Modular Organization of Phylogenetically Conserved Domains Controlling Developmental Regulation of the Human Skeletal Myosin Heavy Chain Gene Family*

Stephane Konig‡, James Burkman, Julie Fitzgerald, Marilyn Mitchell, Leonard Su, and Hansell Stedman§

From the Department of Surgery, University of Pennsylvania Medical System, Philadelphia, Pennsylvania 19104

Received for publication, April 2, 2002, and in revised form, April 22, 2002
Published, JBC Papers in Press, April 22, 2002, DOI 10.1074/jbc.M203162200

The mammalian skeletal myosin heavy chain locus is composed of a six-membered family of tandemly linked genes whose complex regulation plays a central role in striated muscle development and diversification. We have used publicly available genomic DNA sequences to provide a theoretical foundation for an experimental analysis of transcriptional regulation among the six promoters at this locus. After reconstruction of annotated drafts of the human and murine loci from fragmented DNA sequences, phylogenetic footprint analysis of each of the six promoters using standard and Bayesian alignment algorithms revealed unexpected patterns of DNA sequence conservation among orthologous and paralogous gene pairs. The conserved domains within 2.0 kilobases of each transcriptional start site are rich in putative muscle-specific transcription factor binding sites. Experiments based on plasmid transfection in vitro and electroporation in vivo validated several predictions of the bioinformatic analysis, yielding a picture of synergistic interaction between proximal and distal promoter elements in controlling developmental stage-specific gene activation. Of particular interest for future studies of heterologous gene expression is a 650-base pair construct containing modules from the proximal and distal human embryonic myosin heavy chain promoter that drives extraordinarily powerful transcription during muscle differentiation in vitro.

Throughout metazoan evolution, the range of molecular mechanisms used to achieve functional diversity in muscle partially reflects the anatomic complexity of the species under consideration. In vertebrates, the maximal contractile speed and energy utilization rate of individual muscle fibers is largely controlled by the structure of the motor protein myosin (reviewed in Ref. 1). The conventional myosins of striated muscle are hexameric proteins composed of paired trimers of myosin heavy chain (which contains the ATP-splitting motor domain) and one each of the essential and regulatory light chains. The human genome has at least 11 distinct striated myosin heavy chain (MyHC) genes (2), six of which are abundantly expressed in skeletal muscle and are tandemly linked at a single locus on chromosome 17 (3–5). The members of the human skeletal MyHC locus have 38 coding exons and two 5‘ noncoding exons. Unlike striated MyHC isofrom diversification in Drosophila, which reflects a complex pattern of alternative exon splicing (6, 7), the process in vertebrates relies on sequential activation of a family of distinct genes, each encoding a single MyHC (8). By virtue of the extraordinary abundance of these proteins (35% of muscle protein), a large body of experimental data on MyHC isoform switching was assimilated before isolation of the encoding genes (9).

During embryonic myogenesis in mammals, the transcriptionally 5‘-most gene in the skeletal MyHC locus is activated first. During fetal muscle development, the perinatal gene is selectively activated in a large proportion of the muscle cells with progressive down-regulation of the embryonic gene until birth. During subsequent postnatal growth, the perinatal MyHC is progressively replaced by the three “type II” (adult fast twitch) MyHCs. Throughout adult life, reversible transitions are possible in response to cycles of degeneration and regeneration or alterations in the hormonal environment, innervation, and pattern of locomotive recruitment of individual muscle fibers (1, 10). A sixth gene at the skeletal MyHC locus is expressed only in selected extraocular and laryngeal muscle fibers (11). The order of sequential activation during development contrasts with the physical order of the linked genes: 5‘-embryonic, IIa, IIx, IIb, perinatal, extraocular-3‘. A seventh MyHC abundantly expressed in type I (slow twitch) skeletal muscle fibers is encoded by the physically unlinked cardiac β MyHC. At all times, the stoichiometry of myofibril assembly must be maintained such that inactivation or down-regulation of one skeletal MyHC gene must be precisely offset by activation or up-regulation of another.

Two exceptional attributes of the mammalian skeletal MyHC locus, size and tandem gene number, have limited the study of transcriptional regulatory mechanisms within this gene family. Predating the modern era of general information about transcription factor binding sites, studies of a 1.4-kb region of genomic DNA upstream from the rat embryonic MyHC gene suggested a complex pattern of regulation by counteracting cis-elements (12). Studies of a large number of other single copy genes expressed during myogenesis have more recently defined roles and binding specificities for several tran-

* This work was supported in part by grants (to H. S.) from the Association Française contre les Myopathies (AFM), the Muscular Dystrophy Association of America, NIAMS (National Institutes of Health (NIH)), NINDS (NIH), and the Department of Veterans Affairs. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
‡ A postdoctoral fellow of the AFM.
§ To whom all correspondence should be addressed: Rm. 608, Biomedical Research Bldg. II/III, 421 Curie Blvd., Philadelphia, PA 19104-6160. Tel.: 215-898-1432; Fax: 215-573-8606; E-mail: hstedman@mail.med.upenn.edu.

1 The abbreviations used are: MyHC, myosin heavy chain; ANOVA, analysis of variance; contig, group of overlapping clones; ChemE, conserved human embryonic MyHC element; MRF, muscle regulatory factor; HSD, honestly significant difference.
scription factors, providing a basis for general models of promoter and enhancer function during commitment and terminal differentiation. Transcription factors of the MyoD/MRF and MEF-2 families play pivotal and synergistic roles in the activation and maintenance of the myogenic differentiation pathway (13, 14). Although the precise mechanisms by which these factors achieve transcriptional activation remain unclear, there is increasing evidence for a central role of histone modification and chromatin remodeling. In addition to MRF and MEF-2, numerous ubiquitous and muscle-specific transcription factors have been found to play important roles in gene up-regulation during terminal myogenesis. A partial list reflecting the extent of characterization includes AP1, AP2, GR, Oct-1, MEF-2, numerous ubiquitous and muscle-specific transcription factors, providing a basis for general models of promoter and enhancer function during commitment and terminal differentiation. 

