

Statistical methods for detecting natural selection from genomic data

Yoshiyuki Suzuki*

*Graduate School of Natural Sciences, Nagoya City University, 1 Yamanohata,
Mizuho-cho, Mizuho-ku, Nagoya-shi, Aichi-ken 467-8501, Japan*

(Received 12 September 2010, accepted 6 December 2010)

In the study of molecular and phenotypic evolution, understanding the relative importance of random genetic drift and positive selection as the mechanisms for driving divergences between populations and maintaining polymorphisms within populations has been a central issue. A variety of statistical methods has been developed for detecting natural selection operating at the amino acid and nucleotide sequence levels. These methods may be largely classified into those aimed at detecting recurrent and/or recent/ongoing natural selection by utilizing the divergence and/or polymorphism data. Using these methods, pervasive positive selection has been identified for protein-coding and non-coding sequences in the genomic analysis of some organisms. However, many of these methods have been criticized by using computer simulation and real data analysis to produce excessive false-positives and to be sensitive to various disturbing factors. Importantly, some of these methods have been invalidated experimentally. These facts indicate that many of the statistical methods for detecting natural selection are unreliable. In addition, the signals that have been believed as the evidence for fixations of advantageous mutations due to positive selection may also be interpreted as the evidence for fixations of deleterious mutations due to random genetic drift. The genomic diversity data are rapidly accumulating in various organisms, and detection of natural selection may play a critical role for clarifying the relative role of random genetic drift and positive selection in molecular and phenotypic evolution. It is therefore important to develop reliable statistical methods that are unbiased as well as robust against various disturbing factors, for inferring natural selection.

Key words: statistical methods, recurrent natural selection, recent/ongoing natural selection, divergence data, polymorphism data

INTRODUCTION

In the study of molecular and phenotypic evolution, understanding the relative importance of random genetic drift on selectively neutral or nearly neutral mutations and positive selection on advantageous mutations as the mechanisms for driving divergence between populations (species) and maintaining polymorphism within populations (species) has been a central issue. For this purpose, it is critical to detect natural selection. The neutral theory of molecular evolution provides the null hypothesis of selective neutrality, which can be tested with statistical methods using nucleotide and amino acid sequence data for detecting positive and negative selection at the molecular level (Kimura, 1968, 1983).

Detection of positive and negative selection is also of practical use. Identifying negatively selected amino acid sites in proteins among vertebrates may be useful for predicting deleterious mutations associated with human diseases (Ng and Henikoff, 2001; Sunyaev et al., 2001; Chun and Fay, 2009). It has also been reported that disease-associated genes were often positively selected during evolution (Medawar, 1946; Neel, 1962; Clark et al., 2003; Nielsen et al., 2005a; Suzuki, 2006a, but see also Bustamante et al., 2005; The Chimpanzee Sequencing and Analysis Consortium, 2005). In pathogens, detection of positively and negatively selected amino acid sites may be useful for predicting epitopes recognized by host immune systems (Suzuki and Gojobori, 2001) and targets of effective vaccines and drugs (Suzuki, 2004a, 2006b), respectively.

A variety of statistical methods has been developed for detecting positive and negative selection at the molecular

Edited by Fumio Tajima

* Corresponding author. E-mail: yossuzuk@nsc.nagoya-cu.ac.jp

level using nucleotide and amino acid sequence data. These methods may be largely classified into those aimed at detecting recurrent and/or recent/ongoing natural selection by utilizing divergence data between populations and/or polymorphism data within populations.

METHODS FOR DETECTING RECURRENT NATURAL SELECTION USING DIVERGENCE DATA BETWEEN POPULATIONS

Mutations are ultimately the change in genomic nucleotide sequences. They are advantageous, neutral, or deleterious if they confer a greater, equal, or smaller fitness, respectively, to the mutant individuals compared with the average in the population. Since the probability of fixation for advantageous mutations in a population is greater than that for neutral mutations, which in turn is greater than that for deleterious mutations (Kimura, 1977, 1983), positive and negative selection operating recurrently for a certain type of mutation may be inferred when the substitution rate of that mutation is greater or smaller than the substitution rate of the neutral mutation, respectively (Hughes and Nei, 1988).

Amino acid sequence level Point mutations occurring in a protein-coding nucleotide sequence are synonymous or nonsynonymous according to whether they preserve or alter coding amino acids, respectively. If we assume that synonymous mutations are selectively neutral or nearly neutral, natural selection operating at the amino acid sequence level may be inferred by comparing the rate of nonsynonymous substitution (r_N) with that of synonymous substitution (r_S), where the relationships $r_N > r_S$, $r_N < r_S$, and $r_N = r_S$ indicate positive, negative, and no selection, respectively (Hughes and Nei, 1988).

All-sites analysis Natural selection operating over the entire region of a protein may be detected by comparing r_S and r_N between a pair of protein-coding nucleotide sequences. For this purpose, the number of synonymous substitutions per synonymous site (d_S) and that of nonsynonymous substitutions per nonsynonymous site (d_N) are estimated. Since d_S and d_N are expected to be $2r_S t$ and $2r_N t$, respectively, where t denotes the time elapsed from the divergence of two sequences compared, the relationships $d_N > d_S$, $d_N < d_S$, and $d_N = d_S$ indicate positive ($r_N > r_S$), negative ($r_N < r_S$), and no ($r_N = r_S$) selection, respectively (Fig. 1A). Many methods have been developed for estimating d_S and d_N , which are largely classified into the mutation fraction method (Miyata and Yasunaga, 1980; Perler et al., 1980; Nei and Gojobori, 1986; Kondo et al., 1993; Ina, 1995; Zhang et al., 1998; Yang and Nielsen, 2000; Suzuki et al., 2009), the degenerate site method (Li et al., 1985; Li, 1993; Pamilo and Bianchi, 1993; Comeron, 1995), and the codon model method (Goldman and Yang, 1994; Muse and Gaut, 1994; Muse,

1996; Yang and Nielsen, 1998) (for review, see Nei and Kumar, 2000; Suzuki and Gojobori, 2003).

Window analysis In a protein molecule, different amino acid sites may perform different functions, where the direction and magnitude of natural selection may vary. A window analysis method was developed to compare d_S and d_N for a group of codon sites located in close proximity in the linear sequence (Fig. 1Ba). By sliding the window from the 5'-end to the 3'-end of the sequence, regions with $d_N > d_S$ and $d_N < d_S$ are identified as positively and negatively selected, respectively (Clark and Kao, 1991). This method is based on the assumption that functions of amino acid sites located in close proximity in the linear sequence are similar. In fact, however, functions are determined in the context of three-dimensional structure of proteins. Consequently, the three-dimensional window analysis method was developed for detecting natural selection for a group of amino acid sites located in close proximity in the tertiary structure of proteins (Fig. 1Bb) (Suzuki, 2004c; Berglund et al., 2005). This approach appeared to be more sensitive than the classical one-dimensional approach in detecting positive selection. A sliding window may also be used to scan the phylogenetic tree for detecting chronological changes in natural selection (Fig. 1Bc) (Suzuki, 2008b).

Single-site analysis By narrowing the window size for comparison of d_S and d_N to 1, it becomes possible to detect natural selection at single amino acid sites. Here the pairwise analysis of protein-coding nucleotide sequences is no longer appropriate, and many sequences are compared simultaneously using the phylogenetic tree to observe sufficiently large numbers of synonymous and nonsynonymous substitutions for obtaining a statistical significance at single codon sites (Fig. 1C). Parsimony (Fitch et al., 1997; Suzuki and Gojobori, 1999; Suzuki et al., 2009), likelihood (Suzuki, 2004b; Kosakovsky Pond and Frost, 2005; Massingham and Goldman, 2005), and Bayesian (Nielsen and Yang, 1998; Yang et al., 2000) methods have been developed for this purpose. It should be noted that a linkage to advantageous or deleterious mutations does not affect the fixation probability of neutral mutations but decreases and increases those of advantageous and deleterious mutations, respectively, suggesting that the linkage reduces the effect of natural selection (Hill and Robertson, 1966; Birky and Walsh, 1988). Recombination, however, rescues the effect, although it does not affect the fixation probability of neutral mutations.

From the studies using computer simulation, it was reported that the Bayesian method produces more positive results than the parsimony method (Wong et al., 2004). For this reason, the former method has been frequently used for detecting positively selected amino acid sites in various proteins. However, it was also reported that the Bayesian method tends to produce excessive

