Supplementary Material

1 Sequence Data and Multiple Alignments

The five vertebrate, four insect, two worm, and seven yeast genomes used in the analysis are summarized in Table S1, and the four genome-wide multiple alignments are summarized in Table S2. All four multiple alignments are downloadable as supplementary data from http://www.cse.ucsc.edu/∼acs/conservation. Continually updated versions are also browsable and downloadable via the UCSC Genome Browser (http://genome.ucsc.edu).

2 Prediction of Conserved Elements Using PhastCons

2.1 The Model

Let the nonconserved phylogenetic model be defined as $\boldsymbol{\psi}_n = (\mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\beta})$, where $\mathbf{Q}$ is a 4×4 substitution rate matrix indicating the instantaneous rate of replacement of each nucleotide for each other nucleotide, $\boldsymbol{\pi}$ is a vector of background (equilibrium) probabilities for the four bases, $\boldsymbol{\tau}$ is a binary tree representing the topology of the phylogeny, and $\boldsymbol{\beta}$ is a vector of non-negative real numbers representing the branch lengths of the phylogeny in expected substitutions per site (Siepel and Haussler, 2004a, 2005). The conserved phylogenetic model, denoted $\boldsymbol{\psi}_c$, is identical to $\boldsymbol{\psi}_n$ except for the scaling parameter ρ; it is defined as $\boldsymbol{\psi}_c = (\mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}, \rho\boldsymbol{\beta})$. The tree topology $\boldsymbol{\tau}$ is assumed to be known, and the background distribution $\boldsymbol{\pi}$ is assumed to be well approximated by the relative frequencies of the four bases in the data set, which can easily be obtained in a preprocessing step. Therefore, the free parameters of the phylo-HMM are summarized by the parameter vector $\boldsymbol{\theta} = (\mathbf{Q}, \boldsymbol{\beta}, \rho, \mu, \nu)$, where μ and ν define the transition probabilities and initial-state probabilities of the Markov chain for state transitions (Figure 1).

The rate matrix $\mathbf{Q}$ (treated here as a single free parameter, for simplicity) can be defined by any of several parametric DNA substitution models (see, e.g., Whelan et al., 2001; Felsenstein, 2004). For all results discussed in this paper, the five-parameter general reversible (REV) (Tavaré, 1986) substitution model was used. Assuming the REV model and n present-day species (implying an unrooted tree with $2n-3$ branches), the total number of free parameters is $2n + 5$. The tree topologies assumed for each data set were as shown in Figure 2. The topologies for the vertebrates and insects are fairly well established, and there was only one topology possible for the two worms; for the yeasts, we used the topology reported by Rokas et al. (2003).

The model used by phastCons can be shown to be a special case of Felsenstein and Churchill's (1996) phylo-HMM with two rate categories (states), rates $r_1 = \rho$ and $r_2 = 1$, stationary distribution $(f_1, f_2) = (\frac{\nu}{\mu+\nu}, \frac{\mu}{\mu+\nu})$, and autocorrelation parameter $\lambda = 1 - \mu - \nu$.

2.2 Parameter Estimation

Let $\mathbf{x} = (x_1, \ldots, x_L)$ be a multiple alignment of n rows (sequences) and L columns, whose ith column has probabilities $P(x_i|\boldsymbol{\psi}_c)$ and $P(x_i|\boldsymbol{\psi}_n)$ of being "emitted" by the conserved and nonconserved models, respectively. Because the ancestral bases associated with x_i (i.e., the bases at internal nodes of the tree) are unknown, these "emission" probabilities are marginal probabilities, obtained by summing over all possible ancestral bases. Let $\mathbf{z} = (z_1, \ldots, z_L)$ be a *path* through the phylo-HMM, where $z_i \in \{c, n\}$ indicates whether the conserved (c) or nonconserved (n) state is visited at position i. Like the ancestral bases, the path is unobserved, so to compute the marginal probability of the data given the model (the likelihood of the model), one must also sum over paths. The joint probability of the data and a specific path is simply a product over the positions of the alignment of emission probabilities and "transition" probabilities (the probability of each state z_i given the previous state z_{i-1}). The likelihood function is:

$$L(\boldsymbol{\theta}|\mathbf{x}) = P(\mathbf{x}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} \prod_{i=1}^{L} P(x_i|z_i, \boldsymbol{\theta}) P(z_i|z_{i-1}, \boldsymbol{\theta}) \tag{1}$$

where we use the notational convention that $P(z_1|z_0)$ is the probability of beginning with state z_1. The "double marginalization" required to compute the likelihood of a phylo-HMM can be accomplished efficiently using a combination of Felsenstein's (1981) "pruning" algorithm and the "forward" algorithm for HMMs (Felsenstein and Churchill, 1996; Durbin et al., 1998b; see also Siepel and Haussler, 2004a).

In an EM algorithm, a maximum of the likelihood function is found by iteratively maximizing a related but often more tractable quantity, sometimes called the "expected complete log likelihood." The algorithm consists of alternating expectation (E) and maximization (M) steps, which are guaranteed eventually to result in convergence to a local maximum of the likelihood function (Dempster et al., 1977; Durbin et al., 1998b). Let the "complete log likelihood," denoted $l(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z})$, be the log of the joint probability of an observed alignment and a specific path:

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z}) = \log P(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) &= \log \prod_{i=1}^{L} P(x_i|z_i, \boldsymbol{\theta}) P(z_i|z_{i-1}, \boldsymbol{\theta}) \\ &= \log \left[\prod_{x \in \mathcal{X}, z \in \mathcal{Z}} [P(x|z, \boldsymbol{\theta})]^{u_{x,z}} \prod_{z_1, z_2 \in \mathcal{Z}} [P(z_2|z_1, \boldsymbol{\theta})]^{v_{z_1,z_2}} f(\mu, \nu) \right] \\ &\approx \sum_{x \in \mathcal{X}, z \in \mathcal{Z}} u_{x,z} \log P(x|z, \boldsymbol{\theta}) + \sum_{z_1, z_2 \in \mathcal{Z}} v_{z_1,z_2} \log P(z_2|z_1, \boldsymbol{\theta}) \end{aligned} \tag{2}$$

where $\mathcal{X}$ is the set of distinct columns (sometimes called "patterns") in $\mathbf{x}$, $\mathcal{Z}$ is the set of states in the phylo-HMM ($\mathcal{Z} = \{c, n\}$), $u_{x,z}$ is the number of instances in which a column x is emitted by a state z, and v_{z_1,z_2} is the number of instances in which a state z_1 is followed by a state z_2. (The counts $u_{x,z}$ and v_{z_1,z_2} are sufficient statistics for $l(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z})$.) The function $f(\mu, \nu)$, dropped in the third step, is equal to $\frac{\nu}{\mu+\nu}$ if $z_1 = c$ and equal to $\frac{\mu}{\mu+\nu}$ if $z_1 = n$; it can safely be ignored as long as the length of the alignment $L \gg 1$. The expected complete log likelihood is the expectation of $l(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z})$ considering all possible paths, conditional on the data and a particular version of the parameter vector ($\boldsymbol{\theta}^{(t)}$). Using the notation $\langle\cdot\rangle_{\theta^{(t)}}$ to indicate these conditional expectations, and letting $U_{x,z}$ and V_{z_1,z_2} be random variables corresponding to the counts $u_{x,z}$ and v_{z_1,z_2}, the expected complete log likelihood is:

$$\begin{aligned}\langle l(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z})\rangle_{\boldsymbol{\theta}^{(t)}} &\approx \left\langle \sum_{x\in\mathcal{X}, z\in\mathcal{Z}} U_{x,z} \log P(x|z, \boldsymbol{\theta}) + \sum_{z_1,z_2\in\mathcal{Z}} V_{z_1,z_2} \log P(z_2|z_1, \boldsymbol{\theta}) \right\rangle_{\boldsymbol{\theta}^{(t)}} \\ &= \sum_{x\in\mathcal{X}, z\in\mathcal{Z}} \langle U_{x,z}\rangle_{\boldsymbol{\theta}^{(t)}} \log P(x|z, \boldsymbol{\theta}) + \sum_{z_1,z_2\in\mathcal{Z}} \langle V_{z_1,z_2}\rangle_{\boldsymbol{\theta}^{(t)}} \log P(z_2|z_1, \boldsymbol{\theta}) \qquad (3)\end{aligned}$$

where we show approximate equality only because we have dropped the function $f(\mu, \nu)$. (This description of the EM algorithm follows one given by M. Jordan, book in prep.)

The E step of the algorithm consists of computing the "expected counts" $\langle U_{x,z}\rangle_{\boldsymbol{\theta}^{(t)}}$ and $\langle V_{z_1,z_2}\rangle_{\boldsymbol{\theta}^{(t)}}$. This step can be accomplished using a forward/backward procedure, essentially the same as with ordinary HMMs (Durbin et al., 1998b), but with emission probabilities that are computed by summing over possible ancestral bases via Felsenstein's (1981) algorithm. The M step is to compute a new parameter vector $\boldsymbol{\theta}^{(t+1)}$ from the old parameter vector $\boldsymbol{\theta}^{(t)}$ by maximizing the expected complete log likelihood:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \sum_{x\in\mathcal{X}, z\in\mathcal{Z}} \langle U_{x,z}\rangle_{\boldsymbol{\theta}^{(t)}} \log P(x|z, \boldsymbol{\theta}) + \sum_{z_1,z_2\in\mathcal{Z}} \langle V_{z_1,z_2}\rangle_{\boldsymbol{\theta}^{(t)}} \log P(z_2|z_1, \boldsymbol{\theta}) \qquad (4)$$

This maximization problem decomposes into a maximization problem involving only $\mathbf{Q}$, $\boldsymbol{\beta}$, and ρ (corresponding to the first term inside the arg max of equation 4) and a maximization problem involving only μ and ν (corresponding to the second term). The second problem has a straightforward solution, which we show here for completeness, although in practice μ and ν are not estimated by maximum likelihood but are determined by constraints on coverage and smoothness (see below).

$$\mu^{(t+1)} = \frac{\langle V_{c,n}\rangle_{\boldsymbol{\theta}^{(t)}}}{\langle V_{c,c}\rangle_{\boldsymbol{\theta}^{(t)}} + \langle V_{c,n}\rangle_{\boldsymbol{\theta}^{(t)}}}, \qquad \nu^{(t+1)} = \frac{\langle V_{n,c}\rangle_{\boldsymbol{\theta}^{(t)}}}{\langle V_{n,c}\rangle_{\boldsymbol{\theta}^{(t)}} + \langle V_{n,n}\rangle_{\boldsymbol{\theta}^{(t)}}} \qquad (5)$$

The first maximization problem is more difficult. The expression to be maximized is a weighted combination of the log likelihoods of the two phylogenetic models, with weights given by the expected counts $\langle U_{x,c}\rangle_{\boldsymbol{\theta}^{(t)}}$ and $\langle U_{x,n}\rangle_{\boldsymbol{\theta}^{(t)}}$. Because the two phylogenetic models share parameters, this expression does not decompose further. PhastCons performs this maximization numerically, using the BFGS quasi-Newton algorithm (Press et al., 1992).

2.3 The Constraint on Smoothness

As noted in the text, the constraint on smoothness depends on a quantity denoted $L_{\min}$, which is the expected minimum length of a sequence of conserved sites (in the midst of a stretch of nonconserved sites) required for a conserved element to be predicted. Because predictions are based on the Viterbi algorithm, $L_{\min}$ must be such that the expected log likelihood of staying in the nonconserved state of the phylo-HMM is equal to the expected log likelihood of switching to the conserved state and back again. If conserved and nonconserved sites are assumed to be drawn independently from the distributions associated with $\boldsymbol{\psi}_c$ and $\boldsymbol{\psi}_n$, respectively, then $L_{\min}$ is determined by the equation:

$$\begin{aligned}(L_{\min}+1)\log(1-\nu) + &L_{\min}\sum_x P(x|\boldsymbol{\psi}_c)\log P(x|\boldsymbol{\psi}_n)\\ &= \log\nu + \log\mu + (L_{\min}-1)\log(1-\mu) + L_{\min}\sum_x P(x|\boldsymbol{\psi}_c)\log P(x|\boldsymbol{\psi}_c)\end{aligned}$$

where the sums are over all possible distinct alignment columns x. Hence,

$$L_{\min} = \frac{\log\nu + \log\mu - \log(1-\nu) - \log(1-\mu)}{\log(1-\nu) - \log(1-\mu) - H(\boldsymbol{\psi}_c||\boldsymbol{\psi}_n)} \tag{6}$$

where $H(\boldsymbol{\psi}_c||\boldsymbol{\psi}_n) = \sum_x P(x|\boldsymbol{\psi}_c)\log\frac{P(x|\boldsymbol{\psi}_c)}{P(x|\boldsymbol{\psi}_n)}$ is the relative entropy of the distribution associated with $\boldsymbol{\psi}_c$ with respect to the distribution associated with $\boldsymbol{\psi}_n$. The key quantity of interest is the product of $L_{\min}$ with the relative entropy, $L_{\min}H(\boldsymbol{\psi}_c||\boldsymbol{\psi}_n)$, which we call the phylogenetic information threshold (PIT) because it can be interpreted as the expected minimum amount of phylogenetic information required to predict a conserved element. During parameter estimation, this quantity (a function of μ, ν, $\boldsymbol{\psi}_c$, and $\boldsymbol{\psi}_n$) was constrained to be equal for all data sets. A value of $L_{\min}H(\boldsymbol{\psi}_c||\boldsymbol{\psi}_n) = 9.8$ bits was selected, based on the degree of smoothing that had been chosen manually for the vertebrate data set, with feedback from users of the conservation track in the genome browser.

