Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites

Xiaohui Xie[†], Tarjei S. Mikkelsen[†‡], Andreas Gnirke[†], Kerstin Lindblad-Toh[†], Manolis Kellis[†§], and Eric S. Lander[†¶‖††]

[†]Broad Institute of MIT and Harvard, Massachusetts Institute of Technology and Harvard Medical School, Cambridge, MA 02142; [‡]Division of Health Sciences and Technology, [§]Computer Science and Artificial Intelligence Laboratory, and [‖]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and [¶]Whitehead Institute for Biomedical Research, Cambridge, MA 02142

Contributed by Eric S. Lander, March 3, 2007 (sent for review January 26, 2007)

Conserved noncoding elements (CNEs) constitute the majority of sequences under purifying selection in the human genome, yet their function remains largely unknown. Experimental evidence suggests that many of these elements play regulatory roles, but little is known about regulatory motifs contained within them. Here we describe a systematic approach to discover and characterize regulatory motifs within mammalian CNEs by searching for long motifs (12–22 nt) with significant enrichment in CNEs and studying their biochemical and genomic properties. Our analysis identifies 233 long motifs (LMs), matching a total of ≈60,000 conserved instances across the human genome. These motifs include 16 previously known regulatory elements, such as the histone 3′-UTR motif and the neuron-restrictive silencer element, as well as striking examples of novel functional elements. The most highly enriched motif (LM1) corresponds to the X-box motif known from yeast and nematode. We show that it is bound by the RFX1 protein and identify thousands of conserved motif instances, suggesting a broad role for the RFX family in gene regulation. A second group of motifs (LM2*) does not match any previously known motif. We demonstrate by biochemical and computational methods that it defines a binding site for the CTCF protein, which is involved in insulator function to limit the spread of gene activation. We identify nearly 15,000 conserved sites that likely serve as insulators, and we show that nearby genes separated by predicted CTCF sites show markedly reduced correlation in gene expression. These sites may thus partition the human genome into domains of expression.

comparative genomics | conserved noncoding element

Comparative analysis of the human and several other mammalian genomes has revealed that 5% of the human genome is under purifying selection, with less than one-third of the sequences under selection encoding proteins. The vast majority lies in hundreds of thousands of conserved noncoding elements (CNEs). The functional significance of these CNEs is largely unknown. It seems likely that many are involved in gene regulation, and transgenic experiments have identified some CNEs that are capable of driving highly specific spatiotemporal gene expression patterns (1–4). However, little is known about regulatory motifs contained within CNEs or proteins that recognize these elements.

We and others have previously undertaken large-scale efforts to discover conserved motifs in limited subsets of the human genome (5–8), specifically, gene promoters and 3′-UTRs. The approach has been to search for motifs that are preferentially conserved in these regions by using syntenic alignments of human, mouse, rat, and dog sequences (5). Using this approach we have discovered 174 motifs in promoter regions (within 2 kb of the transcriptional start), most of which are involved in transcriptional regulation and in tissue-specific gene expression control, and 105 motifs in 3′-UTRs, implicated in posttranscriptional regulation with half related to microRNA targeting. These studies were limited in scope because gene promoters and 3′-UTRs contain only a small fraction (≈6%) of the CNEs in the genome. In addition, they were limited in power because they involved comparison with only three non-human mammals.

Here we use the recent availability of sequences of 12 mammalian genomes to extend our motif discovery efforts to the entire human genome. We focus specifically on long regulatory motifs, between 12 and 22 nt, which provide a strong signal for motif discovery. We searched for motifs that are enriched in CNE regions relative to the rest of the genome.

We discovered >200 motifs showing striking enrichment in CNE regions. The analysis automatically rediscovered a dozen previously known regulatory elements. More importantly, most of the discovered motifs are new and show properties distinct from typical promoter elements. In particular, one of the novel motifs defines ≈15,000 potential insulator elements in the human genome, highlighting the diverse role of the CNEs in gene regulation.

Results

Creating a Motif Catalog. We began by compiling a data set of 829,730 CNEs in the human genome (totaling 62 Mb or ≈2% of the euchromatic genome), consisting of sequences showing strong conservation in syntenic regions in comparisons involving 12 mammalian genomes [see supporting information (SI) *Text*]. The vast majority of these elements are located at a considerable distance from the transcriptional start sites (TSS) of protein-coding genes (SI Fig. 4). Approximately 95% are located >2 kb away from the TSS of any gene, and half are >100 kb from a TSS. This suggests that only a small portion of the CNEs serve functions specific to core or proximal promoters.

