Genetica (2007) 129:7–18
DOI 10.1007/s10709-006-0029-y

ORIGINAL PAPER

Detecting natural selection on *cis*-regulatory DNA

Matthew W. Hahn

Received: 6 July 2004 / Accepted: 25 June 2005 / Published online: 6 September 2006

Abstract Changes in transcriptional regulation play an important role in the genetic basis for evolutionary change. Here I review a growing body of literature that seeks to determine the forces governing the non-coding regulatory sequences underlying these changes. I address the challenges present in studying natural selection without the familiar structure and regularity of protein-coding sequences, but show that most tests of neutrality that have been used for coding regions are applicable to non-coding regions, albeit with some caveats. While some experimental investment is necessary to identify heritable regulatory variation, the most basic inferences about selection require very little functional information. A growing body of research on *cis*-regulatory variation has uncovered all the forms of selection common to coding regions, in addition to novel forms of selection. An emerging pattern seems to be the ubiquity of local adaptation and balancing selection, possibly due to the greater freedom organisms have to fine-tune gene expression without changing protein function. It is clear from multiple single locus and whole genome studies of non-coding regulatory DNA that the effects of natural selection reach far beyond the start and stop codons.

Keywords Promoter · Natural selection · Neutrality · Positive selection · *cis*-regulatory · Binding site

M. W. Hahn (✉)
Department of Biology and School of Informatics, Indiana University, Bloomington, IN 47405, USA
e-mail: mwh@indiana.edu

Introduction

Determining the evolutionary processes that shape both within-species variation and between-species differences in DNA sequences has been a major goal of population genetics for the past 30 years (Lewontin 1974; Kimura 1983; Gillespie 1991; Li 1997). Research into a variety of organisms and cellular processes has now revealed a large of number of coding sequences under either positive selection or balancing selection (reviewed in Yang and Bielawski 2000; Fay and Wu 2001; Nielsen 2001). While estimates of the proportion of non-synonymous mutations that are under selection reveal much about the process of evolution, coding regions make up only a very small fraction of eukaryotic genomes—this likely means that they are a commensurately small fraction of the nucleotide variation within and between species. Recent work has revealed the important role changes in non-coding, *cis*-regulatory sequences have in phenotypic evolution (reviewed in Carroll 2000; Stern 2000; Davidson 2001; Wray et al. 2003). These *cis*-regulatory regions, or promoters, are needed to control the timing, level, and spatial location of transcription for thousands of proteins, and can have evolutionary dynamics much different from the protein-coding regions they control. Though the study of non-coding sequences presents unique problems for population genetics, it will be critical for a complete picture of the phenotypic and fitness consequences of all genetic variation.

Transcriptional regulation is complex, indirect, idiosyncratic, and context dependent (reviewed in Wray et al. 2003). Because of this, distinguishing functional regulatory nucleotides and phenotypically relevant variants from an examination of sequence

alone is currently, and may always be, impossible. One of the main challenges for studies on regulatory sequences, therefore, is the ascertainment of functional regulatory variation that has a consequence for fitness. This challenge results from three main features of *cis*-regulatory DNA: (1) transcription factor binding site motifs are short (6–10 bp) and thus will appear thousands of times in a genome through random chance alone, (2) the proteins that interact with *cis*-regulatory sequences are themselves expressed in an array of times and places, precluding easy biochemical characterization of all necessary binding sites, and (3) functioning binding sites may often appear and disappear with little consequence for either phenotype or fitness. For all of these reasons, discovery of the genetic basis for heritable differences in gene expression will often require biochemical experiments or in vivo expression assays. While experimental difficulties have limited the number of studies exploring the effects of natural selection on relevant non-coding sequences, this has meant that the few studies that there are (Table 1) often have much better functional evidence than analogous studies of protein variation. Another consequence of the experimental investment necessary is that studies on human variation, where evidence abounds on functional promoter polymorphisms (Rockman and Wray 2002), are over-represented.

Despite all of the challenges inherent in studying non-coding DNA, a significant body of work has arisen documenting many modes of evolution in regulatory sequences: negative (purifying) selection preserving regulatory interactions over long periods of time, positive (directional) selection for sexual signaling and adaptation to local habitat, and balancing selection in host–parasite interactions and along environmental clines. Here I review the best examples of the action of natural selection on regulatory variation, and give an in-depth analysis of the methods researchers have used to detect selection. Though each study covered here often uses multiple tests of the neutral–equilibrium hypothesis, and population genetics can sometimes seem like just a laundry list of significant statistical tests, I try to focus on exemplars that highlight the advantages of statistics that are often best at detecting slightly different evolutionary processes. Finally, there are many different ways to group tests of neutrality: the way used here is simply meant to underscore the different experimental and statistical challenges present in studying non-coding sequences.

Interspecific analyses

Comparative analyses of non-coding regions between species offer some of the best evidence for the importance of *cis*-regulatory sequences. A handful of interspecific comparisons between different taxa have revealed that the number of conserved non-coding nucleotides is roughly similar to or greater than the number of conserved coding nucleotides (Shabalina and Kondrashov 1999; Onyango et al. 2000; Bergman and Kreitman 2001; Frazer et al. 2001; Shabalina et al. 2001; Keightley and Gaffney 2003). This large amount of conserved intergenic sequence suggests that there are just as many functional non-coding regulatory nucleotides as there are coding nucleotides, and that, given equal mutation rates, approximately half of all functional variation is found in non-coding regions. In addition, evidence from multiple whole genomes

Table 1 Genes with population genetic evidence for positive or balancing selection in regulatory regions

Locus/taxon	Reference
ftz/*Drosophila melanogaster*	Jenkins et al. (1995)
desat2/*D. melanogaster*	Takahashi et al. (2001)
Est6/*D. melanogaster*	Odgers et al. (2002)
Ldh-B/*Fundulus heteroclitus*	Crawford et al. (1999); Schulte et al. (1997)
Fy/*Homo sapiens*	Hamblin and Di Rienzo (2000)
CCR5/*H. sapiens*	Bamshad et al. (2002)
TNFSF5/*H. sapiens*	Sabeti et al. (2002)
IL4/*H. sapiens*	Rockman et al. (2003); Sakagami et al. (2004)
F7/*H. sapiens*	Hahn et al. (2004); Sabater-Lleal et al. (2006)
AGT/*H. sapiens*	Nakajima et al. (2004)
MMP3/*H. sapiens*	Rockman et al. (2004)
RET/*H. sapiens*	Emison et al. (2005)
PDYN/*H. sapiens*	Rockman et al. (2005)
HLA-G/*H. sapiens*	Tan et al. (2005)
MHC/*Mus musculus*	Cowell et al. (1998)
tb1/*Zea mays*	Wang et al. (1999)

suggests that even non-functioning intergenic sequences may be under weak purifying selection to avoid containing spurious transcription factor binding sites (Hahn et al. 2003). In order to understand the complete role of natural selection in shaping genomes, therefore, we must consider both coding and non-coding sequences.

