Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Genomics 87 (2006) 733–746
GENOMICS
www.elsevier.com/locate/ygeno

The human desmin locus: Gene organization and LCR-mediated transcriptional control

Jennifer L.Y. Tam [a], Kostas Triantaphyllopoulos [a,b], Helen Todd [a], Selina Raguz [a,1], Ton de Wit [c], Jennifer E. Morgan [d], Terence A. Partridge [d], Eleni Makrinou [a], Frank Grosveld [c], Michael Antoniou [a,*]

[a] *Nuclear Biology Group, Department of Medical and Molecular Genetics, King's College London School of Medicine, King's College London–Guy's Campus, 8th Floor Guy's Tower, Guy's Hospital, London SE1 9RT, UK*

[b] *Cobbold Labs, Department of Medicine, University College London, 7th Floor Jules Thorn Institute, Middlesex Hospital, Mortimer Street, London W1T 3AA, UK*

[c] *Department of Cell Biology, Erasmus Medical Center, 3000 DR Rotterdam, The Netherlands*

[d] *Muscle Cell Biology Group, MRC Clinical Sciences Centre, Imperial College of Science, Technology and Medicine, Hammersmith Hospital, Du Cane Road, London W12 ONN, UK*

Received 23 August 2005; accepted 29 January 2006
Available online 20 March 2006

Abstract

Locus control regions (LCRs) are defined by their ability to confer reproducible physiological levels of transgene expression in mice and therefore thought to possess the ability to generate dominantly a transcriptionally active chromatin structure. We report the first characterization of a muscle-cell-specific LCR, which is linked to the human desmin gene (*DES*). The *DES* LCR consists of five regions of muscle-specific DNase I hypersensitivity (HS) localized between −9 and −18 kb 5′ of *DES* and reproducibly drives full physiological levels of expression in all muscle cell types. The *DES* LCR DNase I HS regions are highly conserved between humans and other mammals and can potentially bind a broad range of muscle-specific and ubiquitous transcription factors. Bioinformatics and direct molecular analysis show that the *DES* locus consists of three muscle-specific (*DES*) or muscle preferentially expressed genes (*APEG1* and SPEG, the human orthologue of murine striated-muscle-specific serine/threonine protein kinase, Speg). The *DES* LCR may therefore regulate expression of SPEG and *APEG1* as well as *DES*.

Keywords: Locus control region; Desmin; Muscle-specific gene expression; Chromatin domains

An analysis of the complete genome sequence of numerous eukaryotes has revealed that chromosomal gene order in these organisms is not random; genes that have similar functions or whose expression needs to be tightly coordinated are generally clustered [1]. This observation adds support to the concept that genes in higher eukaryotes are organized in the form of functional domains [2] or functional expression modules [3], with each domain possessing genetic control elements or undergoing processes that are able to establish and maintain a transcriptionally competent chromatin structure that allows tissue-restricted or ubiquitous gene expression. A major challenge of the postgenomic era is the elucidation of the mechanisms by which a transcriptionally competent open chromatin domain is established and how spatial and temporal patterns of expression are coordinated from within arrays of genes. At present there are only two known classes of transcriptional regulatory elements that appear to be part of dominant chromatin remodeling processes and that may also be involved in coordinating gene expression, namely locus control regions (LCRs) [4] and CpG islands associated with housekeeping genes [5].

LCRs are functionally defined by their ability to confer on a transgene linked *in cis* reproducible physiological levels of expression regardless of integration site [4]. The involvement of LCRs in processes that establish a dominant, transcriptionally competent open chromatin structure at ectopic transgene integration sites is most clearly demonstrated by their ability

* Corresponding author. Fax: +44 207 188 2585.
E-mail address: michael.antoniou@genetics.kcl.ac.uk (M. Antoniou).
[1] Current address: Membrane Transport Biology Group, MRC Clinical Sciences Centre, Imperial College of Science, Technology and Medicine, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK.

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.01.009

to drive physiological and stable levels of expression even from within or near centromeres, showing they can obstruct the spread of heterochromatin and prevent progressive silencing that leads to a variegated expression pattern [6]. LCRs appear to participate in two activities in the regulation of gene expression: the establishment of a transcriptionally competent chromatin structure and a dominant transcriptional activation capacity to confer reproducible physiological levels of expression per transgene copy [2–4]. It has also recently been established that LCRs activate and coordinate transcription by coming into close proximity to the gene that is required to be expressed via a DNA/chromatin looping mechanism [3].

