Detecting Past Positive Selection through Ongoing Negative Selection

Georgii A. Bazykin[1,2] and Alexey S. Kondrashov[1,3,*]

[1]Department of Bioengineering and Bioinformatics, M. V. Lomonosov Moscow State University, Moscow, Russia

[2]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

[3]Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan

*Corresponding author: E-mail: kondrash@umich.edu.

Accepted: 3 August 2011

Abstract

Detecting positive selection is a challenging task. We propose a method for detecting past positive selection through ongoing negative selection, based on comparison of the parameters of intraspecies polymorphism at functionally important and selectively neutral sites where a nucleotide substitution of the same kind occurred recently. Reduced occurrence of recently replaced ancestral alleles at functionally important sites indicates that negative selection currently acts against these alleles and, therefore, that their replacements were driven by positive selection. Application of this method to the *Drosophila melanogaster* lineage shows that the fraction of adaptive amino acid replacements remained approximately 0.5 for a long time. In the *Homo sapiens* lineage, however, this fraction drops from approximately 0.5 before the Ponginae–Homininae divergence to approximately 0 after it. The proposed method is based on essentially the same data as the McDonald–Kreitman test but is free from some of its limitations, which may open new opportunities, especially when many genotypes within a species are known.

Key words: natural selection, amino acid substitutions, polymorphism, divergence, McDonald–Kreitman test, allele frequency spectrum.

After the Beneficial Allele Gets Fixed, Positive Selection Turns into Negative Selection

At any given moment of time, positive selection, which favors currently uncommon derived alleles, affects only a small fraction of sites in the genome and, thus, is much rarer than negative selection, which favors common ancestral alleles (Kimura 1983). A variety of methods are used to detect positive selection, both past (McDonald and Kreitman 1991; Yang and Bielawski 2000; Smith and Eyre-Walker 2002; Bazykin et al. 2004; Eyre-Walker 2006; Huelsenbeck et al. 2006) and ongoing (Nielsen et al. 2007; Novembre and Di Rienzo 2009; Grossman et al. 2010), but neither of these methods is perfect. Here, we propose a method for detecting past positive selection through ongoing negative selection.

After a positive selection-driven allele replacement is over, positive selection transforms into negative selection (fig. 1*a*). Thus, at a site where an allele replacement occurred recently, ongoing negative selection against the ancestral allele (which is incessantly recreated by mutation) indicates, as long as the fitness landscape remains invariant, that this replacement has been driven by positive selection. Past allele replacements can be revealed by comparison of the species in which negative selection is studied to other species; here, we use maximum parsimony to infer them (fig. 1*b* and supplementary fig. S1, Supplementary Material online), but other methods, for example, maximum likelihood-based or Bayesian, can be used as well. In turn, ongoing negative selection at functionally important sites where allele replacements occurred during some time interval in the past can be detected using the polymorphism data. Specifically, we compare the prevalence of ancestral alleles at these sites and at supposedly selectively neutral sites where the allele replacements of the same type (e.g., for the case of single nucleotide substitutions, corresponding to the same ancestral-derived nucleotides pair) also occurred during the same time interval. This last requirement is necessary to control for the difference in the mutation rates across the genome; in particular, sites which underwent recent allele replacements are likely to have locally elevated mutation rate (Asthana et al. 2007; Bazykin et al. 2007; Hodgkinson et al. 2009).

Using Data on Current Negative Selection to Reveal Past Positive Selection

Let us compare functionally important nonsynonymous sites of protein-coding genome regions with synonymous sites, which will be assumed to be selectively neutral. In the McDonald–Kreitman (MK) test, the proportion of positive selection-driven nonsynonymous replacements is estimated, under the assumption that a nonsynonymous mutation can be strongly advantageous, strongly deleterious, or neutral, as $\alpha = 1 - D_sP_n/(D_nP_s)$, where D_n and D_s are the numbers of nonsynonymous and synonymous substitutions and P_n and P_s are the numbers of polymorphic nonsynonymous and synonymous sites within the same sample (McDonald and Kreitman 1991; Smith and Eyre-Walker 2002; Eyre-Walker 2006). In the test proposed here, this proportion is estimated, assuming that a nonsynonymous replacement can be either strongly advantageous or neutral, as $\beta = 1 - p_n/p_s$, where p_n (p_s) is the proportion of sites, among those nonsynonymous (synonymous) sites at which a nucleotide replacement occurred in the past, which currently carry both the derived and the ancestral nucleotides (fig. 1*b*). Note that while α is dependent on all polymorphic sites in the analyzed set of sites, β only takes into account polymorphism at sites at which a nucleotide replacement has previously occurred in a particular segment of the considered lineage. Estimating positive selection through α and through β is based on the same fact that a neutral nonsynonymous mutation contributes as much to polymorphism as a synonymous mutation and a strongly deleterious nonsynonymous mutation contributes nothing (McDonald and Kreitman 1991; Smith and Eyre-Walker 2002; Eyre-Walker 2006).

