Coordinated Genome-Wide Modifications within Proximal Promoter *Cis*-regulatory Elements during Vertebrate Evolution

Ken Daigoro Yokoyama*[,1,2], Jeffrey L. Thorne[3], and Gregory A. Wray[1]

[1]Department of Biology, Duke University

[2]Department of Biochemistry and Molecular Genetics, University of Colorado Denver

[3]Department of Statistical Genetics and Bioinformatics, Bioinformatics Research Center, North Carolina State University

*Corresponding author: E-mail: ken.yokoyama@ucdenver.edu.

Accepted: 22 November 2010

Abstract

There often exists a "one-to-many" relationship between a transcription factor and a multitude of binding sites throughout the genome. It is commonly assumed that transcription factor binding motifs remain largely static over the course of evolution because changes in binding specificity can alter the interactions with potentially hundreds of sites across the genome. Focusing on regulatory motifs overrepresented at specific locations within or near the promoter, we find that a surprisingly large number of *cis*-regulatory elements have been subject to coordinated genome-wide modifications during vertebrate evolution, such that the motif frequency changes on a single branch of vertebrate phylogeny. This was found to be the case even between closely related mammal species, with nearly a third of all location-specific consensus motifs exhibiting significant modifications within the human or mouse lineage since their divergence. Many of these modifications are likely to be compensatory changes throughout the genome following changes in protein factor binding affinities, whereas others may be due to changes in mutation rates or effective population size. The likelihood that this happened many times during vertebrate evolution highlights the need to examine additional taxa and to understand the evolutionary and molecular mechanisms underlying the evolution of protein–DNA interactions.

Key words: gene expression, transcriptional regulation, protein–DNA coevolution, *cis*-regulatory evolution.

Introduction

Changes in the *cis*-regulatory elements that control gene expression can influence a variety of functionally significant traits, including morphology, behavior, and physiology (Wray et al. 2003; Latchman 2004; Wray 2007; Tirosh et al. 2009; Li and Johnson 2010). A single transcription factor can regulate the expression of hundreds of genes, binding to a subset of commonly occurring DNA regulatory motifs in a sequence-specific manner (FitzGerald et al. 2004; Yang et al. 2007). Mutations that alter the binding affinity of a transcription factor are therefore likely to be pleiotropic and, consequently, deleterious in most cases. Indeed, DNA-binding domains are often highly conserved across species and, in many cases, appear to be under stronger evolutionary constraints than other regions of the protein (Hirsch and Aggarwal 1995; Luscombe and Thornton 2002; Rorick and Wagner 2010).

For these reasons, it is often assumed that transcription factor binding motifs remain mostly static over time, an assumption that underlies most cross-species comparisons. This includes studies regarding regulatory element conservation (Kellis et al. 2003, 2004; Xie et al. 2005; Vardhanabhuti et al. 2007), regulatory element turnover (Hare et al. 2008; Li et al. 2008), and losses and gains of regulatory element occurrences (Doniger and Fay 2007). However, analytical approaches have not been previously available. The essential challenge for in silico approaches is that the vast majority of motifs within the genome matching a particular transcription factor consensus sequence are likely to be nonfunctional, coincidental matches. Without the development of appropriate computational tools that restrict attention to

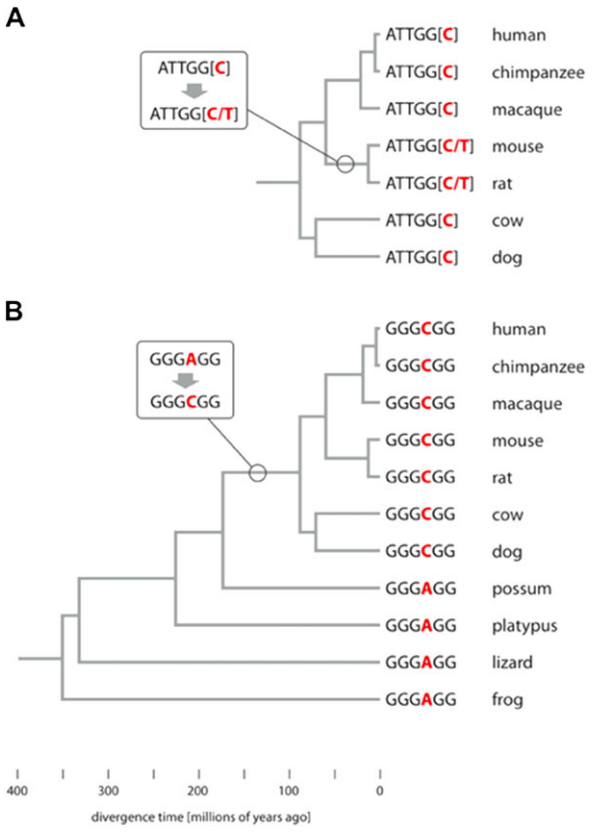

Fig. 1.—Examples of lineage-specific regulatory motif modifications. Shown are lineage-specific changes within the (*A*) NFY binding site and the (*B*) SP1 binding site. Branch lengths are drawn to scale, with evolutionary divergence times as estimated in Hedges et al. (2006).

likely functional motifs, searching for transcription factor affinity changes computationally has proven difficult.

