DNA conservation methods

Journal Club

XJM, Jan 16, 2012

PhyloP:

Pollard et al., **Detection of nonneutral substitution rates on mammalian phylogenies,** *Genome Research* 20, no. 1 (2010): 110-121.

Methods Commonly used to detect conservation in DNA sequences

- **Protein-based methods**

  - Until recently, comparative genomic studies were mostly based on sequences of protein-coding genes.

  - high dN/dS ratios => positive selection (Nielsen and Yang, 1998; Yang and Nielsen, 2002; Clark et al., 2003; Forsberg and Christiansen, 2003; Guindon et al., 2004; Nielsen et al., 2005).

- **Percent identity scoring methods**

  - Early efforts: pairwise alignments, e.g. human vs. mouse (Schwartz et al., 2000; Mayor et al., 2000; Gottgens et al., 2001; Ovcharenko et al., 2004).

  - Later multiple sequence alignments: conservation individually for pairs of sequences (Schwartz et al., 2003) or jointly across all sequences by some form of averaging pairwise percent identity scores (Chapman et al., 2004; Shah et al., 2004; Ovcharenko et al., 2005).

- Heuristic phylogenetic methods
- Molecular evolutionary methods

Methods Commonly used to detect conservation in DNA sequences

Percent identity scoring methods

Method	Score Type	Statistical Significance	Resolution	Predicted Elements	Phylogenetic Tree Used?	Lineage-Specific	Indels	Refs
MultiPipMaker	% ID	X	1 bp	X	X	X	X	Schwartz et al. (2000, 2003)
PhyloVISTA	% ID	X	1 bp	X	X	user-specified node	X	Mayor et al. (2000); Shah et al. (2004)
SynPlot	% ID	X	user-specified window	X	X	X	X	Gottgens et al. (2001); Chapman et al. (2004)
Mulan	% ID	X	user-specified window	X	X	X	X	Ovcharenko et al. (2004, 2005)
binCons	binomial p-value	X	25 bp	maximal windows	phylogenetic averaging	X	X	Margulies et al. (2003)
FootPrinter	parsimony p-value	X	25 bp	maximal windows	✓ (HKY) (for p-values only)	X	X	Blanchette and Tompa (2002) Blanchette et al. (2002); Margulies et al. (2003)
(*no name*)	ω	false positive rate	25 bp	maximal windows	✓ (JC)	X	✓	Cooper et al. (2004)
SCONE	ω	false positive rate	1 bp	maximal windows	✓ (trinucleotide)	X	✓	Asthana et al. (2007)
GERP	rejected subs	false positive rate	1 bp	maximal windows	✓ (HKY)	X	✓	Cooper et al. (2005)
Phylogenetic shadowing	likelihood ratio	X	1 bp	X	✓ (HKY)	X	X	Boffelli et al. (2003)
phastCons	posterior probability	X	1 bp with Markov dependence	Viterbi path	✓ (REV)	X	X	Siepel and Haussler (2004a); Siepel et al. (2005)
DLESS	posterior probability	X	1 bp with Markov dependence	Viterbi path	✓ (any)	✓ (any)	✓	Siepel et al. (2006)

- **Heuristic phylogenetic methods**

  - Margulies et al. (2003) proposed two methods using a phylogenetic tree.

  - *binCons* compares each species to a reference species (e.g., human), and estimates the significance of the observed number of matches in a 25 bp window using neutral sites and a binomial distribution. Species-specific p-values are combined by "phylogenetic averaging".

  - Based on *FootPrinter* program (Blanchette and Tompa, 2002; Blanchette et al., 2002), which treats all species symmetrically by calculating an overall parsimony score reflecting the minimum number of substitutions needed along the tree to account for the observed sequences. The significance of the parsimony score for a single site is assessed relative to the distribution of parsimony scores under a continuous-time Markov model of neutral evolution (Hasegawa et al. 1985), and these p-values are combined over sites in a 25 bp window.

  - Both methods treat sites as independent and take the maximum score for all overlapping windows at each site.

Method	Score Type	Statistical Significance	Resolution	Predicted Elements	Phylogenetic Tree Used?	Lineage-Specific	Indels	Refs
MultiPipMaker	% ID	X	1 bp	X	X	X	X	Schwartz et al. (2000, 2003)
PhyloVISTA	% ID	X	1 bp	X	X	user-specified node	X	Mayor et al. (2000); Shah et al. (2004)
SynPlot	% ID	X	user-specified window	X	X	X	X	Gottgens et al. (2001); Chapman et al. (2004)
Mulan	% ID	X	user-specified window	X	X	X	X	Ovcharenko et al. (2004, 2005)
binCons	binomial *p*-value	X	25 bp	maximal windows	phylogenetic averaging	X	X	Margulies et al. (2003)
FootPrinter	parsimony *p*-value	X	25 bp	maximal windows	✓ (HKY) (for *p*-values only)	X	X	Blanchette and Tompa (2002) Blanchette et al. (2002); Margulies et al. (2003)
(*no name*)	ω	false positive rate	25 bp	maximal windows	✓ (JC)	X	✓	Cooper et al. (2004)
SCONE		false positive		maximal	✓			

- **Molecular evolutionary methods**

  - Continuous time Markov models of neutral sequence evolution have become a standard tool.

  - Using a Jukes-Cantor parameterization, Cooper et al. (2004) estimated substitution rates in 25 bp windows as a percentage of the neutral rate ω.

  - Asthana et al. (2007) used a context-dependent model for tri-nucleotide substitutions plus indels. Their method (*SCONE)*, evaluates the statistical significance of the statistic at a single site compared to its distribution over data sets simulated from the neutral model.

