Detection of nonneutral substitution rates on mammalian phylogenies

Katherine S. Pollard, Melissa J. Hubisz, Kate R. Rosenbloom, et al.

Genome Res. 2010 20: 110-121 originally published online October 26, 2009
Access the most recent version at doi:10.1101/gr.097857.109

Detection of nonneutral substitution rates on mammalian phylogenies

Katherine S. Pollard,[1,4] Melissa J. Hubisz,[2] Kate R. Rosenbloom,[3] and Adam Siepel[2]

[1]*Gladstone Institutes, University of California, San Francisco, San Francisco, California 94158, USA;* [2]*Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14850, USA;* [3]*Center for Biomolecular Science and Engineering, School of Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA*

Methods for detecting nucleotide substitution rates that are faster or slower than expected under neutral drift are widely used to identify candidate functional elements in genomic sequences. However, most existing methods consider either reductions (conservation) or increases (acceleration) in rate but not both, or assume that selection acts uniformly across the branches of a phylogeny. Here we examine the more general problem of detecting departures from the neutral rate of substitution in either direction, possibly in a clade-specific manner. We consider four statistical, phylogenetic tests for addressing this problem: a likelihood ratio test, a score test, a test based on exact distributions of numbers of substitutions, and the genomic evolutionary rate profiling (GERP) test. All four tests have been implemented in a freely available program called phyloP. Based on extensive simulation experiments, these tests are remarkably similar in statistical power. With 36 mammalian species, they all appear to be capable of fairly good sensitivity with low false-positive rates in detecting strong selection at individual nucleotides, moderate selection in 3-bp elements, and weaker or clade-specific selection in longer elements. By applying phyloP to mammalian multiple alignments from the ENCODE project, we shed light on patterns of conservation/acceleration in known and predicted functional elements, approximate fractions of sites subject to constraint, and differences in clade-specific selection in the primate and glires clades. We also describe new "Conservation" tracks in the UCSC Genome Browser that display both phyloP and phastCons scores for genome-wide alignments of 44 vertebrate species.

[Supplemental material is available online at http://www.genome.org.]

In recent years, the technique of scanning aligned genomic sequences for elements that are evolving faster, slower, or by different patterns than would be expected under neutral drift has emerged as a powerful approach for discovering novel functional elements. This technique is particularly useful in mammalian genomes, because of their size, complexity, and relative intractability in experimental investigation. Computational scans of mammalian genomes have been used to identify various classes of functional elements, including protein-coding genes (Guigó et al. 2003; Siepel et al. 2007), RNA genes (Pedersen et al. 2006), enhancers (Nobrega et al. 2003), and micro-RNA target sites (Xie et al. 2005). These methods have become steadily more valuable as deep alignments of orthologous sequences—which until recently covered only a small fraction of the genome (The ENCODE Project Consortium 2007)—have become available genome-wide (Miller et al. 2007; Mammalian Genome Sequencing and Analysis Consortium, in prep.).

A wide variety of methods have been introduced for detecting signatures of nonneutral evolution in aligned genomic sequences, and they can be categorized in various ways (Supplemental material S1; Supplemental Table S1). For example, some methods depend on pre-defined annotations for training (as in gene-finding), while others can be used in a fully "unsupervised" manner with unannotated genomic sequences; some methods make full use of the phylogenetic relationships among the species in question, while others consider only pairwise comparisons; and some methods make use of statistical models of molecular evolution, while others use heuristic scores or invoke parsimony assumptions. In this study, we focus on unsupervised, statistical, phylogenetic methods, which we believe have the greatest promise for general functional element discovery and characterization, even if they are sometimes outperformed by more specialized approaches in particular classification tasks (e.g., Gross et al. 2007).

Within this class of methods, the primary signal used to identify sequences of interest has been conservation or constraint—that is, a reduced rate of evolution compared to what is expected under neutral drift (Boffelli et al. 2003; Margulies et al. 2003; Cooper et al. 2005; Siepel et al. 2005; Asthana et al. 2007). Recently, methods have been introduced for detecting sequences that are experiencing "acceleration," or faster-than-neutral evolution, with particular emphasis on scanning aligned genomic sequences for fast-evolving elements in the human lineage (Pollard et al. 2006b; Prabhakar et al. 2006; Bird et al. 2007) or other mammalian lineages (Haygood et al. 2007; Kim and Pritchard 2007; see also Wong and Nielsen 2004). Most conservation-detection methods have assumed uniform selection pressures across the branches of a phylogeny, but several acceleration-detection methods have allowed for lineage-specific selection (Pollard et al. 2006a; Prabhakar et al. 2006; Bird et al. 2007; Kim and Pritchard 2007). In addition, most conservation-detection methods have been designed to scan entire genomic alignments, using a sliding window (Margulies et al. 2003; Cooper et al. 2004), a hidden Markov model (Siepel et al. 2005), or, as increasingly deep alignments have become available, by measuring conservation at individual nucleotides and then identifying runs of sites with a combined score above an empirically determined threshold (Cooper et al. 2005; Asthana et al. 2007). In contrast, acceleration-detection methods have generally been applied to predefined elements of interest.

In this study, we treat conservation and acceleration in a unified manner and examine the general problem of detecting departures from the neutral rate of substitution in either direction. We consider elements of any length (including single nucleotides)

[4]Corresponding author.
E-mail kpollard@gladstone.ucsf.edu; fax (415) 355-0141.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.097857.109. Freely available online through the *Genome Research* Open Access option.

20:110–121 © 2010 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/10; www.genome.org

and allow for clade-specific selection as well as selection that acts uniformly across the phylogeny. In these respects, our study is similar to that of Kim and Pritchard (2007), but we focus more on questions of methodology, considering various alternative approaches to this problem. In particular, we conduct an extensive series of simulation experiments to compare the performance of four methods for detecting nonneutral substitution rates on a phylogeny: a likelihood ratio test, a score test, a method based on the distribution of the number of substitutions per site (Siepel et al. 2006), and the genomic evolutionary rate profiling (GERP) method (Cooper et al. 2005). We find that all four methods have fairly good power with currently available data for mammals, but they do have clear limitations, especially for short elements and elements experiencing weak or lineage-specific selection. Somewhat surprisingly, the four tests are nearly identical to one another in power, despite their different statistical underpinnings. We have implemented all four methods in a program called phyloP ("phylogenetic *P*-values"), which is freely available as part of the PHAST package (http://compgen.bscb.cornell.edu/phast). We apply phyloP to multiple alignments of 36 species in the ENCODE regions and analyze patterns of conservation/acceleration for various annotation classes and clades of interest. We also introduce new "Conservation" tracks in the UCSC Genome Browser (Kuhn et al. 2009) that display phyloP scores (alongside phastCons scores) for genome-wide alignments of 44 vertebrate species.

Results

Statistical tests and software implementation

The general statistical problem considered in this study is to identify a significant increase or decrease in the rate of substitution in a given genomic element, relative to what would be expected under neutral drift. As in most previous work, we assume a precomputed alignment of orthologous sequences from multiple species, and a corresponding neutral phylogeny with branch lengths in units of expected substitutions per site. To allow for sufficient power for short elements (1–10 bp), we consider tests that group multiple branches of the tree together, focusing on two particular types of tests: "all-branch tests," which examine increases or decreases in rate across all branches of the phylogeny; and "subtree tests," which examine increases or decreases in rate within a particular subtree (clade) of interest, relative to the rate in the remainder of the phylogeny (Supplemental Fig. S1). We consider four alternative methods for testing for nonneutral evolution, which we denote LRT, SCORE, SPH, and GERP. These methods are summarized in Table 1 and described in detail in the Methods and Supplemental material (Supplemental sections S2.1–S2.4). While these four methods share certain features (e.g., several of them make use of a common subroutine that estimates branch-length scaling parameters by maximum likelihood), in general, they have quite different properties—they rely on different test statistics, null distributions, and approximations. A major goal of this study is to compare and contrast their performance empirically, using data simulated from molecular evolutionary models and permuted real data. The four tests were implemented in the phyloP program in the PHAST package (Supplemental section S2.5).