Here, we take a genome-wide approach to assess the prevalence and nature of regulatory motif modifications within vertebrates, using location-specific overrepresentation as a way to screen out most nonfunctional motif occurrences (Yokoyama et al. 2009). We focus upon motifs overrepresented at particular locations within the proximal promoter, which comprises the region within ~200 bp of the transcription start site (TSS) (Lodish et al. 2003). Rises in the frequency of motif occurrence at specific locations relative to the TSS, such as those exhibited by the TATA-box and Inr sequence, reflect a functional role of the motif within that region (Martinez et al. 1994; Lodish et al. 2003; Juven-Gershon et al. 2008). Differences in the frequencies of location-specific motifs among species are often likely to represent compensatory substitutions within *cis*-regulatory elements throughout the genome, driven by a change in the affinity of the corresponding transcription factor (fig. 1). Alternatively, changes in the frequencies of location-specific motifs can be a consequence of changes in effective population size and mutational bias. With a small population size, natural selection is inefficient, and mildly deleterious mutations can be fixed even when they would not occur in a large population. Regardless of the mechanisms underlying *cis*-regulatory modifications, however, coordinated changes in *cis*-regulatory elements can have a substantial impact upon the organism. Such modifications alter the relationship between a transcription factor and its binding sites throughout the genome, with or without a change in the binding affinity of the protein itself.

Based upon comparisons between species, we find that a large fraction of location-specific *cis*-regulatory elements exhibit modifications in consensus binding sequences over the course of vertebrate evolution. Moreover, we show that preferred regulatory consensus sequences can differ even between relatively closely related vertebrate species, and, in many cases, these lineage-specific substitutions have occurred at hundreds of functional sites throughout the genome. These findings challenge the view that regulatory consensus motifs remain mostly static over long intervals of time and highlight a previously unrecognized mechanism driving genome evolution.

Materials and Methods

Data Preparation

We extracted a nonredundant set of human and mouse promoter sequences from the UCSC Genome Browser (http://genome.ucsc.edu; Lander et al. 2001; Waterston et al. 2002; Karolchik et al. 2004). Each data set comprised the window 500 bp prior to and 100 bp after a known TSS in RefSeq (Maglott et al. 2000; Pruitt and Maglott 2001). Orthologous promoter sequences were then determined for nine additional vertebrate species, including cow, dog, rat, chimp, macaque, opossum, platypus, lizard, and frog. Genomic coordinates of the orthologous promoter sequences were first approximated according to the genome-wide multiz28way (hg18) and multiz20way (mm9) alignments (Miller et al. 2007). For each pair of orthologous promoter sequences, we estimated the location of the TSS in the alternative species by aligning the two sequences using dynamic programming (supplementary material S1, Supplementary Material online). We then set the orthologous TSS to be the site aligned to the known human or mouse TSS in RefSeq. We applied stringent quality controls to filter low-confidence orthologs from our data sets. Orthologs were excluded if more than 25% of the aligned columns within the (−10, +10) window contained gaps or if less than 70% of the aligned columns contained matching nucleotides.

For both human and mouse, we determined a set of 6-mer motifs exhibiting location-specific overrepresentation using the Functional Region Evaluation Engine (FREE) (Yokoyama et al. 2009). The program was run genome wide on each species, and each 6-mer exhibiting location-specific overrepresentation at a P value under $P < 1 \times 10^{-15}$ was considered for subsequent analyses. We subsequently

Table 1
Cross-species Nucleotide Co-occurrences at the Sixth Site of the NFY Binding Motif (ATTGGn) within the Region of Overrepresentation

		Human			
		attggA	attggC	attggG	attggT
Mouse	attggA	43	9	14	3
	attggC	6	341	9	38
	attggG	8	3	40	1
	attggT	2	80	5	104

determined the region of overrepresentation for each motif, buffering each window by a maximum of 100 bp.

Modeling Modifications within Location-Specific Regulatory Motifs on Single Branches

For each location-specific motif, we tested for cross-species differences separately at each individual consensus site. For two species (X and Y), cross-species comparisons were conducted by considering motif occurrences targeting orthologous genes. We focused specifically upon putatively functional motif occurrences within the region of overrepresentation. Each motif co-occurrence then contains a specific nucleotide at the chosen site within each of the two species. We then determined the number of each pairwise combination of nucleotides across species genome wide, as illustrated in table 1. Our goal was to search for nonidentical nucleotides (i and j) where species X prefers nucleotide i at the consensus site while species Y prefers nucleotide j at the same site. We set a random variable (T_{ij}) that represents the number of co-occurrences of i and j at the given consensus site in species X and Y, respectively; we denote the observed value of T_{ij} to be t_{ij}.

To detect evolutionary modifications, we compare the value t_{ij} to the number of nucleotide co-occurrences after switching i and j across the two species (i.e., the value t_{ji}). With our null hypothesis that no cross-species differences exist, we would expect that $t_{ij} \approx t_{ji}$. In contrast, in the presence of evolutionary modification, we would expect a significant asymmetry between these two values, producing large positive values for the difference $t_{ij} - t_{ji}$.