2.4 Tuning Parameters

The constraints on coverage and smoothing are met by adjusting two tuning parameters, the expected coverage γ and the expected length ω. The expected coverage γ is equal to the probability of the conserved state of the phylo-HMM at stationarity, or equivalently, the expected fraction of conserved sites at equilibrium. Its value of $\gamma = \frac{\nu}{\mu+\nu}$ is given by the eigenvector corresponding to the largest eigenvalue (which has value 1) of the transition matrix for the two-state Markov chain. The expected length parameter, $\omega = \frac{1}{\mu}$, is simply the expected number of steps for which the Markov chain will remain in the conserved state, given that it is already in that state.

The expected coverage and expected length, if set *a priori*, completely determine the values of μ and ν, and hence, completely define the Markov chain associated with the phylo-HMM. (The parameters μ and ν could as easily be treated directly as tuning parameters, but we find them to be harder to interpret.) Larger values of γ tend to result in a higher coverage by conserved elements and larger average conservation scores, and larger values of ω tend to result in "smoother" conservation scores (i.e. more similar scores at adjacent sites) and less fragmented predictions of conserved elements (implying fewer predictions with larger average lengths). When γ and ω are fixed, only $\mathbf{Q}$, $\boldsymbol{\beta}$, and ρ need to be updated in the M step of the EM algorithm, and the expected counts for state transitions (denoted $\langle V_{z_1,z_2}\rangle_{\boldsymbol{\theta}^{(t)}}$ above) need not be computed in the E step.

It is important to note that γ and ω are prior rather than posterior quantities—they define properties of the model only, which influence final (posterior) inferences but which can be overridden by a sufficiently strong signal in the data. In the absence of any phylogenetic information (e.g., if nothing were aligned to the reference genome), the posterior probability of conservation at every site (the conservation score) would equal γ. Thus, γ can also be interpreted as the prior probability that each site is conserved.

2.5 Parameter Estimation Subject to Constraints

Parameter estimation subject to constraints was accomplished by trial and error. First, values of the tuning parameters γ and ω were selected; then all other parameters were estimated genome-wide using the EM algorithm, these estimates were averaged, and genome-wide predictions were obtained. The coverage of coding regions and the PIT were then computed and compared to their target values. If necessary, γ and ω were adjusted and the cycle was repeated. Convergence typically required four to six iterations.

The final estimates of the parameters were: $\gamma = 0.265$ and $\omega = 12.0$bp ($\mu = 0.0833$, $\nu = 0.0300$) for vertebrates, $\gamma = 0.468$ and $\omega = 28.1$bp ($\mu = 0.0356$, $\nu = 0.0313$) for insects, $\gamma = 0.620$ and $\omega = 53.0$bp ($\mu = 0.0189$, $\nu = 0.0308$) for worms, and $\gamma = 0.400$ and $\omega = 23.0$bp ($\mu = 0.0435$, $\nu = 0.0290$) for yeasts. These parameters, along with the estimated phylogenetic models, implied $H(\boldsymbol{\psi}_c||\boldsymbol{\psi}_n) = 0.608$ bits/site and $L_{\min} = 16.1$bp for vertebrates, $H(\boldsymbol{\psi}_c||\boldsymbol{\psi}_n) = 0.725$ bits/site and $L_{\min} = 13.5$bp for insects, $H(\boldsymbol{\psi}_c||\boldsymbol{\psi}_n) = 0.159$ bits/site and $L_{\min} = 60.6$bp for worms, and $H(\boldsymbol{\psi}_c||\boldsymbol{\psi}_n) = 0.839$ bits/site and $L_{\min} = 11.7$bp for yeasts.

These expected minimum lengths $L_{\min}$ are not quite small enough to allow most individual conserved transcription factor binding sites (TFBSs) to be detected, but they should be small enough (with the exception of worm) to allow small clusters of conserved TFBSs to be detected. As more and more genomes become available, $H(\boldsymbol{\psi}_c||\boldsymbol{\psi}_n)$ will become larger and larger and the same PIT will be achievable with smaller and smaller values of $L_{\min}$. In the limit, individual alignment columns can be identified as conserved or nonconserved and the HMM becomes irrelevant.

2.6 Predictions and Scores

PhastCons predicts conserved elements using the Viterbi algorithm, then assigns log-odds scores to each predicted element. The log-odds score of an element predicted from position i to position j in an alignment of length L ($1 \leq i \leq j \leq L$) is

$$s_{ij} = \log \frac{P(x_i, \ldots, x_j | \boldsymbol{\psi}_c)}{P(x_i, \ldots, x_j | \boldsymbol{\psi}_n)} = \sum_{k=i}^{j} [\log P(x_k | \boldsymbol{\psi}_c) - \log P(x_k | \boldsymbol{\psi}_n)] \quad (7)$$

These scores are closely related to the likelihood ratios used by Boffelli et al. (2003).

2.7 Missing Data and Alignment Gaps

In general, only a subset of the other genomes are aligned to any given region of a reference genome, because of large-scale insertions and deletions, incomplete assemblies, or unalignability due to extreme sequence divergence. Such missing data was handled by marginalizing over missing bases when computing emission probabilities (e.g., Siepel and Haussler, 2004a). This strategy is computationally efficient and conceptually straightforward, requiring no change to the Viterbi, forward/backward, or EM algorithms. It effectively causes these algorithms to consider, at each site, only the sub-tree corresponding to the set of species for which data is available. In the presence of missing data, longer sequences of conserved sites are required for conserved elements to be predicted, and log odds scores tend to decrease. Alignment gaps are currently also treated as missing data, despite that they contain phylogenetic information and are potentially useful in helping to distinguish conserved and nonconserved regions (e.g., Ogurtsov et al., 2004; Siepel and Haussler, 2004b). Making better use of alignment gaps is an area for future work.

2.8 Synteny Filtering

Predicted elements in the vertebrate data set were discarded if they did not fall on the *syntenic net* between human and mouse—a subset of chained local alignments judged to be “syntenic” in the two species (Kent et al., 2003). The remaining predictions can be considered with reasonably high confidence to be based on alignments of orthologous sequence in human and mouse, according to alignment quality and synteny considerations. (This filter is conservative, and causes some orthologous alignments as well non-orthologous alignments to be discarded.) The human and mouse genomes were selected because of their evolutionary distance and assembly quality. A similar filter was applied to the insect predictions; in this case predictions were discarded if they did not fall on either the *D. melanogaster/D. yakuba* syntenic net or the *D. melanogaster/D. pseudoobscura* syntenic net.

3 Parameter Estimates and Predicted Elements

3.1 Variation in Local Parameter Estimates

As noted in the text, parameter estimation was done separately in approximately 1Mb windows, and individual estimates were averaged. The estimates of total nonconserved branch length (the sum of the lengths of all branches in the nonconserved tree) had mean 1.94 and s.d. 0.20 for the vertebrates, mean 1.85 and s.d. 0.14 for the insects, mean 0.614 and s.d. 0.05 for the worms, and mean 2.62 and s.d. 0.20 for the yeasts (units of subst./site). The estimates of ρ had mean 0.325 and s.d. 0.05 for the vertebrates, mean 0.240 and s.d. 0.04 for the insects, mean 0.361 and s.d. 0.04 for the worms, and mean 0.320 and s.d. 0.01 for the yeasts. The parameters estimates for individual windows are available as supplementary data (http://www.cse.ucsc.edu/∼acs/conservation).

3.2 Nonconserved Branch Lengths

In many respects, the "nonconserved" models estimated by phastCons (Figure 2) agree well with independent estimates based on sites that are apparently evolving neutrally. For example, in the vertebrate tree, the subtree including human, mouse, and rat had total branch length equal to 0.66 substitutions/site, compared to a recent genome-wide estimate of 0.65 substitutions/site (Cooper et al., 2004). In addition, the ratio of the distances from the rodent ancestor to human, rat, and mouse was 4.9:1.1:1 compared to Cooper et al.'s (2004) ratio of 4.8:1.1:1.

The branches between the mammals and the non-mammalian vertebrates (chicken and *Fugu*), however, appeared to be substantially underestimated (see, e.g., International Chicken Genome Sequencing Consortium, 2004; Margulies et al., 2005), presumably because because of an alignment-related ascertainment bias (i.e., regions chicken and *Fugu* that align to the mammalian genomes are probably not evolving neutrally, but instead are under at least weak selective pressure). In addition, the mammalian subtree is rooted in a way that suggests more clock-like substitution rates than have been estimated from neutral DNA. (The nonconserved phylogeny implies that the ratio of the total distances from mouse and human to their most recent common ancestor is only about 1.5:1, while the true ratio is probably closer to 3:1; Cooper et al., 2003; Siepel and Haussler, 2004c.) The reason is apparently that substitution rates in conserved sites are more clock-like than those in neutral sites (K. Pollard, pers. comm.), and—due to sparse data from "outgroup" species in nonconserved regions—these sites end up exerting a dominant influence on the estimated position of the root of the mammalian subtree.

With the fly and yeast data sets, the branch lengths estimated by phastCons also appeared to be reasonable for close species but "shrunken" for more distant species. In particular, the two species in the worm set, (*C. elegans* and *C. briggsae*), are quite divergent and are considered to be essentially unalignable in neutrally evolving regions (Kent and Zahler, 2000; Stein et al., 2003). The estimated distance between them of 0.614 subst./site is probably a substantial underestimate.

More accurate nonconserved (neutral) distances might be obtained by the more traditional method of fitting a phylogenetic model to "fourfold degenerate" (4d) sites—i.e., sites at which the encoded amino acid does not depend on which of the four nucleotides is present. This method has certain deficiencies as well—e.g., distances can be underestimated because of purifying selection at 4d sites (due to codon bias or noncoding constraints) and some shrinkage of distance estimates can still occur (due to "saturation" at 4d sites)—but because alignments of coding regions are generally quite good estimates based on 4d sites should be less sensitive than ours to alignment biases.

To gauge the extent to which the nonconserved phylogenies estimated by phastCons reflect neutral evolution, we compared these phylogenies to ones estimated from 4d sites. For each set of multiple alignments, we extracted all columns corresponding to third codon positions of known genes such that all species present had identical bases in the first and second codon positions and those bases defined fourfold degenerate codons. (Missing data was allowed—i.e., each codon did not need to be present in all species.) We then fitted a phylogenetic model to these columns using the phyloFit program in the PHAST package (REV substitution model). The resulting trees are shown in Figure S1, in comparison to the nonconserved trees estimated by phastCons.

The estimated branch lengths of the two sets of trees are fairly similar for shorter branches, but as expected, the longer branches are substantially shorter when estimated by phastCons than when estimated from 4d sites, sometimes by a factor of two or more. Interestingly, the branch lengths between human, mouse, and rat in the vertebrate tree are shorter when estimated from 4d sites than when estimated by phastCons, probably mostly because of constraint at 4d sites, but perhaps also because of accelerated evolution in noncoding regions, e.g., due to positive selection or non-equilibrium base- or word-composition in transposons upon insertion. The opposite effect occurs with the five most closely related yeast species: they are found to be more distant from one another in 4d sites than in phastCons "nonconserved" regions. This may occur because phastCons is underestimating the fraction of the yeast genomes that is conserved—because conserved sites are mixed in with nonconserved sites, the branch lengths in nonconserved regions are shrunken (see below).

There are clearly some distortions in the nonconserved trees estimated by phastCons, but there are also probably some distortions in the ones estimated from 4d sites. In any case, the key question is not whether the phylo-HMM estimated by phastCons perfectly characterizes the substitution processes in conserved and nonconserved regions, but whether it discriminates effectively between these regions. It is possible, for example, that the long branches underestimated by phastCons have little effect on its predictions, because long branches tend to be uninformative, and because the proportion of missing data tends to be high for distant species. Below we experimentally examine the effect on the predicted elements of distortions in the nonconserved phylogenies, and find, indeed, that these distortions do not have a pronounced effect.

3.3 Lengths and Scores of Predicted Conserved Elements

The lengths of predicted elements for all four data sets were approximately geometrically distributed, with average values of 103.8bp for the vertebrates, 120.6bp for the insects, 268.8bp for the worms, and 99.6bp for the yeasts. Elements ranged in length from 5bp[1] to maximum lengths of 4922bp for the vertebrates, 6095bp for the insects, 12646bp for the worms, and 4005bp for the yeasts. The differences in the lengths of the elements predicted for the four data sets primarily reflect differences in the amount of phylogenetic information available rather than biologically interesting characteristics of functional elements in the different species sets. More phylogenetic information at each site (e.g., due to more species, higher alignment coverage, or greater total branch length) allows the detection of shorter elements and makes long elements more likely to be broken up, resulting in an overall decrease in the average lengths of predicted elements (see above).