  - A neutral model is used in a slightly different way by the method GERP (Cooper et al., 2005), which scores conservation based on the **difference (not ratio)** between the estimated number of substitutions at a site and the expected number under an HKY model of neutral evolution.

  - All of these methods identify conserved elements based on runs of sites with conservation scores and then estimate false positive rates on sets of elements.

Method	Score Type	Statistical Significance	Resolution	Predicted Elements	Phylogenetic Tree Used?	Lineage-Specific	Indels	Refs
(*no name*)	ω	false positive rate	25 bp	maximal windows	✓ (JC)	X	✓	Cooper et al. (2004)
SCONE	ω	false positive rate	1 bp	maximal windows	✓ (trinucleotide)	X	✓	Asthana et al. (2007)
GERP	rejected subs	false positive rate	1 bp	maximal windows	✓ (HKY)	X	✓	Cooper et al. (2005)
Phylogenetic shadowing	likelihood ratio	X	1 bp	X	✓ (HKY)	X	X	Boffelli et al. (2003)
phastCons	posterior probability	X	1 bp with Markov dependence	Viterbi path	✓ (REV)	X	X	Siepel and Haussler (2004a); Siepel et al. (2005)
DLESS	posterior probability	X	1 bp with Markov dependence	Viterbi path	✓ (any)	✓ (any)	✓	Siepel et al. (2006)

- **Molecular evolutionary methods (Cont.)**

- Rather than directly evaluating observed data relative to a neutral model, one can also compare the likelihood of the data under a neutral model relative to an alternative (i.e., conserved) model.

- Boffelli et al. (2003) plotted the log likelihood ratio for a "fast" versus "slow" HKY model along the chromosomes.

- The *phastCons* method (Siepel and Haussler, 2003; Siepel et al., 2005) uses a phylogenetic hidden Markov model (phylo-HMM) with "conserved" and "not-conserved" states. The transitions between these states along the chromosome are modeled with a Markov process, and the most likely state path is used to predict conserved elements. At each site, the posterior probability of the data being generated by the conserved state provides a conservation score.

Method	Score Type	Statistical Significance	Resolution	Predicted Elements	Phylogenetic Tree Used?	Lineage-Specific	Indels	Refs
(*no name*)	ω	false positive rate	25 bp	maximal windows	✓ (JC)	X	✓	Cooper et al. (2004)
SCONE	ω	false positive rate	1 bp	maximal windows	✓ (trinucleotide)	X	✓	Asthana et al. (2007)
GERP	rejected subs	false positive rate	1 bp	maximal windows	✓ (HKY)	X	✓	Cooper et al. (2005)
Phylogenetic shadowing	likelihood ratio	X	1 bp	X	✓ (HKY)	X	X	Boffelli et al. (2003)
phastCons	posterior probability	X	1 bp with Markov dependence	Viterbi path	✓ (REV)	X	X	Siepel and Haussler (2004a); Siepel et al. (2005)
DLESS	posterior probability	X	1 bp with Markov dependence	Viterbi path	✓ (any)	✓ (any)	✓	Siepel et al. (2006)

Methods Commonly used to detect conservation in DNA sequences

Method	Score Type	Statistical Significance	Resolution	Predicted Elements	Phylogenetic Tree Used?	Lineage-Specific	Indels	Refs
MultiPipMaker	% ID	X	1 bp	X	X	X	X	Schwartz et al. (2000, 2003)
PhyloVISTA	% ID	X	1 bp	X	X	user-specified node	X	Mayor et al. (2000); Shah et al. (2004)
SynPlot	% ID	X	user-specified window	X	X	X	X	Gottgens et al. (2001); Chapman et al. (2004)
Mulan	% ID	X	user-specified window	X	X	X	X	Ovcharenko et al. (2004, 2005)
binCons	binomial p-value	X	25 bp	maximal windows	phylogenetic averaging	X	X	Margulies et al. (2003)
FootPrinter	parsimony p-value	X	25 bp	maximal windows	✓ (HKY) (for p-values only)	X	X	Blanchette and Tompa (2002) Blanchette et al. (2002); Margulies et al. (2003)
(*no name*)	ω	false positive rate	25 bp	maximal windows	✓ (JC)	X	✓	Cooper et al. (2004)
SCONE	ω	false positive rate	1 bp	maximal windows	✓ (trinucleotide)	X	✓	Asthana et al. (2007)
GERP	rejected subs	false positive rate	1 bp	maximal windows	✓ (HKY)	X	✓	Cooper et al. (2005)
Phylogenetic shadowing	likelihood ratio	X	1 bp	X	✓ (HKY)	X	X	Boffelli et al. (2003)
phastCons	posterior probability	X	1 bp with Markov dependence	Viterbi path	✓ (REV)	X	X	Siepel and Haussler (2004a); Siepel et al. (2005)
DLESS	posterior probability	X	1 bp with Markov dependence	Viterbi path	✓ (any)	✓ (any)	✓	Siepel et al. (2006)

- Protein-based methods
- Percent identity scoring methods
- Heuristic phylogenetic methods
- Molecular evolutionary methods

PhyloP

- PhyloP focuses on unsupervised, statistical, phylogenetic methods, which we believe have the greatest promise for general functional element discovery and characterization.

- The primary signal: conservation or constraint—that is, a reduced rate of evolution compared to what is expected under neutral drift.
- Recent methods for "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; Wong and Nielsen 2004).
- Most conservation-detection methods: scan entire genomic alignments.
- Acceleration-detection methods: applied to predefined elements of interest.
- PhyloP treats conservation and acceleration in a unified manner.