There are still relatively few transcriptional regulatory elements that have been described that strictly meet the criteria of LCR activity. However, bone fide LCRs have invariably been found linked with tissue-specific genes and to consist of a series of elements sensitive to digestion with DNase I [DNase I-hypersensitive (HS) sites] and located either 5′ or 3′ of the gene or genes that they regulate. LCRs located in a 5′ position are the β-globin family [3], chicken lysozyme [7], the human red and green visual pigment [8], the murine MHC Class II Ea gene (*Ea1*) [9], mouse tyrosinase (*Tyr*) [10], the human growth hormone (*GH1*)–chorionic somatomammotropin (*CSHL1*) gene cluster, which shows a dual pituitary and placental specificity [11], and possibly the human serpin gene cluster [12]. Most recently described is the LCR that regulates Th2 cell expression of the *IL4* and *IL13* cytokine genes [13]; this element is located within the *RAD50* gene, which lies between *IL5* and *IL13*. LCRs that are present 3′ of their cognate genes are human *CD2* [14], human IgH [15,16], human apolipoprotein E/C-I [17], and the murine α/δ T cell receptor (*Tcra*/*Tcrd*)–defender against cell death 1 (*Dad1*) locus [18]. The LCR activity associated with the mouse λ5–light chain (*Igl-C5*) locus appears to be distributed both 5′ and 3′ of this gene [19].

Our functional analysis in transgenic mice of a 240-kb genomic clone (240DES) encompassing the human desmin gene (*DES*), with 220-kb 5′ and 10-kb 3′ flanking sequences, is the only report to date showing definitive data demonstrating the existence of muscle-specific LCR activity [20]. We have now extended these initial studies and show that the LCR associated with the *DES* locus (*DES* LCR) is a complex, multiple-component element located within 18 kb 5′ of the *DES* transcriptional start site and is highly conserved between human and other mammalian (rodent, canine, opossum) species. We also show that the *DES* locus consists of three genes, namely *DES*, AY603755 (SPEG; the human orthologue of the murine striated muscle-specific serine/threonine protein kinase gene, Speg), and aortic preferentially expressed gene 1 (*APEG1*), suggesting that the *DES* LCR forms a key part of a functional expression module [3] that may be involved in regulating expression of SPEG and *APEG1* as well as *DES*.

Results

The 5′ DES region possesses LCR activity

Our previous analysis of PAC clone 240DES in mice showed that line DN55, harboring a single copy of this transgene, possessed a deletion that removed almost the entire 3′ flanking sequences past the poly(A)-addition site [20]. However, this deleted transgene was still fully functional with respect to *DES* expression, indicating that the *DES* transcriptional regulatory elements including any LCR are in all likelihood located 5′ of the transcriptional start site.

To test this possibility functionally we isolated cosmid clones spanning *DES* with either 22 kb (22DES; Fig. 1A) or 4.6 kb (4.6DES; Fig. 1A) of 5′ flanking sequences. These two cosmid clones were then used to generate transgenic mice. A total of three transgenic lines were obtained with 22DES, Southern blot analysis of which showed transgene copy numbers of 2, 4, and 5 (data not shown). The copy number of the four transgenic mice harboring 4.6DES ranged from 2 to 28 (data not shown) although only two mice (D0 and D1) were successfully bred on to establish lines. Total RNA was isolated from a range of muscle and nonmuscle tissues from three F1 or F2 generation animals per established line that were at least