We sought to create a catalog of sequence motifs enriched in the CNEs (SI Fig. 5). We began by identifying *k*-mers (for $k \geq 12$) that occur at a significantly higher frequency in the CNE sequences than in the remainder of the genome. We focused only on relatively long *k*-mers, because the expected number *N* of random occurrences in the entire CNE database is small (for example, $n < 8$ for $k = 12$; SI Fig. 6). We identified a total of 69,810 enriched *k*-mers. An example is 5′-GTTGCCATGGAAAC-3′, which appears 698 times in the CNE data set, whereas only 27 sites are expected based on its genome-wide frequency (26-fold enrichment). We noticed that many of enriched *k*-mers were closely related; therefore, we clustered them based on sequence similarity. The 69,810 enriched *k*-mers collapsed into 233 distinct groups, denoted LM1, LM2, etc.

Author contributions: X.X. and E.S.L. designed research; X.X. and E.S.L. performed research; X.X., T.S.M., A.G., K.L.-T., M.K., and E.S.L. contributed new reagents/analytic tools; X.X., T.S.M., A.G., K.L.-T., M.K., and E.S.L. analyzed data; and X.X., T.S.M., A.G., K.L.-T., M.K., and E.S.L. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Abbreviations: CNE, conserved noncoding element; TSS, transcriptional start site; PWM, positional weight matrix; NRSE, neuron-restrictive silencer element.

[††]To whom correspondence should be addressed. E-mail: lander@broad.mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0701811104/DC1.

Table 1. List of top 50 most highly enriched motifs

ID	*k*-mer sequence	No. of allowed mismatches	No. of initial sites in CNEs	Fold enrichment	*Z*-score	Known motif
LM1	GTTGCCATGGAAAC	1	698	25.9	130.0	X-box
LM2	ACCACTAGATGGCA	1	305	22.9	80.4	
LM3	GTTGCTAGGCAACC	1	204	30.7	76.9	
LM4	GCCTGCTGGGAGTTGTAGTT	3	143	26.3	59.2	
LM5	AACTCCCATTAGCGTTAATGG	3	43	68.1	53.5	
LM6	AAAGGCCCTTTTAAGGGCCAC	3	48	46.2	46.3	Histone 3′-UTR
LM7	CAGCAGATGGCGCTGTT	2	97	22.1	44.4	
LM8	ATGAATTATTCATG	1	280	8.8	44.3	
LM9	TCAGCACCACGGACAG	1	82	25.6	44.2	NRSE
LM10	CTGTTTCCTTGGAAACCAG	3	165	9.3	35.2	
LM11	GAAATGCTGACAGACCCTTAA	3	41	30.7	34.5	
LM12	TGGCCTGAAAGAGTTAATGCA	3	51	22.8	32.7	
LM13	TGCTAATTAGCA	0	82	13.1	30.4	CHX10
LM14	ATCCAGATGTTTGGCA	1	33	27.0	28.9	RP58
LM15	CATTTGCATGCAAATGA	2	124	8.5	28.8	
LM16	TTGAGATCCTTAGATGAAAG	3	64	14.6	28.6	
LM17	CATCTGGTTTGCAT	1	117	8.8	28.5	
LM18	CATTTGCATCTGATTTGCAT	3	80	11.8	28.2	
LM19	TGCTAATTAGCAGC	1	88	10.8	28.1	
LM20	TGACAGCTGTCAAA	1	118	8.5	28.0	
LM21	ATTTGCATCTCATTTGC	2	123	8.2	27.9	
LM22	CAGCTGTTAAACAGCTG	2	80	11.4	27.6	
LM23	AGCACCACCTGGTGGTA	2	65	13.4	27.5	
LM24	AGAACAGATGGC	0	70	12.1	26.8	TAL1BETAITF2
LM25	AAAAGCAATTTCCT	1	202	5.3	26.7	
LM26	TAAACACAGCTG	0	83	10.2	26.3	
LM27	CATTTGCATCTCATTAGCA	3	110	8.0	26.1	
LM28	AGAACATCTGTTTC	1	144	6.3	25.5	
LM29	GCTAATTGCAAATG	1	98	8.4	25.3	
LM30	CTTTGAAATGTCAA	1	182	5.3	25.3	
LM31	CTTTTCATCTTCAAAGCACTT	3	57	13.0	25.2	
LM32	CTGACATTTCCAAA	1	174	5.4	25.0	
LM33	GTAATTGGAAACAGCTG	2	69	10.7	24.8	
LM34	GATTTGCATTGCAAATG	2	84	8.8	24.1	
LM35	ACTTCAAAGGGAGC	1	87	8.5	24.1	
LM36	GAAATGCAATTTGC	1	125	6.4	24.1	
LM37	ATGCAAATGAGCCC	1	85	8.5	23.9	
LM38	GCAAATTAGCAGCT	1	82	8.5	23.4	
LM39	GTCTCCTAGGAAAC	1	84	8.4	23.4	
LM40	TCCCATTGACTTCAATGGGA	3	44	14.2	23.4	
LM41	TTTGAAATGCTAATG	1	80	8.6	23.2	
LM42	AAGCCTAATTAGCA	1	69	9.6	23.1	
LM43	CAGGAAATGAAA	0	141	5.6	23.1	
LM44	GTGTAATTGGAAACAGCTG	3	75	8.9	23.0	
LM45	GCTAATTGGATTTG	1	76	8.7	22.9	
LM46	AACAGCTGTTGAAA	1	128	5.9	22.9	
LM47	AGAGTGCCACCTACTGAAT	3	65	9.8	22.7	
LM48	TAATGAGCTCATTA	1	108	6.5	22.6	
LM49	GTAATTAGCAGCTG	1	68	9.3	22.5	
LM50	TGGGTAATTACATTCTG	2	65	9.6	22.5	