PhyloP

- Four methods for detecting nonneutral substitution rates on a phylogeny: a likelihood ratio test (LRT), a score test (SCORE), a method based on the distribution of the number of substitutions per site (SPH), and the genomic evolutionary rate profiling (GERP) method.
- Implemented all four methods in a program called phyloP ("phylogenetic *P*-values"), which is freely available as part of the PHAST package.
- Two 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.

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.

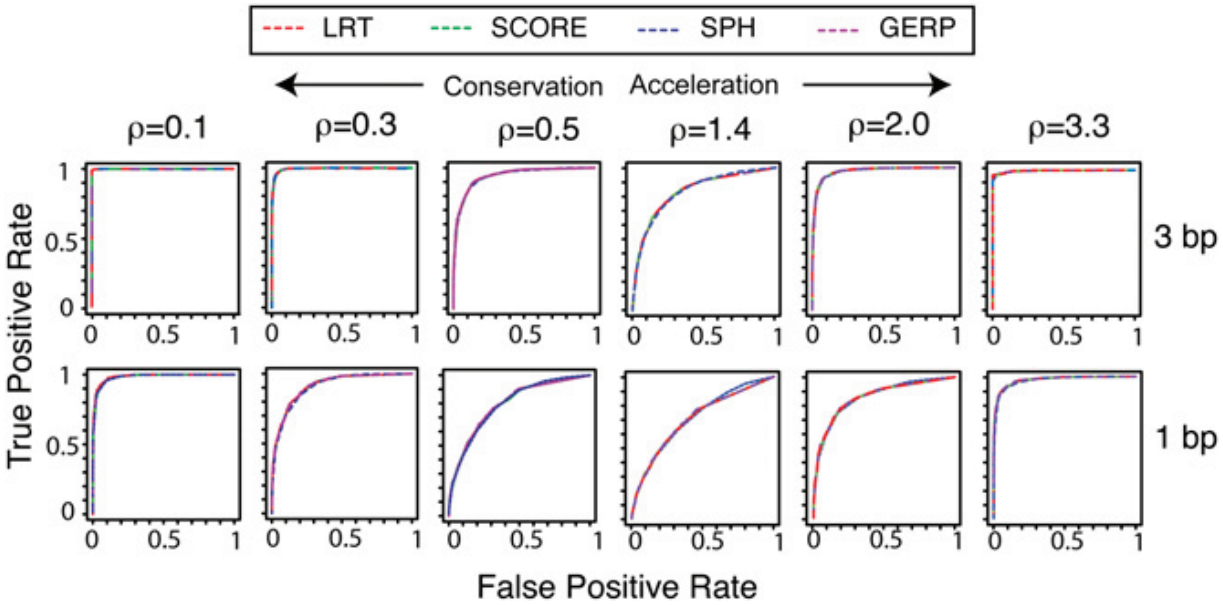


Figure 1. Receiver operating characteristic (ROC) curves showing false-positive versus true-positive rates for the all-branch tests implemented in phyloP: (red) LRT, (green) SCORE, (blue) SPH, and (purple) GERP. Individual plots show results for simulated data sets with either 3-bp (*top*) or 1-bp (*bottom*)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

