Positive Selection in the Human Genome: From Genome Scans to Biological Significance

Joanna L. Kelley\(^1\) and Willie J. Swanson\(^2\)

\(^{1}\)Department of Human Genetics, University of Chicago, Chicago, IL 60637; email: jjkkelley@uchicago.edu
\(^{2}\)Department of Genome Sciences, University of Washington, Seattle, Washington 98195; email: wswanson@gs.washington.edu

Key Words
adaptive evolution, population differentiation, amylase, FOXP2, lactase, Neanderthal

Abstract
Here we review the evidence for positive selection in the human genome and its role in human evolution and population differentiation. In recent years, there has been a dramatic increase in the use of genome-wide scans to identify adaptively evolving loci in the human genome. Attention is now turning to understanding the biological relevance and adaptive significance of the regions identified as being subject to recent positive selection. Examples of adaptively evolving loci are discussed, specifically \(LCT\) and \(FOXP2\). Comprehensive studies of these loci also provide information about the functional relevance of the selected alleles. We discuss current studies examining the role of positive selection in shaping copy number variation and noncoding genomic regions and highlight challenges presented by the study of positive selection in the human genome.
INTRODUCTION

A fundamental question in evolutionary biology is what distinguishes humans from our closest living relatives, chimpanzees? One way to uncover biologically relevant differences in our genetic makeup is to examine regions of the genome that have been subject to positive selection. In this review, we define positive selection as the process that increases the frequency of mutations that confer a fitness advantage to individuals carrying those mutations. The identification of adaptively evolving genes and regulatory regions may help to elucidate specific regions of the genome that have evolved during the past six million years to allow humans to develop language, use tools, and ultimately populate nearly every continent.

Historically, the study of positive selection in human evolution was motivated by a priori assumptions about selective pressure. For example, the Duffy Fy*O blood group mutation was shown to be under positive selection (28, 29), a discovery motivated by the finding that African individuals with the Fy*O blood group were resistant to malaria infection by Plasmodium vivax (52). Recently, the availability of large-scale polymorphism and divergence data has made it possible to scan the human genome for evidence of positive selection without making assumptions about the potential selective advantage at particular loci. Genome-wide scans for adaptations have identified hundreds of putatively selected genes in the human genome (1, 7, 9, 11, 37, 38, 40, 43, 54, 67, 79, 86, 89, 90; also see 5, 55, 68). These scans apply a range of statistical methods to divergence and variation data to investigate features of the genome such as haplotype lengths and allele frequency spectrum differences. As genotyping and large-scale DNA sequencing efforts produce additional data, these scans will become more common and will ultimately provide a catalog of regions that have adaptively evolved in the human lineage and in specific populations. However, this information will not indicate which loci represent true examples of positive selection, the biological relevance of the region under positive selection, or the time of onset of the selective pressures.

METHODS TO IDENTIFY POSITIVE SELECTION

Positive selection may be identified in two fundamental ways: Divergence data are used to identify positive selection between species, whereas polymorphism data are used to identify positive selection within a species. Each method detects selection on a different timescale. Divergence data are used to identify older selective events, whereas polymorphism data are used to identify recent selective events. The use of both divergence and polymorphism data may provide additional support for positive selection acting in a region; however, a signature of selection with one test is sufficient evidence for selection (54).

Divergence Data

For protein-coding regions, a clear signal of positive selection is an excess in the number of nonsynonymous substitutions (i.e., amino acid altering) per nonsynonymous sites \( d_N \) compared with the number of synonymous substitutions (i.e., silent changes) per synonymous sites \( d_S \). This comparison is typically quantified by the ratio of \( d_N/d_S \). In the absence of selection, the \( d_N/d_S \) ratio is expected to equal one because \( d_N \) and \( d_S \) are normalized to the number of sites. An average \( d_N/d_S \) ratio that is less than one is a signature of purifying selection—selection against nonsynonymous substitutions. A high \( d_N/d_S \) ratio can result from either positive selection or a lack of functional constraint. A stringent criterion for positive selection requires an average \( d_N/d_S \) ratio greater than one for the entire gene (92). Analysis of six human class I major histocompatibility complex (MHC) molecules showed them to have a \( d_N/d_S \) ratio averaged across all sites and lineages of 0.5, which is high but less than one (76). However, the subset of sites in the three-dimensional structure that comprise the antigen-recognition site had a \( d_N/d_S \) ratio...
significantly greater than one (34). This local variation in the \( d_N/d_S \) ratio can be examined using codon models that allow for variability in the \( d_N/d_S \) ratio between sites; the best-fitting model is identified using maximum likelihood ratio tests (57, 93, 94). These methods can identify subsets of sites that have been subjected to positive selection even when the \( d_N/d_S \) ratio averaged across all sites is less than one. One additional benefit of these analyses is the ability to identify specific amino acid sites that have been the target of selection (95), which can lead to explicit functional predictions such as identification of binding sites. A limit of these tests is that they cannot detect selection acting on regulatory changes, which have been hypothesized to be important in human evolution (44).

