

Putting the pieces together: overlap among genome-wide analyses

As the number of genome-wide scans for positive selection continues to accumulate, it is both interesting and important to analyze how consistent inferences of selection are across studies. Continuing with our human-centric focus, Table 1 summarizes the pairwise overlap in genes that possess signatures of positive selection identified by recent genome-wide scans in humans for studies with publicly available results. In total, the various genome-wide scans have identified 2316 genes with signatures of positive selection. Because many of these genes are clustered in the genome, probably reflecting a phenomenon known as genetic hitchhiking, the number of selected loci is probably <2316. Overall, the overlap between genes reported to have signatures of positive selection between studies is fairly modest. For example, Voight *et al.* [22] and Wang *et al.* [21] have the greatest number of significant genes shared between studies ($\sim$27% of the significant genes identified in Voight *et al.* were also found by Wang *et al.*), as might be expected because both of these studies searched for positive selection based on extended regions of LD. Interestingly, although $\sim$27% of the significant genes from Carlson *et al.* [14] overlap with Wang *et al.* [21] only 8% overlap with Voight *et al.* [22].

How can the inconsistencies in these results be explained? First, given the varying statistical tests used to detect deviations from neutrality, we should not expect complete concordance between studies. More specifically, different studies are probably detecting different selective events. For example, the LD-based analyses of Wang *et al.* [21] and Voight *et al.* [22] have the greatest power to detect incomplete and on-going selective sweeps. Conversely, tests based on the site frequency spectrum used by Carlson *et al.* [14] have greater power to identify sweeps where the advantageous allele is approaching fixation or completed sweeps in which new mutations are occurring on selected haplotypes that over time will return patterns of genetic variation to equilibrium. In addition, Nielsen *et al.* [35] and Bustamante *et al.* [37] were searching for examples of adaptive evolution between humans and chimpanzee. In short, the statistical tests used in each study are recovering selective events from different time periods and for different stages of the selective sweep. Second, even for tests that should detect similar types of selective events, low statistical power further decreases the probability of overlap. Third, most studies report only the most significant results (i.e. outliers in the 1% empirical distribution). Therefore, the results presented in Table 1 are probably a conservative estimate of overlap between studies. Finally, as will be discussed in more detail in the next section, the false positive rate in genome-wide scans for selection is likely to be high.

Table 1. Pairwise overlap of genes with evidence of positive selection in humans across studies[a,b]

Test statistic		LD	Tajima's D	F_{ST}	d_N/d_S	MK-PRF
Study	Wang *et al.* [21]	Voight *et al.* [22]	Carlson *et al.* [14]	Altshuler *et al.* [13]	Nielsen *et al.* [35]	Bustamante *et al.* [37]
Wang *et al.* [21]	1799	125	47	5	4	40
Voight *et al.* [22]		455	11	4	1	12
Carlson *et al.* [14]			176	5	0	2
Altshuler *et al.* [13]				27	0	1
Nielsen *et al.* [35]					50	14
Bustamante *et al.* [37]						304

[a]The number of genes that possess evidence of positive selection in each study are shaded.
[b]The number of genes with signatures of positive selection that overlap between the studies is also shown.

Would the real targets of positive selection please stand out?

A common theme to identifying putative targets of positive selection employed in many of these studies is to identify outlier loci. However, it is unclear how powerful and robust simple outlier approaches are in finding genes subject to selection. More specifically, basic population genetics theory predicts that evolution is an inherently stochastic, and thus noisy, process (Box 2). As thousands or tens of thousands of genes are studied in a typical genome-wide analysis it is increasingly likely to observe an 'unusually large' test statistic for a neutrally evolving locus. How often will a positively selected gene rise above the background of neutral noise?