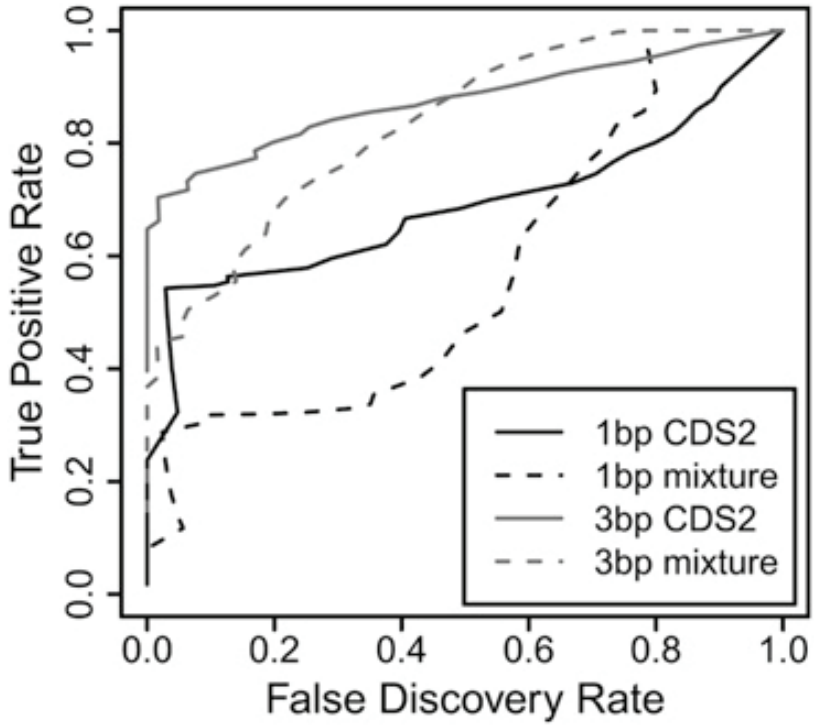
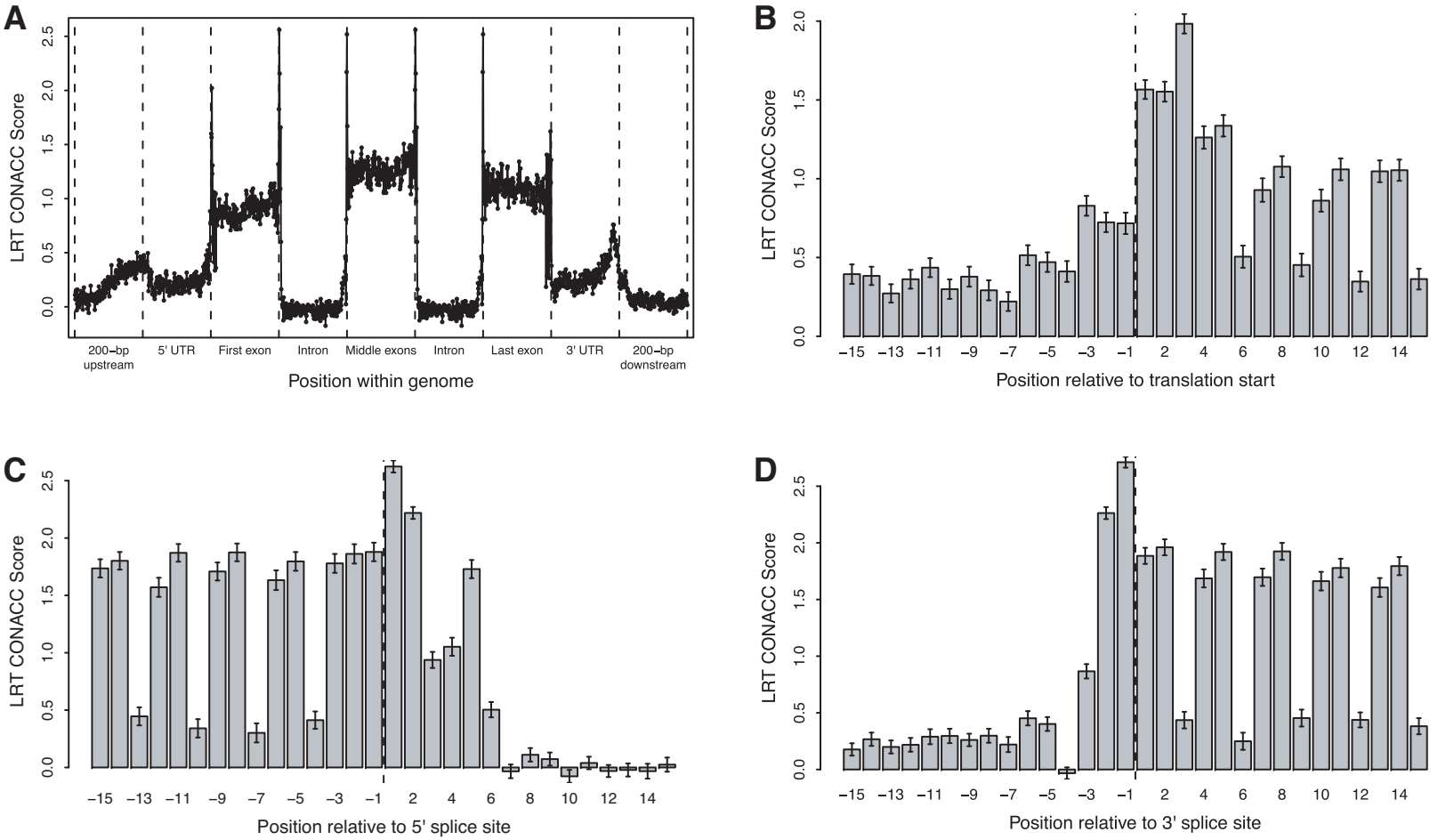


Figure 2. Estimated FDR for all-branch LRT. Estimates of false discovery rate (FDR) versus true-positive rate (TPR) based on two indirect methods, for 1-bp and 3-bp elements. (CDS2) Average TPRs are estimated from second codon position sites; (mixture) average TPRs are estimated by decomposing the genome-wide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


A
LRT CONACC Score
0.0
0.5
1.0
1.5
2.0
2.5
200–bp upstream
5' UTR
First exon
Intron
Middle exons
Intron
Last exon
3' UTR
200–bp downstream
Position within genome
B
LRT CONACC Score
0.0
0.5
1.0
1.5
2.0
–15 –13 –11 –9 –7 –5 –3 –1 2 4 6 8 10 12 14
Position relative to translation start
C
LRT CONACC Score
0.0
0.5
1.0
1.5
2.0
2.5
–15 –13 –11 –9 –7 –5 –3 –1 2 4 6 8 10 12 14
Position relative to 5' splice site
D
LRT CONACC Score
0.0
0.5
1.0
1.5
2.0
2.5
–15 –13 –11 –9 –7 –5 –3 –1 2 4 6 8 10 12 14
Position relative to 3' splice site