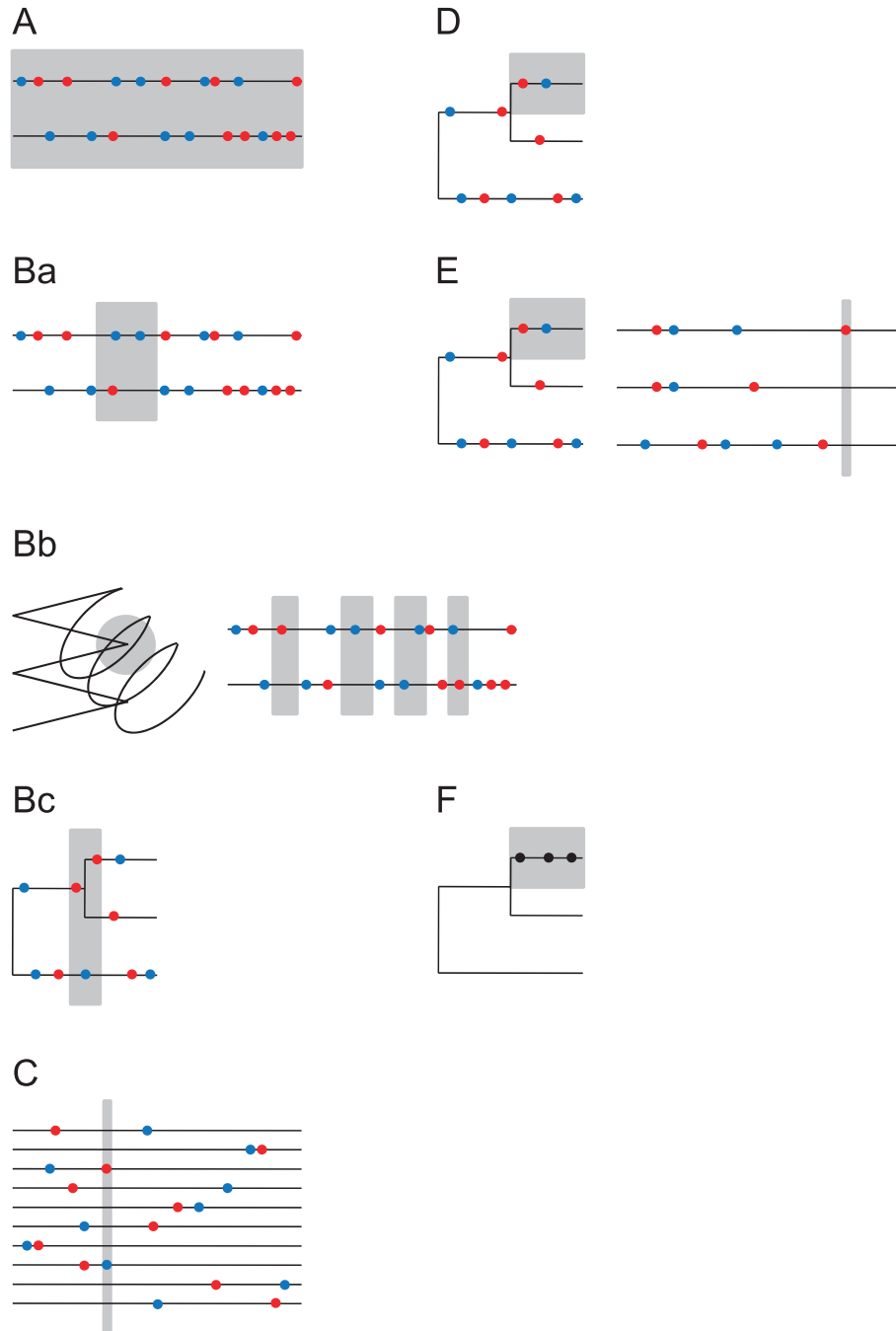


Fig. 1. Schematic diagrams for the divergence data analyzed by the statistical methods for detecting recurrent natural selection with (A) all-sites analysis, (Ba) one-dimensional window analysis, (Bb) three-dimensional window analysis, (Bc) phylogenetic window analysis, (C) single-site analysis, (D) branch method, (E) branch-site method, and (F) accelerated evolution of conserved non-coding sequences. Black horizontal lines in A, Ba, Bb, C, and E denote hypothetical nucleotide sequences. Hypothetical phylogenetic tree and three-dimensional protein structure are drawn in Bc, D, E, and F, and in Bb, respectively. Black, blue, and red dots indicate hypothetical nucleotide, synonymous, and nonsynonymous substitutions, respectively. The regions where the statistical test is conducted are shaded in gray.

false-positives in computer simulation (Suzuki and Nei, 2002; Berlin and Smith, 2005) and real data analysis (Suzuki and Nei, 2001, 2004; Friedman and Hughes, 2007). In particular, it was shown from the analysis of

rhodopsins and opsins in vertebrates (Hughes, 2008b; Yokoyama et al., 2008; Nozawa et al., 2009a, b) and an odorant receptor in primates (Zhuang et al., 2009) that positively selected sites identified by the Bayesian and

likelihood methods mostly disagreed with experimentally determined adaptive sites. Detection of natural selection at a site is also influenced by configuration of natural selection at other sites in the Bayesian method (Suzuki, 2006c). It should be noted that, since these methods are aimed at detecting recurrent positive selection, multiple amino acid substitutions should be observed for a site to be identified as positively selected. In fact, it has been reported from the theoretical study that the minimum number of nonsynonymous substitutions required for detecting positive selection is 4 (Suzuki, 2008a). In reality, however, positive selection is sometimes detected in the Bayesian method even when the number of nonsynonymous substitutions observed is smaller than 4 (Suzuki, 2006c).

Branch analysis In addition to the possibility that different amino acid sites in a protein molecule perform different functions, it is also possible that the function of an amino acid site changes during evolution, due to substitutions at this or other sites of the protein and alterations of micro-environments surrounding the protein. The small-sample test based on the parsimony principle (Zhang et al., 1997), which is mainly applicable to the analysis of closely related sequences, and the likelihood method (Yang, 1998) have been developed for detecting natural selection operating over the entire region of a protein at a branch of the phylogenetic tree by comparing d_S and d_N (Fig. 1D).

Branch-site analysis The branch-site method based on the Bayesian principle has been developed for detecting positively selected amino acid sites at a specific (foreground) branch of the phylogenetic tree taking into account the heterogeneity of r_N/r_S among sites and branches (Fig. 1E) (Yang and Nielsen, 2002; Zhang et al., 2005). Codon sites in protein-coding nucleotide sequences are assumed to be classified into 4 classes, each having different r_N/r_S values at the foreground and other (background) branches. The (selection) model with positive selection at the foreground branch is compared with the null (neutral) model without positive selection using the likelihood-ratio test, and positive selection is inferred when the null model is rejected.

In computer simulation, the false-positive rate of the branch-site method appeared to exceed the significance level under certain conditions (Zhang, 2004; Zhang et al., 2005; Bakewell et al., 2007; Suzuki, 2008a; Nozawa et al., 2009a, b). Positive selection was falsely detected even when the total number of nonsynonymous substitutions over the entire sequence was less than 4, which is the theoretically minimum value required for detecting positive selection (Suzuki, 2008a). The existence of a single codon site with the occurrence of multiple nonsynonymous substitutions in the entire sequence sometimes appeared to be sufficient for erroneously detecting positive selection. This method was also found to produce

spurious results of positive selection when estimates of d_N/d_S were inflated due to episodes of GC-biased gene conversion (Berglund et al., 2009; Galtier et al., 2009; Hurst, 2009) and each of sequencing, annotation, and alignment errors (Mallick et al., 2009; Schneider et al., 2009; Fletcher and Yang, 2010). From the analysis of rhodopsins in vertebrates, it was shown that positively selected sites identified by the branch-site method mostly disagreed with experimentally determined adaptive sites (Nozawa et al., 2009b). Since this method is aimed at detecting amino acid sites where positive selection operated recurrently, multiple amino acid substitutions should be observed for a site at a branch to be identified as positively selected. However, the maximum number of nonsynonymous substitutions observable for a codon site at a branch is 3, which is smaller than the theoretically minimum number of nonsynonymous substitutions (4) required for detecting positive selection (Suzuki, 2008a).

Nucleotide sequence level Along with the advancement of functional genomics, it is becoming increasingly evident that a significant fraction of non-coding nucleotide sequences in the genomes of various organisms are functional, e.g., siRNA, miRNA, and piRNA. In addition, comparative genomics studies revealed that many non-coding sequences are highly conserved among diverse organisms, e.g., ultra-conserved non-coding elements.

Natural selection recurrently operating at the nucleotide sequence level may be detected in a similar way to detecting that at the amino acid sequence level, by comparing the substitution rate of the nucleotide mutation of interest with that of the neutral mutation. Acceleration of nucleotide substitution in a particular lineage of the phylogenetic tree for otherwise highly conserved non-coding elements is also considered as a signature of positive selection.

Comparison with the neutral substitution rate The substitution rate of the nucleotide mutation of interest is compared with that of the neutral mutation; positive or negative selection is inferred if the former rate is greater or smaller than the latter rate, respectively. Mutations at synonymous sites or introns are usually assumed to be selectively neutral. The Bayesian method for detecting natural selection at single amino acid sites (Wong and Nielsen, 2004) and the branch-site method (Haygood et al., 2007), which are described above, have been modified for detecting natural selection at the nucleotide sequence level.

Accelerated evolution of conserved non-coding sequences It has been proposed that acceleration of nucleotide substitution in a particular lineage of the phylogenetic tree for otherwise highly conserved non-coding sequences is a signature of positive selection (Fig. 1F) (Prabhakar et al., 2006). Using this strategy, positive

selection was inferred to have operated in the human lineage for 992 of 110,549 (Prabhakar et al., 2006), 202 of 34,498 (Pollard et al., 2006), and 1,356 of 304,291 (Bird et al., 2007) conserved non-coding elements. Interestingly, however, only 15 elements were shared among these three results, raising a concern about their reliability (Bird et al., 2007).

It should be noted that in this approach the accelerated substitution rate is not required to exceed the neutral substitution rate for detecting positive selection. Therefore, it is usually difficult to distinguish whether acceleration was due to positive selection or relaxation of functional constraint. In addition, the GC-biased gene conversion, which is dependent on the existence of recombination hotspots, may also cause acceleration of nucleotide substitution at a particular lineage, because recombination hotspots are usually short-lived (Prabhakar et al., 2008; Duret and Galtier, 2009).

Problems in detecting recurrent natural selection For detecting recurrent natural selection operating at the amino acid and nucleotide sequence levels, it is customary to test the null hypothesis of equal substitution rates for the mutation of interest and the neutral mutation. Mutations at synonymous sites, pseudogenes, or introns are usually assumed to be neutral. However, the null hypothesis may be rejected even when the mutation of interest is neutral, if the mutation rates are different between the sites compared or the neutral mutation assumed is not neutral in reality (for review, see Filipowski et al., 2007). Alignment errors may also be taken into account in the comparison of d_S and d_N (Wong et al., 2008).

Unequal mutation rates In vertebrates and plants, methylated CpG dinucleotides are known to be hypermutable. Therefore, the rates of synonymous and nonsynonymous mutations may not be the same if the proportions of synonymous and nonsynonymous sites overlapping with CpG dinucleotides are different in these organisms (Subramanian and Kumar, 2006). In addition, the GC-biased gene conversion, which is dependent on the existence of short-living recombination hotspots, may occur episodically in the mammalian genome, where the background mutation pattern is usually AT-biased (Berglund et al., 2009; Galtier et al., 2009). Therefore, the rates of synonymous and nonsynonymous mutations may differ if the GC contents at the synonymous and nonsynonymous sites are different. Primate promoters may be subject to a higher mutation rate than other genomic regions because of biased sequence composition, atypical methylation, and exceptionally accessible chromatin structure (Taylor et al., 2008). Methods for comparing d_S and d_N taking into account the hypermutability at CpG dinucleotides have been developed in both the all-sites and single-site analyses (Suzuki et al., 2009).

Functional constraints on synonymous mutations It has been observed that the substitution rate at synonymous sites was smaller than that at non-coding sites, indicating that synonymous mutations were under functional constraint (Hellmann et al., 2003; Resch et al., 2007; Parsch et al., 2010; Zhou et al., 2010). It was proposed that mRNAs containing codons that are recognized by less abundant tRNAs tends to be mistranslated. Since mistranslated proteins may be misfolded and toxic, natural selection may operate to form the codon usage bias toward codons recognized by more abundant tRNAs (Drummond and Wilke, 2008). In addition, synonymous mutations may affect phenotypes and cause diseases (Cartegni et al., 2002) by influencing splicing, miRNA binding, RNA editing, mRNA stability, and translation pausing, sometimes through changing the secondary structure of mRNAs (Chamary et al., 2006; Kimchi-Sarfaty et al., 2007; Watts et al., 2009). Negative selection on synonymous mutations may also operate locally, e.g., exonic splicing enhancers and silencers near exon-intron junctions, such that the estimates of d_N/d_S may erroneously exceed 1 without positive selection only at specific codon sites (Hurst and Pal, 2001). Negative selection may operate on introns if they contain snoRNAs and miRNAs, or form secondary structures. The first introns and the intronic sites close to exon-intron junctions may be constrained if they possess transcription control signals and splicing control elements, respectively (Keightley and Gaffney, 2003). The effect of natural selection operating at the nucleotide sequence level should be corrected in the comparison of d_S and d_N (Tamura et al., 2004; Subramanian and Kumar, 2003, 2006; Yang and Nielsen, 2008).

