Functional characterization of motif sequences under purifying selection

De-Hua Chen¹, Andrew Ying-Fei Chang², Ben-Yang Liao² and Chen-Hsiang Yeang¹,*

¹Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, ROC and ²Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan, ROC

Received September 18, 2012; Revised and Accepted December 13, 2012

ABSTRACT

Diverse life forms are driven by the evolution of gene regulatory programs including changes in regulator proteins and cis-regulatory elements. Alterations of cis-regulatory elements are likely to dominate the evolution of the gene regulatory networks, as they are subjected to smaller selective constraints compared with proteins and hence may evolve quickly to adapt the environment. Prior studies on cis-regulatory element evolution focus primarily on sequence substitutions of known transcription factor-binding motifs. However, evolutionary models for the dynamics of motif occurrence are relatively rare, and comprehensive characterization of the evolution of all possible motif sequences has not been pursued. In the present study, we propose an algorithm to estimate the strength of purifying selection of a motif sequence based on an evolutionary model capturing the birth and death of motif occurrences on promoters. We term this measure as the 'evolutionary retention coefficient', as it is related yet distinct from the canonical definition of selection coefficient in population genetics. Using this algorithm, we estimate and report the evolutionary retention coefficients of all possible 10-nucleotide sequences from the aligned promoter sequences of 27 748. orthologous gene families in 34 mammalian species. Intriguingly, the evolutionary retention coefficients of motifs are intimately associated with their functional relevance. Top-ranking motifs (sorted by evolutionary retention coefficients) are significantly enriched with transcription factor-binding sequences according to the curated knowledge from the TRANSFAC database and the ChIP-seq data generated from the ENCODE Consortium. Moreover, genes harbouring high-scoring motifs on their promoters retain significantly coherent expression profiles, and those genes are over-represented in the functional classes involved in gene regulation. The validation results reveal the dependencies between natural selection and functions of cis-regulatory elements and shed light on the evolution of gene regulatory networks.

INTRODUCTION

Diverse life forms are largely driven by conservation and variations of the gene regulatory circuits. Recent progress in high-throughput technologies such as next-generation sequencing platforms and DNA microarrays enables biologists to map the regulatory networks and investigate their evolution across multiple species. For instance, studies in evolutionary developmental biology (EvoDevo) compared the gene regulatory networks for animal development and discovered conserved cores responsible for body plan formation and variable modules modifying species-specific phenotypes such as the shapes of limbs or wings [e.g., (1,2)].

One remarkable feature from the gene regulatory networks of multiple species is the conservation of their constituent proteins (3). Most proteins possess multiple functions (pleiotropic), hence are subjected to tight selective constraints. Alterations on protein sequences (e.g., changes on the DNA-binding domain of a transcription factor) may affect many partners (e.g., changes on the bindings of all targets of a transcription factor), thus are likely to be deleterious. In contrast, alterations on cis-regulatory elements have local effects and thus enable the systems to evolve in an incremental fashion. Consequently, evolution of non protein-coding regions in general and cis-regulatory elements in particular plays a critical role in the evolution of the gene regulatory systems.

Early studies of cis-regulatory element evolution focus on identification of conserved transcription factor-binding motifs (4) and detection of conserved regions on gene promoters (5). Sequence conservation alone, however, does
not suffice to account for the evolution of gene regulatory systems. Comparison of known cis-regulatory elements on closely related species indicates high rates of turnover and divergence (6–10). These changes may yield gains and losses of cis-regulatory elements (19), modify the regulatory programs (1,2) or are accompanied by compensatory mutations to maintain stable regulatory programs (11,12).

Like protein-coding regions, evolution of cis-regulatory elements is driven by a variety of mechanisms including sequence substitutions (18), gene duplications (13), tandem repeat insertions and deletions (14). Cis-regulatory elements are added or deleted on the promoters/enhancers according to these mechanisms. Ideally, a complete model for the evolution of cis-regulatory elements should be based on the models of all individual mechanisms for molecular evolution. In practice, mechanisms other than sequence substitutions are hard to formalize. Consequently, the majority of quantitative models for cis-regulatory element evolution are derived from sequence substitution processes. Several studies use simulations to examine the effects of sequence mutations on the rates for cis-regulatory element evolution [e.g., (15,16)]. Others start with sequence substitution models in population genetics and attempt to identify the cis-regulatory elements under selection [e.g., (17–19)]. Despite the fruitful outcomes generated from these studies, they suffer from two major limitations. First, they focus primarily on the deviation of observed sequences from a known regulatory element (e.g., a transcription factor-binding motif) rather than the changes of regulatory element occurrence on promoters. Alterations on motif counts can be more critical for gene regulation than specific sequence variations, as the former modulate the number of transcription factors bound on promoters. Second, all the current studies only examine a collection of known transcription factor-binding motifs. Complete characterization of the evolution of all possible motif sequences of a fixed length is lacking. This characterization, however, is critical for discovering new regulatory elements and comprehending their evolution on genomes.

Previously, we proposed an evolutionary model and an algorithm to quantify the strength of natural selection of a motif sequence (20). The evolution of motif occurrence was formulated as a birth–death process, whereas the rates of motif additions and deletions were derived from substitutions of their constituent sequences. The evolutionary retention coefficient of a motif was defined as a penalty to slow down motif death, and the evolutionary retention coefficient of a motif by comparing the empirical distribution of its occurrences over multiple species with the one generated by a neutral evolutionary model. The evolutionary model of motif occurrences and the algorithm to quantify the strength of natural selection of motifs are described below.

A Poisson process model of sequence substitution

We adopt the simplest model of sequence substitution assuming all nucleotides at all positions and across all lineages transition with an equal rate (31). In an infinitesimal time interval \( dt \), the nucleotide sequence of a position transitions to another base with probability \( \lambda dt \). \( n_i(t) \) denotes the cumulative number of sequence changes at time \( t \). The transitions within the time interval \( [t, t+dt] \) is as follows

\[
P(n_i(t+dt) = (N+1)|n_i(t) = N) = \lambda dt.
\]

and \( n_i(t) \) has a Poisson distribution

\[
P(n_i(t) = N|n_i(0) = 0) = \frac{(\lambda t)^N}{N!} e^{-\lambda t}.
\]

The maximum likelihood estimate of \( \lambda \) is simply the total number of sequence changes divided by the total length of the time interval considered. In this work we estimated \( \lambda \) from the aligned 5 kb promoter sequences of the 27748

MATERIALS AND METHODS

Data sources

Aligned 5 kb upstream sequences of 27748 orthologous gene families from 34 mammalian species were extracted from the UCSC Genome Browser (21). Supplementary Table S1 and Figure S1 report the names and the phylogenetic tree of the selected species.

To validate the functional relevance of the high-scoring motifs, we downloaded external datasets from the following sources: the consensus motifs of transcription factor-binding sequences from the TRANSFAC database (22), 407 ChIP-seq data files from the ENCODE database (23), DNA microarray data of human and mouse tissue gene expressions (24) and RNA-seq data of human tissue gene expressions (25), the annotations and member genes of 3201 Gene Ontology (GO) categories (27) and pathway information from three databases (28–30).

Quantifying the strength of natural selection of motif sequences

We define a motif as a collection of sequences with the same length. Over time motifs are created, annihilated or maintained in a specified region (e.g., a gene promoter) by sequence substitutions of the constituting nucleotides. Motifs undergoing purifying selection would possess slower rates of annihilation than those without selective constraints. Accordingly, we quantify the strength of natural selection of a motif by comparing the empirical distribution of its occurrences over multiple species with the one generated by a neutral evolutionary model. The evolutionary model of motif occurrences and the algorithm of evaluating the evolutionary retention coefficients of motifs are described below.
gene families over the 34 mammalian species. For each position of the aligned promoters in each gene family, we observed the sequences in the terminal nodes (the 34 extant species) of the species tree and inferred the sequences in the internal nodes (ancestral species) by a dynamic programming algorithm (32). We then counted the total number of sequence changes along all branches of the species tree for all positions and all gene families and the total lengths of the time intervals, and calculated $\lambda$ accordingly. From the empirical data, $\lambda = 0.8371$.

**A birth–death model for the evolution of motif occurrences**

A motif $\mathcal{M} \subseteq B^{10}$ (B = {A, G, T}) is defined as a collection of nucleotide sequences of fixed length $l_m$. We first consider the sequence evolution of $l_m$ consecutive positions. There are $4^{l_m}$ possible sequences that can occur in this $l_m$-mer window, and each sequence $s \in B^{l_m}$ can be labelled as either a member of the motif $(s \in \mathcal{M})$ or not $(s \notin \mathcal{M})$. These sequences comprise an undirected graph $G = (V, E)$, where a node $v \in V$ denotes an $l_m$-mer sequence and an edge $e = (v_1, v_2)$ denotes a sequence pair $v_1$ and $v_2$ different at one position. The evolution of $l_m$-mer sequences can be viewed as a Markov random walk on $G$. In an infinitesimal time interval, a sequence can only transition to a neighboring node in $G$ and the rate of transition is $\lambda_{lm}$.