Close analysis of well-characterized regulatory sequences has revealed conservation of functional elements as a common pattern (reviewed in Hardison 2000; Wray et al. 2003). Specific transcription factor binding sites, or clusters of binding sites, can be conserved over millions of years and may still function in very similar roles (e.g., Aparicio et al. 1995; Frasch et al. 1995; Beckers and Duboule 1998; Margarit et al. 1998; Shashikant et al. 1998; Plaza et al. 1999; Hough et al. 2002). There are also examples of rapid divergence in *cis*-regulatory sequences (McGregor et al. 2001; Dermitzakis and Clark 2002), even when transcriptional output is maintained (e.g., Wu and Brennan 1993; Tamarina et al. 1997; Ludwig et al. 2000; Romano and Wray 2003).

The norm in the study of coding sequences for measuring selective constraint is a comparison of the number of substitutions per site in non-synonymous sites (K_a) to synonymous sites (K_s) (Kimura 1977). A K_a/K_s ratio <1 is consistent with a history of negative selection and constraint, although it does not rule out positive selection, while a K_a/K_s ratio >1 indicates strong positive selection, although it does not mean that negative selection is not also acting (Hughes 1999). By analogy, we can measure the ratio of the substitutions per site in binding sites (K_b) and intervening sites (K_i) in regulatory regions, with the same interpretation of results. In a number of well-characterized *cis*-regulatory regions, we can estimate this ratio: for *DQB1* (0.03/0.077) and *HLA-A* (0/0.40) in primates, $K_b/K_i = 0.39$ and 0, respectively. For *even-skipped* (0.018/0.567) in Drosophila and *leghemoglobin* (0.254/0.853) in legumes, $K_b/K_i = 0.32$ and 0.30, respectively (all unpublished results). These and other limited results (Dermitzakis and Clark 2002; Moses et al. 2003) support the idea that negative selection plays a major role in the conservation of functional regulatory sequences, as it does in coding regions. If no determination of the binding site nucleotides can be made, one can still compare the entire inferred regulatory region to some neutral standard such as synonymous mutations in the adjacent coding region (Kohn et al. 2004; Wong and Nielsen 2004). Although classed tests of this form are generally weaker and tell us about selection at a much coarser scale, they still provide evidence of both positive and negative selection on *cis*-regulatory regions (Kohn et al. 2004; Andolfatto 2005).

Tests that use different classes of sequences (such as K_b/K_i) to compare a presumed selected class of mutations to a presumed neutral class of mutations come with some important caveats in the analysis of promoters. First, because functional *cis*-regulatory sequences are often only characterized in a single species, assignment of functional homology to a set of homologous nucleotides in species comparisons is not always warranted (e.g., Moses et al. 1990). Experimentally verified binding sites present in a well-studied, focal species may be absent in other species; conversely, sequences with no known function in the focal species may actually be binding sites in the other species used in a comparison. Both types of errors will lead to mis-classification of nucleotides and to some error in estimates of the strength and direction of selection. This mis-classification will also occur when regulatory regions have been incompletely characterized: many binding sites will be missed.

A second caveat is that the genetic code of binding sites is unknown. We have little, if any, information on the effect on binding affinity of changes in binding sites. Although the classification of any change within a binding site as selected is therefore largely a hypothesis, it may be just as good as considering any amino acid change in a protein to be functionally relevant. Some up-and-coming technologies (e.g., Mukherjee et al. 2004) hold the promise to elucidate the binding affinities of every nucleotide motif for every transcription factor, and thus to give us a genetic code for binding sites (Bulyk et al. 1999, 2001). Unfortunately, results so far reveal an additional complexity not often considered in coding regions: non-additivity of mutations (i.e., AAA and ATT have equivalent binding affinities, but not AAT; Benos et al. 2002). For evolutionary studies this implies that more distantly related sequences may actually function in a more similar manner.

The third major caveat in applying classed tests of selection to promoter sequences concerns the manner in which positive selection acts. While repeated substitution of amino acids in a protein seems like good evidence for positive selection, it is harder to imagine how this might work in a promoter region. This follows from some important features of *cis*-regulatory sequences: binding sites are often not restricted to specific positions, binding sites arise through point mutation quite often, and multiple changes in a binding site often result in the complete loss of binding affinity (reviewed in Wray et al. 2003). None of these reasons preclude natural selection from acting in this manner, they simply suggest that instances where repeated

substitutions due to directional selection are detectable will be rare. We might imagine a situation, though, where selection acts to abolish multiple binding sites in a region (possibly after gene duplication and divergence in expression domains). This may be a situation where an excess of binding site substitutions is easily detected.

Given all of these caveats, there are in fact a few examples of positive selection being detected through a comparison of binding site substitutions to non-binding site substitutions. Most take advantage of within-species variation, and are covered below under *Classed tests of multisite data*, but we will examine two examples of interspecific comparisons here. The *factor VII* (*F7*) locus produces a coagulation factor important for proper hemostasis in humans. Hahn et al. (2004) sequenced the well-characterized *cis*-regulatory region of this locus in humans and the other great apes and showed that there were a disproportionate number of substitutions within binding sites only along the branch leading to humans. K_b/K_i was significantly greater than 1 along this branch ($K_b/K_i = 14.6$), and hence there was good evidence for repeated positive selection as a force influencing the transcriptional regulation of *F7* in humans. Recent population genetic analyses have confirmed this result (Sabater-Lleal et al. 2006). In a slightly different use of this type of test, Rockman et al. (2005) studied divergence in the *cis*-regulatory region of the *prodynorphin* (*PDYN*) locus in humans and other primates. *Prodynorphin* produces endogenous opiates with major effects on human behavior. The authors found five fixed differences along the lineage leading to modern humans within a 68-bp element known to affect the transcription of *PDYN*. This number of fixations is at least ten times as high as that expected from the rate of substitution in either the flanking non-coding DNA or the *PDYN* coding region, and likelihood-based tests showed it to be highly significant. Further population genetic and functional analyses confirmed the interspecific results (Rockman et al. 2005).