Co-occurrence data were taken across the genome, providing a large sample size. Thus, we assume a normal approximation for $T_{ij} - T_{ji}$, whose null distribution is derived from background motif occurrences within a set of intergenic sequences. We set a value b_{ij} analogous to t_{ij}, which represents the number of co-occurrences of nucleotides i and j in species X and Y within the set of intergenic sequences. The significance of functional asymmetry within the region of overrepresentation can then be assessed using a Z-score. The Z-score represents the number of standard deviations by which the observed asymmetry ($t_{ij} - t_{ji}$) deviates from its expected value. This Z-score is given by

$$Z = \frac{t_{ij} - t_{ji}}{\sqrt{(N_t/N_b)(b_{ij} + b_{ji})}}, \quad (1)$$

where N_t is the total number of motif co-occurrences within the region of overrepresentation, and N_b is the number of co-occurrences within the intergenic sequences (supplementary material S2, Supplementary Material online).

Results

Location-Specific Motifs Are Shared across Closely and Distantly Related Species

Scanning for location-specific motifs within the mouse and human genomes resulted in 255 location-specific 6-mers in mouse and 212 location-specific 6-mers in humans. The majority of the predicted 6-mers exhibited locational specificity in both species, with 169 6-mers predicted in both mouse and human. Our approach explicitly accounts for fluctuations in dinucleotide frequencies within the promoter, and therefore, these location-specific motifs are not likely to be the result of fluctuations in GC content near the TSS (for discussion, see Yokoyama et al. 2009). We found that the majority of these 6-mers (80% in human and 79% in mouse) had previously documented regulatory functions, matching known binding sites in the TRANSFAC database at a STAMP E value threshold of $E < 1 \times 10^{-5}$ (Matys et al. 2003; Mahony and Benos 2007).

Comparisons across even more highly diverged species also showed a significant overlap in location-specific motifs. Even between mammals and zebrafish, 18 of the top 20 consensus motifs in humans matched similar location-specific motifs in zebrafish (table 2). This amount of overlap is striking because these predictions were generated independently on different data sets from two highly diverged species. In addition, the region of overrepresentation of these location-specific motifs was also highly preserved. Within the majority of the top-ranking motifs in human, the central location (peak) of overrepresentation differed by less than ten nucleotide sites from that in zebrafish, suggesting strong functional conservation in the location of preference across even distantly related vertebrates.

Location-Specific Motifs Evolve Differently according to Location

A natural question to ask is whether location-specific motifs are subject to different evolutionary constraints depending upon the location in which they occur. Certain *cis*-regulatory elements function specifically at the location at which they preferentially occur (Martinez et al. 1994; Lodish et al. 2003; Juven-Gershon et al. 2008) and both in silico as well as experimental evidence demonstrate that the regulatory function of location-specific motifs can change according to location (Xi et al. 2007; Tharakaraman et al. 2008). Thus, we might expect that motifs exhibiting location-specific

Table 2
Comparisons between the Top 20 Ranked Location-Specific Motifs in Humans and Location-Specific Motifs in Zebrafish

		Human			Zebrafish		
#	TF	Motif	Peak	Width	Motif	Peak	Width
1	SP1	KCCCCKCCCM	−73	100	CCCCTCCY	−67	100
2	TBP	TMTATAAAARGC	−30	6	NSTATAAAAGC	−30	6
3	NFY	AGCCAATSAG	−83	100	AGCCAATCA	−88	100
4	CREB	GTSACGTGA	−44	100	CGTGACGTC	−49	100
5	SP1	GnGGGGGGCGKG	−63	100	GGGAGGGGG	−76	100
6		GTGTGTG	−440	100			
7	NFY	CTGATTGGY	−79	100	CTGATTGGCT	−83	100
8	REST	CRCCATGGMn	+52	100	ACATGGCT	+22	64
9	ZEB1	MAGGTRAGTG	+71	100	GTAAGW	+65	89
10	ETS	SCGGAAGTG	−31	100	MGGAAGT	−21	92
11	ERF2	CAGCGGCSGC	+35	100			
12	HBP	RCGTCAC	−47	100	CACGTG	−50	100
13	E2F	TGGCGG	+26	54	TGGCGG	+18	28
14	ZFP161	YGCGCGC	−29	100	CGCGCGC	−46	100
15	CREB	ACTTCCGG	−20	74	WCTTCCT	−31	97
16	NRF1	TGCGCA	−59	100	GCATGCGCGT	−46	100
17		TCTGCTGCY	+58	100	GCTGCTGC	+49	100
18	NF-muE1	GRTGGC	+29	66	RATGGC	+16	30
19		AAAAAA	−104	100	AAAAAA	−93	100
20	YY1	ASATGG	+17	34	ACATGGCT	+22	54

NOTE.—Regulatory motifs were predicted independently for each species using separate TSS annotations in RefSeq (Maglott et al. 2000; Pruitt and Maglott 2001). The motifs were predicted using FREE (Yokoyama et al. 2009); the central location and width of overrepresentation are given to the right of each motif.