In the present study, we use draft murine genome sequence from several fragmented draft files, initially identified on the basis of BLAST (25) comparisons with the human genome, to reconstruct a working model for the complete murine skeletal MyHC locus. Each of the six human MyHC promoter regions is analyzed in detail using bioinformatic tools, as exemplified by those cited above, to yield annotated maps reflecting both the level of sequence conservation and the precise location of matches to transcription factor matrices. Regions of strong conservation between the orthologous human and murine MyHC promoters occur in patches scattered throughout at least 2 kb immediately upstream of the transcriptional start sites. Paralogous promoter comparisons reveal surprisingly strong homology within 300 bp of the transcriptional start sites, providing evidence for a conserved proximal “core” promoter domain. Transfection of myotubes in vitro and electroporation of myofibers in vivo are used to further dissect the relevant domains of four of these promoters with the findings that 1) isolated proximal promoter regions can retain developmental stage specificity despite homology to a core consensus sequence and 2) an E box-rich distal region of the human embryonic MyHC promoter dramatically amplifies the in vitro activity of all of the proximal core promoters tested in cis. These findings suggest several testable hypotheses regarding potential short and long range interactions between transcriptional control elements within the skeletal MyHC tandem gene locus.

**EXPERIMENTAL PROCEDURES**

**Computer Analysis**—Deduced full-length cDNA sequences derived from an annotated version of the human skeletal myosin heavy chain locus (2) were used to query the high throughput genomic DNA sequence databases accessible through the National Center for Biotechnology Information Homepage (www.ncbi.nlm.nih.gov/BLAST/). Species-specific repeats in the reconstructed human locus were identified using the REPEATMASKER algorithm as implemented on the Washington University server (ftp.genome.washington.edu/cgi-bin/Repeat-Masker). As soon as draft sequences appeared online to confer complete coverage of all six of the genes at the murine locus, we ordered the draft sequence fragments to correspond to the order of the orthologous human sequences. A reconstructed murine sequence draft spanning the entire locus was annotated, initially by using GENSCAN as implemented on the MIT server (genes.mit.edu/GENSCAN) and later by cross-checking with the human annotated sequence to verify exon splice

---

**Table I**

| Primers | Sequence 5' → 3' | Plasmid |
|---------|------------------|--------|
| Embryonic promoter | GGGGGCATTGGCAACGGAGGTTG | KNE, KNS, KNM |
| PCR.ProE/Nhe1-Sens | GGGGAGATCCCTCGTCTCATACCTCTAGTTG | KLY, KNS, KLT, KMA |
| Embryonic cis-elements | TCTTCTCTCCTCTCTCTCAGGC | KOQ |
| PCR-EmbF13-Sens | KOQ |
| PCR-EmbB25-anti | GGGGGGATCTCGTCTCTCTGACCC | KOP |
| PCR-EmbB24-anti | GGGGGGATCTCGTCTCTCTGACCC | KOP |
| PCR-Emb1064/Kpn1-sens | GGGGGGATCTCGTCTCTCTGACCC | KOP |
| PCR-Emb1064/Kpn1-anti | GGGGGGATCTCGTCTCTCTGACCC | KOP |
| PCR-EmbE5/Kpn1-anti | GGGGGGATCTCGTCTCTCTGACCC | KOP |
| PCR-EmbE5S1-anti | GGGGGGATCTCGTCTCTCTGACCC | KOP |
| Site-directed mutagenesisa | GGGGGGATCTCGTCTCTCTGACCC | KOP |
| Mut-EmbE5-Sens | GGGGGGATCTCGTCTCTCTGACCC | KOP |
| Mut-EmbE6-Sens | GGGGGGATCTCGTCTCTCTGACCC | KOP |
| Mut-Emb1-SeqA-Sens (Y1) | GGGGGGATCTCGTCTCTCTGACCC | KOP |
| Mut-Emb.MEF2rev-sens | GGGGGGATCTCGTCTCTCTGACCC | KOP |

*a The introduced mutations are underlined. Only the forward primers are represented; the reverse primers used are the reverse and complement of the represented one.
Developmental Regulation of Human Myosin Gene Family

A

Mouse

Embryonic

Hybrid MyHC Loci

Human

Paralogous MyHC Gene Comparisons

B

Embryonic MyHC Genes

C

IIb MyHC Genes

D

Paralogous MyHC Gene Comparisons

Fig. 1. High resolution physical maps of the syntenic human and murine MyHC loci. A, size and order of the genes and intergenic regions showing slight relative expansion of the human locus. Exact

tections of the draft files used to reconstruct the murine locus are shown. When initially downloaded, each file was composed of 15 independent contigs in random order and orientation. B and C, raised
definition dot matrix analysis of the regions 8 kb upstream of the human and murine embryonic (B) and IIb (C) transcriptional start sites. Dot height ranges from a low of 40% to a high of 100% sequence similarity using the default DNA data base matrix within the MacVector software package. ChemE-1, -2, and -3 correspond to the conserved domains defined under “Results.” Open boxes at the bottom of the plot correspond to repeat domains as annotated in the Human Genome Project. D, standard dot matrix analyses covering 2 kb upstream from each of the following promoters, in the following order: IIa, IIb/x, IIb, perinatal. Not shown is the slightly lower homology to the embryonic promoter, truncated because of the computational matrix size limitations for this version of the MacVector program. There was no detectable homology with the extracocular promoter.