A motif $\mathcal{M}$ constitutes a subset of nodes in $G$ (black nodes in the left diagram of Figure 1), while the remaining nodes are non-motif sequences (white nodes in the left diagram of Figure 1). We are interested in the transition rate from non-motif sequences to motif sequences and vice versa. With a simplifying approximation, we characterize these transitions with two numbers: $r_{01}$ as the fraction of all non-motif $\rightarrow$ motif transitions among all transitions from non-motifs, and $r_{10}$ as the fraction of all motif $\rightarrow$ non-motif transitions among all transitions from motifs. $P_A$, $P_C$, $P_G$ and $P_T$ denote the background frequencies of the four nucleotides obtained from all promoters of the 34 species. $r_{01}$ and $r_{10}$ are calculated by the following formulas:

$$r_{01} = \frac{\sum \{ (v_1, v_2) : v_1 \in \mathcal{M}, v_2 \notin \mathcal{M} \} \delta(v_1, v_2)}{\sum \{ (v_1, v_2) : v_1 \in \mathcal{M}, v_2 \notin \mathcal{M} \} \delta(v_1, v_2)},$$

$$r_{10} = \frac{\sum \{ (v_1, v_2) : v_1 \notin \mathcal{M}, v_2 \in \mathcal{M} \} \delta(v_1, v_2)}{\sum \{ (v_1, v_2) : v_1 \notin \mathcal{M}, v_2 \in \mathcal{M} \} \delta(v_1, v_2)}.$$

Where $\delta(*)$ is an indicator function and $\omega(v_1, v_2)$ is the nucleotide background probability of $v_2$ at the position where $v_1$ and $v_2$ differ. For instance, $\omega(\text{AGGC}, \text{AGTC}) = P_T$. $\omega(v_1, v_2)$'s rescale the weights of transitions according to the frequencies of the destination sequences. For instance, transitions to GC-rich sequences are more likely to occur on mammalian promoters, as they are over-represented in the CpG islands (33).

$n(t)$ denotes the number of motif occurrence at time $t$. In an $l_m$-mer window, $n(t) \in [0,1]$ as the sequence is either a motif or not. The transitions of $n(t)$ hence conform with the following equations

$$P(n(t+1) = 1 | n(t) = 0) = \lambda_{lm} r_{01} dt,$$

$$P(n(t+1) = 0 | n(t) = 1) = \lambda_{lm} r_{10} dt.$$

We now extend the analysis to the entire promoter of length $l_s \gg l_m$. Suppose motif occurrence at time $t$ is $n(t) = n$ and the $n$ occurring motifs are not overlapped. Each motif instantiation can be annihilated with a rate $\lambda_{lm} r_{10}$. Hence the ‘death rate’ on the entire promoter is the rate on an $l_m$-mer window multiplied by $n$:

$$P(n(t+dt) = n-1 | n(t) = n) = \lambda_{lm} r_{10} n dt.$$

There are $l_s - l_m n$ positions unoccupied by motif sequences, and the maximum number of (possibly overlapped) $l_m$-mer windows is $l_s - l_m n - l_m + 1$. Each of these $l_m$-mer windows can generate a new motif. Hence the ‘birth rate’ on the entire promoter is approximately the rate on an $l_m$-mer window multiplied by $l_s - l_m n - l_m + 1$:

$$P(n(t+dt) = n+1 | n(t) = n) = \lambda_{lm} r_{01} (l_s - l_m n - l_m + 1) dt.$$

Equations (6) and (5) specify a birth–death process (34) of motif occurrence on a promoter of length $l_s$. The distribution $P_n(t) \equiv P(n(t) = n)$ of motif occurrences over time can be expressed as a system of differential-difference equations:

$$\frac{dp_{01}(t)}{dt} = \mu_1(t) p_{1}(t) - \mu_0(t) p_{01}(t),$$

$$\frac{dp_{10}(t)}{dt} = \lambda(t) p_{n-1}(t) + \mu n(t) p_{n+1}(t) - \lambda(t) p_{n}(t) - \mu n(t) p_{n}(t),$$

$$\mu(n) = \lambda_{lm} r_{01} n.$$

The system is illustrated by the right diagram of Figure 1.

The aforementioned model assumes that sequences randomly drift and henceforth no selective pressure is exerted on the evolution of motif occurrence. In contrast, purifying selection should penalize decrements of motif occurrence. Therefore, the evolutionary model of motif occurrence under purifying selection largely resembles the model for neutral evolution (Equation 7) except for a modification of the motif death rate:

$$\mu(n) = \frac{\mu(n)}{s}.$$

The motif death rate $\mu'(n)$ under selection slows down the neutral motif death rate $\mu(n)$ by a factor $s > 1$. We term $s$ as the evolutionary retention coefficient of a motif. Notably, this definition is related yet distinct from the canonical definition of selection coefficient in population genetics (35). In population genetics, the selection coefficient is the decline of the relative fitness of a selectively disadvantageous genotype compared with that of a selectively favoured genotype. In a sufficiently large population, the selectively advantageous genotype will appear with a higher frequency than that of a genotype without selection. In this regard, both the canonical selection coefficient and evolutionary retention coefficient aim for capturing the strength of purifying selection from the observed genotypes. However, despite the common goals shared by the two measures, the evolutionary retention coefficient is distinct from the canonical selection coefficient in two
aspects. First, the evolutionary retention coefficient bypasses the abstract notion of relative fitness and directly tackles the consequences of purifying selection—elevation of motif occurrence frequencies. Second, the canonical selection coefficient examines the allele frequencies of a single site, whereas the evolutionary retention coefficient is inferred from the frequencies of motif occurrence in a consecutive region of the genome. We will further clarify the relation between these two scores in simulation studies.

**Estimating the evolutionary retention coefficients from empirical data**

We estimated the evolutionary retention coefficient of a $l_m$-mer motif from the aligned 5 kb promoter sequences of 34 mammalian species with the following procedures. First, we divided a promoter into multiple segments of fixed length $l_s = 30$. Segments with >10% gaps in any species were discarded. This partition reduces the number of valid terms in Equation (7), hence greatly simplifies subsequent estimation.

Second, we treated humans as the reference species and assumed that alterations of motif counts from the reference to another species followed the birth–death process. $t$ denotes the distance between humans and another species $x$ in the phylogenetic tree, $n_0$ and $n_1$ the motif counts in the segments of humans and species $x$, respectively. For each combination of $t$ (or species $x$), $n_0$ and $n_1$, we then counted $f(t,n_0,n_1)$, the total number of segments with $n_0$ and $n_1$ motif instances in the counterparts of humans and species $x$.

Third, the log likelihood of motif occurrences can be expressed as

$$\mathcal{L} = \sum_t \sum_{n_0} \sum_{n_1} f(t,n_0,n_1) \log P(n(t) = n_1|n(0) = n_0) + C. \quad (9)$$

where $C$ is a constant and $P(n(t) = n_1|n(0) = n_0)$ denotes the conditional probability derived from the birth–death model under selection by $s$. Applying the death rate of Equation (8) in Equation (7). Given the relatively short length of segments, we only considered motif occurrences up to 3 and restricted the terms in Equation (7) to $n \leq 3$ accordingly. For each fixed value of evolutionary retention coefficient $s$, we solved $P(n(t) = n_1|n(0) = n_0)$ numerically by the finite difference method for ordinary differential equations. To estimate $s$ that maximizes the log likelihood in Equation (9), we used a binary search to find the optimal $s$ over the interval $[0,0.2]$

**Comparison of selection coefficients and evolutionary retention coefficients in simulated data**

To elucidate the relation between selection coefficients and evolutionary retention coefficients, we simulated haploid sequence evolution with varying selection coefficients and compared the evolutionary retention coefficients estimated from the observed data with the given selection coefficients. Given a promoter sequence of fixed length (30 nucleotides) and a motif (10 nucleotides), we define the relative fitness of the promoter as $f = 1 - (2 - \min(k,2))s$, where $k$ is the number of motif occurrence on the promoter and $s$ the selection coefficient. The sequence containing $\geq 2$ motif instances possesses the highest fitness, whereas the sequences containing 1 and 0 motif instance possess intermediate and low fitness, respectively.

We simulated promoter evolution according to both sequence substitutions of individual positions and purifying selection dictated by motif occurrence. One promoter sequence was randomly generated in the first generation. In each of the following generations, each sequence produced 10 progenies with a Poisson mutation rate of 0.02 per position. Among the progenies from the same cohort, 100 of them were selected with probabilities proportional to their relative fitness, and the remaining sequences were eliminated. This process of sequence substitutions and selection lasted for 100 generations. For each selection coefficient, we generated randomly 20 motif and initial promoter sequences and simulated their evolution separately. Finally, we repeated simulations for the following selection coefficient values: 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.49.