Simulation study

We performed two types of simulation experiments to evaluate the false-positive rates and power of the tests implemented in phyloP. In the first, "parametric" series of experiments, we generated synthetic alignments from "neutral" and "selected" phylogenetic models and then measured how well elements of various sizes could be distinguished based on phyloP scores. In these experiments, neutral sites were generated from a phylogenetic model estimated from fourfold degenerate (4D) sites in 36-species alignments from the ENCODE regions (Methods). Selected models were generated from this model by scaling all branches by a factor ρ (all-branch cases) or the branches in a subtree by a factor λ (subtree cases), for various choices of ρ and λ (Methods). Thus, the simulation scheme reflects the same assumptions as the tests themselves and is expected to produce somewhat optimistic estimates of absolute performance. Nevertheless, it is useful for comparing the relative performance of the different methods under known, precisely characterized conditions. To complement these parametric experiments, we also performed a series of model-free, nonparametric bootstrapping experiments, drawing "neutral" and "selected" alignment columns from 4D and second codon position (CDS2) sites, respectively. (CDS2 sites were chosen because all nucleotide substitutions at these positions are nonsynonymous.) These simulations require fewer assumptions about the nucleotide substitution process, but their usefulness depends on how representative CDS2 sites are of the broader class of genomic sites under selection. In addition, they apply only to the all-branch tests, and not to the subtree tests. In all experiments, we computed phyloP scores for 1000 neutral and 1000 selected genomic elements, keeping track of false-positive rates (FPRs), true-positive rates (TPRs), false discovery rates (FDRs), and running times. We summarize performance using these statistics and the area under the receiver operating characteristic curve (AUC).

The most striking result from these experiments is that the four tests have nearly identical power across a broad range of scenarios (Fig. 1; Table 2; Supplemental Tables S2–S13), despite being based on quite different statistical principles. This is true for nonparametric and parametric simulations alike. Similar levels of power might be expected when signals are strong, but high concordance is also evident under weak conservation or acceleration. The only major differences between methods occur for subtree-specific conservation, for which the LRT method shows somewhat higher power than the SCORE method, which, in turn, shows higher power than the SPH method, especially for short elements (Supplemental Tables S2–S4). The SPH method also displays slightly reduced power for all-branch tests based on a reduced species set in the case of strong conservation (Supplemental Table S9), probably as a result of the highly discrete nature of the test statistic used by the SPH method (see Discussion). In all other cases, the methods are essentially indistinguishable.

Second, the absolute level of power for all four methods—with this 36-species mammalian phylogeny—appears to be fairly good. The parametric simulations indicate that, in the case of strong conservation (all-branch rescaling by a factor $\rho = 0.1$) or acceleration ($\rho = 3.3$), selection at single nucleotides can be identified reasonably reliably—for example, with TPRs of 90% (for $\rho = 0.1$) or 92% (for $\rho = 3.3$) at a per-element FPR of 5% (AUC of 0.97 and 0.99, respectively). At more moderate levels of conservation (e.g., $\rho = 0.3$, as observed on average in phastCons elements) (Siepel et al. 2005), power is weaker for single nucleotides, but for 3-bp elements it is possible to achieve a TPR of 97% at a 5% FPR (AUC = 0.99). Similarly, with acceleration at twice the neutral rate in 3-bp elements, a TPR of 87% at a 5% FPR is achievable (AUC = 0.97). For 10-bp elements (about the size of a typical transcription factor binding site), much milder constraint or acceleration can be detected with good power (e.g., $\rho = 0.7$ with

Table 1. Summary of statistical tests considered in this study

Test	Description[a]	Option[b]	Test statistic	Null[c]	References
Likelihood ratio test	Traditional hypothesis test for parametric models, central in the Neyman-Pearson framework. Here a null model and an alternative model, defined by different rate parameters ($\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, respectively), are both fitted to an alignment $\mathbf{X}$ by maximum likelihood, and twice the difference in their maximized log likelihoods is used as a test statistic.	LRT	$2\left[L\left(\hat{\boldsymbol{\theta}}_1\right) - L\left(\hat{\boldsymbol{\theta}}_0\right)\right]$	χ^2	Huelsenbeck and Rannala 1997; Casella and Berger 2002; Pollard et al. 2006b
Score test	Another traditional hypothesis test, with similar asymptotic properties as the LRT but the advantage that only the null model needs to be fitted to the data. The test statistic in this case is derived from the values of the score function U and the Fisher information matrix I, both evaluated at the maximum likelihood estimate under the null model, $\hat{\boldsymbol{\theta}}_0$.	SCORE	$U^T\left(\hat{\boldsymbol{\theta}}_0\right) I^{-1}\left(\hat{\boldsymbol{\theta}}_0\right) U\left(\hat{\boldsymbol{\theta}}_0\right)$	χ^2	Rao 1948, 2005
Number-of-substitutions test	Test based on the total number of substitutions n during the evolution of the element $\mathbf{X}$, under a phylogenetic model ψ. An exact null distribution is computed by a dynamic programming algorithm that depends on uniformization of the continuous-time Markov chain. The actual number for the observed data is approximated by the posterior mean, which is computed similarly.	SPH[d]	$E[n\|\psi,\mathbf{X}]$	Exact $p(n\|\psi)$	Siepel et al. 2006
GERP-like test	Test based on a statistic called "rejected substitutions," defined as the total branch length of the neutral phylogeny minus the total branch length after maximum likelihood estimation of a scale factor ρ. This test can be used in the all-branch setting but not the subtree setting.	GERP	$T(1-\hat{\rho})$	Empirical	Cooper et al. 2005

[a]See Methods for complete details.
[b]Option to –method argument in phyloP that specifies each test; also used throughout this study as an abbreviation for the test.
[c]Null distribution of test statistic assumed when computing *P*-values. The χ^2 distributions for the LRT and SCORE tests hold asymptotically but are approximate for finite data sets. See Methods for discussion of issues that arise in one-sided tests.
[d]The abbreviation "SPH" stands for "Siepel-Pollard-Haussler," the authors of the conference paper in which the relevant algorithms were introduced.

AUC = 0.94 or $\rho = 1.43$ with AUC = 0.96) (Table 2). At 50 bp, departures from the neutral rate by only ~10% are reliably detectable (data not shown). Detection power increases as ρ decreases or increases relative to the neutral value of $\rho = 1$, but it does so more rapidly for acceleration than for conservation. This behavior appears to result from a boundary effect from complete invariant sites (with no substitutions)—despite their perfect conservation, these sites typically have nonnegligible probability under the null model, limiting the ability of the tests to distinguish even extreme conservation from neutrality. For various element lengths, power in the nonparametric experiments is comparable to that observed in parametric (all-branch) experiments at $\rho = 0.5$.

We attempted to translate these FPRs (for each TPR) into predicted FDRs (expected fractions of predicted elements that are false-positives), which are of particular interest in applications in genomics. If the fraction of sites under selection is γ, then

$$\mathrm{FDR} \approx \left(1 + \frac{(1-\beta)\gamma}{\alpha(1-\gamma)}\right)^{-1}$$

for FPR α and TPR $(1-\beta)$ (Supplemental section S2.8). When γ is small (as it is believed to be in mammalian genomes), a relatively small FPR can still produce a large FDR for fixed TPR, essentially because the pool of sites from which false-positives are drawn is much larger than the pool from which true-positives are drawn. However, predicting the FDR is not straightforward because the true value of γ is unknown and sites under selection obey some unknown distribution of selection scenarios, each with its own TPR. Nevertheless, we were able to obtain crude estimates of FDR as a function of TPR by two indirect methods, focusing on the case of the all-branch LRT. First, we used CDS2 sites as a proxy for sites under selection, estimating TPR(s), for score threshold s, as the fraction of CDS2 elements having score $\geq s$. Second, we estimated

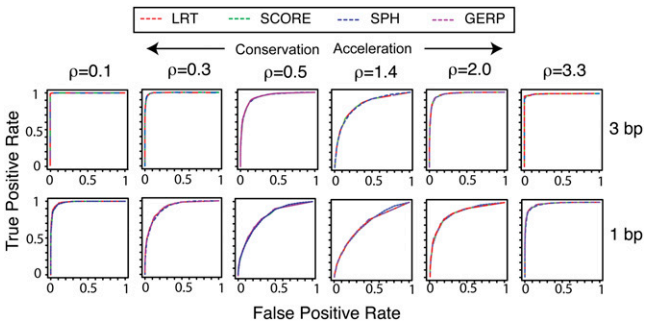

Figure 1. Receiver operating characteristic (ROC) curves showing false-positive versus true-positive rates for the all-branch tests implemented in phyloP: (red) LRT, (green) SCORE, (blue) SPH, and (purple) GERP. Individual plots show results for simulated data sets with either 3-bp (*top*) or 1-bp (*bottom*) elements generated from models with a range of deviations ρ from the neutral rate $\rho = 1.0$ (columns).

a distribution of phyloP scores for sites under selection by decomposing the full score distribution into a neutral component and a selected component, and calculating TPR(s) from this distribution (Supplemental section S2.8). We used maximum-likelihood estimates of γ in the CDS2 case and lower-bound estimates in the mixture decomposition case. These calculations suggest that if single-nucleotide elements are subject to selective effects similar to those in CDS2 sites, more than half can be detected with FDR ≈ 5% (Fig. 2). However, much higher FDRs must be tolerated if a large majority of 1-bp elements are to be detected (e.g., FDR ≈ 50% to detect two-thirds of 1-bp elements, or FDR ≈ 80% to detect 80%). If a broader class of elements subject to weaker selective effects is considered (as inferred by the mixture decomposition method), power is somewhat weaker, with an ~30% TPR at a 5% FDR and an ~40% TPR at a 50% FDR for 1-bp elements. Power for 3-bp elements shows a similar overall pattern, but is considerably higher than for 1-bp elements in the range of interest, with between about one-half (mixture) and three-quarters (CDS2) of elements detectable at 5% FDR. While these estimates are clearly subject to a great deal of uncertainty, they suggest that current alignments are sufficiently informative to allow substantial fractions—but not large majorities—of 1- to 3-bp elements to be detected at low FDRs. Power will improve as more sequence data become available.