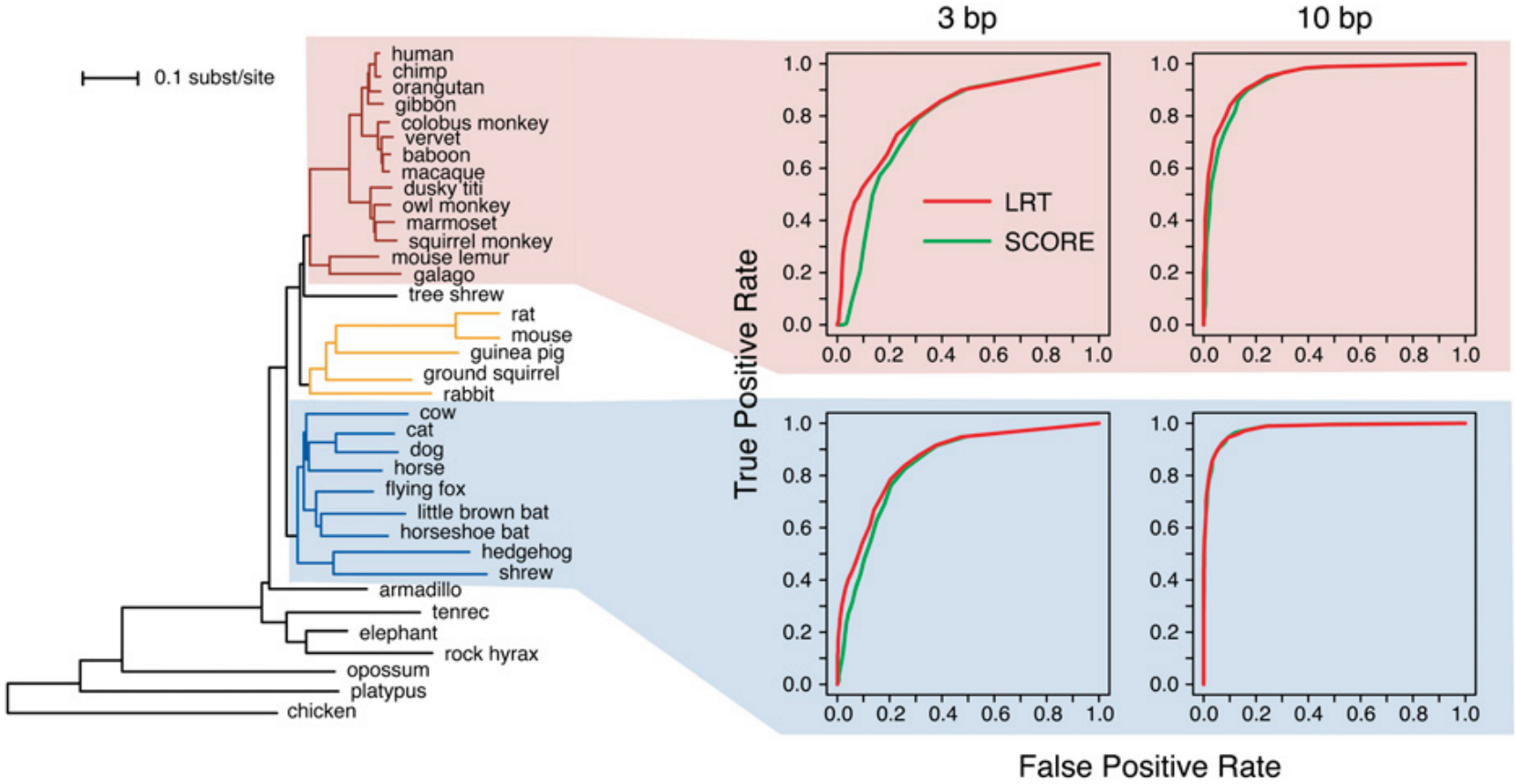
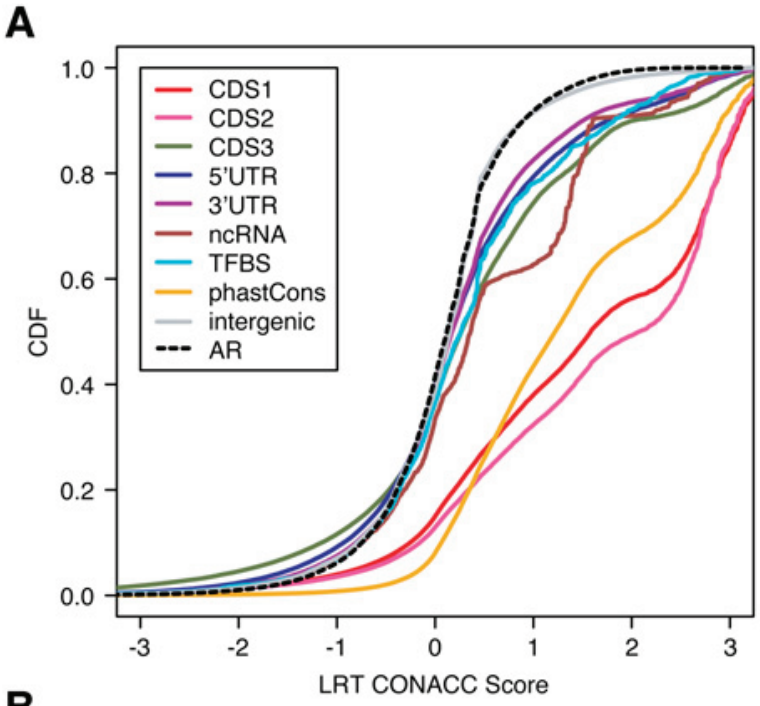