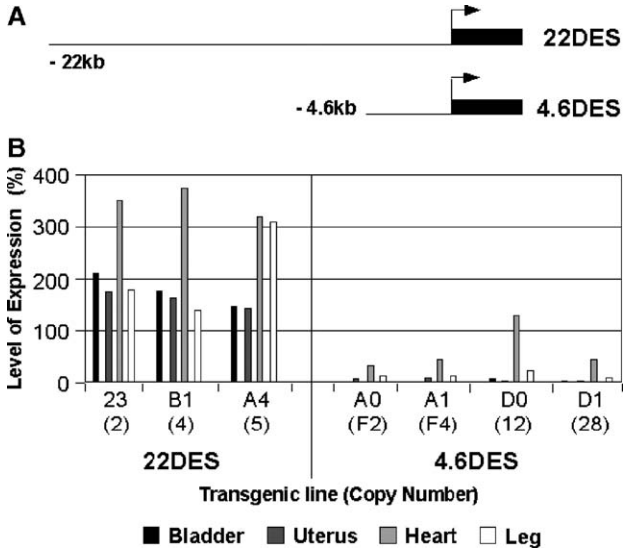

Fig. 1. A cosmid clone encompassing *DES* with 22 kb of 5′ flanking sequences gives reproducible, full physiological levels of expression in transgenic mice. (A) Illustration of the cosmid clones 22DES and 4.6DES with 22 and 4.6 kb of 5′ flanking sequences, respectively, and encompassing the complete *DES* gene. Horizontal arrows denote the direction of transcription. (B) The genomic inserts in 22DES (39.92 kb) and 4.6DES (35.76 kb) were isolated and used to generate transgenic mice. Southern blot analysis was used to determine transgene copy number and integrity (data not shown). Transgenic founder mice 22DES-23, -B1, and -A4 and 4.6DES-D0 and -D1 were successfully bred through to established lines. Total RNA was isolated from a range of muscle (bladder, uterus, leg muscle, heart) and nonmuscle (liver, kidney, spleen, lung) tissues from three F1 or F2 generation animals per established line that were at least 2 months of age. RNA from the same range of tissues was also prepared from the 4.6DES transgenic founder mice (A0, A1) that did not breed through to the F1 stage. These RNA samples were analyzed for transgene expression by an S1-nuclease protection assay and results calculated as the average level of expression of human *DES* as a percentage of endogenous mouse desmin per transgene copy (see Materials and methods). The results from the smooth muscle uterine and bladder, skeletal leg muscle, and heart muscle are shown. No expression was detected in nonmuscle tissues (data not shown). Numbers in parentheses denote transgene copy number and F represents founder transgenic mice that did not breed through to established lines. Reproducible, physiological levels of expression per transgene copy are seen in all muscle cell types with 22DES but not 4.6DES.

2 months of age. RNA from the same range of tissues was also prepared from the 4.6DES transgenic founder mice (A0, A1) that did not breed through to the F1 stage. These RNA samples were analyzed by an S1-nuclease protection assay and results calculated as the average expression of human *DES* as a percentage of endogenous mouse desmin mRNA per transgene copy as previously described [20].

The results show that 22DES gave rise to consistent, full physiological levels of *DES* expression per transgene copy in all muscle tissues analyzed (Fig. 1B, left). Expression in the smooth muscle bladder and uterine samples respectively ranged from 145 to 210% and from 144 to 175% of endogenous mouse desmin per transgene copy, whereas in cardiac muscle expression was 319–373%. Skeletal leg muscle showed expression of 141 and 179% in lines 23 and B1, respectively. However, a large positive position effect was evident in line A4, which showed an average expression of 308% per transgene copy in this tissue. Human *DES* transgene expression was not observed in nonmuscle tissues such as liver, spleen, kidney, and lung (data not shown). In marked contrast expression of the *DES* transgene from 4.6DES was significantly lower in all muscle cell types analyzed (Fig. 1B, right). Expression in the uterine and bladder smooth muscle samples was 2–10%, in cardiac muscle 31–130%, and in skeletal leg muscle 10–22% of murine desmin per transgene copy. Further analysis of transgenic mice harboring a *DES* genomic clone deleted to −1.7 kb from the transcriptional start site gave a pattern of expression similar to that seen with 4.6DES (data not shown). These data show that 22DES but not 4.6DES expresses with the same efficiency and specificity as the 240-kb PAC clone encompassing the human desmin gene that we had previously shown to possess LCR activity [20]. Taken together these results strongly suggest not only that 22DES contains an LCR but also that the region located between −4.6 and −22 kb 5′ of the *DES* transcriptional start site is essential for this function.

Muscle-specific DNase I hypersensitive sites are located 5′ of DES

All LCRs identified to date have been found to consist of a series of tissue-specific DNase I HS sites [4]. As the transgenic mouse expression data with the 22DES and 4.6DES cosmid clones indicated the presence of LCR-type elements between −4.6 and −22 kb from the *DES* Cap site (Fig. 1), we next undertook experiments to determine if muscle-specific DNase I HS sites are located within this 5′ region.

Sufficient quantities of pure skeletal myoblasts to carry out this analysis were obtained by crossing a homozygous fully functional single-copy 240DES transgenic mouse line (DN28) [20] with homozygous ImmortoMouse animals [21]. As the ImmortoMouse contains a gene for temperature-sensitive SV40 large T-antigen (tsA58), this allowed the isolation from these double-transgenic animals clones of myoblast cells with a highly extended life span in tissue culture when grown at the permissive temperature of 33°C [22]. Transgenic myoblast cultures were allowed to grow to confluency and transferred to nonmitogenic medium to stimulate differentiation (fusion) to myotubes to induce *DES* transgene expression. Nuclei isolated from these myotubes were then subjected to DNase I HS site analysis as previously described [23,24]. The presence of DNase I HS sites was investigated within a conveniently located *Xho*I restriction enzyme fragment extending from −1.7 to −18.6 kb from the start of *DES* transcription (Fig. 2D). This region extends from immediately upstream of the known skeletal muscle-specific promoter and enhancer elements [25]. Southern blot analysis employed probes from the 5′ and 3′ ends of the *Xho*I fragment (Fig. 2D).

