Computational identification and functional validation of regulatory motifs in cartilage-expressed genes

Sherri R. Davies, Li-Wei Chang, Debabrata Patra, et al.

Genome Res. 2007 17: 1438-1447 originally published online September 4, 2007
Access the most recent version at doi:10.1101/gr.6224007

Letter

Computational identification and functional validation of regulatory motifs in cartilage-expressed genes

Sherri R. Davies,[1,7] Li-Wei Chang,[2] Debabrata Patra,[1] Xiaoyun Xing,[1] Karen Posey,[3] Jacqueline Hecht,[3,4] Gary D. Stormo,[2,5] and Linda J. Sandell[1,6,8]

[1]*Department of Orthopaedic Surgery, Washington University School of Medicine, St. Louis, Missouri 63110, USA;* [2]*Department of Biomedical Engineering, Washington University, St. Louis, Missouri 63130, USA;* [3]*Department of Pediatrics, University of Texas Medical School at Houston, Houston, Texas 77030, USA;* [4]*Shriners Hospital for Children, Houston, Texas 77030, USA;* [5]*Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110, USA;* [6]*Department of Cell Biology and Physiology, Washington University School of Medicine, St. Louis, Missouri 63110, USA*

Chondrocyte gene regulation is important for the generation and maintenance of cartilage tissues. Several regulatory factors have been identified that play a role in chondrogenesis, including the positive transacting factors of the SOX family such as SOX9, SOX5, and SOX6, as well as negative transacting factors such as C/EBP and delta EFI. However, a complete understanding of the intricate regulatory network that governs the tissue-specific expression of cartilage genes is not yet available. We have taken a computational approach to identify *cis*-regulatory, transcription factor (TF) binding motifs in a set of cartilage characteristic genes to better define the transcriptional regulatory networks that regulate chondrogenesis. Our computational methods have identified several TFs, whose binding profiles are available in the TRANSFAC database, as important to chondrogenesis. In addition, a cartilage-specific SOX-binding profile was constructed and used to identify both known, and novel, functional paired SOX-binding motifs in chondrocyte genes. Using DNA pattern-recognition algorithms, we have also identified *cis*-regulatory elements for unknown TFs. We have validated our computational predictions through mutational analyses in cell transfection experiments. One novel regulatory motif, NI, found at high frequency in the *COL2A1* promoter, was found to bind to chondrocyte nuclear proteins. Mutational analyses suggest that this motif binds a repressive factor that regulates basal levels of the *COL2A1* promoter.

[Supplemental material is available online at www.genome.org.]

Gene expression and its regulation are important for the coordination of various activities of a cell. This complex regulatory circuit involves a multitude of transcription factors (TFs) and their corresponding *cis*-acting regulatory elements whose interaction in a variety of permutational and combinatorial events enable the function and maintenance of tissues. An understanding of the transcriptional regulatory network (Covert et al. 2004) can be attempted using a computational approach in which TF-binding sites can be modeled, searched, and identified in the non-coding sequence through the use of position weight matrices (PWMs) and DNA pattern recognition programs (Stormo 2000). The conservation of functional *cis*-acting elements among the non-coding sequences of orthologous genes denoted by "phylogenetic footprinting" has refined computational approaches to allow for a credible identification of *cis*-acting elements (Tagle et al. 1988; Wasserman et al. 2000; Blanchette and Tompa 2002).

The successful use of DNA pattern recognition programs in inferring regulatory circuits has been evident in the detection of novel regulatory elements involved in heat-shock response (GuhaThakurta et al. 2002a) and in foregut development (Ao et al. 2004) in the nematode *Caenorhabditis elegans*. Position weight matrix models and phylogenetic footprinting have also been used to define transcriptional regulatory mechanisms in mammalian genomes (Qiu et al. 2003; Hu et al. 2004; Nelander et al. 2005; Chang et al. 2006). Such an approach to chondrocyte gene regulation is of interest to add to the understanding of the generation and maintenance of cartilage and its influence on endochondral bone formation derived from the cartilage anlagen (Karsenty and Wagner 2002).

Many genes such as collagen type II (*COL2A1*), collagen type IX (*COL9A1*), collagen type XI (*COL11A2*), and melanoma inhibitory activity (*MIA*) that are normally expressed predominantly and almost exclusively in cartilage are known to share common regulatory factors (Okazaki et al. 2002; Okazaki and Sandell 2004; Furumatsu et al. 2005; Imamura et al. 2005). This suggests that there are common regulatory modules for cartilage-specific genes that determine their expression pattern. The annotation of such sequences would help define the temporal and tissue-specific expression pattern of these genes in chondrocytes. Therefore, in this study, we have taken a computational approach in order to identify TF-binding motifs important to cartilage maintenance and development. Although several reports have used computational methods to detect mammalian regulatory sequence or regulatory modules, only a few of them have experimentally validated their predictions (Blanchette et al. 2006; Wang et al. 2006). In this report, we have successfully

[7]Present address: Division of Oncology, Washington University School of Medicine, St. Louis, MO 63110, USA.
[8]Corresponding author.
E-mail sandelll@wudosis.wustl.edu; fax (314) 454-5900.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6224007.

17:1438–1447 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

Table 1. Statistical analyses of significant transcription factor binding sites in the cartilage gene set

A. Top-ranking transcription factors based on log(P_c/P_a) values

Motif ID	Transcription factor	Class	Log(P_c/P_a)
M00131	HNF-3beta	3.3.2	4.209
M00289	HFH-3	3.3.0	3.353
M00731	Osf2	4.11.1	3.300
M00722	CBF	0.3.2	3.294
M00053	c-Rel	4.1.1	3.264
M00649	MAZ	2.3.0	3.204
M00720	CAC-binding	2.3.0	3.203
M00769	AML	4.11.1	3.161
M00188	AP-1	1.1.1	3.076
M00130	FOXD3	3.3.0	3.054
M00192	GR	2.1.1	3.044
M00724	HNF-3alpha	3.3.0	3.031

B. Other known transcription factors that are known to regulate or potentially regulate cartilage genes

Motif ID	Transcription factor	Class	Log(P_c/P_a)
M00415	AREB6	3.1.4	2.98
M00807	Egr	2.3.2	2.457
M00189	AP-2	1.6.1	2.266
M00074	c-Ets-1	3.5.2	2.034
M00033	p300	N/A	1.856
M00770	C/EBP	1.1.3	1.635
M00042	SOX-5	4.7.1	1.351
M00410	SOX-9	4.7.1	0.842

C. Novel binding motifs over-represented in the cartilage gene set

Motif ID	Transcription factor	Class	Log(P_c/P_a)
N1	Unknown	N/A	2.302
N20	Unknown	N/A	2.292
N23	Unknown	N/A	1.471
N24	Unknown	N/A	2.597
N31	Unknown	N/A	3.382

Putative binding site of known transcription factors in cartilage gene promoters were identified by PATSER. The binding probabilities were calculated based on the scores of the sites. The log(P_c/P_a) values are logarithms of binding probabilities normalized using all the promoters in the genome. The TRANSFAC classification of binding motif is indicated.

identified and validated several SOX-binding motifs that regulate chondrocyte-specific gene expression. We have also successfully identified several novel TF-binding motifs in our cartilage gene set. One of these motifs, N1, was found to bind nuclear proteins, and mutational analyses suggest that it participates in regulating at least three cartilage genes in our set. These data point out the usefulness of computational approaches in annotating functional regulatory motifs in mammalian genomes.

Results

Analyses of previously characterized TF-binding motifs in cartilage characteristic genes

Our analyses included steps to identify known TF-binding motifs, novel binding motifs, and paired SOX-binding sites within cartilage enhancer sequences by examining the conserved sequence in 18 orthologous pairs of human and mouse genes. The results of the analysis of the characterized TF binding motifs from TRANSFAC are shown in Table 1A. Transcription factor-binding motifs were ranked by the log ratio of the probability scores, taking into account multiple predicted binding sites within a promoter, calculated using cartilage genes and background genes, respectively (see Methods). Based on this model, a higher log ratio value indicates a higher probability that the TF will bind to the promoter of these cartilage genes. The log ratios of the probability scores for TFs with documented roles in chondrogenesis are given in Table 1B.