positions of the draft files used to reconstruct the murine locus are shown. When initially downloaded, each file was composed of 15 independent contigs in random order and orientation. B and C, raised
definition dot matrix analysis of the regions 8 kb upstream of the human and murine embryonic (B) and IIb (C) transcriptional start sites. Dot height ranges from a low of 40% to a high of 100% sequence similarity using the default DNA data base matrix within the MacVector software package. ChemE-1, -2, and -3 correspond to the conserved domains defined under “Results.” Open boxes at the bottom of the plot correspond to repeat domains as annotated in the Human Genome Project. D, standard dot matrix analyses covering 2 kb upstream from each of the following promoters, in the following order: IIa, IIb/x, IIb, perinatal. Not shown is the slightly lower homology to the embryonic promoter, truncated because of the computational matrix size limitations for this version of the MacVector program. There was no detectable homology with the extracocular promoter.

Results.

Plasmids—All of the promoter constructions were made in the pGL3-basin plasmid, a promoterless gene encoding the firefly luciferase (Promega). The different promoters were cloned using the NheI and BglII sites of the pGL3-basic plasmid. For this purpose, all of the promoters were amplified by 30 cycles of PCR using the proofreading Puco polymerase (Roche Molecular Biochemicals) according to the manufacturer’s protocol. All primers were designed to have a 58°C annealing temperature and to contain the NheI site in 5’ or the BglII site in 3’ as shown in Table I to allow a unidirectional cloning. The plasmid CMV-βGal (CLONTECH) was used for the quantification of transcription efficiency. The hybrid plasmids between the embryonic and the IIb promoters (KNX, KNF) were obtained by the ligation of two PCR fragments through a Sall site introduced into the primer (Table I). The KO insert was obtained by a recombinant PCR to join both independent PCR fragments.

Cell Culture and Transfection—The C2C12 cell line proliferates at low density in Dulbecco’s modified Eagle’s medium plus 20% fetal calf serum. The differentiation into myotubes was induced by replacing the proliferation medium by Dulbecco’s modified Eagle’s medium + 0.5% fetal calf serum + ITS (insulin, transferrin, and selenium from Roche Molecular Biochemicals at 5 μg/ml each). These cells were transfected using Fugene6 (Roche Molecular Biochemicals). Briefly, at day 1, myoblasts were plated at 200,000 cells/well in a six-well plate. At day 2, 2 h prior to the transfection, the medium was removed and replaced by differentiation medium. Transfection was achieved by incubating 6 μl of FuGene6 in 94 μl of Dulbecco’s modified Eagle’s medium at room temperature. After 5 min, 2 μg of the luciferase reporter plasmid and 1 μg of CMV-βGal plasmid were then incubated for another 15 min. The transfection mixture (100 μl) was added to the cell drop by drop. Cells were collected at day 5 for luciferase assay using 200 μl of 1× lysis buffer per well according to the luciferase assay system (Promega). We used 10 μl of cell extracts for the determination of the luciferase activity. The β-galactosidase activity was measured using the chemiluminescent reporter gene assay system for the detection of the β-galactosidase (Galacthon-Light, Tropix PE biosystems) with 10 μl of cell extracts.

Rat in Vivo Electroporation—Electroporation of rat tibialis anterior muscles was performed as per Ref. 30. All electroporation and subsequent euthanizing and procurement procedures were performed in ac-
liquid nitrogen frozen muscles were pulverized to powder and solubilized with $\frac{1}{1000}$ lysis buffer (luciferase assay system; Promega). After a 5-min incubation and one freeze/thaw cycle, samples were centrifuged, and 20 μl of supernatant were processed for luciferase and β-galactosidase assays.

**Normalization of Reporter Gene Activity and Statistical Analysis**—The luciferase results are expressed as the ratio between the luciferase activity and the β-galactosidase activity. Statistical analysis was performed using a two-tailed Student's t-test. A p-value of less than 0.05 was considered statistically significant.
**Fig. 3. ClustalW DNA sequence alignments for five pairs of orthologous human and murine promoters.** Labeling of individual promoters appears at the extreme left. Larger open boxes circumscribe regions predicted to contain three sets of MEF-2 binding sites. The TATA boxes in the bottom row contribute the most proximal of the alignment probability peaks in the phylogenetic footprints for the distantly related promoters, as shown in Fig. 2. An E box (CAACCTG) in the middle of the most proximal MEF-2 site for the IIb promoters is outlined. The three-base departure from the consensus sequence in this domain is predicted to disrupt the MEF-2 site in the IIb promoters. A box encompassing the IIb sequences partially offset to the left of the most distal MEF-2 consensus domain shows the preferred site for MEF-2 binding of the IIb promoters.
activity and the β-galactosidase activity. Furthermore, for each series of transfections, we transfected the promoterless plasmid pGL3-basic for the quantification of background. The normalized luciferase results obtained with the pGL3-basic plasmid are defined as 1. All of the other activities of the series are then expressed as the fold induction over background. Data sets on levels of marker gene expression were analyzed (JMP, SAS Institute, Inc., Cary, NC) to determine arithmetic means, S.D. values, S.E. values, S.E. values of the means, and the probabilities associated with one-way analysis of variance (ANOVA), Student’s t test, and the Tukey-Kramer HSD test (i.e. the probability of incorrectly rejecting the null hypothesis about the difference between means among multiple groups within a data set). The Tukey-Kramer HSD provides a conservative calculation of statistical significance in the analysis of intergroup comparisons that minimizes the risk of type I error by increasing the quantile multiplied into the S.E. values to create the least significant difference.