Intraspecific analyses

Population genetic studies of intraspecific variation have benefited from a long history of theoretical models that provide expectations under a variety of selective and demographic conditions (Fisher 1930; Haldane 1932; Wright 1969; Kimura and Ohta 1971; Nei 1987; Hartl and Clark 1997). Since the introduction of the neutral theory of molecular evolution (Kimura 1968; King and Jukes 1969), many population geneticists have focused on determining the expectations for variability under complete neutrality of mutations. These expectations have subsequently been used to construct statistical tests of the neutral hypothesis (e.g., Lewontin and Krakauer 1973; Watterson 1977; Hudson et al. 1987; Tajima 1989; McDonald and Kreitman 1991; Fu and Li 1993). Tests of neutrality (also referred to as 'tests for selection') rely variously upon the amount of differentiation between sub-populations, the amount of variation at a locus relative to divergence, the frequency distribution of variants, and the ratio of selected to selectively neutral variation. Some of the tests are comparisons to a theoretical distribution (e.g., Tajima 1989), and therefore often have their assumptions violated by demographic changes in a population, while others are relatively robust to non-equilibrium situations because they compare different loci or classes of sites that are both subject to the same demographic forces (e.g., McDonald and Kreitman 1991). The tests also vary in their statistical power to detect selection or in the types of selection that will lead them to reject the neutral hypothesis.

For the analysis of *cis*-regulatory sequences, an important distinction between the various tests for selection is the amount of experimental evidence they require. Below I present three general categories of population genetic analyses that each requires a different level of experimental investment in order to be informative. As with any scientific endeavor, the more data available, the better; however, for researchers interested in taking advantage of the large amount of experimental and sequence data already available for many promoter regions, knowing both the advantages and the limits of the framework for statistical inference will be invaluable.

Non-classed tests of multisite data

Non-classed tests of multisite data require the least investment in biochemical characterization of *cis*-regulatory sequences. In fact, no information on functional regulatory sequences or selected and neutral classes of mutations is necessary for the most basic of inferences. The disadvantage of this lack of knowledge is that one has limited ability to identify which mutations may be under selection, or even if the target of selection is definitively in the *cis*-regulatory region examined. The types of population genetic studies considered here all require sequences from multiple individuals in a species across hundreds or thousands of bases; individual tests may also require additional data as described below.

One of the simplest tests of neutrality is the HKA test (Hudson et al. 1987). This test relies on the fact that under neutrality levels of variability and divergence will be directly correlated due to a constant neutral mutation rate. The test uses a comparison of polymorphism to divergence at the locus of interest to identical comparisons at a locus or loci that are presumed neutral. The two main types of selection detected by the HKA test, recent directional selection ('selective sweeps') and balancing selection, are well-illustrated using *cis*-regulatory examples. Genetic analyses have identified *teosinte branched1* (*tb1*) as a major quantitative trait locus for modern corn morphology (Doebley et al. 1995). Sequencing across the coding region and upstream regulatory regions in domesticated maize, researchers found a greatly reduced level of polymorphism only in the *cis*-regulatory region, consistent with a selective sweep (Wang et al. 1999). Comparison to an outgroup showed that the coding and non-coding regions had roughly equivalent levels of divergence (in fact the non-coding DNA showed slightly higher levels of divergence), and thus the HKA test rejected the equivalence of the coding and regulatory regions. In this case, a reduced level of variation was attributed to (artificial) selection on regulatory variation at *tb1* contributing to the corn phenotype, though there was no information on the identity of the nucleotide or nucleotides responsible for the change (which may lie far outside the region sequenced). Bamshad et al. (2002) studied polymorphisms in the 5′ *cis*-regulatory region of the *CC chemokine receptor 5* (*CCR5*) locus in humans. Regulatory variation at this locus is associated with varying susceptibility to HIV-1 and time to progression to AIDS (Bamshad et al. 2002). Sequencing of the *CCR5* promoter revealed disproportionately high levels of polymorphism relative to divergence, consistent with balancing selection acting to maintain multiple alleles. Comparison with variation and divergence at a number of other human loci led the HKA test to reject neutrality. Once again, the inference of balancing selection using the HKA test was made without any specific information about the position of binding sites or the phenotypic effects of particular alleles. And like the *tb1* example, *CCR5* shows very different patterns of evolution in the non-coding and coding regions: a partial selective sweep of a deletion allele in the protein-coding sequence is also detectable in European populations (Stephens et al. 1998).

A number of tests of neutrality use predictions about the mutation frequency distribution (i.e., the number of mutations expected at particular frequencies) to detect selection (e.g., Tajima's *D*, Fu and Li's *D*, *F*, *D**, and *F**; Tajima 1989; Fu and Li 1993). An excess of low-frequency mutations can be evidence for a recent selective sweep, and an excess of high-frequency mutations can be evidence for balancing selection. Largely because these methods require data from only one locus, they are highly sensitive to demographic effects such as population bottlenecks, expansion, or subdivision. In addition, multiple types of selection may all give the same statistical pattern and will therefore be difficult to distinguish (Simonsen et al. 1995). Despite all of these problems, these tests are very popular because of their ease of use and statistical power, and when used carefully or in an analysis using many loci they can tell us much about both selective and demographic processes (see next section for more detail). Odgers et al. (2002) studied variation in the promoter of the *Esterase 6* (*Est6*) locus in *Drosophila melanogaster*. Two main *cis*-regulatory haplotypes, differing at 14 nucleotide sites of unknown individual effect, were shown to differ in their ability to drive expression of *Est6* in the male ejaculatory duct. Population surveys from around the world revealed a great excess of high-frequency mutations within most populations. Tajima's *D* statistic and Fu and Li's *D* and *F* statistics were positive and significant, consistent with balancing selection maintaining the two expression-level haplotypes (Fu and Li's *D** and *F** do not depend on outgroup sequences to polarize mutations, but have less power to detect selection as a result). Once again, these results did not require, but were certainly informed by, expression assays.

Fay and Wu's *H* statistic (Fay and Wu 2000) is similar to the previous statistics measuring the mutation frequency distribution, but is most sensitive to the excess of high-frequency derived mutations that may 'hitchhike' along with a selective sweep. Takahashi et al. (2001) mapped intraspecific differences in cuticular hydrocarbon pheromone levels to a 16-bp deletion in the *cis*-regulatory region of the *desaturase 2* (*desat2*) locus of *D. melanogaster*. Analysis of polymorphism surrounding this functional variant showed an excess of high-frequency derived mutations and a significant *H* test. These pheromones may be involved in mate choice in *D. melanogaster*, suggesting that this regulatory variant is under positive selection for signaling between the sexes (Takahashi et al. 2001).