While it is difficult to compare element lengths across species groups, they can be meaningfully compared within a group. Subsets of vertebrate conserved elements that primarily overlap (>50% of bases) known CDSs, UTRs, introns or intergenic (unannotated) regions, or ARs have noticeably different length distributions (Figure S2A). Elements in ARs tend to be shortest (median length 50bp), elements in introns and intergenic regions tend to be slightly longer (median length 64bp), elements in UTRs tend to be longer still (median length 70bp), and elements in CDSs tend to be longest (median length 87bp). (The length distributions for $5'$ and $3'$ UTR elements were similar, as were those for intronic and intergenic elements.) In addition, the length distribution for CDS elements has a distinctive shoulder, reflecting the fact that most vertebrate exons have lengths in the range of 100–150bp.

The distributions of log-odds scores are similar to the distributions of element lengths, due to a strong correlation between lengths and scores ($r^2 = 0.77$), but the distributions of length-normalized scores (Figure

[1] The lengths of predicted elements were measured in the coordinate frame of the reference genome. There were a few dozen elements of <5bp, but all appeared to be longer in the coordinate frame of the multiple alignment due to insertions or deletions.

S2B) accentuate the differences between the various classes of conserved elements. Elements in ARs are typically not only shorter but less highly conserved (i.e., lower scoring) than elements in introns and intergenic regions, which are less highly conserved than UTR elements, which are less highly conserved than CDS elements. Note that the log-odds scores tend to decrease with increasing "missing data" in the alignments (i.e., due to the absence or unalignability of sequences orthologous to the reference genome); this partly explains the low scores of ARs and other conserved elements in introns and intergenic regions, because these sequences are typically not present or not alignable in non-mammalian vertebrates such as chicken and fish.

Consistent with these differences in length and score distributions, the composition of the set of conserved elements is dependent on element score (Figure S3). In particular, the fractions of conserved elements in coding regions, UTRs, and other mRNAs generally increases with element score, while the fractions in introns and unannotated regions generally decrease with score. Note that the fraction in $3'$ UTRs is particularly large among the highest scoring elements (see text). The percentage of bases in ARs also decreases sharply with element score (results not shown).

In the fly, worm, and yeast data sets, the distributions of lengths and length-normalized scores, and the dependence of element composition on score, were similar to those for the vertebrates, except that the UTR distributions for fly and worm were strongly influenced by the fact that the UTRs in these species tend to be much shorter than vertebrate UTRs (data not shown).

3.4 Functional Enrichment of Genes Associated with HCEs

Several well-studied human genes overlapped by HCEs are shown in Table S3, and functional enrichments of insect, worm, and yeast genes overlapped by HCEs in coding regions are shown in Table S4. These tables are referred to and discussed briefly in the text.

3.5 Vertebrates HCEs and Segments Rich in Conserved Noncoding Sequence

About 1.5% of bases in known coding regions fall in phastCons high-CNF segments, similar to the fraction of bases in coding regions genome-wide. Thus, these segments are not significantly depleted for genes, in contrast to human/chicken high-CNF segments, which showed nearly an eight-fold depletion. This difference may be partly due to mammal-specific regulatory elements proximal to coding regions, e.g., in promoter regions, UTRs, or introns, because the elements predicted by phastCons need not be present in non-mammalian vertebrates.

For twelve phastCons high-CNF segments, the level of non-coding conservation with chicken is below the genome average. Interestingly, those twelve show a distinctly higher G+C fraction than the phastCons high-CNF segments having high conservation with chicken (>45% vs. <40%). Some of these segments appear to represent regions with mammal-specific conservation, such as one on chromosome X that contains *PLAC1*, a gene not found in chicken or *Fugu* which has placenta-specific expression (Cocchia et al., 2000). Others appear related to weaknesses in the chicken assembly. Conversely, all of the human/chicken high-CNF segments have above-average CNF_{pc}. The four human-chicken high-CNF segments with the lowest CNF_{pc} (<6%; i.e., less than twice the genome-wide average of 3.4%) contain or are contained in the trans-dev genes *DACH2*, *IRX1*, *SALL3*, and *ZFHX4*.

To test whether HCEs were correlated with phastCons high-CNF segments, the high-CNF segments were redefined such that the HCEs were excluded when computing the CNF_{pc}. The new set was reduced in number from 101 to 69 and reduced in genome-wide coverage from 2.1% to 1.6% (average length 645kb and CNF_{pc} 13.3%). However, this new set still included or overlapped 13% of all HCEs and 18% of intronic/intergenic HCEs—8-fold and 13-fold enrichments, respectively (the set included 1.5% of all bases and 1.4% of intronic/intergenic bases).

It is possible that the observed correlation between phastCons HCEs and phastCons high-CNF segments is partly a consequence of mutation rates, because (unlike human/chicken high-CNF segments) the phastCons high-CNF segments appear to have a slightly suppressed rate of neutral substitution, as estimated from human/mouse ancestral repeats (ARs) ($t_{\text{AR}} = 0.439$ subst./site [s.d. 0.027] in phastCons high-CNF segments; $t_{\text{AR}} = 0.464$ subst./site [s.d. 0.023] genome-wide). However, this apparent difference in mutation rates is small compared to the difference in substitution rates estimated by phastCons for conserved and nonconserved regions (Figure 2). Also, it could be an artifact of an increased fraction within high-CNF segments of constrained (functional) sites in ARs. This issue remains unresolved.

4 Robustness of Results to Procedure for Parameter Estimation

Many variations are possible on the methods used to estimate the parameters of the model and to calibrate the model across species groups. In this section, a few alternative approaches are discussed and are shown to yield generally similar results to those presented in the main part of the paper. Some deficiencies of these alternative methods are also discussed.

4.1 Full Maximum Likelihood Estimation

Perhaps the most natural and straightforward approach to parameter estimation would be to estimate all parameters, including the HMM state-transition parameters μ and ν, from the data by maximum likelihood. This approach avoids the need for *a priori* constraints on coverage and smoothness and does not require any assumptions about similarities across species groups in the way coding regions (or any other functional elements) evolve.

Maximum likelihood estimates of μ and ν, corresponding values of ω and γ, and various properties of conserved elements predicted using these estimates are shown in Table S5 (rows labeled “MLE”), along with corresponding results for other parameter estimation methods (see below). These results were obtained by averaging parameter estimates based on approximately 1Mb windows, then making predictions genome-wide based on these global averages, as described in the text. For the insect, worm, and yeast data sets, the parameter estimates were based on the entire genome-wide alignments, but in the interest of time, the vertebrate parameter estimates were based not on the entire data set but on a random sample of 100 1Mb windows. (The predictions were still done genome-wide.)

The maximum likelihood estimates of μ and ν turn out not to be ideal for our purposes, particularly for the larger and less gene-dense genomes (i.e., the vertebrates and insects). As shown in Table S5, the m.l.e.’s of μ and ν are quite small relative to estimates based on other methods, resulting in large values of the expected-length parameter ω and a tendency to predict small numbers of long conserved elements. For example, compared to the method described in the text (represented by the “65%” rows in Table S5), maximum likelihood parameter estimation yields, in vertebrates, roughly half as many conserved elements with roughly twice the average length. The quantity $L_{\min}$, the expected minimum length of a sequence of conserved sites required to predict a conserved element, is also nearly twice as large. Associated with large values of ω is a high degree of “smoothing” in the conservation plot. We have found that the conservation plot and predicted conserved elements based on the m.l.e. method do not match biologists’ intuition about what is conserved and what is not conserved: surprisingly long sequences of apparently conserved sites can receive low conservation scores and not be predicted as conserved elements, and surprisingly long sequences of apparently nonconserved sites, in the midst of conserved sites, can receive high conservation scores and be included in predicted conserved elements. This “oversmoothing” from maximum likelihood estimation appears to occur because the geometric length distributions of conserved and nonconserved elements implicitly assumed by the model do not match the data well. Particularly in large, gene-sparse genomes, an abundance of very long nonconserved sequences and an abundance of conserved elements of intermediate length (see Figure S2A) presumably push both μ and ν downward, into a region of the parameter space that has higher likelihood but is of less biological relevance. This effect is strongest for the vertebrates, but a similar effect occurs with the insects, and there is a slight tendency for reduced numbers of predictions and/or longer average lengths with the worms and yeasts as well.

The maximum likelihood estimates of the coverage parameter γ also differ somewhat from estimates based on the method described in the text (the “65% CDS coverage” method), but the differences are less pronounced than with ω (Table S5). Because γ acts as a prior in the 65% coverage method, differences between the m.l.e.'s for γ and the values used in our main analysis might be taken to imply unjustified biases toward greater or lower coverage, and hence, toward higher false positive or false negative error rates (see below). However, the effects of these differences on the overall (posterior) coverage and the coverage of coding regions are fairly small with the vertebrate and insect data sets. With the worm data set, the coverage constraint on coding regions did force γ and the posterior coverages to be considerably higher than under the maximum likelihood method, perhaps implying a somewhat elevated false positive rate. Because of the clear “oversmoothing” effect of the m.l.e.'s, however, it is difficult to put too much stake in these differences. With the yeast data set, the coverage constraint may have produced a bias in the other direction, toward reduced coverage and an elevated false negative rate.

Interestingly, despite the oversmoothing effect of the m.l.e.'s, the composition of conserved elements by annotation type is essentially the same under both maximum likelihood estimation and the 65% coverage method (Figure S4). This comparison suggests that our conclusions about the fractions of conserved elements in coding vs. noncoding regions and about differences in these fractions across species groups are fairly robust.

4.2 Alternative Coverage Constraints

The target coverage of 65% was chosen essentially arbitrarily, using Chiaromonte et al.'s (2003) human/mouse analysis as a guideline. One might ask what happens to the predicted elements if an alternative target is chosen. To gauge the sensitivity of our results to the target coverage, we repeated the parameter estimation and prediction steps for all groups using alternative targets of 55% and 75%. As in the m.l.e. analysis, parameter estimation was done genome-wide (by averaging estimates for 1Mb windows) except for the vertebrates, where a random sample of 100 1Mb windows were used; predictions were made genome-wide for all species groups. The estimation procedure and smoothing constraint (PIT = 9.8 bits) were as described in the text.

For all species groups, the tuning parameter γ (expected coverage) must be steadily increased to reach target coverages increasing from 55% to 65% and then to 75%, as would be expected (Table S5). Less obviously, the tuning parameter ω (expected length) also must be increased, causing an accompanying increase in the average length of predicted conserved elements. (This appears to be due to a decrease in $H(\boldsymbol{\psi}_c||\boldsymbol{\psi}_n)$ as γ increases.) The overall coverage of each reference genome by conserved elements changes significantly as the target CDS coverage is increased from 55% to 75%, e.g., from 2.8% to 8.1% in vertebrates and from 36.9% to 53.1% in insects. The composition of the set of conserved elements, however, generally does not change dramatically (Figure S5). Indeed, for the insect, worm, and yeast data sets, the composition of conserved elements is remarkably insensitive to the target coverage. The greatest change occurs with the vertebrate data set, where the fraction of bases in coding regions is smallest, and hence, the sensitivity of the composition of the set to false positive or false negative predictions (or to the definition of what constitutes a conserved element) in noncoding regions is greatest. Nevertheless, in all cases the vast majority (64.2% to 78.1%) of vertebrate conserved elements are found to lie outside the exons of known or suspected protein-coding genes. The finding that the fraction of conserved bases that fall in coding regions increases from vertebrates to insects to worms to yeasts also does not seem to be highly sensitive to the target coverage.

4.3 Prediction Based on Local Parameter Estimates

For our main analysis, we have used a global parameter estimation scheme, fitting a model to each complete data set and then using it genome-wide for prediction. Potentially important factors, however, such as G+C content and the neutral rate of substitution, are known to vary across the genomes of vertebrates and other species (e.g., Mouse Genome Sequencing Consortium, 2002; Hardison et al., 2003; International Chicken Genome Sequencing Consortium, 2004; Stein et al., 2003). Variation in evolutionary rates is of particular concern—by assuming uniform substitution rates across each reference genome, we might effectively be applying different thresholds for conservation in different regions, and the predicted conserved elements might be enriched for slowly mutating regions.

To test whether our results were strongly influenced by failing to take such variation into account, we estimated parameters separately in 100 randomly selected 1Mb intervals from the vertebrate data set (using the 65% coverage target), then, instead of averaging these estimates, we predicted conserved elements in each interval using the locally estimated parameters. Thus, the predicted conserved elements reflected local G+C content and neutral substitution rates.

These predictions based on local parameter estimates were generally similar to predictions based on global parameter estimates, with about 90% overlap at the base level. The coverage of the "local" predictions (4.6%) was similar to but slightly lower than that of the "global" predictions (4.9%; these values are prior to synteny filtering); the same was true of the coverage of coding regions (64.7% vs. 68.4%). The composition of the elements by annotation type was not substantially different.

One problem with using locally estimated parameters is that they may cause the sensitivity to conserved bases in coding regions (which tend to be G+C rich) to depend on the G+C richness and/or overall gene-density of the local region. Errors in the estimates due to sparse data can also be a problem, because the model is relatively parameter rich. These factors presumably account for a portion of the differences observed between the local and global predictions. In general, it is more difficult with a local estimation scheme to ensure that predictions and conservation scores are comparable from one region to another. For this reason, we chose to use global parameter estimates for our main analysis.