RESULTS

To provide a basis for phylogenetic analysis of putative transcriptional control domains within the skeletal MyHC locus, we periodically screened high throughput DNA sequence databases until a representation of the entire murine locus could be identified in draft form. Three overlapping bacterial artificial chromosomes spanned the entire locus as depicted in Fig. 1A, although the data contained in the relevant accession files was highly fragmented at the time of this analysis. Based on synteny with the human locus, we reconstructed a single file that contained a properly ordered and annotated draft of the murine locus. As expected, coding regions exhibit greater than 90% sequence homology, without gaps. Although the orthologous genes occupy approximately the same relative proportion of the length of the entire locus, the murine locus is only 80% of the size of the human locus. From the standpoint of transcriptional regulation, the regions upstream of the first coding exon of the orthologous and paralogous genes are of greatest interest. We identified transcriptional start sites by comparison with full-length cDNA sequences or by analyzing homologous intergenic regions for TBP binding site and initiator consensus sequences using Promoter Prediction by Neural Network NNPP (Lawrence Berkeley National Laboratory).

An initial screen of 8 kb upstream of each transcriptional start site revealed a pattern exemplified by the orthologous embryonic and Iib promoters in which there is strong conservation interrupted by blocks of nonconserved sequence, which in most cases precisely coincide with the boundaries of species-specific repeats as annotated previously using the REPEATERMASKER algorithm. There are blocks of >100-bp DNA sequence that attain 60% conservation (note peak height in the raised relief image) out to at least 8 kb upstream of both the human embryonic and Iib MyHC transcriptional start sites (Fig. 1, B and C; embryonic MyHC upstream 8 × 8 kb and Iib MyHC upstream 8 × 8 kb). The conserved regions identified as conserved human embryonic MyHC element (ChemE)-1, -2, and -3 are defined below. With the sole exception of the extracutaneous MyHC promoter pair, the pattern of conservation within 2 kb of the start sites provided striking evidence for selection against mutation by insertion or deletion. This pattern suggests that the order of and/or the physical distance between transcription factor binding sites spread over at least 2 kb is crucial to the proper developmental regulation of MyHC gene activation repression. To address the possibility of a shared core promoter structure, we extended the pairwise comparison to include paralogous human MyHC promoters (Fig. 1D). Homology among promoter paralogs is uniquely detected within the most proximal 300 bp, as shown for the IIA, IId/x, Iib, and perinatal genes.

A limitation of this analysis for the identification of transcriptional regulatory elements is its dependence on user-defined parameters such as window size, DNA scoring matrix, and gap penalty. These biases have been largely eliminated through the recent introduction of a Bayesian statistical method based on a Gibbs sampling algorithm (24, 26). The relationship between level of sequence conservation and distance upstream from the transcriptional start sites is quantitatively displayed in Bayesian "phylogenetic footprint" plots for five of the promoter pairs (Fig. 2). The z axis in each of the three-dimensional plots indicates the local probability of mouse-human DNA sequence alignment based on a modification of the Bayes block aligner algorithm (24). The overall pattern of gene conservation as a function of distance from the transcriptional start site is similar for the perinatal, IIA, IId/x, and Iib genes. Relative to these four genes, the embryonic gene has a comparatively larger block of conserved sequence in the interval −600 to −1000 (ChemE-2) and a smaller block of conserved sequence in the interval +1 to −300 (ChemE-1). In addition, 4.1 kb upstream of the human embryonic gene transcriptional start site is a 0.5-kb block of extraordinarily conserved noncoding sequence (ChemE-3), a pattern unique among the six genes at this locus. Finally, sequence conservation among the paralogous genes is typified by the human IIA versus IId/x pairing, with four distinct spikes of homology within 300 bp of the transcriptional start.

A position weight matrix-based search for binding sites in a transcription factor data base was filtered to identify candidate protein-DNA interactions in the phylogenetically conserved portion of each sequence pair, as shown schematically above the footprints (Fig. 2, lettering above individual plots). Three groups of transcription factors known to participate in the regulation of muscle-specific gene expression are identified as having high concentrations of sites within these conserved domains: MRF, MEF-2, and SRF. Additional transcription factors identified by this approach include AP-2, AhR, CCAAT enhancer-binding protein, GR, HEB, Oct-1, RREB-1, Sp1, STAT4, TEF-1, and YY1. As previously demonstrated for other muscle-specific promoters, the concentration of predicted sites is higher in the conserved than in the divergent sequence domains (24). The overall distribution of these sites is not appreciably conserved among the paralogous promoters except...
in the most proximal 300 bp of sequence.

We used ClustalW (27) to align all of the proximal promoters for each orthologous gene pair (Fig. 3). This alignment draws attention to four core domains and provides a consensus sequence for these and the intervening regions. A computational analysis of each of the individual sequences and the consensus sequence identifies several "high affinity" matches to the Transfac matrices for MEF-2, MyoD, Oct-1, CCAAT enhancer-binding protein, and TBP, the former of which is outlined on the sequence alignment. In this alignment, the embryonic promoter pair is distinguished from the other four pairs in having the highest affinity MEF-2 site (28) nearest to the transcriptional start site (proximal domain 2 as compared with proximal domain 3 for the others). In fact, this site achieves the highest binding score of any predicted MEF-2 site within 2 kb of an MyHC promoter. Three MEF-2 consensus sequences are identified in the IIa, IIb/x, and perinatal promoters. MEF-2 binding is predicted from position weight matrices for the corresponding regions of the other two promoters, but the binding strength falls below the default limits at the most distal site for the embryonic promoter and at the most proximal site for the IIb promoter.