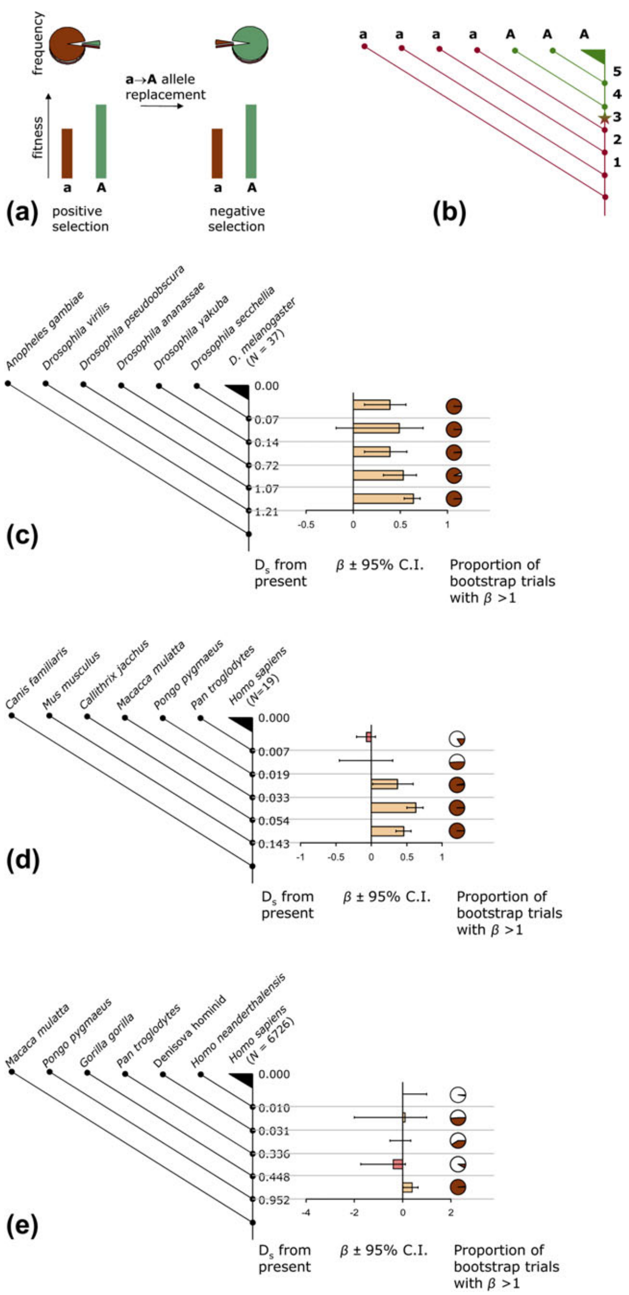

Fig. 1.—Test for positive selection based on polymorphism at sites of ancestral divergence. (*a*) Change in the mode of selection as a result of an allele replacement. A fitness landscape that initially causes positive selection in favor of a rare allele "A" (left) causes negative selection against a rare allele "a" after "a"→"A" allele replacement is accomplished (right). Fitnesses are shown by vertical bars and allele frequencies are shown by pie charts. (*b*) Approach to measuring past positive selection. Extant species (dots) were used to infer allele replacements ("a"→"A") that occurred at different segments (1–5; in the example shown, 3) of the ancestral lineage (see also supplementary fig. S1, Supplementary Material online). At sites of such replacements, the species for which polymorphism data is available (shown as a triangle) was used to assess the frequency of the ancestral variant ("a"). Such frequencies at nonsynonymous and synonymous sites were compared with measure β. (*c*–*e*) Results of the test for positive selection for substitutions which occurred in the lineage of the *Drosophila melanogaster* nuclear genome (*c*), *Homo sapiens* nuclear genome (*d*), or *H. sapiens* mitochondrial genome (*e*). The considered phylogeny is shown together with the times of the beginning and end of each segment of the lineage, measured in units of D_s from present. The species for which polymorphism data has been analyzed is shown as a triangle, with the number of available haploid genotypes *N* shown next to the species name. For each of the five considered segments, presented are the values of β together with 95% confidence intervals as horizontal bars, and the proportion of bootstrap replicates with $\beta > 0$ as pie charts.