**Population Data**

The detection of recent positive selection requires the identification of regions displaying evidence of a selective sweep. A selective sweep is the increase in frequency of a beneficial allele that confers a fitness advantage to its carrier. Neutral alleles closely linked to the beneficial allele also rise in frequency, which is called the hitchhiking effect (Figure 1b). A characteristic molecular signature of a selective sweep is the elimination of nucleotide variation in the region of the genome close to the beneficial allele (Figure 1c). As new mutations subsequently arise in the swept region, the excess of rare alleles skews the site frequency spectrum. An incomplete or partial sweep, where the favored allele does not reach fixation, does not lead to as dramatic a deviation in the site frequency spectrum (71). However, the region surrounding an incomplete selective sweep will have a mix of long haplotypes with the selected allele and ancestral haplotypes of varying lengths (Figure 1b). Thus, a consideration of different aspects of the data allows researchers to identify selective events that vary in their extent and timing (68).

Signatures of adaptive events that occurred since the last ice age or during human migration out of Africa should still be identifiable from samples of extant human populations. A selective sweep takes approximately \( 2\ln(2N)/s \) generations, where \( N \) is the effective population size and \( s \) is the selection coefficient (72). Using the estimated human effective population size of 10,000 (67) and 25 years per generation, sweeps with strong selection coefficients will be complete in \( \sim 10,000 \) years (\( s = 5\% \)). In addition, the signature of selection will be identifiable for an additional amount of time, depending on the mutation and recombination rate in the region (71). Alleles with a lower

---

**Figure 1**

Selective sweep with recombination. The figure shows eight chromosomal regions with neutrally segregating alleles (red); the gray chromosome is the chromosome with the adaptive allele (blue). (b) A snapshot of the region during the sweep. (c) The result after a complete sweep in the region.
Derived allele: mutation from a common ancestor, usually determined by comparison to an outgroup
iHS: integrated haplotype score
SNP: single-nucleotide polymorphism

derived selection coefficient will take longer to reach fixation in the population. For example, an allele with \( s = 1\% \) will take \( \sim 50,000 \) years to fix in the population. The time to completion of a selective sweep is largely dependent on the strength of selection. Therefore, sweeps that are presently incomplete are either recent sweeps with a strong selection coefficient or old sweeps with a weaker coefficient.

Site Frequency Spectrum

As depicted in Figure 1c, following a complete selective sweep the site frequency spectrum is shifted relative to the neutral expectation (71). The swept region has very little variation, if any. The amount of variation remaining in the region is highly dependent on the recombination distance from the selected site. As new mutations accumulate after a complete selective sweep, there will be an excess of rare alleles in the swept region as compared with neutral regions that are unlinked to selected sites. Tajima's D is the most sensitive test for identifying regions with an excess of common alleles or an excess of rare alleles (78). However, Tajima's D is also sensitive to population demographics (64, 77). A test used to identify skews in the site frequency spectrum that is less affected by population demography is the composite likelihood test (42, 56, 90). The composite likelihood test uses the background pattern of variation to compare the likelihood of a neutral model versus a selective sweep model for a given genomic region. During a selective sweep, the hitchhiking effect drags variants to high or low frequency. Therefore, high-frequency derived alleles are a hallmark signature of a selective sweep. An excess of derived alleles can be measured using Fay and Wu's H (20).

Population Differentiation

Population differentiation, or allele frequency variation between populations, is largely determined by genetic drift in a population. When a locus is subjected to positive selection in a geographically restricted population, the allele frequencies around the selected locus change rapidly, leading to a high degree of population differentiation in the region (49). Therefore, a high degree of population differentiation (measured by \( F_{ST} \), the fixation index) can be an indication of positive selection (2). However, \( F_{ST} \) alone does not indicate that a region has been adaptively evolving (22).

Haplotype Length

Hitchhiking due to positive selection results in an increase in the frequency of the haplotype on which the selected allele occurs. The selected haplotype is longer than expected because insufficient time has elapsed to allow recombination to reduce its length. Long high-frequency haplotypes are therefore indicative of the action of positive selection, whereas nonselected haplotypes vary in length. The observation of unusually frequent, long haplotypes within a population is measured by determining the relative haplotype homozygosity (67) or integrated haplotype score (iHS) (86). These statistics identify incomplete selective sweeps by requiring that the selected and nonselected haplotypes be present within the test population. More recently, test statistics have been developed to compare haplotype lengths between populations in an effort to identify complete selective sweeps (43, 69, 79).

GENOME-WIDE SCANS

Polymorphism-based scans aimed at detecting recent positive selection apply various combinations of the test statistics outlined above to publicly available genotype data (9, 37, 38, 40, 43, 69, 79, 86, 89). The two main genotype resources are Perlegen Sciences and the International HapMap Project, which have typed more than 1.5 and 3.1 million single-nucleotide polymorphisms (SNPs), respectively, in individuals with European, Asian, and African ancestry (33, 37, 38). Although the datasets are extensive, SNP genotype data are not free of ascertainment and data-collection biases. The ascertainment schemes, SNP density, and
sample sizes differ between the two datasets. For example, for the Perlegen resource (33), SNPs for genotyping were ascertained by sequencing from an ethnically diverse panel of 24 individuals from the DNA Polymorphism Discovery Resource (13). By discovering polymorphisms in a diverse panel, common polymorphisms are more likely to be discovered; however, rare polymorphisms are less likely to be discovered and genotyped (12). In contrast, for the HapMap dataset, SNPs were specifically chosen to represent the most common polymorphisms in the human genome (37). Many of the test statistics were developed for application to sequence data, not genotype data. Therefore, it is important to take into account ascertainment biases when using these datasets to identify adaptively evolving regions because biases can skew population genetic measures (12). For example, if rare alleles are missed owing to either the SNP discovery method or the resequencing technology, statistics that rely on the site frequency spectrum will be skewed.