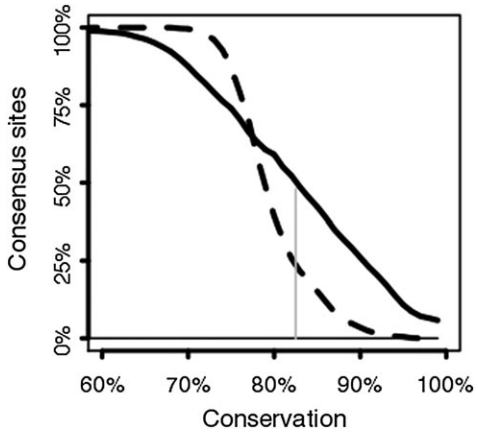

FIG. 2.—Conservation frequencies of location-specific regulatory motifs across mouse and human according to location. The *x* axis denotes the frequency of conservation at a given consensus site of an individual regulatory motif, whereas the *y* axis gives the cumulative number of consensus sites at or above the given frequency of conservation. The solid plot shows the amount of nucleotide conservation within the region of overrepresentation, whereas the dashed plot shows the amount of conservation within intergenic regions. Note that most consensus sites tend to be more conserved within the location of overrepresentation than outside this region. For instance, half of all consensus sites have 83% or more conservation within the region of overrepresentation (vertical gray line), whereas only 22% of the consensus sites are conserved at the same threshold within the intergenic sequences.

overrepresentation would be subject to different evolutionary constraints within the region of overrepresentation than outside this region.

In order to test whether location-specific motifs evolve differently according to their location, we assessed the amount of nucleotide conservation within our predicted motifs between human and mouse. The frequency of conservation was determined both inside the region of overrepresentation as well as within a set of intergenic sequences that serve as a neutral proxy. We found that most consensus sites were more conserved within the region of overrepresentation than within the intergenic regions (fig. 2), suggesting that location-specific motifs are under stronger evolutionary constraints within the region of overrepresentation than their location-independent occurrences.

Although the majority of location-specific regulatory motifs appear to have been subject to stronger purifying selection within the region of overrepresentation, certain motifs showed less conservation within their preferred location compared with the intergenic sequences. To determine whether these were simply random changes or whether they reflected lineage-specific nucleotide biases, we compared differences in nucleotide biases between human and mouse. We found that substitutions within the region of overrepresentation were frequently nucleotide-specific, systematically changing to and from particular nucleotides in a lineage-specific manner (fig. 3). In contrast, substitutions outside this region tended to accumulate randomly, without significant preferences regarding the nucleotides fixed over the course of evolution, a pattern consistent with drift.

Taken together, these results suggest that location-specific regulatory motifs are subject to different evolutionary constraints within the region of overrepresentation than within intergenic regions. Within the region of overrepresentation, the majority of motifs were found to be subject to greater amounts of purifying selection, whereas the remaining motifs generally exhibit strong nucleotide preferences within specific lineages. In contrast, the same motifs generally tend to be subject to weaker evolutionary constraints within intergenic sequences, accumulating random substitutions without any biases in nucleotide substitutions.

Many *Cis*-regulatory Elements Have Undergone Evolutionary Changes within Mammals

We scanned for patterns of evolutionary modifications within the comprehensive list of predicted 6-mer motifs exhibiting location-specific overrepresentation. For each motif, we searched for cross-species differences in nucleotide preference by focusing upon putatively functional occurrences within the region of overrepresentation. Fixing one nucleotide per species at a given consensus site, we scanned

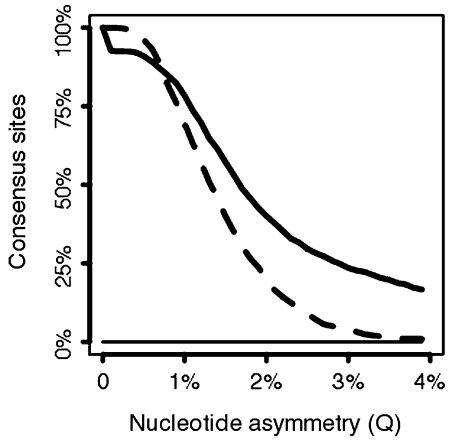

Fig. 3.—Differences in nucleotide preference across human and mouse according to location. Shown are the cumulative fractions of motif consensus sites (*y* axis) exhibiting differences, that is, asymmetries, in nucleotide preference (*x* axis). For a given consensus site, we consider the fraction of co-occurrences (q_{ij}) of nonidentical nucleotides *i* and *j* at a consensus site across human and mouse, respectively. Large values for $q_{ij} - q_{ji}$ then indicate a strong preference for nucleotides *i* and *j* to occur in human and mouse, respectively, but not vice versa. Asymmetries *Q* along the *x* axis are then defined as $Q = \sum_{i \neq j} |q_{ij} - q_{ji}|$, where the sum is across all nonidentical nucleotides. Results are shown for motif occurrences within the region of overrepresentation (solid plot) and occurrences within intergenic regions (dashed plot). We find that *Q* values are significantly greater within the region of overrepresentation, reflecting lineage-specific differences in nucleotide substitution biases. In contrast, *Q* values are significantly smaller in intergenic regions, reflecting random fixation of nucleotides without lineage-specific preferences.