In the 5′ direction from the start of *DES* transcription, a series of HS sites are clearly discernable as bands of 7.8 kb, 9.7 kb, 11.1/11.3-kb doublet, and 13.6 kb when using the 3′ *Pst*I probe (Fig. 2B). Analysis with the 5′ *Xho*I–*Sac*I probe gives a weak but discernable band of 8.3 kb and more prominent fragments of 7.1, 5.4, 4.9, and 1.7 kb (Fig. 2A). There is a good correlation between the location of HS sites of 7.8, 9.7, 11.1, and 11.3 kb seen with the 3′ probe respectively with those of 8.3, 7.1, 5.4, and 4.9 kb upon hybridization with the 5′ probe. The bands of 13.6 (Fig. 2B) and 1.7 kb (Fig. 2A) are detected only with their respective probes. These data taken together demonstrate the presence of DNase I HS sites at positions −9.5 kb (HS1), −11.4 kb (HS2), −13.2 kb (HS3i), −13.7 kb (HS3ii), −15.3 kb (HS4), and −16.9 kb (HS5) from the *DES* transcriptional start site (Fig. 2D, HS1–5).

Similar experiments with the nonmuscle human myelogenous leukemia cell line K562 [26] showed the absence of any DNase I HS sites within the 16.9-kb *Xho*I region, indicating that HS sites 1–5 are muscle specific (Fig. 2C). This suggests that the regions of DNase I hypersensitivity we have identified 5′ of *DES* may highlight the location of transcriptional regulatory elements and may at least in part constitute an LCR.

The DES locus consists of three closely linked muscle preferentially expressed genes

As LCRs are implicated as playing a major role in coordinating expression from within gene clusters [2–4], we placed the location of the upstream DNase I HS site regions that may constitute the *DES* LCR within the context of the locus as a whole. The Human Genome Browser Gateway (http://zeon.well.ox.ac.uk/cgi-bin/hgGateway?org=Human&db=hg12&hgsid=61887), NCBI Map Viewer (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606), and Ensembl Genome Browser (http://www.ensembl.org/Homo_sapiens/index.html) databases show a number of genes within the immediate vicinity of *DES*. Specifically, they show that 30.5 kb 5′ of *DES* is the transcriptional start site of aspartyl aminopeptidase (*DNPEP*), a housekeeping gene [27]. The 5′ end of a gene designated as aortic preferentially expressed gene 1 (*APEG1*) [28] is located 16.9 kb 3′ of *DES*. Another housekeeping gene, GDP mannose pyrophosphorylase (*GMPPA*) is positioned 32 kb 3′ of *APEG1*. In addition, the Ensembl Genome Browser shows there are a further two gene sequences that lie between *APEG1* and *GMPPA*; Q9P2P9 (ENSG00000072195), 26.4 kb in length, which is indicated as encoding a fragment of "striated muscle preferentially expressed protein," and a novel Ensembl

