Available online at www.sciencedirect.com

SCIENCE DIRECT

Genomics 86 (2005) 295 – 305

GENOMICS

www.elsevier.com/locate/ygeno

Detection of potential *GDF6* regulatory elements by multispecies sequence comparisons and identification of a skeletal joint enhancer

Matthew E. Portnoy[a], Kelly J. McDermott[b], Anthony Antonellis[a], Elliott H. Margulies[a], Arjun B. Prasad[a], NISC Comparative Sequencing Program[a,c], David M. Kingsley[d], Eric D. Green[a,c], Douglas P. Mortlock[b,*]

[a]*Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA*
[b]*Department of Molecular Physiology and Biophysics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*
[c]*NIH Intramural Sequencing Center (NISC), National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA*
[d]*Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA 94305, USA*

Received 21 March 2005; accepted 24 May 2005
Available online 24 June 2005

Abstract

The identification of noncoding functional elements within vertebrate genomes, such as those that regulate gene expression, is a major challenge. Comparisons of orthologous sequences from multiple species are effective at detecting highly conserved regions and can reveal potential regulatory sequences. The *GDF6* gene controls developmental patterning of skeletal joints and is associated with numerous, distant *cis*-acting regulatory elements. Using sequence data from 14 vertebrate species, we performed novel multispecies comparative analyses to detect highly conserved sequences flanking *GDF6*. The complementary tools WebMCS and ExactPlus identified a series of multispecies conserved sequences (MCSs). Of particular interest are MCSs within noncoding regions previously shown to contain *GDF6* regulatory elements. A previously reported conserved sequence at −64 kb was also detected by both WebMCS and ExactPlus. Analysis of *LacZ*-reporter transgenic mice revealed that a 440-bp segment from this region contains an enhancer for *Gdf6* expression in developing proximal limb joints. Several other MCSs represent candidate *GDF6* regulatory elements; many of these are not conserved in fish or frog, but are strongly conserved in mammals.

Keywords: Gdf6; Enhancer elements; Bacterial artificial chromosome; Transgenes; Sequence analysis

The use of comparative sequence analysis for identifying the functional portions of complex genomes has great potential for facilitating the detection of noncoding functional elements. Pair-wise sequence alignments and comparisons can be used for this purpose, for example, by simply assessing the percentage sequence identity across windows of a defined size [1]. Such approaches have proven effective at identifying *cis*-acting regulatory sequences, including those associated with developmentally regulated vertebrate genes, which tend to have multiple, distantly located regulatory elements [2–4]. However, the use of pair-wise sequence comparisons is limited. There is often too much aligning sequence between pairs of closely related species, forcing the use of somewhat arbitrary thresholds (e.g., >70% identity across 100 bp). Meanwhile, there is often too little aligning sequence between pairs of more distantly related species, especially within noncoding regions. To overcome the limitations of simple pair-wise analyses, approaches for performing multi-species sequence comparisons have been developed [5–7]; these provide a more powerful means of identifying the most highly conserved genomic regions. It is thus of great interest to apply these new approaches for studying genomic regions thought to contain complex, *cis*-acting regulatory elements.

GDF6 is a member of the BMP (bone morphogenetic protein) gene family, a group of genes that encode secreted

* Corresponding author. Fax: +1 615 343 8619.
E-mail address: mortlock@chgr.mc.vanderbilt.edu (D.P. Mortlock).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.05.003

signaling factors. Like other BMP genes, *GDF6* is expressed in many anatomical locations during embryonic development, including numerous skeletal joints [8,9], and it is required for normal formation of limb, ear, and skull joints [9]. These findings are consistent with the known roles of BMP members in mediating many developmental processes, such as the regional control of bone growth and shape [8,10,11] and soft tissue development [12,13]. As with other developmentally important genes, the localized effects of BMPs seem to be largely controlled by modular arrangements of *cis*-acting regulatory sequences that control the expression of BMP genes in specific anatomical locations. These regulatory sequences can reside far away from the gene [14,15]; for example, mouse bacterial artificial chromosome (BAC)-transgene studies revealed that regulatory sequences are distributed across a region of more than 100 kb encompassing the *GDF6* gene and that these elements mediate *GDF6* transcription in limb joints, digits, retina, genitalia, laryngeal cartilages, skull bones, and other tissues [4].

To localize and identify further individual *GDF6 cis*-acting regulatory elements, we performed extensive multispecies sequence comparisons. Previously, we performed pair-wise comparisons using PipMaker to indicate conserved *GDF6* sequences [16]. Here, we describe the use of two multispecies comparative approaches for detecting highly conserved regions in and around *GDF6*. These methods detect a developmentally regulated *GDF6* enhancer that can direct gene expression in proximal limb joints in vivo. Our findings suggest that multispecies conserved sequence (MCS) analysis may be a sensitive approach for detecting other *GDF6* regulatory elements.

Results

Multispecies sequences and alignments

The 209-kb mouse BAC clone RPCI23-117O7 contains the entire *Gdf6* gene and extensive flanking regions [4]. Using the previously established sequence of this BAC as a reference, we have generated (~2.7 Mb) or obtained (~1.5 Mb) sequences of the orthologous genomic regions from 13 additional vertebrates (Fig. 1a). A previous PipMaker analysis of the mammalian *GDF6* sequences has been described [16]. To examine the degree of noncoding conservation in other vertebrates, we obtained sequences from additional species and performed MultiPipMaker analysis on the entire data set. For nine species (chimpanzee, baboon, cow, pig, cat, dog, rat, platypus, and zebrafish), appropriate BACs were isolated and sequenced as previously described [16,17]. For four species (human, chicken, *Fugu,* and *Xenopus*), the orthologous sequence was obtained from the whole-genome sequence assembles available at the UCSC Genome Browser (http://genome.ucsc.edu/) [18].

Pair-wise alignments were generated between the mouse reference sequence and each of the other species' sequences using BLASTZ [19], with the results visualized using MultiPipMaker [20]. Sequence coverage of the region immediately encompassing *Gdf6* was nearly complete for most species, with minor exceptions (Fig. 1b). The two *GDF6* exons are highly conserved across vertebrates, as are several intronic noncoding regions (Fig. 1b). As reported previously [16] much of the noncoding DNA flanking the *GDF6* gene is conserved among mammals. However, a more limited set of flanking regions is also conserved in chicken and *Xenopus.* The mouse–zebrafish and mouse–*Fugu* alignments in regions flanking *GDF6* appear to be spurious (i.e., related to simple repeat-like sequences; data not shown); however, true mouse–fish conserved orthologous sequences were found within the *GDF6* exons and intron.