Numerical Simulations

In order to investigate the proposed test and to compare it with the MK test, we analyzed the results of simulated molecular evolution. The data on divergence and polymorphism were generated by evolving a Wright–Fisher population along the phylogenetic tree corresponding to the actual phylogenetic tree of a clade within genus *Drosophila* (for details, see Materials and Methods). The genome was assumed to consist of many unlinked diallelic synonymous and nonsynonymous sites. All synonymous sites where assumed to be neutral. A nonsynonymous site evolved under one of the three modes of selection: neutrality, constant selection always favoring one of the alleles, or switching selection. In the latter case, the absolute value of the selection coefficient remained constant, but its sign switched at random moments of time, which led to episodes of positive selection favoring a previously inferior low-frequency allele.

We combined nonsynonymous sites under these three selection modes in different proportions and studied the behavior of α and β. Both tests performed well in detecting the fraction of positively selected substitutions when switching selection sites were combined with neutral sites (fig. 2*a*), with sites of very weak constant selection (fig. 2*b*), or with sites of strong constant selection (fig. 2*c*). α is sensitive to slightly deleterious alleles segregating within a population, so that admixture of constant selection sites with small selection coefficients leads to negative values of α; this can be remedied by excluding low-frequency polymorphisms (fig. 2*d*). When positive selection is present, the same effect leads to underestimation of the fraction of positively selected substitutions by α (fig. 2*e*–*g*). Both β and α with excluded low-frequency polymorphism give a reasonably good approximation of the fraction of the positively selected substitutions under all these scenarios (fig. 2*a*–*g*).

Positive Selection in Fruit Fly and Human Revealed by the Proposed Test

Supplementary table S1 (Supplementary Material online) presents data on nonsynonymous and synonymous allele replacements in the lineages of *Drosophila melanogaster* and *Homo sapiens*, each split into five segments, and on nonsynonymous and synonymous polymorphisms due to the presence of ancestral alleles at sites where these replacements occurred. So far, these are the only two species with extensive genome-level data on intraspecies variation available. We can see that our test implies that in the *D. melanogaster* lineage, the fraction of nonsynonymous replacements that were driven by positive selection remained steady at approximately 0.5, in agreement with the estimates obtained by the MK test (Eyre-Walker 2006) (fig. 1*c*).

Inferring substitutions by comparing with a single reference sequence may be erroneous when the reference sequence carries a low-frequency allele. In our approach, such errors are unlikely, because in each analysis, both the ancestral and the derived allele are observed in multiple sequences (supplementary fig. S1, Supplementary Material online). In rare instances, however, mistaking a polymorphic variant for a fixed one can lead to misidentification of the exact segment where the substitution has occurred. To assess this effect, we repeated the analysis for segment 5 of the *Drosophila* phylogeny using only the sites at which the nucleotide of *D. sechellia* coincided with the nucleotide of a sister species *D. simulans.* The results were similar (supplementary table S1, Supplementary Material online).

In contrast to *Drosophila,* in the human lineage, the fraction of positively selected substitutions declined, after Ponginae–Homininae divergence around 20 Ma (Ruff 2003), from approximately 0.5 to 0 (fig. 1*d*). This decline was probably caused by the increased body size (Ruff 2003) and the associated decrease of the effective population size (Popadin et al. 2007) and the efficiency of positive selection in the course of hominid evolution. Also, a decline of N_e may bias β downward, due to fixations of slightly deleterious alleles that were previously kept rare by more efficient selection, because after such a fixation, slightly beneficial ancestral alleles will segregate at elevated frequencies, compared with that after a selectively neutral fixation (Charlesworth and Eyre-Walker 2007). Thus, $\beta = 0$ in the hominid segments of the human lineage does not mean that there were no adaptive nonsynonymous replacements at that time. The MK test also does not reveal a large role of positive selection in the evolution of proteins in hominids (Eyre-Walker 2006).

Finally, our analysis does not reveal any positive selection within human mitochondrially encoded proteins, except early in the evolution of Hominidae (fig. 1*e*). Partially, this may be due to a small number of sites where allele replacements occurred recently. Also, because of the large number of known human mitochondrial genotypes, characterizing their variation by the fraction of polymorphic sites leads to a loss of information. However, distributions of ancestral allele frequencies at polymorphic sites of recent nonsynonymous and synonymous allele replacements are also very similar (fig. 3*a*), arguing against a major role of positive selection in the evolution of mitochondrially encoded proteins in the human lineage during the last approximately 20 Myr.