Reverse-translation of aligned amino acid sequences In the analysis of natural selection by comparing d_S and d_N , homologous codons are often identified by reverse-translating aligned amino acid sequences into codon sequences. In this method, however, homologous codons may be mis-identified when frame-shifts occurred (Mills et al., 2006) or amino acid sequences were misaligned (Wong et al., 2008; Liu et al., 2009; Fletcher and Yang, 2010), which may lead to overestimation of the d_N/d_S ratio (Wong et al., 2008; Mallick et al., 2009; Schneider et al., 2009) and false detection of positively selected amino acid sites (Vamathevan et al., 2008; Wong et al., 2008; Mallick et al., 2009; Schneider et al., 2009; Fletcher and Yang, 2010). In a large-scale analysis of protein-coding nucleotide sequences from vertebrate species, it was observed that 1–9% of codon sites that were identified as homologous with reverse-translation apparently contained non-homologous codons, where the d_N/d_S ratio was unduly high, and the overall d_N/d_S ratio was 5–43% overestimated with reverse-translation (Y. Suzuki, unpublished). Therefore, reverse-translation of aligned amino acid sequences is not recommendable in the study of

natural selection by comparing d_S and d_N . Information from nucleotide sequences (Fletcher and Yang, 2010) as well as alignment errors (Wong et al., 2008) may be taken into account for obtaining reliable estimates of d_N/d_S ratio.

METHODS FOR DETECTING RECENT/ONGOING NATURAL SELECTION USING POLYMORPHISM DATA

In the previous section, statistical methods for detecting recurrent natural selection were introduced. These methods are based on the molecular evolutionary principles and designed to test whether the number of possibly adaptive substitutions is significantly greater or smaller than that of neutral substitutions within the same evolutionary time interval. In this approach, a number of possibly adaptive substitutions is required for detecting positive selection. In fact, at the amino acid sequence level, recurrent positive selection has been identified mainly for interacting molecules changing continuously under counteracting selective pressures to construct and destruct the interaction, e.g., pathogens and host immune systems, and eggs and sperms. However, most of evolutionary innovations, which are considered as candidates for adaptive evolution, appear to have occurred only once in the evolutionary history. In these cases, it is difficult to detect positive selection by counting the number of such events. It is required to detect positive selection that has operated on one or a few occasions, and for this purpose, another approach based on the population genetic principles is necessary.

In a population, the pattern of polymorphisms for a genomic region evolving under the selective neutrality is expected to reach equilibrium, which is determined mainly by the rate of neutral mutation and the effective population size (N_e) (Kimura, 1983). However, when an advantageous mutation arises in a population, positive selection may operate to rapidly raise its frequency until fixation. During this process, frequencies of neutral or nearly neutral mutations linked to the selected mutation may also increase passively, unless these mutations are uncoupled with the selected mutation by recombination. This evolutionary mechanism is called the genetic hitchhiking (Maynard Smith and Haigh, 1974). By the genetic hitchhiking, haplotypes containing a selected mutation rapidly increase their frequencies, causing a strong linkage disequilibrium around the selected site.

As the selected mutation reaches fixation, the genomic region around the selected site becomes highly homogeneous, causing an excess of rare alleles and high frequency derived alleles. This condition is called the selective sweep. The size of selective sweep depends on the strength of positive selection and the recombination rate around the selected site; the size is expected to be

larger when positive selection is stronger and the recombination rate is smaller, because the selected mutation spreads more rapidly in the population with a smaller chance for the occurrence of recombination. It has been reported that fixation of single advantageous mutation can influence the pattern of neutral or nearly neutral polymorphisms within a region of length $\sim 0.01 s/r$, where s (> 0) is the selection coefficient and r is the recombination rate per nucleotide site (Kaplan et al., 1989).

In addition, since a population usually comprise subpopulations in nature, adaptation of subpopulations to local environments due to emergences of new advantageous mutations (hard sweep) or changes in the fitness effect of standing variations upon exposure to new environments (soft sweep) may facilitate differentiation of genetic composition among subpopulations.

Statistical methods have been developed for detecting recent/ongoing positive selection using these properties; i.e., the site frequency spectrum, the population differentiation, and the linkage disequilibrium. The first two classes of methods are aimed at detecting recently completed positive selection, whereas the last class of methods is aimed at detecting ongoing incomplete positive selection. All of these methods are considered to be effective in detecting positive selection that has occurred $< \sim 4N_e$ generations ago (Tajima, 1983; Garrigan et al., 2010).

Site frequency spectrum When the selective sweep is completed or nearly completed, the site frequency spectrum around the selected site is expected to be skewed toward an excess of rare alleles and high frequency derived alleles with a deficiency of intermediate frequency alleles, compared with the neutral expectation (Fig. 2A). Methods have been developed to detect this skew. Most of these methods utilize the fact that the parameter $\theta = 4N_e\mu$, where μ denotes the neutral mutation rate, can be estimated by using various summary statistics representing different parts of the site frequency spectrum. Since estimates of θ obtained from different summary statistics should be the same under the selective neutrality, the null hypothesis of no difference is tested between estimates. If a departure from 0 is statistically significant, the selective neutrality is rejected. Another approach compares the neutral and hitchhiking models by computing the likelihood of obtaining the observed site frequency spectrum. These methods are primarily effective for detecting recently completed selective sweeps, but less so for detecting ongoing incomplete selective sweeps and soft sweeps (Przeworski et al., 2005; Voight et al., 2006). These methods are also known to be sensitive to perturbation of steady-state (equilibrium) allelic frequency distribution (Garrigan et al., 2010).

Tajima's D If nucleotide sequences of a neutrally evolving genomic region is sampled from a panmictic popula-

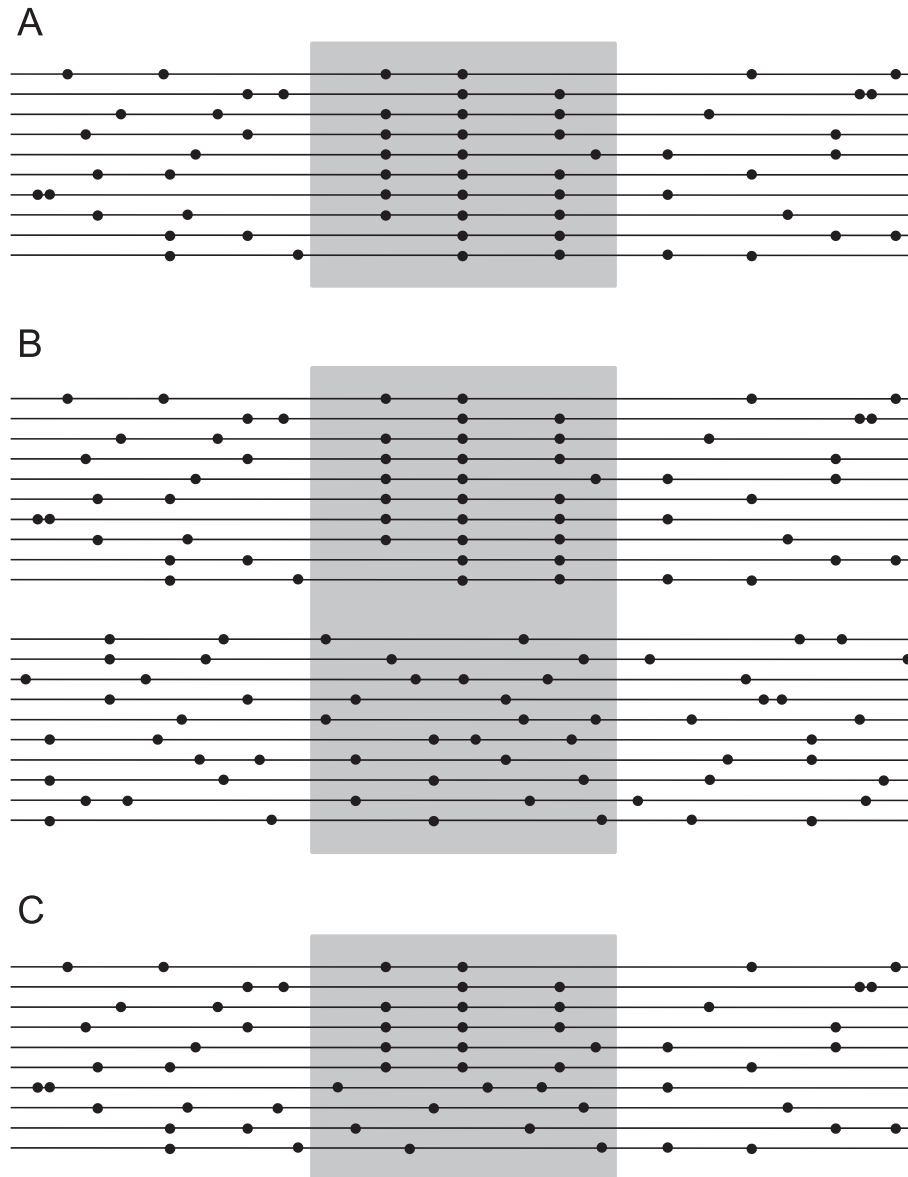


Fig. 2. Schematic diagrams for the polymorphism data analyzed by the statistical methods for detecting recent/ongoing natural selection with (A) site frequency spectrum, (B) population differentiation, and (C) linkage disequilibrium. Black lines and dots denote hypothetical nucleotide sequences and mutations, respectively. The regions where the statistical test is conducted are shaded in gray.

tion at equilibrium, the average number of nucleotide differences over all possible pairs of sequences (π) (nucleotide diversity, Nei and Li, 1979) and the proportion of segregating sites (s) are both expected to be expressed as a linear function of θ . (Here s is defined as the proportion of segregating sites, whereas the same symbol is defined as the selection coefficient in other parts of the paper.) Therefore, θ can be estimated using π and s , where the estimates may be denoted as $\hat{\theta}_{\pi}$ and $\hat{\theta}_s$, respectively. It should be noted that π and s mainly represent intermediate and low frequency alleles, respectively. Under the selective neutrality, $\hat{\theta}_{\pi}$ and $\hat{\theta}_s$ are expected to be the same. However, they may differ if the

site frequency spectrum is deviated from that expected under the selective neutrality. Tajima's D (Tajima, 1989) is defined as the difference $\hat{\theta}_{\pi} - \hat{\theta}_s$ divided by its standard deviation. A significantly positive value of Tajima's D indicates an excess of intermediate frequency alleles, which may be caused by balancing selection, population bottleneck, or population subdivision. A significantly negative value of Tajima's D indicates an excess of rare alleles, which may be caused by selective sweep, population expansion, or purifying selection.