WebMCS and ExactPlus analyses

Since MultiPipMaker essentially displays separate pair-wise comparisons, we then performed two multispecies comparative analyses (WebMCS and ExactPlus) to prioritize conserved regions by assessing multiple alignments

Fig. 1. Multispecies comparative sequence analysis of the genomic region encompassing *GDF6*. (a) Venn diagram showing the major cladistic relationships of the vertebrate species whose sequences were analyzed. (b) A low-resolution overview of MultiPipMaker analysis of the *GDF6*-containing sequences from 14 species, with the mouse sequence used as the reference. The horizontal colored bars at the top of the overview plot (red, yellow, green, blue, and magenta) indicate positions of five contiguous genomic regions in mouse that were previously identified by BAC-transgene analysis to contain different subsets of tissue-specific *Gdf6* regulatory enhancers [4]. The relative positions and orientations of mouse *Gdf6* and two pseudogenes (*Uqcrb* and *Gapdh*) are indicated, as is the 2.9-kb interval previously shown to contain a proximal joint enhancer (PJE) [4]. For each species, portions of the mouse reference sequence that align to that species' sequence are indicated by green and red bars (reflecting regions with >50 or >75% identity with the mouse sequence, respectively). Gray bars indicate known gaps in the sequence data that are greater than 1 kb. (c) UCSC Genome Browser-based view depicting the positions of MCSs around the mouse *Gdf6* gene (mm4/NCBI build 32, October 2003; shown is the interval chr4:9,641,000–9,850,732, which corresponds to bases 50,001–209,733 of mouse BAC RP23-117O7; GenBank No. AC058786). The top 10 tracks (labeled "EP") depict the positions of MCSs detected with ExactPlus; these are labeled using the format "EP: #-#-#," in which the three numbers indicate the initial seed length (in bases), minimum number of species whose sequence needs to match the initial seed region, and minimum number of additional species' sequences used to extend the initial match in either direction (see Results for details). The two WebMCS tracks depict MCSs corresponding to the top 5% (WebMCS-95) and 2% (WebMCS-98) most conserved sequence. The "Transgenes" track shows the positions of the previously tested 2.9-kb PJE fragment [4] and the 440-bp fragment tested in this study (see Figs. 3 and 4). The "Known Genes" track (top) shows the positions of *Gdf6,* a Riken clone transcript that probably originates from the *Uqcrb* pseudogene [4], and a LINE-containing transcript (GenBank No. U156547). The "Spliced ESTs" and "Non-Mouse mRNAs" tracks show data from the UCSC Genome Browser for positions of spliced mouse ESTs and regions having protein-coding homology to nonmouse mRNAs, respectively. The "RepeatMasker" track displays the locations of repetitive elements identified by the RepeatMasker program.

simultaneously. Both programs are designed to detect MCSs, although the underlying algorithms used by each are quite distinct. Note that we use the term MCS to refer to a conserved region detected by multispecies sequence comparisons, regardless of the specific method used to perform those comparisons. Both WebMCS and ExactPlus use the same MultiPipMaker-generated multisequence alignment as input and generate similar output files that can be uploaded to the UCSC Genome Browser for visualization (Fig. 1c).

WebMCS uses a previously described binomial-based approach [6] to derive a "conservation score" for each base in a reference sequence by analyzing windows across a multispecies sequence alignment. WebMCS can be implemented to detect different amounts of conserved sequence; for example, WebMCS-95 and WebMCS-98 identify the top 5% and 2% mostly highly conserved bases in the reference sequence, respectively. ExactPlus finds small blocks of bases (or "seeds") of a designated size such that each base in a block is identical across a defined minimum number of

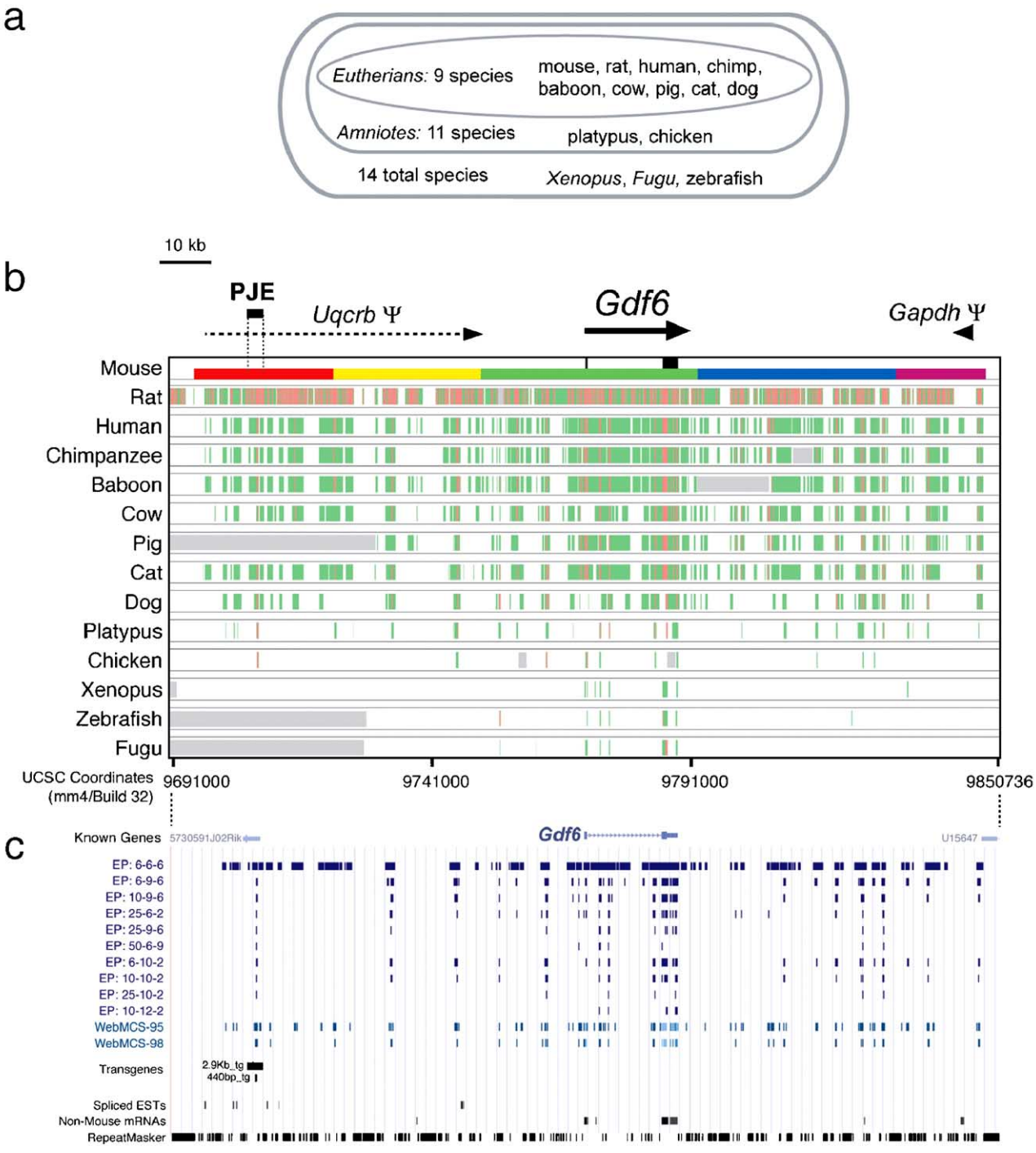

Table 1
MCS detection with WebMCS

MCS detection method	No. MCSs detected	Total MCS bases	Avg. MCS length (bases)	No. of coding MCSs[a]	No. of noncoding MCSs[b]	Coding bases overlapping MCSs[c]	Coding bases missed[d]	Noncoding MCS bases[e]	Sensitivity of detecting coding bases[f]	Specificity of detecting coding bases[g]
WebMCS-95	153	10,488	69	8	145	1111	251	9377	0.816	0.106
WebMCS-98	58	4201	72	9	49	829	533	3372	0.609	0.197

[a] Number of MCSs that overlap protein-coding sequence by at least 1 base. Note that the only protein-coding sequence in the region resides in *GDF6* exons 1 and 2.
[b] Number of MCSs that do not overlap protein-coding sequence by at least 1 base.
[c] MCS bases that overlap the 1362 bases of *GDF6* coding bases.
[d] *GDF6* coding bases not overlapping MCSs.
[e] MCS bases not overlapping *GDF6* coding sequence (total MCS bases − coding bases).
[f] Coding bases overlapping MCSs/(coding bases overlapping MCSs + coding bases missed).
[g] Coding bases overlapping MCSs/total MCS bases.

species (Antonellis et al., manuscript in preparation). The seeds can then be extended in either direction based on identity across a separately defined minimum number of species (see Materials and methods for details). The extension step was designed in an attempt to detect ancient, strongly conserved sequences that could represent the core of a larger functional element. For example, in regulatory elements such as enhancers, core transcription factor-binding sites may be highly conserved while flanking sequences may have evolved considerably.