Many of the genome-wide scans for selection rely on using an empirical distribution of the test statistic to determine significance. Regions that fall in the tail of the distribution are considered candidate selection loci. Prior to the availability of genome-wide data, coalescent simulations were the most sophisticated method for determining whether or not a locus had significant evidence of positive selection [see Rosenberg & Nordborg (66) for a review of coalescent theory]. Coalescent simulations rely on assumptions about the underlying demographic histories, mutation rates, and recombination rates. An empirical distribution has the advantage in that it eliminates the need to assume the underlying demographic history, but the disadvantage is that it requires the assumption that positive selection is prevalent in human evolution. Choosing outliers using an empirical distribution assumes that an identifiable proportion of loci have been recently subject to positive selection (40) and will miss loci that are under positive selection but do not fall in the tail of the distribution (80). Moreover, although the set of candidate selection loci chosen from the tail of an empirical distribution may be enriched for true positives, it will also contain false positives.

The identification of regions by genome-wide scans depends on the test statistic used. The overlap between lists of identified regions is minimal, ranging from 8% to 27% (5; also see 55 for a discussion of the correlation between scans). Moreover, even when the test statistic identifies similar selective events (e.g., scans that identify unusually long haplotypes), the correspondence between candidate loci has so far been limited. A question that frequently arises is how prevalent is positive selection in the human genome? Evidence from divergence data (7) and site frequency spectrum data (90) suggests that approximately 10% of the genome is under positive selection. This estimate implies that selective sweeps are recurrent and frequent; therefore, parsing out recent and past selective events will be difficult. New theoretical models will be required to understand how the site frequency spectrum and haplotype lengths will be affected by multiple, overlapping sweeps in a region.

To summarize the functional representation of candidate loci, we have compiled population-specific lists of putatively selected genes. We further grouped the lists into three categories on the basis of the type of selective event: complete selective sweeps, intermediate sweeps, and fixed differences. We analyzed the PANTHER (Protein ANalysis THrough Evolutionary Relationships) classifications (81) to identify the categories of biological processes that are represented by candidate selection loci (Figure 2). The classes of genes identified are remarkably similar between the different sets of loci, suggesting that similar processes are recurrently selected. However, evaluation of the PANTHER classifications for over- and underrepresentation (82) of specific processes in the complete sweep versus partial sweep candidate loci lists revealed few differences. The differences include, but are not limited to, an overrepresentation of loci involved in mRNA transcription in Asian complete sweeps, chemosensory perception and olfaction in
African complete sweeps, and cell cycle and signal transduction in European complete sweeps.

The rest of this review focuses on examples of existing techniques and highlights areas of recent interest where new techniques are needed. First, we outline two examples of loci that are under positive selection that have proposed biological relevance: lactase persistence and the lactase gene (LCT) (population-specific selection) and speech and the forkhead box P2 gene (FOXP2) (human-specific changes). We discuss areas that have only recently been considered, and where methods for detecting positive selection are being developed, specifically in relation to copy number variation and noncoding genomic regions.

**Population-Specific Selection: Lactase Persistence and LCT**

Lactase persistence is an example of positive selection acting in different populations, on several variants, over the past 7000 years. Lactase persistence is postulated to be an adaptive feature because it is a human-specific trait, differs between populations, and is correlated with the domestication of cattle (74). The ability to digest lactose is speculated to be adaptive in populations that have domesticated cattle because the selective advantage enables individuals to obtain nutrition from the milk of domesticated cows.

In Northern Europeans, a common haplotype that has been described contains cis-regulatory polymorphisms that affect LCT transcript abundance and thus lactase persistence (19, 46, 59, 84). The polymorphisms found to correlate with lactase persistence occur on a haplotype with high linkage disequilibrium extending over 1 Mb (62). Poulter and colleagues (62) speculate the C/T-13,910 allele to be causal; however, some individuals with lactase persistence lack the causal allele, suggesting that additional uncharacterized cis- or trans-acting polymorphisms exist.

Although the derived SNP was found to be important in lactase persistence, the evolutionary history of the allele had yet to be fully characterized. The LCT region appears to have undergone a selective sweep 2000–20,000 years ago (4), coinciding with the domestication of cattle. Bersaglieri and colleagues (4) used population differentiation (FST), the long-range haplotype test (67), and a novel p excess statistic to look at the empirical distribution of population differentiation as evidence of positive selection acting in the region. The haplotype on which the derived C/T-13,910 SNP occurs is much longer than expected, given its frequency. The derived C/T-13,910 SNP allele frequency differs greatly between Europeans (77%), African Americans (~14%), and Chinese (0%), which is consistent with lactase persistence in each population. The absence of the European allele in non-European populations suggests the allele arose after the colonization of Europe. The high selection coefficient (between 0.014 and 0.15) distinguishes LCT as one of the most strongly selected loci in the human genome (4, 86). The length of the haplotype and the linkage disequilibrium in the region preclude the definitive identification of the -13910T allele as the causal variant.