Fay and Wu's H A significantly negative value of Tajima's D is an indicator of selective sweep, population expansion, or purifying selection, as indicated above. It

should be noted that for computing Tajima's D , it is not necessary to distinguish whether the alleles examined are ancestral or derived. However, if ancestral and derived alleles are distinguished, an excess of high-frequency derived alleles may be considered as evidence for selective sweep. Fay and Wu (2000) derived an estimator of θ mainly representing high-frequency derived alleles ($\hat{\theta}_H$), and defined Fay and Wu's H as the difference $\hat{\theta}_\pi - \hat{\theta}_H$. A significantly negative value of Fay and Wu's H indicates an excess of high-frequency derived alleles, which may be caused by selective sweep.

Fu and Li's D , D^* , F , and F^* Fu and Li (1993) derived another estimators of θ based on the genealogy of samples. They found that the expected numbers of mutations on the interior and exterior branches of the genealogical tree are both expressed as a linear function of θ under the selective neutrality. Therefore, θ can be estimated using the number of mutations on interior or exterior branches. Differences between these and other estimates were defined as Fu and Li's D , D^* , F , and F^* taking into account the availability of outgroup sequences for determining the root of the genealogical tree, and were used to test the selective neutrality.

Composite-likelihood ratio test In the above methods, the site frequency spectrum for a genomic region was represented by summary statistics for examining its deviation from the neutral expectation. However, by assuming a model for allele frequency probability, it is possible to directly compute the probability of obtaining the observed site frequency spectrum. Kim and Stephan (2002) developed a method for comparing the neutral and hitch-hiking models for a genomic region by using the composite-likelihood ratio test. For constructing the models, the effect of population demographic history may be explicitly implemented if it is estimated accurately (Li and Stephan, 2006). Alternatively, the demographic effect may be incorporated into the model by assuming the empirical site frequency spectrum for the rest of the genome as the allele frequency probability of the neutral model (Nielsen et al., 2005b).

Population differentiation When a population comprises subpopulations, adaptations of different subpopulations to various local environments may enhance genetic differentiation among them. In contrast, if balancing selection is taking place in subpopulations or directional selection operates for the entire population, genetic differentiation among subpopulations may be suppressed compared with the neutral expectation. The measure of population differentiation, such as the fixation index (F_{ST}), may be useful for testing whether the observed degree of genetic differentiation among subpopulations is significantly greater or smaller than the neutral expectation (Fig. 2B). Similarly to the case for the methods depending on site frequency spectrum, this

approach is effective for detecting recently completed or nearly completed selective sweeps.

F_{ST} F_{ST} is defined as the proportion of reduction in heterozygosity for subpopulations (H_S) compared with the total population (H_T); i.e., $F_{ST} = (H_T - H_S)/H_T$ (Wright, 1921), or as the proportion of reduction in nucleotide diversity for subpopulations (π_S) compared with the total population (π_T); i.e., $F_{ST} = (\pi_T - \pi_S)/\pi_T$ (Hudson et al., 1992). A significantly larger value of F_{ST} compared with the neutral expectation indicates that subpopulations are highly genetically differentiated, suggesting local selective sweeps in subpopulations. A significantly smaller value of F_{ST} indicates that subpopulations are highly genetically homogeneous, suggesting balancing selection or population-wide selective sweep. When quantitative characters are analyzed, $V_{ST} = (V_T - V_S)/V_T$ may be computed, where V_T and V_S denote the variance for the entire population and the average variance over all subpopulations weighted by population size, respectively (Redon et al., 2006).

It should be noted that the F_{ST} value may be inflated by the SNP ascertainment bias (Gardner et al., 2007). It has also been reported that the assumption on population structure, such as the finite island model or the hierarchical island model, affects the tests of F_{ST} values (Foll and Gaggiotti, 2008; Excoffier et al., 2009; Hermisson, 2009).

Linkage disequilibrium When positive selection is operating on an advantageous mutation and its frequency is rapidly increasing in a population together with linked neutral or nearly neutral mutations, genomic region around the selected site is characterized by existence of haplotypes with the selected mutation, which are unusually extended, homozygous, and frequent, and other haplotypes without the selected mutation, which more or less reflect the neutral background (Fig. 2C). Because of the existence of these haplotypes with different properties, the region exhibits a strong linkage disequilibrium, which decays more slowly than expected under the selective neutrality. As the selected mutation reaches fixation, the linkage disequilibrium diminishes because haplotypes without the selected mutation disappear and the region becomes highly homogeneous.

To detect ongoing incomplete selective sweeps, statistical methods have been developed for identifying unusually extended haplotype homozygosity and slow linkage disequilibrium decay. The former class of methods includes the LRH (long-range haplotype) test (Sabeti et al., 2002) and the iHS (integrated haplotype score) test (Voight et al., 2006), whereas the latter class includes the LDD (linkage disequilibrium decay) test (Wang et al., 2006). These tests are effective in detecting ongoing incomplete selective sweeps, but not in detecting recently completed selective sweeps.

The LRH and iHS tests can be modified for detecting recently completed selective sweeps. Haplotypes from two populations are compared to detect genomic regions with unusually extended haplotype homozygosity in either one population. These tests include the XP-EHH (cross population extended haplotype homozygosity) test (Sabeti et al., 2007), $\ln(Rsb)$ (standardized log ratio of integrated EHH for a SNP site) test (Tang et al., 2007), rMHH (ratio of most frequent haplotype homozygosity) test, and rHH (ratio of haplotype homozygosity) test (Kimura et al., 2007).

LRH test We first focus on a genomic locus (core locus) that is short enough to ensure the occurrence of no recombination (Sabeti et al., 2002). All haplotypes at this locus are determined in the sample and are called the core haplotypes. The purpose of this method is to find a core haplotype that is associated with a combination of higher frequency and greater EHH compared with other core haplotypes. Here the EHH for a core haplotype at a particular distance from the core locus is defined as the homozygosity of haplotypes extended from the core locus to that distance, carrying the core haplotype of interest at one end. For each core haplotype, the EHH values are recorded at various distances, and the relative EHH (rEHH) is computed at each distance as the ratio of the EHH value for the core haplotype of interest to the EHH value for other core haplotypes combined. A core haplotype with a combination of significantly higher frequency and greater rEHH values compared with other core haplotypes is considered to have spread in the population more rapidly than expected under the selective neutrality. The statistical significance for the combination of higher frequency and greater rEHH values is examined by conducting computer simulation under the assumption of various population demographic histories, and by constructing empirical distribution of these values using randomly selected core loci. This test is effective for detecting ongoing incomplete selective sweeps, where the frequency and rEHH values can be compared among segregating core haplotypes. It should be noted that in the comparison of core haplotypes at a core locus, one core haplotype serves as an internal control of local recombination rate for other core haplotypes.

iHS test The iHS test is a modification of the LRH test. We first focus on a SNP (core SNP) site, and determine all the alleles (core alleles) at this site. For each core allele, the decay of EHH relative to the genetic distance from the core SNP site is plotted, and the area under the decay curve is computed by integrating EHH until EHH drops below 0.05 in both directions (Voight et al., 2006). This integrated EHH (iHH) is denoted as iHH_A or iHH_D according to whether the core allele is ancestral or derived, respectively. The log-ratio of iHH_A to iHH_D is standardized using its genome-wide empirical distribution, and the standardized log-ratio value is called

the iHS. Therefore, the iHS indicates the degree of peculiarity for haplotype configuration extended from the core SNP site in the entire genome. Genomic regions containing a large number of core SNP sites with extremely positive or negative values of iHS are considered as positively selected. Similarly to the case for the LRH test, the iHS test is effective for detecting ongoing incomplete selective sweeps.

LDD test When an advantageous mutation is rapidly increasing its frequency in a population due to positive selection, the genomic region around the selected allele at a SNP site is expected to be in a state of strong linkage disequilibrium, which is absent around other alleles at the same site. Therefore, the existence of such distinct alleles at a SNP site is considered as evidence for ongoing incomplete positive selection. For detecting such a SNP site, the pattern of LDD was modeled for selected alleles, and this model was used for comparing the patterns of LDD between different alleles at each SNP site by computing the average log-likelihood ($ALnLH$) (Wang et al., 2006). In this method, errors usually introduced into the analysis of linkage disequilibrium through the inference of haplotypes from genotype data were eliminated by only using genotype data of homozygous individuals at the SNP site of interest.

XP-EHH, $\ln(Rsb)$, rMHH, and rHH tests Although the LRH, iHS, and LDD tests are effective for detecting ongoing incomplete selective sweeps, they lack a power to detect selective sweeps that have been completed or nearly completed. To detect recently completed selective sweeps using a similar approach, Tang et al. (2007) and Sabeti et al. (2007) modified the iHS test to develop the XP-EHH test and the $\ln(Rsb)$ test, respectively, where the total iHH value for a SNP site was compared between subpopulations. The LRH test was also modified to compare EHH between subpopulations by using rMHH and rHH (Kimura et al., 2007).

Problems in detecting recent/ongoing natural selection For detecting recent/ongoing natural selection, polymorphism data within populations are analyzed. However, since the polymorphism data are influenced by population demographic history, which is usually unknown but is considered to be highly complicated, it is generally difficult to obtain the null distribution of test statistics and conduct statistical tests. For this reason, it has been customary to adopt the outlier approach, where extreme values in the empirical distribution of test statistics are considered as signatures of positive or negative selection. The rationale for this approach is that the demographic effect is genome-wide, whereas the selection effect is locus-specific, and most of background genomic nucleotide sites are considered to evolve neutrally. However, extreme values in the empirical distribution may not necessarily support the rejection of

selective neutrality because the null and empirical distributions may be different. In addition, it is impossible to estimate the prevalence of positive and negative selection using this approach.

These facts raise a concern about the reliability of the results obtained from the methods adopting the outlier approach. An indirect way to assess the reliability is to examine overlaps between the results obtained from the analyses of similar data using different methods. The results may be considered relatively reliable or unreliable according to whether the proportion of overlap is large or small, respectively.