Recapitulating the bioinformatic analysis, five of the six tandemly linked skeletal MyHC promoters exhibit similar overall structures in which focally conserved sequences, rich in putative muscle-specific transcription factor binding sites, are distributed throughout 1.5–2 kilobases upstream of a broadly conserved core promoter domain of 250–300 bp. We chose a convenient nomenclature to describe the three larger blocks of conserved sequence in the embryonic promoter: conserved human embryonic MyHC elements 1, 2, and 3 (ChemE-1, -2, and -3). These observations provide a basis for interpreting the results of an experimental dissection of proximal and distal promoter elements in vitro and in vivo. As an initial screen for transcriptional activity in vitro, we used large constructs based on cosmid-cloned fragments of three human skeletal MyHC genes (Fig. 4). As expected, the SV40 promoter drove high level marker gene expression in C2C12 myoblasts and myotubes (612- and 146-fold increases over background, respectively). Cis-acting sequences spanning at least 4 kb 5′ of the embryonic, IIb/x, and IIb MyHC gene translational start sites drove marker gene expression in myoblasts at levels barely above background. In myotubes, however, the embryonic MyHC gene construct uniquely increased marker gene expression to a level 160-fold higher than background, exceeding even that of the SV40 control. In contrast, marker gene expression was the same in myoblasts and myotubes for the IIb/x and IIb MyHC gene constructs (i.e. just above background). The surprisingly
high marker gene expression from the embryonic construct suggested that the 30-kb fragment encompassed most of the elements necessary for developmental stage-specific transcriptional activation. The relative inactivity of the adult MyHC promoters in this assay recapitulates the activities of the endogenous genes (29) and may reflect the absence of developmental stage-specific activators, the presence of repressors, or a combination thereof.

To address the possibility that repressive elements distal to the broadly conserved core domains were responsible for the inactivation of the adult MyHC promoters, we assayed marker gene expression in C2C12 myoblasts and myotubes using plasmids containing the proximal 300 bp of the embryonic (ChemE-1), perinatal, IId/x, and IIb promoters. The results of this series of experiments (Fig. 5) again showed strong activation in myotubes of the proximal 300-bp portion of the embryonic MyHC promoter and very low (near background) activity of the non-embryonic promoters. Further truncation of the embryonic promoter resulted in an abrupt loss of activity between –231 and –174 bp. The –174 bp construct is predicted to retain the high affinity MEF-2 site between –140 and –170, but lose a predicted YY1 site between –174 and –200 and a STAT4 site between –210 and –230. There is an apparent 15-fold reduction in the activity of the –300 bp embryonic MyHC promoter construct relative to the 30-kb construct. These results suggest that the extraordinary strength of the intact embryonic MyHC promoter reflects the contribution of cis-acting elements both within and beyond the most proximal 300 bp and that the relative inactivity in myotubes of the larger nonembryonic MyHC promoter constructs at least partially reflects the structure of the most proximal 300 bp.

To further localize the critical elements in transcriptional control of the embryonic MyHC gene, we compared the activities of five additional truncation constructs in vitro. As shown in Fig. 6, there is progressive and dramatic loss of activity with truncation below –2 kb to –500 bp and –230 bp. The 2-kb construct has the highest marker gene expression in this assay, slightly exceeding that for both the 6.2-kb fragment ending at the transcriptional start site (the only construct containing ChemE-3) and the 4-kb fragment ending at the translational start site (i.e., including introns 1 and 2). These results suggest that putative transcription factor binding sites identified in phylogenetically conserved sequences between –230 bp and –2 kb (within and surrounding ChemE-2) greatly contribute to the overall activity of the endogenous promoter during terminal differentiation. Several candidates are shown in Fig. 2, notably the cluster of E boxes between –700 and –1500. The E boxes have potentially important differences based on both their se-
sequence similarity to consensus sequences identified in other promoters and on their proximity to other sites identified through our bioinformatic analysis. The most proximal two of the six E boxes in this interval are predicted to bind the myogenic helix-loop-helix factors with highest affinity, whereas the one at −860 is adjacent to a predicted MEF-2 site.

To study the role of this phylogenetically conserved region in the muscle-specific activation of this promoter, we constructed a series of internal deletion mutants as shown in Fig. 7. Marker gene expression in a series of C2C12 transfections was only slightly reduced when regions distal to −1.5 kb and between −230 and −700 bp were deleted. The KLX construct retains all six of the E boxes cited above, which we arbitrarily labeled E1 through E6 in order from distal to proximal. When the 800-bp fragment containing these E boxes was further divided into subfragments containing E1–E4 and E3–E6, it was possible to identify E3–E6 as the most important of these cis-acting regions for transcription in C2C12 myotubes. This 512-bp subfragment (−699 to −1211) wholly contains and experimentally defines the boundaries of ChemE-2.

To further characterize the pattern of transcriptional activation from the ChemE-2 subfragment, we tested its ability to synergize with heterologous minimal promoters (Fig. 8). The SV40 minimal promoter, activated in cis by a wide range of enhancers, was minimally affected by juxtaposition to the ChemE-2 subfragment in C2C12 myotube transfections. In contrast, all of the MyHC proximal promoters previously tested were further activated by ChemE-2. The relative magnitude of this cis-activation was stronger for the IId/x and perinatal promoters than for the embryonic promoter (37.3-, 31.8-, and 14.1-fold, respectively), and was similar for the IIb (12.3-fold) and embryonic promoters. Nevertheless, the absolute activity of the ChemE-2 subfragment fused to the embryonic minimal promoter was highest by almost 10-fold in this assay. We interpret this result to indicate that the overall developmental stage- and fiber type-specific activity of each of the endogenous MyHC promoters reflects the input of multiple DNA-transcription factor interactions in both the highly conserved proximal promoter and the isogene-specific distal promoter regions. Strong activation by distal elements can partially override weak activation or repression from more proximal elements.