Identification of paired SOX-binding motifs in cartilage characteristic genes

One TF we expected to detect with a high log ratio in the analyses above is SOX9, a master regulator of chondrogenesis (Wagner et al. 1994; Wheatley et al. 1996; Lefebvre et al. 1997; Ng et al. 1997; Bi et al. 1999; Wegner 1999; Akiyama et al. 2002). However, the log ratio values for both the TRANSFAC SOX9 and SOX5 motifs in cartilage genes were relatively low compared to other motifs (0.842 for SOX9 and 1.351 for SOX5) given the known functional roles these factors play in vivo. SOX transcription factors bind to *cis*-regulatory sequences through a conserved HMG domain. The SOX subfamily of HMG proteins is remarkable as they appear to play distinct and essential roles in many different developmental processes, yet the DNA-binding sites they recognize are highly degenerate. Their *trans*-activating functions and specificities appear to be highly dependent on both the orientation of, and the spacing between, their binding sites and the binding sites of other cofactors (Wegner 1999; Peirano and Wegner 2000; Bridgewater et al. 2003). This suggests that the SOX motif present in TRANSFAC may not model functional SOX-binding sites present in chondrocyte genes precisely and accurately. Therefore, to better model the binding sites involved in cartilage gene regulation, we constructed a cartilage-specific SOX-binding profile from enhancer elements known to specify cartilage gene expression in vivo (Fig. 1A,B). We collected a set of 31 previously documented SOX-binding sites (Supplemental Table 1) from three cartilage-specific genes. The consensus sequence of this profile is CTTT GWW, which is slightly different from the TRANSFAC motifs for SOX5 (ATTGTT) or SOX9 (CYATTGTT). These three SOX-binding

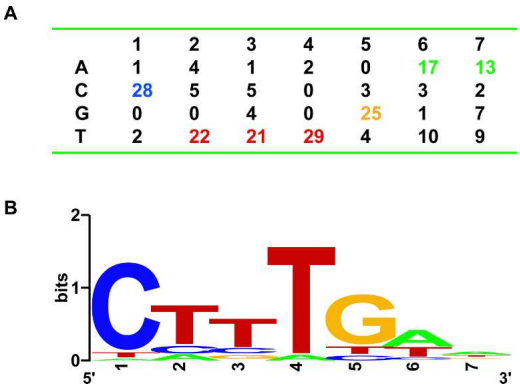

Figure 1. The position weight matrix and the sequence logo of the cartilage-specific SOX-binding profile constructed from experimentally validated SOX sites. Thirty-one experimentally validated SOX-binding sites from three genes were collected from the literature and used to model a SOX-binding profile for searching the cartilage gene set. (*A*) An illustration of the position weight matrix developed using these sites. (*B*) The sequence logo describing the position weight matrix (Schneider and Stephens 1990). For sequences and species information, see Supplemental Table 1.

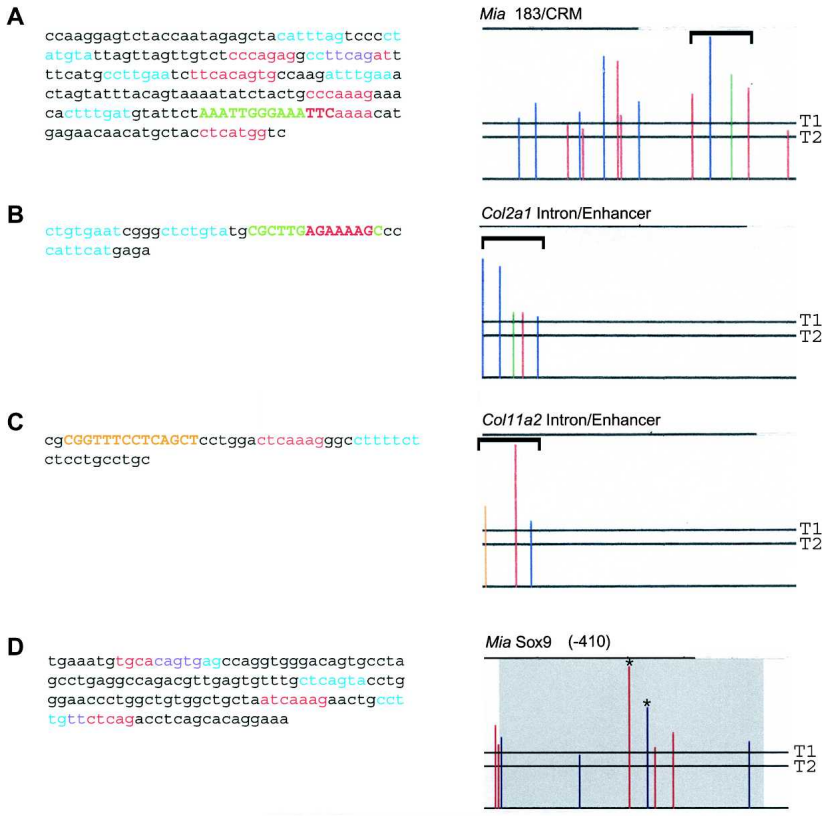

Figure 2. Computational prediction of cartilage regulatory elements. Weight matrices for the SOX- and C/EBP beta binding sites were used to search functional cartilage enhancers. (*A*) The 183-bp *Mia* (mouse) element located at −2251 to −2058. (*B*) *Col2A1* 48-bp intron enhancer (mouse). (*C*) *Col11A2* intron enhancer (mouse). (*D*) *Mia* proximal promoter (mouse). Colors represent motifs and their orientations as follows. (Blue) SOX sites on the forward strand; (red) SOX sites on the reverse strand; (green) C/EBP beta sites on the forward strand; (yellow) C/EBP beta sites on the reverse strand; (purple) overlapping SOX sites in opposite orientation. Bold letters in any color represent C/EBP beta motifs. Chondrocyte regulatory motifs containing SOX pairs and C/EBP beta are indicated by brackets. Gray shaded area in *D* indicates that the sequence region is conserved at >60% between human and mouse. (*) Sites mutated in transient transfections (−410m) and (−395m). T1 and T2 are statistical thresholds for C/EBP beta and SOX sites, respectively.

profiles are short and contain degenerate positions. Therefore, to reduce the false-positive discovery rate, we used evolutionary conservation to search for SOX sites conserved in human and mouse cartilage genes. This approach significantly decreased the number of predicted SOX sites in cartilage-expressed genes (data not shown).

To test the sensitivity of our SOX model, we used this motif to predict experimentally identified SOX-binding sites in our training set of cartilage regulatory regions of the *Mia*, *Col2a1*, and *Col11a2* genes (Fig. 2A–C). Although using the SOX profile to identify the contributing SOX-binding sites seems circular, the purpose of this test is to identify (1) any SOX site in the training set that is significantly different from other SOX sites and (2) additional unknown SOX sites in the known cartilage regulatory regions. All of the sites used to train the motif were also predicted from our search algorithm with the exception of one in the 183-bp *Mia* gene promoter element, indicating that our search criteria were functioning appropriately. When the orientation and spacing of the predicted sites were examined, we found at least one tandem pair of sites with opposite orientations, with a short spacer sequence in between (3–8 bp), for each of the regulatory elements used to develop our model motifs (Fig. 2A–C).

Multiple sites were predicted in the 183-bp element, located at −2251 to −2058 of the *Mia* promoter, an essential part of the cartilage regulatory module (Okazaki et al. 2006). Indeed, it has been shown in previous studies that the E class of SOX factors, which includes SOX9, can exhibit homodimeric, cooperative DNA binding that is dependent on closely spaced and oppositely oriented binding sites (Peirano and Wegner 2000; Sock et al. 2003). Next, we searched the entire *Mia* gene promoter sequence to see if our algorithm could predict a known SOX9-binding site at −410 (Xie et al. 1999). In addition to predicting our previously identified SOX9 binding site, a novel, closely spaced, and oppositely orientated site was identified immediately downstream at −395 (Fig. 2D). We mutated both the −410 and −395 sites individually and analyzed the effects using luciferase reporter constructs in transient transfection experiments in RCS cells (Fig. 3). As previously reported, the full-length −2251-bp construct that is required for cartilage-specific expression exhibited strong activity (Xie et al. 1999). In contrast, the negative control construct truncated to −401 bp (and missing the −410 SOX9 site) had virtually no activity. Mutation of two core nucleotides in the newly identified motif at −395 bp similarly ablated activity of the full-length promoter construct (−2251 bp) despite the presence of the −410 SOX9-binding site that has previously been characterized (Fig. 3) (Xie et al. 1999). Thus, both the −410 and −395 sites appear to be equally essential to the full-length promoter activity.