for "long motif." For each of these motifs we derived a positional weight matrix (PWM) representation reflecting the distribution of 4 nt at each position. The enrichment of each motif in the CNE data set was expressed in an enrichment score (see *Methods*). The top 50 motifs are shown in Table 1, and a full list of 233 motifs is given in SI Table 3. The motifs range in size from 12 to 22 bases.

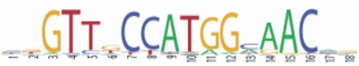
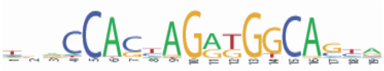
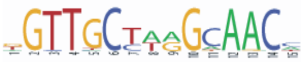
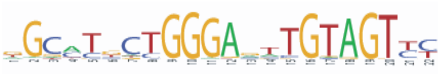
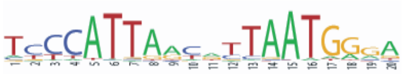
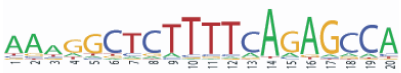
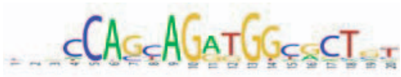
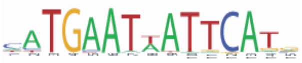
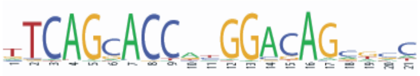
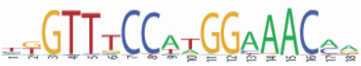
For each of 233 discovered motifs, we searched the entire human genome to identify conserved instances; that is, we identified all human sites matching the PWM and then found those sites that show clear cross-species conservation (see *SI Text*). We found a total of 60,019 conserved instances, with roughly half residing within the CNE data set and roughly half in the remainder of the genome. Importantly, the approach of focusing on motifs enriched in the CNE data set identified many motif instances elsewhere in the genome.

To assess the significance of these results, the procedure was repeated with matched control motifs. For each of the 233 motifs we created a control motif by permuting the columns of the PWM while preserving the occurrence of CpG dinucleotides. These control motifs have only 3,081 conserved instances, which is 20-fold lower than for the discovered motifs. These results indicate that only a small fraction of the 60,019 instances of the discovered motifs are likely to have occurred purely by chance.

The number of conserved instances is highly uneven across the motifs (range 37–7,549, with mean of 266 and median of 61). Most motifs (67%) have <100 conserved instances (SI Table 3). But, remarkably, the two motifs with the highest enrichment scores, LM1 and LM2, both have >5,000 conserved instances in the human genome (Table 2), suggesting a widespread functional role for these elements.

Characterizing the Discovered Motifs. ***Known regulatory elements.*** Among the 233 discovered motifs, 16 match known regulatory elements (Table 1). For example, the LM9 motif is nearly identical

Table 2. Properties of the top 10 discovered motifs

ID	Motif profile	No. of conserved instances	False positive rate*	Conservation rate,† %	Fold increase in conservation rate‡	Correlation between cross-species conservation and motif profile	Positional bias around TSS§
LM1		5,332	0.050	29.3	9.5	0.92	
LM2		7,549	0.048	29.4	14.0	0.91	
LM3		844	0.048	40.1	14.3	0.94	
LM4		1,877	0.046	20.3	13.5	0.89	20.3
LM5		224	0.042	19.4	16.3	0.87	
LM6		79	0.026	20.1	10.1	0.81	25.5
LM7		6,302	0.048	21.6	10.3	0.72	
LM8		608	0.047	17.2	9.6	0.68	
LM9		1,443	0.039	11.8	8.4	0.90	6.1
LM10		5,914	0.050	14.5	6.6	0.77	

*The proportion of conserved instances expected to have occurred by chance.

†The proportion of instances detected in human that are also conserved in orthologous regions of other mammals.

‡Compared to the conservation rates of control motifs.

§Fold enrichment on the number of motif sites located within 1 kb of TSS over those for control motifs. Only motifs with fold enrichment above 4 are shown.

to the consensus sequence of the neuron-restrictive silencer element (NRSE). The NRSE is recognized by the transcription factor REST (RE1 silencing transcription factor), which plays a pivotal role in repressing the expression of neuronal genes in nonneuronal tissues (9–11). Between 800 and 1,900 NRSE sites have been estimated to exist in the human genome (12, 13), which is consistent with our count of the number of conserved instances (1,443). It is reassuring to note that our procedure identified the LM9 motif without any prior knowledge, recovering the correct size of NRSE and showing nearly perfect similarity to NRSE along all of its 21 positions (SI Fig. 7).