A third major category of statistical tests of neutrality using polymorphism data from across a locus aims to compare the age of an allele to its frequency. Because the variance in expected frequency of older alleles is quite large, these age-of-allele tests focus on detecting large increases in the frequency of relatively young mutations. These tests seek to identify an excess

of identical haplotypes that would result from an incomplete or ongoing selective sweep (e.g., Hudson et al. 1994; Slatkin 2000; Sabeti et al. 2002; Toomajian et al. 2003; Voight et al. 2006). The tests do not require any prior knowledge of selected mutations, though *a posteriori* identification of a haplotype of interest can lead to reduced power for statistical inference (Hudson et al. 1994). While age-of-allele tests are not easily represented by simple summary statistics, there are a handful of tests based on summary statistics that use expectations on the number of haplotypes observed given a number of polymorphisms to test for selection (e.g., Strobeck's S, Fu's F_s, Depaulis and Veuille's H and K; Strobeck 1987; Fu 1997; Depaulis and Veuille 1998). These summary statistics may be significant under similar selective conditions as age-of-allele tests: an incomplete selective sweep of a single allele and its associated haplotype may greatly reduce the total number of haplotypes in the population.

Two examples of selection on functional *cis*-regulatory mutations may serve to better demonstrate the use of age-of-allele tests. The *matrix metalloproteinase 3* (*MMP3*) locus encodes an enzyme important for the degradation of extracellular matrix in humans. A single nucleotide insertion/deletion polymorphism in the promoter of this gene is associated with large differences in transcriptional output and has been shown to affect transcription factor binding (Humphries et al. 2002). Rockman et al. (2004) studied variation in the frequency of this functional polymorphism in multiple human populations and across 11.9 kb of *MMP3* in two populations. The single-mutation analysis (covered in detail below under Sect. 'Single-mutation tests') showed that the deletion allele had risen to unexpectedly high frequencies among Europeans. Examination of the polymorphism data across the locus revealed that the deletion allele was also on a haplotype identical in 22 of 46 sequenced chromosomes in Europeans, even though there were 35 single nucleotide polymorphisms among all the individuals. Hudson's haplotype test (Hudson et al. 1994) uses coalescent simulations to give the probability of seeing 22 identical haplotypes when 35 mutations are present in a genealogy; this test, as well as the summary statistic tests of total haplotype number, was highly significant for *MMP3*, supporting the hypothesis that the deletion allele has been under directional selection in Europe (Rockman et al. 2004).

Hudson's test requires relatively little sequence data: information on polymorphism is only needed at the locus of interest. Other age-of-allele tests require sequence data from multiple loci up to hundreds of kilobases away from the gene of interest in order to measure how far the selected haplotype extends (e.g., Sabeti et al. 2002; Voight et al. 2006). Sabeti et al. (2002) used their own extended haplotype homozygosity (EHH) test to look for a selective sweep around a *cis*-regulatory polymorphism associated with protection against malaria in the *TNFSF5* locus. Polymorphic sites up to 500 kb away showed linkage disequilibrium with the functional regulatory variant and showed significant EHH. This pattern is consistent with directional selection on the promoter variant causing an increase in frequency of its haplotype.

Classed tests of multisite data

As discussed earlier under Sect. 'Interspecific analyses', tests that compare classes of mutations (i.e., in binding vs. non-binding sites) come with several important caveats in analyses of non-coding DNA. But these types of tests can be very powerful in detecting selection on *cis*-regulatory polymorphism, and, unlike non-classed tests of neutrality, classed tests come much closer to identifying those mutations that are actually under selection. In order to classify mutations as being either within binding site nucleotides or in intervening nucleotides, some amount of biochemical characterization of the *cis*-regulatory region must be done. The simplest experiments consist of degradation of the promoter DNA after a nuclear extract has been washed over it to allow for transcription factor binding. These 'footprinting' assays allow the experimenter to identify those nucleotides that are protected from degradation by protein binding, and classification of nucleotides can be done (Carey and Smale 2000). However, it should be noted that proteins may often protect nucleotides beyond those necessary for binding simply because of their bulk. There may therefore be some mis-classification of nucleotides and mutations when footprinting is the only method used.

The most common classed test of polymorphism data is the McDonald–Kreitman test (McDonald and Kreitman 1991). McDonald and Kreitman suggested a comparison of the ratio of polymorphism to fixed differences of synonymous and non-synonymous mutations. Under neutrality, the ratio of the number of non-synonymous to synonymous polymorphisms should be equal to the ratio of the number of non-synonymous to synonymous fixed differences. An excess of non-synonymous fixed differences can then lead to a rejection of the neutral hypothesis without meeting the extremely restrictive criterion for detecting positive selection using interspecific data alone, i.e., $K_a/K_s > 1$. This test has been used to test for an excess of within binding site substitutions by a number of

researchers (Jenkins et al. 1995; Ludwig and Kreitman 1995; Crawford et al. 1999). [See also Jordan and McDonald (1998) for an interesting but incorrect application of this method to regulatory sequences.] Jenkins et al. (1995) studied variation in the *cis*-regulatory region of the *fushi tarazu* (*ftz*) locus in *D. melanogaster*. They found an excess of fixed differences in nucleotides identified as being responsible for transcription factor binding relative to substitutions in non-binding sites. These results are consistent with repeated positive selection leading to the fixation of regulatory mutations.

As discussed earlier, the McDonald–Kreitman test and tests of the form K_b/K_i are both liable to be slightly inaccurate because of mis-identification of nucleotides as binding or non-binding in the ingroup or outgroup species used (more closely related outgroup species are therefore better for these analyses). There are classed tests of neutrality, though, that do not depend on outgroup comparisons or the number of fixed differences. Hughes and Nei (1988) used the ratio of pairwise non-synonymous to synonymous differences per site *within* species, denoted π_a/π_s, to detect selection in mouse and human major histocompatibility (MHC) loci. The ratio of π_a/π_s was significantly greater than 1, suggesting the action of overdominant, balancing selection. Cowell et al. (1998) applied this logic to test for selection in variation within and between binding sites in the *cis*-regulatory regions of the same MHC loci. They found an analog of π_a/π_s in regulatory regions to be greater than 1, and therefore evidence for balancing selection, in mouse but not man. Their result and further research into this and related regions (Mitchison and Roes 2002; Tan et al. 2005), suggests that selection is acting to maintain both multiple protein variants and multiple expression patterns at histocompatibility loci.