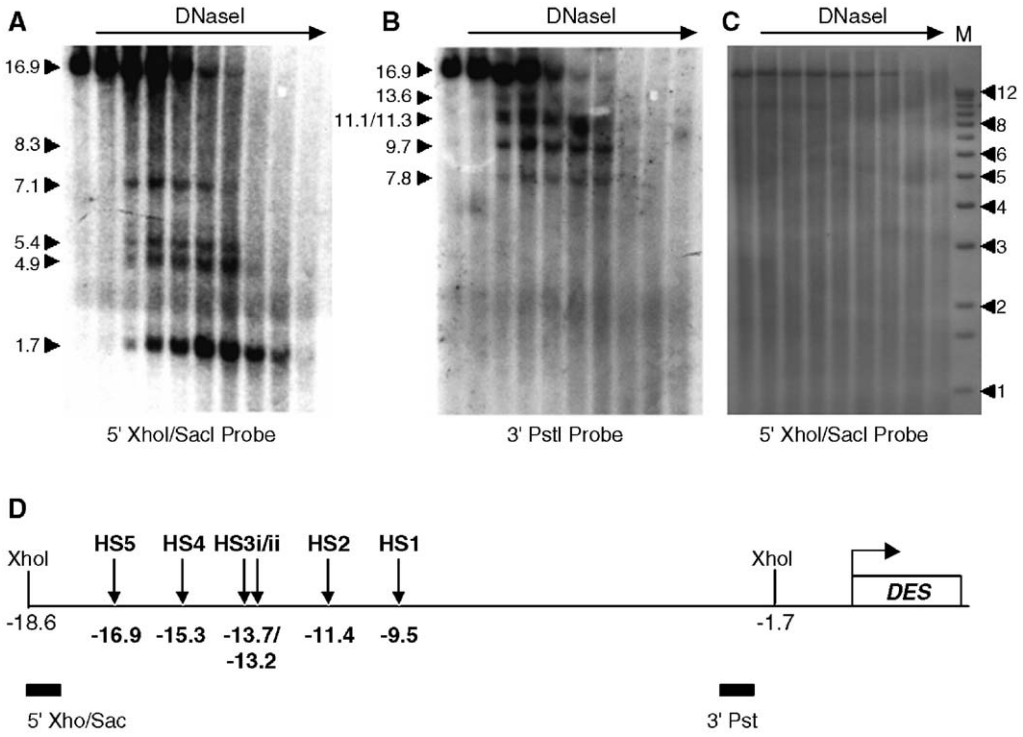

Fig. 2. Muscle-specific DNase I hypersensitive sites are located 5′ of the *DES* promoter region. DNase I HS site mapping was conducted within the 5′ region of *DES* to locate potential transcriptional regulatory elements that may constitute at least in part an LCR. Cultures of skeletal myoblast cell clones from double-transgenic mice homozygous for a single-copy integration of a fully functional 240-kb PAC clone spanning the *DES* locus (line 240DES-DN28) [20] and homozygous for a temperature-sensitive large SV40 T-antigen tsA58 gene (ImmortoMouse) [21,22] were induced to undergo fusion to myotubes, and nuclei were isolated and subjected to DNase I HS site mapping (see Materials and methods). Genomic DNA from DNase I-treated nuclei was digested with *Xho*I and Southern blotted. The presence of DNase I HS sites was assessed within a 16.9-kb region extending from −1.7 to −18.6 kb from the *DES* transcriptional start site. Blots were sequentially probed with (A) a 5′ 626-bp *Xho*I–*Sac*I fragment (position −18.6 to −17.974 kb) and (B) a 3′ 940-bp *Pst*I fragment (position −2163 to −3103 bp). Arrowheads mark the position and size in kilobases of the parental 16.9-kb *Xho*I fragment and subfragments resulting from digestion at DNase I HS sites. The horizontal arrow above the lanes denotes increasing amount of DNase I, with the sample in the first lane receiving no enzyme. (C) The same DNase I HS site mapping procedure applied to nuclei isolated from the human myelogenous leukemia cell line K562, which acted as a nonmuscle negative control. M, 1 kb ladder size markers. (D) Illustration of the 5′ region of *DES* showing the positions of DNase I HS sites (HS1–5; vertical arrows) identified in panels A and B and their distance in kilobases from the *DES* transcriptional start site. The positions of the *Xho*I sites at −1.7 and −18.6 kb and the position of the 5′ *Xho*I/*Sac*I and 3′ *Pst*I probes (black rectangles) within the 16.9-kb *Xho*I region used in Southern blots A and B, respectively, are also shown. The arrowhead denotes the direction of *DES* transcription.

predicted 1.5 kb gene fragment (ENSG00000144585) with homology to the myosin light chain kinase family.

This gene organization at the *DES* locus is similar to that seen in the equivalent position in mouse (Figs. 3B, H; data not shown) except for the annotation with respect to *APEG1*. In the mouse, the Human Genome Browser Gateway and Ensembl Genome Browser show *Apeg1* as a gene of 57 kb with its 5′ end 7.6 kb 3′ of *Des* and its 3′ end 3.6 kb 5′ of *Gmppa*. This contrasts with the indicated extent of *APEG1*, which is shown as being only 23.2 kb in length starting 16.9 kb 3′ of *DES* and terminating 32 kb 5′ of *GMPPA*.