Another example is LM6, which is a well studied RNA motif that is present exclusively in the 3′-UTRs of genes encoding histone proteins. In histone mRNAs this sequence is known to fold into a stem-loop structure involved in posttranscriptional regulation, playing a role similar to poly(A) tails on typical mRNAs (14, 15).

Conservation properties. The discovered motifs have two notable conservation properties. First, they show a much higher conservation rate than the control motifs, even outside the CNEs where they were discovered. The conservation rate was defined as the ratio of conserved instances to total instances in the human genome. All of the discovered motifs have a conservation rate that is 2-fold higher than for their matched controls, and 65% have a rate that is 5-fold higher (SI Table 4). If the conservation rate is computed based only on motif instances outside the CNEs, 96% of the discovered motifs have a conservation rate that is 2-fold higher than for their controls, and 63% have a rate that is 5-fold higher.

Second, the motifs show similar patterns of cross-species conservation and within-species conservation (Fig. 1 *a* and *b*). For each motif we asked whether the most conserved positions across the various motif instances within human (and thus those most likely to be involved in motif recognition) are also the positions within individual instances that show the highest conservation across species (and thus are most constrained in their evolution). To measure the within-species conservation of a motif we used the information content (I_k) of its PWM at the position k. To quantify its cross-species conservation we identified its instances located within the CNEs and calculated the proportion (M_k) of the instances with bases not mutated in the orthologous regions of the mouse or dog genomes at position k of the motif. The correlation coefficient between I and M for each of the discovered motifs is shown in Fig. 1*c*. We found that nearly all motifs (95%) show a positive correlation, and 53% have correlation coefficient >0.5. This suggests that the discovered motifs are indeed functional. The results also suggest that these motifs retain similar recognition properties across species.

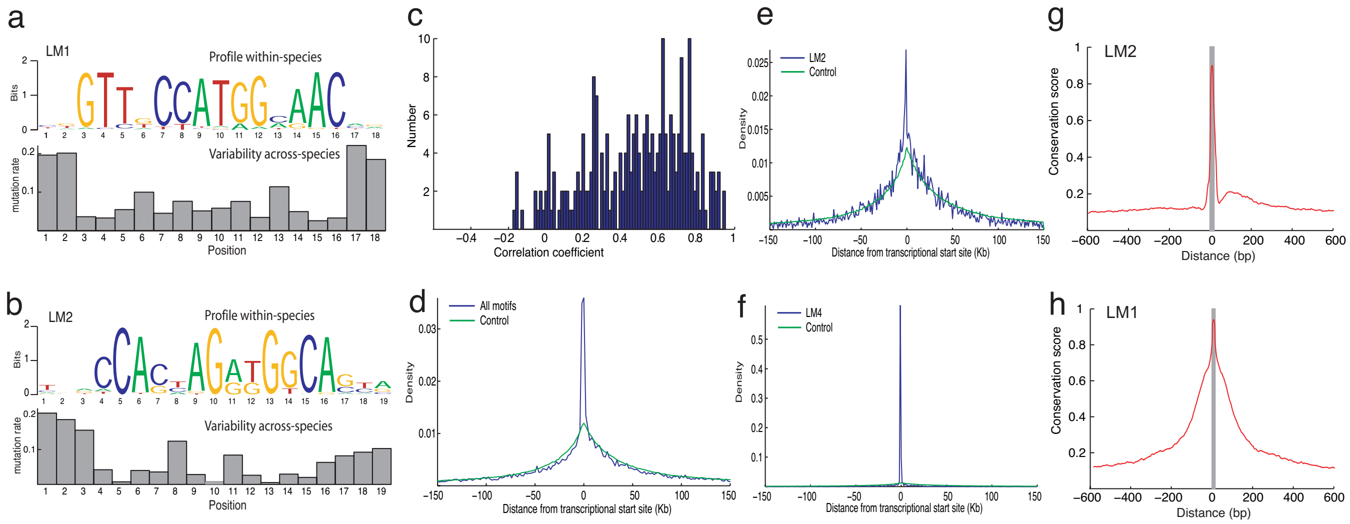


Fig. 1. A summary on properties of the discovered motifs. (*a* and *b*) Motif profile within species and variability across species for LM1 (*a*) and LM2 (*b*). Positions with high information content are less variable than those with low information content in cross-species comparison. (*c*) The correlation coefficients between motif profile and across-species conservation pattern for all discovered motifs. (*d–f*) The location of motif sites relative to TSS for all motifs (*d*), LM2 (*e*), and LM4 (*f*). They demonstrate that most of the discovered motifs, in particular LM2, are broadly distributed relative to TSS, not much different from the distribution of control sites randomly drawn from the genome (green lines). (*h*) Sequences surrounding LM1 sites are also conserved, in contrast to those surrounding LM2 sites (*g*). Gray bars show locations of the motifs. Conservation scores are phastCons scores (28) averaged over motif sites.