To assess qualitatively how WebMCS and ExactPlus perform on the *GDF6* data set, we ran each with different input parameters and quantified the number of MCSs and amount of conserved sequence detected in each case. WebMCS-95 and WebMCS-98 detect 153 (average size of 69 bases) and 58 (average size of 72 bases) MCSs, respectively, within the roughly 209,000-base multisequence alignment (Table 1). Of these, 8 and 9 MCSs, respectively, overlap the two *GDF6* protein-coding exons (Fig. 2). Not surprisingly, ExactPlus detects different

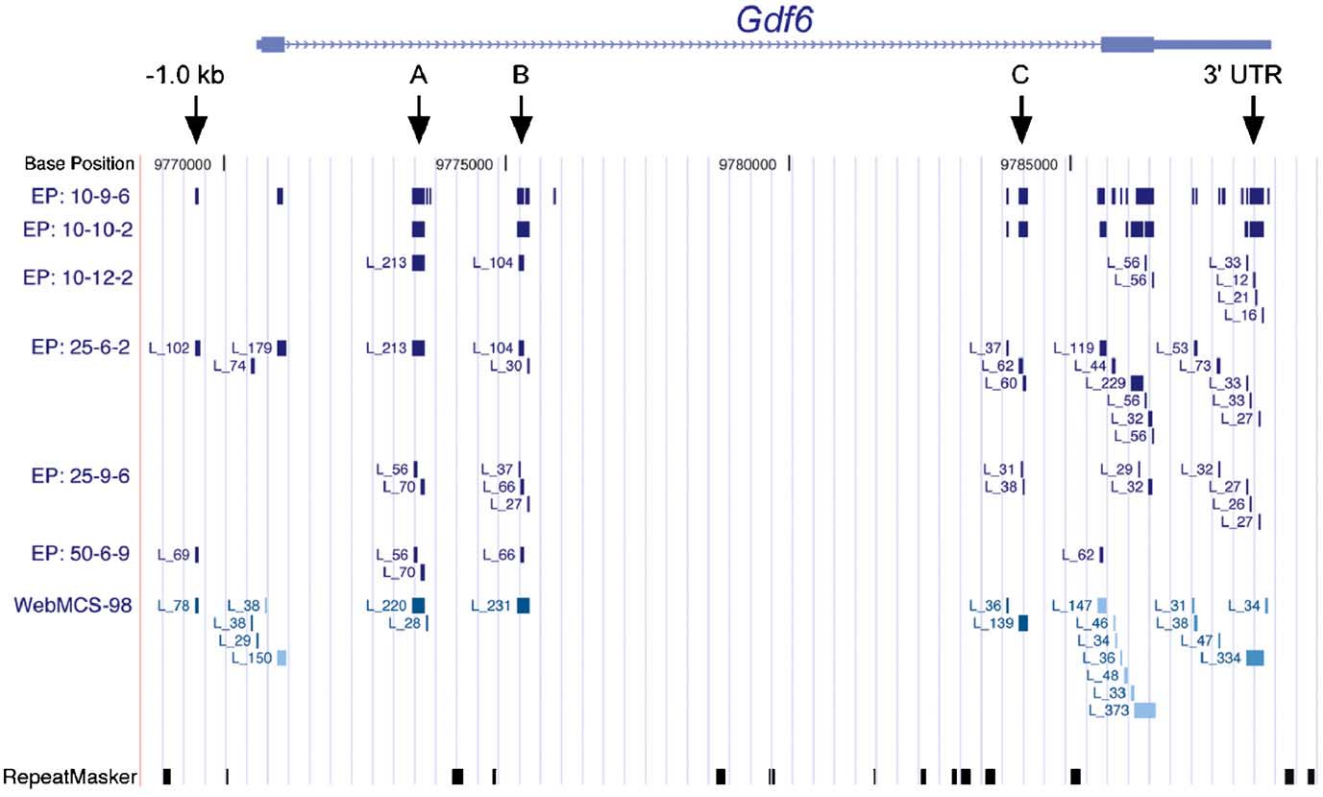

Fig. 2. MCSs in the immediate region spanning the *GDF6* transcription unit. An expanded 21-kb region from Fig. 1c highlights the interval spanning from 2 kb upstream to 1 kb downstream of the *GDF6* gene in the UCSC Genome Browser-based view. WebMCS-98 and selected ExactPlus data tracks are shown below the *GDF6* gene structure (see Fig. 1 for details). Arrows indicate MCSs located outside the *GDF6* exons (see text for details). Individual MCSs in certain tracks have been labeled with "L_#" to indicate the length of the MCS in bases. In the WebMCS-98 track, the light and dark shading of MCSs is used to represent MCSs that do and do not overlap, respectively, a *GDF6* exon by at least 1 base.

numbers of MCSs depending upon the parameters used for initial seed size and for the minimum number of species required for seeds and extended bases. For example, using low-stringency parameters (6-base seeds initiated with 6 species and extended in 6 species, designated as 6-6-6), ExactPlus detects 1621 MCSs (Fig. 1c and Table 2). If the seed length is increased to 10 bases and the number of initial species is increased to 9 (10-9-6), only 136 MCSs are detected (Fig. 1c and Table 2). Interestingly, using the latter parameters, nearly all (98.8%) of the MCS bases detected by ExactPlus overlap those detected by WebMCS-95 (Table 2).

As expected, increasing the ExactPlus initial seed length to 25 or 50 bases greatly reduces the number of identified MCSs; however, greater than 97% of the resulting MCS bases overlap with those detected by WebMCS-95 (Fig. 1c and Table 2). Several distal flanking regions are still detected by ExactPlus using a 50-base seed, including: (1) a region at −63.2 kb within the previously mapped limb joint regulatory segment [4], (2) a region at −1.0 kb, (3) two regions in the *GDF6* intron, (4) portions of the *GDF6* exon 2 coding region and the 3′ UTR, and (5) two regions at +53.7 and +57.7 kb (Figs. 1c, 2, and 3; note that all coordinates are given with respect to the *GDF6* 5′ end). With sequences from 9 eutherian mammals included in our data set, the initial requirement of 10 species' sequences to match the seed might discriminate between eutherian-specific conserved sequences and more ancient conserved sequences. For example, a conserved region (approximately 75 bases) roughly 1 kb upstream of the first *GDF6* exon is nearly identical in all eutherian mammals but not in any noneutherian species (Figs. 1b, 1c, and 2); this may reflect a eutherian-specific regulatory sequence(s).

Increasing the ExactPlus minimum-species number at initial seed to 12 results in the identification of regions that are conserved in all mammals and at least one of *Xenopus*, *Fugu*, or zebrafish. Interestingly, this identifies two MCSs in the *GDF6* intron (Figs. 1c and 2, Table 2), indicating ancient conserved sequences. This is consistent with the recent findings that noncoding sequences that are highly conserved between mammals and fish often reside near developmentally regulated genes [21,22].