Lactase persistence is prevalent in pastoral populations in Africa; however, the European variant is absent or at low frequency. This suggests that there is either a different causal variant in Europe or independent origins of lactase persistence in Africa. Tishkoff and colleagues (83) undertook a genotype-phenotype

---

**Figure 2**

Biological process representation of complied candidate selection loci. Distribution of candidate selection in PANTHER (Protein ANalysis THrough Evolutionary Relationships) biological process categories for loci identified for each population in scans for complete (a) and incomplete (b) selective sweeps, as well as loci identified under positive selection in the human lineage, using divergence data (c). Loci for analysis are pooled for complete sweeps (9, 40, 43, 69, 79, 90), intermediate sweeps (38, 86), and fixed differences (7, 54).
association study in 470 Africans from several different populations (43 ethnic groups) to identify the causal alleles for lactase persistence. Using a lactose tolerance test, they classified individuals as having lactase persistence, intermediate persistence, or nonpersistence. Lactase persistence was at the highest frequency in pastoral populations and lowest in a hunter-gatherer population, supporting the hypothesis that lactase persistence is correlated with cattle domestication. Three novel variants (G/C-14010, T/G-13915, and C/G-13907) were found to have a significant association with lactase persistence in various pastoral populations. The three newly discovered alleles occur on different haplotype backgrounds, which are distinct from the European-selected haplotype. The three SNPs account for ∼20% of phenotypic variation, suggesting that other genetic and/or environmental factors affect lactase persistence in the populations.

To identify the effects of each SNP on promoter activity, the LCT core promoter was fused to 2 kb of each of the ancestral and derived haplotypes. In the in vitro transcription assay, all fusion constructs show nominal enhancer activity over the promoter’s baseline activity. Promoters driven by the derived alleles showed an 18%–30% increase in expression over the fusions containing ancestral alleles. There was no expression difference between the derived fusions, as was the case for ancestral fusions. The linkage disequilibrium in the region extends for more than 1 Mb for haplotypes with any of the derived alleles. The iHS score for the C-14010 allele was highly significant compared with the iHS empirical distribution for HapMap Yoruba populations. The variants appear to have arisen in the past 1200–23,200 years, with selection coefficients in the range of 0.01 to 0.15. From association study, expression analysis, and presence of an extended haplotype, the G/C-14010 allele appears to be the causal SNP for lactase persistence in several African populations.

The Saudi Arabian population also has a high proportion of lactose tolerant individuals (15). The G-13915 derived allele was found to correlate with disaccharidase activities and the lactase:sucrase ratio, which are a proxy for lactase persistence in an urban Saudi Arabian population (36). Although Imtiaz and colleagues (36) did not specifically search for evidence of positive selection in the population, the presence of lactase persistence may shed light on the origin of some of the African haplotypes.

Knowing where and how transcription factors bind to the region may reveal how SNP differences affect promoter binding and drive differences in expression. Several of the alleles implicated in lactase persistence fall in the Oct-1 transcription factor-binding site (48) (Figure 3). The studies discussed above represent a detailed analysis of the evolution of lactase persistence and derived alleles that are associated with lactase persistence. Lactase persistence has independently evolved at least twice in geographically distinct populations. The study of LCT is a good example of linking positive selection and biological significance. The causal allele was predicted and functional

**Figure 3**

Locations of transcription factor-binding sites and predicted adaptive alleles upstream of LCT, the lactase gene. Three alleles were identified as potentially causal alleles in the African pastoral populations, whereas C/T-13910 was predicted to be the causal allele in Northern Europeans. Additionally, the T/G-13915 allele is correlated with lactase persistence in the Saudi Arabian population. The transcription factors and the sequence they bind in a supershift assay (48) are: HNF-4α (−13854 to −13830), HNF-3α and FOX (−13872 to −13848), Oct-1 and GAGA (−13933 to −13909), and Cdx-2 (−14040 to −14016).
follow-up demonstrated the significance of the variants. These types of detailed statistical and functional analysis are needed for characterizing loci identified by genome-wide scans.

**HUMAN-SPECIFIC CHANGES: FOXP2**

Cognition and the ability to communicate through language are attributes that set humans apart from other species. Because of its complexity, the evolution of cognition is undoubtedly influenced by many interacting factors: genetic, epigenetic, and environmental. The identification of a family with the inability to speak provided the first insight into the presence of a gene (*FOXP2*) that influences human speech (47, 50). It is important to note that the phenotype of individuals affected by *FOXP2* mutations is more complex than the presence or absence of speech. *FOXP2* is a transcription factor that is highly conserved and is in the 5% tail of most conserved sequences in comparisons between human and mouse coding regions (18). Sequence comparison between extant primates and several other mammals revealed very few changes, notably two on the human lineage, one of which is predicted to have functional consequence because it creates a phosphorylation site (18). *FOXP2* appears to be adaptively evolving along the human lineage (96) and has evidence of a recent selective sweep on the basis of intronic variation revealed by Tajima’s D comparison with 313 other genes (73) and comparison with standard, neutral coalescent simulations (18). Using a likelihood approach with the limited amount of available variation data, the selective sweep was dated to between 0 and 120,000 years ago, around the emergence of anatomically modern humans (39).