Palindromes. A significant proportion (17%) of the 233 motifs are palindromes, forming perfect or nearly perfect matches to their reverse complement over nearly their entire length. For example, LM3 consists of GTTGCY juxtaposed with its reverse complement, RGCAAC, with a central W, itself a self-palindrome (W = A/T). The proportion of palindromes is much higher than for random control sequences (0.13%) (see *SI Text*) and is similar to the proportion seen for the 16 known motifs (18%). The enrichment is especially pronounced among the 20 top-scoring motifs, with 45% being palindromic. Notably, the palindromic motifs are also symmetric in the information content of each base, and weakly specified positions are symmetrically placed with weakly specified positions on the two motif halves. Palindromicity can be indicative of DNA sequences that bind by a protein homodimer. Alternatively, palindromicity can sometimes reflect RNA sequences that form stem-loop structures, as illustrated by the LM6 motif in the 3′-UTRs of histone genes.

Distance from transcriptional starts. Most of the discovered motifs show little or no enrichment near genes. More than 93% have 80% of their conserved instances located >10 kb away from the TSS of any gene (Fig. 1*d*). A typical example is the LM2 motif (Fig. 1*e*). Most of these motifs are likely not to be related to core and proximal promoter functions, but may instead encode distal regulators, insulators, or other functions.

There are five cases, however, with a strong preference for being located near gene starts. A striking example is the LM4 motif, for which ≈60% of the conserved instances lie within 1 kb of the TSS (26-fold enriched over random expectation) and the modal distance is 75 bases upstream of the TSS (Fig. 1*f*). Another example is LM100, a palindromic sequence for which 45% of conserved instances lie within 2 kb of a TSS. These motifs are likely to be related to core and proximal promoter functions.

Local conservation context. We studied the conservation context of the discovered motifs. Because CNE sequences used to discover the motifs tend to occur in large blocks (N50 length = 110 bases, where N50 length is the length x such that 50% of all CNE bases lie in CNEs with the size $\geq x$), conserved motif occurrences lying within CNEs would be expected to be embedded within blocks of conserved sequence. This is indeed the case. For each motif M we examined the block of conserved sequence surrounding the each conserved occurrence in a CNE and defined $d_1(M)$ to be the N50 length of the block. The median value of $d_1(M)$ is 112 bases, with an interquartile range of 88–140 bases.

More revealingly, we examined the corresponding value $d_2(M)$ defined for conserved motif instances that lie outside the CNEs data set. The median value of $d_2(M)$ is 96 bases (interquartile range of 61–133 bases), which is similar to $d_1(M)$. This indicates that the discovered motifs typically function as part of regulatory modules containing many other regulatory elements. These results suggest that CNE motifs here may provide a useful initial entry point for studying the function of diverse large CNEs, including ultra-conserved elements that have been shown to have enhancer function.

Although most motifs appear to function in concert with others, we found eight striking examples among the 233 motifs that appear to act in isolation. This is true both for conserved occurrences within and outside the CNE data set. These motifs are LM9 (NRSE), LM6 (the histone 3′-UTR element), LM4 (a promoter-proximal motif), and four unknown motifs: LM2, LM7, LM23, and LM194. (We show below that LM2, LM7, and LM23 correspond to CTCF binding sites.) The median lengths of surrounding conserved sequences for these motifs are all less than five flanking bases on each side. For example, LM2 has only a median two flanking conserved bases on each side (Fig. 1*g*), whereas LM1 (Fig. 1*h*) has a median of 31 flanking conserved bases on each side.

LM1 Defines RFX Binding Sites.

The most highly enriched motif LM1 is similar to the X-box motif, which has been extensively studied in yeast and nematodes (16–18). In yeast, more than three dozen X-box sites have been identified, and these sites have been shown to be bound by the Crt1 protein, an effector of the DNA damage checkpoint pathway (19). In *Caenorhabditis elegans*, >700 X-box sites have been computationally predicted, and several dozen of these sites have been demonstrated to be recognized by the DAF-19 protein, which is known to regulate genes involved in the development of sensory cilia (16, 18).