Our third main finding is that the subtree tests (LRT, SCORE, and SPH only) have substantially lower power than the all-branch tests, yet do have reasonable power for slightly longer elements, provided subtrees of adequate size are considered. In our experiments, we considered three different clades, with different numbers of species and branch lengths: the primate (14 species, short branches), glires (five species, longer branches), and laurasiatherian (10 species, longer branches) clades (Fig. 3). In all cases, power is poor for individual nucleotides, except in the case of extreme clade-specific acceleration (subtree rescaling by a factor $\lambda = 10$) (Supplemental Tables S2–S4). At 3 bp, power is improved for moderate to strong departures from neutrality ($\lambda \leq 0.3$, $\lambda \geq 3.33$) but generally remains poor (Fig. 3). By 10 bp, however, power has improved considerably, with elements under moderate clade-specific conservation ($\lambda = 0.3$) showing power comparable to that seen for the all-branch test for 3-bp elements with $\rho = 0.5$ (primates; AUC of 0.93–0.95) or $\rho = 0.3$ (laurasiatherians; AUC of 0.98–0.99). Predictably, power is generally highest for the laurasiatherians and lowest for the primates, with the glires clade being intermediate between them.

To further examine the sensitivity of our results to modeling assumptions, we applied phyloP to two additional sets of synthetic alignments, simulated in more realistic ways. First, we generated data under a model that allows for rate variation across sites (Yang 1994), using parameters estimated from AR and CDS2 sites for neutral and selected sites, respectively (Supplemental section S2.7.1). Second, we relaxed the assumption (made by all subtree tests in phyloP) that all branches in a subtree of interest use one substitution rate, while all other branches use another, by introducing various amounts of "noise" to the branch-length scaling factors during data generation (Supplemental section S2.7.2). We applied phyloP to these alignments and measured its performance, exactly as above (i.e., the tests were not altered to reflect the new assumptions). These experiments indicated that simplifications in the parametric experiments do tend to inflate absolute estimates of power somewhat, but in general, the effect is not dramatic, and relative performance is mostly unaffected (Supplemental sections S3.5.1 and S3.5.2).

Finally, we compared and evaluated several other aspects of performance, including running time, two-sided versus one-sided tests, the effect of considering subsets of species, and the accuracy of reported P-values. The running times of the LRT and GERP methods were comparable, while the SCORE method was considerably faster, and the SPH method was considerably slower—by more than an order of magnitude in some cases. Results of the other experiments were largely consistent with expectations (Supplemental sections S3.3–S3.7; Supplemental Tables S6–S14).

Table 2. Area under the ROC curve for phyloP one-sided all-branch tests

	1 bp				3 bp				10 bp			
Rate (ρ)	GERP	LRT	SCORE	SPH	GERP	LRT	SCORE	SPH	GERP	LRT	SCORE	SPH
Conservation												
0.1	0.99	0.99	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.3	0.91	0.91	0.91	0.89	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00
0.5	0.80	0.79	0.79	0.79	0.94	0.94	0.94	0.94	1.00	1.00	1.00	1.00
0.7	0.67	0.66	0.66	0.66	0.81	0.80	0.80	0.80	0.94	0.94	0.94	0.94
0.9	0.56	0.55	0.55	0.56	0.61	0.60	0.60	0.61	0.69	0.68	0.68	0.68
Acceleration												
1.11	0.57	0.56	0.56	0.57	0.61	0.60	0.60	0.61	0.69	0.68	0.68	0.69
1.43	0.71	0.70	0.70	0.70	0.84	0.83	0.83	0.83	0.96	0.96	0.96	0.96
2	0.86	0.86	0.86	0.85	0.97	0.97	0.97	0.97	1.00	1.00	1.00	1.00
3.33	0.98	0.98	0.97	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

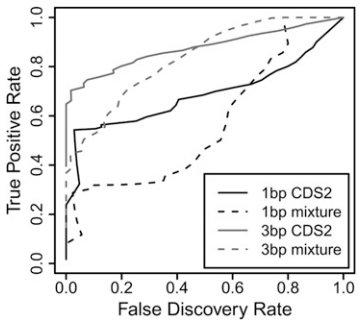

Figure 2. Estimated FDR for all-branch LRT. Estimates of false discovery rate (FDR) versus true-positive rate (TPR) based on two indirect methods, for 1-bp and 3-bp elements. (CDS2) Average TPRs are estimated from second codon position sites; (mixture) average TPRs are estimated by decomposing the genome-wide score distribution into components corresponding to neutral and selected sites. Details are given in Supplemental section S2.8.

Analysis of ENCODE regions

Having established with simulated data that phyloP performs fairly well for a range of realistic parameter settings, we next applied the method to real biological data. Here we again made use of the alignments for the 44 ENCODE regions (Margulies et al. 2007) (see Methods), which, at present, constitute the largest published comparative genomic data set for mammals.

First, we analyzed distributions of phyloP scores for various classes of sites, focusing on the LRT method and three tests—the all-branch test and subtree tests for the primate and the glires clades. These scores were produced by running phyloP in "CONACC" mode, which produces positive scores for predicted conservation and negative scores for predicted acceleration (see Methods). In the all-branch case, we computed single-nucleotide scores, but for the subtree tests, which have less power, we computed scores in a 10-bp sliding window. We considered various annotation types, including known protein-coding genes and noncoding RNAs (ncRNAs), putative transcriptional fragments of unknown function (Un.TxFrags), sequence-specific regulatory factor binding regions (RFBRSeqsp), and predicted transcription factor binding sites within ChIP/chip-identified regions (TFBS). For the protein-coding genes, we separately considered coding regions (CDSs; positions CDS1, CDS2, and CDS3), 5′and 3′ untranslated regions (UTRs), 5′and 3′ flanking regions (200 bp upstream of the 5′ UTR and downstream from the 3′ UTR, respectively), and introns. For comparison, we also considered scores in putatively conserved phastCons elements and putatively neutral ancestral repeats (ARs).

The distributions of all-branch scores show clear differences by annotation class, generally in expected ways (Fig. 4A; Supplemental Fig. S6). For example, CDS1 and CDS2 sites are strongly enriched for high conservation scores, with CDS2 sites being slightly more conserved than CDS1 sites, while CDS3 sites and sites in 5′ UTRs, 5′ flanks, 3′ UTRs, and 3′ flanks (in decreasing order) show clear, but more modest, enrichments for high scores. Non-protein-coding functional elements (ncRNAs and TFBSs) show levels of conservation intermediate between CDS1/CDS2 sites and UTRs, and the bulk distributions for intronic, intergenic, and AR sites are all quite similar. Un.TxFrag sites show no significant enrichment for constraint, as observed previously (The ENCODE Project Consortium 2007). Interestingly, CDS3 sites are the only class to show an excess of fast-evolving sites, most likely as a result of an enrichment for hypermutable CpGs in coding regions (Eöry et al. 2009). Basewise phyloP scores can also be summarized at individual positions within functional elements, by averaging across elements of the same type. Such "conservation profiles" for protein-coding genes and transcription factor binding sites highlight several known features of these elements (Fig. 4B; Supplemental Fig. S5), providing further validation that phyloP scores capture biologically meaningful signals in the data.