It should be noted that natural selection operating at the genomic region with lower recombination rate is more easily detectable than that with higher recombination rate in all the methods presented above, because the signature of natural selection is more extensively preserved in the former case than in the latter case. Therefore, functional categorization of positively selected genes may be influenced by the recombination rate of genomic regions where genes belonging to specific functional categories exist.

Outlier approach Teshima et al. (2006) examined the reliability of the outlier approach using computer simulation. Genomic regions were ranked according to Tajima's D , π , θ_s , Fay and Wu's H , Fu and Li's D , and haplotype homozygosity for detecting positive selection. It was observed that the false-discovery rate and the false-negative rate can be excessively large especially when population bottleneck occurred and positive selection operated on standing variations. Kelley et al. (2006) also conducted computer simulation based on Tajima's D and found that the false-discovery rate can be high when positive selection is weak, the fraction of truly positively selected loci is small, and the fraction of test statistic values considered as evidence for positive selection is large.

Little overlap among results from different studies It has been reported that the fraction of overlap between positively selected regions identified in different studies is generally small (Mallick et al., 2009). When 713 positively selected genes in human identified by the iHS test (Voight et al., 2006) were compared with 90 positively selected genes identified by the LDD test (Wang et al., 2006), only 9 genes were shared between the two results (Nielsen et al., 2007). It should be noted that both of these tests are aimed at detecting ongoing incomplete selective sweeps using the linkage disequilibrium. Akey (2009) compared positively selected regions in the human genome identified by 9 studies (Akey et al., 2002; Carlson et al., 2005; Kelley et al., 2006; Voight et al., 2006; Wang et al., 2006; Kimura et al., 2007; Sabeti et al., 2007; Tang et al., 2007; Williamson et al., 2007). Among the total of 5,110 distinct regions identified as positively selected in any one of these studies, only 722 regions (14.1%) were identified in two or more studies, 271

regions (5.3%) in three or more studies, and 129 regions (2.5%) in four or more studies. When 81 previously proposed candidates for positively selected genes were examined using the measures such as function-altering mutations (Bustamante et al., 2005; Nielsen et al., 2005a), heterozygosity, excess of rare alleles, and LRH (The International HapMap Consortium, 2005), only 25 genes were ranked in the top 1% of the empirical distribution by at least one test (Sabeti et al., 2006). It has been proposed to combine several test statistics such as iHS, XP-EHH, and F_{ST} using a composite likelihood statistic to narrow candidates for variants under population specific positive selection (Grossman et al., 2010).

METHODS FOR DETECTING RECURRENT OR RECENT/ONGOING NATURAL SELECTION USING DIVERGENCE AND POLYMORPHISM DATA

In the previous sections, the divergence data between populations and the polymorphism data within populations were analyzed separately for detecting recurrent and recent/ongoing natural selection, respectively. However, these data may be mutually helpful for detecting both types of natural selection. The MK (McDonald and Kreitman) test (McDonald and Kreitman, 1991) and the HKA (Hudson, Kreitman, and Aguade) test (Hudson et al., 1987) have been developed for facilitating detection of recurrent and recent/ongoing natural selection by contrasting the divergence and polymorphism data.

MK test Let us assume that we analyze a set of allelic sequences without occurrence of recombination from one population and reference sequences from another population. We also assume that we know the exact sites evolving neutrally in the sequences and we are interested in testing the selective neutrality at other (tested) sites. In the MK test (McDonald and Kreitman, 1991), the numbers of sites with fixed differences between populations (D) and those with polymorphic differences within populations (P) are counted for both the tested and neutral sites, and the D/P ratio at the tested sites ($D_{\text{Tested}}/P_{\text{Tested}}$) is contrasted with that at the neutral sites ($D_{\text{Neutral}}/P_{\text{Neutral}}$) (Fig. 3A). If the tested sites evolve neutrally, these ratios are expected to be the same, i.e., $D_{\text{Tested}}/P_{\text{Tested}} = D_{\text{Neutral}}/P_{\text{Neutral}}$. However, if positive selection has operated recurrently at the tested sites, advantageous mutations are repeatedly fixed in the population quickly without remaining as polymorphisms to any large extent (Kimura, 1983). Therefore, the ratio is expected to be inflated, such that $D_{\text{Tested}}/P_{\text{Tested}} > D_{\text{Neutral}}/P_{\text{Neutral}}$. It should be noted that the relationship $D_{\text{Tested}} > D_{\text{Neutral}}$ or $P_{\text{Tested}} > P_{\text{Neutral}}$ may not be necessarily observed even under positive selection. In contrast, if balancing selection operates to maintain polymorphisms, or slightly del-

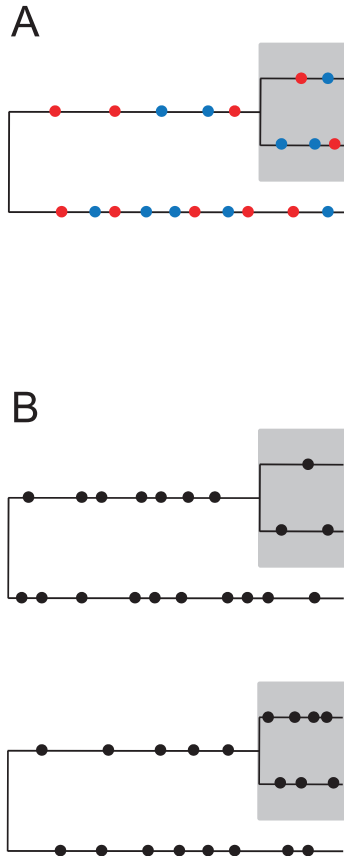


Fig. 3. Schematic diagrams for the divergence and polymorphism data analyzed by the statistical methods for detecting recurrent or recent/ongoing natural selection with (A) MK test and (B) HKA test. Black, blue, and red dots on hypothetical phylogenetic trees indicate hypothetical nucleotide, synonymous, and nonsynonymous mutations (for polymorphism) or substitutions (for divergence), respectively. Polymorphism data are shaded in gray, whereas divergence data are not.

eterious mutations exist in low frequency in the population under purifying selection at the tested sites, the ratio is expected to be deflated, such that $D_{\text{Tested}}/P_{\text{Tested}} < D_{\text{Neutral}}/P_{\text{Neutral}}$. In the MK test, the null hypothesis of $D_{\text{Tested}}/P_{\text{Tested}} = D_{\text{Neutral}}/P_{\text{Neutral}}$ is tested by the G -test, χ^2 test, or Fisher's exact test using a 2×2 contingency table for D_{Tested} , D_{Neutral} , P_{Tested} , and P_{Neutral} . It was suggested to use d_{Tested} , d_{Neutral} , π_{Tested} , and π_{Neutral} instead of D_{Tested} , D_{Neutral} , P_{Tested} , and P_{Neutral} (Graur et al., 1991).

If we denote the fraction of positively selected substitutions among all fixed substitutions at the tested sites as α , the number of neutral substitutions is described as $(1 - \alpha)D_{\text{Tested}}$. P_{Tested} is considered to mainly reflect neutral mutations, because advantageous and deleterious mutations are likely to be quickly fixed or eliminated from the population, respectively, and are unlikely to contribute to polymorphisms to any large extent. Therefore, the equation $(1 - \alpha)D_{\text{Tested}}/P_{\text{Tested}} = D_{\text{Neutral}}/P_{\text{Neutral}}$ is roughly expected to hold, which leads to $\alpha = 1 - (D_{\text{Neutral}}/P_{\text{Neutral}})/$

$(D_{\text{Tested}}/P_{\text{Tested}})$. The α value can be estimated from D_{Tested} , D_{Neutral} , P_{Tested} , and P_{Neutral} using this equation (Charlesworth, 1994).

It is sometimes convenient to divide polymorphic differences at the tested (P_{Tested}) and neutral (P_{Neutral}) sites into classes according to whether minor alleles are rare ($P_{\text{Tested,rare}}$ and $P_{\text{Neutral,rare}}$) or common ($P_{\text{Tested,common}}$ and $P_{\text{Neutral,common}}$). Since rare alleles are likely to contain slightly deleterious mutations, whereas common alleles are not, it is considered that $P_{\text{Tested,common}}/P_{\text{Neutral,common}}$ and $1 - P_{\text{Tested,common}}/P_{\text{Neutral,common}}$ indicate the proportions of neutral and deleterious mutations at the tested sites, respectively. $P_{\text{Tested,rare}}/P_{\text{Neutral,rare}} - P_{\text{Tested,common}}/P_{\text{Neutral,common}}$ is considered to reflect the proportion of slightly deleterious mutations (Fay et al., 2001).

In the above argument, the fitness effect of each mutation is assumed to be discrete, i.e., advantageous, neutral (and slightly deleterious), or deleterious. However, the fitness effect is likely to be continuous in reality. In this case, the Poisson random field model may be used for estimating the parameter values of fitness distribution from the 2×2 contingency table for D_{Tested} , D_{Neutral} , P_{Tested} , and P_{Neutral} (Sawyer and Hartl, 1992) or from the distribution of derived allele frequencies in a sample (Boyko et al., 2008).

Amino acid sequence level In the analysis of protein-coding nucleotide sequences, positive selection operating at the amino acid sequence level may be detected by contrasting the D/P ratio at nonsynonymous sites (D_N/P_N) with that at synonymous sites (D_S/P_S) under the assumption that the latter sites evolve neutrally. It has been reported that the proportion of positively selected amino acid substitutions was 54%, 56%, and 57% in fruit fly (Begun et al., 2007), enteric bacteria (Charlesworth and Eyre-Walker, 2006), and mouse (Halligan et al., 2010), respectively. In an extreme case, 94.0%–94.3% of amino acid substitutions were inferred as positively selected in fruit fly (Sawyer et al., 2003, 2007), although the average scaled selection coefficient ($N_e s$) for the substitutions was estimated to be only 2.5–5.1. (Here s denotes the selection coefficient.) It should be noted that a mutation is considered to be neutral or nearly neutral if its $N_e s$ value is ≤ 2 (Ohta, 2002). In contrast, no amino acid substitution was inferred as positively selected in *Arabidopsis* (Foxe et al., 2008; Gossmann et al., 2010) and yeast (Doniger et al., 2008; Liti et al., 2009). In human, the proportion of positively selected amino acid substitutions was estimated to be 0–11.9% (The Chimpanzee Sequencing and Analysis Consortium, 2005; Gojobori et al., 2007). The existence of $\sim 10\%$ of positively selected amino acid substitutions appears to be consistent with the neutral theory of molecular evolution (Ohta and Kimura, 1971).