In each simulation, the coalescent tree of the 100 observed sequences was recorded. We chose the sequence closest to other observed sequences as the reference and evaluated their phylogenetic distances according to the structures and branch lengths of the coalescent tree. The evolutionary retention coefficient of each designated motif can be consequently inferred from the simulated sequences.
An exhaustive evaluation of evolutionary retention coefficients of all 10-mers on mammalian promoters

Using the aforementioned algorithm, we estimated the evolutionary retention coefficients of all $4^{10} = 1048576$ 10-mer sequences on the aligned 5 kb promoter sequences of 34 mammalian species. To accelerate computations, we ran the estimation procedures on two PC cluster systems simultaneously. Eight jobs were assigned in parallel to the HP DL360 G7 servers containing dual Intel Xeon E5520 CPUs with 2.27 GHz and 24 GB main memory, and 10 jobs were assigned in parallel to the HP BL460C servers containing Intel Xeon CPUs with 3.16 GHz and 16 GB main memory. The total running time was 6048 hours. Notably, although the theoretical framework we present can handle more general motifs (i.e., a collection of nucleotide sequences), in this work we only investigate the single sequence motifs (i.e., occurrences of a particular 10-mer sequence), as their evolutionary retention coefficients can be exhaustively calculated with limited computing resources.

Functional validation of motif sequences under selective pressure

We incurred the following four tests to validate the functional relevance of motif sequences under selection.

**Enrichment of TRANSFAC motifs**

First, we demonstrated that evolutionary retention coefficients were correlated with enrichment of transcription factor-binding motifs extracted from TRANSFAC (22). A non-parametric statistical test was used to evaluate the enrichment of transcription factor-binding motifs in high-scoring sequences. In brief, all the 1048576 10-mers sequences were sorted by their evolutionary retention coefficients in a descending order. 168397 of these sequences matched completely or partially with transcription factor-binding motifs in TRANSFAC. We defined $F_1(x)$ over the normalized rank $x \equiv \frac{\text{rank}}{m} \in [0,1]$ resembling the cumulative distribution function (CDF) of TRANSFAC motifs over the sorted 10-mer sequences.

$$F_1(x) = \frac{\# \text{TRANSFAC motifs in sequences} \rightarrow [x \cdot 4^{10}]}{\# \text{TRANSFAC motifs in sequences}}$$

(10)

$F_1(x)$ should have a high area under the curve if TRANSFAC motifs are enriched in the top-ranking sequences. In contrast, the null hypothesis stipulates that TRANSFAC motifs are evenly distributed along the sorted sequences and the corresponding CDF is $F_0(x) = x$. The maximum deviation between $F_1(x)$ and $F_0(x)$ gives rise to a statistic of the Kolmogorov–Smirnov test. This method is similar to the Gene Set Enrichment Analysis (GSEA) (36).

**Enrichment of protein-binding sites from the ENCODE data**

Second, we demonstrated that the top-ranking motifs were enriched in the protein-binding sites of the human genome reported from the ChIP-seq data generated by the ENCODE consortium (23). 407 ChIP-seq data files were extracted from the ENCODE website (http://hgdownload. cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHai bTifs/). Each file reports the sequences of fragments containing the binding sites of one protein in one cell type. For each motif, we constructed a simple null model assuming its occurrence on the entire human genome followed a Poisson process. The rate of motif occurrence per position is $\frac{\eta}{L}$, where $N$ denotes the number of motif occurrence in the entire genome and $L$ denotes the genome length. Suppose in a ChIP-seq file, the total length of fragments is $l$ and the number of motif occurrence is $n$. Then the Poisson rate of motif occurrence in the designated fragments is $\eta = \frac{nl}{L}$ and the $P$-value for motif enrichment is

$$P = \sum_{m=n}^{N} \frac{\eta^m}{m!} e^{-\eta}.$$  

(11)

**Coherence of expression profiles in human and mouse genes**

Third, we showed that human and mouse genes harbouring high-scoring motif sequences tended to have coherent expression profiles compared with genes harbouring low-scoring motif sequences. We used both oligonucleotide microarray data (24) and RNA-seq data (25) in defining expression levels and co-expression of human and mouse genes. For the microarray data, we obtained the expression information of human genes and mouse genes from the Gene Atlas V2 dataset (http://symatlas.gnf.org/SymAtlas/). This dataset comprises oligonucleotide microarray data in 63 human and 58 mouse normal tissues sampled from animal bodies. We assigned the expression data from probe sets to corresponding Ensembl genes following (37,38). The expression levels of a gene in a specific tissue were averaged among replicates. For the RNA-seq data, we obtained that of 11 human tissues from GEO Series GSE13652 from the University of Toronto (25) (brain/liver/muscle/cerebral cortex) and GSE12946 from MIT (26) (adipose/breast/colon/heart/lung/lymph node/testes). The raw 32-mer RNA-seq sequence reads were mapped to human genome (Ensembl version v56), and RNA-seq-based gene expression levels were calculated according to (39,40).

The expression profile divergence between two genes in the human or mouse genome was defined by $1 - r$, where $r$ is the Pearson’s correlation coefficient of expression levels across the tissues. In the present study, we specifically examined co-expression of genes that are not paralogs or genes located on different chromosomes. The chromosomal coordinates and annotations of paralogous relationships of human and mouse genes based on Ensembl v62 were obtained through BioMart (http://www.biomart .org/).

**Functional enrichment of genes harbouring high-scoring motifs**

Fourth, we showed that human genes harbouring high-scoring motif sequences were enriched with certain functional classes. Four sources pertaining to functional
information of human genes were extracted: the GO categories (27), the curated pathway databases of Reactome (28), Biocarta (29) and the NCI-Nature curations (30). For a given motif, we extracted the genes harbouring the motif on their promoters and calculated hyper-geometric $P$-values of enrichment for each GO category and pathway. The enriched functional classes for both top-ranking motifs (evolutionary retention coefficient $\geq 3.0$, 231 motifs) and control motifs (231 motifs surrounding the median of the sorted list) were reported.

RESULTS

Evolutionary retention coefficients are correlated with selection coefficients in simulated data

We first demonstrate the resemblance between canonical selection coefficients and motif evolutionary retention coefficients with simulated data. For each of 11 pre-determined selection coefficient values, we simulated the evolution of 20 phylogenies, where each phylogeny constituted 100 generations and 100 observed promoter sequences in their leaves. We then inferred the motif evolutionary retention coefficient from the observed promoter sequences of each simulated phylogeny. The left part of Figure 2 shows the (pre-determined) canonical selection coefficients and the (inferred) motif evolutionary retention coefficients of the 220 phylogeny instances. Overall, the two scores exhibit a positive correlation coefficient ($r = 0.678$). Because direct evaluation of relative fitness in a population is often challenging, the high correlation between the two quantities indicates that the motif retention coefficient is a reasonable measure for the selective strength of motifs.

Summary of evolutionary retention coefficients of 10-mer motif sequences

We evaluated the evolutionary retention coefficients of all $4^{10} = 1048576$ 10-mer sequences among the 5 kb promoters of 27,748 gene families in 34 mammalian species. The right part of Figure 2 shows the empirical distribution of evolutionary retention coefficients, and Supplementary Table S2 reports the sorted evolutionary retention coefficients of these sequences. As expected, the majority of sequences possess low evolutionary retention coefficients: the median value is 0.508 and the scores of about 80% of the sequences (834,552 of 1,048,576) are below 1.0. We considered the first 231 (0.022%) sequences with evolutionary retention coefficients $\geq 3.0$ as the top-ranking motif sequences and employed further analysis to these sequences.

Motifs with high evolutionary retention coefficients have slower death rates, thus should exhibit high level conservation on promoters. For each motif, we define a conservation measure as the probability of its presence on the promoter of a mammalian species, conditioned on its presence on the orthologous promoter of humans. Figure 3 displays the conditional probabilities of motif presence of the 231 top-ranking motifs (top panel) and 231 control motifs (bottom panel) with evolutionary retention coefficients near the global median and with $\geq 50$ instances on human promoters. The species (indices in the horizontal axis) are sorted by their phylogenetic distances to humans. All (both high-scoring and control) motifs have high level conservation between humans and anthropoid primates (chimpanzees, gorillas, orangutans, macaques, indices 2–5) and marmosets (Callithrix jacchus, index 6), and the conditional probabilities drop abruptly beyond marmosets. For instance, the median conditional probability between humans and chimpanzees (index 2), orangutans (index 4) and marmosets (index 6) are 0.861, 0.697 and 0.320 respectively, whereas the median conditional probability between humans and the Philippine tarsiers (Tarsius syrichta, index 7) drops below 0.074. However, the top-ranking motifs retain considerably higher level conservation than control motifs in all selected mammals. For instance, the median conditional probabilities of the top-ranking motifs between humans and guinea pigs (Cavia porcellus, index 14), horses (Equus caballus, index 21) and African elephants (Loxodonta africana, index 28) are 0.074, 0.138 and 0.070 respectively, whereas those of the control motifs are 0.030, 0.072 and 0.031 respectively.