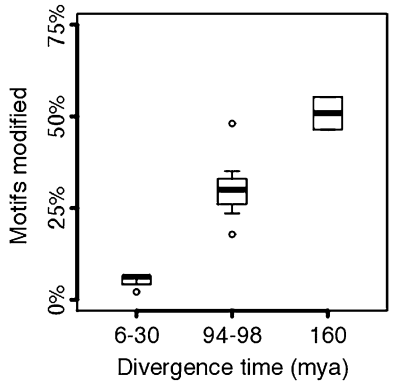

Fig. 4.—Prevalence of evolutionary modifications within location-specific consensus motifs according to divergence time. *y* axis values represent the fraction of regulatory consensus motifs exhibiting evolutionary changes in sequence following species divergence (*Z*-score > 5). Species comparisons were conducted in a pairwise fashion, each comparison producing a single set of modified motifs. The sets of modified motifs are separated according to divergence time between the corresponding pair of species (*x* axis) (Hedges et al. 2006). Each barplot shows the median fraction of modified consensus motifs (center line), the first and third quartile (bar extremes), and the most extreme comparisons within 80% of the interquartile range (standard bar). Circles represent single outlier points.

for cross-species differences by comparing the frequency of nucleotide co-occurrence with that obtained after switching nucleotides across species. Large differences in these two co-occurrence frequencies indicate that each nucleotide is found frequently in only one of the species but is far less preferred along the other lineage.

To quantify the significance of cross-species differences, we considered a "background" frequency of substitution toward each pair of nonidentical nucleotides. As motif occurrences within the intergenic regions have generally been subject to weaker evolutionary constraints, we used orthologous motif occurrences within the intergenic regions to estimate these background frequencies of substitution. The significance of cross-species differences within the region of overrepresentation was then assessed according to the background frequency of substitution (supplementary material S2, Supplementary Material online). The statistical significance of this comparison was quantified using a *Z*-score. Significant *Z*-scores then indicate regulatory motif modifications whose rate of change cannot be adequately explained by the background rate of substitution.

Using this approach, we conducted pairwise comparisons across an array of eight mammalian lineages (supplementary table 1, Supplementary Material online). We found that a large fraction of location-specific regulatory motifs have been subject to evolutionary changes in consensus sequence. For instance, we find that close to a third of all location-specific consensus motifs have been modified on either the human or mouse lineages following their divergence ($Z > 5$) (supplementary tables 2 & 3, Supplementary Material online). Comparisons between other species with divergence times similar to that of human and mouse (Hedges et al. 2006) produced similar numbers of modified consensus sequences, with predicted modifications generally occurring within 24–35% of all location-specific consensus motifs (fig. 4). There was also a strong correlation between divergence time and the prevalence of consensus sequence modification. Only 6% of all consensus motifs exhibited differences within primates, at divergence times less than 30 My (Hedges et al. 2006). In contrast, about half of all location-specific consensus motifs exhibited differences between eutherians and the more distantly related possum lineage (divergence times of 160 My [Hedges et al. 2006]).

To assess the expected number of false positives, we conducted simulation analyses using randomized co-occurrence data. For each consensus site and pair of nonidentical nucleotides, we randomized the data assuming no difference in nucleotide preferences across species (supplementary material S3, Supplementary Material online). Comparisons were then conducted in the same manner using the randomized co-occurrence data. The simulation analyses showed that with our *Z*-score threshold ($Z > 5$), we expect about one false positive prediction across all motifs and

species comparisons, demonstrating that multiple testing is unlikely to be the explanation for our findings.

We also assessed the quality of our inferred promoter orthologs. Because it has been observed that TSS locations are not always conserved across species (Frith et al. 2006), we tested for the effects of TSS turnover on our results. Comparing RefSeq promoter data in mouse with the mouse promoters inferred from the human–mouse alignments showed that location-specific motif predictions were nearly identical across the two data sets, both in sequence as well as the location and width of overrepresentation (supplementary table 4, Supplementary Material online). This was true even for motifs such as the TATA-box and Inr sequence, which are generally constrained to very precise locations within the promoter (six sites and one site, respectively). Thus, comparisons between known and inferred promoter data sets are likely to be reliable because independent analyses on the inferred promoters produced consistent location-specific motif predictions, both in consensus sequences as well as the location and width of over representation.

Predictions of Motif Modification Overlap Using a Second Statistical Model

To test the robustness of our motif predictions, we performed a second statistical assessment of motif modification using a different set of underlying assumptions. This second model considers only motif occurrences within the region of overrepresentation without the use of the background occurrences. Thus, this second approach does not compare motif evolution within the proximal promoter with that within intergenic regions. This model is instead based upon the binomial distribution, assuming that, in the absence of consensus motif modifications, there should be an equal probability of co-occurrence for nonidentical nucleotides across lineages. Again, we consider a pair of nonidentical nucleotides, fixing one of the nucleotides within each of the species. The number of occurrences should then follow the binomial distribution with probability of success being 1/2, whereas the number of trials is estimated using the total number of nucleotide co-occurrences, regardless of the species in which they occur. Statistical significance of modification can then be determined by calculating a *P* value using a two-tailed test.

Results from this alternative statistical model confirmed the high prevalence of evolutionary motif modification (supplementary material S4, Supplementary Material online). For instance, approximately 27% of all location-specific motifs exhibited evolutionary modifications between human and mouse (False discovery rate < 0.05). Many of these predicted motifs overlapped with those predicted during the previous analysis, with 15 of the top 20 motif predictions also producing significant *Z*-scores in the previous scan for modifications. The majority of the species comparisons with similar divergence times as human and mouse predicted between 12% and 33% of all location-specific motifs to exhibit modifications. These results confirm the high prevalence of evolutionary modifications within our set of predicted motifs and suggest that regulatory element modifications can be detected using statistical methods based upon different underlying assumptions.