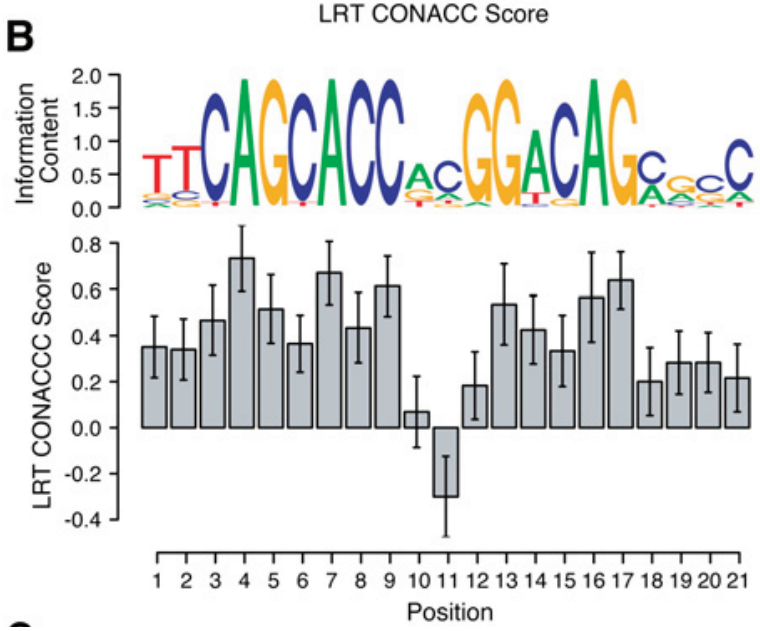
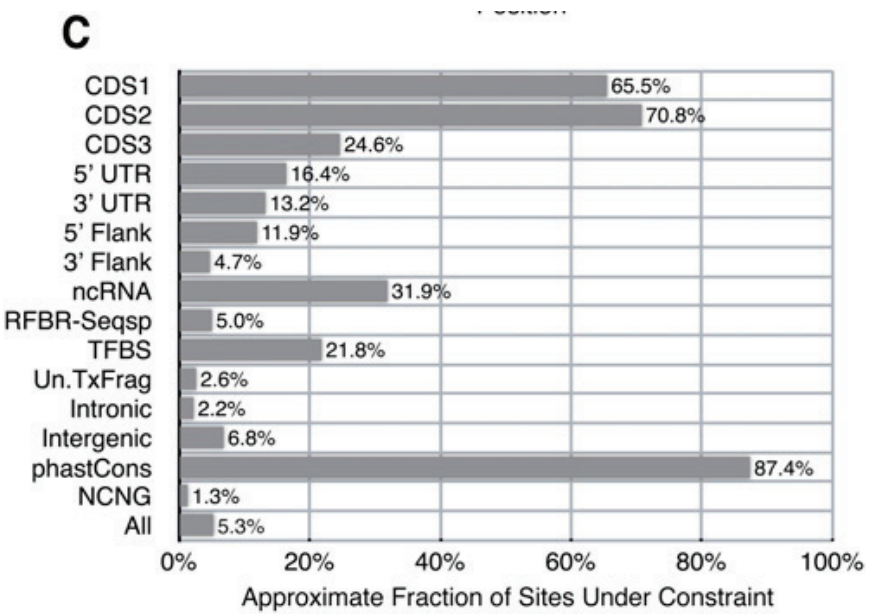