Notably, in Figure 3 all but one of the top 23 motifs have relatively low level conservation compared with the remaining top-ranking motifs. By examining those sequences (Table 1), we found they all belonged to the Alu-J repeat elements (41). They exhibit background level conservation between humans and most other species but undergo a massive number of insertions on the promoters of gray mouse lemurs (Microcebus murinus, index 8) and small-eared galagos (Otolemur garnettii, index 9) (Supplementary Table S3). These insertions violate the Poisson process model of sequence substitutions, increase the counts of $f(t, 0, n_1)$ where $n_1 > 0$, and therefore elevate the evolutionary retention coefficients.

High-scoring motifs are enriched with transcription factor-binding sites

Transcription factor-binding sites likely accommodate some motif sequences under selective pressure. We confirmed the dependency of transcription factor-binding sites and evolutionary retention coefficients with two external datasets. First, we verified that sequences with higher evolutionary retention coefficients tended to match the transcription factor-binding motifs reported in the TRANSFAC database (22). A simple check on sequences sorted by evolutionary retention coefficients provides obvious evidence: 77 of the top 231 10-mer sequences match TRANSFAC motifs, whereas only 35 of the 231 sequences in the middle and 22 of the 231 sequences in the bottom of the sorted list match TRANSFAC motifs. Supplementary Table S4 reports the TRANSFAC match on top-ranking, middle and bottom control motifs.

In addition to observations on small subsets of sequences, we also quantified this dependency on the entire sorted list. Denote $X$ a random variable indicating the match of sorted sequences with TRANSFAC motifs. $P(X = x)$ indicates the probability that a sequence with
Figure 2. Left: The scatter plot of canonical selection coefficients and evolutionary retention coefficients on simulated data. Each point denotes the scores obtained from 100 simulated sequences derived from one common ancestor over 100 generations. Right: Empirical distribution of selection coefficients among the $4^{10} = 1048576$ 10-mer sequences. The probabilities are displayed in a log scale.