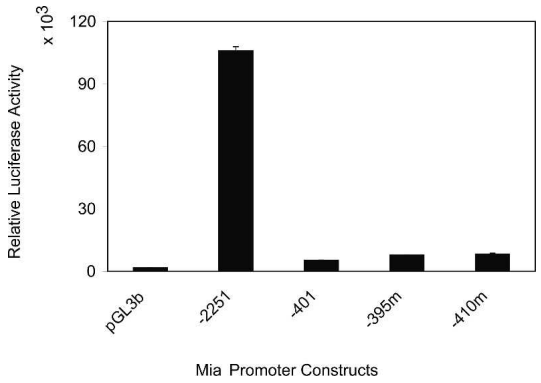

Figure 3. Mutational analyses of computationally predicted paired SOX-binding motifs in the *Mia* promoter. RCS cells were transiently transfected (in triplicate experiments) with either the full-length *Mia* (−2251 bp mouse) promoter (required for chondrogenic expression), truncated promoter (−401), or with mutations in the computationally predicted SOX-binding motifs (CA to GG) (−410m and −395m).

Since experimental data suggested that pairing of SOX motifs is an important part of the chondrocyte regulatory module, we refined our search criteria to include not just the conservation between the mouse and human sequences, but also the close positioning of the sites that are 3–8 bp apart. A summary of all the conserved sites predicted in the promoters of all the searched cartilage genes is shown in Supplemental Table 2. As the table indicates, the conservation between human and mouse orthologous sequences significantly reduced the number of motifs that would be predicted to be functional. There were only three genes in our cartilage set (*COL9A2*, *COMP*, *FGF18*) that did not have any conserved pair of SOX sites. To test the specificity of our search criteria, we also collected the promoters of 13 previously reported liver-specific genes (Krivan and Wasserman 2001) and performed the exact same search. Only six of the 13 liver-specific genes have conserved paired SOX sites in their promoters (Supplemental Table 3).

In addition to SOX factors, our previous results have implicated the repressor C/EBP beta as a TF that may work in these regulatory sequences to restrict cartilage gene expression (Okazaki et al. 2006). Therefore, we also used the C/EBP beta binding profile in TRANSFAC (M00109) to search the cartilage gene sequences to see if we could predict similar functional cartilage regulatory modules (Fig. 2). When the positions of both the conserved SOX sites and C/EBP beta motifs were considered, we found that seven of the 18 genes in our list possessed at least one of these putative regulatory modules containing both paired SOX sites and C/EBP beta sites in close proximity. When liver-specific genes were searched, only two genes had a similar regulatory module in their promoters (Supplemental Table 3).

Since the SOX-binding module seemed a predominant feature in our gene set, we then decided to test these predicted sites for functional activity in another well-characterized cartilage gene promoter, *COL9A1*, that is known to be activated by SOX9 in chondrocytes (Zhang et al. 2003) (Fig. 4). We found 10 conserved motifs in the 10-kb upstream sequence for *COL9A1*. Six of these motifs were in the proximal region that could readily be tested using transfection reporter assays (Fig. 4A). Each site was individually mutated by 1–2 bp (CA to GG, depending on the endogenous sequence) in *COL9A1* p846Luc constructs (Zhang et al. 2003) and tested for promoter activity in the RCS cells (Fig. 4B, M1–M6). Similar to the *Mia* gene promoter, all mutations of the core CA to GG resulted in nearly complete loss of promoter activity in the RCS cells. Site 4, represented by mutation M4, has previously been characterized and demonstrated to bind SOX9 in electromobility gel shift assays (EMSA), and also served as a positive control in these experiments (Zhang et al. 2003). In contrast to the RCS cells, the mutations did not have a significant effect on the promoter activity in the NIH3T3 cells, a fibroblast cell line that does not express *SOX9* mRNA (Zhang et al. 2003; Davies et al. 2004) (Fig. 4C). This suggested that the newly identified SOX-binding motifs were tissue-specific and each site is as functionally important to promoter activity as site 4. A similar mutation in a nonconserved SOX site (M7) did not have the same effect in RCS cells (Fig. 4B).

Identification of a novel motif NI in cartilage-expressed genes

Although the TRANSFAC database curates ~500 vertebrate TF-binding profiles, there are estimated to be ~2000 TF proteins in

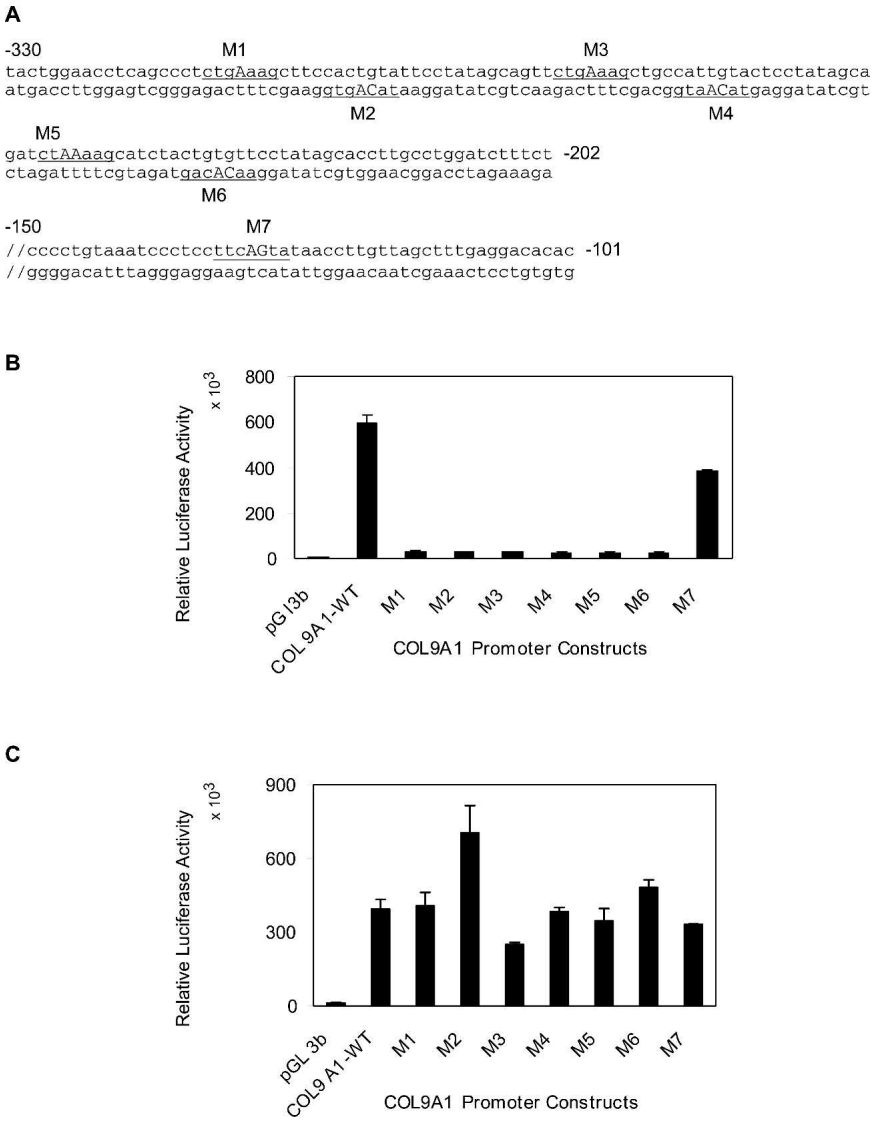

Figure 4. Mutational analyses of predicted SOX-binding motifs in the *COL9A1* promoter. (*A*) The proximal promoter region of the *COL9A1* human gene showing the conserved and computationally predicted SOX-binding motifs underlined (M1–M7). Mutated nucleotides are indicated by capital letters. RCS cells (*B*) or NIH3T3 cells (*C*) were transiently transfected (in triplicates) with the *COL9A1*-WT promoter construct (p846) or with the p846 construct with mutations (CA to GG) in the predicted SOX-binding motifs M1, M2, M3, M4, M5, M6, and M7.