By decomposing the score distributions for each annotation class into "neutral" and "selected" components, it is possible to obtain lower-bound estimates for the fractions of sites that have experienced long-term selective constraint (Methods). By this approach, we estimate that 5.3% of all sites in the ENCODE regions show evidence of conservation, in good agreement with previous findings (Chiaromonte et al. 2003; Lunter et al. 2006; The ENCODE Project Consortium 2007). Furthermore, we estimate that about two-thirds of CDS1 and CDS2 sites have evolved under constraint, as well as about one-third of ncRNA sites, one-fourth of CDS3 sites, one-fifth of TFBS sites, and 12%–16% of sites in UTRs and 5′ flanking regions (Fig. 4C). Not surprisingly, the estimated fraction of constrained sites is highest for phastCons elements (87.4%). Consistent with previous findings (Asthana et al. 2007), we estimate that a nonnegligible fraction (1.3%) of bases outside of phastCons elements or annotated CDSs, UTRs, and ncRNAs are conserved, suggesting that many unannotated functional sites may remain, even within the ENCODE regions. In general, these estimated fractions are remarkably concordant with estimates from a recent genome-wide pairwise analysis of hominid and

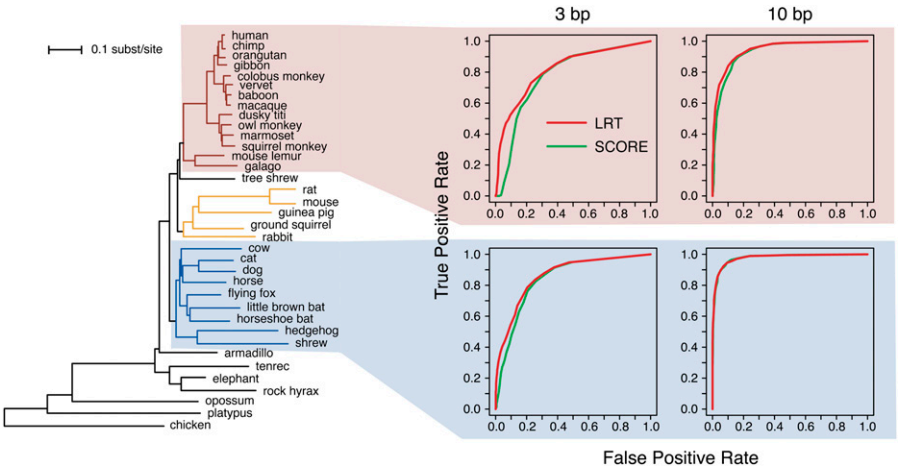

Figure 3. Subtree ROC curves. (*Left*) Phylogenetic tree used in this study, with branch lengths drawn in proportion to the values estimated from 4D sites. Three subtrees are highlighted: (maroon) primates, (gold) glires, and (blue) laurasiatherians. (*Right*) ROC curves for the LRT (red) and SCORE (green) subtree tests as applied to 3-bp and 10-bp elements under clade-specific selection in the primates (*top*) and laurasiatherians (*bottom*). (The SPH method did not perform as well, and the subtree test is not supported with the GERP method.) Results are shown for the case in which $\rho = 1.0$ and $\lambda = 0.3$, meaning that the clade of interest is evolving at approximately one-third the neutral rate, while the rest of the tree is neutrally evolving.

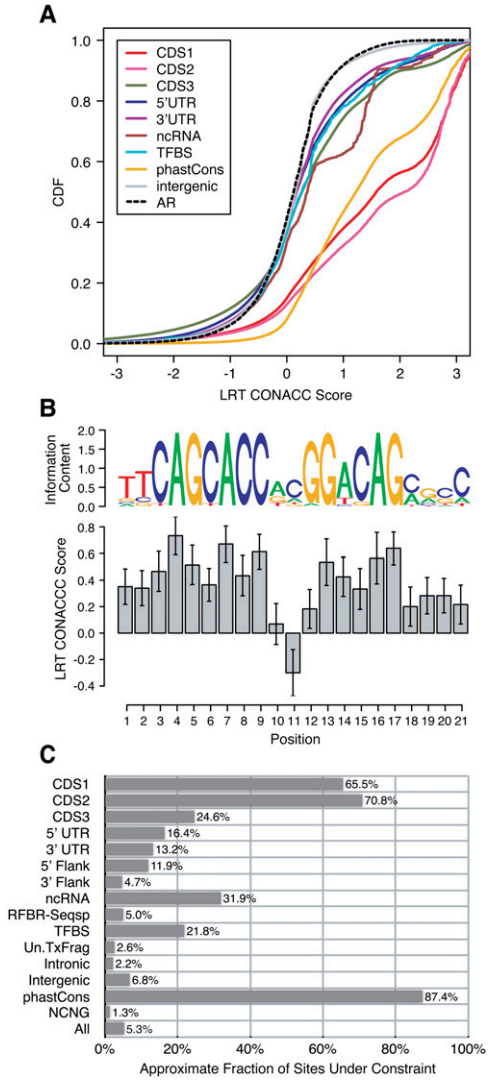

Figure 4. Distributions of all-branch scores. (*A*) Cumulative distribution functions (CDFs) for phyloP scores in sites of different annotation classes, based on the LRT method and 36-species multiple alignments for the ENCODE regions. Positive scores indicate conservation, and negative scores indicate acceleration (CONACC mode) (see Methods). Curves are shown for first, second, and third codon positions (CDS1, CDS2, CDS3), 5′ and 3′ UTRs, noncoding RNAs (ncRNAs), predicted transcription factor binding sites (TFBS), conserved elements identified by phastCons, intergenic sites, and ancestral repeats (AR). (See Supplemental Fig. S6 for additional annotation classes.) (*B*) Average conservation scores as a function of genomic position within 52 predicted NRSF binding sites in the ENCODE regions. Binding sites were predicted at ChIP/chip peaks using the motif from TRANSFAC (FDR = 20%) (Supplemental section S2.9). A sequence logo representation of the motif is shown for comparison. Notice the general correlation between information content and cross-species conservation across the positions of the motif (see Moses et al. 2003). (*C*) Estimated fractions of sites under selection for each annotation class. Classes include those from *A*, plus 5′ and 3′ flanking regions of genes, sequence-specific regulatory binding regions (RFBR-Seqsp), putative transcriptional fragments of unknown function (Un.TxFrags), intronic sites, and nonconserved nongenic (NCNG) sites. These are estimates of lower bounds computed by a simple mixture-decomposition method (see Methods) and should be considered approximate. All classes show a highly significant enrichment for conserved sites relative to the AR distribution by a one-sided Mann-Whitney U test ($P \approx 0$) except the 3′ flank, intronic, Un.TxFrags, and NCNG categories (all $P \approx 1$).

murid genomes, based on quite different methods (Eöry et al. 2009).

Unlike the distributions of all-branch scores, the distributions of subtree scores for the primate clade are quite similar across annotation types, suggesting that the null hypothesis of equal evolutionary rates between this clade and the remainder of the tree holds fairly well (Fig. 5A). The glires clade, however, shows much more pronounced differences in subtree score distributions (Fig. 5B), suggesting an increased tendency for clade-specific selection. In particular, the CDS, phastCons, 5′ UTR, and 5′ flank classes (in decreasing order) show clear shifts toward higher scores in the glires. This trend holds if a series of strict filters is applied to the alignments, indicating that it is not an artifact of missing data or nonorthologous alignments (Supplemental section S2.10). The observed difference between the primates and glires distributions also does not appear to result from differences in power in these two clades (Supplemental section S3.8). The shift toward larger glires-subtree scores in functional elements appears to be driven by increases in negative selection rather than decreases in positive selection, because it is strongest for sites that are evolving at or beneath the neutral rate outside the glires subtree (Supplemental Fig. S8). A possible explanation for this shift would be increased strength of selection owing to larger effective population sizes in the glires (Keightley et al. 2005; see also Kosiol et al. 2008).

Finally, we used clade-specific phyloP scores to test for accelerated evolution in conserved (and hence likely functional) elements within the ENCODE regions, again focusing on the primate and glires clades. We used phastCons and strict alignment-quality filters to identify a set of 16,449 conserved regions for primate analysis and 19,498 for glires analysis (see Methods). These elements were scored for clade-specific acceleration in the primates or glires groups relative to the rest of the tree using the subtree LRT. At FDR ≈ 5%, we identified 216 primate accelerated regions (PARs) and 3529 glires accelerated regions (GARs). The two lists of accelerated regions are generally similar in terms of genomic locations, but a slightly larger proportion of PARs fall in coding sequences of GENCODE genes (7.4% vs. 4.5% of GARs). Known and predicted RNA genes overlap nine PARs and 83 GARs. The most significantly accelerated PARs and GARs are described in Supplemental Tables S15–S16 and Supplemental Figure S9. Interestingly, the glires clade shows a pronounced excess of accelerated regions over a large range of nominal *P*-value thresholds, again suggesting the possibility of increased selection in this clade. However, differences in the starting set of elements, in the power of the subtree tests, and asymmetries in the human-referenced alignments may also contribute to this observation.