Distributions of frequencies of ancestral alleles are also similar at polymorphic sites of recent nonsynonymous and synonymous allele replacements in human nuclear genes (fig. 3*b* and *c*), which is consistent with the assumption that nonsynonymous replacements were either strongly advantageous or neutral. In contrast, in *D. melanogaster* there is a substantial excess of singletons at polymorphic nonsynonymous sites (fig. 3*d*). This suggests that some of the allele replacements were driven by positive selection so weak (Watterson 1975) that the ancestral allele is present in the sample of 37 genotypes, albeit at a low frequency.

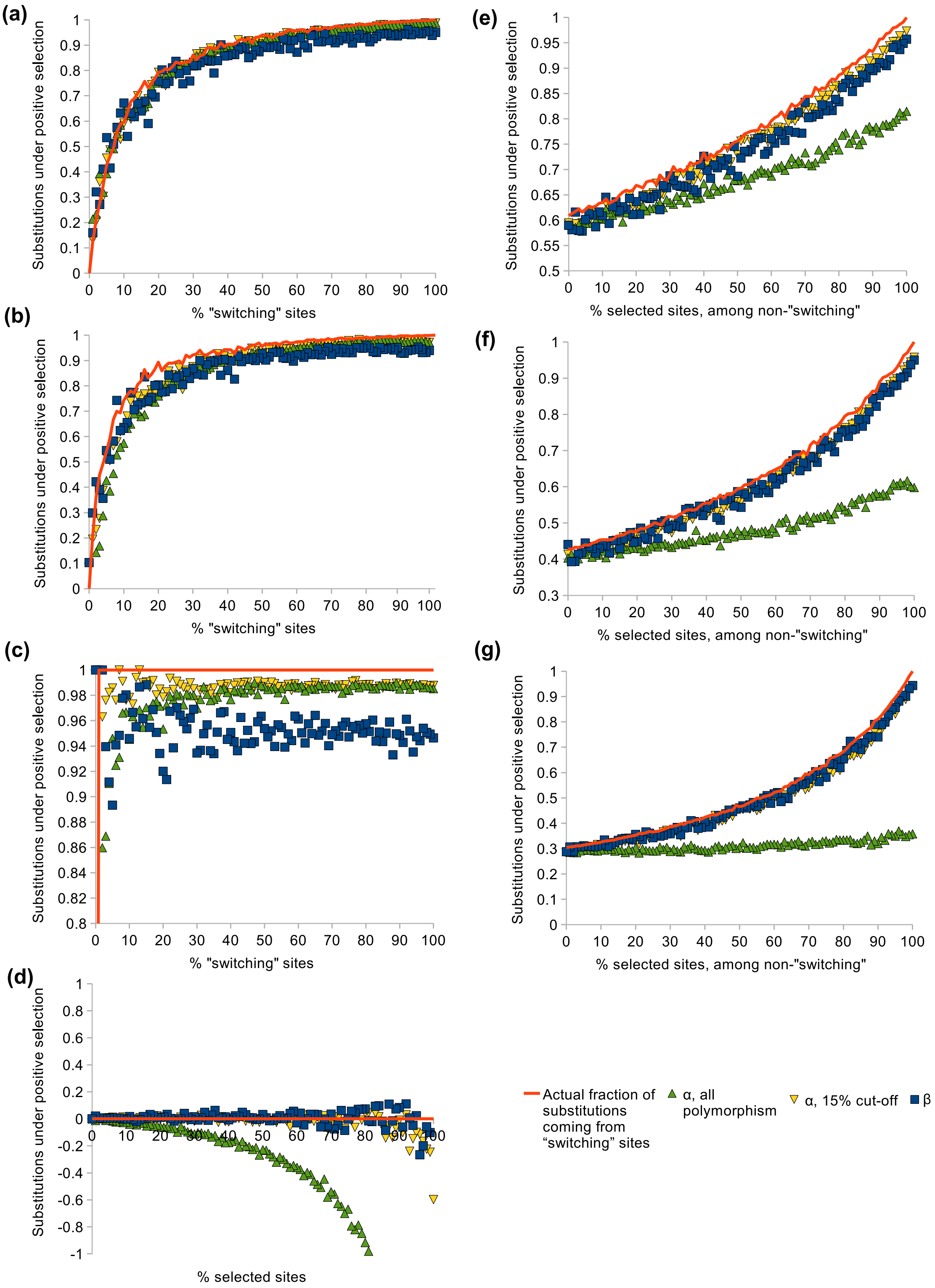

Fig. 2.—Performance of the MK test and the proposed test on mixtures of sites with different modes of selection. Orange solid line, the actual fraction of substitutions coming from the sites under switching selection; green triangles, α; yellow triangles, α with low-frequency (<15%) polymorphisms excluded; blue squares, β. (*a*–*c*) Horizontal axis: the fraction of switching sites ($s = \pm 10^{-3}$); the remaining sites are neutral (*a*) or are under very weak (*b*, $s = 10^{-5}$) or strong (*c*, $s = 10^{-3}$) constant selection. (*d*) Horizontal axis: the fraction of sites under weak constant selection ($s = 10^{-4}$); the