To further dissect the proximal promoter regions, we took advantage of the low level homology between the embryonic and IIb promoters in the construction of chimeric reporter plasmids (Fig. 9). The ChemE-2 subfragment was used in juxtaposition to a series of progressively truncated embryonic proximal promoter fragments with the finding that these serial deletions were associated with the progressive loss of promoter activity. When the deleted regions were replaced by the homologous domains from the IIb promoter, there was neither further
ther loss nor compensatory restoration of activity. The absence of developmental stage-specific transcriptional repression associated with the IIb for embryonic chimeric substitutions suggests the complete absence of inhibitory DNA-protein interactions in this region. In other words, relative to the embryonic deletion constructs, insertional replacement with IIb sequences had essentially no effect on activity, such that the IIb DNA served essentially as neutral spacer sequence in the context of the C2C12 transcriptional milieu. The noninterchangeability of homologous domains from the IIb and embryonic promoters reveals the potential complexity of activating DNA-protein and protein-protein interactions in the proximal promoters.

In vivo, the adult isoforms are highly expressed in the fast twitch muscles of the rodent hind limb. As a simple test of the activity of these promoter constructs in vivo, we used an electroporation protocol to partially augment the efficiency of somatic gene transfer. While less efficient than the use of recombinant vectors that allow binding of the myocyte surface by viral capsid proteins, the plasmid electroporation approach avoids the complicating use of viral DNA sequences in cis with the MyHC promoter sequences. The results of this series of experiments are shown in Fig. 10. We tested the embryonic and IIb constructs with and without the ChemE-2 subfragment at 4 and 7 days, and without ChemE-2 at 21 days with the expectations that (a) localized regeneration after electrical injury (31) would initially activate both of the endogenous genes (embryonic and IIb), (b) activation of the endogenous embryonic MyHC gene would be transient as regeneration subsided, and (c) augmentation from the ChemE-2 subfragment might be required to achieve detectable signal above background. This experiment showed that the embryonic and IIb core promoters were comparably active in vivo and that they were both up-regulated by the ChemE-2 subfragment (although to a lesser extent than seen in vitro). The comparative activity of the IIb promoter constructs at 4, 7, and 21 days probably reflects the role of trans-acting factors present in vivo but absent from the C2C12 system in vitro. The trends observed also suggest that any repression attributable to elements within the ChemE-2 subfragment is outweighed by activation from elements within the IIb core domain. At 21 days, however, the embryonic and IIb core promoters showed minimal and maximal activity, respectively, a result that contrasts sharply with the initial relative activity of these constructs in vitro but parallels that of the endogenous genes in vivo. Thus, the core promoter constructs qualitatively recapitulate the expression patterns of the cognate genes, providing tools for use in a detailed dissection of the comparative roles of trans-activation and trans-repression during development and regeneration.

DISCUSSION

Despite the critical importance of the sarcomeric MyHC in skeletal muscle function and the extraordinary abundance of

FIG. 8. The ChemE-2 sequence cis-activates heterologous MyHC proximal promoters in myotubes in vitro. Exceptionally high levels of expression are only obtained with the embryonic proximal promoter in the presence of the ChemE-2 element, yielding the only statistically significant ($p < 0.05$) results in an all pairs comparison using the Tukey-Kramer HSD test. The relative levels of expression observed between the IIb, IIx, and perinatal proximal promoters in the presence of the ChemE-2 are not statistically different from one another ($p > 0.05$), although there is a consistent upward trend when each of these is associated with the ChemE-2 (ratios shown to the right of paired bars on the graph). C2C12 cell extracts were prepared 72 h after induction of differentiation. Cell extracts were assayed for luciferase and β-galactosidase as described under “Experimental Procedures.” The results are expressed as means ± S.E. Normalized results are expressed as a $\times$-fold induction over the activity of the pGL3-basic promoter (defined as 1). By one-way ANOVA, $F = 9.78$ and $p < 0.0001$. 
this protein, the size and complexity of the human skeletal MyHC locus has delayed detailed investigation of the transcriptional control of its six tandemly linked genes. During the preparation of this manuscript, we became aware of a report on some recent in vitro studies using plasmid constructs based on the murine IIa, IId/x, and IIb proximal promoters (32). Our data on early developmental isoform switching complement these studies (which instead focused on possible mechanisms for fiber-type regulation of adult MyHC isoform switching). We have used newly available draft DNA sequence for the human and murine genomes as a basis for modeling important transcription factor binding sites in the cis-acting regulatory sequences. Analysis of paralogous skeletal MyHC genes reveals striking homology in the most proximal 300 bp of the embryonic, IIa, IId/x, IIb, and perinatal promoters, centered around three predicted MEF-2 sites. Our previous analysis of the evolutionary relationships among these genes, based wholly on comparisons of coding sequence for the “molecular yardstick-like” rod domain of the protein, suggests that this MEF-2 site rich “core” structure to the promoter dates back to a common ancestral gene that existed during early vertebrate evolution, before the human avian species split (2). Consistent with this prediction is the result of a phylogenetic footprint analysis of the human IId/x promoter with that from the most closely related chicken MyHC gene (Fig. 2H). The high probability regions match the core consensus at all three of the MEF-2 sites identified in the IIa, IId/x, and perinatal promoters. In contrast, neither the embryonic nor the IId/x promoters contain E boxes within 300 bp of the transcriptional start sites, and CANNTG consensus sequences identified in the other promoters are located at sporadic positions (IIa −52; IIb −61 and −215; perinatal −80 and −118; chicken “J02714” −266).