Krause and colleagues (45) sequenced the two human-specific derived alleles in Neanderthal DNA. The sequence results revealed that the Neanderthal sequence contains the two human-specific variants. Several possible explanations for the observations exist. To date, it has proven quite difficult to extract Neanderthal DNA from bone fragments. Contamination is a pervasive feature of Neanderthal sequence products (88). If the presence of the variants is not an artifact of contamination or DNA damage, the sites may have been fixed or segregating in the ancestral population prior to the human–Neanderthal split. Perhaps because the phenotype is so complex, other epistatic mutations had to occur, leading to the selective sweep in humans. The presence of the derived alleles in both humans and Neanderthals may reflect admixture between *H. neanderthalensis* and *H. sapiens*. Alternatively, the dating method used to estimate the age of the selective sweep in humans may be flawed.

The *FOXP2* case is an interesting example of positive selection because it incorporates both polymorphism and divergence data. Using a disease model provides some insight into the putative functional relevance. Finally, Neanderthal data demonstrate the need for more reliable methods to date the onset of selective pressures.

**Positive Selection on Copy Number**

The small number of amino acid differences between humans and chimpanzees has led to the suggestion that the observed phenotypic diversity may be due to regulatory changes (44). Expression differences exist between populations (35) and can confer different fitness advantages and thus be positively selected. Both copy number variation and noncoding substitutions can affect transcript levels. Therefore, positive selection can potentially act on copy number and on noncoding regions.

Copy number is a highly variable feature of the human genome and copy number polymorphisms exist within and between populations (35, 65, 70, 85). Recently, Perry and colleagues (60) correlated salivary amylase gene (*AMY1*) copy number with dietary starch prevalence. Amylase is an enzyme involved in the metabolism of starch. *AMY1* copy number varies considerably (27, 35), and copy number is positively correlated with salivary amylase protein expression (60). Seven populations...
were analyzed; four populations historically consumed a low-starch diet and the other three consumed a high-starch diet. Mean \textit{AMY1} copy number was higher in the high-starch populations (60). Geographic population distribution and therefore common ancestry were not the best predictors of \textit{AMY1} copy number. Dietary habits were found to more accurately predict copy number, suggesting that positive selection may be acting on \textit{AMY1} copy number. To assess how copy number variation at the \textit{AMY1} locus compares with other copy number variable loci in the human genome, Perry and colleagues (60) compared all copy number variable loci in the Yakut population to the same loci in the Japanese HapMap sample. This empirical distribution of copy number variation revealed that \textit{AMY1} is more differentiated between the two populations than other copy number variable loci. The observed degree of population differentiation further supports the hypothesis of positive selection acting on \textit{AMY1} copy number in high-starch diet populations. The same level of analysis was not conducted for the other high- and low-starch diet populations. The speculated adaptive phenotype is the production of additional \textit{AMY1} protein through increased copy number. The hypothesized fitness advantage of having higher salivary amylase expression is the increased digestion of starchy foods. The higher production of amylase may translate to higher amylase levels in the stomach and intestine (21) for continual enzymatic activity throughout digestion. In addition to the human population data, the fact that chimpanzees do not have multiple \textit{AMY1} copies suggests that \textit{AMY1} copy number has adaptively evolved in human populations. These lines of reasoning suggest that \textit{AMY1} copy number has been under positive selection in recent human evolutionary history (60).

Additional information is necessary to provide evidence of positive selection acting on \textit{AMY1} copy number. For example, detailed sequence analysis of each variant would facilitate the identification of a promoter for each \textit{AMY1} copy and help determine whether each copy generates a transcript and functional protein product. Copy variants may have amino acid differences that affect function; for example, although bonobos appear to have additional \textit{AMY1} copies compared with chimpanzees, the additional copies contain disrupted coding sequences (60). Characterizing copy number variation in additional populations, which may not be categorized into starch or nonstarch dietary preference, will provide additional information about selective pressure and perhaps support the hypothesis of positive selection acting on copy number.

There are several examples of copy number variation associated with disease susceptibility. Copy number variants may confer some selective advantage; however, this hypothesis has not been tested. For example, individuals with a low copy number, relative to the population average, of the CC chemokine ligand 3–like 1 gene (\textit{CCL3L1}), have an increased susceptibility to HIV/AIDS (25). Low \textit{CCL3L1} copy number, along with specific CCR5 mutations, act together to increase susceptibility to HIV infection. Length polymorphisms in the mucin 6 gene (\textit{MUC6}) are associated with \textit{Helicobacter pylori} infection (53). Individuals with shorter \textit{MUC6} alleles are more susceptible to infection by the bacterium. The previous examples are those in which infection susceptibility depends on copy number. Further analysis may reveal evidence for recent positive selection.