remaining sites are neutral. (*e*–*g*) The fraction of switching sites ($s = \pm 10^{-3}$) is 10% (*e*), 5% (*f*), or 3% (*g*); horizontal axis: the fraction of sites under weak constant selection ($s = 10^{-4}$); the remaining sites are neutral.

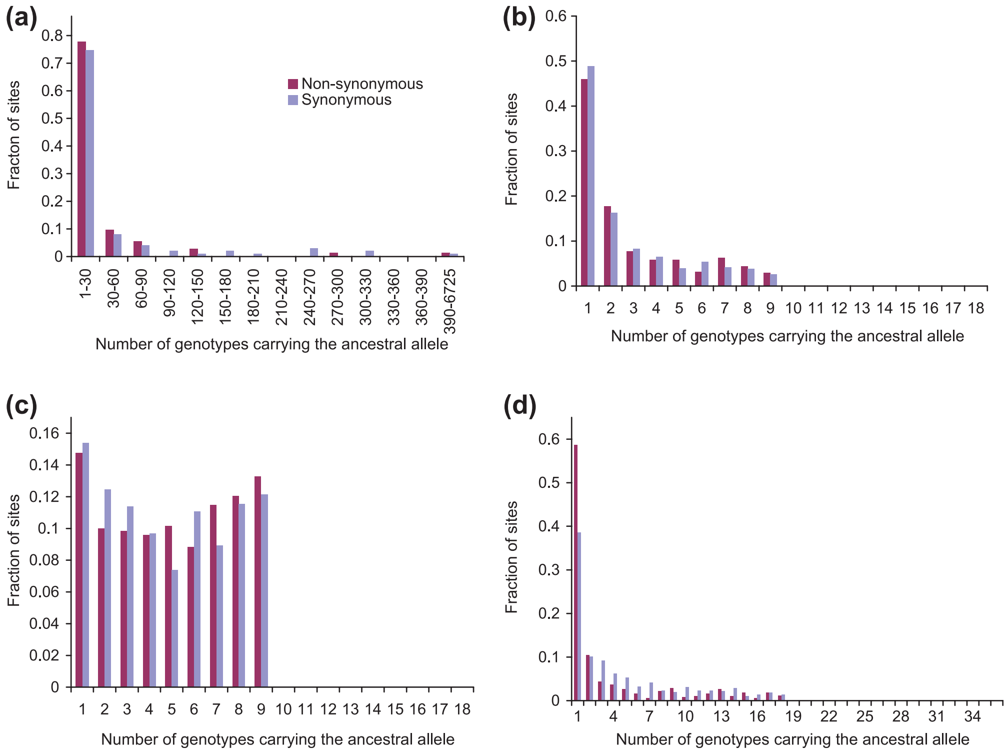

Fig. 3.—Distributions of the number of genotypes carrying the ancestral alleles at polymorphic sites. Only sites carrying the derived allele in >50% of the genotypes, and the ancestral allele in some of the genotypes, were taken into account. (*a*) Data for 6,726 human mitochondrial genotypes, for replacements that occurred in any of the five segments (Mann–Whitney *U* test for frequency of ancestral nonsynonymous vs. synonymous allele, $n = 159$, $P = 0.053$). (*b*) and (*c*) Data for 19 human nuclear genotypes, for replacements that occurred in segments 1–3 (*b*; $n = 1184$, $P = 0.355$) and 5 (*c*; $n = 1870$, $P = 0.108$); the excess of high-frequency polymorphism in this case is due to unfinished replacements. (*d*) Data for 37 *Drosophila melanogaster* genotypes, for replacements that occurred in segments 1–4 ($n = 1892$, $P = 1.03 \times 10^{-8}$). Nonsynonymous sites, purple; synonymous sites, blue. For the complete data set, see supplementary figures S2–S4 (Supplementary Material online).