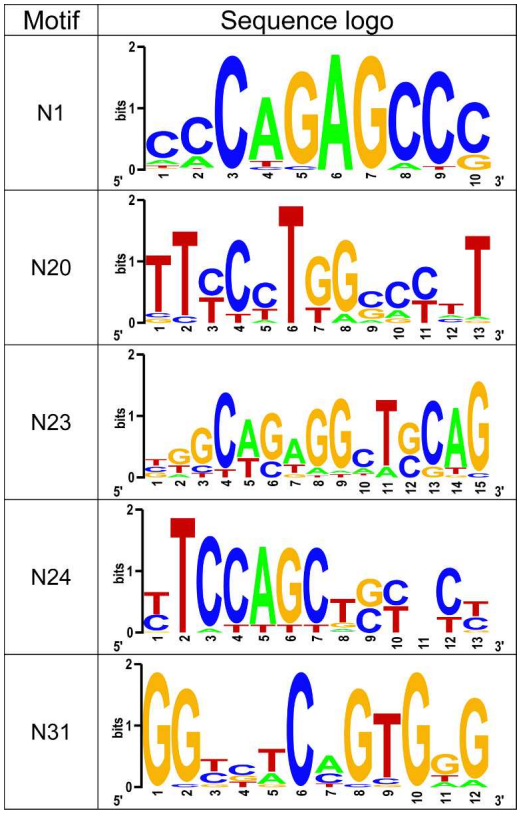

Figure 5. Sequence logos of computationally predicted novel motifs over-represented in cartilage genes. These motifs were identified using the program CONSENSUS, which searches for enriched motifs in the cartilage-specific promoters.

the human genome. Therefore, the binding profiles of many TFs are still not available, and the computational analysis described above using the TRANSFAC motifs may have missed other chondrocyte *cis*-regulatory elements. To circumvent this problem, we applied the DNA pattern recognition algorithm CONSENSUS (Stormo and Hartzell 1989; Hertz and Stormo 1999) to identify potential regulatory motifs in chondrocyte genes that have not been identified in TRANSFAC. The input data set for CONSENSUS is the conserved sequences of the 18 cartilage genes. We analyzed the human and mouse sequence sets separately and used the ALLR statistics (Wang and Stormo 2003) to find nonredundant candidate motifs that were conserved between human and mouse and over-represented in cartilage promoters. By this procedure, 87 conserved motifs were identified. Comparison and consolidation between similar motifs generated 23 nonredundant conserved motifs. The log ratio of the probability scores was calculated for each motif (Table 1C), and the sequence logo for the five top ranking motifs that are not present in the TRANSFAC database is shown in Figure 5. The distribution of all five novel motifs in cartilage-characteristic genes and their genomic location is shown in Supplemental Table 4.

The location of the N1 motif in the promoter regions of both the *COL2A1* and *COL11A2* genes made this motif a good candidate for experimental validation of its function. The consensus sequence determined for this motif is CCAGAGCCC. Significantly, evolutionarily conserved N1 motifs were detected in the proximal promoter region of 11 of the 18 cartilage genes analyzed in both human and mouse (Table 2). In some genes, multiple conserved N1 motifs were detected. For example, in the *COL2A1* promoter, four conserved N1 motifs were identified within −1 kb (N-125/126, N-135, N-147, and N-991), and one conserved N1 motif was identified at the +16 (N+16) position relative to the transcriptional start site.

To assess the function of the N1 motif and to determine its relevance in the regulation of chondrogenesis, oligonucleotides (15 bp in length) were generated from the promoter sequences of the cartilage genes in which this motif was recognized (see Table 2) and used as probes in electromobility gel shift assays using nuclear extracts from the RCS cells to determine protein-binding capacity (Fig. 6). Significantly, a major DNA–protein complex with similar mobility was generated for all the 16 oligonucleotides tested (Fig. 6, O1–O16, arrowhead), suggesting that these protein–DNA complexes probably have an identical TF binding to the N1 motif. In all of these cases, the binding could be competed by the addition of nonradiolabeled oligonucleotides to the reaction, demonstrating the specific nature of this binding. For the oligonucleotides O6, O13, O14, and O16, an additional, faster migrating complex (double arrowhead) was also observed, suggesting additional TFs may bind to these sequences as well. Four conserved N1 sites were identified in the *COL2A1* promoter within −1 kb from the transcription start site (O6, O7, O8, O9). The N1 site at −125/126 was found to be present on both strands of the *COL2A1* promoter sequence at the indicated positions. Interestingly, this motif, **CCCC**GGAGCCC, partially overlaps a previously identified EGR1 site (with overlap shown in bold/underlined) that mediates repression of the *COL2A1* gene by interleukin 1 beta (IL1B) (Tan et al. 2003).

Table 2. Distribution of the novel N1 motif in the cartilage gene set

Motif	Sequence	Genes	Position	Orientation
O1	cgg**TCCGGAGCCA**gg	*ACAN*	−504	+
O2	cac**CCCAGGGACC**ga	*ACAN*	−438	−
O3	tgg**GCCAGGGCCA**gt	*ACAN (COL11A1)*	−1602	+
O4	ccg**CCCACAGCCA**cc	*COL11A2*	−84	−
O5	ggg**CCCAGAGCCC**cc	*COL11A2 (FGF18, SOX9)*	−306	−
O6	gac**CCGGCAGCCC**ag	*COL2A1*	16	−
O7	gcc**CCCGGAGCCC**gc	*COL2A1*	−126	−
O8	ct**GCCAGTGCCC**gca	*COL2A1*	−147	−
O9	ctg**GCCAGGGCCG**ca	*COL2A1*	−991	+
O10	tca**CCCAAAGCCA**tg	*COL9A1*	−9189	−
O11	cgc**CCCGCAGCCC**ct	*FGF18*	293	+
O12	ctg**ACCAGAGTCC**tg	*FGF18*	−862	+
O13	ggg**CACAGAGCTG**cc	*HAPLN1*	−48	+
O14	gtg**CCCTGAGCCC**tg	*MATN1 (CILP)*	−1856	+
O15	tt**CCCAGAGAGC**cca	*MATN1*	−4994	−
O16	ggg**GCCCGAGCCC**gg	*IHH*	−769	+

The novel N1 motif was highly represented in the cartilage gene set. O1 to O16 represent 15-mer oligonucleotides containing the N1 motif that were generated for electromobility gel shift assays. The predicted motifs are indicated in bold with the corresponding gene and location on their respective promoters indicated (given in base pairs from the annotated transcription start site of the first gene listed). The direction of orientation is listed in the last column with "+" corresponding to the sense strand and "−"corresponding to the antisense strand of the DNA.

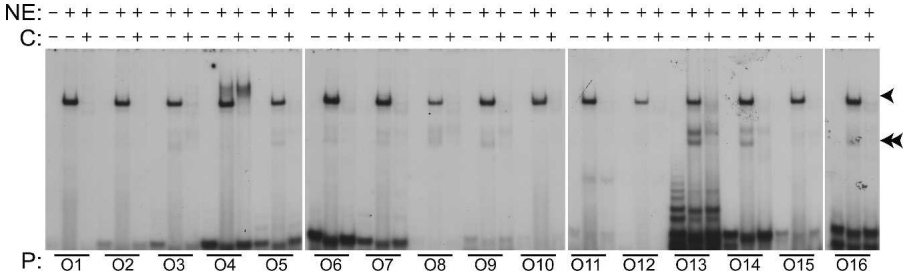

Figure 6. Nuclear protein binding to predicted novel (N1) motifs. Electromobility gel shift assays were performed to assess the ability of the motifs to bind proteins. Oligonucleotides (O1–O16) representing motifs (mouse) identified within the cartilage gene set were radiolabeled with [^{32}P]ATP, incubated with nuclear extracts (NE) from rat chondrosarcoma cells, and the protein-binding moieties were separated by SDS-PAGE (4%). In some reactions, additional unlabeled oligonucleotide (C) was also added at 100× to evaluate specificity binding to the motifs. The presence or absence of nuclear extract (NE) and unlabeled oligonucleotides (C) is indicated in the rows corresponding to the labels by + and −, respectively. The radiolabeled oligonucleotides used in each reaction are indicated by lines above the numbered oligonucleotide from Table 2. Predominant protein-binding complexes are indicated by single and double arrowheads.