Nucleotide sequence level The MK test can also be applied to detection of natural selection operating at the nucleotide sequence level by contrasting the D/P ratio at

nucleotide sites of interest with that at synonymous sites, pseudogenes, or introns, which are assumed to evolve neutrally. The proportion of positively selected nucleotide substitutions in non-coding nucleotide sequences was estimated to be 7–61% in fruit fly (Andolfatto, 2005; Begun et al., 2007), although the $N_e s$ value for each nucleotide substitution was estimated to be 0.143–0.900. In contrast, no signature of adaptive evolution was observed in yeast (Doniger et al., 2008) and human (Keightley et al., 2005).

These results apparently indicate that the role of positive selection may be different among lineages. In particular, according to the observations that positive selection was pervasive for the protein-coding and non-coding nucleotide sequences in fruit fly, the selection theory of molecular evolution has been proposed (Hahn, 2008; Sella et al., 2009). It is known that the effect of natural selection is positively correlated with N_e . However, the above difference in α , with the relationship {fruit fly; enteric bacteria; mouse} > {human; Arabidopsis; yeast}, does not appear to be consistent with the difference in N_e , because N_e 's of fruit fly, enteric bacteria, and mouse are $\sim 10^6$, $\sim 10^9$, and $\sim 4 \times 10^5$, respectively, and those of human, Arabidopsis, and yeast are $\sim 10^4$, $\sim 10^5$, and $\sim 10^7$, respectively (enteric bacteria > yeast > fruit fly > mouse > Arabidopsis > human). The codon usage bias, which can inflate α because of the existence of slightly deleterious alleles in P_s , is also unlikely to be responsible for the variation of α , because the codon usage bias has been reported to be strong in enteric bacteria and yeast, intermediate in fruit fly and Arabidopsis, and weak in mouse and human ({enteric bacteria; yeast} > {fruit fly; Arabidopsis} > {mouse; human}).

HKA test Under the selective neutrality, both the expected amounts of polymorphism within populations and divergence between populations are expressed as a linear function of mutation rate. Therefore, if we analyze allelic sequences without occurrence of recombination from one population and reference sequences from another population, the D/P ratio is expected to be the same across genomic loci even if the mutation rate varies, as long as all analyzed loci share the same genealogy. In the HKA test (Hudson et al., 1987), the D/P ratio is compared among genomic loci by using the χ^2 test (Fig. 3B).

If the null hypothesis of equal D/P ratios for all loci is rejected, it is inferred that polymorphism of some loci is enhanced by balancing selection or reduced by directional selection. However, it is not always easy to identify the exact loci where positive selection operated. Positively selected loci may be identified by examining marginal χ^2 deviations or by conducting pairwise HKA tests as well as the likelihood-ratio test of different mutation rates for polymorphism and divergence at a specific locus (Wright and Charlesworth, 2004).

Problems in detecting recurrent or recent/ongoing natural selection Because of the occurrence of recombination, different genomic loci may have followed different genealogical histories, exhibiting various ratios of the time elapsed for divergence to polymorphism, which may result in different D/P ratios among genomic loci even under the selective neutrality. In addition, population bottlenecks and expansions may alter the effect of positive and negative selection genome-wide during evolution, changing the D/P ratio in a site-specific manner according to natural selection operating at each site. The strength of natural selection may also change in a site-specific manner during evolution, because of changes in environmental conditions.

Genealogy When multiple polymorphic sites are included in the sequences analyzed by the HKA test, it is required that all of these sites share a single genealogy or are unlinked. Otherwise, the D/P ratio may not be expected as constant among genomic loci under the selective neutrality, and positive selection may be falsely identified in the HKA test. This effect may also affect the MK test; a combination of differences in genealogies and functional constraints among sites may cause the relationship $D_{\text{Tested}}/P_{\text{Tested}} \neq D_{\text{Neutral}}/P_{\text{Neutral}}$ even when tested sites evolve neutrally. In particular, positive selection may be falsely identified if the ratios of $(P_{\text{Tested}} + P_{\text{Neutral}})/(D_{\text{Tested}} + D_{\text{Neutral}})$ (size of genealogy) and $(D_{\text{Tested}} + P_{\text{Tested}})/(D_{\text{Neutral}} + P_{\text{Neutral}})$ (selective constraint) are negatively correlated among protein-coding nucleotide sequences (Shapiro et al., 2007; Nei et al., 2010). In an analysis of fruit fly, the proportion of positively selected amino acid substitutions reduced from 47.4% to 28.9% when the genealogical effect was corrected (Shapiro et al., 2007).

Population bottleneck Population bottleneck and expansion weaken and strengthen the effect of natural selection, respectively. The MK test may produce spurious signatures of positive selection when a population passed through bottlenecks during evolution (McDonald and Kreitman, 1991; Ohta, 1993; Hughes, 2008a). Slightly deleterious mutations may be fixed by random genetic drift at tested sites during bottlenecks because these mutations may behave as if they are neutral under weakened purifying selection. Consequently, D_{Tested} may be inflated such that the relationship $D_{\text{Tested}}/P_{\text{Tested}} > D_{\text{Neutral}}/P_{\text{Neutral}}$ may be observed without positive selection. The effect of bottlenecks is expected to be genome-wide. The HKA test may also produce significant results without natural selection when a population is subdivided (Ingvarsson, 2004).

Change in functional constraint In both the MK and HKA tests, the D/P ratio is affected if the direction and magnitude of natural selection change during evolution. The relationship $D_{\text{Tested}}/P_{\text{Tested}} > D_{\text{Neutral}}/P_{\text{Neutral}}$ may be observed in the MK test without positive selection if functional constraint is recently enhanced at tested

sites and P_{Tested} is suppressed (Hughes et al., 2008). The equality of the D/P ratio among genomic loci may also be rejected without positive selection in the HKA test if D or P is specifically enhanced or suppressed for a particular locus due to changes in functional constraints. In addition, even if the selection coefficient fluctuates randomly around 0 at tested sites, P_{Tested} is expected to be reduced such that directional selection may be falsely identified in the MK and HKA tests (Nei and Yokoyama, 1976; Nei et al., 2010). It should be noted that the effect of change in natural selection is expected to be locus specific, which is distinct from the effect of bottleneck.

PERSPECTIVES

A variety of statistical methods has been developed for detecting recurrent and/or recent/ongoing natural selection using the divergence and/or polymorphism data. These methods are usually designed to be highly sensitive to the signatures of positive selection. A large number of papers has been published reporting the inference of positive selection using these methods. The occurrence of positive selection is usually concluded only from the results of the statistical methods without experimental verification, partly due to difficulty in proving selective advantages in nature under experimental settings. In addition, statistical methods have been used even when they appeared to be inapplicable, e.g., when d_S and d_N were considered to be saturated (Studer et al., 2008). Biological significance of the results was speculated even when they were obtained without conducting statistical tests (Uddin et al., 2008; Goodman et al., 2009; Goodman and Sterner, 2010) or the proportion of positive results was smaller than the significance level adopted. Positively selected genes identified were sometimes randomly associated with functions and gene duplications (Studer et al., 2008), and little overlap was observed among positively selected regions obtained from the analyses of similar data using different methods. Eventually, pervasive positive selection was identified for the protein-coding and non-coding nucleotide sequences in the genomic analysis of some organisms, and the selection theory of molecular evolution was proposed.

To conclude the occurrence of positive selection only from the results of statistical methods, it is required that the statistical tests are unbiased as well as robust against various disturbing factors such as sampling errors and violations of assumptions. However, many of the methods have been criticized by using computer simulation and real data analysis to produce excessive false-positives and to be sensitive to sampling errors and violations of assumptions as well as sequencing, annotation, and alignment errors. Importantly, some of these methods have been invalidated experimentally, e.g., positively selected amino acid sites identified by the Bayesian and likelihood

methods for single-site analysis and by the branch-site method mostly disagreed with experimentally determined adaptive sites. These observations indicate that many of the statistical methods for detecting positive selection are unreliable and a significant fraction of the results published in a large number of papers is false-positives. Therefore, the relative role of random genetic drift and positive selection on molecular and phenotypic evolution is still controversial.

In addition, there may be an alternative interpretation for the signals considered as evidence for positive selection in the methods for detecting recent/ongoing positive selection and recurrent or recent/ongoing positive selection; these signals may indicate evidence for fixations of deleterious mutations by random genetic drift rather than those of advantageous mutations by positive selection. It should be noted that the expected fixation time for deleterious mutations that are destined to be fixed in a population is shorter than that for neutral mutations, although the fixation probability for the former mutations may be lower than that for the latter mutations. In particular, it has been reported that the expected fixation times for advantageous and deleterious mutations are the same (symmetric) if absolute values of selection coefficients as well as dominance coefficients are the same (Maruyama and Kimura, 1974). Therefore, a strong linkage disequilibrium may be observed around deleterious mutations when they are going to be fixed, and the relationship $D_{\text{Tested}}/D_{\text{Neutral}} > P_{\text{Tested}}/P_{\text{Neutral}}$ may be obtained when many deleterious mutations are fixed at the tested sites. It would be important to distinguish whether these signals are derived from fixations of advantageous mutations due to positive selection or those of deleterious mutations due to random genetic drift.

In accordance with the advancement in the sequencing technology, genomic diversity data are rapidly accumulating in various organisms, e.g., The 1000 Genomes Project in human, The 1001 Genomes Project in Arabidopsis, and The Influenza Genome Sequencing Project. These data should be useful for clarifying the relative importance of random genetic drift and positive selection on molecular and phenotypic evolution, and for identifying disease-associated genes and mutations in humans as well as in pathogens. For these purposes, detection of positive and negative selection operating at the amino acid and nucleotide sequence levels may play a critical role. It is therefore important to develop reliable statistical methods that are unbiased as well as robust against various disturbing factors, for detecting natural selection.

The author thanks Masatoshi Nei, Fumio Tajima, Masafumi Nozawa, Yuki Kobayashi, and two anonymous reviewers for valuable suggestions and comments. This work was supported by KAKENHI 20580007.