Conserved sequences in the GDF6 promoter, intron, and 3′ UTR

The above results indicate the presence of several highly conserved sequences close to the *GDF6* gene that may be candidates for ancient cis-acting regulatory elements. Indeed, in zebrafish and *Xenopus*, *Gdf6* is transcribed in the dorsal retina and dorsal neural tube [23,24], and similar patterns of mouse *Gdf6* expression were implied by BAC-transgene experiments [4]. Therefore we scrutinized the area proximal to *GDF6* in more detail. Fig. 2 depicts the positions of MCSs detected by WebMCS and ExactPlus within the region immediately encompassing the *GDF6* transcription unit. At −1.0 kb relative to the *GDF6* translation start site an MCS was detected by WebMCS-98 and ExactPlus (using two sets of ExactPlus parameters, 10-9-6 and 50-6-9; this MCS is labeled "L_69" in the EP: 50-6-9 parameter track within Fig. 2). WebMCS-98 and ExactPlus (with 25-6-2 parameters) also detect an MCS just upstream of the mRNA 5′ end (the leftmost L_38 in WebMCS-98 and L_74, respectively), suggesting conservation within the *GDF6* promoter; this region was not detected by ExactPlus using a shorter initial seed length and a larger number of initial and extension species (EP: 10-9-6). Both of these regions are within a genomic interval shown to have neural tube and brain enhancer activity affecting *GDF6* expression [4].

Three notably large ExactPlus-detected MCSs reside within the *GDF6* intron (labeled A, B, and C in Fig. 2). These may function to regulate *GDF6* expression in dorsal

Table 2
Relationship of MCSs detected by ExactPlus and WebMCS-95

MCS detection method	No. MCSs detected	Total MCS bases	No. MCSs overlapping with WebMCS-95[a]	MCS bases overlapping with WebMCS-95[b]	WebMCS-95 sensitivity[c]	WebMCS-95 specificity[d]
WebMCS-95	153	10,488	N/A	N/A	N/A	N/A
ExactPlus 6-6-6	1621	16,002	575	7787	0.487	0.742
ExactPlus 6-9-6	320	5089	275	4660	0.916	0.444
ExactPlus 10-9-6	136	3118	133	3080	0.988	0.294
ExactPlus 25-6-2	42	2973	42	2890	0.972	0.276
ExactPlus 25-9-6	21	901	21	901	1.000	0.086
ExactPlus 50-6-9	8	543	8	543	1.000	0.052
ExactPlus 6-10-2	114	4348	110	4140	0.952	0.395
ExactPlus 10-10-2	57	2844	57	2823	0.993	0.269
ExactPlus 25-10-2	7	637	7	637	1.000	0.061
ExactPlus 10-12-2	8	511	8	511	1.000	0.049

N/A, not applicable.
[a] Number of ExactPlus MCSs that overlap with WebMCS-95 MCSs by at least 1 base.
[b] Number of ExactPlus MCS bases that overlap with WebMCS-95 MCS bases.
[c] Fraction of ExactPlus MCS bases that overlap with WebMCS-95: MCS bases overlapping with WebMCS-95 MCSs/total ExactPlus MCS bases.
[d] Fraction of ExactPlus MCS bases also detected by WebMCS-95: MCS bases overlapping with WebMCS-95/10,488.

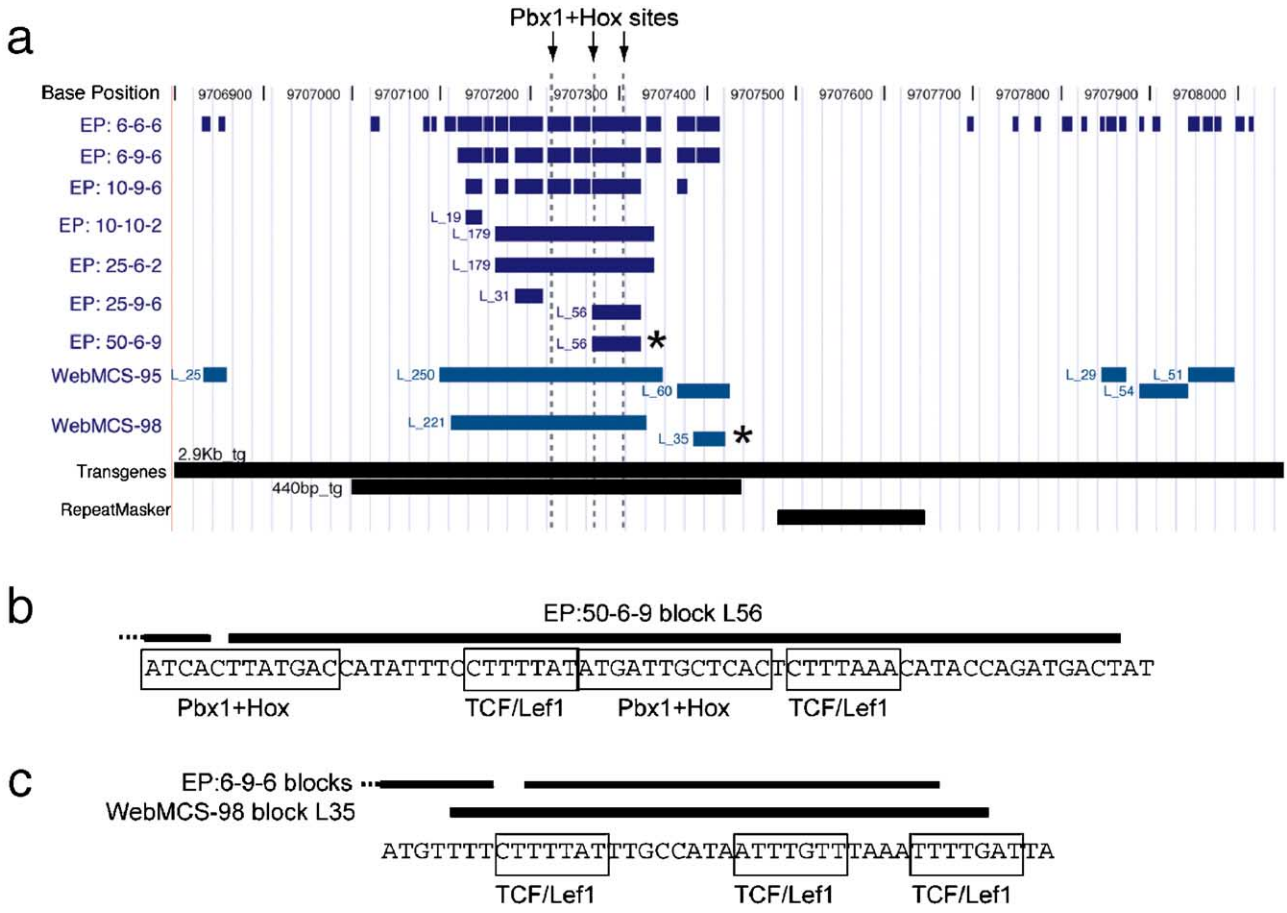

Fig. 3. Highly conserved regions and potential transcription factor binding sites in the PJE region. (a) UCSC Genome Browser-based view highlighting the positions of MCSs in a 1.2-kb interval within the PJE region (see Figs. 1b and 1c for orientation). Individual MCSs in certain tracks have been labeled with "L_#" to indicate the length of the MCS in bases. "2.9Kb_tg" and "440bp_tg" represent the locations of the 2.9-kb transgene [4] and 440-bp transgene (PJE-440; see Results), respectively. Arrows indicate locations of three potential Pbx1/Hox binding motifs (see Results). (b and c) Putative transcription factor binding motifs within specific MCSs (indicated in (a) by asterisks). Coverage across the motif regions is shown for selected MCSs that were detected using different parameters, as indicated.

retina or genitalia [4]. In addition to the MCSs overlapping the coding portion of exon 2, an MCS was also found in the 3′ UTR just upstream of the noncanonical ATTAAA polyadenylation signal. This conserved sequence may play a role in the posttranscriptional regulation of *GDF6* mRNA. We were unable to detect evidence for an RNA secondary structure within this MCS (e.g., long hairpins or stem–loops). Interestingly, a highly conserved sequence has been found in the 3′ UTR of another BMP family gene, *BMP2*, where it may regulate mRNA stability [25].