Both Site Degeneracy and Preferred Consensus Nucleotides Change Over the Course of Evolution

Inspection of the results showed that, in many cases, motifs with evolutionary modifications exhibited differences in the amount of degeneracy. For instance, the NFY binding site exhibited more degeneracy at the sixth site in mouse than in humans (table 1). Although both species favor the "C" consensus nucleotide at this site, we found that substitutions to the degenerate "T" nucleotide have occurred about twice as frequently in mouse than in humans. We observe a total of 433 occurrences of the motif with the preferred C in humans, where 92 co-occurred with a degenerate nucleotide (i.e., A, G, or T) at this site in mouse. In contrast, only 53 of the NFY binding site occurrences with a C in mouse contained a degenerate nucleotide at this site in human. Other species comparisons showed that the T nucleotide was found frequently at this site only within the mouse and rat lineages but not for other eutherians, suggesting a branch-specific gain of degeneracy along the rodent lineage (fig. 1).

We note that there is more than one possible explanation for these observed nucleotide-pair asymmetries. First, it is possible that substitutions toward the T nucleotide have accelerated within the rodent branch, causing the asymmetry in C/T co-occurrences across species. However, it is also possible that substitutions toward the T nucleotide in mouse have occurred in a nearly neutral rate, whereas the same site is under strong purifying selection within other eutherian lineages. In either case, we can infer cross-species differences in the frequency of degenerate nucleotides, regardless of the underlying mechanism driving this asymmetry.

Although we found that most of our regulatory motifs with significant *Z*-scores exhibited differences in the amount of degeneracy, many other motifs differed in the most commonly occurring nucleotide consensus. Between human and mouse, about 16–19% of all location-specific motifs differed in their most common nucleotide sequences. Many of these motifs, although not all, produced highly significant *Z*-scores. Similar numbers of motifs (approximately 13–22%) showed cross-species differences in their preferred nucleotide sequences across other eutherians. Thus, despite the relatively small amount of divergence time between the various eutherian lineages, regulatory element modification appears to have been rather common during eutherian evolution. Note that our analysis underestimates the actual

Table 3
Motif Co-occurrence Frequencies across Each Strand of the GC/GA Box in Its Well-Studied Form (gggCgg) and Its Inferred Ancestral Form (gggAgg)

	Mouse (Forward Strand)			Mouse (Reverse Strand)		
		gggCgg (%)	gggAgg (%)		ccGccc (%)	ccTccc (%)
Human	gggCgg	49	4	ccGccc	45	4
	gggAgg	5	20	ccTccc	5	23
Possum	gggCgg	28	5	ccGccc	28	5
	gggAgg	17	20	ccTccc	15	23
Platypus	gggCgg	16	2	ccGccc	12	3
	gggAgg	11	27	ccTccc	13	29
Lizard	gggCgg	32	7	ccGccc	23	3
	gggAgg	18	15	ccTccc	27	13
Frog	gggCgg	21	3	ccGccc	26	6
	gggAgg	21	24	ccTccc	22	20

frequency of such changes because we considered only a fraction of all mammalian species.

The GC Box Regulatory Motif Has Been Modified along the Eutherian Branch

One notable example of such a modification is the GC box motif. The commonly described form of this element (gggCgg) in fact represents an altered version of its ancestral sequence in mammals because it is consistently found among non-eutherians as the gggAgg consensus sequence. As the term "GC box" was derived from its well-studied consensus sequence, we refer to the predicted ancestral sequence as the "GA box." There is a striking pattern of GC/GA box co-occurrences between eutherians and non-eutherians, respectively, with the ancestral form commonly appearing in lineages ranging from possum to frog. This pattern was found upon both strands of this regulatory element (table 3). The consistency of the preferred GA box motif within non-eutherians indicates that this regulatory element has been modified following the split with possum but prior to the divergence of the various eutherian lineages (fig. 1). Separate analyses conducted upon zebrafish showed significant amounts of location-specific overrepresentation of the ancestral form but no locational specificity of the GC box form, suggesting that the common eutherian version of this regulatory element is largely nonfunctional along the zebrafish lineage.

As this regulatory motif is highly prevalent throughout the genomes of many vertebrates, the conversion of the ancestral form to its common eutherian version represents sequence modifications across hundreds of functional sites genome wide. Over 9% of all orthologous target genes contained a GC/GA box co-occurrence between mouse and possum, respectively, whereas the reverse co-occurrence was only about half as common. As it has been estimated that possum and eutherians share approximately 15,000 orthologous genes (Mikkelsen et al. 2007), we estimate a genome-wide difference of ~600 more genes containing a GC/GA box co-occurrence in eutherians and non-eutherians, respectively, than vice versa. The rate of modification for this element was particularly high relative to the background rate, producing a Z-score of $Z = 8.5$ in the mouse–possum comparison.