It is worth noting that we observe a significant negative correlation ($r^2 = 0.28$), in non-overlapping 1Mb windows across the human genome, between the estimated rate of substitution between human and mouse in ancestral repeats (t_{AR}) and the fraction of bases of the human genome that fall in predicted conserved elements. This correlation might be taken to imply that variation in the neutral rate of substitution has a significant effect on the threshold for conservation. An alternative explanation, however, might be that there is a negative correlation between t_{AR} and the density of legitimate functional elements, perhaps both because functional elements in highly mutable regions will have a stronger tendency to be lost during evolution, and because ARs in the vicinity of functional elements are more likely to contain functional elements themselves. Whether or not such a correlation exists is difficult to know without much more complete annotation of functional elements.

4.4 Prediction Based on Fourfold Degenerate Sites

As noted above, there are some distortions—probably mostly due to ascertainment biases from alignment—in the nonconserved trees estimated by phastCons. To test whether these distortions have a significant influence on the set of predicted conserved elements and on the main conclusions of the paper, we predicted a new set of conserved elements for each species group, replacing the nonconserved model estimated by phastCons with a model estimated from 4d sites. (To avoid base-composition biases, we used only the branch lengths and not the [extremely G+C rich] background distribution estimated from 4d sites; the tuning parameters were left unchanged.) Table S6 compares these conserved elements to ones based on the 65%-coverage target.

In general, the use of the 4d model, with the longer branches to distant species, tends to increase the sensitivity (and decrease the specificity) to conserved elements—the longer branches in the nonconserved phylogeny make moderately conserved sequences more likely to be considered conserved. In addition, the relative entropy $H(\boldsymbol{\psi}_c||\boldsymbol{\psi}_n)$ increases substantially with the 4d model, resulting in shorter expected minimum lengths L_{min} and, generally, in shorter average lengths of actual predicted elements. With the exception of the vertebrates, these factors combine to produce somewhat higher coverage by conserved elements, of the entire genome and of coding regions in particular.

Oddly, however, the use of the 4d model causes the total coverage to drop for the vertebrates. Despite the drop in overall coverage, the CDS coverage stays roughly the same, so that the fraction of conserved bases in coding regions increases from 18% to 24%. This shift toward coding bases in conserved elements apparently occurs because (1) the decreased branch lengths in the primate/rodent subtree cause moderately conserved sequences present only in these species, which are primarily in noncoding regions, to be less likely to be considered conserved; and (2) the increased branch lengths to chicken and *Fugu* cause moderately conserved sequences present in all or most species, which are primarily in coding regions, to be more likely to be considered conserved. The longer average length of conserved elements for the vertebrates, despite the reduced $L_{\min}$, is probably a consequence of this shift toward coding sequences.

For the insects, worms, and yeasts, the increase in coverage suggests that the calibration procedure described in the text may be causing the fractions of these genomes that are conserved to be underestimated, by some 4% for the insects, 10% for the worms, and 12% for the yeasts, consistent with some other indications noted above. (Interestingly, for the insects and yeasts, the coverage levels based on the 4d model agree fairly well with those based on the m.l.e.'s.) Nevertheless, for the insects and yeasts, the fraction of conserved elements in coding regions stays roughly the same, indicating (as above) that by erring on the side of specificity (rather than sensitivity) in our main analysis, we have not substantially altered the composition of the set of conserved elements. The fraction of conserved elements in coding regions increases somewhat for the worms under the 4d model, apparently because the short nonconserved branch lengths and high degree of smoothing imposed by our main estimation procedure cause more potentially conserved bases to be omitted from conserved elements in noncoding regions than in coding regions.

In summary, our analysis based on 4d sites provides some indication that the procedure for parameter estimation reported in the main text may have led to underestimates of the fractions of insect, worm, and yeast genomes that are conserved, by roughly 4–12%. It also suggests the possibility that we have overestimated the fraction of vertebrate genomes that is conserved, but the evidence for this is weaker: it depends on the "true" neutral branch lengths for the primate/rodent subtree being closer to the ones estimated from 4d sites than to the larger estimates obtained from ancestral repeats and other noncoding regions, which is perhaps unlikely. Regardless, the composition of conserved elements by annotation type does not seem to be too strongly influenced by our methods for parameter estimation. Based on the 4d-site analysis, we still find vertebrates to have the lowest fraction of conserved elements in coding regions, followed by insects, then worms, then yeasts.

4.5 Sensitivity of HCEs to Methods

We have focused above on the sensitivity of the entire set of conserved elements on the parameter estimation methods, and have ignored the subset of highly conserved elements (HCEs), which are the focus of much of the paper. The reason for this emphasis is that, as might be expected, the set of HCEs is much less sensitive to parameter estimation methods than is the complete set of elements: most differences between the sets discussed above occur among elements that are moderately conserved. The set of HCEs does change somewhat from one method to another, mostly because differences in the degree of "smoothing" cause certain HCEs to be broken up in some cases and certain HCEs to be merged in others. The overall properties of the set of HCEs, however, appear to be highly robust.

4.6 False Positive and False Negative Rates

The issues examined above ultimately come down to questions of false negative and false positive rates: which methods do best at identifying elements that truly are conserved and at avoiding identifying elements that truly are not conserved? The difficulty in addressing these questions comes from the fact that we have no annotations of regions agreed to be conserved or nonconserved, nor even a consensus on what it means to be (or not be) "conserved." Even setting aside issues of mutation vs. selection, there are clearly gradations of conservation (i.e., of unexpectedly high similarity across species), and there is a large "gray area" of sequences that might reasonably be called conserved by one definition but not by another. Evaluating false negative rates is particularly difficult, because even known functional elements such as protein-coding or RNA genes contain many bases that are not conserved, and we do not in general have good information on exactly which bases these are. Thus, we have no set of known conserved sites against which to test our false negative rate (sensitivity).

We can, however, perform a very simple test of false positive rates (specificity) by asking the question: given a synthetic multiple alignment that approximates neutrally evolving DNA, what fraction of sites are predicted, by any given method, to be in conserved elements? This fraction should be representative of a certain kind of false positive rate: the rate of predictions that are completely "wrong," i.e., the rate at which predictions are made that consist completely of nonconserved sites. The results of this test will not, however, reflect false positive bases in the middle or at the extremities of legitimate conserved elements (see below), and they will not address the issue of sites in the conservation "gray area."

This test was performed for each species group, using a nonparameteric method based on 4d sites. Synthetic data sets were created by independently drawing alignment columns (with replacement) from a large collection of columns extracted from 4d sites, then concatenating these columns into multiple alignments. For each species group, a synthetic alignment of 1 million columns was created. False positive rates were then estimated by running phastCons on each synthetic alignment, using the parameters estimated from real data by three different methods: full maximum-likelihood estimation and estimation based on targets of 65% and 75% coverage of coding regions. Note that, while some fraction of 4d sites is presumably conserved, if this fraction is reasonably small, then the probability of encountering a sequence of such sites in the synthetic data set that is sufficiently long to support the prediction of a conserved element should be very small. Furthermore, the test will be conservative in that it should provide an upper bound on the false positive rate. (This will be true as long as no 4d sites, or very few 4d sites, evolve *faster* than the neutral rate; conceivably, hot spots of mutation or context-dependent substitution could produce some fast-evolving 4d sites.)

As shown in Table S7, the false positive rates estimated in this way turn out to be quite low—less than 0.4% in all cases but one (the insect group with the 75% coverage target, for which the rate was about 1%). As expected, the false positive rates are somewhat higher for the 75% coverage target than for the 65% coverage target. They tend to be very low under maximum likelihood estimation, apparently because of the high degree of "smoothing" imposed by the m.l.e.'s, which causes longer sequences of conserved sites to be required for a conserved element to be predicted. Clearly, however, these estimates do not tell the whole story. For example, as the degree of smoothing increases, the rate at which completely erroneous elements are predicted will decrease, but the rate at which short sequences of nonconserved bases will erroneously be incorporated into larger conserved elements will increase. Nevertheless, this test indicates that completely "wrong" predictions—i.e., apparent conserved elements that arise simply from random variation in nonconserved sites—are probably a negligible factor in our analysis.[2]

There are various problems with extending this simulation-based analysis to estimate false negative rates or false positive rates at the edges or in the middle of predicted elements. To address such questions, one would need to include conserved elements in the synthetic data sets, which would require making assumptions about the distribution of conserved alignment columns, the distribution of lengths of conserved elements, and the degree to which conserved elements tend to cluster together. We have little data on which to base these assumptions, especially if we want to avoid the circularity of basing them on results obtained with our own methods or methods similar to them.

One simple computation we can do is to find the expected maximum length of a sequence of nonconserved sites within a (sufficiently large) predicted conserved element—a kind of complement to $L_{\min}$ that we will call $L_{\max}$. Proceeding as in the derivation of $L_{\min}$, we obtain:

$$L_{\max} = \frac{\log \nu + \log \mu - \log(1-\nu) - \log(1-\mu)}{\log(1-\mu) - \log(1-\nu) - H(\boldsymbol{\psi}_n || \boldsymbol{\psi}_c)} \tag{8}$$

where $H(\boldsymbol{\psi}_n||\boldsymbol{\psi}_c) = \sum_x P(x|\boldsymbol{\psi}_n) \log \frac{P(x|\boldsymbol{\psi}_n)}{P(x|\boldsymbol{\psi}_c)}$ is the relative entropy of the distribution associated with the nonconserved model $\boldsymbol{\psi}_n$ with respect to the distribution associated with the conserved model $\boldsymbol{\psi}_c$. Values

[2] Note that this analysis—because it depends on an independence assumption—does not rule out the possibility that our results are significantly impacted by, say, cold spots of mutation, or local regions of hyper-efficient repair.

of $L_{\max}$ are shown in Table S8 for all four species groups and for three estimation methods. In general, $L_{\max}$ is somewhat smaller than $L_{\min}$ (cf. Table S5) but it is still substantial in size—e.g., sequences of as many as (an expected) 11 nonconserved bases could be included in a predicted conserved element for the vertebrate group and the 65% coverage target. In theory, even if there were no completely wrong predictions (as described above), the false positive rate at the base level could approach $\frac{L_{\max}}{L_{\min}+L_{\max}}$, or about 40% for the vertebrates and the 65% coverage target. In practice, however, this rate is likely to be much lower.

If there were sufficient data to determine whether or not individual bases were conserved, and if every base in a conserved element was required to be conserved, then false positive bases within conserved elements could be eliminated. It is not clear, however, that this would be ideal. For example, if conserved elements could not contain nonconserved bases, then most conserved elements in coding regions would be only two bases long. Perhaps it is better to tolerate some nonconserved bases, so that conserved elements correspond more directly to what we tend to think of as functional elements (e.g., protein-coding exons, RNA genes, and transcription-factor binding sites). Thus, some of the "false positives" implied by having $L_{\max} > 0$ (and some of the "false negatives" implied by having $L_{\min} > 1$) may actually be desirable. Still, the optimal value for $L_{\max}$ may well be smaller than the current values—perhaps 4 or 5 bases. It will be possible to achieve smaller values of $L_{\max}$ as more sequence data becomes available.

4.7 Inherent Differences Between Species Groups

Decreased levels of sensitivity and/or specificity are inevitable in species-poor groups (such as worm) as compared to species-rich groups (such as yeast). What happens under the calibration method used for the main analysis is that species-poor groups—or more generally, groups with low phylogenetic information—require higher levels of smoothing (larger $L_{\min}$ and $L_{\max}$), which causes more short conserved elements to be missed and more short intervals of nonconserved sites to be pulled into predicted conserved elements. Essentially, the conserved elements are redefined from one group to another, in an attempt to compensate for differences in numbers of species, phylogenetic difference, missing data, etc. For this reason, all comparisons between groups in our analysis must be interpreted with caution.

Apart from what is *predicted* to be conserved across species, what is *actually* conserved is also a function of evolutionary divergence. All other things being equal, species sets with more and more ancient most recent common ancestors will tend to have smaller and smaller sets of conserved elements, because of turnover of functional elements over evolutionary time (Smith et al., 2004). Thus, the correlation between conservation and function is partly determined by the extent to which such turnover has occurred, and differences in conserved elements between groups could reflect differences in turnover as well as differences in functional elements. This issue may particularly influence apparent differences in the fraction of conserved bases in coding versus noncoding regions. It is likely that turnover of certain noncoding functional elements, such as transcription factor binding sites, occurs at a higher rate than turnover of protein-coding sequences. As a result, the fraction of noncoding sites in conserved elements may be considerably higher for primates only, for example, than for all vertebrates or all mammals (Boffelli et al., 2003). Similarly, our sets of highly conserved elements (HCEs) are relative to the particular species we have considered. As the sequences of more genomes become available, it will be possible to begin to address the issue of turnover of conserved elements by running phastCons and programs like it on subsets of genomes, e.g., on just the primates.