Our experimental findings with the set of core promoter constructs revealed that reporter expression roughly paralleled the relative levels of the corresponding isomyosins in myotubes in vitro (29) and in myofibers in vivo (33). In a detailed analysis of chimeric substitutions, none of the IIb sequence motifs identified in the sequence alignments could substitute for the ho-
mologous portions of the embryonic promoter without a significant loss of activity in vitro. This implies that the divergent sequences intervening between the consensus-matching portions of the core contribute to the process of promoter switching during development. Some of the divergent sequence falls within the three MEF-2 sites predicted in this region, raising the possibility that subtle differences in MEF-2 binding affinity could play a role in this process.

Mutation of the E box at −61 bp in a series of murine Iib promoter constructs has been shown to eliminate MyoD binding at this site in vitro and to inactivate transcription from these constructs in vivo (34). Our observation that the proximal embryonic promoter, in the complete absence of E box consensus CANNTG sites, drives much stronger marker gene expression than the Iib promoter in C2C12 cells supports the possibility that MRFs have acquired a modulatory role in the proximal promoters. Evidence from a wide range of studies has implicated tripartate interactions involving DNA, MEF-2, and additional transcription factors in the control of other muscle-specific promoters during terminal differentiation (35). To the extent that the proximal promoters control the timing of skeletal MyHC gene activation and repression, our bioinformatic analysis suggests that the underlying control mechanisms evolved around an ancestral “footprint” with three members of the MEF-2 family recruiting other transcription factors into multisubunit complexes. In the embryonic promoter, the most proximal predicted MEF-2 site uniquely overlaps with a predicted SRF site, raising the possibility that interaction or competition between these transcription factors plays a role in developmental regulation. Documented competition between SRF and YY1 at CArG boxes within the skeletal (36) and cardiac (37) α-actin promoters provide precedents for regulation by this mechanism.

The sequences between −300 and −2 kb upstream of each of the transcriptional start sites for the five human genes listed above reveal a multifocal pattern of conserved sequence motifs and evidence for strong selection against insertions or deletions (indels). This suggests that intermediate range positional relationships among groups of transcription factor binding sites has been preserved by functional constraint. Our finding of an E box-rich region devoid of high affinity MEF-2 sites (ChemE-2) that dramatically increased the expression of the
core proximal promoter suggests that MRF and MEF-2 transcription factors, respectively, recruit other factors to physically distinct regions of the promoter to set the stage for synergistic interaction across a distance of several hundred base pairs. Despite evidence for evolutionary selection against indels between these two regions of the promoter, our deletional constructs preserve the full activity of the unaltered gene. The ability of the ChemE-2 subfragment to cis-activate the nonembryonic proximal promoters raises the possibility that this element functions in the native locus as either an enhancer or, by analogy to the β-globin locus, as a locus control region (38, 39). In either case, our in vivo studies provide preliminary evidence that this region lacks the ability to repress the nonembryonic core promoters during postembryonic stages of development. Our largest embryonic gene constructs encompassed an even more distal upstream region (ChemE-3) that attracted considerable interest because of the inordinate size of the conserved sequence domain and the local prevalence of muscle-specific transcription factor binding sites. In the C2C12 system, the transcriptional activity of the embryonic constructs was maintained despite deletion of this conserved block of sequence. Further studies will address a range of possible roles for both of these putative regulatory elements.

The skeletal myosin heavy chains are among the most abundant proteins in the bodies of vertebrates, and they must accumulate with other contractile proteins in precisely controlled stoichiometric ratios. This gene dosage effect is best documented in Drosophila, where compensatory changes in actin number improve muscle structure by restoring proper stoichiometric ratios. This gene dosage effect is best documented in vertebrates, and they must achieve a set of competitive values of the total MyHC mRNA mix without disrupting overall protein stoichiometry. Alternatively, if all of the genes are autonomously regulated by their own upstream regulatory sequences, down-regulation of one gene must be precisely offset by up-regulation of another. Although it is premature to suggest detailed models for long range interaction between transcription factors binding the endogenous MyHC genes, two candidate control regions are identified in the current studies (ChemE-2 and -3). One of these is incorporated in the only minimal length skeletal MyHC promoter construct to show robust autonomous activity in vitro, perhaps providing a cassette powerful enough for functional expression of recombinant MyHCs. Further studies will establish whether the other promoter regions identified at this locus can autonomously drive heterologous protein expression in adult tissues at levels sufficient for assembly of the contractile apparatus, a sine qua non for isolation of the entire control region.

In summary, the foregoing studies illustrate the complementarity of readily available tools for in vitro, in vivo, and in silico analysis, as applied to the dissection of the transcriptional regulatory circuitry controlling the developmental program for the six-membered human skeletal MyHC gene family.