Figure 3. Subtree ROC curves. (*Left*) Phylogenetic tree used in this study, with branch lengths drawn in proportion to the values estimated from 4D sites. Three subtrees are highlighted: (maroon) primates, (gold) glires, and (blue) laurasiatherians. (*Right*) ROC curves for the LRT (red) and SCORE (green) 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.





NRSF motif



- 44 ENCODE regions (Margulies et al. 2007), which constitute the largest published comparative genomic data set for mammals.

- LRT method

- Positive scores for predicted conservation and negative scores for predicted acceleration

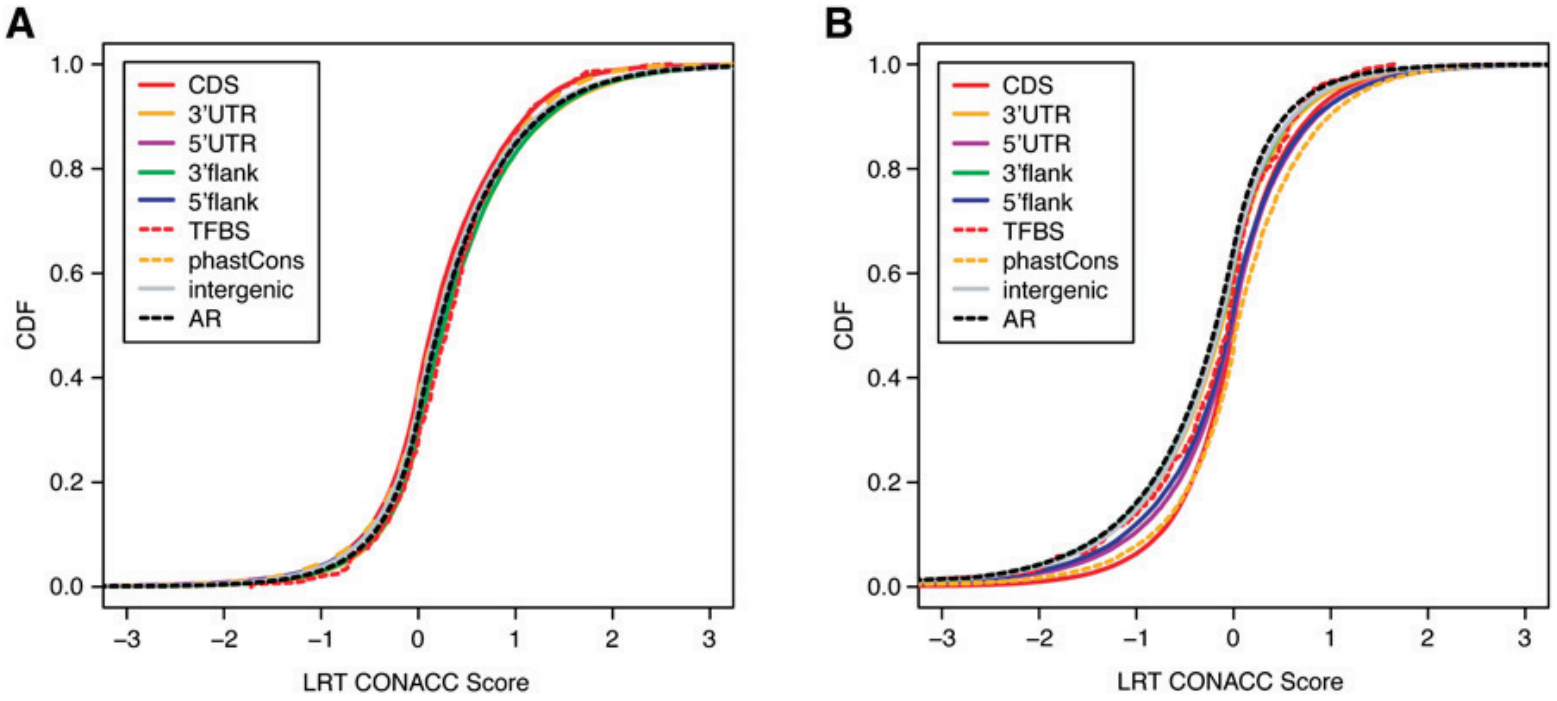


Figure 5. Distributions of subtree scores for the primate and glires clades. Cumulative distribution functions (CDFs) of scores for selected annotation classes as computed by the subtree test for the primate (*A*) and glires (*B*) clades. As in previous figures, CONACC scores computed by the LRT method 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*).

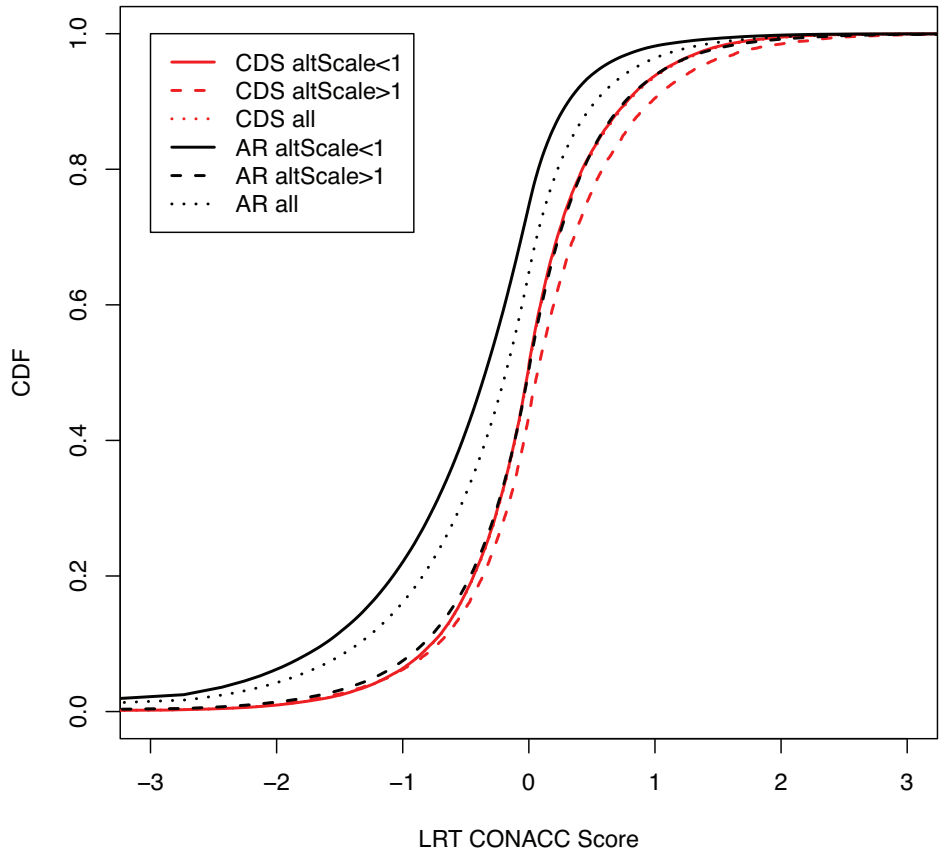


Figure S8: **Distributions of glires subtree scores for "fast" and "slow" sites.** CDFs for subtree scores in CDS and AR sites are shown, as in Figure 5B. In addition, separate curves are shown for just those 10-mers evolving slower than the neutral rate outside the glires subtree ($\hat{\rho} < 1$; here labeled "altScale<1") and for just those 10-mers evolving faster than the neutral rate ($\hat{\rho} > 1$; here "altScale>1"). Observe that the difference between the CDS and AR distributions is much more pronounced for the slow-evolving sites than for the fast-evolving sites, suggesting the general shift toward large scores in the CDS sites is driven by increased negative selection rather than loss of positive selection.

Identify clade-specific accelerated evolution in conserved elements within the ENCODE regions

- 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
- 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%, identified 216 primate accelerated regions (PARs) and 3529 glires accelerated regions (GARs).
- The glires clade shows a pronounced excess of accelerated regions over a large range of nominal *P*-value thresholds, 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.