Figure 3. Conservation of motif occurrence between humans and another species $P(\text{motif occurs in a species } | \text{motif occurs in humans})$ for the top 231 motifs and 231 control motifs from the middle of the ranked list. The horizontal axis denotes the species index with an increasing distance from humans (same as the species order in Supplementary Table S1). The vertical axis denotes the motif index from high selection coefficients (top) to low selection coefficients (bottom). The top-ranking and control motifs are separated by a white line. Colours in the heat map denote the levels of conditional probabilities between 0 (black) and 1 (bright red).
| Rank | Sequence                  | Coeff | Annotation        | Rank | Sequence                  | Coeff | Annotation        |
|------|---------------------------|-------|-------------------|------|---------------------------|-------|-------------------|
| 1    | AGCAACCTCA                | 7.790699 | Alu-J          | 2    | TGAGGTTGCT                | 7.050924 | Alu-J          |
| 4    | ACAGCAACCT                | 6.27913  | Alu-J          | 5    | GCAACCTCAA                | 6.140494 | Alu-J          |
| 7    | CGGCACTTTT                | 5.898189 | YY1, E2F, RB    | 8    | AGGATCGTCT                | 5.765456 | Alu-J          |
| 10   | TGTGCTTTGAG               | 5.459984 | Alu-J          | 11   | GAGTTGCTCT                | 5.459200 | Alu-J          |
| 13   | ACCTCAAAC                 | 5.206502 | Alu-J          | 14   | GCTTTCTTTT                | 5.017748 | Alu-J          |
| 16   | TGCGCTTTGAG               | 4.853342 | Alu-J          | 17   | GCCCGCTTCC                | 4.805634 | Alu-J          |
| 20   | CGGTTGCTGT                | 4.280663 | Alu-J          | 21   | GTGCTTTTTT                | 4.253287 | Alu-J          |
| 23   | TCCACCCAGCA               | 4.722712 | Alu-J          | 24   | GCTGTGACTGT               | 4.716143 | Alu-J          |
| 25   | GCTGTGACTG                | 4.48658  | Alu-J          | 26   | GCCGCCGCCG                 | 4.400560 | Alu-J          |
| 28   | CGCCGCACCA                | 3.81426  | Alu-J          | 29   | TCACTGCTGT                 | 4.315666 | Alu-J          |
| 31   | TAGCTACCAAA               | 4.232268 | Alu-J          | 32   | TGCTTACTTGT                | 4.215245 | NF1            |
| 34   | AGTTTGAGCT                | 4.089617 | Alu-J          | 35   | CTGCTGAAGCA                | 4.085705 | Alu-J          |
| 37   | GTTTGAGCTT                | 3.977175 | NF1            | 38   | TCTTTGCTCT                | 3.881167 | NF1            |
| 40   | CCCCCCCC                   | 3.819712 | Alu-J          | 41   | GTCTTCTTCTC                | 3.795781 | AP4            |
| 43   | CGTTGCTGT                 | 3.725561 | E12            | 44   | TTTTACTTCTT                | 3.719894 | Alu-J          |
| 46   | TCTGCTCAGCA               | 3.682684 | Alu-J          | 47   | TTATCTTCTT                 | 3.662582 | Alu-J          |
| 49   | AGCCATGCAA                 | 3.653545 | NF1            | 50   | GCTCTTTCTTC                | 3.649808 | SPI            |
| 52   | TGATTCTTCT                 | 3.631703 | Alu-J          | 53   | GCCGCACATCC                | 3.625968 | E2F, RB, YY1   |
| 55   | GCTGCTTTGCA               | 3.601241 | Alu-J          | 56   | GCGGCACATCC                | 3.581966 | SPI, EGR1,      |
| 58   | AAACCTTGTGTT               | 3.549357 | Alu-J          | 59   | TTTAATAAAAT                | 3.523619 | AP2            |
| 61   | GCACGCTGTCA                | 3.496142 | AP4, E12,       | 62   | AAGTTTCTTGT                | 3.492596 | AP4            |
| 64   | TATTTCCTCCCA               | 3.481676 | MITF, E47      | 65   | GGGGGGCCC                   | 3.481322 | EGR1, AP2      |
| 67   | TTTATTTCTT                | 3.470639 | Alu-J          | 68   | GTTTATTTCTT                | 3.460942 | Alu-J          |
| 70   | CCCCCCCCC                   | 3.451922 | Alu-J          | 71   | AATTCTTACTT                | 3.448670 | NF1            |
| 73   | CCAATACGGC                 | 3.442177 | NFY            | 74   | GCTACACATC                 | 3.427437 | NF1            |
| 76   | GCGATTGCT                  | 3.406631 | NF1            | 77   | GGAATATCAGTC                | 3.394995 | NF1            |
| 79   | ATTATTTTCG                 | 3.383930 | Alu-J          | 80   | GGATTACAGTC                 | 3.378569 | NF1            |
| 82   | CTGCTTTGCT                 | 3.374680 | NFY            | 83   | GGATGAGAGG                  | 3.368325 | NF1            |
| 85   | TTAAAATGTC                 | 3.364714 | Alu-J          | 86   | TGGTATTTCTT                | 3.326666 | NF1            |
| 88   | CACACGCTGCT                | 3.353610 | MITF, AP4,      | 89   | TTATTTATCTT                | 3.336673 | NF1            |
| 91   | CTGTGTTTGT                  | 3.331653 | Alu-J          | 92   | CTGACGTGTC                  | 3.322349 | MITF, AP4,      |
| 94   | CTAGCTTCC                  | 3.305246 | Alu-J          | 95   | GCTACTGTC                   | 3.303905 | MITF, E47,      |
| 97   | CAAATTTGTC                 | 3.301032 | Alu-J          | 98   | TATCTTCTT                  | 3.297026 | MITF, E47,      |
| 100  | CATTTTCTAC                 | 3.282516 | Alu-J          | 101  | GTTCTTCTT                  | 3.274815 | MITF, E47,      |
| 103  | GCGACAGCAGG                | 3.262684 | E12            | 104  | GACGCTGCTC                 | 3.256540 | MITF, E47,      |
| 106  | CAGCTGCTGT                 | 3.250314 | Alu-J          | 107  | GTCAGCTGCT                 | 3.246076 | MITF, E47,      |
| 109  | ATTTATTTCTT                | 3.240528 | Alu-J          | 110  | AATCTTCTT                   | 3.236959 | MITF, E47,      |
| 112  | TATTTTCTC                 | 3.230302 | NF-AT1         | 113  | GCGCTGCTC                  | 3.228710 | MITF, AP4,      |
| 115  | TACATTTCTC                | 3.218336 | FOXMI, STAT5A  | 116  | AATTACATG                   | 3.215724 | MITF, E47,      |
| 118  | TATTTTCTT                 | 3.191850 | Alu-J          | 119  | TTTTATTTCTT                | 3.196940 | MITF, E47,      |
| 121  | TTTATTTTAT                 | 3.190555 | Alu-J          | 122  | TTTTATTTA                  | 3.189907 | MITF, E47,      |
| 124  | TACTGCTCAG                 | 3.185563 | Alu-J          | 125  | TGACGTGCT                  | 3.182055 | MITF, AP4,      |
| 127  | GCTGCTGCTT                | 3.175051 | Alu-J          | 128  | ACCTCAATC                  | 3.171738 | MITF, E47,      |
| 130  | GCGACGAGTT                 | 3.167511 | E12, HEB       | 131  | CTATACTATT                 | 3.166592 | MITF, E47,      |
| 133  | AAATATTAGT                | 3.163564 | Alu-J          | 134  | GCCCGGCAGC                 | 3.162647 | MITF, E47,      |
| 136  | ATTACATGAA                  | 3.155954 | Alu-J          | 137  | TTTATTAATC                 | 3.152800 | MITF, E47,      |
| Rank | Sequence       | Coeff  | Annotation          | Rank | Sequence       | Coeff  | Annotation          | Rank | Sequence       | Coeff  | Annotation          |
|------|----------------|--------|---------------------|------|----------------|--------|---------------------|------|-------------------|--------|---------------------|
| 139  | ATTTCCTCTG     | 3.146366 |                     | 140  | AGCCTAGCTC     | 3.144451 |                     | 141  | TTTATTTTCA       | 3.143995 |                     |
| 142  | ATTTTCATTT     | 3.143630 |                     | 143  | TTATTTTCTT     | 3.142355 |                     | 144  | TATATTGATT       | 3.139713 |                     |
| 145  | TTCTGCTACT     | 3.136256 |                     | 146  | GTTTTCTCTIC     | 3.134619 |                     | 147  | TGCAGCAGCT       | 3.131622 |                     |
| 148  | TATATTCTT      | 3.126722 |                     | 149  | GTGCTATGG       | 3.125634 | RFX1                | 150  | AAATATTTGT       | 3.124547 |                     |
| 151  | TAATTAATAA     | 3.123641 | IPF1                | 152  | AACTGGTTTG      | 3.123551 |                     | 153  | AAATCCCTTT       | 3.122826 |                     |
| 154  | TGTTTCTTCA     | 3.122192 | FOXA1, FOXA2, FOXO3 | 155  | CTTTCAAACCT     | 3.120292 |                     | 156  | ACTTTCCTTC       | 3.119390 |                     |
| 157  | CTCTTTTCTT     | 3.118754 |                     | 158  | TAATTAGCAT      | 3.118754 |                     | 159  | TATTTTCCGT       | 3.116765 |                     |
| 160  | ATATTAATT      | 3.115681 |                     | 161  | TTTCATCTT      | 3.115049 |                     | 162  | TGTGAGCTAG       | 3.115219 |                     |
| 163  | TTTGIGATT      | 3.104319 |                     | 164  | AATGTTTCA       | 3.103509 |                     | 165  | CAGCCTACTG       | 3.100721 |                     |
| 166  | TCTTTTTTCTT    | 3.094613 |                     | 167  | TTTCAAGCTT      | 3.091116 |                     | 168  | TTTCCTAAAT       | 3.088428 |                     |
| 169  | TTCTCTTCTT     | 3.082702 |                     | 170  | TTCTTCTCTT      | 3.081630 | PU.1                | 171  | TTTCCTCTTT       | 3.081630 |                     |
| 172  | CTGGAATTTT     | 3.081004 |                     | 173  | TATAAATAAAA     | 3.079754 | PBX1A, HNF3         | 174  | ATATTTTTGA       | 3.078415 |                     |
| 175  | AATGCAAAAT     | 3.077612 | POU2F1             | 176  | ATTAATTAAA      | 3.077522 |                     | 177  | ATTAATTACA       | 3.076987 |                     |
| 178  | AGTTAATAA      | 3.074847 |                     | 179  | ATTTTTCTT       | 3.073777 |                     | 180  | ATTTCTTATT       | 3.071193 |                     |
| 181  | ATTTTCCTCT     | 3.070392 |                     | 182  | CAGCAGCAGC      | 3.069680 |                     | 183  | TCATTATTT       | 3.069680 |                     |
| 184  | TCTATTAATT     | 3.069235 |                     | 185  | AAATTATTTT      | 3.068968 | NFK6, HNF-1alpha   | 186  | AAATACATT       | 3.067367 |                     |
| 187  | TTATTTTGAGA    | 3.064788 |                     | 188  | AAATTTTGTG      | 3.062832 |                     | 189  | GAATGTGTTT       | 3.060349 |                     |
| 190  | TGGACCGTIG     | 3.060260 |                     | 191  | AGGAGCTGCT      | 3.058663 |                     | 192  | ATTTACATTG       | 3.057776 |                     |
| 193  | TTGATTCTGG     | 3.054852 |                     | 194  | CTCTTGGCTT      | 3.050515 | PU.1                | 195  | ATATTTATTT       | 3.049188 |                     |
| 196  | TTATAAAAT      | 3.047686 |                     | 197  | TTTATAATT       | 3.047597 |                     | 198  | TTTTTCTTTG       | 3.045389 |                     |
| 199  | ATTAACCTT      | 3.044859 |                     | 200  | CTGTTTCTCTT      | 3.043270 |                     | 201  | TATTCCTAAAT       | 3.042653 |                     |
| 202  | CCATGGCAAC      | 3.039038 | RFX1               | 203  | CAGCAGCTGG      | 3.038245 | AP4, E12, E47       | 204  | CATTATTTAT       | 3.037628 |                     |
| 205  | ATCATCACAT     | 3.033579 |                     | 206  | CATTATAAAT      | 3.033404 |                     | 207  | TACAATCCCT       | 3.030766 |                     |
| 208  | CATCTGTAA      | 3.030415 |                     | 209  | TGAACATCATC     | 3.027692 |                     | 210  | CCAAGCAGCT       | 3.027523 |                     |
| 211  | GTAATTTTATT    | 3.024358 |                     | 212  | ATTTAAATTT      | 3.023919 | POU3F1, POU3F2     | 213  | GAAATTTTTCTT     | 3.023832 |                     |
| 214  | CTTCTGCGAGG    | 3.022717 | NRF1, ELF1         | 215  | CTATTCTCTT      | 3.021641 | STAT5A             | 216  | AAACTGGTTA       | 3.020502 |                     |
| 217  | CTCTAAACATT    | 3.018313 |                     | 218  | TCTCTATT       | 3.017001 |                     | 219  | TGTGATGTCA       | 3.016913 | ATF6, API, CREB, NRF2 |
| 220  | TGTTTCTTCA     | 3.013766 |                     | 221  | TCTCTGAGCC      | 3.012805 |                     | 222  | TTTATTTATT       | 3.010883 |                     |
| 223  | TCAGCTTACGC    | 3.008266 |                     | 224  | TCTTCAGCT       | 3.007917 |                     | 225  | TCTCTCAAGC       | 3.007481 |                     |
| 226  | TTATTTTACTT    | 3.005650 |                     | 227  | TATATTGGA       | 3.003472 |                     | 228  | TAATTTTCTA       | 3.003124 |                     |
| 229  | TTCTGATTTG     | 3.002344 | NFY                | 230  | TTCTAAATAA      | 3.002340 |                     | 231  | ACATTTTATT       | 3.001905 |                     |
normalized rank \( x(0 \leq x \leq 1) \) matches TRANSFAC motifs. If evolutionary retention coefficients are uncorrelated with the presence of TRANSFAC motifs, then all TRANSFAC motifs should be evenly distributed along the normalized ranks, and \( X \) follows a uniform distribution. Therefore, enrichment of TRANSFAC motifs on high-scoring sequences is quantified by the deviation of the empirical distribution of \( X \) from a uniform distribution.

Figure 4 plots the empirical CDF of \( X \) (\( F_1(x) \) in Equation 10) and the CDF of a uniform distribution (\( F_0(x) \)). \( F_1(x) \) lies above \( F_0(x) \) for all \( x \in [0,1] \), indicating that sequences with high evolutionary retention coefficients are more likely to match TRANSFAC motifs than those with low evolutionary retention coefficients. The \( P \)-value of the Kolmogorov–Smirnov test \( < 10^{-325} \).

Second, we showed that the top-ranking motifs were enriched in the protein-binding DNA fragments reported from the ENCODE data (23). We downloaded 407 files from the ENCODE website. Each file reports the protein-binding DNA fragments generated by one ChIP-seq experiment with a specified antibody and cell type. The 407 files cover 59 proteins (transcription factors, RNA polymerase II, nucleosome-binding proteins, etc.), where multiple ChIP-seq experiments with distinct cell types and replicates were undertaken for each protein. For each motif, we quantified the significance of its enrichment in an ENCODE file with a null model assuming that its occurrence followed a Poisson process with a rate \( \lambda = \frac{\pi L}{l} \), where \( N \) denoted the number of motif occurrence in the entire genome, \( L \) the genome length and \( l \) the total fragment length in the file.