Conserved sequences regulate GDF6 expression

In the embryonic limb bud, *GDF6* is transcribed in a stripe-like pattern that marks the locations of interzones, histologically distinct regions that give rise to skeletal joints [8,9]. Previous transgenic studies indicated that a region far upstream (>60 kb) of *GDF6* is required for its expression in proximal limb joints during embryonic development [4]. Furthermore, a 2.9-kb fragment from this region can drive expression of a LacZ-containing minigene cassette in elbow, knee, shoulder, and hip joints [4]. The presumed enhancer in this fragment was called the PJE (proximal joint element). Two potential heterodimeric binding sites for Pbx1/Hox transcription factors [26] have been detected in this region, consistent with a role for this sequence in proximal limb patterning [16,27].

We therefore examined WebMCS and ExactPlus results for this region. Both methods detect multiple MCSs in the 2.9-kb region (Fig. 3). Both programs also identify a highly conserved region of approximately 300 bp within the 2.9-kb region found to contain the PJE (Fig. 3a). This conserved region is present in all mammalian sequences (except pig, for which sequence coverage is lacking in this region), but not *Fugu*, zebrafish, or *Xenopus* (Fig. 1b). Thus, this likely reflects a mammal-specific conserved element. There is also a less conserved region several hundred bases downstream that is detected by WebMCS-95 and ExactPlus with low-stringency parameters (6-6-6); however, this region is not detected by WebMCS-98 or ExactPlus with higher stringency parameters (Fig. 3a).

Within the highly conserved approximately 300-bp region reside several stretches of high conservation, as detected by ExactPlus. One block of 56 bp was detected using either 25:9:6 or 50:6:9 parameters; this may reflect a core enhancer element (Fig. 3b). Closer inspection of this block revealed that it overlaps with the second of the two previously reported Pbx1/Hox binding motifs [16]. A third Pbx1/Hox binding motif was also found. Interestingly, this Pbx1/Hox motif is flanked by two CTTT(T/A)A(T/A) motifs, which are similar to the consensus Lef1/TCF1 binding site [28,29]. WebMCS-98 detects one long (L_221) and one short (L_35) MCS in this region. While the former contains the three Pbx1/Hox motifs, the latter contains three imperfect matches to the Lef1/TCF consensus sequence (Fig. 3c). These results are particularly striking given that the Lef1/TCF factors function in the Wnt ligand signaling pathway [30,31] and that Wnt14 has been proposed to be a key regulator of early joint development [32,33]. These data suggest that Hox and Pbx factors may interact with the Wnt signaling pathway to regulate joint-specific gene expression.

To test if the highly conserved L_56 region indicates the core of a *cis*-acting sequence that enhances *GDF6* expression in the proximal limb joint, a 440-bp segment encompassing L_56 was subcloned into an *Hsp68* promoter–LacZ minigene construct [4,34]. Transgenic mouse embryos containing this construct (PJE-440) were analyzed midgestation for LacZ expression, with the results summarized in Fig. 4 and Table 3. Of seven independently generated transgenic embryos, several had staining in the eye, which we have found to be a common expression artifact of the

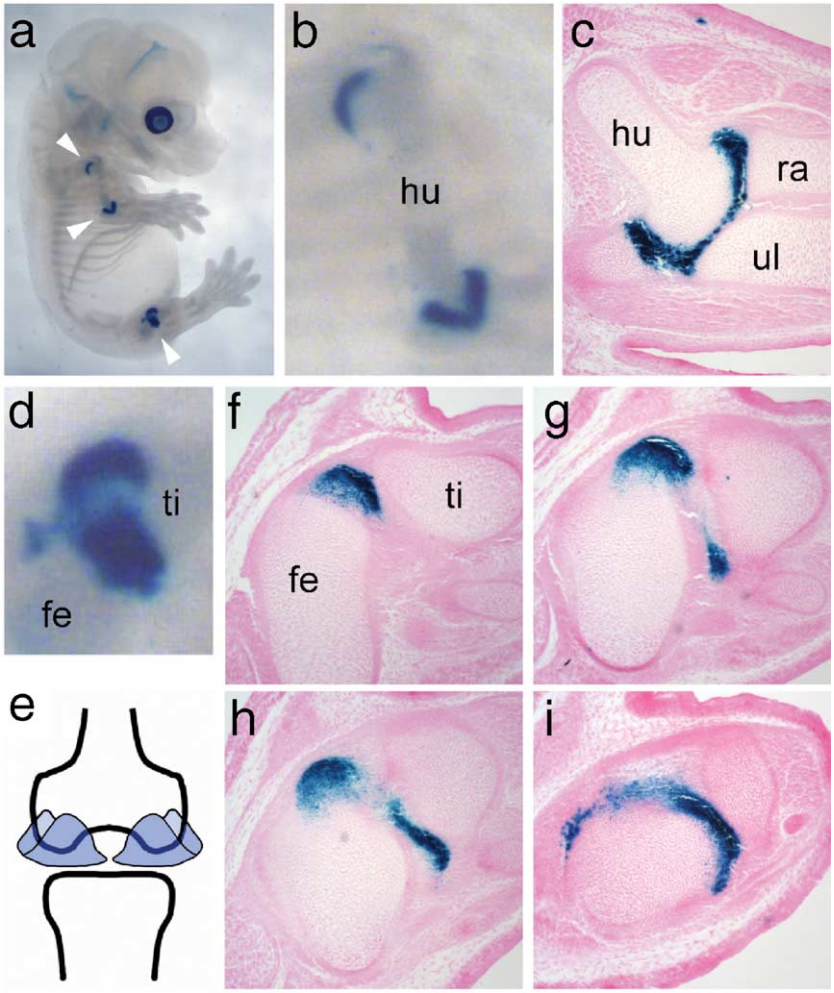

Fig. 4. A conserved sequence located 64 kb upstream of *GDF6* functions as a transcriptional enhancer in proximal limb joints. The PJE-440 construct (containing a highly conserved 440-bp portion of the PJE region cloned into a heat-shock promoter/LacZ reporter vector) was injected into one-cell mouse embryos. Transgenic embryos were harvested at 14.5 dpc, cut in half down the midline, stained with X-gal, and cleared with glycerol for imaging. (a) A representative embryo is shown, with white arrowheads indicating LacZ expression in shoulder, elbow, and knee joints. The hip joint was also stained but not retained in the photographed half of the embryo after dissection (data not shown). LacZ expression in the eye is a cryptic effect of the transgene vector and is likely not associated with the 440-bp sequence (see Results). The faint staining in the brain is due to background (i.e., nontransgenic) β-galactosidase activity. (b) Close-up view of the upper arm, showing shoulder and elbow joint staining at the ends of the humerus. (c) Section through the elbow of a transgenic embryo reveals that staining is specific to cells in the humeroradial and humeroulnar joints. (d) Close-up view of the knee joint showing two separate major domains of staining, plus a faint area of more proximal staining near the future patella. (e) Schematic drawing of the two major staining domains relative to the femur and tibia. (f–i) Near-adjacent serial sections through the knee of a transgenic embryo. hu, humerus; ra, radius; ul, ulna; fe, femur; ti, tibia.