Allele frequencies are not taken into account by β, which thus underestimates the role of weak positive selection. More detailed analysis which takes into account the distribution of allele frequencies, similar to that recently proposed for the MK test (Charlesworth and Eyre-Walker 2008; Eyre-Walker and Keightley 2009), is needed to address this problem. Our test can also underestimate the role of positive selection in nucleotide substitutions confined to the terminal segment of the lineage (segment 5 in fig. 1*b*) because a substantial fraction of them have not yet been accomplished, leading to an excess of ancestral polymorphisms (compare fig. 3*b* and *c*).

Comparison of the Proposed Test to the MK Test

The MK test and the test proposed here are both based on data on interspecies divergence and intraspecies polymorphism and apparently produce consistent estimates. Still, there is a number of substantial differences between the two tests. First, the proposed test does not compare rates of evolution at nonsynonymous versus synonymous sites; instead, it uses past synonymous replacements only to avoid a bias due to cryptic variation in mutation rates (Asthana et al. 2007; Bazykin et al. 2007; Hodgkinson et al. 2009). Second, the proposed test only deals with nonsynonymous sites that underwent a recent allele replacement, thus avoiding altogether possible complications due to mildly deleterious alleles that segregate at sites that remained under constant, negative selection for a long time and, thus, have not underwent allele replacements. In contrast, in the MK test, such alleles may lead either to underestimation (Charlesworth and Eyre-Walker 2008) or overestimation (Eyre-Walker 2002) of the role of positive selection, depending on the peculiarities of population demography. Third, the proposed approach can estimate not only the fraction of adaptive allele replacements, but also the strength of past positive selection which drove them, through the strength of ongoing negative selection (Keightley and Eyre-Walker 2007). This possibility will become especially important with the increase in the amount of genome-level data on intraspecies genetic variation, which will greatly increase the power of quantifying negative selection.

More generally, all existing tests for positive selection are involved with serious problems. Positive selection at a site does not act constantly; instead, under realistic assumptions, it probably changes into negative selection immediately after an allele replacement is accomplished and stays such for a while (e.g., Kryazhimskiy and Plotkin 2008; Mustonen and Lässig 2009). The implications of this for the D_n/D_s test have been discussed (Kryazhimskiy et al. 2008). The MK test makes the same assumption implicitly. Indeed, it assumes that the fraction of sites under negative selection is independent of the action of positive selection (Smith and Eyre-Walker 2002); in reality, polymorphism is probably reduced at sites of past positive selection. The test we propose uses this exact feature of past positive selection to detect it.

Materials and Methods

Numerical Simulation

A Wright–Fisher population of constant size $N = N_e = 10^5$ was evolved under a free recombination model, with two possible alleles in each locus, and the mutation rate of $\mu = 10^{-7}$ between them. At a generation, each locus was characterized by the selection coefficient *s* and (for the switching selection mode) the waiting time for this coefficient to switch sign. At each generation, a switch occurred with the same probability; we used a waiting time of 2×10^8 generations, so that switches were rare. After an initial burn-in of 10^7 generations, evolution then proceeded for another 7×10^7 generations, with cladogenesis events occurring according to the prescribed phylogeny shown in figure 1c. After each cladogenesis event, both derived species inherited the ancestral allele frequency and selection coefficient. A total of 1.4×10^6 loci were simulated: 10^5 for each of the four constant selection coefficients used and 10^6 for switching selection.

The results in figure 2 were obtained using the simulated equivalent of segment 5 of the *D. melanogaster* lineage in figure 1c. A site was categorized as “switched” if it had undergone a switch of the selection coefficient at this segment. Polymorphism was assessed in the terminal species (corresponding to *D. melanogaster)* by drawing 37 random alleles from the population. For the MK test with the low-frequency cutoff, all polymorphism with minor allele occurring in fewer than 6 of these 37 individuals was excluded. To make α and β more comparable, we used polarized MK, so that only substitutions in the simulated equivalent of *D. melanogaster* lineage since its divergence from *D. simulans* were considered; *D. erecta* was used to infer the ancestral state. Similarly, β was calculated at sites of substitutions in *D. melanogaster* lineage since its divergence from *D. simulans,* which were inferred using *D. erecta* as an outgroup.