Conservation tracks in UCSC Genome Browser

PhyloP scores for genome-wide multiple alignments of 44 vertebrate species (including 32 mammals) have been incorporated into a new "Conservation Track" in the UCSC Genome Browser (http://genome.ucsc.edu, hg18 assembly). This track shows phyloP scores for individual sites alongside conservation scores and conserved elements produced by phastCons, for all species, the eutherian mammals only, and the primates only (Fig. 6). The phyloP and phastCons scores provide complementary measures of nonneutral substitution rates, with phyloP capturing both conservation and acceleration and operating independently at each site, and phastCons measuring conservation only in a way that considers "runs" of conserved sites (through the use of an HMM). A separate track shows phyloP subtree scores for the primate and glires clades (data not shown).

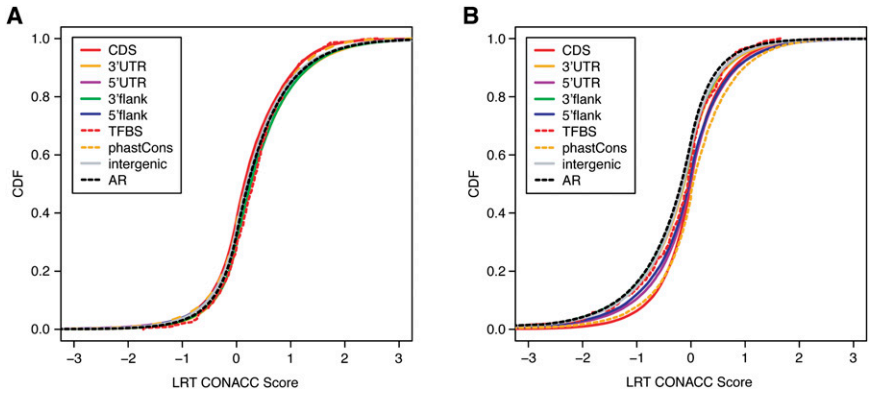

Figure 5. Distributions of subtree scores for the primate and glires clades. Cumulative distribution functions (CDFs) of scores for selected annotation classes as computed by the subtree test for the primate (*A*) and glires (*B*) clades. As in previous figures, CONACC scores computed by the LRT method are shown, but in this case, scores are computed in a 10-bp sliding window. In both figures most distributions are significantly different from the AR distribution by a two-sided Mann-Whitney *U* test even when the curves appear very similar, because the data sets are generally quite large (exceptions are phastCons and TFBS in *A* and 5′ flank and TFBS in *B*).

Discussion

Methods for detecting signatures of selection from rates and patterns of substitution have a long history in the field of molecular evolution (e.g., Kimura 1977; Miyata et al. 1980). In recent years, methods of this kind have also become important tools in applied genomics because of their usefulness in detecting and characterizing functional elements. As more and more genomic sequence data become available, it should be possible to take this line of research to its logical conclusion and characterize selection pressures at very high resolution—perhaps even at the level of individual nucleotides. In this study, we examine the problem of detecting nonneutral substitution rates from aligned genomic sequences, focusing on what is possible with currently available data for the eutherian mammals. Our contributions consist of four main components: (1) a detailed comparison of four alternative approaches to this problem; (2) estimates of the absolute power of these methods; (3) an analysis of patterns of conservation/acceleration in the ENCODE regions; and (4) the release of a software tool, called phyloP, and an associated track in the UCSC Genome Browser, which we expect to be useful resources for the comparative genomics community.

The four methods considered here show surprisingly little difference in statistical power, given their quite different theoretical foundations. The LRT and SCORE methods use test statistics based on the full likelihood function and might be expected to make better use of the pattern of substitution—such as the fact that transitions occur at much higher rates than transversions—than the GERP and SPH methods, which simply work with estimators of the number of substitutions. In addition, the GERP method considers only a point estimate of the number of substitutions, ignoring its variance, and the SPH method makes use of a highly discrete test statistic (an integral number of substitutions), which should (and does to an extent) limit its power, especially for short elements. However, in practice, these methodological differences seem to be relatively unimportant in distinguishing between neutral and selected sites. Instead, it seems that information about substitution rates can be accessed by a variety of means, provided good use is made of the phylogeny and substitution model. This argument may extend to methods that make only partial use of a phylogeny and/or a continuous-time Markov substitution model, such as binCons, the parsimony-based *P*-value method (Margulies et al. 2003), and SCONE (Asthana et al. 2007). Indeed, SCONE and GERP have been found to have similar performance in experiments similar to the ones performed here (Asthana et al. 2007).

Regardless of which method is used, the ability to detect constraint or acceleration depends in a predictable way on the amount of "signal" in the data. Power increases with the magnitude of the departure from the neutral model (as measured by ρ or λ), the length of the element, and the number of species affected. These results are qualitatively consistent with the predictions of theoretical models (Eddy 2005; McAuliffe et al. 2005; Stone et al. 2005) and with previous empirical studies (Cooper et al. 2003; Margulies et al. 2003). However, they are supported in this case by a somewhat more extensive set of experiments, considering both parametric and nonparametric methods, conservation and acceleration, all-branch and clade-specific selection, and richer phylogenetic models. Our results suggest that, while it is premature to claim single-nucleotide resolution in the detection of nonneutral substitution rates, elements 1–3 bp in length can be detected with reasonable power—for example, 30%–75% TPRs at 5% FDRs. Similarly, moderately strong clade-specific selection can be detected at the level of 10-bp elements. Even in scenarios in which power is weak, useful information can be obtained by pooling together similar sites from across the genome (as in Fig. 4). Of course, power will steadily improve as additional genomes are sequenced.

The similarity in power of the methods considered here could be taken to suggest that little is to be gained by further methodological work on the problem of detecting selection from aligned sequences. However, these methods are all based completely on substitution rates and ignore other sources of information about natural selection, such as patterns of substitution (Moses et al. 2004; Pedersen et al. 2006) or rates and patterns of insertion and deletion (Kellis et al. 2003; Siepel and Haussler 2004a; Lunter et al. 2006). A recently introduced method, called SiPhy, attempts to exploit the pattern of substitution by using an LRT similar to phyloP's all-branch test, except that it treats the equilibrium nucleotide frequencies as free parameters to be estimated at each site for the alternative model (together with a branch-length scaling parameter) (Garber et al. 2009). In principle, this approach should increase power for subtle selective pressures that influence base preferences but have only a mild effect on the overall substitution rate. However, there are risks associated with the use of a richer alternative model. Because SiPhy assumes constant equilibrium frequencies for its null model, it essentially performs a compound test of both rate and pattern and will therefore tend to predict more elements (and have increased TPRs and FPRs) in regions of the genome with unusual base composition. Compared with rate-based methods, SiPhy may also be more influenced by phenomena more directly associated with mutation and repair than with natural selection, such as transcription-coupled repair (Green et al. 2003), biased gene conversion (Marais 2003; Dreszer et al. 2007),