We evaluated the enrichment \( P \)-values for the top-ranking motifs in each ENCODE file. About 12% of the motif-file combinations (11 063 of 940 17) exhibit significant enrichment (\( P \leq 10^{-20} \)). To reduce errors generated by individual ChIP-seq experiments, we grouped the results of the same proteins together and counted the fractions of files in each group with significant enrichment. There are 182 motif–protein combinations with at least 5 ENCODE files and significant enrichment \( P \)-values (\( \leq 10^{-20} \)) in at least 80% of the constituent ENCODE files. The number of enriched motif–protein combinations drops considerably in control motifs. Among the 231 control motifs in the middle of the sorted list, 80 motif–protein combinations retain the same level of enrichment. Furthermore, among the 231 control motifs in the bottom of the sorted list, only 38 motif–protein combinations retain the same level of enrichment. Intriguingly, both TRANSFAC and ENCODE data indicate that levels of enrichment shrink by half from the top to the middle and from the middle to the bottom of the sorted list.

Table 2 shows the 182 motif–protein combinations with significant enrichment. Six motifs are enriched in the ChIP-seq data of at least 10 proteins. These motifs are heavily biased toward GC-rich sequences: GCGCCCTGC GC (index 27), GGGGGCGGGGC (index 33), GCCCCCG GC (index 56), GGGGGGGGGG (index 65), GCGCAT GC GC (index 76) and GGGGGCCGCC (index 134). The GC-rich sequences match the binding motifs of several proteins such as SP1 (42), AP2 (43), NRF1 (44) and E2F1 (44). Reciprocally, the ChIP-seq files of seven proteins contain at least 10 enriched motif sequences: ELF1, GABP, YY1, ERG1, RAD21, POL2 and PAX5. Some of these proteins (such as SP1, AP2, POL2, YY1, RAD21) ubiquitously regulate many genes, thus their binding motifs yield high evolutionary retention coefficients.

Four motif–protein combinations enriched in ENCODE files correspond to exact match with the TRANSFAC data. Motifs 33 (GGGGCGGGGC) and 36 (GGGGCCCCGGG) match the SP1-binding motif in TRANSFAC. Motif 7 (CCGCATCCTTT) matches the YY1-binding motif, and motif 170 (CTTCCCTCCTTT) matches the PU.1-binding motif.

**Genes harbouring high-scoring motifs tend to retain functional coherence**

Genes sharing the same protein-binding sequences on their promoters are likely co-regulated by the same transcription factors. Consequently, we expect genes harbouring high-scoring motifs to possess functional coherence. We validated this prediction with two tests using external data. First, using the method described in (39,40), we evaluated the divergence of expression profiles of genes from two human expression datasets and one mouse expression dataset. The distribution of expression divergence on genes harbouring the top 5000 motifs was compared with the distribution on the genes harbouring the bottom 5000 motifs. Intriguingly, genes harbouring the top 5000 motifs have consistently lower expression divergence than genes harbouring bottom 5000 motifs across all three datasets. The Wilcoxon test \( P \)-values of the deviation between the two gene sets are significant across the three datasets: \( 2.047 \times 10^{-14} \), \( 2.414 \times 10^{-4} \) and \( 1.670 \times 10^{-12} \) respectively. Furthermore, by ruling out the two confounding factors for co-expression—co-localization of genes on the same chromosomes and paralogous genes sharing the same ancestry—the deviation of expression divergence between genes harbouring top 5000 and bottom 5000 motifs remains pronounced. Table 3 reports the significance of the deviation of expression divergence between gene pairs harbouring top 5000 and bottom 5000 motifs. The deviation of expression divergence suggests that genes harbouring motifs of high evolutionary retention coefficients tend to retain functional coherence.

Second, we extracted the human genes harbouring each of the top 231 motif sequences and assessed their over-representations in 3201 GO categories and 889 pathways from three sources. Supplementary Table S5 reports the functional categories and pathways significantly enriched (hyper-geometric \( P \leq 0.001 \)) with each top-ranking motif. There are 45 motif–functional class pairs with significant enrichment. In contrast, there are only 9 significant motif–functional class pairs among the 231 control motifs in the middle of the sorted list.

By examining the functional enrichment results in Supplementary Table S5, we found that many top-ranking motifs were highly enriched in functional classes
related to transcriptional regulation such as nucleosome assembly (motif CCAGCTCCAG, $P$-value $3.264 \times 10^{-10}$), transcription factor activity (motif CCCCTCCCCC, $P$-value $8.773 \times 10^{-10}$) and chromatin modification (motif CCGCCGCGGC, $P$-value $6.109 \times 10^{-6}$). To justify this observation, we categorized the GO terms into four classes: regulators (transcription factors and signalling proteins, 2563 genes), enzymes (6947 genes), transporters (1087 genes) and structural proteins (571 genes), and evaluated the enrichment $P$-values of top-ranking and control motif targets in each class. Figure 5 reports the enrichment of motif targets on the four major categories. Strikingly, among the targets of the top 231 motif sequences, 20 are significantly enriched with known regulators ($P \leq 0.01$). In contrast, the targets of only one motif are enriched with enzymes, transporters and structural proteins, respectively. Furthermore, among the targets of the 231 control motifs in the middle of the sorted list, only 3 have significant enrichment in regulators. These results suggest that regulators tend to harbour motifs under stronger selective pressure on their promoters.

**Motif sequences with high evolutionary retention coefficients are derived by diverse causes**

Beyond statistical validations on the motif sequences sorted by evolutionary retention coefficients, we also examined the individual top-ranking motifs and annotated them with known regulatory sequences or repeat elements. Table 1 reports the functional annotations of the top 231 motifs. Several remarkable features emerge from the annotations. First, 26 10-mers constitute two blocks of 13 consecutive nucleotides (TAGCTCACAGCAACCTC and AGTTTGAGGTTGCTGTGAGCTA, respectively). These two blocks have background level conservation between humans and other species but undergo a massive number of insertions in gray mouse lemurs and small-eared galagos. Second, another ten 10-mers constitute a block of 19 consecutive nucleotides (TGCAGCAGCTGCTGCTGCT). Unlike Alu-J this block does not hit human repeats or gene sequences with significant blast E-values. This block largely coincides with many binding sites of MITF and AP4 according to the cisRED database of genome-wide regulatory module and element predictions.

![Figure 4](https://academic.oup.com/nar/article-abstract/41/4/2105/2414983/2115)

**Figure 4.** Enrichment of TRANSFAC motifs in high-scoring sequences. The blue curve shows the distribution of TRANSFAC motif occurrences along the normalized rank of the sorted 10-mer sequences $[F_1(x)$ in equation 10]. The red curve shows the CDF of a uniform distribution $[F_0(x)]$. 

Table 1 reports the functional annotations of the top 231 motifs. Several remarkable features emerge from the annotations. First, 26 10-mers constitute two blocks of 13 consecutive nucleotides (TAGCTCACAGCAACCTC and AGTTTGAGGTTGCTGTGAGCTA, respectively). These two blocks match exactly the Alu-J repeat elements. As shown in Figure 2, they have background level conservation between humans and other species but undergo a massive number of insertions in gray mouse lemurs and small-eared galagos. Second, another ten 10-mers constitute a block of 19 consecutive nucleotides (TGCAGCAGCTGCTGCTGCT). Unlike Alu-J this block does not hit human repeats or gene sequences with significant blast E-values. This block largely coincides with many binding sites of MITF and AP4 according to the cisRED database of genome-wide regulatory module and element predictions.