5 Specification of Phylo-SCFGs for Folding Potential Scores

This section will provide a more detailed description of the two phylo-SCFGs (θ_{sp} and θ_{nsp}) used to calculate the folding potential scores. We start by defining the notation used to specify the phylo-SCFGs. The reader is referred to Durbin et al. (1998), Knudsen and Hein (1999), and Pedersen et al. (2004) for more information on the methodology of SCFGs and phylo-SCFGs in particular.

An SCFG can be specified by a three-tuple $\theta = (W, t, e)$, where W denotes a set of states, t denotes a set of transition probabilities, and e denotes a set of emission distributions. It is convenient for algorithmic reasons to define a set of state types which determines the emission and transition properties of states. We will here make use of the state types proposed for grammars describing RNA secondary structures (Durbin et al., 1998a): pair-emitting (P), left-emitting (L), right-emitting (R), start (S), bifurcate (B), and end (E). Each emitting state has an associated emission distribution. The pair-emitting state emits two correlated symbols, while the left-emitting and the right-emitting states emit a single symbol. The emission distributions of phylo-SCFGs are specified by a set of phylogenetic models. We will start below by defining the transition graphs (given by W, t) of θ_{sp} and θ_{nsp} and then will define their phylogenetic models.

5.1 The Transition Graphs

The transition graph of θ_{sp} is very similar to the transition graph of the phylo-SCFG employed by RNA-decoder (Pedersen et al., 2004). It can be decomposed into three components: pairing, non-pairing, and high-level. The pairing component contains a pair-emitting state and is capable of modeling arbitrary RNA secondary structures (see right part of Figure S6). The non-pairing component contains a single left-emitting state and can therefore only model non-pairing regions (see left part of figure S6). The high-level component models the sequence of non-pairing and pairing regions (i.e. regions containing RNA secondary structure). The high-level component is identical to the high-level sub-grammar used by RNA-decoder and therefore is not shown here.

The transition graph of θ_{nsp} consists solely of the non-pairing component.

5.2 The Phylogenetic Models

The phylo-SCFGs have three emitting states: one pair-emitting (in the pairing component), and two left-emitting (one in each of the pairing and non-pairing components). The two left-emitting states use the same emission distribution. The phylo-SCFGs therefore only employ two phylogenetic models: a dinucleotide model (ψ^{di}) in the pair emitting state and a single-nucleotide model (ψ^{s}) in the two left-emitting states. The phylogenetic models are specified by a four-tuple given by $\psi = (Q, \pi, \tau, \beta)$ as explained for the phylo-HMM used by phastCons. The two phylogenetic models use the same (pre-estimated) phylogenetic tree (i.e., the same tree topology and branch lengths).

The dinucleotide rate matrix is parameterized by the general-time reversible model (REV), but with a number of added constraints to lower the number of free parameters: the rates of substitution to all non-pairing dinucleotides (the following dinucleotides are pairing in RNA: A-U, U-A, G-C, C-G, G-U, and U-G) are fixed at a constant value. A similar approach is taken with the equilibrium distribution: all non-pairing dinucleotides are fixed at a small constant value ($\sim 10^{-9}$). This approach lowers the number of free parameters in the parameterization of Q^{di} and π^{di} from 134 to 19, and has the effect of making the rate of substitution from pairing to non-pairing dinucleotides very low.

The substitution process of the single-nucleotide model is calculated as an average of the substitution process observed in the left and right position of the dinucleotide model. This is done by calculating a left and a right marginalized version of π^{di} as well as Q^{di} (Yang et al., 1998) and then defining π^{s} and Q^{s} as the average of these. This approach was chosen in order to remove any differences between the pairing component and the non-pairing component apart from the characteristics unique to RNA secondary structures: correlation between the bases as well as their substitution pattern in regions which form functional stems.

5.3 Algorithms

5.3.1 Parameter Estimation

The parameters of the phylo-SCFGs were estimated from a training set consisting of 150 multiz alignments (Blanchette et al., 2004) of sequences from eight species of vertebrates. Each alignment consists of a phastCons conserved element overlapping a well-conserved functional RNA derived from the Rfam database (Griffiths-Jones et al., 2003). The alignments were annotated with the RNA secondary structures given by Rfam.

Only the parameters of θ_{sp} need to be estimated, since θ_{nsp} is defined as the non-pairing component of θ_{sp}. The free parameters consist of: the set of state-transition probabilities, the rate matrix of ψ^{di}, and the equilibrium frequencies of ψ^{di}. τ and β, which specify the phylogenetic tree, are not treated as free parameters, but given by the nonconserved phylogeny estimated by phastCons, which was scaled to reflect the overall rate of substitution in the non-pairing regions of the training set. The free parameters were estimated using two different maximum likelihood approaches. The transition probabilities were estimated using the inside-outside algorithm (Durbin et al., 1998a). The rate matrix and the equilibrium frequencies of ψ^p were estimated using the BFGS algorithm, as implemented in the OPT++ package (Meza, 1994).

5.3.2 Likelihood Calculations

The likelihood of an alignment x given a phylo-SCFG is the sum over all possible derivations from the grammar which could lead to x. Each derivation corresponds to an annotation y of the alignment; the likelihood can therefore be written: $P(x|\theta) = \sum_y P(x, y|\theta)$. The inside algorithm (Durbin et al., 1998a) efficiently performs this summation. Only a single derivation is possible from the simple grammar of θ_{nsp}; this derivation annotates every position as being non-pairing. A huge number of different derivations are possible from θ_{sp}, each corresponding to an annotation of an RNA secondary structure.

The folding potential score (FPS) of a given alignment x is defined as a log likelihood ratio between the two models: $s(x) = \log P(x|\theta_{sp}) - \log P(x|\theta_{nsp})$. Little evolutionary information is present in alignments when only few sequences are present and spurious RNA secondary structures will often be predicted in these cases (Rivas and Eddy, 2000; Pedersen et al., 2004). Alignments are therefore assigned a partial annotation, which forces positions with more than 50% gaps or missing data to be treated as non-pairing. The likelihood then becomes a sum over all structures compatible with the given partial annotation.

6 Results of FPS Analysis

Sets of local alignments corresponding to the different classes of HCE elements were defined and their distributions of FPSs were compared. The alignments correspond to tilings of the HCEs with overlapping 150bp long intervals (step size 50bp). Null sets of UTR elements were constructed for comparative purposes (see Methods). Summary statistics of the FPS distributions and results of tests for statistical differences between the distributions are presented below for both the vertebrate sets and the insect sets. Interpretation of the results is given in the text of the article.

6.1 Vertebrates

The mean, the median, and the score at the 95th percentile are calculated for each set (see first part of Table S9). Summary statistics were also calculated for the reverse complements of the UTR sets (see second part of Table S9). Note that all sets, apart from the intergenic, contain strand-specific alignments.

Presence of CpG islands in $5'$ UTRs has a big effect on their FPSs. UTR sets which exclude all alignments overlapping a CpG-island (as defined by the CpG-island track of the UCSC human genome browser) were therefore constructed and their summary statistics calculated (see Table S10).

Differences between the overall distributions of FPSs between the sets were measured (Wilcoxon rank sum test). We specifically test if one distribution is right-shifted relative to another and thus has a stronger signal for presence of functional RNA secondary structures. In order to fulfill the independence assumption of the test, we divide each set into three subsets of non-overlapping alignments. (Because the 150bp windows overlap by 50bp, sets consisting of every third window do not contain overlaps.) The comparison of two sets (A and B) is therefore composed of three pair-wise comparisons of subsets (A_1 vs. B_1, A_2 vs. B_2, and A_3 vs. B_3). These three subset tests are correlated; P-values reported in the main text are the least significant of the three. The test results of comparisons between different HCE sets as well as between UTR HCE sets and UTR null sets are reported in Table S11. Tests involving $5'$ UTR sets excluding CpG islands are reported in Table S12.

A second series of tests was performed in which FPSs for vertebrate $3'$ and $5'$ UTRs were compared to FPSs for their reverse complements. Because RNA secondary structures are not completely strand symmetric (G can pair with U, but C cannot pair with A), a significant difference between the distributions of FPSs for the forward and reverse orientations should be observed in the presence of an enrichment for true structures. This was tested for each of the four UTR sets (see Table S13). These tests supported the existence of a strong enrichment for secondary structure in $3'$ UTR HCEs, a lesser but still strong enrichment in $3'$ UTRs without HCEs, and a slight enrichment in $5'$ UTR HCEs (on the border of statistical significance). No enrichment was seen in $5'$ UTRs without HCEs. Note that the order of the comparison is switched for the CDS set, which has a significantly stronger signal for presence of functional RNA structures on the reverse strand.

6.2 Insect

The summary statistic for the insect sets are given in Table S14. The results of the comparisons of FPS distributions of different insect sets are given in Table S15. Note that neither the reverse complement sets nor the UTR sets excluding CpG-islands were generated for the insects.

Tables

Table S1: Summary of genomes and assemblies

Species	Group	UCSC assembly	Reference
H. sapiens	vertebrate	hg17	International Human Genome Sequencing Consortium, 2001
M. musculus	vertebrate	mm5	Mouse Genome Sequencing Consortium, 2002
R. norvegicus	vertebrate	rn3	Rat Genome Sequencing Project Consortium, 2004
G. gallus	vertebrate	galGal2	International Chicken Genome Sequencing Consortium, 2004
F. rubripes	vertebrate	fr1	Aparicio et al., 2002
D. melanogaster	insect	dm1	Adams et al., 2000
D. yakuba	insect	droYak1	(In prep.)
D. pseudoobscura	insect	dp2	Richards et al., 2005
A. gambiae	insect	anoGam1	Holt et al., 2002
C. elegans	worm	ce2	C. elegans Sequencing Consortium, 1998
C. briggsae	worm	cb1	Stein et al., 2003
S. cerevisiae	yeast	sacCer1	http://www.yeastgenome.org
S. castelli	yeast	–	Cliften et al., 2003
S. kluyveri	yeast	–	Cliften et al., 2003
S. kudriavzevii	yeast	–	Cliften et al., 2003
S. mikatae	yeast	–	Kellis et al., 2003
S. bayanus	yeast	–	Kellis et al., 2003
S. paradoxus	yeast	–	Kellis et al., 2003

Table S2: Summary of multiple alignments

Group	n[a]	Reference Genome	Aligned Genomes (Coverage)[b]	Tot. Cov.[c]
vertebrate	5	*H. sapiens*	*M. musculus* (35.4%), *R. norvegicus* (34.0%), *G. gallus* (3.5%), *F. rubripes* (1.6%)	40.0%
insect	4	*D. melanogaster*	*D. yakuba* (85.1%), *D. pseudoobscura* (58.1%), *A. gambiae* (14.6%)	86.9%
worm	2	*C. elegans*	*C. briggsae* (43.8%)	43.8%
yeast	7	*S. cerevisiae*	*S. paradoxus* (96.6%), *S. mikatae* (93.0%), *S. kudriavzevii* (89.7%), *S. bayanus* (89.5%), *S. castelli* (63.4%), *S. kluyveri* (56.1%)	96.6%

[a] Number of species.

[b] Genomes aligned to reference genome and fraction of bases in reference genome covered by local alignments with each one.

[c] Fraction of bases in reference genome covered by alignments with at least one other genome.

Table S3: Some Human Genes Overlapped by Highly Conserved Elements

Gene	N[a]	Lengths[b]	Regions Overlapped	Product[c]
NOVA1	1	4922	CDS, 3′ UTR	RNA-binding protein, may regulate neuron-specific splicing. Associated with breast, fallopian, and lung cancer.
ARHGAP5	1	3738	CDS, 5′ UTR	Rho GTPase-activating protein. Negatively regulates RHO GTPases, which mediate cytoskeleton changes by stimulating hydrolysis of bound GTP.
PURA	1	3211	CDS, UTRs	Sequence-specific DNA-binding protein, implicated in the control of DNA replication and transcription. Associated with myelodysplastic syndrome and acute myelogenous leukemia.
ELAVL4	3	691–3117	CDS, UTRs	RNA-binding protein, binds to 3′ UTRs, may play a role in neural development.
FOXG1	1	2634	CDS, introns, UTRs	Forkhead transcription factor, may play a role in brain development.
ZFHX1B	23	351–2476	CDS, introns, 3′ UTR	Zinc-finger/homeodomain protein, probable transcriptional repressor, mutated in Hirschsprung disease syndrome.
BCL11A	4	740–2451	CDS, introns, 3′ UTR	Zinc-finger protein, associated with leukemia.
SYNCRIP	2	798–2369	CDS, 3′ UTR	RNA-binding protein, involved in RNA processing and transport. May modulate RNA-editing of APOB.
NIPBL	3	644–2106	CDS, UTRs	Probably plays a role in developmental regulation, also involved in sister chromatid cohesion and probably places a structural role in chromatin. Mutated in Cornelia de Lange syndrome.
ATBF1	7	738–2030	CDS, introns, 3′ UTR	AT-motif-binding transcriptional activator of alpha-fetoprotein (AFP) gene, multiple homeodomains and zinc finger motifs.
FMR1	3	593–1548	CDS, intron, 3′ UTR	RNA-binding protein, possibly involved in the transport of mRNA from the nucleus to the cytoplasm. Mutated in fragile X syndrome.
UBE3A	2	729–1261	CDS, 3′ UTR	E3 ubiquitin-protein ligase, part of the ubiquitin protein degradation system. Maternally inherited deletion of this imprinted gene causes Angelman Syndrome.
CPEB2	3	745–1260	3′ UTR	Highly similar to *CPEB*, an mRNA-binding protein that binds to the cytoplasmic polyadenylation element and modulates translational repression and mRNA localization. *CPEB1*, *CPEB3*, and *CPEB4* also have HCEs in their 3′ UTRs.
FOXP2	24	441–1056	CDS, introns, UTRs	Forkhead transcription factor, mutant in developmental verbal dyspraxia, possible "speech gene."
SHH	1	793	CDS, 3′ UTR	Signaling molecule homologous to the sonic hedgehog development gene in *Drosophila*, which is critical in embryogenesis.
MYC	1	789	CDS, 3′ UTR	DNA-binding protein and transcriptional activator, involved by translocation in the generation of Burkitt lymphoma.
PAX6	5	435–733	CDS, introns, 3′ UTR	Paired-box-containing transcription factor, expressed in developing nervous system and in developing eyes. Mutations cause the ocular diseases aniridia and Peter's anomaly.