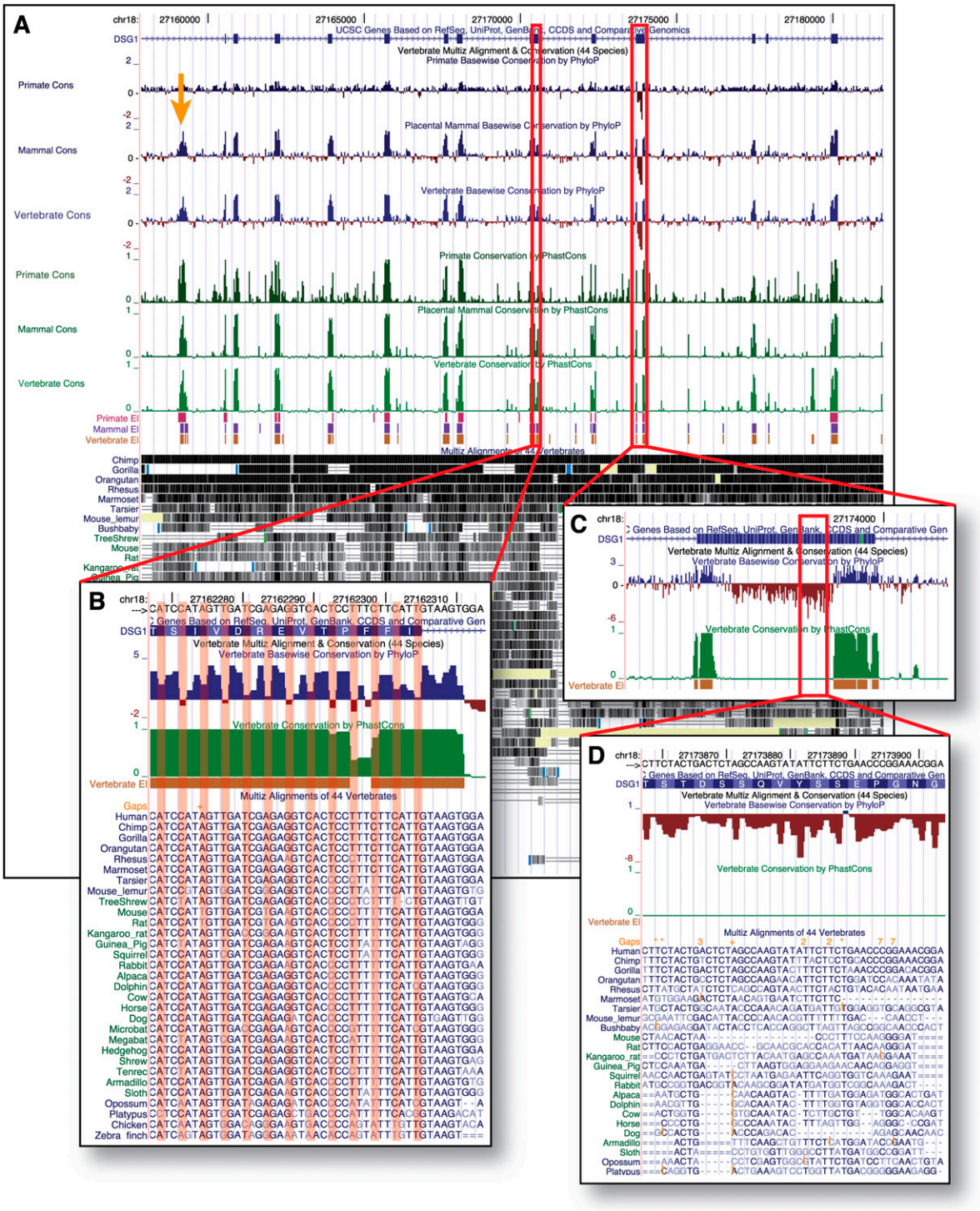

Figure 6. Conservation track in UCSC Genome Browser. A portion of the desmoglein 1 (*DSG1*) gene on human chromosome 18 shown with the new Conservation track, including a 44-way vertebrate alignment and nine conservation subtracks. The subtracks display phyloP scores (in blue and red), phastCons scores (green), and phastCons-predicted conserved elements (pink, purple, and mustard) for all species, the 32 placental mammals, and the nine primates (*bottom* to *top* within each group). (*A*) The phyloP and phastCons scores are broadly similar when the display is zoomed out, with scores near zero for most noncoding regions but elevated in exons (thick blue bars at *top*) as well as in conserved noncoding elements (orange arrow). (*B*) At finer resolution, however, phyloP reveals significantly more variation from base to base than does the hidden Markov model–based phastCons. In this coding exon, codon position effects are clearly evident from phyloP but not from phastCons. (*C*,*D*) The phyloP tracks also indicate accelerated evolution (with negative scores, shown in red), while phastCons measures conservation only. Here an exon with a striking fast-evolving segment is shown. Interestingly, cDNA data from other mammals suggest that this exon derives from a fusion of two ancestral exons, with the fast-evolving segment corresponding to the ancestral intron.

and methylation of cytosines (Ehrlich and Wang 1981). It is likely that factors such as these are at least partly responsible for the nearly twofold increase in the number of evolutionarily "constrained" sites detected by this method in the ENCODE regions (Garber et al. 2009). Nevertheless, pattern-based methods for detecting selection do have the potential to improve power and are worthy of further investigation.

The rate-based methods considered here also have several limitations worth noting. First, the phylogenetic models on which they are based, while rich in some respects, are highly simplified in others. For example, these models ignore regional variation (Wolfe et al. 1989; Mouse Genome Sequencing Consortium 2002) and context dependencies (Hwang and Green 2004; Siepel and Haussler 2004b) in neutral substitution rates, variation in G+C content (Hardison et al. 2003), transcription-associated mutational asymmetry (Green et al. 2003), and differences between clades in selection on 4D sites (Eöry et al. 2009). Second, the tests (and our parametric experiments) assume constant levels of directional selection, producing sustained increases or decreases in evolutionary rate over long periods of evolutionary time. While these assumptions appear to be reasonable for some types of functional elements (such as conserved protein-coding genes), they undoubtedly do not hold in many cases. Finally, these methods all depend on accurate alignments of mammalian genomes. Genome-wide multiple alignment remains a challenging, unsolved problem, and alignment error can have a substantial influence on predictions of constrained elements (Margulies et al. 2007). New methods offer some hope that it may be possible to address problems of functional element identification while integrating or sampling over alignments, thereby mitigating the effects of alignment error from a single fixed alignment (Satija et al. 2009). However, at present, these methods require orders of magnitude more computational time than methods that assumed fixed alignments and are not feasible for use on a genome-wide scale. Still, it may be possible to use heuristic methods to substantially improve the speed of such methods (Bradley et al. 2009; Paten et al. 2009), or to quantify alignment uncertainty and then use this information in downstream functional element identification (Lunter et al. 2008). In short, many opportunities remain for improving the biological realism, statistical power, and robustness of methods for identifying functional elements from comparative sequence data.

Methods

Statistical tests

The statistical tests considered in this study can all be placed in the following general framework. Let ψ_N be a neutral phylogenetic model, consisting of a tree topology, a vector of branch lengths β_N, a set of equilibrium nucleotide frequencies, and a substitution rate matrix. ψ_N can be estimated from large quantities of genomic data and is assumed to be known. Let $\psi(\theta)$, for a vector of non-negative branch-length scaling parameters θ (of the same dimension as β_N), be a scaled phylogenetic model identical to ψ_N except that it has branch lengths $\beta_\theta = \theta \cdot \beta_N$, a Hadamard (pointwise) product of θ and β_N. We consider two parameterizations of θ : (1) the uniform scaling vector, $\theta(\rho) = \rho\mathbf{1}$, which scales all branches by a single (non-negative) scalar parameter ρ; and (2) the subtree scaling vector $\theta(\rho, \lambda; u)$, which scales all branches by ρ and additionally scales all branches in the subtree beneath a specified node u by a second non-negative scalar parameter λ, that is, $\theta(\rho, \lambda; u) = (\theta^{(i)} : i = 1, \ldots, |\beta_N|)$ such that $\theta^{(i)} = \rho\lambda$ if branch i is in the subtree beneath u, or $\theta^{(i)} = \rho$ otherwise. Notice that these parameterizations are nested, with $\theta(\rho, \lambda = 1; u) = \theta(\rho)$ for all u.

For a given alignment $\mathbf{X}$ of length L, assumed to have independent columns all distributed according to some $\psi(\theta)$, the two-sided all-branch test compares a null hypothesis $H_0 : \theta = \mathbf{1}$ with an alternative hypothesis $H_1 : \theta = \theta(\rho), \rho \geq 0, \rho \neq 1$. The two-sided subtree test, for a given node u (and associated subtree), compares a null hypothesis $H_0 : \theta = \theta(\rho), \rho \geq 0$ with an alternative hypothesis $H_1 : \theta = \theta(\rho, \lambda; u)$, $\rho \geq 0, \lambda \geq 0, \lambda \neq 1$. Thus, the all-branch test can be thought of as a test of $\rho = 1$ and the subtree test as a test of $\lambda = 1$ (with $\rho \geq 0$ as a free parameter). One-sided tests can similarly be defined for conservation (full tree: $\rho < 1$, subtree: $\lambda < 1$) or acceleration (full tree: $\rho > 1$, subtree: $\lambda > 1$).

Likelihood ratio tests

The LRTs are based on the test statistic $T = 2\left[L(\hat{\theta}_1) - L(\hat{\theta}_0)\right]$, where $\hat{\theta}_0$ and $\hat{\theta}_1$ are the maximum likelihood estimates of the parameter vectors associated with the null and alternative hypotheses, respectively. These estimates are obtained numerically, as described below. For our two-sided tests, the regularity conditions required for T to have an asymptotic χ^2_1 null distribution hold. For our one-sided tests, however, the null hypothesis is at the boundary of the parameter space under the alternative hypothesis, which causes the asymptotic distribution to become a 50:50 mixture of a χ^2_1 distribution and a point mass at zero (Self and Liang 1987). With both one-sided and two-sided tests, the asymptotic distributions are used to compute approximate P-values. Additional details are in Supplemental section S2.2.

The LRT-based scores used in the analysis of the ENCODE regions and in the conservation tracks are computed as $-\log_{10} P$, where P is a two-sided P-value. In order to distinguish conservation and acceleration scores, the scores are negated if the estimated ρ (or λ) suggest faster-than-neutral evolution (Supplemental section S2.2). This scoring system is produced by running phyloP with the option --mode CONACC.