Data

Multiple alignments of genome assemblies of six insects species to *D. melanogaster* (dm3) were obtained from UCSC Genome Bioinformatics Site (Kuhn et al. 2009) (http://genome.ucsc.edu). Complete genotypes of 37 inbred strains of *D. melanogaster* (Jordan et al. 2007) were obtained from *Drosophila* Population Genomics Project website (http://www.dpgp.org/). The set of FlyBase (Tweedie et al. 2009) canonical splicing variants (BDGP release 5) was used to map *D. melanogaster* protein-coding genes onto the alignment. Multiple alignment of each coding region was then obtained by joining the aligned segments corresponding to exons of FlyBase canonical genes in *D. melanogaster*. Codons masked by RepeatMasker, not aligned, or containing gaps or non-ACGT characters, as well as codons within six nucleotides of any of such codons, were excluded from the analysis.

Multiple alignments of genome assemblies of six vertebrate species to *H. sapiens* (hg18) were obtained from UCSC Genome Bioinformatics Site (Kuhn et al. 2009) (http://genome.ucsc.edu). Data on variation of human nuclear genotypes were obtained from nine diploid human genotypes downloaded from Galaxy bioinformatics platform (Taylor et al. 2007; Schuster et al. 2010) (http://usegalaxy.org) and the reference human genome, resulting in 19 haploid genotypes. The following individual diploid genotypes were used: KB1 (454 method) (Schuster et al. 2010), ABT (SOLiD method) (Schuster et al. 2010), NA18507 (Bentley et al. 2008), NA19240 (Drmanac et al. 2010), Craig Venter (Levy et al. 2007), NA12891, NA12892, Chinese individual (Wang et al. 2008), and Korean individual (Ahn et al. 2009). The canonical splicing variants of UCSC hg18 Known Genes (Hsu et al. 2006) were used to map *H. sapiens* protein-coding genes onto the alignment. Multiple alignment of each coding region was then obtained by joining the aligned segments corresponding to exons of knownGene canonical genes in *H. sapiens*.

The sequences of 6,726 complete human mitochondrial genotypes were obtained from GenBank (Benson et al. 2009) by using “Homo sapiens [orgn] and complete genome” as a query with the Limits option set to mitochondrial DNA in the Entrez retrieval system (Baxevanis 2008). The sequences of six non-*H. sapiens* primate species were obtained from GenBank (Benson et al. 2009). All sequences were aligned to the revised Cambridge sequence (Andrews et al. 1999) using ClustalW (Thompson et al. 1994), and coding sequences for 12 protein-coding genes (excluding ND6, which is encoded on a different strand than the rest of the genes and has biased nucleotide composition; Yang et al. 1998) were extracted from the alignments.

Lengths of internal segments of insect and primate mitochondrial phylogenetic trees were taken from references Heger and Ponting (2007) and Krause et al. (2010). Lengths of internal segments of vertebrate phylogenetic tree were taken from UCSC Genome Bioinformatics Site (Kuhn et al. 2009). All lengths are in the units of the inferred per site number of synonymous substitutions (D_s).

Analysis

Codon sites with gaps or missing data in any of the six non-*D. melanogaster* (or non-*H. sapiens*) species, in any of the 37

D. melanogaster genotypes, in any of the 19 *H. sapiens* nuclear genotypes, or in more than 1,726 of 6,726 *H. sapiens* mitochondrial genotypes, were excluded from analysis. At each codon site, nucleotide sites were classified as "nonsynonymous" ("synonymous") only when each of the four nucleotides at this site corresponded to a different (the same) amino acid in all the seven species (supplementary fig. S1, Supplementary Material online). An allele replacement was assigned to 1 of the 5 segments of the ancestral lineage of a polymorphic species using maximum parsimony (fig. 1*b* and supplementary fig. S1, Supplementary Material online). Nucleotide sites at which parsimony implied more than one allele replacement, or at which multiple timings of the allele replacement were equally parsimonious, were excluded from further analysis.

Polymorphism was measured at sites of past allele replacement. We included in analysis only those sites where the frequency of the derived allele ("A" in fig. 1*b* and supplementary fig. S1, Supplementary Material online) was above 50%; such sites constituted the vast majority of all sites for allele replacements in segments 1–4. By contrast, when only the species whose variation is studied carries the derived allele, the distribution of allele frequencies contains a much higher fraction of intermediate values (segment 5 vs. segments 1–4 in fig. 1*b*).

Among sites with allele replacements that occurred within a particular segment, each of the 12 possible pairs of ancestral and derived nucleotides was considered separately. This was done to control for different rates of different mutations. For each pair of nucleotides involved in a past nonsynonymous (synonymous) allele replacement, we estimated, in the species used to study variation, the fraction of nonsynonymous (synonymous) sites with the ancestral variant present. These fractions were then averaged across the 12 possible nucleotide pairs, separately for nonsynonymous and synonymous sites, to produce the values in supplementary table S1 (Supplementary Material online) and in figure 1*c*–*e* and to calculate β. Among polymorphic sites with the ancestral variant present in the species used to study variation, the distributions of frequencies of the ancestral nonsynonymous versus synonymous variants were compared using the Mann–Whitney *U* test. The confidence intervals for β were obtained by bootstrapping the data. In each of the 1000 bootstrap replicates, we randomly selected with replacement the same number of sites from the observed set of sites with an allele replacement and calculated β for this resampled distribution.