Table 3
PJE-440 transgene expression in mouse embryos

Embryo	Proximal limb joints	Ectopic expression
1	+++	Some phalangeal joints, limb tendons, brain, neural tube, stomach
37	+++	Retina
38	–	Widespread ectopic staining
65	+	
69	–	Brain (weak)
72	+++	Retina
82	++	Retina, forebrain, neural tube
88	+[a]	None

[a] Weak staining in elbow joint only.

Hsp68–LacZ vector backbone (D.P.M., D.M.K, unpublished observations; Catherine Guenther, personal communication) (Fig. 4a and Table 3). However, five of the embryos showed strong LacZ expression in the nascent elbow, knee, shoulder, and hip joints (Figs. 4a, 4b, and 4d), in a pattern essentially identical to that previously observed with the 2.9-kb segment [4]. This indicates that the 440-bp fragment can function as a modular regulatory element capable of activating a heterologous promoter specifically in proximal limb joints.

Histological sections of PJE-440-containing embryos confirmed that the LacZ expression is restricted to the joints (Figs. 4c and 4f–4i), generally within the articular cartilage and also in intervening cells between the cartilage elements and at a developmental stage prior to cavitation of the joint cavity. Interestingly, in the knee joint, two arc-shaped domains of LacZ expression are apparent. Sectioning revealed that these domains curve around the lateral and medial condyles of the distal femur where it articulates with the tibia (Figs. 4d–4i). We also tested a larger construct that, in addition to the 440-bp segment, contains the less well conserved region about 400 bp downstream (see EP: 6-6-6 and WebMCS-95 blocks, right side of Fig. 3a). This larger construct confers LacZ expression in proximal limb joints in a pattern indistinguishable from that of the PJE-440 construct (data not shown). While the functional relevance (if any) of this less well conserved region is unclear, the 440-bp region seems to contain the *GDF6* PJE.

Discussion

Previous human and mouse sequence comparisons suggested the presence of numerous conserved noncoding regions within and flanking the *GDF6* gene [4,16]. These regions represent tantalizing candidates for serving as *cis*-acting regulatory elements that mediate the complex expression of *GDF6* [8,9]. Here, we report an extension of those studies that has involved the generation and comparison of the sequence of the genomic region encompassing *GDF6* in multiple additional vertebrates. Two analytical methods, WebMCS and ExactPlus, were used to analyze these multispecies sequences, allowing the identification of a large set of MCSs. The majority of these reside within noncoding regions. More detailed functional analysis of one MCS region identified an enhancer (the PJE) that mediates *GDF6* expression in limb joints. This has refined the PJE location from 2.9 kb to 440 bp (Figs. 3 and 4), although both WebMCS and ExactPlus suggest the critically conserved core of the PJE is closer to 300 bp (Fig. 3; see below).

Our findings indicate that ExactPlus and WebMCS represent complementary approaches for identifying conserved sequences, with degree of overlap depending on input parameters. Though virtually all MCSs detected by highly stringent implementation of ExactPlus tend to overlap with WebMCS elements, when smaller seeds and matching-species parameters are used some important differences are notable. For example, ExactPlus is more likely to detect small sequence blocks (e.g., on the scale of transcription factor-binding sites) than WebMCS, which analyzes sequences in 25-base windows [6]. We also suggest that ExactPlus may be a useful alternative to methods that search for consensus motifs for transcription factor binding sites. For example, the potential nonconsensus TCF/Lef binding sites depicted in Fig. 3 may fall into this category. At the same time, regulatory elements often contain multiple transcription factor-binding sites that function redundantly, with the amount of sequence conservation at any one site being variable among species. In these situations, WebMCS should be particularly effective since it assesses conservation with sliding windows and does not require 100% sequence identity. In other words, WebMCS probably tolerates plasticity better than ExactPlus.

The amount and distribution of conserved sequences across the *GDF6* genomic region are broadly similar among the nonrodent eutherian mammals (see Fig. 1b). A very different pattern is seen for platypus (a noneutherian mammal), with conservation largely confined to the most stringently defined MCSs. Comparisons with the orthologous chicken sequence reveal similar findings, though fewer regions of conservation are noted. Conservation between the mouse and the nonamniote species (*Xenopus tropicalis*, *Fugu*, and zebrafish) is mostly confined to *GDF6* coding sequences and a minimal number of noncoding regions near *GDF6* and was well detected by both WebMCS and ExactPlus. Across the *GDF6* region, the platypus sequence appears to be the most effective for identifying noncoding MCSs, particularly for segments flanking the gene. These results are consistent with a recent larger study investigating the utility of noneutherian mammal (marsupial and monotreme) sequences for identifying conserved genomic regions [35].

MCS analysis indicates the PJE is well conserved among amniote species. The discovery of the PJE also provides new insights into potential roles for *GDF6* in the patterning

of embryonic limb joints. *GDF6* transcription in the embryonic elbow joint has been previously documented [4,9]. However, because the *GDF6* mRNA is transcribed at low levels we have found it difficult to characterize its expression in the other proximal limb joints. In contrast, the more sensitive transgenic LacZ assay has proven valuable for characterizing the function of the *GDF6* PJE. Taken together with our comparative sequence analyses, the LacZ expression data strongly suggest a conserved role for *GDF6* in proximal joint patterning. Interestingly, in the knee joint PJE-regulated transgene expression was restricted to subregions within the joint cavity. Given the ability of GDF proteins to stimulate chondrogenesis [36–38], the transgenic expression pattern suggests that *GDF6* regulates growth of the adjacent femoral condyles.

MCS analysis was useful for stimulating hypotheses for candidate PJE-binding factors. The PJE contains putative binding sites for Pbx1/Hox heterodimers, factors that both pattern proximal limb tissue [27]. Studies in *Drosophila* suggest that Hox-binding regulatory modules typically require additional inputs from interacting signaling pathways [39]. We hypothesize that the PJE integrates positional information provided by the Pbx/Hox factors with other signaling pathways that specify skeletal joints. Lef1/TCF factors function in the Wnt signaling pathway, raising the intriguing possibility that Hox and Pbx patterning factors interact with Wnt proteins to specify expression in developing joints. Indeed, Wnt14 is thought to direct early synovial joint development [32]. Further testing should determine if Pbx, Hox, and/or TCF/Lef proteins bind the PJE in vitro or in vivo. We also note that the long tracts of sequence conservation in the PJE are not easily explained by the detected Pbx/Hox and Lef1/TCF binding sites, so the binding of additional transcription factors is probably important for PJE function. Further analysis of ExactPlus data may help characterize binding sites for such factors.