Discussion

Our analyses reveal that *cis*-regulatory consensus motifs are not nearly as static as commonly assumed. The assumption that regulatory consensus sequences change little over time underlies many studies conducting cross-species comparisons (Kellis et al. 2003, 2004; Xie et al. 2005; Hare et al. 2008; Li et al. 2008). This assumption is often based upon the high level purifying selection within most DNA binding domains, whose conservation across species is often stronger than other parts of the protein factors (Hirsch and Aggarwal 1995; Luscombe and Thornton 2002; Rorick and Wagner 2010). Although some transcription factors may preferentially bind to consensus motifs that are highly conserved across species, the findings presented here suggest that this is not always the case. Our observations show that a significant number of regulatory consensus sequences differ even within eutherians, suggesting that compensatory motif modifications are far more common than previously recognized.

We focus here upon motifs that are found overrepresented at particular locations within the proximal promoter region. We predicted location-specific motifs accounting for dinucleotide fluctuations in the promoter. Our method effectively eliminates incorrect predictions due to rises in GC content near the TSS (Yokoyama et al. 2009). This is supported by the observation that our GC-rich motifs are found overrepresented at locations other than the TSS, usually ~40–100 bp either upstream or downstream of the TSS, whereas GC content rises directly across the start of transcription (Yokoyama et al. 2009). The vast majority (~80%) of our location-specific motifs match known transcription factor binding sites in TRANSFAC (Matys et al. 2003). Thus, the genome-wide trends presented here are likely to reflect evolutionary changes in true *cis*-regulatory elements.

Locational specificity allows us to distinguish functional motif occurrences that are likely to play a role in gene regulation from the remaining background motif occurrences exhibiting weaker evolutionary constraints. It has previously been observed that the position of location-specific motifs within or outside the region of overrepresentation has a significant effect upon regulatory function (Xi et al. 2007; Tharakaraman et al. 2008). This unique characteristic therefore offers a convenient way by which to study regulatory element evolution. In contrast, distinguishing between functional and nonfunctional occurrences of

motifs without locational specificity, such as those occurring in distal enhancers, is less straightforward because there is no clear way to distinguish functional from nonfunctional motif occurrences comprehensively in silico. Although there is an inherent interest in understanding the evolution of these other regulatory elements, in the absence of effective computational approaches, the prevalence of compensatory mutations within such elements remains unclear.

One can imagine a variety of evolutionary mechanisms operating within regulatory consensus sequences. In the simplest scenario, differences in binding site motifs among species may reflect evolutionary changes in transcription factor binding affinities. Such changes would likely favor compensatory mutations at functional motif locations throughout the genome, preserving the trans-factor's ability to bind near the same set of target genes and resulting in nucleotide-specific substitution patterns. Changes in a binding site consensus motif do not necessarily imply that the regulatory function of the transcription factor has become altered. In fact, genome-wide conversions to a new preferred consensus sequence may reflect preservation of protein function specifically in genes targeted by the modified *cis*-regulatory elements. These cases represent a dichotomy between functional preservation and sequence conservation, as it is likely that sites converted to the modified sequence element would continue to recruit the trans-acting factor, preserving the original function. In contrast, sites conserved in sequence would likely lose the ability to bind the altered protein factor, altering the target gene's expression pattern. This phenomenon may, in part, explain recent findings that conservation in expression patterns is largely uncorrelated with sequence conservation in nonexonic sequences (Chan et al. 2009).

Although one-to-one modification between a transcription factor and its binding affinity is a convenient explanation of our observations, it is possible that the mechanisms underlying regulatory motif modification may be more complex in some cases. For example, multiple forms of a trans-factor can often exist within a single species, and these paralogous or alternatively spliced proteins may share similar, yet distinct, DNA binding sequence preferences. Losses, gains, and modifications of paralogous transcription factors over the course of evolution have been shown to contribute to changes in morphology (e.g., Huntley et al. 2006; Nowick and Stubbs 2010). In other cases, coordinated changes in *cis*-regulatory elements may reflect regulatory element "handover," where a protein factor regulating a set of target genes is replaced by an alternate transcription factor. Such trans-factor handovers have recently been shown to be quite common in regulatory networks (Li and Johnson 2010).

It is possible that other mechanisms may also play a role in motif frequency changes, including posttranslational modifications of proteins, mutational biases, and changes in effective population size (Berg et al. 2004; Chernatynskaya et al. 2009; Ezkurdia et al. 2009). For instance, methylation of the cytosine ring in CG dinucleotides causes increased rates of biased mutation (Baele et al. 2008, 2010; Illingworth and Bird 2009) and may well influence evolutionary changes in GC-rich motifs. The combination of such mutational biases and changes in effective population size may account for some of the cross-species differences observed in our regulatory motifs. Motif frequencies are the outcome of a balance between mutation, natural selection, and genetic drift. The probability that a new mutation fixes depends on the product of the effective population size and the difference in relative fitness between the mutant and mutated motif.

Although some *cis*-regulatory modifications may occur without changes in protein binding affinities, the genome-wide binding site modifications we document here are likely to have functional consequences. If a protein's binding affinities remain unchanged, *cis*-regulatory modifications (e.g., due to mutational bias or changes in effective population size) will naturally alter the protein's ability to bind to the modified sites. Such cases can cause genome-wide changes in gene expression due to loss or modification of transcription factor binding. Additional studies will be needed in order to identify the relative contribution of selection and drift and of direct and indirect molecular interactions to evolutionary changes in binding site motifs throughout the genome. Regardless of the underlying evolutionary and molecular mechanisms, it is clear that coordinated changes in location-specific motifs have occurred many times during vertebrate evolution and embody a previously unexplored aspect of genome evolution.