To test the functional significance of the N1 motif in the *COL2A1* gene and validate the computational search, a mutational analysis of the N1 motif was performed. To avoid complications in the analysis caused by the binding of other factors to the site at −125/126, the more distal N1 motif at −147 in a *COL2A1* promoter-luciferase reporter construct was mutated to test its function in the RCS cell line in which strong *COL2A1* expression is normally observed. This mutational analysis involved deletion of the four most conserved bases from the N1 motif (CCA**GTGC**CC to CCA----CC) at the −147 position. Deletion of these four core bases increased the basal *COL2A1* promoter activity in our reporter construct as measured by relative luciferase activity (Fig. 7A), suggesting that this is probably a motif that binds to a repressor resulting in negative regulation. As IL1B is known to repress chondrogenesis, the response of the mutated *COL2A1* promoter was also tested by measuring its response to IL1B (Fig. 7A). As expected, the wild-type *COL2A1* promoter demonstrated a reduction in its basal activity (by 62%) in the presence of IL1B. The mutant promoter activity was also repressed (by 54%–63%), suggesting that the regulation to inflammatory mediators was still active. Interestingly, the repression of the mutant promoter in the presence of IL1B results in its activity being equal to the activity of the wild-type promoter in the absence of IL1B again, further suggesting that the −147 N1 motif probably binds to a negatively regulating TF.

The functional significance of the N1 motifs found in the promoter of an additional cartilage characteristic gene, *COMP*, was also analyzed. When oligonucleotides representing an N1 motif in the *COMP* promoter were tested by EMSA as above, a high-molecular-weight DNA–protein complex similar to that of the other tested oligonucleotides was observed. This DNA–protein complex was competed away with non-radiolabeled oligonucleotides representing the wild-type N1 motif, but not by oligonucleotides with mutation(s) of the core nucleotides within the N1 motif (data not shown; for details, see Supplemental Methods). Furthermore, to test the function of this motif in vitro, analogous mutations in the human *COMP* gene promoter-luciferase reporter construct were made, mutating the N1 sequence motif from CACCACAGCCC to CACTGTGGCCC to form N1Mut (mutational changes are underlined) (Fig. 7B). The RCS cell line transfected with this mutant construct also showed an increase in activity compared to the wild-type promoter, again demonstrating a negative transacting *cis*-regulatory function for this motif.

Discussion

In this study, we have taken a computational approach to annotate and validate important *cis*-regulatory motifs in cartilage-specific genes. The use of standard molecular biology techniques such as systematic sequence deletions and other mutagenesis approaches to define cartilage-specific regulatory regions would have been time-consuming given the size of the promoter region and the number of genes that we have used in this study. To circumvent this problem, we used computational methods that utilized position weight matrices and phylogenetic footprinting to increase the sensitivity of our predictions. Using this approach, we screened our cartilage-specific gene set for previously characterized TF-binding motifs from the TRANSFAC database and novel TF-binding sites with customized binding motifs. Statistical

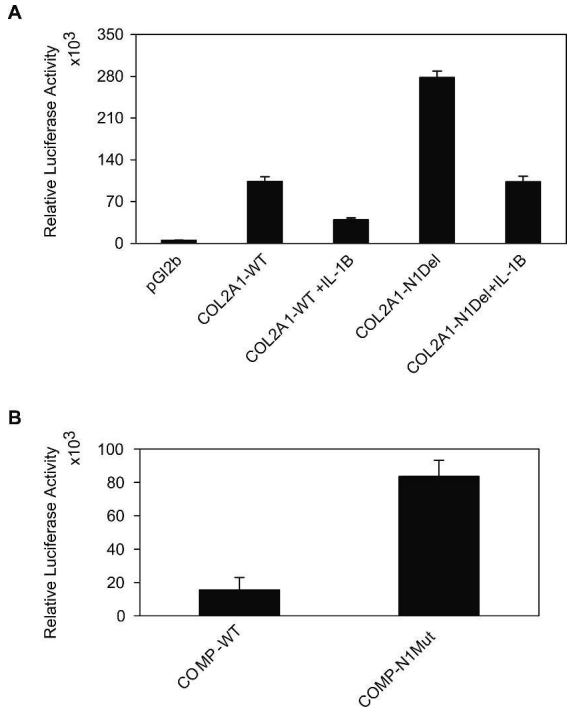

Figure 7. Mutation of the N1 motif increases basal promoter activity in *COL2A1* and *COMP* promoters. (*A*) RCS cells were transiently transfected (in triplicate experiments) with either wild-type *COL2A1* human promoter (*COL2A1*-WT) or the mutated (*COL2A1*-N1Del) promoter-luciferase reporter constructs in the presence (+) or absence (−) of IL1B, to evaluate the *cis*-regulatory function of the predicted N1 motif. The core sequence (underlined) in the N1 motif (CCAGTGCCC) was deleted to CCA----CC. (*B*) RCS cells were transfected with wild-type (*COMP*-WT) or mutant (*COMP*-N1Mut) *COMP* luciferase-reporter promoter constructs (human). The N1 motif in *COMP* promoter was converted from CACCACAGCCC to CACTGTGGCCC to form N1Mut (the mutational changes are underlined). Luciferase activity was normalized as described in Methods.

analysis of motif enrichment was performed to predict those that are most likely to regulate the chondrocyte genes.

Several TRANSFAC motifs were highly represented in our gene set for factors that play critical roles in patterning required for skeletal development (Nissen et al. 2003), cartilage differentiation (Inada et al. 1999; Miller et al. 2002), or cartilage gene regulation (Takagi et al. 1998; Xie et al. 1998). AREB6 (encoded by the gene *ZEB1* and also called delta EF1), C/EBP, and possibly other transcription factor motifs highly represented in our data set are negative regulators of cartilage-specific gene expression (Table 1B). In addition to down-regulation of the gene after the tissue has been established, negative regulation may be particularly relevant to tissue-specific genes in order to repress them in all other tissues. Indeed, we have demonstrated that C/EBP is responsible for suppression of *Mia* and *COL2A1* in C2C12 muscle cells in vitro, and potentially many other tissues in vivo (Okazaki et al. 2006). This analysis, however, did not identify any high-ranking SOX motifs despite the established importance of the SOX5, SOX6, and SOX9 proteins in chondrogenesis (Lefebvre et al. 2001). We used evolutionary conservation and the constraint of the orientation and the distance between paired SOX sites to refine our search further. This approach identified several highly conserved SOX sites in the 15 cartilage-expressed genes (Supplemental Table 2). The potential function of paired SOX-binding sites is supported by previous studies that have shown that cooperative binding is important for SOX9 activity in cartilage. For example, an inability of SOX9 to dimerize results in campomelic dysplasia even though its other functions are unaltered (Sock et al. 2003).

Using this information, we were able to refine our computational search and predict novel SOX-binding motifs for experimental validation more reliably. We found many high-scoring SOX-binding sites in our cartilage genes using the customized SOX-binding profile, but only some of them were present in a tandem pair with opposite orientations. We hypothesized that this is the preferred architecture of the chondrocyte regulatory module. Our analysis identified five novel SOX-binding motifs in the *COL9A1* promoter in addition to one that is already published. Mutations in any of these SOX-binding motifs significantly decreased promoter activity in luciferase reporter assays, validating our computational approach. We also identified a novel SOX-binding motif in the cartilage-specific *Mia* gene and similarly validated its function by mutational analyses. Although we did not test it for functional activity, our model also predicted an additional SOX site (+130 bp), oppositely orientated and adjacent to, a documented site (+120 bp) in the *HAPLN1* gene (also called cartilage linking protein 1 gene) (Kou and Ikegawa 2004) that again supports the hypothesis that this is the preferred architecture for the chondrocyte gene regulatory module.