Individual instances of the X-box motif in vertebrates have been reported, but no systematic survey of X-box motifs in the human genome has been conducted. Approximately three dozen such sites have been reported to be bound by RFX family proteins, which are

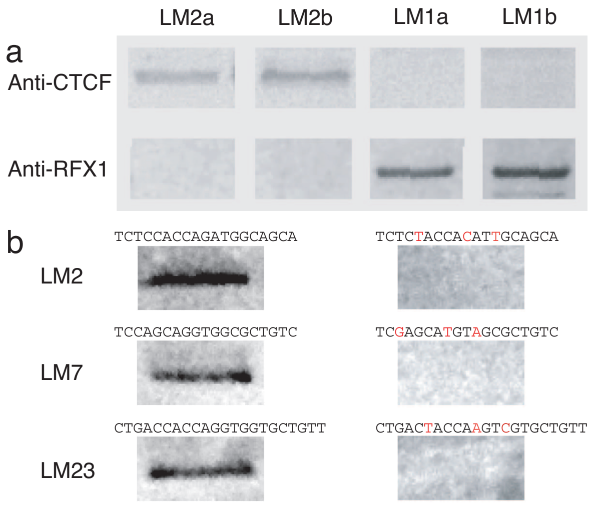


Fig. 2. Confirmation of CTCF and RFX1 binding by *in vitro* affinity capture. (*a*) CTCF was specifically captured by probes LM2a and LM2b constructed for the LM2 motif, whereas RFX1 was specifically captured by probes LM1a and LM1b constructed for the LM1 motif. (*b*) The binding of CTCF to LM2, LM7, and LM23 (*Left*), but not to their corresponding mutant motifs with three core bases altered (*Right*). See *Methods* for probes used in the experiments.

homologous to both Crt1 and DAF-19 and contain a highly conserved winged helix DNA binding domain. The biochemically characterized consensus sequence for RFX binding shows similarity to the LM1 motif (20), although it contains less information.

To test whether LM1 binds RFX proteins, we performed an affinity-capture experiment (see *SI Text*). A biotinylated double-stranded DNA probe containing multiple copies of the LM1 motif was incubated with HeLa cell nuclear extract and then captured with streptavidin. The bound protein was electrophoresed, blotted, and probed with an antibody against RFX1, a prototypical member of the RFX family, revealing that the protein indeed specifically binds LM1 (Fig. 2*a*).

LM2 Defines a Common Insulator Site Across the Human Genome. The most interesting case among the 233 discovered motifs is LM2. It has the largest number of conserved instances (7,549) in the genome, with the vast majority being located far from TSSs (Fig. 1*e*). The LM2 motif is 19 bases in length and does not match the reported consensus sequence of any known motif.

We obtained a hint regarding the possible function of the LM2 motif by using proteomic experiments in which HeLa cell nuclear extract was subjected to affinity capture with a biotinylated double-stranded DNA probe containing multiple copies of the LM2 motif, and the resulting material was analyzed by protease digestion and mass spectrometry. These affinity-capture experiments suggested that the CTCF protein binds the LM2 motif (unpublished data).

CTCF, a protein containing 11 zinc-finger domains, is a major factor implicated in vertebrate insulator activities (21–23). An insulator is a DNA sequence element that prevents a regulatory protein binding to the control region of one gene from influencing the transcription of neighboring genes. When placed between an enhancer and a promoter, an insulator can block the interaction between the two. Several dozen insulator sites have been characterized, and almost all have been shown to contain CTCF binding sites. In some cases, the CTCF site has been directly shown to be both necessary and sufficient for enhancer blocking activities in heterologous settings. The known CTCF sites show considerable sequence variation, and no clear consensus sequence has been derived (22). The well studied CTCF sites in the IGF2/H19 locus show similarity to the LM2 motif (24), although the similarity score is below our threshold used for detecting LM2 sites.

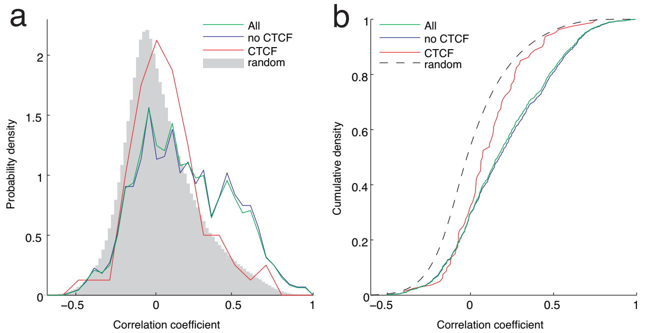


Fig. 3. Genes separated by predicted CTCF sites are less correlated in gene expression. Correlation coefficient between neighboring gene pairs is shown in terms of probability density (*a*) and cumulative distribution (*b*). Green line, correlation between all neighboring genes; red line, correlation between genes separated by at least one CTCF site; gray shading, correlation between randomly chosen gene pairs.

To test directly whether CTCF binds the LM2 motif we analyzed the material obtained by affinity capture with a biotinylated double-stranded DNA probe containing multiple copies of the LM2 motif by immunoblotting with an antibody against the human CTCF protein (see *SI Text*). The results confirmed that CTCF does indeed bind the LM2 motif (Fig. 2). By contrast, mutation of the three core positions with the highest information content (positions 5, 10, and 13 of LM2) (Fig. 1*b*) completely abolished the binding of the CTCF protein.