[a] Number of HCEs overlapping gene.

[b] Lengths of HCEs in human (bp).

[c] Based on descriptions in RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq), SwissProt (http://us.expasy.org/sprot), and OMIM (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM). See these resources for primary references.

Table S4: Selected Gene Ontology (GO) Categories of Insect, Worm, and Yeast Genes Overlapped by Highly Conserved Elements (CDS only)

Group	Term	Description	N^a	exp.[b]	obs.[c]	P^d
insect	GO:0000166	nucleotide binding	762	234.0	423	2.1e-51
	GO:0007275	development	1214	372.8	575	3.6e-41
	GO:0009653	morphogenesis	702	215.5	377	8.1e-41
	GO:0009887	organogenesis	605	185.8	330	2.8e-37
	GO:0007552	metamorphosis	281	86.2	161	4.4e-21
	GO:0007399	neurogenesis	296	90.9	167	7.1e-21
	GO:0016462	pyrophosphatase activity	290	89.0	164	1.2e-20
	GO:0005515	protein binding	496	152.3	244	3.5e-19
	GO:0007267	cell-cell signaling	131	40.2	82	3.1e-14
	GO:0007268	synaptic transmission	119	36.5	76	5.9e-14
	GO:0030154	cell differentiation	237	72.7	125	6.0e-13
	GO:0004672	protein kinase activity	223	68.4	119	7.7e-13
	GO:0016310	phosphorylation	279	85.6	141	1.5e-12
	GO:0007017	microtubule-based process	90	27.6	58	3.4e-11
	GO:0048477	oogenesis	206	63.2	102	8.3e-09
	GO:0030528	transcription regulator activity	324	99.5	148	4.9e-09
	GO:0019953	sexual reproduction	300	92.1	133	2.7e-07
	GO:0004386	helicase activity	78	23.9	44	2.0e-06
	GO:0005244	voltage-gated ion channel activity	31	9.5	22	4.5e-06
worm	GO:0042302	structural constituent of cuticle	156	11.7	71	3.9e-39
	GO:0006811	ion transport	603	45.3	113	5.6e-21
	GO:0000166	nucleotide binding	1012	76.0	139	1.8e-13
	GO:0006520	amino acid metabolism	111	8.3	27	3.1e-08
	GO:0007155	cell adhesion	63	4.7	19	9.0e-08
	GO:0006412	protein biosynthesis	290	21.7	48	1.3e-07
	GO:0008158	hedgehog receptor activity	28	2.1	11	2.6e-06
	GO:0016787	hydrolase activity	1193	89.6	129	6.1e-06
	GO:0009451	RNA modification	37	2.7	11	5.6e-05
yeast	GO:0000166	nucleotide binding	619	98.1	251	9.7e-58
	GO:0003735	structural constituent of ribosome	201	31.8	89	1.4e-22
	GO:0004386	helicase activity	94	14.9	53	1.1e-19
	GO:0006520	amino acid metabolism	222	35.2	85	9.5e-17
	GO:0016787	hydrolase activity	765	121.3	196	2.4e-14
	GO:0006096	glycolysis	32	5.0	22	2.6e-11
	GO:0006006	glucose metabolism	75	11.8	32	2.3e-08
	GO:0016301	kinase activity	216	34.2	61	1.7e-06
	GO:0003723	RNA binding	335	53.1	85	2.3e-06
	GO:0000287	magnesium ion binding	86	13.6	31	3.3e-06
	GO:0009451	RNA modification	85	13.4	30	7.8e-06
	GO:0016310	phosphorylation	190	30.1	53	1.2e-05
	GO:0006333	chromatin assembly or disassembly	32	5.0	15	3.5e-05

[a] Number of genes in background set assigned to category.

[b] Expected number of genes overlapped.

[c] Observed number of genes overlapped.

[d] P-value. Values of less than 5e−5 can be considered significant (see Methods).

Table S5: Predicted Conserved Elements and Estimated Parameters Under Four Different Estimation Methods

Group	Method	Total no.[a]	Ave. len.[b]	Cov.[c]	CDS cov.[d]	μ	ν	ω	γ	$L_{\min}$
vert.	MLE	561,103	216.1	4.2%	68.8%	0.018	0.004	55.4	0.191	30.4
	55%	1,058,855	75.3	2.8%	56.8%	0.125	0.029	8.0	0.187	12.9
	65%[e]	1,157,180	103.5	4.2%	66.1%	0.083	0.030	12.0	0.265	16.0
	75%	1,381,978	167.5	8.1%	76.6%	0.043	0.031	23.0	0.415	22.6
insect	MLE	352,620	173.1	48.3%	73.3%	0.018	0.019	56.1	0.522	17.1
	55%	502,974	92.9	36.9%	57.0%	0.050	0.029	20.0	0.370	13.5
	65%	467,232	120.6	44.5%	68.3%	0.036	0.031	28.1	0.468	13.5
	75%	427,815	156.9	53.1%	78.4%	0.025	0.039	40.0	0.610	12.7
worm	MLE	71,419	258.8	18.4%	43.1%	0.017	0.014	58.6	0.455	83.7
	55%	108,588	181.2	19.6%	45.8%	0.037	0.033	27.0	0.470	56.4
	65%	98,415	268.9	26.4%	56.4%	0.019	0.031	53.0	0.620	60.6
	75%	87,228	357.5	31.1%	62.5%	0.010	0.030	100.0	0.750	65.6
yeast	MLE	57,610	134.1	63.6%	77.8%	0.021	0.034	47.1	0.615	11.5
	55%	71,726	78.8	46.5%	58.0%	0.067	0.023	15.0	0.260	12.5
	65%	62,640	107.9	55.6%	68.9%	0.043	0.029	23.0	0.400	11.7
	75%	62,754	124.1	64.0%	78.1%	0.025	0.041	40.0	0.620	10.6

[a] Total number of predicted conserved elements; here and below, only elements passing synteny filters are considered (where applicable)

[b] Average length of a predicted conserved element (bp)

[c] Genome-wide coverage by predicted conserved elements

[d] Coverage of coding regions by predicted conserved elements. These numbers differ slightly from the coverage targets (55%, 65%, and 75%) because of the adjustment for alignment coverage in coding regions and because of the removal of nonsyntenic predictions.

[e] Slight differences with the numbers reported in the main text result from the use of only a portion of the genome in parameter estimation (a random sample of 100 1Mb intervals).

Table S6: Conserved Elements Based on PhastCons Nonconserved Model (65% Coverage Target) vs. Elements Based on 4d Model

Group	Method	Total no.[a]	Ave. len.[b]	Cov.[c]	CDS cov.[d]	CDS frac.[e]	$H(\boldsymbol{\psi}_c \| \boldsymbol{\psi}_n)$	$L_{\min}$
vert.	65%	1,157,180	103.5	4.2%	66.1%	18.0%	0.611	16.0
	4d	797,777	109.3	3.0%	64.2%	24.0%	0.854	11.0
insect	65%	467,232	120.6	44.5%	68.3%	26.4%	0.725	13.5
	4d	554,823	110.0	48.3%	75.0%	26.8%	1.032	9.5
worm	65%	98,415	268.9	26.4%	56.4%	54.9%	0.159	60.6
	4d	195,062	188.0	36.6%	69.3%	48.7%	0.403	25.4
yeast	65%	62,640	107.9	55.6%	68.9%	86.1%	0.836	11.7
	4d	94,615	86.8	67.6%	81.8%	84.1%	1.914	5.3

[a] Total number of predicted conserved elements; here and below, only elements passing synteny filters are considered (where applicable)

[b] Average length of a predicted conserved element (bp)

[c] Genome-wide coverage by predicted conserved elements

[d] Coverage of coding regions by predicted conserved elements.

[e] Fraction of conserved elements in coding regions, at the base level

Table S7: Estimated False-Positive Rates for PhastCons Under Three Parameter Estimation Methods (Nonparameteric Test Based on 4d Sites)

Group	65%	75%	MLE
vertebrate	0.00279[a]	0.00362	0.00005
insect	0.00286	0.01026	0.00152
worm	0.00000	0.00000	0.00000
yeast	0.00006	0.00042	0.00023

[a] False-positive rates at the base level; 0.00279 means that an average of 2.79 out of every 1,000 nonconserved bases were incorrectly predicted as belonging to conserved elements.

Table S8: Expected Maximum Length of a Sequence of Nonconserved Sites Within a Predicted Conserved Element (L_{max})

Group	65%	75%	MLE
vertebrate	11.0	19.1	25.3
insect	11.5	12.0	15.3
worm	64.1	83.9	67.3
yeast	10.3	10.4	10.6

Table S9: Summary Statistics for FPSs of HCE Subsets and UTR Null-Sets

		original			*reverse complement*		
data set	size^a	mean	median	95%^{b}	mean	median	95%^{b}
$3'$ UTR HCE	5614	0.31	0.05	2.69	0.13	−0.14	2.29
$5'$ UTR HCE	689	−0.20	−0.43	1.70	−0.24	−0.51	1.74
$3'$ UTR null	35651	−0.17	−0.41	2.09	−0.29	−0.52	1.84
$5'$ UTR null	6326	−0.20	−0.44	2.11	−0.24	−0.48	2.01
CDS HCE	16081	−0.33	−0.59	1.60	–	–	–
intron HCE	10679	0.30	0.08	2.36	–	–	–
intergenic HCE	36594	0.19	0.00	2.15	–	–	–

[a] Number of overlapping 150bp windows tested.

[b] Score at the 95th percentile.

Table S10: Summary Statistics for FPSs of UTR Sets Excluding CpG Islands

		original			*reverse complement*		
data set	size^a	mean	median	95%^{b}	mean	median	95%^{b}
$3'$ UTR HCE	5524	0.32	0.06	2.70	0.13	−0.14	2.130
$5'$ UTR HCE	319	−0.18	−0.41	1.89	−0.32	−0.58	1.53
$3'$ UTR null	34846	−0.17	−0.40	2.09	−0.29	−0.52	1.84
$5'$ UTR null	3453	−0.32	−0.58	1.93	−0.34	−0.58	1.83

[a] Number of overlapping 150bp windows tested.

[b] Score at the 95th percentile.

Table S11: Tests of Distribution Differences Between FPSs of Different HCE Sets

		subset sizes						P-values [a]		
set A	set B	A_1	A_2	A_3	B_1	B_2	B_3	test 1	test 2	test 3
3′ UTR HCE	3′ UTR null	2065	1861	1688	13235	11790	10626	1.70e-67	2.36e-71	8.75e-66
5′ UTR HCE	5′ UTR null	297	215	177	2949	1994	1383	0.26	0.16	0.14
3′ UTR HCE	5′ UTR HCE	2065	1861	1688	297	215	177	5.33e-13	3.68e-10	1.14e-8
3′ UTR null	5′ UTR null	13235	11790	10626	2949	1994	1383	2.80e-2	6.80e-2	3.08e-2
3′ UTR HCE	intergenic HCE	2065	1861	1688	6632	6632	6092	3.16e-2	4.43e-2	5.55e-2
introns HCE	3′ UTR HCE	3862	3545	3241	2065	1861	1688	0.13	3.52e-2	2.36e-2
introns HCE	5′ UTR HCE	3862	3545	3241	297	215	177	8.09e-17	9.05e-14	6.90e-12
intergenic HCE	5′ UTR HCE	6632	6632	6092	297	215	177	5.04e-13	3.16e-09	1.06e-8
3′ UTR HCE	CDS HCE	2065	1861	1688	5746	5339	4937	2.36e-116	1.80e-113	7.14e-109
5′ UTR HCE	CDS HCE	297	215	177	5746	5339	4937	5.80e-3	4.49e-3	8.56e-3
introns HCE	CDS HCE	3862	3545	3241	5746	5339	4937	4.43e-215	2.99e-224	2.59e-217
intergenic HCE	CDS HCE	6632	6632	6092	5746	5339	4937	9.28e-227	7.39e-221	7.50e-233

[a] The null hypothesis is that set A is not right-shifted relative to set B. A significant P-value therefore indicates that set A is enriched for RNA structures compared to set B. Three tests are performed as explained in the text.