This difference in the extent of the *APEG1* and *Apeg1* transcription units prompted us to conduct a detailed comparative sequence analysis of this region. First, we analyzed BAC clone AC053503 with the Nucleotide Identification (NIX) program of the former UK Medical Research Council Human Genome Mapping Project (HGMP) facilities using Genescan [now available through Ensembl (http://www.ensembl.org/)] and the former Celera public database. These programs predicted the presence of a 62.1-kb gene starting approximately 8.3 kb downstream of *DES* and from which the partial cDNA clone KIAA1297 (Accession No. AB037718) of the Kazusa cDNA sequencing project [29] appears to be derived (Fig. 3A; data not shown). This is homologous to what is present in the mouse and other mammals (Fig. 3B). In addition, we found by BLAST analysis (http://www.ncbi.nlm.nih.gov/BLAST/) of AB037718 that this mRNA corresponds to the human orthologue of the murine striated muscle preferentially expressed protein gene (Speg; Accession Nos. AF215896 and AAG34791), which like *APEG1* is a serine/threonine protein kinase [30]. In addition, a compilation of cDNA sequences that extends AB037718 in the 5′ direction by 2.5 kb (AY603755) and that is designated as human striated muscle preferentially expressed protein (SPEG) mRNA has been reported [31].

Although there have been extensive experimental analyses of the murine *Apeg1* [28] and Speg [30] genes and their transcripts, none has as yet been reported on the corresponding human locus. We therefore conducted multiple muscle-tissue Northern blot analysis to determine whether AY603755 displays a preferential striated (skeletal and cardiac) muscle expression pattern, which would confirm and add further evidence that it is the human orthologue of Speg. Based on published data on the

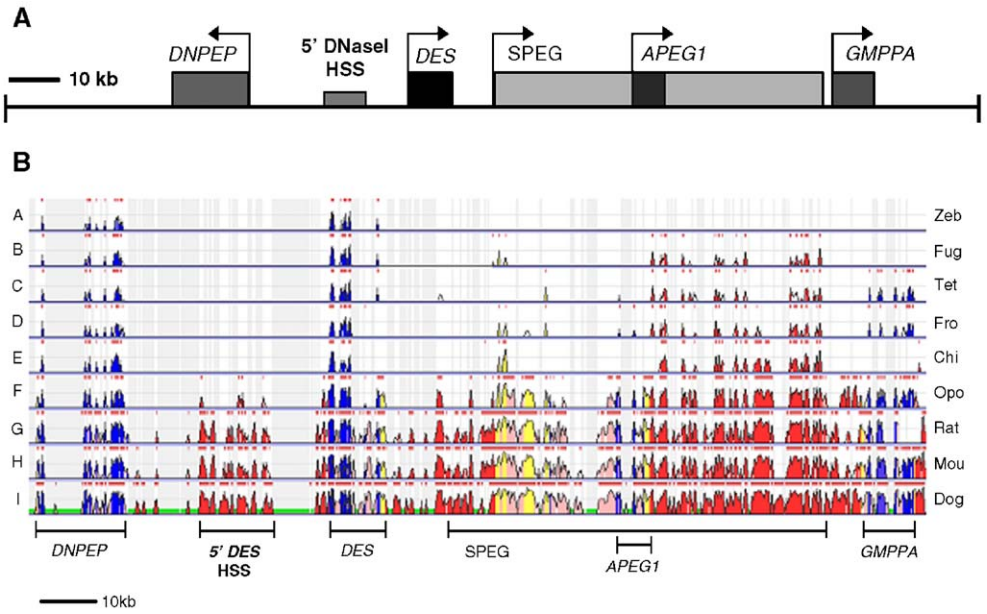

Fig. 3. Gene order at the *DES* locus and the DNA sequence underlying the muscle-specific DNase I HS sites 5′ of *DES* is highly conserved between human and mammalian species. (A) Illustration of the gene organization at the *DES* locus ascertained by bioinformatics and direct molecular analysis (see Materials and methods; Fig. 4). Large shaded rectangles denote identified genes; *DNPEP*, aspartyl aminopeptidase; *DES*, desmin; *APEG1*, aortic preferentially expressed gene 1 [28]; *GMPPA*, GDP mannose pyrophosphorylase; SPEG, human orthologue of murine striated muscle-specific serine/threonine protein kinase (Speg) [30]. 5′ DNase I HSS, region encompassing the muscle-specific DNase I HS sites (Fig. 2) that constitute the *DES* LCR. (B) Multiple vertebrate genome comparison of the *DES* locus employing the ECR Browser. The conservation profiles of the human *DES* locus region in comparison with the zebrafish (A, Zeb), *Fugu* (B, Fug), tetraodon fish (C, Tet), frog (D, Fro), chicken (E, Chi), opossum (F, Opo), rat (G, Rat), mouse (H, Mou), and dog (I, Dog) genomes are shown. Plots were constructed by using a sliding window of 30 bp throughout the alignment with a minimum identity of 80%. A conserved alignment is shown in blue if it overlaps with a coding exon; yellow, untranslated exon region; pink, intron; red, intergenic region. Red bars at the top of each layer of sequence comparison provide an overview of the distribution of evolutionarily conserved regions and are used to highlight underlying regions of alignment. Green bars at the base of each layer indicate the locations of repetitive elements. The positions of the various genes within the *DES* locus (A) as well as the 5′ DNase I hypersensitive sites (5′ DNase I HSS; see Fig. 2) that potentially constitutes the *DES* LCR are indicated. *Note*: SPEG is currently not highlighted within the database but is shown as conserved intergenic sequences.