Score tests

The score tests are based on test statistics of the form $S = U^T\left(\hat{\boldsymbol{\theta}}_0\right) I^{-1}\left(\hat{\boldsymbol{\theta}}_0\right) U\left(\hat{\boldsymbol{\theta}}_0\right)$, where U is the score function (the vector of partial derivatives of the log likelihood) and I is the Fisher information matrix. Both U and I are defined with respect to $\boldsymbol{\theta}_1$ but evaluated at the maximum-likelihood estimate under the null hypothesis, $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_0$. S is known, like T above, to have an asymptotic χ^2_j null distribution, where j is the difference in the number of free parameters of $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ ($j = 1$ here), and this asymptotic distribution is used by phyloP in computing P-values. Notably, the score test has the same local power as the LRT, being a most powerful test for small deviations from the null hypothesis (in this case, weak conservation or acceleration). However, the score test requires fitting only the null model to the data, instead of both the null and alternative models, which results in a substantial savings in computation. Indeed, with our all-branch test, no estimation is required because the null model has no free parameters. The subtree case requires estimation of a single scale factor ρ to fit the null model to the data. For both types of tests, computation of the Fisher information matrix appears to be intractable, so we approximate it by Monte Carlo sampling. Additional details are in Supplemental section S2.3.

SPH tests

The all-branch SPH test is based on a test statistic (denoted n) equal to the number of substitutions that have occurred along the branches of a phylogeny in an alignment $\mathbf{X}$, assuming a given neutral model ψ (Siepel et al. 2006). The exact null distribution of this statistic, $P(n|\psi)$, can be approximated arbitrarily closely by a uniformization procedure and a recursive dynamic programming

algorithm that resembles Felsenstein's "pruning" algorithm. A closely related algorithm can be used to compute an estimate of n for an observed alignment X, and this estimate can be compared to the null distribution to compute a P-value. In the case of subtree tests, a similar procedure is used, but the joint distribution for the number of substitutions in a subtree of interest and the rest of the tree is considered (Siepel et al. 2006; Supplemental section S2.4). For various reasons, the P-values computed by this procedure tend to be somewhat conservative (Supplemental Fig. S4).

GERP–like tests

GERP makes use of a statistic called "rejected substitutions" (*RS*), which is defined as the number of substitutions expected under a neutral model minus the number "observed" (estimated) for a particular alignment $\mathbf{X}$ (Cooper et al. 2005)—that is, the expected number of mutations that would have been fixed under neutrality but instead were "rejected" by purifying selection. For a given neutral model ψ_N and alignment X, GERP estimates a scaling parameter ρ for ψ_N by maximum likelihood (as in the LRT and SCORE tests), and estimates RS as $RS = T - \hat{\rho}T = T(1 - \hat{\rho})$, where T is the total branch length of the neutral model and $\hat{\rho}$ is the estimated scale factor. While the other test statistics considered are conservative in the presence of missing data, RS can be quite sensitive to it, because a branch of length t for which no data are available will still contribute $\hat{\rho}t$ to the overall value of the test statistic. Therefore, a separate value of T is computed for each alignment $\mathbf{X}$, by considering just the branches of the phylogeny for which aligned nucleotides are available. In addition, RS is set to zero if bases are available for fewer than three species. The GERP program (http://mendel.stanford.edu/sidowlab/downloads/gerp) assumes the use of the HKY85 substitution model and combines the steps of estimating the neutral substitution model and computing the desired RS values. To facilitate comparisons with the other tests, we reimplemented the core functionality of GERP within phyloP, treating it as analogous to the other all-branch tests in all respects. (It is not supported for subtree tests.) Like the GERP program, phyloP simply outputs the raw RS values and allows P-values to be computed separately in post-processing, if desired. To generate the ROC curves, we applied varying thresholds to RS for one-sided tests of conservation, to $-RS$ for one-sided tests of acceleration, and to $|RS|$ for two-sided tests. A comparison of phyloP's GERP mode with the latest version of GERP (version 2.1b) showed that the two programs were very similar in performance (Supplemental section S3.1).

Parameter estimation

Three of the four tests above depend on numerical estimation of the scale factors ρ and/or λ by maximum likelihood for each alignment segment $\mathbf{X}$. This is accomplished using the Newton-Raphson method for one-dimensional optimization of ρ and the BFGS method for two-dimensional optimization of (ρ, λ). In practice, this optimization is the rate-limiting step in most analyses, so various techniques were used to improve its efficiency (Supplemental section S2.1). It is worth noting that these estimated scale factors may be of interest for other reasons. For example, under a model in which all mutations are either deleterious or neutral and deleterious mutations are rapidly eliminated by natural selection, $\hat{\rho}$ is an estimator of the fraction of mutations that are neutral (in a given element $\mathbf{X}$), and $(1 - \hat{\rho})$ is an estimator of the fraction that are deleterious (e.g., Kondrashov and Crow 1993).

Multiple testing

It should be emphasized that phyloP computes all P-values independently, disregarding correlations between tests. Adjustments for multiple hypothesis tests are needed when jointly interpreting the reported P-values for a collection of sites or elements.

Alignments and neutral model

Alignments of the 44 ENCODE regions were produced with the TBA program (Blanchette et al. 2004), as described by Margulies et al. (2007), but using an expanded set of sequences (June 2008 freeze; 33 eutherian mammals vs. 21 previously analyzed). The neutral model was estimated from 4D sites in these alignments, using the phyloFit program in PHAST. After estimation, the model was adjusted to maintain the estimated nucleotide exchangeabilities but ensure a stationary distribution equal to the genome-wide average (Supplemental section S2.6). The same neutral model was used for the simulation experiments and the analysis of real data.

Simulations

Parametric simulations were based on alignment columns generated by forward sampling from phylogenetic models, using the program phyloBoot in PHAST. Neutral alignment columns (for evaluating false-positive rates) were generated from the estimated neutral model, and selected alignment columns (for evaluating true-positive rates) were generated from versions of this model in which all branches were scaled by a parameter ρ, or the branches in a subtree of interest were scaled by a parameter λ. For both ρ and λ, the following set of scale factors was considered: $\{q/10, 10/q : q \in \{1, 3, 5, 7, 9\}\} = \{0.1, 0.3, 0.5, 0.7, 0.9, 1.11, 1.43, 2.00, 3.33, 10.00\}$. The simulated data did not contain alignment gaps or missing data. In some cases, data were generated with rate variation across sites (Supplemental section S2.7.1) or by adding "noise" to the scale factors, so they did not exactly match the assumptions of the subtree test (Supplemental section S2.7.2). Nonparametric experiments were based on alignment columns drawn with replacement (also using phyloBoot) from sets of columns from ancestral repeats (neutral) and second codon positions (selected).

Annotations for ENCODE regions

Protein-coding gene annotations were based on 408 nonoverlapping genes from the GENCODE set (Harrow et al. 2006). ncRNAs consisted of eight well-characterized structural and regulatory RNAs from the ENCODE regions (*SNORA70* [also known as *U70*], *SNORA36A* [also known as *ACA36*], *SNORA56* [also known as *ACA56*], *MIR192*, *MIR194-2*, *MIR196B*, *MIR483*, and *H19*). RFBR-Seqsp and Un.TxFrag annotations were obtained from The ENCODE Project Consortium (2007). Specific TFBS sites were predicted by our own methods, using three transcription factors for which ChIP-chip data and binding motifs were available (22 MYC [also known as c-Myc], 52 REST [also known as NRSF], and 21 STAT1 sites) (Supplemental section S2.9). For ARs, we extracted RepeatMasker (http://www.repeatmasker.org) annotations corresponding to repeat families and classes previously identified as ancestral to the eutherian mammals (Mouse Genome Sequencing Consortium 2002). In all cases, annotations were defined in human (hg18) genome coordinates, and then mapped to the multiple alignments.

Estimation of fractions of sites under selection

Fractions of sites under selection were estimated by a method similar to that used by Chiaromonte et al. (2003), but based on empirical cumulative distribution functions (CDFs) instead of estimated density functions. Specifically, the distribution of phyloP scores for each annotation class a was assumed to be a mixture of

neutral and selected components, $F_a(s) = (1 - \pi_a)G(s) + \pi_a H_a(s)$, where F_a is the CDF for all sites in class a (a function of scores s), G is the CDF for the sites that are neutrally evolving, H_a is the CDF for the sites under selection, and π_a ($0 \leq \pi_a \leq 1$) is the fraction of sites in class a that are under selection. Owing to nonnegativity of H_a, $F_a(s) \geq (1 - \pi_a)G(s)$ for all s, so a lower bound for π_a is given by

$$\hat{\pi}_a = 1 - \min_s \frac{F_a(s)}{G(s)}.$$

This lower bound was estimated by substituting the empirical CDFs for ARs and for all sites of each annotation class a for G and H_a, respectively, excluding the smallest scores (<-1.5) so that the estimated bounds were not determined by the extreme left tails of the empirical CDFs (which reflect sparse data). For various reasons, these estimates should be considered crude—for example, they may be influenced by differences between the ARs and the various annotation classes in base composition, substitution patterns, or amounts of missing data. Nevertheless, they agree well with estimates obtained by quite different methods (Eöry et al. 2009).