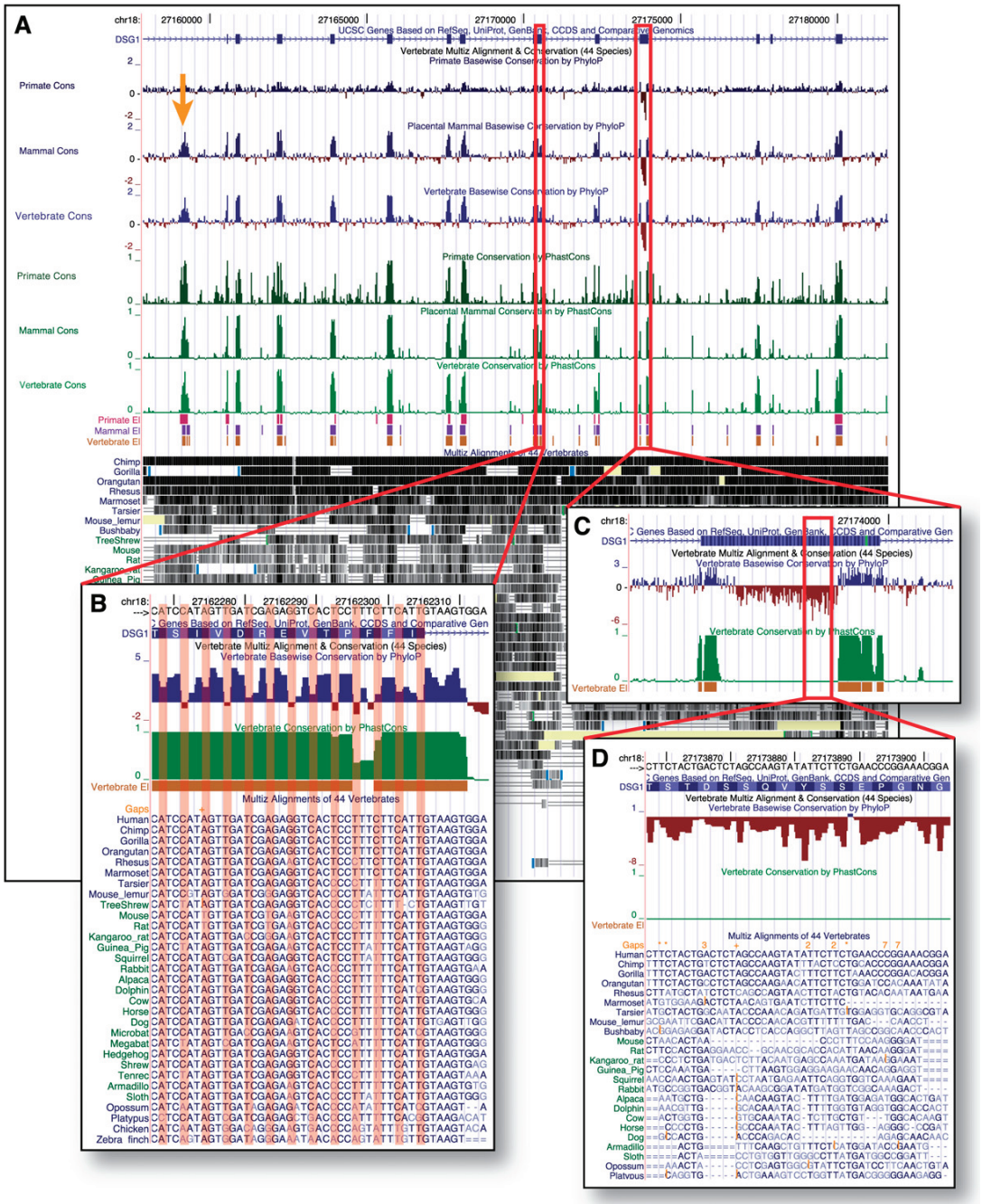


Figure 6. Conservation track in UCSC Genome Browser. A portion of the desmoglein 1 (*DSG1*) gene on human chromosome 18 shown with the new Conservation track, including a 44-way vertebrate alignment and nine conservation subtracks. The subtracks display **phyloP scores (in blue and red)**, **phastCons 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.

Discussion

- 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—e.g., 30%–75% TPRs at 5% FDRs. Similarly, moderately strong clade-specific selection can be detected at the level of 10-bp elements.
- Power will steadily improve as additional genomes are sequenced.
- The similarity in power of the four methods suggest that little is to be gained by further methodological work on this problem.
- 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).

SiPhy (Garber et al. 2009)

- Exploit the pattern of substitution by using an LRT.
- Advantage: In principle should increase power for subtle selective pressures that influence base preferences but have only a mild effect on the overall substitution rate.
- Risk: 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 to rate-based models, SiPhy may also be more influenced by phenomena associated with mutation and repair than with natural selection, e.g. transcription-coupled repair (Green et al. 2003), biased gene conversion (Marais 2003; Dreszer et al. 2007), and methylation of cytosines (Ehrlich and Wang 1981).
- Nearly twofold increase in the number of evolutionarily "constrained" sites detected by SiPhy in the ENCODE regions (Garber et al. 2009).
- Nevertheless, pattern-based methods have the potential to improve power and are worthy of further investigation.

Other limitations for rate-based methods

- Ignore regional variation and context dependencies in neutral substitution rates, variation in G+C content, transcription-associated mutational asymmetry, and differences between clades in selection on 4D sites.
- 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.
- Depend on accurate alignments of mammalian genomes. Alignment error can matter.
- It may be possible to integrate or sample 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. It may be possible to use heuristic methods to substantially improve the speed of computation (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).

XJM's summary of PhyloP vs. PhastCon

	PhyloP	PhastCon
Dependencies	Nucleotides are independent	Nucleotides have dependencies through HMM
Conservation variations from base to base	More variations	Less variations
Type of natural selection	Both conservation and acceleration	Conservation only
Package	in PHAST package	in PHAST package
Methods	have four implemented methods including GERP	-
Authors/Groups & year of publication	PhyloP: Katherine Pollard and Adam Siepel 2010 GERP: Arend Sidow 2005	Adam Siepel and David Haussler, 2005

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

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.