To formalize this argument, we performed coalescent simulations with the program SelSim [38] to mimic a small-scale genome-wide analysis by simulating 5000 unlinked loci consisting of 4950 neutral and 50 positively selected genes. For each locus, we calculated Tajima's D and asked how many selected genes were found in the first percentile of the empirical distribution across all loci. The results based on 200 simulation replicates are summarized in Figure 3, and suggest reasons for encouragement and caution. The encouraging news is that even the simple outlier approach considered here results in an enriched set of genes that contain targets of positive selection. The cautionary, and obvious, result is that we should not expect all outliers to have experienced selective pressures, and thus outliers identified as extreme observations in empirical distributions are almost certainly some combination of both true and false positives. As one would expect, the magnitude of selection is an important parameter in ultimately determining power and specificity. For example, when selection is strong ($s=0.10$) there is $>80\%$ power to detect at least half of all genes under selection. However, when selection is of moderate intensity ($s=0.01$), the power to detect at least half of all genes under selection is essentially zero. Although these simulations are obvious simplifications of real genomes because we have not taken into account variation in rates of mutation, recombination and selection coefficients across loci, nor have we considered demographic perturbations that real populations have probably experienced, they do highlight the inherent difficulties in distinguishing between the confounding effects of genetic drift and natural selection. Clearly, more theoretical work needs to be performed and new methods developed [39] to extract the maximum amount of information possible from empirical patterns of genetic variation.

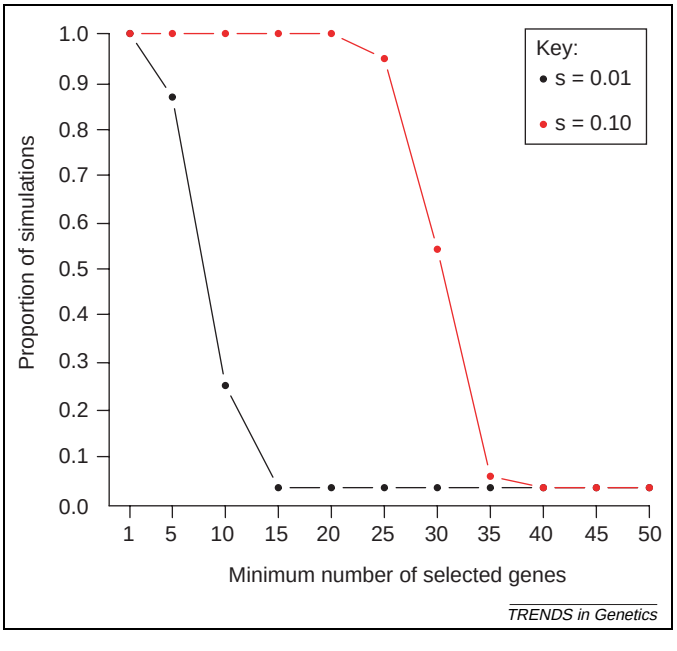

Figure 3. The power of a simple outlier approach to detect genes under positive selection. Patterns of genetic variation were simulated for 5000 unlinked loci, 4950 of which were evolving neutrally and 50 of which were targets of positive selection. For the positively selected loci, we modeled an incomplete selective sweep in which the advantageous allele has reached a population frequency of 90%. Tajima's D was calculated for each locus and the number of selected genes falling in the first percentile of the empirical distribution of Tajima's D was recorded. This process was repeated 200 times and power was measured as the proportion of simulations that recovered $\times$ ($1 \leq \times \leq 50$) or more selected genes. The black and red curves denote results assuming a selective coefficient (s) of 0.01 and 0.10, respectively. Other parameters of the simulation include the population mutation and recombination rates, which were set to 10.

In addition, for genome-wide analyses of selection that calculate measures of significance at individual loci, it remains unclear how to correct for multiple hypothesis tests efficiently to minimize false positives while maintaining reasonable power to detect deviations from neutrality. Usually, reported significance levels are not corrected for multiple tests, creating ambiguity about which results can be ascribed to positive selection and which are due to chance. More methodological research, with a particular focus on developing approaches that are robust to issues such as non-independence among loci, is clearly needed to address this important issue.

Can genome-wide scans for selection facilitate disease mapping studies?