Amylase copy number is an interesting potential example of positive selection acting on increased copy number, which is associated with enhanced starch digestion. Statistical methods to correlate copy number with positive selection have not yet been developed. Accurate and high-throughput detection and typing of copy number variants in multiple populations coupled with the development of new statistics may provide insight into copy number variation in human adaptation.

**Positive Selection on Noncoding Genomic Regions**

Noncoding polymorphisms that have conferred selective advantage in recent human history are...
identifiable by polymorphism-based genome-wide scans. For example, several scans have identified gene deserts with evidence for positive selection, which may reflect selection acting on uncharacterized regulatory regions (8, 86, 90). $d_N/d_S$ methods are not applicable to noncoding regions of the DNA because it is not possible to partition each region into synonymous and nonsynonymous sites. Sophisticated methods have been developed to identify positive selection in noncoding regions on longer timescales (32, 41, 63). By comparing *cis*-regulatory containing regions with intronic sequence, Haygood and colleagues (32) found that neural and nutritional genes, which are essential to human differentiation from primates, have been subject to positive selection in the human lineage. By comparing substitution rates in conserved noncoding sequences (CNSs), Prabhakar and colleagues (63) found an excess of human-specific substitutions in CNSs near neuronal genes. CNSs are enriched for regulatory regions (31), which suggests that positive selection may be acting on neuronal regulatory regions (63).

Currently, promoter and enhancer annotation is limited; therefore, it is hard to parse out which regions of the noncoding genome are functionally relevant and which are evolving at neutral rates. Several approaches have been utilized to find human-specific or general regions of noncoding DNA that are under positive selection or have an elevated accumulation of substitutions along a branch (32, 41, 61, 63). These methods are particularly challenging to develop because of context-dependent mutation rate differences (6).

**FUTURE DIRECTIONS AND CHALLENGES**

Genomic scans of polymorphism and divergence data have led to the identification of hundreds of loci predicted to have been subject to positive selection in recent human history. Many of the regions identified by polymorphism-based genome-wide scans show population-specific signatures of selection. Using these results, researchers can begin to evaluate how selective signatures detected in the human genome relate to postagricultural selective pressures and to what extent they reflect older events that occurred along the human lineage after the human-chimpanzee split.

A major advantage of scanning the genome for positive selection is the ability to create an empirical distribution that is not sensitive to assumptions about the populations’ demographic history. However, regardless of the type of distribution that is used, simulated or empirical, many selected loci will be inherently missed (80). Missed loci include those that have a lower selection coefficient or were selected on a different timescale and have recovered from the sweep to a point that is not in the tail of an empirical distribution. Conversely, false positives may be prevalent in the outliers of a distribution. Discriminating between true and false positives will require additional sequence analysis and, ultimately, finding the functional significance of the adaptive change.

Perhaps inflating the false-positive and false-negative rates, demography and ascertainment biases are confounders to studying positive selection with genome-wide scans. Demographic events perturb the site frequency spectrum in a very similar way to positive selection, making it difficult to distinguish between selection and changes in the underlying population structure (64). Ascertainment biases in the genotype data may also confound analyses by skewing the site frequency spectrum (12). The correspondence between candidate-selected loci is limited, even for methods that identify selective events on similar timescales. As it stands, each method is designed to deal with ascertainment biases and correct for potentially confounding demographic histories in a slightly different way. Each method is often evaluated with a highly specific set of simulated data, which is inconsistent between studies. Perhaps a consistent set of simulated data should be used to compare the multitude of scanning methods currently available.

Once a candidate region has been identified to be under positive selection the identification...
Paleosequencing:
DNA sequencing from extinct organisms

of the specific adaptive alleles in selected loci can be difficult, especially in regions whose signature of selection may extend over large genomic regions. However, identifying the causative SNP is imperative for dating the onset of selection in the region and understanding the biological relevance of the change. A simple strategy has been developed to identify the selected allele (69) on the basis of the following assumptions: i) the allele is newly arisen, ii) the allele is frequent only in the population with evidence for a selective sweep, and iii) the allele is annotated as nonsynonymous or in a conserved noncoding region. This strategy provides a starting point for identifying the causative alleles, yet its assumptions are conservative. For example, these strategies will miss cases where selection is acting on standing variation, migration has occurred between populations, or regulatory changes do not occur in conserved noncoding regions.

Once the specific adaptive allele has been identified, it becomes possible to determine the age of the selected allele as a proxy for the onset of selection. Precisely dating the onset of selective pressure has important implications for our understanding of human evolution. Modern humans are estimated to have arisen in Africa approximately 100,000 years ago and subsequently colonized the globe (39). Therefore, an accurate estimate of the age of selected alleles frames the anthropological context of selection and facilitates the identification of population-specific selective events. In this respect, the human-specific substitutions thought to be the target of selection in FOXP2 are relevant. On the basis of variation among extant humans, the amino acid substitutions thought to be the target of selection were originally dated as having arisen in the past 200,000 years (18). This date was contradicted by evidence that the modern human allele is also found in Neanderthals, suggesting the allele predates the human-Neanderthal split ~400,000 years ago (45). This disparity highlights the need to assess the reliability of current methods in estimating the age of selected alleles and to develop new methods that integrate extant population data and Neanderthal data.