Acknowledgments

We thank Dmitry Pervouchine, Andrey Mironov, and Konstantin Popadin for valuable comments. This work was supported by the Russian Ministry of Science and Education (grant number 11.G34.31.0008 and contract number P916) and the Molecular and Cellular Biology Program of the Russian Academy of Sciences.

Supplementary Material

Supplementary figures S1–S4 and table S1 are *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Literature Cited

Ahn SM, et al. 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res. 19:1622–1629.

Andrews RM, et al. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. 23:147.

Asthana S, et al. 2007. Widely distributed noncoding purifying selection in the human genome. Proc Natl Acad Sci U S A. 104:12410–12415.

Baxevanis AD. 2008. Searching NCBI databases using Entrez. Curr Protoc Bioinformatics Chapter 1:Unit 1.3.

Bazykin GA, Kondrashov FA, Ogurtsov AY, Sunyaev S, Kondrashov AS. 2004. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. Nature 429:558–562.

Bazykin GA, et al. 2007. Extensive parallelism in protein evolution. Biol Direct. 2:20.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2009. GenBank. Nucleic Acids Res. 37:D26–D31.

Bentley DR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.

Charlesworth J, Eyre-Walker A. 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. Proc Natl Acad Sci U S A. 104:16992–16997.

Charlesworth J, Eyre-Walker A. 2008. The McDonald-Kreitman test and slightly deleterious mutations. Mol Biol Evol. 25:1007–1015.

Drmanac R, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327:78–81.

Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. Genetics 162:2017–2024.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. Trends Ecol Evol. 21:569–575.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 26:2097–2108.

Grossman SR, et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327:883–886.

Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. Genome Res. 17:1837–1849.

Hodgkinson A, Ladoukakis E, Eyre-Walker A. 2009. Cryptic variation in the human mutation rate. PLoS Biol. 7:e1000027.

Hsu F, et al. 2006. The UCSC known genes. Bioinformatics 22:1036–1046.

Huelsenbeck JP, Jain S, Frost SW, Pond SL. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proc Natl Acad Sci U S A. 103:6263–6268.

Jordan KW, Carbone MA, Yamamoto A, Morgan TJ, Mackay TF. 2007. Quantitative genomics of locomotor behavior in Drosophila melanogaster. Genome Biol. 8:R172.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251–2261.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Krause J, et al. 2010. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature 464:894–897.

Kryazhimskiy S, Bazykin GA, Plotkin J, Dushoff J. 2008. Directionality in the evolution of influenza A haemagglutinin. Proc R Soc B Biol Sci. 275:2455–2464.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. PLoS Genet. 4:e1000304.

Kuhn RM, et al. 2009. The UCSC Genome Browser Database: update 2009. Nucleic Acids Res. 37:D755–D761.

Levy S, et al. 2007. The diploid genome sequence of an individual human. PLoS Biol. 5:e254.

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

Mustonen V, Lässig M. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends Genet. 25:111–119.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. Nat Rev Genet. 8:857–868.

Novembre J, Di Rienzo A. 2009. Spatial patterns of variation due to natural selection in humans. Nat Rev Genet. 10:745–755.

Popadin K, Polishchuk LV, Mamirova L, Knorre D, Gunbin K. 2007. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. Proc Natl Acad Sci U S A. 104:13390–13395.

Ruff CB. 2003. Long bone articular and diaphyseal structure in Old World monkeys and apes. II: estimation of body mass. Am J Phys Anthropol. 120:16–37.

Schuster SC, et al. 2010. Complete Khoisan and Bantu genomes from southern Africa. Nature 463:943–947.

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature 415:1022–1024.

Taylor J, Schenck I, Blankenberg D, Nekrutenko A. 2007. Using galaxy to perform large-scale interactive data analyses. Curr Protoc Bioinformatics. Chapter 10:Unit 10.5.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Tweedie S, et al. 2009. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res. 37:D555–D559.

Wang J, et al. 2008. The diploid genome sequence of an Asian individual. Nature 456:60–65.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7:256–276.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol. 15:1600–1611.

Associate editor: Takashi Gojobori