REFERENCES

- Akey, J. M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* **19**, 711–722.
- Akey, J. M., Zhang, G., Zhang, K., Jin, L., and Shriver, M. D. (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**, 1149–1152.
- Bakewell, M. A., Shi, P., and Zhang, J. (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc. Natl. Acad. Sci. USA* **104**, 7489–7494.
- Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., et al. (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* **5**, e310.
- Berglund, A.-C., Wallner, B., Elofsson, A., and Liberles, D. A. (2005) Tertiary windowing to detect positive diversifying selection. *J. Mol. Evol.* **60**, 499–504.
- Berglund, J., Pollard, K. S., and Webster, M. T. (2009) Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol.* **7**, e1000026.
- Berlin, S., and Smith, N. G. C. (2005) Testing for adaptive evolution of the female reproductive protein ZPC in mammals, birds and fishes reveals problems with the M7–M8 likelihood ratio test. *BMC Evol. Biol.* **5**, 65.
- Bird, C. P., Stranger, B. E., Liu, M., Thomas, D. J., Ingle, C. E., Beazley, C., Miller, W., Hurles, M. E., and Dermitzakis, E. T. (2007) Fast-evolving noncoding sequences in the human genome. *Genome Biol.* **8**, R118.
- Birky, C. W. Jr., and Walsh, J. B. (1988) Effects of linkage on rates of molecular evolution. *Proc. Natl. Acad. Sci. USA* **85**, 6414–6418.
- Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* **4**, e1000083.
- Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Gnanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., et al. (2005) Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157.
- Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., and Nickerson, D. A. (2005) Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* **15**, 1553–1565.
- Cartegni, L., Chew, S. L., and Krainer, A. R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.* **3**, 285–298.
- Chamary, J. V., Parmley, J. L., and Hurst, L. D. (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7**, 98–108.
- Charlesworth, B. (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.* **63**, 213–227.
- Charlesworth, J., and Eyre-Walker, A. (2006) The rate of adaptive evolution in enteric bacteria. *Mol. Biol. Evol.* **23**, 1348–1356.
- Chun, S., and Fay, J. C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561.
- Clark, A. G., and Kao, T.-H. (1991) Excess nonsynonymous substitution at shared polymorphic sites among self-incompatibility alleles of Solanaceae. *Proc. Natl. Acad. Sci. USA* **88**, 9823–9827.
- Clark, A. G., Gnanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., et al. (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960–1963.
- Comeron, J. M. (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.* **41**, 1152–1159.
- Doniger, S. W., Kim, H. S., Swain, D., Corcuera, D., Williams, M., Yang, S.-P., and Fay, J. C. (2008) A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet.* **4**, e1000183.
- Drummond, D. A., and Wilke, C. O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352.
- Duret, L., and Galtier, N. (2009) Comment on “Human-specific gain of function in a developmental enhancer”. *Science* **323**, 714c.
- Excoffier, L., Hofer, T., and Foll, M. (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* **103**, 285–298.
- Fay, J. C., and Wu, C.-I. (2000) Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413.
- Fay, J. C., Wyckoff, G. J., and Wu, C.-I. (2001) Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234.
- Filipski, A., Prohaska, S., and Kumar, S. (2007) Molecular signatures of adaptive evolution. In: *Evolutionary Genomics and Proteomics* (eds.: M. Pagel, and A. Pomiankowski), pp. 241–254. Sinauer Associates, Inc., Massachusetts.
- Fitch, W. M., Bush, R. M., Bender, C. A., and Cox, N. J. (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. USA* **94**, 7712–7718.
- Fletcher, W., and Yang, Z. (2010) The effect of insertions, deletions and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* **27**, 2257–2267.
- Foll, M., and Gaggiotti, O. (2008) A genome-scan method to identify selection loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977–993.
- Foxe, J. P., Dar, V.-u.-N., Zheng, H., Nordborg, M., Gaut, B. S., and Wright, S. I. (2008) Selection on amino acid substitutions in *Arabidopsis*. *Mol. Biol. Evol.* **25**, 1375–1383.
- Friedman, R., and Hughes, A. L. (2007) Likelihood-ratio tests for positive selection of human and mouse duplicate genes reveal nonconservative and anomalous properties of widely used methods. *Mol. Phylogenet. Evol.* **42**, 388–393.
- Fu, Y.-X., and Li, W.-H. (1993) Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.
- Galtier, N., Duret, L., Glemin, S., and Ranwez, V. (2009) GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.* **25**, 1–5.
- Gardner, M., Williamson, S., Casals, F., Bosch, E., Navarro, A., Calafell, F., Bertranpetit, J., and Comas, D. (2007) Extreme individual marker F_{ST} values do not imply population-specific selection in humans: the *NRG1* example. *Hum. Genet.* **121**, 759–762.
- Garrigan, D., Lewontin, R., and Wakeley, J. (2010) Measuring the sensitivity of single-locus “neutrality tests” using a direct perturbation approach. *Mol. Biol. Evol.* **27**, 73–89.

- Gojobori, J., Tang, H., Akey, J. M., and Wu, C.-I. (2007) Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. *Proc. Natl. Acad. Sci. USA* **104**, 3907–3912.
- Goldman, N., and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736.
- Goodman, M., and Sterner, K. N. (2010) Phylogenomic evidence of adaptive evolution in the ancestry of humans. *Proc. Natl. Acad. Sci. USA* **107**, 8918–8923.
- Goodman, M., Sterner, K. N., Islam, M., Uddin, M., Sherwood, C. C., Hof, P. R., Hou, Z.-C., Lipovich, L., Jia, H., Grossman, L. I., and Wildman, D. E. (2009) Phylogenomic analyses reveal convergent patterns of adaptive evolution in elephant and human ancestries. *Proc. Natl. Acad. Sci. USA* **106**, 20824–20829.
- Gossmann, T. I., Song, B.-H., Windsor, A. J., Mitchell-Olds, T., Dixon, C. J., Kapralov, M. V., Filatov, D. A., and Eyre-Walker, A. (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol. Biol. Evol.* **27**, 1822–1832.
- Graur, D., Li, W.-H., Whittam, T. S., Nei, M., McDonald, J. H., and Kreitman, M. (1991) Neutral mutation hypothesis test. *Nature* **354**, 114–116.
- Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., et al. (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883–886.
- Hahn, M. W. (2008) Toward a selection theory of molecular evolution. *Evolution* **62**, 255–265.
- Halligan, D. L., Oliver, F., Eyre-Walker, A., Harr, B., and Keightley, P. D. (2010) Evidence for pervasive adaptive evolution in wild mice. *PLoS Genet.* **6**, e1000825.
- Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K. D., and Wray, G. A. (2007) Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat. Genet.* **39**, 1140–1144.
- Hellmann, I., Zollner, S., Enard, W., Ebersberger, I., Nickel, B., and Paabo, S. (2003) Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13**, 831–837.
- Hermisson, J. (2009) Who believes in whole-genome scans for selection? *Heredity* **103**, 283–284.
- Hill, W. G., and Robertson, A. (1966) The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294.
- Hudson, R. R., Kreitman, M., and Aguade, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.
- Hudson, R. R., Slatkin, M., and Maddison, W. P. (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589.
- Hughes, A. L. (2008a) Near neutrality: leading edge of the neutral theory of molecular evolution. *Ann. NY Acad. Sci.* **1133**, 162–179.
- Hughes, A. L. (2008b) The origin of adaptive phenotypes. *Proc. Natl. Acad. Sci. USA* **105**, 13193–13194.
- Hughes, A. L., and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170.
- Hughes, A. L., Friedman, R., Rivallier, P., and French, J. O. (2008) Synonymous and nonsynonymous polymorphisms versus divergences in bacterial genomes. *Mol. Biol. Evol.* **25**, 2199–2209.
- Hurst, L. D. (2009) A positive becomes a negative. *Nature* **457**, 543–544.
- Hurst, L. D., and Pal, C. (2001) Evidence for purifying selection acting on silent sites in *BRCA1*. *Trends Genet.* **17**, 62–65.
- Ina, Y. (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* **40**, 190–226.
- Ingvarsson, P. K. (2004) Population subdivision and the Hudson-Kreitman-Aguade test: testing for deviations from the neutral model in organelle genomes. *Genet. Res.* **83**, 31–39.
- Kaplan, N. L., Hudson, R. R., and Langley, C. H. (1989) The “hitchhiking effect” revisited. *Genetics* **123**, 887–899.
- Keightley, P. D., and Gaffney, D. J. (2003) Functional constraints and frequency of deleterious mutations in non-coding DNA of rodents. *Proc. Natl. Acad. Sci. USA* **100**, 13402–13406.
- Keightley, P. D., Lercher, M. J., and Eyre-Walker, A. (2005) Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3**, e42.
- Kelley, J. L., Madeoy, J., Calhoun, J. C., Swanson, W., and Akey, J. M. (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* **16**, 980–989.
- Kim, Y., and Stephan, W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765–777.
- Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., and Gottesman, M. M. (2007) A “silent” polymorphism in the *MDR1* gene changes substrate specificity. *Science* **315**, 525–528.
- 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, New York, Melbourne.
- Kimura, R., Fujimoto, A., Tokunaga, K., and Ohashi, J. (2007) A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS ONE* **2**, e286.
- Kondo, R., Horai, S., Satta, Y., and Takahata, N. (1993) Evolution of hominoid mitochondrial DNA with special reference to the silent substitution rate over the genome. *J. Mol. Evol.* **36**, 517–531.
- Kosakovsky Pond, S. L., and Frost, S. D. (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22**, 1208–1222.
- Li, H., and Stephan, W. (2006) Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet.* **2**, e166.
- Li, W.-H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**, 96–99.
- Li, W.-H., Wu, C.-I., and Luo, C.-C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**, 150–174.
- Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., et al. (2009) Population genomics of domestic and wild yeasts. *Nature* **458**, 337–341.
- Liu, K., Raghavan, S., Nelesen, S., Linder, C. R., and Warnow, T. (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* **324**, 1561–1564.
- Mallick, S., Gnerre, S., Muller, P., and Reich, D. (2009) The dif-