Classed tests can also take advantage of deviations in the frequency spectrum of mutations to detect selection, while controlling for demographic effects. Hahn et al. (2002) suggested comparing statistics such as Tajima's *D* and Fu and Li's *D* calculated separately for non-synonymous and synonymous polymorphisms; demographic processes will affect both types of mutations, but certain forms of selection will only affect non-synonymous mutations. Their Heterogeneity test compares the observed difference in *D* statistics to differences generated by coalescent simulations to find the probability of seeing a difference as great as the one observed. Crawford et al. (1999) studied variation in the *lactose dehydrogenase* (*Ldh-B*) *cis*-regulatory region in the killifish, *Fundulus heteroclitus*. A cline in water temperature along the Atlantic coast corresponds with differences in expression level of *Ldh-B*. Application of the Heterogeneity test to the promoter sequence of this locus (D. Crawford, personal communication) reveals a highly significant difference in frequency spectra between binding and non-binding mutations: Tajima's *D* among non-binding mutations is –1.61, while among binding mutations it is +1.31. A pattern of high-frequency mutations consistent with balancing selection and concordant with the environmental cline is thus revealed, despite a background allele frequency spectrum that is skewed toward low-frequency mutations.

Single-mutation tests

The final type of test of the neutral hypothesis requires frequency data among sub-populations on a single mutation of interest and on several selectively neutral loci. Tests are then structured to compare differentiation among sub-populations at the mutation of interest to the differentiation at neutral loci. Biochemical characterization of the focal *cis*-regulatory region can be limited to showing that the alleles at the site of interest differentially bind a transcription factor through electromobility shift assays. Or, if there are only a few non-coding mutations, this type of test offers the opportunity to distinguish *the* mutation contributing to differences in fitness with little experimental evidence.

The level of differentiation in allele frequencies between sub-populations can be measured by F_{ST} (Wright 1951). Low values of F_{ST} indicate little population differentiation, while high values indicate large amounts of differentiation. High F_{ST} values can indicate positive selection driving changes in allele frequencies in individual populations, while low values can indicate balancing selection maintaining allele frequencies between populations (Lewontin and Krakauer 1973). The amount of differentiation expected between populations in the absence of natural selection is a function of the time since divergence and the effective population size. As such, there is no one value of F_{ST} that indicates whether natural selection has acted between various populations without knowledge of the expected variation due to drift. In order to assess the role of natural selection in causing differences in frequency between populations, therefore, one can compare the F_{ST} at a functional regulatory site to a distribution of F_{ST}s among neutral variants genotyped in the same individuals. This type of single-mutation test has been previously used for inferring selection in coding regions (Karl and Avise 1992; Taylor et al. 1995). Researchers have also used

analytical and simulation methods to identify loci with unusually high or low levels of differentiation (Bowcock et al. 1991; Beaumont and Nichols 1996), or have simply aimed to identify loci in the tails of a large distribution of F_{ST}s (Akey et al. 2002).

Application of single-mutation tests of differentiation to *cis*-regulatory mutations has revealed a number of loci undergoing local adaptation or balancing selection among populations (Bamshad et al. 2002; Hamblin et al. 2002; Rockman et al. 2003; Hahn et al. 2004; Rockman et al. 2004; Rockman et al. 2005); here I will discuss two examples of this type of analysis. Hamblin et al. (2002) studied variation at the *Duffy* (*FY*) blood group locus. A single *cis*-regulatory nucleotide mutation eliminates expression of *FY* and confers resistance to malaria (Tournamille et al. 1995). The site of the null-expression allele, which is fixed in sub-Saharan Africa but is at extremely low frequencies in all other populations, shows an F_{ST} higher than that found in the surrounding sequence and in ten non-functional non-coding regions scattered throughout the genome. There is thus good evidence for local adaptation of human sub-populations to the malarial selective agent through selection on this mutation (Hamblin and Di Rienzo 2000; Hamblin et al. 2002). Rockman et al. (2003) studied a polymorphism in humans whose derived allele creates a new binding site in the *cis*-regulatory region of the *Interleukin-4* (*IL4*) locus. This added binding site increases inducibility of expression of *IL4* and leads to faster reactions to immune system challenges by foreign bodies. While increased inducibility is favored when the body is challenged by harmful pathogens (HIV positive individuals show increased survival with the added binding site; Nakayama et al. 2002), the body may over-react to harmless foreign bodies (asthma and atopic dermatitis, among other diseases, are associated with the added binding site). Rockman et al. (2003) examined the differentiation among multiple human populations of this functional regulatory polymorphism as well as 18 mutually unlinked single nucleotide polymorphisms located far from any protein-coding region. Comparison of F_{ST} at the site of interest to the neutral F_{ST}s revealed a patchy pattern of local selection: certain populations showed increases in frequency of the derived allele well beyond the extent of neutral changes, while other populations showed little change. This pattern may indicate differing trade-offs in selective benefits of the new binding site among human populations (see also Sakagami et al. 2004).

Conclusions and future directions

Despite the many challenges inherent in studying the effects of natural selection on non-coding DNA, a large body of work on this topic is rapidly accumulating. Population genetic studies of *cis*-regulatory variation have now revealed rich and varied histories of adaptation for a number of loci, and have been able to link selective and phenotypic effects for a subset of these. Studies of non-coding DNA will not only enable us to complete our accounting for the effects of selection across every nucleotide of a genome, but has also revealed novel forms of selection (e.g., selection against spurious binding sites; Hahn et al. 2003) and types of mutations not before considered functional (e.g., microsatellites acting as binding sites; Rockman and Wray 2002).

Statistical tests of neutrality that have been traditionally used on coding regions can also be used on non-coding regions as long as several caveats are considered. An understanding of the way *cis*-regulatory DNA interacts with transcription factors and co-factors, and the limits of molecular biology for teasing apart these interactions, is important for understanding the appropriate application of these tests. For the near future the standard of evidence for demonstrating selection on non-coding regions will be higher than for coding; there are several very good examples of rejections of the neutral hypothesis in non-coding regions not discussed here because it is not clear that regulatory sequences are the actual targets of selection (e.g., Makova et al. 2001; Fullerton et al. 2002; Wooding et al. 2002; Bersaglieri et al. 2003; Macdonald and Long 2005). While conservative assessments of the location of selected nucleotides are warranted until a better understanding of promoter structure and function is gained, it is almost certainly true that a significant number of studies of coding regions have actually detected the signature of selection on flanking *cis*-regulatory sequences. There are also some very compelling studies of functional regulatory variation with obvious selective effects, but for which either full population genetic studies have not been done (e.g., Bettencourt et al. 2002; Daborn et al. 2002) or where the action of natural selection cannot be confidently assigned to a specific locus or mutation (e.g., Nurminsky et al. 1998; Michalak et al. 2001; Olsen et al. 2002; Oota et al. 2004; Schlenke and Begun 2004). Future avenues of population genetic research into *cis*-regulatory variation may seek to construct statistical tests of neutrality specifically for non-coding features (such as microsatellites).