Genome-wide scans for selection might facilitate disease-related research. For example, Bustamante *et al.* [37] found that genes with a reduced rate of non-synonymous substitution (indicating purifying selection) were significantly correlated with being involved in mendelian diseases. Furthermore, Nielsen *et al.* [35] found that many of the genes they identified as potential targets of positive selection were involved in cancer-related processes. Signatures of selection have also been reported in genes that are believed to contribute to complex diseases, such as cardiovascular disease [matrix metallopeptidase 3 (*MMP3*) [40], cytochrome P450 (*CYP3A*) [41], angiotensinogen (*AGT*) [42]] type 2 diabetes (*CAPN10* [43]) and asthma [interleukin 13 (*IL13*) [44], *IL4* [45,46] and *IL1A* [11]]. Thus, if selection has differentially acted on disease-related genes then genome-wide analyses should help to inform disease mapping studies. Although the potential correlations between selection and disease have primarily been explored in humans, it might also prove useful in model organisms, such as dogs.

A more comprehensive understanding of how and where selection has acted on genomes would also provide an important conceptual framework for interpreting patterns of disease in an evolutionary context. For example, the 'thrifty' genotype [47] and sodium-retention hypotheses [48] were proposed to explain the high

Table 2. Genes with evidence of geographically restricted selection in humans

Gene	Putative selective pressure	Phenotype or disease associations	Refs
AGT	Climate (thermoregulation)	Hypertension	[42]
CYP3A	Climate (salt avidity)	Hypertension	[41]
SLC24A5	Climate (UV exposure)	Skin pigmentation	[16]
FY	Pathogen (*Plasmodium vivax*)	Malaria resistance	[15]
IL4	Pathogen (unknown)	Asthma	[45,46]
IL13	Pathogen (unknown)	Asthma	[44]
CASP12	Pathogen (unknown)	Sepsis	[65,66]
NAT2	Diet (agriculture)	Bladder cancer and/or adverse drug reactions	[67]
LCT	Diet (milk)	Lactose tolerance or intolerance	[20]
TRPV6	Diet (milk)	Prostate cancer	[11]
MMP3	Diet (unknown)	Coronary heart disease	[40]

prevalence of type 2 diabetes and essential hypertension, respectively. These hypotheses argue that diseases of 'civilization' result from a mismatch between the present day environment and genotypes that were advantageous during earlier periods of human history. Recently, Di Rienzo and Hudson [49] developed an explicit population genetic model derived from the general logic of the thrifty genotype and sodium-retention hypotheses. They proposed the ancestral-susceptibility model, in which disease susceptibility alleles are ancestral and derived variants are protective. In this model, ancestral alleles were adapted to historical environmental conditions, which become maladaptive as humans changed lifestyles and dispersed into new environmental niches. Young *et al.* [50] provide compelling support for this hypothesis, and based on patterns of polymorphism in five genes involved in blood pressure regulation, argue that susceptibility to hypertension is ancestral. In addition, they argue that the geographically restricted selective pressures (e.g. local adaptation) during the out-of-Africa dispersal contributed to differential susceptibility of hypertension among human populations. As discussed earlier, pervasive signals of geographically restricted selection have been found in recent genome-wide scans for selection in humans [14,18,21,22,51], which is consistent with an increasing number of more focused single gene analyses [15,16,20,40–46]. Thus, it seems clear that local adaptation has contributed to recent human evolutionary history (Table 2), and genome-wide scans for selection will continue to be an important tool in finding additional instances of local adaptation (and for ruling out demographic explanations), which in turn will provide strong candidate genes for diseases in which prevalence varies as a function of ethnicity [52].

Conclusions and future prospects

The continuing deluge of genome-wide catalogs of genetic variation poses both great opportunities and challenges. The opportunities afforded by such resources are clear – a more global and comprehensive understanding of genomic patterns of polymorphism and the evolutionary forces that have shaped them. As discussed earlier, several genome-wide scans for positive selection have already been performed and are beginning to provide tantalizing insights into how, where and why adaptive evolution has influenced extant patterns of genetic variation. Although these initial genome-wide scans must now yield to the more difficult and time consuming process of detailed follow-up studies, several general themes are beginning to emerge including the, perhaps, surprising frequency of positive selection, signatures of selection occurring in regions of the genome previously thought to be non-functional (e.g. pseudogenes and non-coding DNA) and the existence of geographically restricted selective pressures.