Third, three 10-mers (motifs 24, 26, 28) are three phases of the GCC-repeat sequences, and they coincide with many binding sites of ERG1 and DEAF1 according to cisRED. Fourth, four 10-mers (motifs 82, 32, 37, 38) form a 13-nucleotide consecutive block (CTCTGATTG) and coincide with NF-Y binding sites. Three additional 10-mers also coincide with NF-Y binding sites. Fifth, seven 10-mers are dominated by Cs and Gs (motif 27, GCGCCTGCGC; motif 33, GGGGCGGGGC; motif 36, GGGGCGGGGG; motif 56, GCCCCGCCCC;...
Table 2. Enrichment of top-ranking motifs in ENCODE ChIP-seq data

| Rank | Sequence | Protein | Fraction | Rank | Sequence | Protein | Fraction | Rank | Sequence | Protein | Fraction |
|------|----------|---------|----------|------|----------|---------|----------|------|----------|---------|----------|
| 27   | GCCGCTTCGGC | ATF3 | 0.8750(8/7) | 76   | GCCGCTTCGGC | ATF3 | 0.8750(8/7) | 130  | GCCGCTTCGGC | CTCF | 0.8333(12/10) |
| 7    | GCCGCCATCTT | ELF1 | 0.8333(6/5) | 24   | GCCGCCATCTT | ELF1 | 1.0000(6/6) | 26   | GCCGCCATCTT | ELF1 | 1.0000(6/6) |
| 27   | GCCGCCCATCTT | ELF1 | 1.0000(6/6) | 28   | GCCGCCCATCTT | ELF1 | 1.0000(6/6) | 33   | GCCGCCCATCTT | ELF1 | 1.0000(6/6) |
| 36   | GCCGCCCATCTT | ELF1 | 1.0000(6/6) | 43   | GCCGCCCATCTT | ELF1 | 0.8333(6/5) | 53   | GCCGCCCATCTT | ELF1 | 0.8333(6/5) |
| 55   | GCCGCCCATCTT | ELF1 | 0.8333(6/5) | 56   | GCCGCCCATCTT | ELF1 | 1.0000(6/6) | 65   | GCCGCCCATCTT | ELF1 | 0.8333(6/5) |
| 70   | GCCGCCCATCTT | ELF1 | 0.8333(6/5) | 76   | GCCGCCCATCTT | ELF1 | 0.8333(6/5) | 90   | GCCGCCCATCTT | ELF1 | 0.8333(6/5) |
| 103  | GCCGCCCATCTT | ELF1 | 0.8333(6/5) | 130  | GCCGCCCATCTT | ELF1 | 0.8333(6/5) | 134  | GCCGCCCATCTT | ELF1 | 0.8333(6/5) |
| 27   | GCCGCCCATCTT | ELF1 | 1.0000(6/6) | 194  | GCCGCCCATCTT | ELF1 | 0.8333(6/5) | 214  | GCCGCCCATCTT | ELF1 | 0.8333(6/5) |
| 24   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 26   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 27   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) |
| 27   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 70   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 76   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) |
| 78   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 103  | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 130  | GCCGCCCATCTT | ELF1 | 0.8000(10/8) |
| 134  | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 182  | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 27   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) |
| 70   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 27   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 55   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) |
| 76   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 56   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 65   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) |
| 134  | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 70   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 134  | GCCGCCCATCTT | ELF1 | 0.8000(10/8) |
| 78   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 182  | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 27   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) |
| 2116 | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 78   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 134  | GCCGCCCATCTT | ELF1 | 0.8000(10/8) |
| 27   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 2116 | GCCGCCCATCTT | ELF1 | 0.8000(10/8) | 27   | GCCGCCCATCTT | ELF1 | 0.8000(10/8) |

(continued)
motif 65, GGGCGGGGCC; motif 70, CCCCGCCCCC; motif 134, GGCCCCGCCC), and most of these motif sequences coincide with the binding sites of SP1, AP2 and ERG1. Sixth, two 10-mers (motif 7, CCGCCATCTT; motif 53, GCCGCCATCT) form a consecutive block (GCCGCCATCTT) and largely coincide with the binding sites of YY1 (45).

**DISCUSSION**

Evolution of *cis*-regulatory elements is an essential and critical aspect of the evolution of the gene regulatory systems. Prior models and studies focus primarily on the sequence evolution of selected known *cis*-regulatory elements but do not characterize the evolution of all possible regulatory sequences. In this work, we propose a model to quantify the strength of purifying selection for motif sequences of a fixed length and estimate the evolutionary retention coefficients of all 10-mer sequences from the aligned promoters of 34 mammalian species. A series of validation tests confirm the functional relevance of the proposed evolutionary retention coefficients. High-scoring motifs are enriched with transcription factor-binding sites according to curated information from TRANSFAC and ChIP-seq experimental data from ENCODE. Furthermore, genes harbouring high-scoring motifs retain more coherent expression profiles in human and mouse and are over-represented in the functional categories and pathways involved in transcriptional regulation.

For each motif–protein combination, the fraction of ENCODE files with significant enrichment ($P < 10^{-10}$) is reported.

**Table 3.** Comparison of the distributions of expression divergence between gene pairs harbouring top 5000 motifs and bottom 5000 motifs

| Wilcoxon test | Human 62-tissues Affymetrix | Human 11-tissues RNAseq | Mouse 58-tissues Affymetrix |
|---------------|----------------------------|-------------------------|----------------------------|
| All possible pairs | $2.047 \times 10^{-14}$ | $2.414 \times 10^{-4}$ | $1.67 \times 10^{-12}$ |
| Inter-chromosomal pairs | $< 10^{-300}$ | $9.565 \times 10^{-5}$ | $8.138 \times 10^{-12}$ |
| Non-paralogous pairs | $3.042 \times 10^{-14}$ | $5.669 \times 10^{-5}$ | $1.141 \times 10^{-12}$ |

The Wilcoxon test $P$-values between the two distributions are shown. The three rows report the $P$-values derived from different criteria for selecting gene pairs: all possible gene pairs, gene pairs on distinct chromosomes and non-paralogous gene pairs. The three columns report the $P$-values derived from three datasets of mRNA expressions: Affymetrix data of 62 human tissues, RNAseq data of 11 human tissues and Affymetrix data of 58 mouse tissues.

motif 65, GGGCGGGGCC; motif 70, CCCCGCCCCC; motif 134, GGCCCCGCCC), and most of these motif sequences coincide with the binding sites of SP1, AP2 and ERG1. Sixth, two 10-mers (motif 7, CCGCCATCTT; motif 53, GCCGCCATCT) form a consecutive block (GCCGCCATCTT) and largely coincide with the binding sites of YY1 (45).
factor belonging to the GLI-Kruppel class of zinc finger proteins. RAD21 encodes a nuclear protein involved in the repair of DNA double-strand breaks, as well as in chromatid cohesion during mitosis. AP2 (TFAP2A) encodes a transcription factor that interacts with enhancer elements. Furthermore, the high-scoring motifs interacting with these proteins are highly biased toward GC-rich sequences. Ubiquitous presence of these motifs on many target genes probably accounts for their high evolutionary retention coefficients. In contrast, binding motifs of the transcription factors with small numbers of specific targets will not exhibit high evolutionary retention coefficients, as there are only a few instances on promoters.

A bias toward abundant sequences among the high-scoring motifs may be relieved by adjusting the rates of the birth–death process to fit the background frequencies of motifs. Currently, the rates of drifting into and out of a motif depend primarily on the relative volume of the motif and frequencies of single nucleotides in sequence space \( r_{01} \) and \( r_{10} \). To further reduce this bias, we can adjust the weights of transitions in Equation (3) according to the background frequencies of motifs in the entire genomes.

The birth–death model of neutral evolution (Equation 7) is based on the simplest model of sequence substitution assuming all nucleotides transition with an equal rate. This assumption is incongruent with various observed biases in sequence evolution. For instance, on single nucleotides, the rates of transitions (purine → purine or pyrimidine → pyrimidine) are higher than those of transversions (purine → pyrimidine or pyrimidine → purine). On dinucleotides, the mutation rates of CpG → TpG tend to be higher as methylated cytosines deaminate to form thymines. The current model can be extended to incorporate these biases with a price of complexity. In an extended version, the two parameters specifying relative transition rates from motifs to non-motifs and vice versa \( r_{01} \) and \( r_{10} \) depend not only on motif complexity but also on its constituting sequences. Transversions between purines and pyrimidines are penalized while substitutions from CpG to TpG are rewarded. For the computational cost for implementing and running this extended model, we decided to leave this task in the future work.

Conservation of motif occurrences can be viewed as an aggregation of two factors: (i) the fraction of functional motif instances among all motif occurrences and (ii) the level of conservation among the functional motif instances. The function of a motif in eukaryotic genomes depends on various contextual factors such as the presence of other transcription factor-binding sites and enhancers, nucleosome positions and chromatin configurations. Hence only a fraction of motif occurrences are likely to

---

**Figure 5.** Enrichment of four functional classes—regulators, enzymes, structural proteins and transporters—among the genes harbouring the top-ranking and control motifs. The horizontal axis denotes the four functional classes. The vertical axis denotes the motif index from high selection coefficients (top) to low selection coefficients (bottom). The top-ranking and control motifs are separated by a yellow line. Colours in the heat map denote the magnitudes of log 10(hyper geometric \( P \)-values) from −6 (bright red) to 0 (black).
be functional and are subjected to selective constraints. Moreover, among these functional instances, differential levels of conservation may be manifested on distinct sites. Some regulatory subsystems are less tolerant with dysregulation and thus undergo a stronger selective pressure, whereas others may be more flexible and thus retain a lower level of conservation. Disentangling these two factors from sequence data alone is very challenging, as the contextual information often cannot be determined by sequences. Alternatively, functional information such as ChIP-seq data can provide additional clues about the first factor. By examining the overlaps of selected motifs and transcription factor-binding sites, we can estimate the fraction of functional motif instances. The level of conservation of a few transcription factor-binding motifs has been investigated in the prior studies mentioned in Introduction.