Table S12: Tests of Distribution Differences Between FPSs of UTR Sets Excluding CpG Islands

		subset sizes						P-values [a]		
set A	set B	A_1	A_2	A_3	B_1	B_2	B_3	test 1	test 2	test 3
5′ UTR null	RC[b] 5′ UTR null	1601	1083	769	1601	1083	769	0.22	0.78	0.52
5′ UTR HCE	RC 5′ UTR HCE	138	98	83	138	98	83	0.30	3.66e-2	0.14
5′ UTR HCE	5′ UTR null	138	98	83	1601	1083	769	4.67e-2	1.75e-2	1.19e-2
3′ UTR HCE	5′ UTR HCE	2031	1832	1661	138	98	83	5.57e-8	2.74e-5	2.28e-4
3′ UTR null	5′ UTR null	12937	11526	10383	1601	1083	769	3.42e-9	3.52e-7	3.93e-6

[a] The null hypothesis is that set A is not right-shifted relative to set B. A significant P-value therefore indicates that set A is enriched for RNA structures compared to set B. Three tests are performed as explained in the text.

[b] reverse complement

Table S13: Tests of Distribution Differences Between FPSs of UTR Sets and Their Reverse Complements

		subset sizes						P-values [a]		
set A	set B	A_1	A_2	A_3	B_1	B_2	B_3	test 1	test 2	test 3
3′ UTR HCE	RC[b] 3′ UTR HCE	2065	1861	1688	2065	1861	1688	1.22e-6	1.77e-6	2.82e-07
5′ UTR HCE	RC 5′ UTR HCE	297	215	177	297	215	177	0.28	0.19	0.29
3′ UTR null	RC 3′ UTR null	13235	11790	10626	13235	11790	10626	3.02e-16	3.99e-17	1.61e-17
5′ UTR null	RC 5′ UTR null	2949	1994	1383	2949	1994	1383	3.98e-2	0.31	0.39

[a] The null hypothesis is that set A is not right-shifted relative to set B. A significant P-value therefore indicates that set A is enriched for RNA structures compared to set B. Three tests are performed as explained in the text.

[b] reverse complement.

Table S14: Summary Statistics for FPSs of HCE Subsets and UTR Null-Sets

data set	size	mean	median	[a]95%
3′ UTR HCE	971	-0.55	-0.71	1.33
5′ UTR HCE	93	-1.04	-1.34	0.42
3′ UTR null	25699	-0.59	-0.84	1.81
5′ UTR null	10734	-0.86	-1.13	1.28
CDS HCE	43070	-0.49	-0.70	1.14
intron HCE	928	-0.86	-1.02	0.70
intergenic HCE	5782	-0.95	-1.10	0.82

[a] score at the 95th percentile

Table S15: Tests of Distribution Differences Between FPSs of Different HCE Subsets and UTR Null-Sets

		subset sizes						*P*-values [a]		
set A[b]	set B	A_1	A_2	A_3	B_1	B_2	B_3	test 1	test 2	test 3
3′ UTR HCE	3′ UTR null	373	291	239	9823	8469	7407	1.66e-3	9.54e-3	1.78e-2
5′ UTR HCE	5′ UTR null	43	30	20	4567	3456	2711	0.96	0.71	0.48
3′ UTR HCE	5′ UTR HCE	373	291	239	43	30	20	1.98e-6	2.24e-3	5.70e-2
3′ UTR null	5′ UTR null	9823	8469	7407	4567	3456	2711	1.28e-38	1.12e-34	1.24e-29
3′ UTR HCE	intergenic HCE	373	291	239	1027	1027	961	6.91e-15	6.07e-12	1.53e-10
intron HCE	3′ UTR HCE	335	311	282	373	291	239	1.00	1.00	1.00
intron HCE	5′ UTR HCE	335	311	282	43	30	20	1.74e-3	8.00e-2	0.37
intergenic HCE	5′ UTR HCE	1027	1027	961	43	30	20	5.70e-2	3.72	0.68
3′ UTR HCE	CDS HCE	373	291	239	15835	14344	12861	0.91	0.93	0.93
5′ UTR HCE	CDS HCE	43	30	20	15835	14344	12861	1.00	1.00	0.99
intron HCE	CDS HCE	335	311	282	15835	14344	12861	1.00	1.00	1.00
intergenic HCE	CDS HCE	1027	1027	961	15835	14344	12861	1.00	1.00	1.00

[a] The null hypothesis is that set A is not right-shifted relative to set B. A significant *P*-value therefore indicates that set A is enriched for RNA structures compared to set B. Three tests are performed as explained in the text.

[b] The order of the comparisons have been kept the same as for vertebrates. This has resulted in some *P*-values being close to one.

Figures

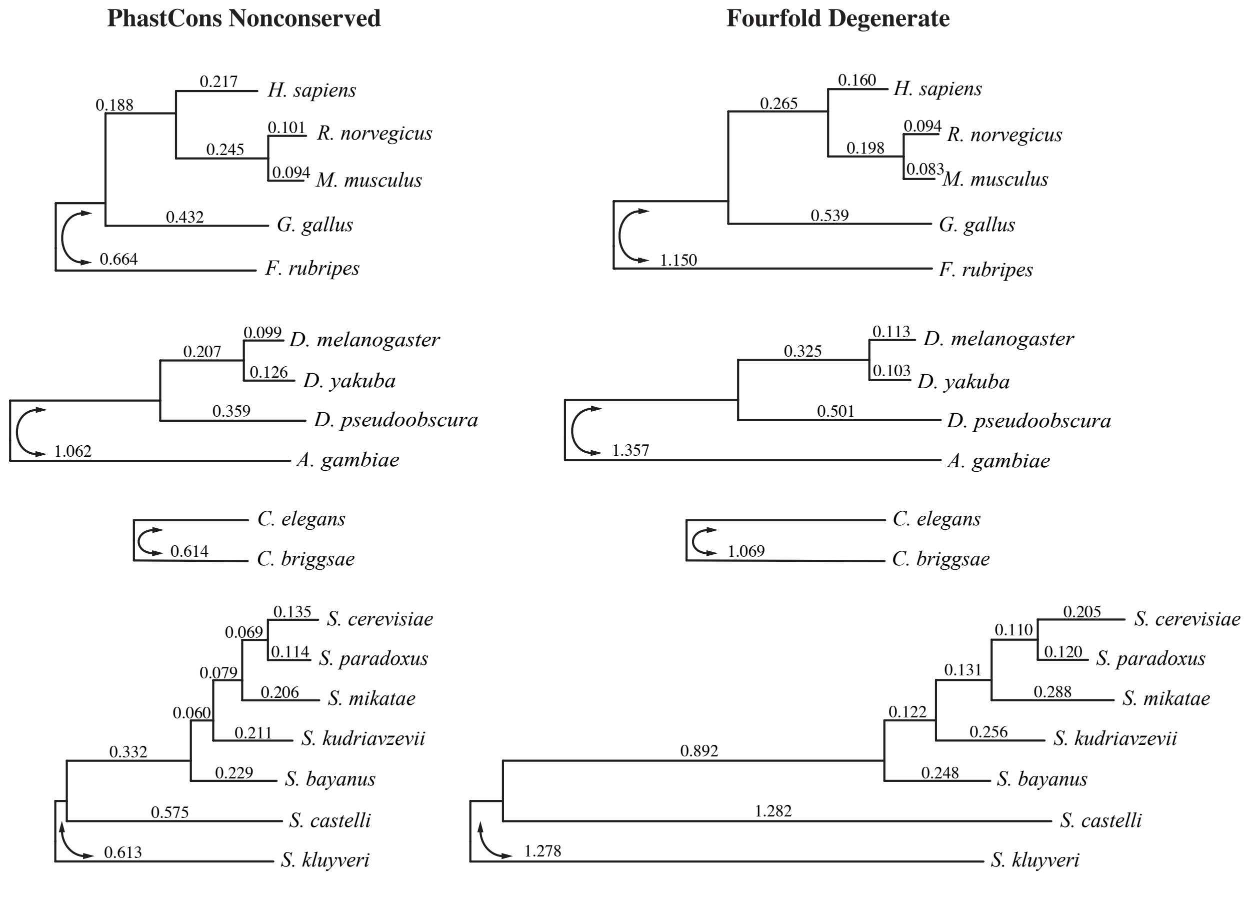

Figure S1: Nonconserved phylogenies estimated by phastCons (left) compared to phylogenies estimated from fourfold degenerate sites (right). Horizontal distances are to scale.

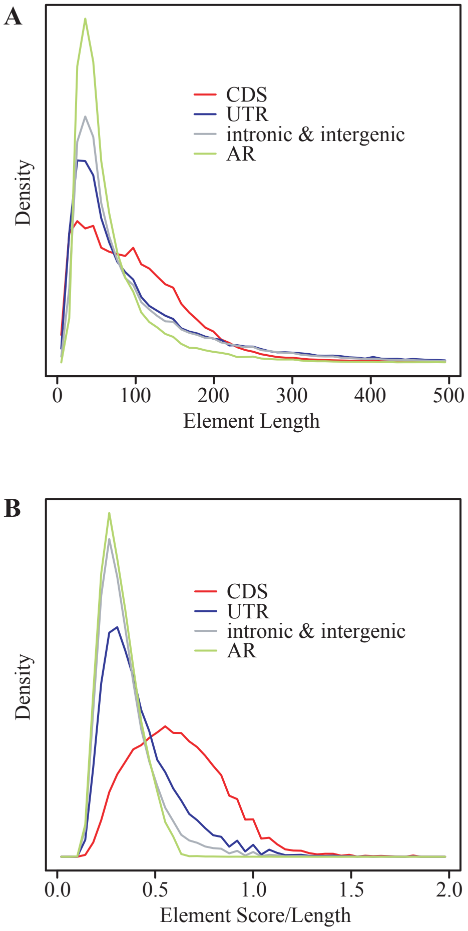

Figure S2: Distributions of (A) lengths and (B) length-normalized log-odds scores (score/length) of vertebrate conserved elements that primarily overlap ($>$ 50% of bases) known CDSs, UTRs, introns or intergenic regions, and ancestral repeats (ARs). The distributions for $5'$ and $3'$ UTRs and for introns and intergenic regions were essentially indistinguishable at this resolution. The entire distributions are not shown (the tails have been truncated).

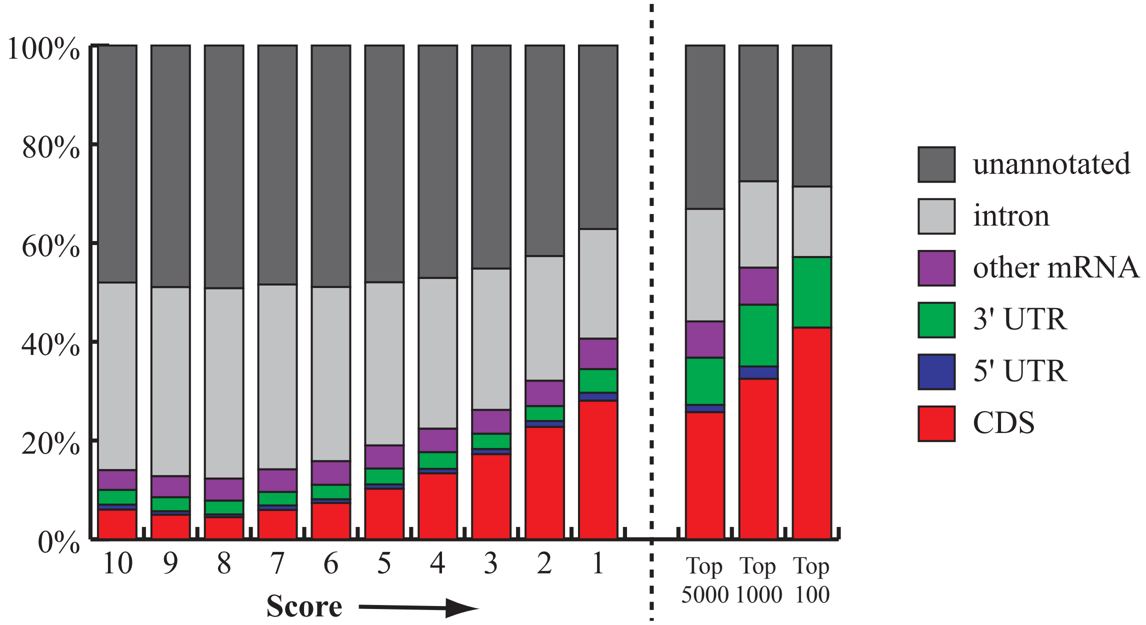

Figure S3: Coverage of vertebrate conserved elements by feature type for various score classes. The numbers 1–10 indicate disjoint classes of equal numbers of conserved elements, ranging from the highest scoring 10% of elements (1) to the lowest scoring 10% of elements (10). Subsets of class 1 consisting of the top-scoring 5000, 1000, and 100 elements are also shown.