Another challenge to defining biologically functional SOX-binding sites through computational approaches is that in vivo, some SOX sites function only in the context of other *cis*-acting motifs and TFs (Kamachi et al. 2000). This is evident in chondrogenesis, where SOX5 and SOX6 have been shown to synergistically activate the *COL2A1* enhancer in collaboration with SOX9 (Lefebvre and de Crombrugghe 1998). This implies that although a predicted SOX-binding motif could be a good match to the binding profile, it may not be biologically functional. Our previous studies suggest that a C/EBP beta binding site may also be a part of the chondrocyte regulatory module (Okazaki et al. 2006). C/EBP beta binding sites were found in close proximity to the paired SOX binding sites in the enhancers of the *Mia*, *Col2A1*, and *Col11A2* genes (Fig. 2). When C/EBP beta sites were also considered across the entire gene set, we found SOX-C/EBP beta modules in the promoters of seven cartilage-specific genes (Supplemental Table 2). In contrast, we found only two SOX-C/EBP beta binding modules in 13 liver-expressed genes (Krivan and Wasserman 2001) using these same search criteria (Supplemental Table 3). This suggests that our approach was reasonably specific. Both of the two binding profiles used in this study, SOX and C/EBP beta, are short and contain degeneracy in several positions. Therefore, without experimental validation, we cannot say for certain that some of the identified paired SOX-binding sites or C/EBP beta sites might not be false positives. Such a comprehensive understanding of the chondrocyte module may help improve our computational predictions. For example, the performance of our prediction may be improved by including binding profiles of additional cooperative TFs in the chondrocyte regulatory modules. Although we have used this approach on cartilage-specific genes, this general model could equally be applied to other mammalian systems by adding in additional motifs with known regulatory function (Fig. 8).

One drawback of performing computational promoter analysis based on the TRANSFAC database is that currently we do not have a complete collection of the binding profiles of all transcription factors. Although recent studies have proposed methods to computationally identify mammalian TF-binding profiles using comparative genomics approaches (Tan et al. 2005; Xie et al. 2005), the latest version of TRANSFAC has only the weight matrix models for about a quarter of the TFs in the human genome. Therefore, to identify enriched motifs in cartilage genes that have not been found in TRANSFAC, we also applied DNA pattern discovery algorithms on our cartilage gene set. Using this computational approach, we were able to identify and validate the function of one novel binding motif, N1, that appeared to participate in regulating the basal promoter activity of the *COL2A* gene. In addition to *COL2A1*, the N1 motif was also identified and validated in the promoter regions of other cartilage-

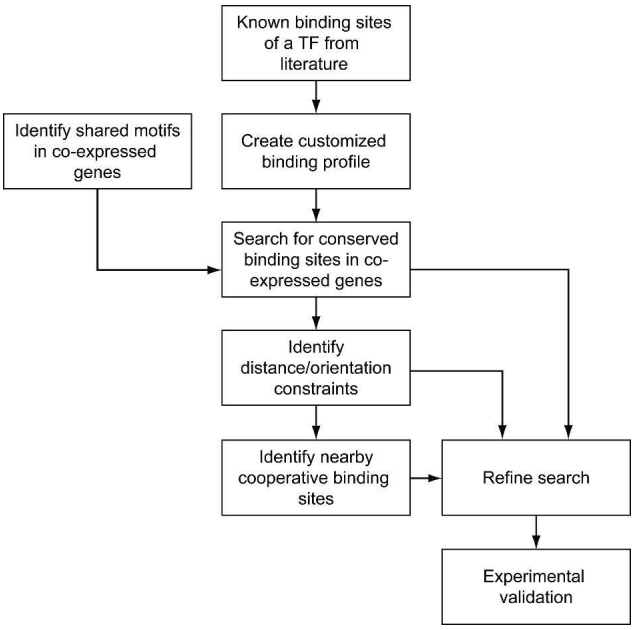

Figure 8. Proposed workflow of mammalian transcriptional regulatory sequence identification (see text for discussion).

expressed genes such as *COMP* and *COL9A1* by mutational analyses (data not shown), supporting its role as a *cis*-regulatory motif. We do not know the identity of the factor that binds to N1, as it does not correspond to any known TF-binding profile in the TRANSFAC database or any *cis*-acting regulatory elements known to drive cartilage-specific gene expression. Several N1 motifs were close to or overlapping with previously identified binding sites for SP1, EGR1 (Tan et al. 2003), or Maf (Huang et al. 2002). However, we were not able to verify binding of any of these proteins to the N1 motif using specific antibodies in gel shift experiments (data not shown). Given that the mutation of the motif appeared to increase basal promoter activity of several cartilage genes, it is likely that there is a repressive factor in RCS cells that binds to this motif. The identity of the N1-binding protein may be uncovered when the TRANSFAC database is more complete.

In summary, there are several challenges in the use of computational approaches to identify TF-binding motifs in mammalian organisms. One major challenge is the fact that mammalian species have very long intergenic sequence, and it is very difficult to detect regulatory sequences in such a large search space. In our approach, we used phylogenetic footprinting methods (Wasserman et al. 2000; Loots et al. 2002) and the constraints of the orientation and the distance between binding sites within a regulatory module to refine the search of regulatory sequences. Using an integration of multiple types of information indeed increases the performance of the computational prediction. Our results suggest that this appears to be particularly important for predicting functional SOX or HMG motifs that have inherently high degeneracy. Therefore, it is anticipated that the performance of computational approaches will continue to improve when additional data sets or novel computational models become available. This could include better evolutionary conservation models based on a larger set of completely sequenced genomes, or a better regulatory sequence analysis model that incorporates additional sequence signals (e.g., chromosome remodeling). Meanwhile, combining computational and experimental methods continues to be an advantageous approach, as they provide complementary and valuable information and/or validation that neither can achieve alone.

Methods

Definition of promoter sequences

A set of 18 orthologous human and mouse genes with documented expression in cartilage either during development or maintenance of the tissue was selected for this analysis: *ACAN*, *CILP*, *COL11A1*, *COL11A2*, *COL2A1*, *COL9A1*, *COL9A2*, *COL9A3*, *COMP*, *HAPLN1*, *CTGF*, *FGF18*, *IHH*, *MATN1*, *MATN3*, *MIA* (Cdrap), *PRG4*, and *SOX9*. An additional set of 13 previously reported liver-specific genes was also collected for testing specificity. The promoter sequences from these gene sets were selected for the analyses as follows. For most genes, the promoter sequence was defined as the 10-kb upstream and the 5-kb downstream sequence according to the transcriptional start site. For some genes, the promoter sequence was truncated when an upstream gene was encountered (e.g., the *COL11A2* gene) or when the translation start site was reached. Focusing on the 15-kb genomic sequence around the transcription start site for each gene in our analysis is keeping within the limit of current DNA recognition pattern programs. For each human gene, the mouse ortholog was determined by the NCBI's HomoloGene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene). The mouse ortholog was further verified for the reciprocal best match of the protein sequences using WU-BLAST (http://blast.wustl.edu/). Promoter sequences of human and mouse genes were retrieved from the UCSC genome browser (http://www.genome.ucsc.edu). Repetitive elements in the promoter sequences were masked by the program RepeatMasker (http://www.repeatmasker.org) with the slow and sensitive mode.

Determination of conserved regions in promoters

The conserved sequence regions in human and mouse orthologous promoters were determined by sequence alignment. Human and mouse promoters were aligned using WU-BLAST with a customized scoring matrix that considers the transition and the transversion rates and the background base frequencies estimated using all human and mouse promoters. This scoring matrix was designed to search for local alignments of an average of 60% sequence identity. The "-link" parameter of WU-BLAST was used to find the highest-scoring set of consistent alignments.

Identification of evolutionarily conserved TF-binding motifs

The program PATSER (Stormo et al. 1982) was used to search 436 vertebrate-specific weight matrix models collected from the TRANSFAC 7.2 database (Wingender et al. 1996; Matys et al. 2003) in the cartilage gene promoters. A short stretch of sequence was identified as a putative binding site of a TF if its score was higher than the cutoff score calculated by PATSER (Staden 1989). Both strands of the sequence were searched in our analysis. Only binding sites that were conserved in both human and mouse promoters according to the sequence alignment were deemed conserved sites (Loots et al. 2002). Only conserved sites were used in the following statistical analysis.