It is puzzling that the top-ranking motifs are over-represented on the promoters of regulators (transcription factors and other DNA-binding proteins, signalling proteins) but not on other functional categories (enzymes, transporters, structural proteins). The results suggest that the regulatory circuits of regulators possess elements with high selective constraints, whereas those of other proteins do not. We provide two speculations to explain this observation. First, many regulators are involved in processes with pervasive impacts such as chromatin modification, nucleosome assembly and multicellular organismal development. Constituent genes of these processes are thus subjected to tighter selective constraints and are regulated by motifs with high evolutionary retention coefficients. Second, many motifs overrepresented in regulators are the GC-rich binding sequences of the aforementioned proteins. Regulatory programs of regulators may result from the combinatorial interactions between these generic motifs and other process-specific motifs. In contrast, the regulatory programs of other proteins may be dominated by process-specific motifs. Further experimental data and analysis are required in order to verify these speculations.

In the present study, the log likelihood function is obtained by comparing motif occurrences between a reference species (human) and all the other species (Equation 9). This is not an exact form of the joint log likelihood function, as it assumes the motif occurrences of other species are independent conditioned on human data and ignores their dependencies owing to a shared phylogeny. A more accurate form is to sum up the log conditional probabilities along all branches of the phylogenetic tree:

$$\mathcal{L} = \sum_{b \in T} \sum_{l(b)=t} (n_0, n_1) \log P(n_{0}(t) = n_1 | n_{0}(0) = n_0) + C.$$  

(12)

where bl denotes a branch in the phylogenetic tree T with length l(bl) = t, n_0(0) and n_0(t) denote the motif occurrences at initial and terminal time points of the branch, respectively. Motif occurrences on ancestral nodes of the phylogenetic tree are not directly observed. Hence a dynamic programming algorithm can be applied to either reconstruct the maximum likelihood promoter sequences of ancestors or marginalize over the probability distributions of all possible sequence configurations (32). However, this accurate formulation requires reconstruction of 5 kb upstream sequences (or their probability distributions) of 27 748 orthologous gene families from 34 mammalian species. Owing to its tremendous computational cost, we decided to implement the simplified approximation for an exhaustive screening of all 10-mer motif sequences on all orthologous gene families. For subsequent studies on specific motifs and selected gene families, the accurate version in Equation (12) should be adopted.

In the present study, we consider each motif as one unique 10-mer nucleotide sequence. The formulation of the motif evolutionary model, however, does not impose this restriction. A motif can be a collection of sequences represented by one or multiple strings of 15 IUPAC symbols (e.g., the Tp53 binding motif is NGRCGWGYCY, where R denotes A or G, W denotes A or T, Y denotes C or T and N denotes any base). The choice of investigating single sequence motifs is based on the concern of computational efficiency. Exhaustive evaluations of selection coefficients of all composite motifs are beyond the computing capacity accessible by the authors. For instance, there are $15^{10} = 5.767 \times 10^{14}$ motifs represented by 10-mers of IUPAC symbols without gaps. The number of these composite motifs is 54 994 folds as the number of 10-mer single sequence motifs, which would take about 333 million CPU hours using the current computing infrastructure. This number will grow exponentially when combinations of 10-mer IUPAC strings and gaps are considered. Therefore, simplifications without exhausting all possible sequences are required when extending the analysis into composite motif sequences.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–5 and Supplementary Figure 1.

ACKNOWLEDGEMENTS

We thank Ker-Chau Li, Robert Shin-Sheng Yuan and Guan-I Wu for logistic assistance.

FUNDING

Academia Sinica (to C.H.Y. and D.H.C.); National Science Council in Taiwan [100-2118-M-001-008-MY2 to C.H.Y.]; National Science Council, Taiwan [100-2118-M-001-025-MY2]. Funding for open access charge: National Science Council, Taiwan [100-2118-M-001-008-MY2].

Conflict of interest statement. None declared.

REFERENCES

1. Carroll, S.B. (2005) Evolution at two levels: on genes and form. PLoS Biol., 3, e245.
2120 Nucleic Acids Research, 2013, Vol. 41, No. 4

2. Davidson,E.H. and Erwin,D.H. (2006) Gene regulatory networks and the evolution of animal body plans. Science, 311, 796–800.
3. King,M.C. and Wilson,A.C. (1975) Evolution at two levels in humans and chimpanzees. Science, 188, 107–116.
4. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory motifs. Nature, 423, 241–254.
5. Siepel,A. and Haussler,D. (2004) Combining phylogenetic and hidden Markov models in sequence analysis. J. Comput. Biol., 11, 413–428.
6. Emberly,E., Rajewsky,N. and Siggia,E.D. (2003) Conservation of regulatory elements between two species of Drosophila. BMC Bioinf., 4, 57.
7. Tautz,D. (2000) Evolution of transcriptional regulation. Cur. Opin. Genet. Dev., 10, 575–579.
8. Dermitzakis,E.T. and Clark,A.G. (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol., 19, 1114–1121.
9. Dermitzakis,E.T., Bergman,C.M. and Clark,A.G. (2003) Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. Mol. Biol. Evol., 20, 703–714.
10. Chen,K. and Rajewsky,N. (2007) The evolution of gene regulation by transcription factors and microRNAs. Nature Genet., 39, 93–103.
11. Ludwig,M.Z., Bergman,C., Patel,N.H. and Kreitman,M. (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature, 403, 564–567.
12. Ruvinsky,I. and Ruvkun,G. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
13. Roth,C., Rastogi,S., Arvestad,L., Dittmar,K., Light,S., Ekman,D. and Liberles,D.A. (2007) Evolutio after gene duplication: models, mechanisms, sequences, systems, and organisms. J. Exp. Zool. B, Mol. Dev. Evol., 308, 58–73.
14. Sinha,S. and Siggia,E.D. (2005) Sequence turnover and tandem repeats in cis-regulatory modules in Drosophila. Mol. Biol. Evol., 22, 874–885.
15. Stone,J.R. and Wray,G.A. (2001) Rapid evolution of cis-regulatory sequences via local point mutations. Mol. Biol. Evol., 18, 1764–1770.
16. MacArthur,S. and Brookfield,J.F.Y. (2004) Expected rates and modes of evolution of enhancer sequences. Mol. Biol. Evol., 21, 1064–1073.
17. Berg,J., Willmann,S. and Lassig,M. (2004) Adaptive evolution of transcription factor binding sites. BMC Evol. Biol., 4, 42.
18. Mustonen,V. and Lassig,M. (2005) Evolutionary population genetics of promoters: predicting binding sites and functional phyllogenies. Proc. Natl Acad. Sci. USA, 102, 15936–15941.
19. Doniger,S.W. and Fay,J.C. (2007) Frequent gain and loss of functional transcription factor binding sites. PLoS Comput. Biol., 3, e99.
20. Yeang,C.H. (2010) Quantifying the strength of natural selection by reduced expression. Trends Genet., 26, 477–480.
21. Balasubramanian,K., Wang,Y. and Ebert,B.L. (2008) A total gene atlas of the mouse and human protein-coding transcripts. Proc. Natl Acad. Sci. USA, 105, 6062–6067.
22. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Belenconew,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet., 40, 1413–1415.
23. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature, 456, 470–476.
24. Kim,J.D. and Kim,J. (2009) YY1's longer DNA-binding motifs. Trends Genet., 25, 477–480.
25. Robertson,A.G., Bilenky,M., Lin,K., He,A., Yuen,W., Duppin,M., Varhol,R., Teague,K., Griffith,O.L., Zhang,X. et al. (2006) cisEID: A database system for genome scale computational discovery of regulatory elements. Nucleic Acids Res., 34, D68–D73.
26. Watanabe,K., Kudo,K., Ogasawara,T. and Watanabe,A. (2006) A gene atlas of the mouse and human protein-coding transcripts. Proc. Natl Acad. Sci. USA, 103, 6062–6067.
27. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Belenconew,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet., 40, 1413–1415.
28. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature, 456, 470–476.
29. The Gene Ontology Consortium. (2008) The gene ontology project in 2008. Nucleic Acids Res., 36, D440–D444.
30. Ingason,A., Kellis,M., Siggia,E.D., de Bono,M., Jondrev,D., Day,M., Honig,L., Tung,Y., Wray,G.A. et al. (2012) A total gene atlas of the mouse and human protein-coding transcripts. Proc. Natl Acad. Sci. USA, 109, D387–D393.
31. Schaffert,S.H., Ruan,J., Nusbaum,C., Fredman,L., Liu,Y., Zhang,L. et al. (2008) A total gene atlas of the mouse and human protein-coding transcripts. Proc. Natl Acad. Sci. USA, 105, 1806–1813.
32. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
33. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
34. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
35. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
36. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
37. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
38. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
39. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
40. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
41. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
42. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
43. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
44. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.
45. Ruvkun,G. and Ruvinsky,I. (2003) Functional tests of enhancer conservation between distantly related species. Development, 130, 5133–5142.