In addition to skeletal joints, *GDF6* is transcribed in the developing skull, larynx, and digits, and previous BAC-transgenic data suggest that *GDF6* is expressed in neural tube, retina, teeth, and other tissues [4]. Interestingly, these structures show varying degrees of morphological diversity across vertebrates. Thus, *GDF6* enhancers might be expected to vary considerably with respect to sequence conservation. Studies of other developmentally regulated genes have shown that the more conserved the sequence between highly divergent species (for example, mammals and fish), the more likely it is to be functionally important [21,22]. However, other studies indicate that many sequences that serve to regulate gene expression are not conserved between mammals and fish [40,41]. At least three *GDF6* noncoding MCSs are conserved between mouse and frog and/or fish, making them candidates for being ancient *GDF6* enhancers. Other *GDF6*-associated MCSs, such as the PJE, are not found in fish or frog. We speculate that this reflects the evolution of new *GDF6* regulatory capabilities correlating with morphological adaptations in amniotes (e.g., limb joint adaptations compatible with terrestrial mobility). Further transgenic experiments to compare the regulatory functions of candidate *GDF6* enhancers in mammals, frog, and fish will be useful for investigating this possibility.

Materials and methods

BAC clones and sequences

The *Gdf6*-containing mouse BAC RPCI23-117O7 sequence (GenBank No. AC058786) served as the reference for comparative analyses and corresponds to UCSC Genome Browser coordinates chr4:9,641,000–9,850,732 (mm4/NCBI build 32, October 2003). Orthologous human and rat genomic sequences were retrieved from the respective genome-wide data sets [16,42,43]. Minimally overlapping *GDF6*-containing BACs from rat, chimpanzee, baboon, cow, pig, cat, dog, platypus, and zebrafish were identified [17] and sequenced as part of the NISC Comparative Sequencing Program [5] (www.nisc.nih.gov). Accession numbers for all of the above BAC sequences were previously reported [16] except for zebrafish BAC CH211-216G21 (GenBank No. AC139623). Sequences were finished to a grade that is of intermediate quality between phase I (full shotgun) and phase III (contiguous, near perfect); this is called "comparative-grade finished sequence" and is an enhanced version of phase II finished sequence [44]. Additionally, sequence from an orthologous *Gdf6*-containing *X. tropicalis* BAC (CH216-129O13, GenBank No. AC147884) was identified and downloaded from the DOE/JGI Web site (http://genome.jgi-psf.org/xenopus). Additional orthologous sequences were obtained from the UCSC Genome Browser as follows: (1) chimpanzee, used to fill in gaps in the chimpanzee BAC sequence (NCBI chimpanzee draft genome sequence build 1, November 2003; chr7:99,408,595–99,511,759); (2) chicken (build galGal2, Chicken Genome Sequencing Consortium, February 2004; chr2:124,800,000–124,950,000; also see www.genome.wustl.edu/projects/chicken); (3) *Fugu rubripes* (*Fugu* build fr1, August 2002, chrUn:140,629,874–140,862,748; International *Fugu* Genome Consortium; also see www.fugu-sg.org).

Comparative analysis and MCS tools

Multispecies sequence comparisons were performed using MultiPipMaker [20], as described [16]. MCSs were detected using ExactPlus (Antonellis et al., manuscript in preparation; http://research.nhgri.nih.gov/projects/exactplus) or WebMCS [6] (http://research.nhgri.nih.gov/MCS). Briefly, ExactPlus finds small blocks of bases (or "seeds") of a designated size such that each base in a block is identical across a defined minimum number of species in the MultiPipMaker alignment. The seeds can then be extended in either

direction in a base-by-base fashion, on the condition that each extended base must be identical in a defined minimum number of species (note that the minimum number of species can be different in the initial seed and subsequent extension steps). The extension steps attempt to detect the presence of ancient, strongly conserved sequences that represent the core of a larger functional element. For example, in regulatory elements such as enhancers, the core transcription factor-binding sites can be highly conserved, while the immediate flanking sequences may have evolved considerably. For all analyses, the presumptive mouse *Gdf6* transcription start site was assumed to be the 5′ end of a *Gdf6* mRNA sequence (GenBank No. AJ537425).

Transcription factor site analysis

MacVector software was used to scan the mouse PJE sequence for previously reported consensus binding sites for Pbx1/Hox heterodimers [26] and Lef/TCF. The Pbx1/Hox binding sequence ATGATTTA(C/T)GAC was previously determined using heterodimers of Pbx1 protein with Hox proteins of Hox groups 9 and 10 [26]. The Lef/TCF binding consensus used was CTTTG(A/T)(A/T), to reflect both the reported TCF1 binding site CTTTGTT [29] and the reported Lef1 binding site CCTTTG(A/T)(A/T) [28]. Consensus binding sites were scanned using the MacVector Nucleic Acid Subsequence analysis tool set to permit mismatches anywhere within the consensus sequence. Up to three mismatches were permitted for the Hox/Pbx1 consensus and up to one mismatch was allowed for the Lef/TCF consensus.

Generation of the PJE-440 transgene construct

A 440-bp segment was amplified by PCR from the *Gdf6*-containing mouse BAC C-bGeo [4] using primers 5′-GTGAGGCCAAACAGGCCAATCCCTGTATTACAA-GGACTCAAATTCT-3′ and 5′-GTGAGGCCTGTTTGGC-CTATACCAACACCTATATAATCAAAATTTAAACA-3′. The resulting product was cloned into the *Sfi*I site of pSfi-HspLacZ [4], a derivative of pHsp68-LacZ [15], linearized with *Sal*I, and purified prior to pronuclear injection into mouse embryos [15].

Generation and analysis of transgenic mouse embryos

Transgene constructs were microinjected into FVB or C57BLB6/D2 F1 hybrid mouse embryos using standard pronuclear injection methods by the Vanderbilt University Transgenic Mouse/ES Cell Shared Resource, in accordance with protocols approved by the Vanderbilt University Institutional Animal Care and Use Committee. Transgenic embryos were verified by PCR from yolk sac or tail DNA samples. Whole-mount X-gal staining was performed as described [4]. For histological analysis, X-gal-stained embryos were dehydrated in ethanol/1× PBS, equilibrated with isopropanol/paraffin, and embedded in paraffin overnight. Sections (10–12 μm) were cut using a microtome, transferred to glass slides, dewaxed briefly with xylene, and counterstained with eosin prior to microscopic imaging.

Acknowledgments

We thank numerous people associated with the NISC Comparative Sequencing Program, in particular Robert Blakesley, Gerry Bouffard, Jennifer McDowell, Baishali Maskeri, Nancy Hanson, the many dedicated mapping and sequencing technicians, and other staff. We also thank Laura Selenke and Lissett Ramirez for expert technical assistance and Karen Deal, Maureen Gannon, Anna Means, and Laura Wilding for generously sharing equipment and advice. We also thank Ronald Chandler for helpful discussions. Kelly J. McDermott was supported by NIH Genetics Training Grant 1T32GM62758-03. David M. Kingsley was supported by NIH Grant 5R37AR042236-12. Douglas P. Mortlock was supported by NIH Grant 1R01HD47880-01. Transgenic mice were generated by the Vanderbilt University Transgenic and ES Cell Shared Resource, which is supported by the Vanderbilt Cancer, Diabetes, Kennedy, and Vision Centers.