murine Speg gene [30] and our analysis employing the former Celera public database (data not shown), AY603755 lacks the sequence of the first 358-bp exon (Accession No. DQ395348; see Supplementary Fig. 1). This additional 5′ exon places the start of the transcription unit within which AY603755 is contained 8.3 kb 3′ of *DES* rather than 16.9 kb as indicated in the current database (Fig. 3A). We therefore decided to hybridize the Northern blot with a probe corresponding to the putative missing first exon of AY603755 (see Materials and methods) to confirm its authenticity. The results show that, again as can be deduced from the murine situation [30], multiple RNA isoforms are detected with this probe (Fig. 4). However, an approximately 11-kb mRNA was detected only in skeletal muscle and, to a lesser degree, in cardiac tissue and appears to correspond to Spegβ [30]. A predominant novel mRNA of about 5 kb is expressed at markedly higher levels in skeletal leg and cardiac muscles. This confirms the striated muscle preferentially expressed nature of AY603755 and therefore in all likelihood does correspond to the human equivalent of Speg [30,31].

Murine *Apeg1* was originally reported as a 4.5-kb gene with a preferential aortic vascular smooth muscle expression pattern [28]. *Apeg1* was subsequently shown to be present within and constitute one isoform of a much larger transcription unit that also encodes the striated muscle preferentially expressed proteins SPEGα and β as well as the brain isoform BPEG [30]. Our comparative sequence analysis (Fig. 3B; data not shown) shows that the human equivalent of *Apeg1* is 5.5 kb in length and is similarly contained within the larger transcription unit encoding the human SPEG protein and from which AY603755 is derived. Consequently, this places the 5′ end of *APEG1* 26.4 kb downstream of the predicted SPEG gene transcriptional start site (Fig. 3A). Furthermore, the entire AB037718 cDNA corresponds to the Q9P2P9 gene and Q9P2P9 as well as AY603755 may lack additional 3′ sequences contained within the Ensembl Genome Browser predicted gene fragment ENSG00000144585. As a result both Q9P2P9 and AY603755 sequences are indicated in the current database to terminate 5.3 kb from the *GMPPA* start site instead of 1.1 kb (Figs. 3A and 3B). Collectively, these data show that the *DES* locus consists of a functional chromatin cluster of three closely linked, muscle-specific (*DES*) or muscle preferentially expressed (SPEG; *APEG1*) genes within a 79-kb interval (Fig. 3A).

Gene order and the 5′ DNase I HS site regions at the DES locus are highly conserved across mammalian species

We conducted comparative genome analysis between human and other species using the ECR Browser (http://ecrbrowser.dcode.org/) [32]. This demonstrated that order, relative position, and transcriptional direction of genes, as well as the 5′ DNase I

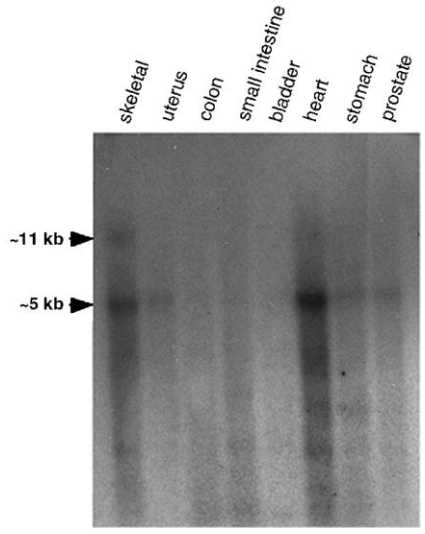

Fig. 4. Confirmation of the AY603755 cDNA, murine Speg orthologous transcription unit 3′ of *DES*. A commercially available Northern blot of poly $(A)^+$-RNA samples from various human muscle tissues was probed with a PCR-generated 400-bp unique sequence DNA fragment spanning a predicted first exon (Fig. 3B; see Supplementary Fig. 1) located 5′ of the human SPEG compilation sequence AY603755 and 8.3 kb downstream of *DES* (see Materials and methods). This probe highlights the presence of different mRNA isoforms starting from this position. A major 5-kb species is present at markedly elevated levels within skeletal and cardiac tissues. The largest, 11-kb RNA also shows a striated (skeletal, cardiac) muscle-specific expression pattern. Arrows denote the approximate sizes of the main isoforms.