Given the sequence diversity among reported CTCF sites, we searched for additional motifs in our catalog that show substantial similarity to LM2. The motifs LM7 and LM23 are nearly identical in their first 14 positions, diverging only in the last four or five bases (SI Fig. 8). The two additional motifs also have an unusually large number of conserved instances (6302 for LM7 and 3758 for LM23). Affinity-capture experiments using probes containing copies of the LM7 and LM23 motifs demonstrated that both motifs bind CTCF, whereas mutation of the three core positions with the highest information content completely abolish binding (Fig. 2*b*). The three motifs, LM2, LM7, and LM23, will be referred to as a “supermotif,” LM2*.

Altogether the LM2* motif has 14,987 conserved instances in the human genome (which is 20-fold higher than for the corresponding control motifs). Strikingly, this comprises approximately one-fourth of the 60,019 sites for the complete catalog of 233 motifs. We propose that the vast majority of these sites are CTCF-binding sites and function as insulators.

Although the predicted CTCF sites tend to be located far from gene starts, they are not randomly distributed across the genome. Instead, their distribution closely follows the distribution of genes, with a correlation coefficient of 0.6 (SI Fig. 9). This is consistent with the notion that the sites are related to gene regulation, rather than, for example, chromosomal structure.

We sought to test whether the predicted CTCF sites actually serve as functional insulators. Although it is possible to perform insulator assays on individual instances in a heterologous context, we were interested to assess the function of many CTCF sites in their natural context. If the predicted CTCF sites actually function as insulators, we reasoned that the presence of a CTCF site between two genes might “decouple” their gene expression.

It is known that divergent gene pairs, transcribed in opposite directions with transcription start sites close to each other, tend to show correlated gene expression patterns (25, 26). We therefore assembled a data set of 963 divergent gene pairs with intergene distance <20 kb and with expression values measured across 75 human tissues (27). As expected, the divergent gene pairs are more closely correlated in gene expression than randomly chosen gene pairs (Fig. 3). When the cases are divided into gene pairs separated by a CTCF site (CTCF pairs, 80 cases) and those not separated by a CTCF site (non-CTCF pairs, 883 cases), the former show correlations that are essentially equivalent to the random background. Overall, 37% of non-CTCF pairs are strongly correlated (correlation coefficient $\rho > 0.3$). This proportion is 2-fold higher than the proportion of random genes pairs (12%) showing similarly strong correlation. By contrast, the proportion of CTCF pairs with similarly strong correlation is 16%, which is close to that seen for random gene pairs. This difference persists after correcting for small difference in the lengths of CTCF-containing and CTCF-non-containing intergenic regions (SI Fig. 10). This provides strong evidence that the majority of the predicted CTCF sites do indeed function as insulators.

Finally, we examined the frequency of the CTCF motif LM2* across various vertebrate genomes. The three motifs all occurred frequently in all eutherian mammals, opossum, chicken, and the pufferfish *Tetraodon*. The motif shows a similar total number of instances across all vertebrate species despite a 5-fold variation in genome size (SI Fig. 11). This is consistent with the LM2* motif being related to gene number (which is fairly constant across these species) rather than genome size.

Discussion

Our analysis provided an initial systematic catalog of regulatory motifs in the conserved regions of the entire human genome. The 233 discovered motifs are highly enriched in the CNE sequence, with all being at least 5-fold enriched relative to the rest of the genome. These motifs match 60,019 conserved instances in the human genome, with a typical motif having ≈100 conserved instances. Among the 233 discovered motifs, only 16 could be recognized as previously known regulatory elements, indicating that much more still remains to be learned about the function of CNE.

The most interesting unknown motif is LM2, which has ≈7,500 conserved instances in the genome, more abundant than any other discovered motif. We used affinity-capture assays to demonstrate that LM2, as well as two other closely related motifs, LM7 and LM23, are specifically bound by the CTCF protein, which is involved in insulator function. Together, the three motifs match nearly 15,000 conserved instances in the human genome, corresponding to approximately one-fourth of all matching instances for the entire set of discovered motifs. Although we cannot rule out that CTCF protein can also bind to other, highly dissimilar sites, our findings suggest that a few dominant CTCF motifs are extremely enriched throughout the human genome.

The results here are, of course, only a step toward comprehensive catalog of regulatory motifs across the human genome. In particular, our analysis used stringent threshold to identify only the most highly enriched motifs in the CNEs and therefore have omitted short motifs (e.g., 6–8 nt). Additionally, the current study focused primarily on motifs present in most mammals, and therefore many lineage-specific motifs, such as those unique to primates, still remain to be discovered. The power of motif discovery can be boosted not just by considering the enrichment of sequences in the human CNEs, but by exploiting their detailed conservation pattern across different species. With the availability of genome sequences from an increasing number of related mammals, it should be possible to create a complete dictionary of human motifs in the years ahead.