However, many challenges remain including the development of increasingly sophisticated and computationally efficient statistical methods to tease apart the effects of drift and selection, identifying the causal gene and genetic variation therein driving the signature of selection observed across large genomic regions, functionally characterizing the suspected targets of selection, determining the relative contributions of adaptive evolution in protein coding versus regulatory regions, estimating important parameters such as the magnitude and timing of selective events and, ultimately, making inferences about the historical forces exerting selective pressures. Despite these difficulties, we are optimistic that genome-wide analyses are poised to answer long-standing questions in evolutionary biology, population genetics, the molecular basis of phenotypes and perhaps human health.

Acknowledgements

We thank Dayna Akey, James Ronald, Nicholas Akey, Robert Moyzis and Jonathan Prichard for helpful discussions related to this work. J.M.A. is supported by research starter grants from the UW Division of Nutritional Sciences and the NSF (DEB-0512279).

References

1 Gillespie, J.H. (1991) *The Causes of Molecular Evolution*, Oxford University Press
2 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press
3 Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* 217, 624–626
4 King, J.L. and Jukes, T.H. (1969) Non-darwinian evolution. *Science* 164, 788–798
5 Ohta, T. (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246, 96–98
6 Kreitman, M. (2000) Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.* 1, 539–559
7 Nielsen, R. (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity* 86, 641–647
8 Smith, J.M. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35
9 Ronald, J. and Akey, J.M. (2005) Genome-wide scans for loci under selection in humans. *Hum. Genomics* 2, 113–125
10 Przeworski, M. *et al*. (2000) Adjusting the focus on human variation. *Trends Genet.* 16, 296–302
11 Akey, J.M. *et al*. (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2, e286
12 Hinds, D.A. *et al*. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* 307, 1072–1079
13 Altshuler, D. *et al*. (2005) A haplotype map of the human genome. *Nature* 437, 1299–1320