The accumulation of data on *cis*-regulatory variation means that we may soon see consistent differences between selection on coding and non-coding regions, and that we will start to be able to make generalizations about the evolution of non-coding DNA itself. For instance, the proportion of functional nucleotides, and the effects of selection, in 5′-flanking, 3′-flanking, and intronic regions containing transcription factor binding sites is an outstanding question in studies of transcriptional regulation (Wray et al. 2003). Several methods may begin to reveal patterns of genome-wide variation in regulatory nucleotides. One promising method compares allele frequencies between populations at hundreds of thousands of loci across the genome, and therefore offers the opportunity to make generalizations about the differences between coding and non-coding differentiation (e.g., Akey et al. 2002; The International HapMap Consortium 2005). Another possible approach to revealing genome-wide patterns of selection on regulatory variation is suggested by the method of Kern et al. (2002). These researchers looked at levels of polymorphism surrounding fixed differences in *Drosophila simulans*. Fixations driven by positive selection are expected to show reduced variation due to hitchhiking, and a comparison of fixations in different selective classes (e.g., non-synonymous vs. synonymous) may reveal significant differences in levels of surrounding variation. An application of this method to an organism (such as *Homo sapiens*) with many mapped regulatory regions will allow for a similar comparison of variation around fixations within and between binding sites.

I have argued here that differences in *cis*-regulatory sequences are a large component of the relevant variation seen by natural selection. The important role for these differences in evolution was predicted by a number of prescient biologists who saw that modular control of transcriptional regulation had the power to decouple protein function from the context in which the proteins were expressed and used (Wallace 1963; Zuckerkandl 1963; Britten and Davidson 1969; Wilson 1975). Indeed, this decoupling may be responsible for the abundance of examples of local adaptation and balancing selection presented here: the ability to control discrete aspects of the expression profile may allow populations to fine-tune expression of proteins in order to adjust to varying biotic and abiotic environmental conditions. In addition, the decoupling of regulatory variation from protein function allows combinations of coding and *cis*-regulatory variation to be constructed such that particular activity variants of proteins (e.g., fast or slow catalysis) may be combined with particular expression variants (e.g., high or low expression) to fine-tune not only the overall expression level of proteins but the expression level of specific protein alleles (e.g., Romey et al. 1999, 2000). Future studies of regulatory variation will surely reveal further examples where this malleability in gene expression has facilitated the action of adaptive natural selection.

Acknowledgments G. Wray, M. Rockman, D. Begun, D. Des Marais, A. Kern, S. Nuzhdin, M. Rausher, J. Stajich, N. Johnson, and two anonymous reviewers all gave constructive comments and criticism.

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814

Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437(7062):1149–1153

Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C et al. (1995) Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. Proc Natl Acad Sci USA 92:1684–1688

Bamshad MJ, Mummidi S, Gonzalez E, Ahuja SS, Dunn DM et al. (2002) A strong signature of balancing selection in the 5' *cis*-regulatory region of *CCR5*. Proc Natl Acad Sci USA 99:10539–10544

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond B Biol Sci 263:1619–1626

Beckers J, Duboule D (1998) Genetic analysis of a conserved sequence in the *HoxD* complex: regulatory redundancy or limitations of the transgenic approach? Dev Dyn 213:1–11

Benos PV, Bulyk ML, Stormo GD (2002) Additivity in protein–DNA interactions: how good an approximation is it? Nucleic Acids Res 30:4442–4451

Bergman CM, Kreitman M (2001) Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Res 11:1335–1345

Bersaglieri T, Drake J, Vanderploeg T, Sabeti PC, Reich DE et al. (2003) Signatures of strong positive selection at the lactase gene. Am J Hum Genet 73:188–188

Bettencourt BR, Kim I, Hoffmann AA, Feder ME (2002) Response to natural and laboratory selection at the *Drosophila hsp70* genes. Evolution 56:1796–1801

Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L et al. (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci USA 88:839–843

Britten RJ, Davidson EH (1969) Gene regulation for higher cells: a theory. Science 165:349–357

Bulyk ML, Gentalen E, Lockhart DJ, Church GM (1999) Quantifying DNA–protein interactions by double-stranded DNA arrays. Nat Biotechnol 17:573–577

Bulyk ML, Huang XH, Choo Y, Church GM (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc Natl Acad Sci USA 98:7158–7163

Carey M, Smale ST (2000) Transcriptional regulation in eukaryotes: concepts, strategies, and techniques. Cold Spring Harbor Laboratory Press, Cold Spring Harbor