Enrichment of binding motifs

Assuming that the score of a weight matrix model is proportional to the free energy of binding (Stormo and Fields 1998), the mathematical formula of the binding probability of a site and its score has been described using the Boltzmann's distribution (Stormo 2000). For a weight matrix of a factor F and a binding site x scored $s(x)$, the probability of F binding this site is given by

$$P = Ae^{s(x)}$$

where A is a factor-specific constant related to several aspects including the binding specificity and the concentration of the TF. Therefore, for a sequence having a set of binding sites, denoted by X, the probability score of the factor F binding to this sequence via any of these sites is defined by (GuhaThakurta et al. 2002b)

$$P = \sum_{x \in X} e^{s(x)}$$

Given all the putative sites and their scores, one can calculate the probability score of one factor binding to a set of promoters by

$$P_{\text{genes}} = \left(\prod_{i=1}^{N} P_i \right)^{\frac{1}{N}}$$

To evaluate the over-representation of a transcription factor-binding site in the cartilage promoters, the probability score calculated using the cartilage promoters was compared to the probability score calculated using the promoters of all 14,127 human genes. For each transcription factor binding site, a

$\log(P_{cartilage\ genes}/P_{all\ genes})$ [denoted by $\log(P_c/P_a)$] value was calculated. This value is a measure of the over-representation of the binding sites of a transcription factor in the cartilage promoters (GuhaThakurta et al. 2002b; Hu et al. 2004).

Construction of a cartilage-specific SOX-binding profile

To construct a model of SOX-binding motifs that is specific to cartilage gene regulation, a set of 17 experimentally validated SOX-binding sites from three cartilage-characteristic mouse genes (*Col2A1*, *Col11A2*, and *Mia*) was collected (Supplemental Table 1). Human and rat orthologs of these genes were identified using the HomoloGene database, and genomic sequences of orthologous genes were aligned using the program WU-BLAST. Sequences in orthologous genes that were aligned to the experimentally validated sites were identified and included in the list of known SOX-binding sites. In this procedure, a total of 31 SOX-binding sites were compiled and then used to build the SOX-binding profile (Fig. 2A–C).

Identification of paired SOX-binding sites in cartilage genes

The conserved SOX-binding sites in cartilage promoters that are conserved between human and mouse were identified using the customized SOX-binding profile and the program PATSER (Stormo et al. 1982) as described above. To identify paired SOX-binding sites, the following criteria were applied: The spacer sequence between the two SOX sites is 3–8 bp, and the two SOX sites in each pair must have opposite orientations.

Identification of novel regulatory motifs in cartilage-characteristic genes

The DNA pattern recognition program CONSENSUS version 6C (Hertz and Stormo 1999) was used to search for shared motifs in the promoters of cartilage genes. Conserved sequences from human and mouse were searched separately on both strands of the DNA. This process generated two sets of candidate motifs from human and mouse. In both sets, motifs that were similar to characterized transcription factor-binding profiles in TRANSFAC were removed. Motifs that were similar to each other were also identified and merged by a dynamic programming algorithm using ALLR as the scoring scheme (Wang and Stormo 2003). These steps yielded a nonredundant list of novel regulatory motifs. The two lists of novel motifs were compared, and motifs that are shared in human and mouse were identified. We then searched for these novel motifs in promoters of 14,127 human and mouse orthologous gene pairs. The probability scores and the $\log(P_c/P_a)$ were calculated as described above.

Cell culture

The LTC-rat chondrosarcoma cells (RCS) (Kucharska et al. 1990; Mukhopadhyay et al. 1995; King and Kimura 2003) and mouse 3T3 embryo fibroblasts were used for in vitro functional testing of motifs (see Supplemental Methods for detailed culture conditions and preparation of nuclear extracts).

Electromobility gel shift assays (EMSA) and transient transfections

Validation of functional activity of predicted motifs was determined using EMSA assays (see Supplemental Methods for experimental details). Mutational analyses of the predicted SOX- or N1-binding sites in the human *COL9A1*, *COL2A1*, *COMP*, and mouse *Mia* promoters were further tested in transient transfection assays. Mutant constructs were created by using the Quikchange Site-Directed Mutagenesis kit (Stratagene) (see Supplemental Methods for primer sequences and details).

Acknowledgments

We thank Hua Yu for her technical assistance in preparation of mutant constructs. We also thank David Stokes (Thomas Jefferson University, Philadelphia) for providing us with the *COL9A1* promoter construct (p846). This work was supported by NIH grants RO1 AR36994, R01 AR45550, and RO1 AR50847 to L.J.S. and NIH grants HG00249 and GM63340 to G.D.S.

References

Akiyama, H., Chaboissier, M.C., Martin, J.F., Schedl, A., and de Crombrugghe, B. 2002. The transcription factor SOX9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of SOX5 and SOX6. *Genes & Dev.* **16:** 2813–2828.

Ao, W., Gaudet, J., Kent, W.J., Muttumu, S., and Mango, S.E. 2004. Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* **305:** 1743–1746.

Bi, W., Deng, J.M., Zhang, Z., Behringer, R.R., and de Crombrugghe, B. 1999. SOX9 is required for cartilage formation. *Nat. Genet.* **22:** 85–89.

Blanchette, M. and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12:** 739–748.

Blanchette, M., Bataille, A.R., Chen, X., Poitras, C., Laganiere, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D., et al. 2006. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* **16:** 656–668.

Bridgewater, L.C., Walker, M.D., Miller, G.C., Ellison, T.A., Holsinger, L.D., Potter, J.L., Jackson, T.L., Chen, R.K., Winkel, V.L., Zhang, Z., et al. 2003. Adjacent DNA sequences modulate SOX9 transcriptional activation at paired SOX sites in three chondrocyte-specific enhancer elements. *Nucleic Acids Res.* **31:** 1541–1553.

Chang, L.W., Nagarajan, R., Magee, J.A., Milbrandt, J., and Stormo, G.D. 2006. A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. *Genome Res.* **16:** 405–413.

Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., and Palsson, B.O. 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429:** 92–96.

Davies, S.R., Li, J., Okazaki, K., and Sandell, L.J. 2004. Tissue-restricted expression of the Cdrap/Mia gene within a conserved multigenic housekeeping locus. *Genomics* **83:** 667–678.

Furumatsu, T., Tsuda, M., Taniguchi, N., Tajima, Y., and Asahara, H. 2005. Smad3 induces chondrogenesis through the activation of SOX9 via CREB-binding protein/p300 recruitment. *J. Biol. Chem.* **280:** 8343–8350.

GuhaThakurta, D., Palomar, L., Stormo, G.D., Tedesco, P., Johnson, T.E., Walker, D.W., Lithgow, G., Kim, S., and Link, C.D. 2002a. Identification of a novel *cis*-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. *Genome Res.* **12:** 701–712.

GuhaThakurta, D., Schriefer, L.A., Hresko, M.C., Waterston, R.H., and Stormo, G.D. 2002b. Identifying muscle regulatory elements and genes in the nematode *Caenorhabditis elegans*. *Pac. Symp. Biocomput.* **2002:** 425–436.

Hertz, G.Z. and Stormo, G.D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15:** 563–577.

Hu, Y., Wang, T., Stormo, G.D., and Gordon, J.I. 2004. RNA interference of achaete-scute homolog 1 in mouse prostate neuroendocrine cells reveals its gene targets and DNA binding sites. *Proc. Natl. Acad. Sci.* **101:** 5559–5564.

Huang, W., Lu, N., Eberspaecher, H., and De Crombrugghe, B. 2002. A new long form of c-Maf cooperates with SOX9 to activate the type II collagen gene. *J. Biol. Chem.* **277:** 50668–50675.

Imamura, T., Imamura, C., Iwamoto, Y., and Sandell, L.J. 2005. Transcriptional co-activators CREB-binding protein/p300 increase chondrocyte Cd-rap gene expression by multiple mechanisms including sequestration of the repressor CCAAT/enhancer-binding protein. *J. Biol. Chem.* **280:** 16625–16634.