REFERENCES

1. Schiaffino, S., and Reggiani, C. (1996) Physiol. Rev. 76, 371–422
2. Desjardins, P., Burkhman, J., Shragger, J., Allmond, L., and Stedman, H. (2002) Mol. Biol. Evol. 19, 375–383
3. Leinwand, L. A., Fournier, R. E. K., Nadal-Ginard, B., and Shows, T. (1983) Science 221, 766–768
4. Yoon, S. J., Seiler, S., Kacherlapati, R., and Leinwand, L. (1992) Proc. Natl. Acad. Sci. U. S. A. 89, 12078–12082
5. Shragger, J., Desjardins, P., Burkhman, J., Konig, S., Stewart, D., Su, L., Shah, M., Bricklin, E., Tewari, M., Hoffman, R., Ricke, M., Julienne, E., Rubinstein, N., and Stedman, H. (2000) J. Muscle Res. Cell Motil. 21, 345–355
6. Hastings, G. A., and Emerson, C. P. J. (1991) J. Cell Biol. 114, 263–276
7. Bernstein, S., and Milligan, R. (1997) J. Mol. Biol. 271, 1–6
8. Mahdavi, V., Strehler, E. E., Periasamy, M., Wieczorek, D., Izumo, S., Grund, S., Strehler, M. A., and Nadal-Ginard, B. (1986) in Molecular Biology of Muscle Development (Emerson, C. F. D., Nadal-Ginard, B., and Siddique, M. A., eds) pp. 345–361, Alan R. Liss, New York
9. Whalen, R., Sell, S., Butler-Browne, G., Schwartz, K., Bouveret, P., and Pinsent-Harston, I. (1981) Nature 292, 805–809
10. Pette, D., and Staron, R. (1997) Int. Rev. Cytol. 170, 143–223
11. Wieczorek, D., Periasamy, M., Butler-Browne, G., Whalen, R., and Nadal-Ginard, B. (1985) J. Cell Biol. 101, 618–629
12. Bouvagnet, P. P., Strehler, P. E., White, G. E., Strehler-Page, M. A., Nadal-Ginard, B., and Mahdavi, V. (1987) Mol. Cell. Biol. 7, 4377–4389
13. Weintraub, H., Davis, R., Tassott, S., Thayer, M., Krause, M., Benezra, R., Blackwell, T. K., Turner, D., Rupp, R., and Hellenburg, S. (1991) Science 251, 761–766
14. McKinsey, T. A., Zhang, C. L., and Olson, E. N. (2001) Curr. Opin. Genet. Dev. 11, 497–504
15. Andreucci, J. J., Grant, D., Cox, D. M., Tomc, L. K., Pyeves, R., Goldhamer, D. J., Rodrigues, N., Bedard, P. A., and McDermott, J. C. (2002) J. Biol. Chem. 277, 16426–16432
16. Perkins, K. J., Burton, E. A., and Davies, K. E. (2001) Nucleic Acids Res. 29, 4843–4850
17. Sun, X., Fischer, D. R., Pritts, T. A., Wray, C. J., and Hasselgren, P. O. (2002) Am. J. Physiol. Regul. Integr. Comp. Physiol. 282, E509–E518
18. Santalucia, T., Moreno, H., Palacin, M., Yacob, M. H., Brand, N. J., and Zorzano, A. (2001) J. Mol. Biol. 314, 195–204
19. Gupta, M., Kogut, P., Davis, F. J., Belaguli, N. S., Schwartz, R. J., and Gupta, M. P. (2001) J. Biol. Chem. 276, 10413–10422
20. Lakich, M. M., Diagana, T. T., North, D. L., and Whalen, R. G. (1998) J. Biol. Chem. 273, 15217–15226
21. McGrew, M. J., Bogdanova, N., Hasegawa, K., Hughes, S. H., Kitis, R. N., and Rosenthal, N. (1996) Mol. Cell. Biol. 16, 4524–4534
22. Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., and Schacherer, F. (2000) Nucleic Acids Res. 28, 316–319
23. Schwartz, S., Zhang, Z., Frazer, K. A., Smitt, A., Rieder, C., Bouch, J., Gibbs, R., Hardison, R., and Miller, W. (2000) Genome Res. 10, 577–586
24. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W., and Lawrence, C. E. (2000) Nat. Genet. 22, 225–228
25. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) J. Mol. Biol. 215, 403–410
26. Zhu, J., Liu, J. S., and Lawrence, C. E. (1998) Bioinformatics 14, 25–39
27. Tindel, J. D., Higgins, D. G., and Gibson, T. J. (1994) Nucleic Acids Res. 22, 4673–4680
28. Fickett, J. W. (1996) Mol. Cell. Biol. 16, 437–441
29. Naumann, K., and Pette, D. (1994) Differentiation 55, 203–211
30. Aihara, H., and Miyazaki, J. (1998) Nat. Biotechnol. 16, 867–870
31. Bier, M., Hammer, S. M., Canaday, D. J., and Lee, R. C. (1999) Bioelectromagnetics 20, 194–201
32. Allen, D. L., Sartoris, C. A., Sacuro, L. K., and Leinwand, L. A. (2001) J. Biol. Chem. 276, 43524–43533
33. Huey, K. A., and Bodine, S. C. (1998) Eur. J. Biochem. 256, 45–50
34. Wheeler, M. T., Snyder, E. C., Varnum, M. N., and Swoap, S. J. (1999) J. Mol. Biol. 282, 10422–10433
35. Molkentin, J. D., Black, B. L., Martin, J. F., and Olson, E. N. (1995) Cell 83, 1125–1136
36. Martin, A. A., Gualberto, A., Kalman, M. F., Lowry, J., and Walsh, K. (1997) DNA Cell Biol. 16, 653–661
37. Chen, C. Y., and Schwartz, R. J. (1997) Mol. Endocrinol. 11, 812–822
38. Grosveld, F., van Assendelft, G., Groves, D., and Kollia, G. (1987) Cell 51, 975–985
39. Schubeler, D., Groudine, M., and Bender, M. A. (2001) Proc. Natl. Acad. Sci. U. S. A. 98, 11432–11437
40. Beall, C., Sepanski, M., and Fyrberg, E. (1989) Genes Dev. 3, 131–140