Supplementary Material

Supplementary materials and tables 1–4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We would like to thank David Garfield, Courtney Babbitt, Lisa Warner, and other members of the Wray laboratory as well as members of the Uwe Ohler laboratory for their comments. This work was supported by the National Institute of Health (GM090201, GM070806 to J.L.T. and 5P50-GM-081883 to G.A.W.) and the National Science Foundation (0614509 to G.A.W).

Literature Cited

Baele G, Van de Peer Y, Vansteelandt S. 2008. A model-based approach to study nearest-neighbor influences reveals complex substitution patterns in non-coding sequences. Syst Biol. 57:675–692.

Baele G, Van de Peer Y, Vansteelandt S. 2010. Modelling the ancestral sequence distribution and model frequencies in context-dependent models for primate non-coding sequences. BMC Evol Biol. 10:244.

Berg J, Willmann S, Lassig M. 2004. Adaptive evolution of transcription factor binding sites. BMC Evol Biol. 4:42.
Chan ET, et al. 2009. Conservation of core gene expression in vertebrate tissues. J Biol. 8:33.
Chernatynskaya AV, Deleeuw L, Trent JO, Brown T, Lane AN. 2009. Structural analysis of the DNA target site and its interaction with Mbp1. Org Biomol Chem. 7:4981–4991.
Doniger SW, Fay JC. 2007. Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol. 3:e99.
Ezkurdia I, et al. 2009. Progress and challenges in predicting protein-protein interaction sites. Brief Bioinform. 10:233–246.
FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C. 2004. Clustering of DNA sequences in human promoters. Genome Res. 14:1562–1574.
Frith MC, et al. 2006. Evolutionary turnover of mammalian transcription start sites. Genome Res. 16:713–722.
Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. 2008. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet. 4:e1000106.
Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 22: 2971–2972.
Hirsch JA, Aggarwal AK. 1995. Structure of the even-skipped homeodomain complexed to AT-rich DNA: new perspectives on homeodomain specificity. EMBO J. 14:6280–6291.
Huntley S, et al. 2006. A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res. 16:669–677.
Illingworth RS, Bird AP. 2009. CpG islands—"a rough guide". FEBS Lett. 583:1713–1720.
Juven-Gershon T, Hsu JY, Theisen JW, Kadonaga JT. 2008. The RNA polymerase II core promoter—the gateway to transcription. Curr Opin Cell Biol. 20:253–259.
Karolchik D, et al. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32:D493–D496.
Kellis M, Patterson N, Birren B, Berger B, Lander ES. 2004. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. J Comput Biol. 11:319–355.
Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 423:241–254.
Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.
Latchman DS. 2004. Eukaryotic transcription factors. Boston: Elsevier Academic Press.
Li H, Johnson AD. 2010. Evolution of transcription networks—lessons from yeasts. Curr Biol. 20:R746–R753.
Li XY, et al. 2008. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol. 6:e27.
Lodish H, et al. 2003. Molecular cell biology. 4th ed. New York: W.H. Freeman and Company.
Luscombe NM, Thornton JM. 2002. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol. 320:991–1009.
Maglott DR, Katz KS, Sicotte H, Pruitt KD. 2000. NCBI's LocusLink and RefSeq. Nucleic Acids Res. 28:126–128.
Mahony S, Benos PV. 2007. STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. 35:W253–W258.
Martinez E, Chiang CM, Ge H, Roeder RG. 1994. TATA-binding protein-associated factor(s) in TFIID function through the initiator to direct basal transcription from a TATA-less class II promoter. EMBO J. 13:3115–3126.
Matys V, et al. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 31:374–378.
Mikkelsen TS, et al. 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature. 447:167–177.
Miller W, et al. 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res. 17:1797–1808.
Nowick K, Stubbs L. 2010. Lineage-specific transcription factors and the evolution of gene regulatory networks. Brief Funct Genomics. 9:65–78.
Pruitt KD, Maglott DR. 2001. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 29:137–140.
Rorick MM, Wagner GP. 2010. The origin of conserved protein domains and amino acid repeats via adaptive competition for control over amino acid residues. J Mol Evol. 70:29–43.
Tharakaraman K, Bodenreider O, Landsman D, Spouge JL, Marino-Ramirez L. 2008. The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. Nucleic Acids Res. 36:2777–2786.
Tirosh I, Barkai N, Verstrepen KJ. 2009. Promoter architecture and the evolvability of gene expression. J Biol. 8:95.
Vardhanabhuti S, Wang J, Hannenhalli S. 2007. Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. Nucleic Acids Res. 35:3203–3213.
Waterston RH, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520–562.
Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 8:206–216.
Wray GA, et al. 2003. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol. 20:1377–1419.
Xi H, et al. 2007. Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. Genome Res. 17:798–806.
Xie X, et al. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature. 434:338–345.
Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E. 2007. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene. 389:52–65.
Yokoyama KD, Ohler U, Wray GA. 2009. Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships. Nucleic Acids Res. 37:e92.

Associate editor: Chung-I Wu