Carroll SB (2000) Endless forms: the evolution of gene regulation and morphological diversity. Cell 101:577–580
Cowell LG, Kepler TB, Janitz M, Lauster R, Mitchison NA (1998) The distribution of variation in regulatory gene segments, as present in MHC class II promoters. Genome Res 8:124–134
Crawford DL, Segal JA, Barnett JL (1999) Evolutionary analysis of TATA-less proximal promoter function. Mol Biol Evol 16:194–207
Daborn PJ, Yen JL, Bogwitz MR, Le Goff G, Feil E et al. (2002) A single P450 allele associated with insecticide resistance in *Drosophila*. Science 297:2253–2256
Davidson EH (2001) Genomic regulatory systems: development and evolution. Academic Press, San Diego
Depaulis F, Veuille M (1998) Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol Biol Evol 15:1788–1790
Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol 19:1114–1121
Doebley J, Stec A, Gustus C (1995) *teosinte branched1* and the origin of maize: evidence for epistasis and the evolution of dominance. Genetics 141:333–346
Emison ES, Mccallion AS, Kashuk CS, Bush RT, Grice E et al. (2005) A common sex-dependent mutation in a *RET* enhancer underlies Hirschsprung disease risk. Nature 434:857–863
Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413
Fay JC, Wu CI (2001) The neutral theory in the genomic era. Curr Opin Genet Dev 11:642–646
Fisher RA (1930) The genetical theory of natural selection. The Clarendon, Oxford
Frasch M, Chen XW, Lufkin T (1995) Evolutionary-conserved enhancers direct region-specific expression of the murine *Hoxa-1* and *Hoxa-2* loci in both mice and *Drosophila*. Development 121:957–974
Frazer KA, Sheehan JB, Stokowski RP, Chen XY, Hosseini R et al. (2001) Evolutionarily conserved sequences on human chromosome 21. Genome Res 11:1651–1659
Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147:915–925
Fu Y-X, Li W-H (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709
Fullerton SM, Bartoszewicz A, Ybazeta G, Horikawa Y, Bell GI 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
Galas DJ, Schmitz A (1978) DNAase footprinting: simple method for detection of protein–DNA binding specificity. Nucleic Acids Res 5:3157–3170
Gillespie JH (1991) The causes of molecular evolution. Oxford University Press, New York
Hahn MW, Rausher MD, Cunningham CW (2002) Distinguishing between selection and population expansion in an experimental lineage of bacteriophage T7. Genetics 161:11–20
Hahn MW, Stajich JE, Wray GA (2003) The effects of selection against spurious transcription factor binding sites. Mol Biol Evol 20:901–906
Hahn MW, Rockman MV, Soranzo N, Goldstein DB, Wray GA (2004) Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the *Factor VII* locus in humans. Genetics 167:867–877
Haldane JBS (1932) The causes of evolution. Longman, New York
Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66:1669–1679
Hamblin MT, Thompson EE, Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70:369–383
Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 16:369–372
Hartl DL, Clark AG (1997) Principles of population genetics. Sinauer Associates, Sunderland
Hough RB, Avivi A, Davis J, Joel A, Nevo E et al. (2002) Adaptive evolution of small heat shock protein/αB-crystallin promoter activity of the blind subterranean mole rat, *Spalax ehrenbergi*. Proc Natl Acad Sci USA 99:8145–8150
Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159
Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994) Evidence for positive selection in the superoxide dismutase (*Sod*) regions of *Drosophila melanogaster*. Genetics 136:1329–1340
Hughes AL (1999) Adaptive evolution of genes and genomes. Oxford University Press, Oxford
Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class-I loci reveals overdominant selection. Nature 335:167–170
Humphries SE, Martin S, Cooper J, Miller G (2002) Interaction between smoking and the stromelysin-1 (MMP3) gene 5A/6A promoter polymorphism and risk of coronary heart disease in healthy men. Ann Hum Genet 66
Jenkins DL, Ortori CA, Brookfield JFY (1995) A test for adaptive change in DNA sequences controlling transcription. Proc R Soc Lond B 261:203–207
Jordan IK, McDonald JF (1998) Interelement selection in the regulatory region of the *copia* retrotransposon. J Mol Evol 47:670–676
Karl SA, Avise JC (1992) Balancing selection at allozyme loci in oysters: implications from nuclear RFLPs. Science 256:100–102
Keightley PD, Gaffney DJ (2003) Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proc Natl Acad Sci USA 100:13402–13406
Kern AD, Jones CD, Begun DJ (2002) Genomic effects of nucleotide substitutions in *Drosophila simulans*. Genetics 162:1753–1761
Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626
Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267:275–276
Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge
Kimura M, Ohta T (1971) Theoretical aspects of population genetics. Princeton University Press, Princeton
King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798
Kohn MH, Fang S, Wu C-I (2004) Inference of positive and negative selection on the 5′ regulatory regions of *Drosophila* genes. Mol Biol Evol 21:374–383
Lewontin RC (1974) The genetic basis for evolutionary change. Columbia University Press, New York
Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of selective neutrality of polymorphisms. Genetics 74:175–195

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland

Ludwig MZ, Kreitman M (1995) Evolutionary dynamics of the enhancer region of even-skipped in *Drosophila*. Mol Biol Evol 12:1002–1011

Ludwig MZ, Bergman C, Patel NH, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403:564–567

Macdonald SJ, Long AD (2005) Prospects for identifying functional variation across the genome. Proc Natl Acad Sci USA 102:6614–6621

Makova KD, Ramsay M, Jenkins T, Li WH (2001) Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter. Genetics 158:1253–1268

Margarit E, Guillen A, Rebordosa C, Vidal-Taboada J, Sanchez M et al. (1998) Identification of conserved potentially regulatory sequences of the SRY gene from 10 different species of mammals. Biochem Biophy Res Commun 245:370–377

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351:652–654

McGregor AP, Shaw PJ, Hancock JM, Bopp D, Hediger M et al. (2001) Rapid restructuring of *bicoid*-dependent *hunchback* promoters within and between Dipteran species: implications for molecular coevolution. Evol Dev 3:397–407

Michalak P, Minkov I, Helin A, Lerman DN, Bettencourt BR et al. (2001) Genetic evidence for adaptation-driven incipient speciation of *Drosophila melanogaster* along a microclimatic contrast in ''Evolution Canyon,'' Israel. Proc Natl Acad Sci USA 98:13195–13200

Mitchison NA, Roes J (2002) Patterned variation in murine MHC promoters. Proc Natl Acad Sci USA 99:10561–10566

Moses K, Heberlein U, Ashburner M (1990) The Adh gene promoters of *Drosophila melanogaster* and *Drosophila orena* are functionally conserved and share features of sequence structure and nuclease-protected sites. Mol Cell Biol 10:539–548

Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB (2003) Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evol Biol 3

Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D et al. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat Genet 36:1331–1339

Nakajima T, Wooding S, Sakagami T, Emi M, Tokunaga K et al. (2004) Natural selection and population history in the human angiotensinogen gene (AGT): 736 complete AGT sequences in chromosomes from around the world. Am J Hum Genet 74:898–916

Nakayama EE, Meyer L, Iwamoto A, Persoz A, Nagai Y et al. (2002) Protective effect of interleukin-4 -589T polymorphism on human immunodeficiency virus type 1 disease progression: relationship with virus load. J Infect Dis 185:1183–1186

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. Heredity 86:641–647

Nurminsky DI, Nurminskaya MV, Deaguiar D, Hartl DL (1998) Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature 396:572–575

Odgers WA, Aquadro CF, Coppin CW, Healy MJ, Oakeshott JG (2002) Nucleotide polymorphism in the *Est6* promoter, which is widespread in derived populations of *Drosophila melanogaster*, changes the level of Esterase 6 expressed in the male ejaculatory duct. Genetics 162:785–797

Olsen KM, Womack A, Garrett AR, Suddith JI, Purugganan MD (2002) Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. Genetics 160:1641–1650