- difficulty of avoiding false positives in genome scans for natural selection. *Genome Res.* **19**, 922–933.
- Maruyama, T., and Kimura, M. (1974) A note on the speed of gene frequency changes in reverse direction in a finite population. *Evolution* **28**, 161–163.
- Massingham, T., and Goldman, N. (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**, 1753–1762.
- Maynard Smith, J., and Haigh, J. (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35.
- McDonald, J. H., and Kreitman, M. (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654.
- Medawar, P. B. (1946) Old age and natural death. *Mod. Quart.* **2**, 30–49.
- Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190.
- Miyata, T., and Yasunaga, T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino-acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16**, 23–36.
- Muse, S. V. (1996) Estimating synonymous and nonsynonymous substitution rates. *Mol. Biol. Evol.* **13**, 105–114.
- Muse, S. V., and Gaut, B. S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**, 715–724.
- Neel, J. V. (1962) Diabetes mellitus: a “thrifty” genotype rendered detrimental by “progress”? *Am. J. Hum. Genet.* **14**, 353–362.
- Nei, M., and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426.
- Nei, M., and Kumar, S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford, New York.
- Nei, M., and Li, W.-H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76**, 5269–5273.
- Nei, M., and Yokoyama, S. (1976) Effects of random fluctuation of selection intensity on genetic variability in a finite population. *Jpn. J. Genet.* **51**, 355–369.
- Nei, M., Suzuki, Y., and Nozawa, M. (2010) The neutral theory of molecular evolution in the genomic era. *Annu. Rev. Genomics Hum. Genet.* **11**, 265–289.
- Ng, P., and Henikoff, S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874.
- Nielsen, R., and Yang, Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.
- Nielsen, R., Bustamante, C., Clark, A. G., Gnanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., et al. (2005a) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170.
- Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005b) Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**, 1566–1575.
- Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A. G. (2007) Recent and ongoing selection in the human genome. *Nat. Rev. Genet.* **8**, 857–868.
- Nozawa, M., Suzuki, Y., and Nei, M. (2009a) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc. Natl. Acad. Sci. USA* **106**, 6700–6705.
- Nozawa, M., Suzuki, Y., and Nei, M. (2009b) Response to Yang et al.: problems with Bayesian methods of detecting positive selection at the DNA sequence level. *Proc. Natl. Acad. Sci. USA* **106**, E96.
- Ohta, T. (1993) Amino acid substitution at the *Adh* locus of *Drosophila* is facilitated by small population size. *Proc. Natl. Acad. Sci. USA* **90**, 4548–4551.
- Ohta, T. (2002) Near-neutrality in evolution of genes and gene regulation. *Proc. Natl. Acad. Sci. USA* **99**, 16134–16137.
- Ohta, T., and Kimura, M. (1971) On the constancy of the evolutionary rate of cistrons. *J. Mol. Evol.* **1**, 18–25.
- Pamilo, P., and Bianchi, O. (1993) Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Mol. Biol. Evol.* **10**, 271–281.
- Parsch, J., Novozhilov, S., Saminadin-Peter, S. S., Wong, K. M., and Andolfatto, P. (2010) On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Mol. Biol. Evol.* **27**, 1226–1234.
- Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., and Dodgson, J. (1980) The evolution of genes: the chicken preproinsulin gene. *Cell* **20**, 555–566.
- Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J. S., Bejerano, G., Baertsch, R., et al. (2006) Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* **2**, e168.
- Prabhakar, S., Noonan, J. P., Paabo, S., and Rubin, E. M. (2006) Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**, 786.
- Prabhakar, S., Visel, A., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., Plajzer-Frick, I., Morrison, H., FitzPatrick, D. R., Afzal, V., et al. (2008) Human-specific gain of function in a developmental enhancer. *Science* **321**, 1346–1350.
- Przeworski, M., Coop, G., and Wall, J. D. (2005) The signature of positive selection on standing genetic variation. *Evolution* **59**, 2312–2323.
- Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., et al. (2006) Global variation in copy number in the human genome. *Nature* **444**, 444–454.
- Resch, A. M., Carmel, L., Marino-Ramirez, L., Ogurtsov, A. Y., Shabalina, S. A., Rogozin, I. B., and Koonin, E. V. (2007) Widespread positive selection in synonymous sites of mammalian genes. *Mol. Biol. Evol.* **24**, 1821–1831.
- Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837.
- Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Vailly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., and Lander, E. S. (2006) Positive natural selection in the human lineage. *Science* **312**, 1614–1620.
- Sabeti, P. C., Vailly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., et al. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918.
- Sawyer, S. A., and Hartl, D. L. (1992) Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176.
- Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D., and Hartl, D. L. (2003) Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive

- selection. *J. Mol. Evol.* **57**, S154–S164.
- Sawyer, S. A., Parsch, J., Zhang, Z., and Hartl, D. L. (2007) Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **104**, 6504–6510.
- Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. H., and Graur, D. (2009) Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol. Evol.* **1**, 114–118.
- Sella, G., Petrov, D. A., Przeworski, M., and Andolfatto, P. (2009) Pervasive natural selection in the *Drosophila* genome? *PLoS Genet.* **5**, e1000495.
- Shapiro, J. A., Huang, W., Zhang, C., Hubisz, M. J., Lu, J., Turissini, D. A., Fang, S., Wang, H.-Y., Hudson, R. R., Nielsen, R., et al. (2007) Adaptive genic evolution in the *Drosophila* genomes. *Proc. Natl. Acad. Sci. USA* **104**, 2271–2276.
- Studer, R. A., Penel, S., Duret, L., and Robinson-Rechavi, M. (2008) Pervasive positive selection on duplicated and non-duplicated vertebrate protein coding genes. *Genome Res.* **18**, 1393–1402.
- Subramanian, S., and Kumar, S. (2003) Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res.* **13**, 838–844.
- Subramanian, S., and Kumar, S. (2006) Higher intensity of purifying selection on >90% of the human genes revealed by the intrinsic replacement mutation rates. *Mol. Biol. Evol.* **23**, 2283–2287.
- Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., Kondrashov, A., and Bork, P. (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591–597.
- Suzuki, Y. (2004a) Negative selection on neutralization epitopes of poliovirus surface proteins: implications for prediction of candidate epitopes for immunization. *Gene* **328**, 127–133.
- Suzuki, Y. (2004b) New methods for detecting positive selection at single amino acid sites. *J. Mol. Evol.* **59**, 11–19.
- Suzuki, Y. (2004c) Three-dimensional window analysis for detecting positive selection at structural regions of proteins. *Mol. Biol. Evol.* **21**, 2352–2359.
- Suzuki, Y. (2006a) Ancient positive selection on CD155 as a possible cause for susceptibility to poliovirus infection in simians. *Gene* **373**, 16–22.
- Suzuki, Y. (2006b) Natural selection on the influenza virus genome. *Mol. Biol. Evol.* **23**, 1902–1911.
- Suzuki, Y. (2006c) Statistical properties of the methods for detecting positively selected amino acid sites. *Gene* **365**, 125–129.
- Suzuki, Y. (2008a) False-positive results obtained from the branch-site test of positive selection. *Genes. Genet. Syst.* **83**, 331–338.
- Suzuki, Y. (2008b) Phylogenetic window analysis for detecting chronological changes in natural selection. *Open Evol. J.* **2**, 13–30.
- Suzuki, Y., and Gojobori, T. (1999) A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **16**, 1315–1328.
- Suzuki, Y., and Gojobori, T. (2001) Positively selected amino acid sites in the entire coding region of hepatitis C virus subtype 1b. *Gene* **276**, 83–87.
- Suzuki, Y., and Gojobori, T. (2003) Analysis of coding sequences. In: *The phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny* (eds.: M. Salemi, and A.-M. Vandamme), pp. 283–311. Cambridge University Press, Cambridge.
- Suzuki, Y., and Nei, M. (2001) Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **18**, 2179–2185.
- Suzuki, Y., and Nei, M. (2002) Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **19**, 1865–1869.
- Suzuki, Y., and Nei, M. (2004) False positive selection identified by ML-based methods: examples from the *Sig1* gene of diatom *Thalassiosira weissflogii* and the *tax* gene of a human T-cell lymphotropic virus. *Mol. Biol. Evol.* **21**, 914–921.
- Suzuki, Y., Gojobori, T., and Kumar, S. (2009) Methods for incorporating the hypermutability of CpG dinucleotides in detecting natural selection operating at the amino acid sequence level. *Mol. Biol. Evol.* **26**, 2275–2284.
- Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.
- Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.
- Tamura, K., Subramanian, S., and Kumar, S. (2004) Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21**, 36–44.
- Tang, K., Thornton, K. R., and Stoneking, M. (2007) A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* **5**, e171.
- Taylor, M. S., Massingham, T., Hayashizaki, Y., Carninci, P., Goldman, N., and Semple, C. A. (2008) Rapidly evolving human promoter regions. *Nat. Genet.* **40**, 1262–1263.
- Teshima, K. M., Coop, G., and Przeworski, M. (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Res.* **16**, 702–712.
- The Chimpanzee Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87.
- The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature* **437**, 1299–1320.
- Uddin, M., Goodman, M., Erez, O., Romero, R., Liu, G., Islam, M., Opazo, J. C., Sherwood, C. C., Grossman, L. I., and Wildman, D. E. (2008) Distinct genomic signatures of adaptation in pre- and postnatal environments during human evolution. *Proc. Natl. Acad. Sci. USA* **105**, 3215–3220.
- Vamathevan, J. V., Hasan, S., Emes, R. D., Amrine-Madsen, H., Rajagopalan, D., Topp, S. D., Kumar, V., Word, M., Simmons, M. D., Foord, S. M., et al. (2008) The role of positive selection in determining the molecular cause of species differences in disease. *BMC Evol. Biol.* **8**, 273.
- Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006) A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72.
- Wang, E. T., Kodama, G., Baldi, P., and Moyzis, R. K. (2006) Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc. Natl. Acad. Sci. USA* **103**, 135–140.
- Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W. Jr., Swanstrom, R., Burch, C. L., and Weeks, K. M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711–716.
- Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., and Nielsen, R. (2007) Localizing recent adaptive evolution in the human genome. *PLoS Genet.* **3**, e90.
- Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008) Alignment uncertainty and genomic analysis. *Science* **319**, 473–476.
- Wong, W. S. W., and Nielsen, R. (2004) Detecting selection in

- noncoding regions of nucleotide sequences. *Genetics* **167**, 949–958.
- Wong, W. S. W., Yang, Z., Goldman, N., and Nielsen, R. (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168**, 1041–1051.
- Wright, S. (1921) Systems of mating. I. The biometric relations between parent and offspring. *Genetics* **6**, 111–123.
- Wright, S. I., and Charlesworth, B. (2004) The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics* **168**, 1071–1076.
- Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573.
- Yang, Z., and Nielsen, R. (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* **46**, 409–418.
- Yang, Z., and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43.
- Yang, Z., and Nielsen, R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19**, 908–917.
- Yang, Z., and Nielsen, R. (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.* **25**, 568–579.
- Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M. K. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449.
- Yokoyama, S., Tada, T., Zhang, H., and Britt, L. (2008) Elucidation of phenotypic adaptations: molecular analyses of dim-light vision proteins in vertebrates. *Proc. Natl. Acad. Sci. USA* **105**, 13480–13485.
- Zhang, J. (2004) Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol. Biol. Evol.* **21**, 1332–1339.
- Zhang, J., Kumar, S., and Nei, M. (1997) Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Mol. Biol. Evol.* **14**, 1335–1338.
- Zhang, J., Nielsen, R., and Yang, Z. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479.
- Zhang, J., Rosenberg, H. F., and Nei, M. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. USA* **95**, 3708–3713.
- Zhou, T., Gu, W., and Wilke, C. O. (2010) Detecting positive and purifying selection at synonymous sites in yeast and worm. *Mol. Biol. Evol.* **27**, 1912–1922.
- Zhuang, H., Chien, M.-S., and Matsunami, H. (2009) Dynamic functional evolution of an odorant receptor for sex-steroid-derived odors in primates. *Proc. Natl. Acad. Sci. USA* **106**, 21247–21251.