Inada, M., Yasui, T., Nomura, S., Miyake, S., Deguchi, K., Himeno, M.,

Sato, M., Yamagiwa, H., Kimura, T., Yasui, N., et al. 1999. Maturational disturbance of chondrocytes in Cbfa1-deficient mice. *Dev. Dyn.* **214:** 279–290.
Kamachi, Y., Uchikawa, M., and Kondoh, H. 2000. Pairing SOX off: With partners in the regulation of embryonic development. *Trends Genet.* **16:** 182–187.
Karsenty, G. and Wagner, E.F. 2002. Reaching a genetic and molecular understanding of skeletal development. *Dev. Cell* **2:** 389–406.
King, K.B. and Kimura, J.H. 2003. The establishment and characterization of an immortal cell line with a stable chondrocytic phenotype. *J. Cell. Biochem.* **89:** 992–1004.
Kou, I. and Ikegawa, S. 2004. SOX9-dependent and -independent transcriptional regulation of human cartilage link protein. *J. Biol. Chem.* **279:** 50942–50948.
Krivan, W. and Wasserman, W.W. 2001. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* **11:** 1559–1566.
Kucharska, A.M., Kuettner, K.E., and Kimura, J.H. 1990. Biochemical characterization of long-term culture of the Swarm rat chondrosarcoma chondrocytes in agarose. *J. Orthop. Res.* **8:** 781–792.
Lefebvre, V. and de Crombrugghe, B. 1998. Toward understanding SOX9 function in chondrocyte differentiation. *Matrix Biol.* **16:** 529–540.
Lefebvre, V., Huang, W., Harley, V.R., Goodfellow, P.N., and de Crombrugghe, B. 1997. SOX9 is a potent activator of the chondrocyte-specific enhancer of the pro alpha1(II) collagen gene. *Mol. Cell. Biol.* **17:** 2336–2346.
Lefebvre, V., Behringer, R.R., and de Crombrugghe, B. 2001. L-Sox5, Sox6 and Sox9 control essential steps of the chondrocyte differentiation pathway. *Osteoarthritis Cartilage* **9:** S69–S75.
Loots, G.G., Ovcharenko, I., Pachter, L., Dubchak, I., and Rubin, E.M. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* **12:** 832–839.
Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31:** 374–378.
Miller, J., Horner, A., Stacy, T., Lowrey, C., Lian, J.B., Stein, G., Nuckolls, G.H., and Speck, N.A. 2002. The core-binding factor β subunit is required for bone formation and hematopoietic maturation. *Nat. Genet.* **32:** 645–649.
Mukhopadhyay, K., Lefebvre, V., Zhou, G., Garofalo, S., Kimura, J.H., and de Crombrugghe, B. 1995. Use of a new rat chondrosarcoma cell line to delineate a 119-base pair chondrocyte-specific enhancer element and to define active promoter segments in the mouse pro-alpha 1(II) collagen gene. *J. Biol. Chem.* **270:** 27711–27719.
Nelander, S., Larsson, E., Kristiansson, E., Mansson, R., Nerman, O., Sigvardsson, M., Mostad, P., and Lindahl, P. 2005. Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals. *BMC Genomics* **6:** 68. doi: 10.1186/1471-2164-6-68.
Ng, L.J., Wheatley, S., Muscat, G.E., Conway-Campbell, J., Bowles, J., Wright, E., Bell, D.M., Tam, P.P., Cheah, K.S., and Koopman, P. 1997. SOX9 binds DNA, activates transcription, and coexpresses with type II collagen during chondrogenesis in the mouse. *Dev. Biol.* **183:** 108–121.
Nissen, R.M., Yan, J., Amsterdam, A., Hopkins, N., and Burgess, S.M. 2003. Zebrafish foxi one modulates cellular responses to Fgf signaling required for the integrity of ear and jaw patterning. *Development* **130:** 2543–2554.
Okazaki, K. and Sandell, L.J. 2004. Extracellular matrix gene regulation. *Clin. Orthop. Relat. Res.* **427** Suppl**:** S123–S128.
Okazaki, K., Li, J., Yu, H., Fukui, N., and Sandell, L.J. 2002. CCAAT/enhancer-binding proteins β and δ mediate the repression of gene transcription of cartilage-derived retinoic acid-sensitive protein induced by interleukin-1β. *J. Biol. Chem.* **277:** 31526–31533.
Okazaki, K., Yu, H., Davies, S.R., Imamura, T., and Sandell, L.J. 2006. A promoter element of the CD-RAP gene is required for repression of gene expression in non-cartilage tissues in vitro and in vivo. *J. Cell. Biochem.* **97:** 857–868.
Peirano, R.I. and Wegner, M. 2000. The glial transcription factor SOX10 binds to DNA both as monomer and dimer with different functional consequences. *Nucleic Acids Res.* **28:** 3047–3055.
Qiu, P., Qin, L., Sorrentino, R.P., Greene, J.R., Wang, L., and Partridge, N.C. 2003. Comparative promoter analysis and its application in analysis of PTH-regulated gene expression. *J. Mol. Biol.* **326:** 1327–1336.
Schneider, T.D. and Stephens, R.M. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* **18:** 6097–6100.
Sock, E., Pagon, R.A., Keymolen, K., Lissens, W., Wegner, M., and Scherer, G. 2003. Loss of DNA-dependent dimerization of the transcription factor SOX9 as a cause for campomelic dysplasia. *Hum. Mol. Genet.* **12:** 1439–1447.
Staden, R. 1989. Methods for discovering novel motifs in nucleic acid sequences. *Comput. Appl. Biosci.* **5:** 293–298.
Stormo, G.D. 2000. DNA binding sites: Representation and discovery. *Bioinformatics* **16:** 16–23.
Stormo, G.D. and Fields, D.S. 1998. Specificity, free energy and information content in protein–DNA interactions. *Trends Biochem. Sci.* **23:** 109–113.
Stormo, G.D. and Hartzell III, G.W. 1989. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci.* **86:** 1183–1187.
Stormo, G.D., Schneider, T.D., Gold, L., and Ehrenfeucht, A. 1982. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.* **10:** 2997–3011.
Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., and Jones, R.T. 1988. Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203:** 439–455.
Takagi, T., Moribe, H., Kondoh, H., and Higashi, Y. 1998. δEF1, a zinc finger and homeodomain transcription factor, is required for skeleton patterning in multiple lineages. *Development* **125:** 21–31.
Tan, L., Peng, H., Osaki, M., Choy, B.K., Auron, P.E., Sandell, L.J., and Goldring, M.B. 2003. Egr-1 mediates transcriptional repression of COL2A1 promoter activity by interleukin-1β. *J. Biol. Chem.* **278:** 17688–17700.
Tan, K., McCue, L.A., and Stormo, G.D. 2005. Making connections between novel transcription factors and their DNA motifs. *Genome Res.* **15:** 312–320.
Wagner, T., Wirth, J., Meyer, J., Zabel, B., Held, M., Zimmer, J., Pasantes, J., Bricarelli, F.D., Keutel, J., Hustert, E., et al. 1994. Autosomal sex reversal and campomelic dysplasia are caused by mutations in and around the SRY-related gene SOX9. *Cell* **79:** 1111–1120.
Wang, T. and Stormo, G.D. 2003. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19:** 2369–2380.
Wang, H., Zhang, Y., Cheng, Y., Zhou, Y., King, D.C., Taylor, J., Chiaromonte, F., Kasturi, J., Petrykowska, H., Gibb, B., et al. 2006. Experimental validation of predicted mammalian erythroid *cis*-regulatory modules. *Genome Res.* **16:** 1480–1492.
Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. 2000. Human–mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26:** 225–228.
Wegner, M. 1999. From head to toes: The multiple facets of SOX proteins. *Nucleic Acids Res.* **27:** 1409–1420.
Wheatley, S., Wright, E., Jeske, Y., McCormack, A., Bowles, J., and Koopman, P. 1996. Aetiology of the skeletal dysmorphology syndrome campomelic dysplasia: Expression of the SOX9 gene during chondrogenesis in mouse embryos. *Ann. N. Y. Acad. Sci.* **785:** 350–352.
Wingender, E., Dietze, P., Karas, H., and Knuppel, R. 1996. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24:** 238–241.
Xie, W.F., Kondo, S., and Sandell, L.J. 1998. Regulation of the mouse cartilage-derived retinoic acid-sensitive protein gene by the transcription factor AP-2. *J. Biol. Chem.* **273:** 5026–5032.
Xie, W.F., Zhang, X., Sakano, S., Lefebvre, V., and Sandell, L.J. 1999. *Trans*-activation of the mouse cartilage-derived retinoic acid-sensitive protein gene by SOX9. *J. Bone Miner. Res.* **14:** 757–763.
Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.
Zhang, P., Jimenez, S.A., and Stokes, D.G. 2003. Regulation of human COL9A1 gene expression. Activation of the proximal promoter region by SOX9. *J. Biol. Chem.* **278:** 117–123.

Received December 19, 2006; accepted in revised form July 18, 2007.