PAR/GAR analysis

A set of conserved elements was identified for each of the clade-specific acceleration tests (primates and glires), by running phastCons on alignments in which the species in the clade of interest had been removed, then applying several filters to eliminate potential alignment and assembly errors (Supplemental section S2.11). Each set of filtered elements was scored with the one-sided acceleration LRT for the appropriate subtree. Nominal P-values were adjusted for multiple comparisons using the FDR-controlling method of Benjamini and Hochberg (1995).

Acknowledgments

We thank Elliott Margulies for providing the multiple sequence alignments and estimated neutral model for the ENCODE regions; Jim Booth for suggesting the score test as an alternative to the likelihood ratio test; Hiram Clawson for setting up the new Conservation tracks in the UCSC Genome Browser; David Haussler and Jim Kent for feedback and support in track development; and Andre Martins for helping with the analysis of transcription factor binding sites. This work was supported by the National Institute of General Medical Sciences (grant GM82901) and by early career awards from the Alfred P. Sloan Foundation, the David and Lucile Packard Foundation, and the National Science Foundation (grant DBI-0644111).

References

Asthana S, Roytberg M, Stamatoyannopoulos JA, Sunyaev S. 2007. Analysis of sequence conservation at nucleotide resolution. *PLoS Comput Biol* **3:** e254. doi: 10.1371/journal.pcbi.0030254.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* **57:** 289–300.

Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, Beazley C, Miller W, Hurles ME, Dermitzakis ET. 2007. Fast-evolving noncoding sequences in the human genome. *Genome Biol* **8:** R118. doi: 10.1186/gb-2007-8-6-r118.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14:** 708–715.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299:** 1391–1394.

Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009. Fast statistical alignment. *PLoS Comput Biol* **5:** e1000392. doi: 10.1371/journal.pcbi.1000392.

Casella G, Berger RL. 2002. *Statistical inference*. Duxbury, Pacific Grove, CA.

Chiaromonte F, Weber RJ, Roskin KM, Diekhans M, Kent WJ, Haussler D. 2003. The share of human genomic DNA under selection estimated from human–mouse genomic alignments. *Cold Spring Harb Symp Quant Biol* **68:** 245–254.

Cooper GM, Brudno M, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. 2003. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res* **13:** 813–820.

Cooper GM, Brudno M, Stone EA, Dubchak I, Batzoglou S, Sidow A. 2004. Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res* **14:** 539–548.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15:** 901–913.

Dreszer TR, Wall GD, Haussler D, Pollard KS. 2007. Biased clustered substitutions in the human genome: The footprints of male-driven biased gene conversion. *Genome Res* **17:** 1420–1430.

Eddy SR. 2005. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol* **3:** e10. doi: 10.1371/journal.pbio.0030010.

Ehrlich M, Wang RY. 1981. 5-Methylcytosine in eukaryotic DNA. *Science* **212:** 1350–1357.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Eöry L, Halligan DL, Keightley PD. 2009. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Mol Biol Evol* (in press). doi: 10.1093/molbev/msp219.

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25:** 54–62.

Green P, Ewing B, Miller W, Thomas PJ, NISC Comparative Sequencing Program, Green ED. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* **33:** 514–517.

Gross SS, Do CB, Sirota M, Batzoglou S. 2007. CONTRAST: A discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol* **8:** R269. doi: 10.1186/gb-2007-8-12-r269.

Guigó R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C, et al. 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc Natl Acad Sci* **100:** 1140–1145.

Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* **13:** 13–26.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen C-K, Chrast J, Lagarde J, Gilbert JGR, Storey R, Swarbreck D, et al. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol* **7:** S4. doi: 10.1186/gb-2006-7-s1-s4.

Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet* **39:** 1140–1144.

Huelsenbeck J, Rannala B. 1997. Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science* **276:** 227–232.

Hwang D, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci* **101:** 13994–14001.

Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol* **3:** e42. doi: 10.1371/journal.pbio.0030042.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.

Kim SY, Pritchard JK. 2007. Adaptive evolution of conserved noncoding elements in mammals. *PLoS Genet* **3:** 1572–1586.

Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267:** 275–276.

Kondrashov AS, Crow JF. 1993. A molecular approach to estimating the human deleterious mutation rate. *Hum Mutat* **2:** 229–234.

Kosiol C, Vinar T, da Fonseca R, Hubisz M, Bustamante C, Nielsen R, Siepel A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet* **4:** e1000144. doi: 10.1371/journal.pgen.1000144.

Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, et al. 2009. The UCSC Genome Browser Database: Update 2009. *Nucleic Acids Res* **37:** D755–D761.

Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2:** e5. doi: 10.1371/journal.pcbi.0020005.

Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J. 2008. Uncertainty in homology inferences: Assessing and improving genomic sequence alignment. *Genome Res* **18:** 298–309.

Marais G. 2003. Biased gene conversion: Implications for genome and sex evolution. *Trends Genet* **19:** 330–338.
Margulies EH, Blanchette M. NISC Comparative Sequencing Program, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res* **13:** 2507–2518.
Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, Keefe D, Schwartz AS, Hou M, et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res* **17:** 760–774.
McAuliffe JD, Jordan MI, Pachter L. 2005. Subtree power analysis and species selection for comparative genomics. *Proc Natl Acad Sci* **102:** 7900–7905.
Miller W, Rosenbloom K, Hardison R, Hou M, Taylor J, Raney B, Burhans R, King D, Baertsch R, Blankenberg D, et al. 2007. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* **17:** 1797–1808.
Miyata T, Yasunaga T, Nishida T. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc Natl Acad Sci* **77:** 7328–7332.
Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB. 2003. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol* **3:** 19. doi: 10.1186/1471-2148-3-19.
Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB. 2004. MONKEY: Identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol* **5:** R98. doi: 10.1186/gb-2004-5-12-r98.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302:** 413. doi: 10.1126/science.1088328.
Paten B, Herrero J, Beal K, Birney E. 2009. Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics* **25:** 295–301.
Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2:** e33. doi: 10.1371/journal.pcbi.0020033.
Pollard K, Salama S, King B, Kern A, Dreszer T, Katzman S, Siepel A, Pedersen J, Bejerano G, Baertsch R, et al. 2006a. Forces shaping the fastest evolving regions in the human genome. *PLoS Genet* **2:** e168. doi: 10.1371/journal.pgen.0020168.
Pollard KS, Salama SR, Lambert N, Lambot M-A, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, et al. 2006b. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443:** 167–172.
Prabhakar S, Noonan JP, Paabo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314:** 786. doi: 10.1126/science.1130738.
Rao CR. 1948. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc Camb Philol Soc* **44:** 50–57.
Rao CR. 2005. Score test: Historical review and recent developments. In *Advances in ranking and selection, multiple comparisons, and reliability* (eds. N Balakrishnan et al.), pp. 3–20. Birkhäuser, Boston, MA.
Satija R, Novak A, Miklos I, Lyngso R, Hein J. 2009. BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC. *BMC Evol Biol* **9:** 217. doi: 10.1186/1471-2148-9-217.
Self S, Liang K. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc* **82:** 605–610.
Siepel A, Haussler D. 2004a. Computational identification of evolutionarily conserved exons. In *Proc. 8th Int'l Conf. on Research in Computational Molecular Biology,* pp. 177–186. ACM Press, New York.
Siepel A, Haussler D. 2004b. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* **21:** 468–488.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.
Siepel A, Pollard K, Haussler D. 2006. New methods for detecting lineage-specific selection. In *Proc. 10th Int'l Conf. on Research in Computational Molecular Biology,* pp. 190–205. Springer-Verlag, Berlin, Germany.
Siepel A, Diekhans M, Brejova B, Langton L, Stevens M, Comstock C, Davis C, Ewing B, Oommen S, Lau C, et al. 2007. Targeted discovery of novel human exons by comparative genomics. *Genome Res* **17:** 1763–1773.
Stone EA, Cooper GM, Sidow A. 2005. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu Rev Genomics Hum Genet* **6:** 143–164.
Wolfe KH, Sharp PM, Li W-H. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283–285.
Wong WSW, Nielsen R. 2004. Detecting selection in noncoding regions of nucleotide sequences. *Genetics* **167:** 949–958.
Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* **39:** 306–314.

Received June 26, 2009; accepted in revised form October 5, 2009.