Methods

We started by enumerating a list of candidate k-mers with $12 \leq k \leq 22$ and counting the number (C) of matching instances of each k-mer present in the CNE data set (SI Fig. 5). A sequence was declared a match to a k-mer if the number of mismatches between the sequence and the k-mer was less than a threshold M (where $M = 0$ for $k \leq 13$; $M = 1$ for $k = 14$, 15, or 16; $M = 2$ for $k = 17$ or 18; and $M = 3$ for $k \geq 19$). For k-mers with $C \geq 30$ we identified all matching instances in the entire human genome and assessed their enrichment in the CNEs using two scores: (*i*) fold enrichment: $\mathrm{SNR} = C/\mu$; and (*ii*) Z-score $= (C - \mu)/sqrt(\mu)$, quantifying the significance of the enrichment. Here, μ is the expected number of matching instances within the CNE data set based on the observed frequency of matches in the overall genome. Finally, we collected all k-mers with $\mathrm{SNR} \geq 5$ and Z-score ≥ 10, resulting in a total of 69,810 k-mers significantly enriched in the CNEs. The probability of a k-mer having a Z-score $\geq$10 at random is $<2 \times 10^{-12}$, and, after Bonferroni correction, the expected number of such k-mers is $<10^{-4}$. These k-mers were further clustered and grouped into 233 distinct motifs according to the procedure described in ref. 5. For each motif we selected the k-mer with the highest Z-score to represent the motif.

Please see *SI Text* for additional methods.

We thank David Jaffe, Manuel Garber, and Sarah Calvo for insightful comments and suggestions. We gratefully acknowledge Xiaolan Zhang, Li Wang, Jacob Jaffe, and Steve Carr for assistance with affinity-capture experiments. This work was supported in part by grants from the National Human Genome Research Institute (to E.S.L.) and from the Broad Institute.

1. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, *et al.* (2006) *Nature* 444:499–502.
2. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, James Kent W, Haussler D (2006) *Nature* 441:87–90.
3. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, *et al.* (2005) *PLoS Biol* 3:e7.
4. Dermitzakis ET, Reymond A, Antonarakis SE (2005) *Nat Rev Genet* 6:151–157.
5. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) *Nature* 434:338–345.
6. Elemento O, Tavazoie S (2005) *Genome Biol* 6:R18.
7. Ettwiller L, Paten B, Souren M, Loosli F, Wittbrodt J, Birney E (2005) *Genome Biol* 6:R104.
8. Jones NC, Pevzner PA (2006) *Bioinformatics* 22:e236–242.
9. Ballas N, Mandel G (2005) *Curr Opin Neurobiol* 15:500–506.
10. Chong JA, Tapia-Ramirez J, Kim S, Toledo-Aral JJ, Zheng Y, Boutros MC, Altshuller YM, Frohman MA, Kraner SD, Mandel G (1995) *Cell* 80:949–957.
11. Schoenherr CJ, Anderson DJ (1995) *Science* 267:1360–1363.
12. Mortazavi A, Thompson EC, Garcia ST, Myers RM, Wold B (2006) *Genome Res* 16:1208–1221.
13. Bruce AW, Donaldson IJ, Wood IC, Yerbury SA, Sadowski MI, Chapman M, Gottgens B, Buckley NJ (2004) *Proc Natl Acad Sci USA* 101:10458–10463.
14. Williams AS, Marzluff WF (1995) *Nucleic Acids Res* 23:654–662.
15. Pandey NB, Marzluff WF (1987) *Mol Cell Biol* 7:4557–4559.
16. Blacque OE, Perens EA, Boroevich KA, Inglis PN, Li C, Warner A, Khattra J, Holt RA, Ou G, Mah AK, *et al.* (2005) *Curr Biol* 15:935–941.
17. Zaim J, Speina E, Kierzek AM (2005) *J Biol Chem* 280:28–37.
18. Efimenko E, Bubb K, Mak HY, Holzman T, Leroux MR, Ruvkun G, Thomas JH, Swoboda P (2005) *Development (Cambridge, UK)* 132:1923–1934.
19. Huang M, Zhou Z, Elledge SJ (1998) *Cell* 94:595–605.
20. Emery P, Strubin M, Hofmann K, Bucher P, Mach B, Reith W (1996) *Mol Cell Biol* 16:4486–4494.
21. Bell AC, West AG, Felsenfeld G (1999) *Cell* 98:387–396.
22. Ohlsson R, Renkawitz R, Lobanenkov V (2001) *Trends Genet* 17:520–527.
23. Gaszner M, Felsenfeld G (2006) *Nat Rev Genet* 7:703–713.
24. Bell AC, Felsenfeld G (2000) *Nature* 405:482–485.
25. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM (2004) *Genome Res* 14:62–66.
26. Li YY, Yu H, Guo ZM, Guo TQ, Tu K, Li YX (2006) *PLoS Comput Biol* 2:e74.
27. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al.* (2004) *Proc Natl Acad Sci USA* 101:6062–6067.
28. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, *et al.* (2005) *Genome Res* 15:1034–1050.