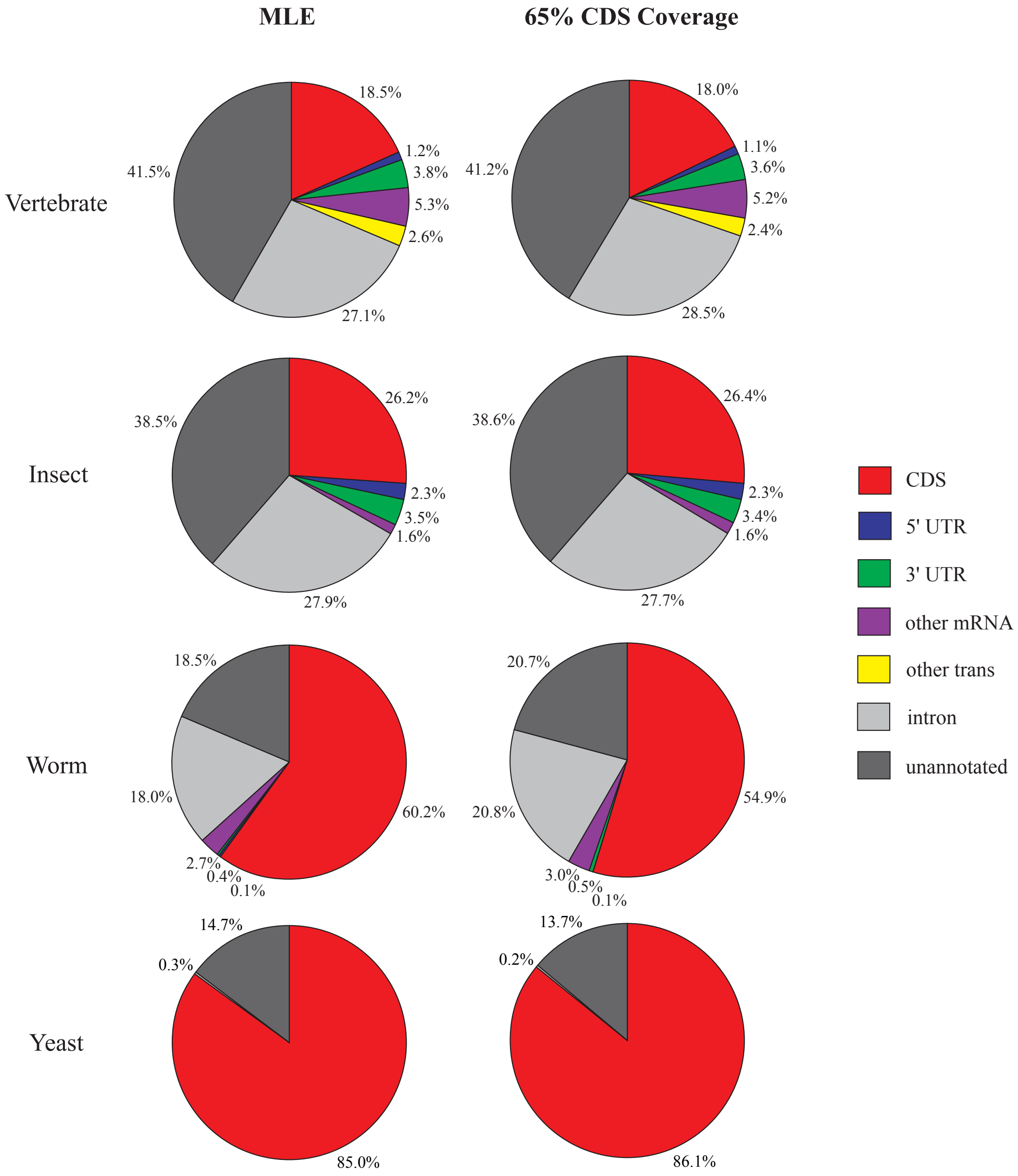

Figure S4: Composition of predicted conserved elements for all four species groups under two parameter estimation methods for μ and ν (γ and ω): maximum likelihood estimation (left column) and the method described in the text, with a target of 65% coverage of coding regions by conserved elements.

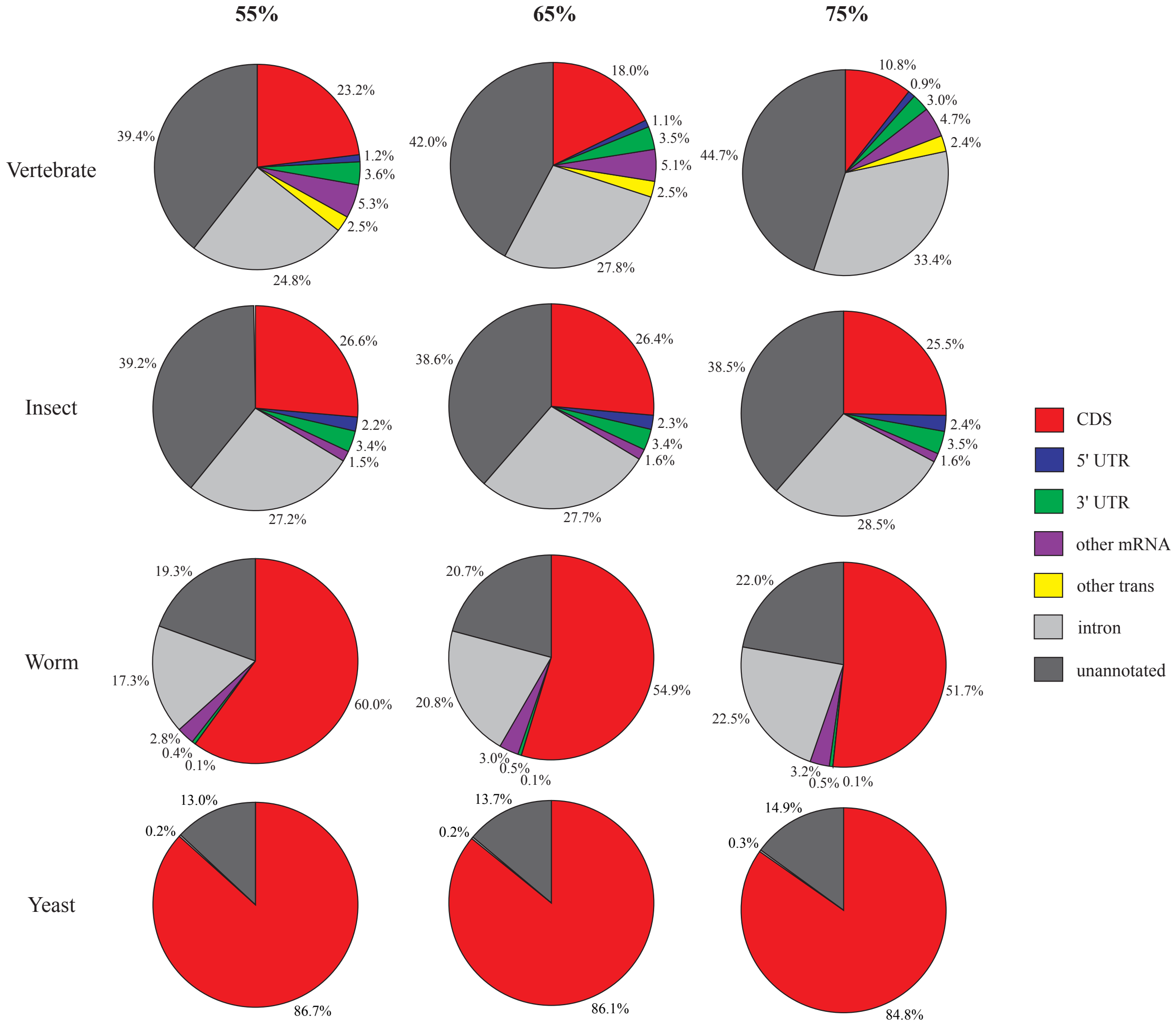

Figure S5: Composition of predicted conserved elements for all four species groups with target CDS coverages of 55%, 65%, and 75%.

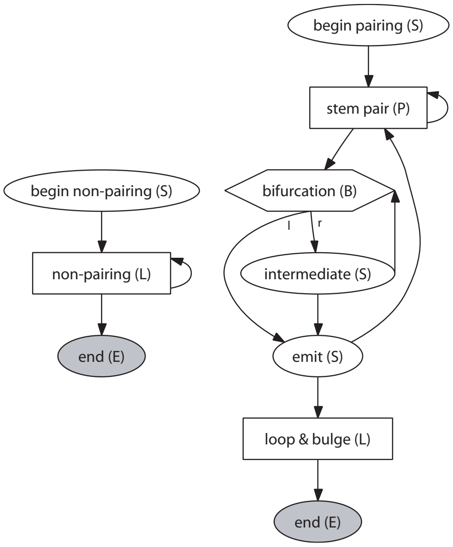

Figure S6: Transition graphs of the non-pairing component (left) and the pairing component (right) of the phylo-SCFGs. The state types are given in parenthesis. The transition from the bifurcation state leads to two states, a left (l) and a right (r), as indicated on the graph. See Pedersen et al. (2004) for more detail on the interpretation of these graphs.

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., et al., 2000. The genome sequence of *drosophila melanogaster*. *Science*, **287**:2185–2195.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., *et al.*, 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, **297**(5585):1301–1310.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., *et al.*, 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, **14**(4):708–715.

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

C. elegans Sequencing Consortium, 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**:2012–2018.

Chiaromonte, F., Weber, R. J., Roskin, K. M., Diekhans, M., Kent, W. J., and Haussler, D., 2003. The share of human genomic DNA under selection estimated from human-mouse genomic alignments. In *Cold Spring Harbor Symp Quant Biol*, volume 68, pages 245–254.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A., and Johnston, M., 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**:71–76.

Cocchia, M., Huber, R., Pantano, S., Chen, E. Y., Ma, P., Forabosco, A., Ko, M. S., and Schlessinger, D., 2000. PLAC1, an Xq26 gene with placenta-specific expression. *Genomics*, **68**(3):305–312.

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

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

Dempster, A., Laird, N., and Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, **39**:1–38.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G., 1998a. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G., 1998b. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences. *J Mol Evol*, **17**:368–376.

Felsenstein, J., 2004. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Felsenstein, J. and Churchill, G. A., 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol*, **13**:93–104.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R., 2003. Rfam: an rna family database. *Nucleic Acids Res.*, **31**(1):439–441.

Hardison, R. C., Roskin, K. M., Yang, S., Diekhans, M., Kent, W. J., 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**(1):13–26.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M. C., Wides, R., *et al.*, 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, **298**(5591):129–149.

International Chicken Genome Sequencing Consortium, 2004. Sequencing and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**:695–716.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**:860–921.

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

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D., 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci USA*, **100**:11484–11489.

Kent, W. J. and Zahler, A. M., 2000. Conservation, regulation, synteny, and introns in a large-scle *C. briggsae-C. elegans* genomic alignment. *Genome Res*, **10**:1115–1125.

Knudsen, B. and Hein, J., 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**(6):446–454.

Margulies, E. H., NISC Comparative Sequencing Program, Maduro, V. V. B., Thomas, P. J., Tomkins, J. P., et al., 2005. Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. *Proc Natl Acad Sci USA*, . In press.

Meza, J. C., 1994. OPT++: An object-oriented class library for nonlinear optimization. Technical Report SAND94-8225, Sandia National Laboratories.

Mouse Genome Sequencing Consortium, 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**:520–562.

Ogurtsov, A. Y., Sunyaev, S., and Kondrashov, A. S., 2004. Indel-based evolutionary distance and mouse-human divergence. *Genome Res*, **14**(8):1610–1616.

Pedersen, J. S., Meyer, I. M., Forsberg, R., Simmonds, P., and Hein, J., 2004. A comparative method for finding and folding rna secondary structures within protein-coding regions. *Nucleic Acids Res*, **32**(16):4925–4936.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T., 1992. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, second edition.

Rat Genome Sequencing Project Consortium, 2004. Genome sequence of the Brown Norway Rat yields insights into mammalian evolution. *Nature*, **428**:493–521.

Richards, S., Liu, Y., Bettencourt, B. R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M. J., Chen, R., Meisel, R. P., *et al.*, 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. *Genome Res*, **15**(1):1–18.

Rivas, E. and Eddy, S. R., 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding rnas. *Bioinformatics*, **16**(7):583–605. (eng).

Rokas, A., Williams, B. L., King, N., and Carroll, S. B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**(6960):798–804.

Siepel, A. and Haussler, D., 2004a. Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol*, **11**:413–428.

Siepel, A. and Haussler, D., 2004b. Computational identification of evolutionarily conserved exons. In *Proc. 8th Int'l Conf. on Research in Computational Molecular Biology*, pages 177–186.

Siepel, A. and Haussler, D., 2004c. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol*, **21**:468–488.

Siepel, A. and Haussler, D., 2005. Phylogenetic hidden Markov models. In Nielsen, R., editor, *Statistical Methods in Molecular Evolution*. Springer.

Smith, N. G. C., Brandstrom, M., and Ellegren, H., 2004. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics*, **84**(5):806–813.

Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., *et al.*, 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol*, **1**(2):E45.

Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, **17**:57–86.

Whelan, S., Liò, P., and Goldman, N., 2001. Molecular phylogenetics: State-of-the-art methods for looking into the past. *Trends Genet*, **17**:262–272.

Yang, Z., Nielsen, R., and Hasegawa, M., 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol*, **15**(12):1600–1611.