Paleosequencing the genome of H. neanderthalensis will provide insight into the selective events specific to the evolution of modern humans. The Neanderthals’ most recent common ancestor to modern humans dates to ~400,000 years ago (58). The Neanderthal genome sequence could validate hypotheses about polymorphisms that have been predicted to be recent (<100,000 years ago) and uniquely human; it will also address the controversy surrounding the possibility of admixture between modern humans and Neanderthals (16, 87). However, these paleosequencing efforts are far from perfect. Human and bacterial DNA contamination, degradation due to double-strand breaks, chemical modification of the bases, and the limited specimen availability make analyses more difficult (26, 58, 88).

The sequencing of additional genomes provides more data for comparative genomics. For example, the macaque genome enables high-confidence ancestral state reconstruction for primates and provides a close outgroup to humans and chimpanzees (24). These data help to identify bursts of adaptive evolution along the human lineage. However, as compared with human, mouse, and chimpanzee sequence alignments for identifying human-specific positive selection (11), the replacement of the mouse sequence with the macaque sequence does not appear to provide much additional statistical power (24). For polymorphism studies, the Human Genome Diversity Project (HGDP) (10) provides a more extensive panel of individuals from several additional populations to understand population demography and selective events (14).

In addition to the vast amount of sequence data being generated for comparative genomics, high-throughput genomics is a cutting-edge technology to survey many individuals across the genome (see sidebar, Genome Scans and Balancing Selection). With 454 (51) and Solexa/Illumina (3) sequencing technologies, the costs of large population-based studies
of human evolution are greatly diminished. Although the new technology reduces the ascertainment bias introduced from genotyping, the methods are not free of problems. Currently, the error rates are higher than traditional Sanger sequencing methods, including systematic errors in 454 sequencing that are not resolved by increasing read depth (97; J. Shendure, personal communication). Using high-throughput genomic techniques to sequence multiple individuals requires additional statistical consideration, including how read lengths and error types affect population genetic inferences.

One of the most important challenges presented by genome-wide scans is the determination of the biological relevance of a locus predicted to be under positive selection. Several methods can be used to determine the function of a locus. One method is to identify disease-causing mutations, as in the case of FOXP2. However, although disease mutations provide insight into the most extreme function of the gene, they are not necessarily indicative of how subtle allele changes will affect normal gene function. Using mouse models is another way to uncover gene function; however, for the loci that are most relevant to human population-specific adaptation, an appropriate mouse model or mouse phenotype may not exist.

In our view, the two most pressing issues in understanding positive selection in the human genome are accurately identifying the causative allele and understanding the functional relevance of the allele that has been subjected to positive selection. What are the anthropological and biological implications of the candidate selected locus? For some cases, such as lactase persistence and disease resistance, this is evident; in other cases, especially when the phenotype is complex, it is much more difficult. Identifying instances of human population-specific adaptation enables population geneticists to elucidate selective pressures in recent human history, characterize disease genes, and understand how human populations have adapted to new environmental challenges. However, before we can gain a clear understanding of these features, the challenges posed by genome-wide scans must be overcome.

**SUMMARY POINTS**

1. Genome-wide scans for evidence of positive selection in the human genome have identified hundreds of candidate selection genes.

**GENOME SCANS AND BALANCING SELECTION**

Genome-wide scans for positive selection have focused on identifying genes that have undergone selective sweeps, but miss other signatures of selection such as balancing selection, which acts to maintain polymorphisms. The focus on selective sweeps is likely due to single nucleotide polymorphism (SNP) ascertainment bias, which skews the site frequency spectrum toward higher frequency SNPs, confounding the ability to identify balancing selection. Reproductive genes are often subject to positive selection in comparisons between species (17, 75, 91). However, polymorphism-based genome-wide scans fail to identify reproductive proteins previously demonstrated to be adaptively evolving. Recent polymorphism surveys using complete resequencing data indicate that two reproductive proteins show the unusual feature of rapid divergence between species as well as a high number of amino acid polymorphisms within species, consistent with balancing selection (23, 30). The regions under selection between species are the same regions that demonstrate high levels of amino acid polymorphisms within humans, suggesting potential adaptive functionality related to fertilization. The high levels of polymorphisms make it unlikely that these loci would be identified by genome-wide scans that use genotyping data. However, with high-throughput sequencing technology and the prospect of complete resequencing of multiple individuals we can look forward to the identification of loci under different selective pressures, such as balancing selection.
2. Lactase persistence has independently evolved at least twice in geographically distinct populations. Alleles upstream of LCT (lactase gene) that are responsible for lactase persistence have recently been subjected to positive selection. Studies of LCT mutations, including the prediction of the causal allele and functional follow-up, provide a comprehensive example of how loci identified by polymorphism-based genome-wide scans should be verified.

3. FOXP2 (forkhead box P2 gene) adaptively evolved along the human lineage and highlights the challenges presented by dating the selected alleles and using Neanderthal sequence data.

4. High AMY1 (amylase gene) copy number correlates with high starch dietary preference. AMY1 is the first locus studied suggesting that positive selection may act on copy number variation.