References

[1] G.G. Loots, et al., Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons, Science 288 (2000) 136–140.
[2] L.A. Lettice, et al., A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly, Hum. Mol. Genet. 12 (2003) 1725–1735.
[3] L.A. Pennacchio, Insights from human/mouse genome comparisons, Mamm. Genome 14 (2003) 429–436.
[4] D.P. Mortlock, C. Guenther, D.M. Kingsley, A general approach for identifying distant regulatory elements applied to the Gdf6 gene, Genome Res. 13 (2003) 2069–2081.
[5] J.W. Thomas, et al., Comparative analyses of multi-species sequences from targeted genomic regions, Nature 424 (2003) 788–793.
[6] E.H. Margulies, M. Blanchette, N.C.S. Program, D. Haussler, E.D. Green, Identification and characterization of multi-species conserved sequences, Genome Res. 13 (2003) 2507–2518.
[7] A. Siepel, D. Haussler, Combining phylogenetic and hidden Markov models in biosequence analysis, J. Comput. Biol. 11 (2004) 413–428.
[8] E.E. Storm, et al., Limb alterations in brachypodism mice due to mutations in a new member of the TGF beta-superfamily, Nature 368 (1994) 639–643.
[9] S.H. Settle Jr., et al., Multiple joint and skeletal patterning defects caused by single and double mutations in the mouse Gdf6 and Gdf5 genes, Dev. Biol. 254 (2003) 116–130.
[10] J.A. King, E.E. Storm, P.C. Marker, R.J. Dileone, D.M. Kingsley, The role of BMPs and GDFs in development of region-specific skeletal structures, Ann. N.Y. Acad. Sci. 785 (1996) 70–79.
[11] D.M. Kingsley, et al., The mouse short ear skeletal morphogenesis locus is associated with defects in a bone morphogenetic member of the TGF beta superfamily, Cell 71 (1992) 399–410.
[12] N. Jena, C. Martin-Seisdedos, P. McCue, C.M. Croce, BMP7 null

mutation in mice: developmental defects in skeleton, kidney, and eye, Exp. Cell Res. 230 (1997) 28–37.
[13] S. Settle, et al., The BMP family member Gdf7 is required for seminal vesicle growth, branching morphogenesis, and cytodifferentiation, Dev. Biol. 234 (2001) 138–150.
[14] R.J. DiLeone, L.B. Russell, D.M. Kingsley, An extensive 3′ regulatory region controls expression of Bmp5 in specific anatomical structures of the mouse embryo, Genetics 148 (1998) 401–408.
[15] R.J. DiLeone, G.A. Marcus, M.D. Johnson, D.M. Kingsley, Efficient studies of long-distance Bmp5 gene regulation using bacterial artificial chromosomes, Proc. Natl. Acad. Sci. USA 97 (2000) 1612–1617.
[16] D.P. Mortlock, M.E. Portnoy, R.L. Chandler, N.C.S. Program, E.D. Green, Comparative sequence analysis of the *Gdf6* locus reveals a duplicon-mediated chromosomal rearrangement in rodents and rapidly diverging coding and regulatory sequences, Genomics 84 (2004) 814–823.
[17] J.W. Thomas, et al., Parallel construction of orthologous sequence-ready clone contig maps in multiple species, Genome Res. 12 (2002) 1277–1285.
[18] W.J. Kent, et al., The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006.
[19] S. Schwartz, et al., PipMaker—A Web server for aligning two genomic DNA sequences, Genome Res. 10 (2000) 577–586.
[20] S. Schwartz, et al., MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences, Nucleic Acids Res. 31 (2003) 3518–3524.
[21] M.A. Nobrega, I. Ovcharenko, V. Afzal, E.M. Rubin, Scanning human gene deserts for long-range enhancers, Science 302 (2003) 413.
[22] A. Woolfe, et al., Highly conserved non-coding sequences are associated with vertebrate development, PLoS Biol. 3 (2004) e7.
[23] M. Rissi, J. Wittbrodt, E. Delot, M. Naegeli, F.M. Rosa, Zebrafish radar: a new member of the TGF-beta superfamily defines dorsal regions of the neural plate and the embryonic retina, Mech. Dev. 49 (1995) 223–234.
[24] C. Chang, A. Hemmati-Brivanlou, Xenopus GDF6, a new antagonist of noggin and a partner of BMPs, Development 126 (Suppl.) (1999) 3347–3357.
[25] K.L. Abrams, J. Xu, C. Nativelle-Serpentini, S. Dabirshahsahebi, M.B. Rogers, An evolutionary and molecular analysis of Bmp2 expression, J. Biol. Chem. 279 (2004) 15916–15928.
[26] W.F. Shen, S. Rozenfeld, H.J. Lawrence, C. Largman, The Abd-B-like Hox homeodomain proteins can be subdivided by the ability to form complexes with Pbx1a on a novel DNA target, J. Biol. Chem. 272 (1997) 8198–8206.
[27] L. Selleri, et al., Requirement for Pbx1 in skeletal patterning and programming chondrocyte proliferation and differentiation, Development 128 (2001) 3543–3557.
[28] K. Giese, A. Amsterdam, R. Grosschedl, DNA-binding properties of the HMG domain of the lymphoid-specific transcriptional regulator LEF-1, Genes Dev. 5 (1991) 2567–2578.
[29] M. van de Wetering, M. Oosterwegel, D. Dooijes, H. Clevers, Identification and cloning of TCF-1, a T lymphocyte-specific transcription factor containing a sequence-specific HMG box, EMBO J. 10 (1991) 123–132.
[30] X. He, A Wnt–Wnt situation, Dev. Cell 4 (2003) 791–797.
[31] R. DasGupta, E. Fuchs, Multiple roles for activated LEF/TCF transcription complexes during hair follicle development and differentiation, Development 126 (1999) 4557–4568.
[32] C. Hartmann, C.J. Tabin, Wnt-14 plays a pivotal role in inducing synovial joint formation in the developing appendicular skeleton, Cell 104 (2001) 341–351.
[33] X. Guo, et al., Wnt/{beta}-catenin signaling is sufficient and necessary for synovial joint formation, Genes Dev. 18 (2004) 2004–2417.
[34] R. Kothary, et al., Inducible expression of an hsp68–lacZ hybrid gene in transgenic mice, Development 105 (1989) 707–714.
[35] E.H. Margulies, et al., Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes, Proc. Natl. Acad. Sci. USA 102 (2005) 3354–3359.
[36] P.H. Francis-West, et al., Mechanisms of GDF-5 action during skeletal development, Development 126 (Suppl.) (1999) 1305–1315.
[37] R. Merino, et al., Expression and function of Gdf-5 during digit skeletogenesis in the embryonic chick leg bud, Dev. Biol. 206 (1999) 33–45.
[38] E.E. Storm, D.M. Kingsley, GDF5 coordinates bone and joint formation during digit development, Dev. Biol. 209 (1999) 11–27.
[39] K.A. Guss, C.E. Nelson, A. Hudson, M.E. Kraus, S.B. Carroll, Control of a genetic regulatory network by a selector gene, Science 292 (2001) 1164–1167.
[40] B. Gottgens, et al., Transcriptional regulation of the stem cell leukemia gene (SCL)—Comparative analysis of five vertebrate SCL loci, Genome Res. 12 (2002) 749–759.
[41] B. Gottgens, et al., Analysis of vertebrate SCL loci identifies conserved enhancers, Nat. Biotechnol. 18 (2000) 181–186 (Erratum appears in Nat. Biotechnol. 18 (2000) 1021).
[42] R.A. Gibbs, et al., Genome sequence of the Brown Norway rat yields insights into mammalian evolution, Nature 428 (2004) 493–521.
[43] E.S. Lander, et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[44] R.W. Blakesley, et al., An intermediate grade of finished genomic sequence suitable for comparative analyses, Genome Res. 14 (2004) 2235–2244.