Onyango P, Miller W, Lehoczky J, Leung CT, Birren B et al. (2000) Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain. Genome Res 10:1697–1710

Oota H, Pakstis AJ, Bonne-Tamir B, Goldman D, Grigorenko E et al. (2004) The evolution and population genetics of the *ALDH2* locus: random genetic drift, selection, and low levels of recombination. Ann Hum Genet 68:93–109

Plaza S, Saule S, Dozier C (1999) High conservation of *cis*-regulatory elements between quail and human for the *Pax-6* gene. Dev Genes Evol 209:165–173

Rockman MV, Wray GA (2002) Abundant raw material for *cis*-regulatory evolution in humans. Mol Biol Evol 19:1991–2004

Rockman MV, Hahn MW, Soranzo N, Goldstein DB, Wray GA (2003) Positive selection on a human-specific transcription factor binding site regulating *IL4* expression. Curr Biol 13:2118–2123

Rockman MV, Hahn MW, Soranzo N, Loisel DA, Goldstein DB, Wray GA (2004) Positive selection on *MMP3* regulation has shaped heart disease risk. Curr Biol 14:1531–1539

Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, Wray GA (2005) Ancient and recent positive selection transformed opioid *cis*-regulation in humans. PLoS Biol 3:e387

Romano LA, Wray GA (2003) Conservation of *Endo16* expression in sea urchins despite divergence in both *cis* and *trans*-acting components of transcriptional regulation. Development 130:4187–4199

Romey MC, Guittard C, Chazalette JP, Frossard P, Dawson KP et al. (1999) Complex allele [-102T>A+S549R(T>G)] is associated with milder forms of cystic fibrosis than allele S549R[T>G] alone. Hum Genet 105:145–150

Romey MC, Pallares-Ruiz N, Mange A, Mettling C, Peytavi R et al. (2000) A naturally occurring sequence variation that creates a YY1 element is associated with increased cystic fibrosis transmembrane conductance regulator gene expression. J Biol Chem 275:3561–3567

Sabater-Lleal M, Soria JM, Bertranpetit J, Almasy L, Blangero J, Fontcuberta J, Calafell F (2006) Human *F7* sequence is split into three deep clades that are related to FVII plasma levels. Hum Genet 118:741–751

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837

Sakagami T, Witherspoon DJ, Nakajima T, Jinnai N, Wooding S et al. (2004) Local adaptation and population differentiation at the *interleukin 13* and *interleukin 4* loci. Genes Immun 5:389–397

Schlenke TA, Begun DJ (2004) Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. Proc Natl Acad Sci USA 101:1626–1631

Schulte PM, Gomez-Chiarri M, Powers DA (1997) Structural and functional differences in the promoter and 5′ flanking region of *Ldh-B* within and between populations of the teleost *Fundulus heteroclitus*. Genetics 145:759–769

Shabalina SA, Kondrashov AS (1999) Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. Genet Res 74:23–30

Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS (2001) Selective constraint in intergenic regions of human and mouse genomes. Trends Genet 17:373–376

Shashikant CS, Kim CB, Borbely MA, Wang WCH, Ruddle FH (1998) Comparative studies on mammalian *Hoxc8* early enhancer sequence reveal a baleen whale-specific deletion of a *cis*-acting element. Proc Natl Acad Sci USA 95:15446–15451

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141:413–429

Slatkin M (2000) Allele age and a test for selection on rare alleles. Philos Trans R Soc Lond B Biol Sci 355:1663–1668

Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW et al. (1998) Dating the origin of the *CCR5*-Δ32 AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62:1507–1515

Stern DL (2000) Perspective: evolutionary developmental biology and the problem of variation. Evolution 54:1079–1091

Stone JR, Wray GA (2001) Rapid evolution of *cis*-regulatory sequences via local point mutations. Mol Biol Evol 18:1764–1770

Strobeck C (1987) Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics 117:149–153

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Takahashi A, Tsaur SC, Coyne JA, Wu CI (2001) The nucleotide changes governing cuticular hydrocarbon variation and their evolution in *Drosophila melanogaster*. Proc Natl Acad Sci USA 98:3920–3925

Tamarina NA, Ludwig MZ, Richmond RC (1997) Divergent and conserved features in the spatial expression of the *Drosophila pseudoobscura esterase-5B* gene and the *esterase-6* gene of *Drosophila melanogaster*. Proc Natl Acad Sci USA 94:7735–7741

Tan Z, Shon AM, Ober C (2005) Evidence of balancing selection at the HLA-G promoter region. Hum Mol Genet 14:3619–3628

Taylor MFJ, Shen Y, Kreitman ME (1995) A population genetic test of selection at the molecular level. Science 270:1497–1499

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

Toomajian C, Ajioka RS, Jorde LB, Kushner JP, Kreitman M (2003) A method for detecting recent selection in the human genome from allele age estimates. Genetics 165:287–297

Tournamille C, Colin Y, Cartron JP, Le Van Kim C (1995) Disruption of a GATA motif in the *Duffy* gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat Genet 10:224–228

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:e72

Wallace B (1963) Genetic diversity, genetic uniformity, and heterosis. Can J Genet Cytol 5:239–253

Wang RL, Stec A, Hey J, Lukens L, Doebley J (1999) The limits of selection during maize domestication. Nature 398:236–239

Watterson GA (1977) Heterosis or neutrality? Genetics 85:789–814

Wilson AC (1975) Evolutionary importance of gene regulation. Stadler Symp 7:117–134

Wong WSW, Nielsen R (2004) Detecting selection in noncoding regions of nucleotide sequences. Genetics 167:949–958

Wooding SP, Watkins WS, Bamshad MJ, Dunn DM, Weiss RB et al. (2002) DNA sequence variation in a 3.7-kb noncoding sequence 5′ of the *CYP1A2* gene: implications for human population history and natural selection. Am J Hum Genet 71:528–542

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M et al. (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20:1377–1419

Wright S (1951) The genetical structure of populations. Ann Eugen 15:323–354

Wright S (1969) The theory of gene frequencies. The University of Chicago Press, Chicago

Wu CY, Brennan MD (1993) Similar tissue-specific expression of the *Adh* genes from different *Drosophila* species is mediated by distinct arrangements of *cis*-acting sequences. Mol Gen Genet 240:58–64

Yang ZH, Bielawski JP (2000) Statistical methods for detecting molecular evolution. Trends Ecol Evol 15:496–503

Zuckerkandl E (1963) Perspectives in molecular anthropology. In: Washburn SL (ed) Classification and human evolution. Aldine, Chicago, pp. 243–272