5. Researchers are currently developing methods that utilize divergence data to identify noncoding regions that have been subject to positive selection along the human lineage.

6. The next major step in studying adaptive evolution in the human genome will be to take candidate loci from genome scans and understand the biological significance of the selected allele.

FUTURE ISSUES

1. Next-generation DNA sequencing technologies will improve our ability to detect positive selection, particularly balancing selection. However, statistical methods must be developed to analyze the vast amount of data.

2. Increasing numbers of test statistics are available for application to genome-wide data. With the limited correspondence between genes identified by scans, a method to compare the sensitivity and specificity of scanning methods is becoming increasingly necessary.

3. Determining the biological relevance of a selected allele is challenging, especially when the phenotype is complex and the selective pressure was exerted thousands of years ago.

4. What is the appropriate strategy for identifying the causal allele?

5. Theoretical models and statistical methods are required to study selection acting on standing variation.

6. Accurately dating the selected allele requires more sophisticated methods than those that currently exist, including data from paleosequencing of Neanderthal DNA.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to acknowledge Cindy Desmarais, Kiran Dhillon, Carole Kelley, Graham McVicker, Molly Przeworski, and Matt Sandel for thoughtful contributions to this review. J.L.K. was
supported by NSF DIG 0709660 and W.J.S. by NSF grant DEB-0716761 and NIH grants HD042563 and HD054631. In this review we tried to focus on a few specific examples; we apologize for not highlighting all the exciting examples of human adaptive evolution.

LITERATURE CITED


58. Describes the Neanderthal genome.

**RELATED RESOURCES**

Haplotter: [http://hg-wen.uchicago.edu/selection/haplotter.htm](http://hg-wen.uchicago.edu/selection/haplotter.htm)
Contents

Human Telomere Structure and Biology
Harold Rietman ................................................................. 1

Infectious Disease in the Genomic Era
Xiaonan Yang, Hongliang Yang, Gangqiao Zhou, and Guo-Ping Zhao ............. 21

ENU Mutagenesis, a Way Forward to Understand Gene Function
Abraham Acevedo-Arozena, Sara Wells, Paul Potter, Michelle Kelly,
Roger D. Cox, and Steve D.M. Brown ........................................ 49

Clinical Utility of Contemporary Molecular Cytogenetics
Bassem A. Bejjani and Lisa G. Shaffer .............................................. 71

The Role of Aminoacyl-tRNA Synthetases in Genetic Diseases
Anthony Antonellis and Eric D. Green ........................................... 87

A Bird’s-Eye View of Sex Chromosome Dosage Compensation
Arthur P. Arnold, Yuichiro Itoh, and Esther Melamed ................................ 109

Linkage Disequilibrium and Association Mapping
B. S. Weir ................................................................................. 129

Positive Selection in the Human Genome: From Genome Scans
to Biological Significance
Joanna L. Kelley and Willie J. Swanson ............................................. 143

The Current Landscape for Direct-to-Consumer Genetic Testing:
Legal, Ethical, and Policy Issues
Stuart Hogarth, Gail Javitt, and David Melzer .................................... 161

Transcriptional Control of Skeletogenesis
Gerard Karsenty ............................................................................ 183

A Mechanistic View of Genomic Imprinting
Ky Sha ....................................................................................... 197

Phylogenetic Inference Using Whole Genomes
Bruce Rannala and Zibeng Yang ...................................................... 217
Transgenerational Epigenetic Effects
Neil A. Youngson and Emma Whitelaw .......................................................... 233

Evolution of Dim-Light and Color Vision Pigments
Shozo Yokoyama ............................................................................................ 259

Genetic Basis of Thoracic Aortic Aneurysms and Dissections: Focus on Smooth Muscle Cell Contractile Dysfunction
Dianna M. Milewicz, Dong-Chuan Guo, Van Tran-Fadulu, Andrea L. Lafont, Christina L. Papke, Sakiko Inamoto, and Hariyadarshi Panunu .............................................. 283

Cohesin and Human Disease
Jinglan Liu and Ian D. Krantz ......................................................................... 303

Genetic Predisposition to Breast Cancer: Past, Present, and Future
Clare Turnbull and Nazneen Rahman ............................................................. 321

From Linkage Maps to Quantitative Trait Loci: The History and Science of the Utah Genetic Reference Project
Stephen M. Prescott, Jean Marc Lalouel, and Mark Leppert .......................... 347

Disorders of Lysosome-Related Organelle Biogenesis: Clinical and Molecular Genetics
Marjan Huizing, Amanda Helip-Wooley, Wendy Westbroek, Meral Gunay-Aygun, and William A. Gahl ................................................................. 359

Next-Generation DNA Sequencing Methods
Elaine R. Mardis ............................................................................................. 387

African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping
Michael C. Campbell and Sarah A. Tishkoff .................................................. 403

Indexes
Cumulative Index of Contributing Authors, Volumes 1–9 ................................. 435
Cumulative Index of Chapter Titles, Volumes 1–9 ........................................... 438

Errata
An online log of corrections to *Annual Review of Genomics and Human Genetics* articles may be found at http://genom.annualreviews.org/