14 Carlson, C.S. *et al*. (2005) Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15, 1553–1565
15 Hamblin, M.T. *et al*. (2002) Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* 70, 369–383
16 Lamason, R.L. *et al*. (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* 310, 1782–1786
17 Hearn, T. *et al*. (2002) *ALMS1*, a large gene with a tandem repeat encoding 47 amino acids, causes Alstrom syndrome. *Nat. Genet.* 31, 79–83
18 Weir, B.S. *et al*. (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Res.* 15, 1468–1476
19 Weir, B.S. and Hill, W.G. (2002) Estimating *F*-statistics. *Annu. Rev. Genet.* 36, 721–750
20 Bersaglieri, T. *et al*. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* 74, 1111–1120
21 Wang, E.T. *et al*. (2006) Global landscape of recent inferred darwinian selection for *Homo sapiens*. *Proc. Natl. Acad. Sci. U. S. A.* 103, 135–140
22 Voight, B.F. *et al*. (2006) A map of recent positive selection in the human genome. *PLoS Biol.* 4, e72
23 Sabeti, P.C. *et al*. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837
24 Innan, H. and Kim, Y. (2004) Pattern of polymorphism after strong artificial selection in a domestication event. *Proc. Natl. Acad. Sci. U. S. A.* 101, 10667–10672
25 Wright, S.I. *et al*. (2005) The effects of artificial selection on the maize genome. *Science* 308, 1310–1314
26 Wang, R.L. *et al*. (1999) The limits of selection during maize domestication. *Nature* 398, 236–239
27 Wang, H. *et al*. (2005) The origin of the naked grains of maize. *Nature* 436, 714–719
28 Ostrander, E.A. *et al*. (2005) The canine genome. *Genome Res.* 15, 1706–1716
29 Pollinger, J.P. *et al*. (2005) Selective sweep mapping of genes with large phenotypic effects. *Genome Res.* 15, 1809–1819
30 Schlotterer, C. (2003) Hitchhiking mapping–functional genomics from the population genetics perspective. *Trends Genet.* 19, 32–38
31 Shiang, R. *et al*. (1994) Mutations in the transmembrane domain of FGFR3 cause the most common genetic form of dwarfism, achondroplasia. *Cell* 78, 335–342
32 Martinez, S. *et al*. (2000) Achondroplastic dog breeds have no mutations in the transmembrane domain of the FGFR-3 gene. *Can. J. Vet. Res.* 64, 243–245
33 Lindblad-Toh, K. *et al*. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438, 803–819
34 Parker, H.G. *et al*. (2004) Genetic structure of the purebred domestic dog. *Science* 304, 1160–1164
35 Nielsen, R. *et al*. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 3, e170
36 Clark, A.G. *et al*. (2003) Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. *Science* 302, 1960–1963
37 Bustamante, C.D. *et al*. (2005) Natural selection on protein-coding genes in the human genome. *Nature* 437, 1153–1157
38 Spencer, C.C. and Coop, G. (2004) SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics* 20, 3673–3675
39 Nielsen, R. *et al*. (2005) Genomic scans for selective sweeps using SNP data. *Genome Res.* 15, 1566–1575
40 Rockman, M.V. *et al*. (2004) Positive selection on MMP3 regulation has shaped heart disease risk. *Curr. Biol.* 14, 1531–1539
41 Thompson, E.E. *et al*. (2004) CYP3A variation and the evolution of salt-sensitivity variants. *Am. J. Hum. Genet.* 75, 1059–1069
42 Nakajima, 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.* 74, 898–916
43 Fullerton, S.M. *et al*. (2002) Geographic and haplotype structure of candidate type 2 diabetes susceptibility variants at the calpain-10 locus. *Am. J. Hum. Genet.* 70, 1096–1106
44 Zhou, G. *et al*. (2004) Haplotype structure and evidence for positive selection at the human IL13 locus. *Mol. Biol. Evol.* 21, 29–35
45 Sakagami, T. *et al*. (2004) Local adaptation and population differentiation at the interleukin 13 and interleukin 4 loci. *Genes Immun.* 5, 389–397
46 Rockman, M.V. *et al*. (2003) Positive selection on a human-specific transcription factor binding site regulating IL4 expression. *Curr. Biol.* 13, 2118–2123
47 Neel, J.V. (1962) Diabetes mellitus: a 'thrifty' genotype rendered detrimental by 'progress'? *Am. J. Hum. Genet.* 14, 353–362
48 Gleibermann, L. (1973) Blood pressure and dietary salt in human populations. *Ecol. Food Nutr.* 2, 143–156
49 Di Rienzo, A. and Hudson, R.R. (2005) An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends Genet.* 21, 596–601
50 Young, J.H. *et al*. (2005) Differential susceptibility to hypertension is due to selection during the out-of-Africa Expansion. *PloS Genet.* 1, e82
51 Akey, J.M. *et al*. (2002) Interrogating a high density SNP map for signatures of natural selection. *Genome Res.* 12, 1805–1814
52 Halder, I. and Shriver, M.D. (2003) Measuring and using admixture to study the genetics of complex diseases. *Hum. Genomics* 1, 52–62
53 Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595
54 Fu, Y.X. and Li, W.H. (1993) Statistical test of neutrality of mutations. *Genetics* 133, 693–709
55 Fay, J.C. and Wu, C.I. (2000) Hitchhiking under positive darwinian selection. *Genetics* 155, 1405–1413
56 Weir, B.S. and Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution Int. J. Org. Evolution* 38, 1358–1370
57 Hudson, R.R. *et al*. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116, 153–159
58 McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351, 652–654
59 Nielsen, R. and Yang, Z. (1998) Likelihood models for detecting positive selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148, 929–936
60 Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426
61 Suzuki, Y. and Gojobori, T. (1999) A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 16, 1315–1328
62 Kingman, J.F.C. (1982) 'The coalescent'. *Stochastic Process. Appl.* 13, 235–248
63 Hudson, R.R. (1983) Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37, 203–217
64 Rosenberg, N.A. and Nordborg, M. (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* 3, 380–390
65 Wang, X. *et al*. (2006) Gene losses during human origins. *PLoS Biol.* 4, e52
66 Xue, Y. *et al*. (2006) Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am. J. Hum. Genet.* 78, 659–670
67 Patin, E. *et al*. (2006) Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes. *Am. J. Hum. Genet.* 78, 423–436