HS site regions (Fig. 2) at the *DES* locus, are highly conserved in the rat (Figs. 3B, G), mouse (Figs. 3B, H), and dog (Figs. 3B, I) genomes. Gene type and order at the desmin locus are also conserved between humans/mammals and the North American opossum marsupial (Figs. 3B, F). However, the regions of the 5′ DNase I HS sites are less well conserved in the opossum, with only components corresponding to HS1 and HS3 as well as a short element 5′ of HS5 (Fig. 2) showing homology above the 50% threshold of analysis (Figs. 3B, F; see Supplementary Fig. 2). Although the relative positions of *DES* and *DNPEP* are preserved between human and chicken, frog, *Fugu*, zebrafish, and tetraodon fish (Figs. 3B, A–E), these loci also show a number of striking comparative differences, with the region corresponding to the 5′ half of the SPEG gene and *APEG1* as well as the 5′ DNase I HS site regions being absent (Figs. 3B, A–E). In addition, the *GMPPA* gene appears to be absent from the 3′ end of the locus in chicken, *Fugu*, and zebrafish (Figs. 3B, A, B, and E).

The conserved sequence within the 5′ DNase I HS site regions

Comparative genomic DNA sequence analysis across species has been successfully used to highlight known and potential transcriptional regulatory elements [33–36]. We therefore performed a detailed human–rodent comparative sequence analysis of the 5′ *DES* region using a number of programs to confirm the identity of regions of homology and determine their precise sequence. As the sequence of BAC clones spanning the desmin locus in human (AC053503), mouse (AC114651), and rat (AC121633) is available, regions up to 20 kb upstream of the transcriptional start site were initially used for dot-plot analysis employing the Jemboss software package provided by the former UK Medical Research Council HGMP facilities. Regions of homology between both human and mouse and human and rat were obtained within the area immediately 5′ of the desmin gene, which included the previously identified promoter as well as skeletal [25,37] and vascular smooth [38] muscle enhancer elements. The only other regions of homology identified by this analysis were short stretches of sequence distributed between approximately −9 and −18 kb from the *DES* transcriptional start site, again in both human–mouse and human–rat comparisons (data not shown).

To improve the resolution and confirm the results obtained by dot-plot analysis, a sequence comparison was then undertaken using the mVISTA program (http://www-gsd.lbl.gov/vista/) with a window size set at 50 bp and a high homology threshold of 75%. A comparison of the human and mouse sequences showed peaks of homology not only as expected at the exon and promoter regions (data not shown), but also between −9 and −18 kb from the *DES* Cap site (Fig. 5B). Furthermore, there is a good correlation between the peaks of human–mouse homology and the position of the muscle-specific DNase I HS sites (Fig. 5B, vertical arrows), with the slight exception of HS5, which in all likelihood reflects a lack in the resolution of the Southern blot analysis (Fig. 2). Similar results were obtained with a comparison of human and rat sequences (Fig. 3B and data not shown).

We next determined precisely the regions of human–rodent homology that underlie the muscle-specific DNase I HS sites, by a DNA sequence comparison of the same 20-kb upstream regions employing BLAST followed by CLUSTALW (http://www.ebi.ac.uk/clustalw/) to obtain a multisequence alignment. Within the area −9 to −18 kb we found 12 elements of high (83–100%) human–rodent homology ranging in size from 32 to 411 bp (Fig. 6), which precisely correspond to the peaks of homology revealed by the mVISTA analysis (Fig. 5C).

The 5′ DES regions of high human–rodent homology are areas with the potential to bind a high density of transcription factors specific for all muscle cell types

All DNase I HS site elements that constitute LCRs and that have been analyzed to date consist of clusters of tissue-specific and ubiquitous transcription factor binding sites [4]. To obtain initial insight into the complement of factors that may bind to, and mediate the function of, the human–rodent regions of homology (Figs. 5 and 6) corresponding to the muscle-specific HS sites 5′ of *DES* and that may constitute the *DES* LCR (Fig. 1), we scrutinized the sequence of these elements for potential transcription factor binding sites that have been demonstrated to regulate muscle-specific gene expression [39] using a number of prediction programs (rVISTA, ConSite, TFSEARCH, PATCH; see Materials and methods).

These programs do indeed predict the presence of both tissue-specific and ubiquitous transcription factor binding sites on both forward (+) and reverse (−) strands of the DNA that are

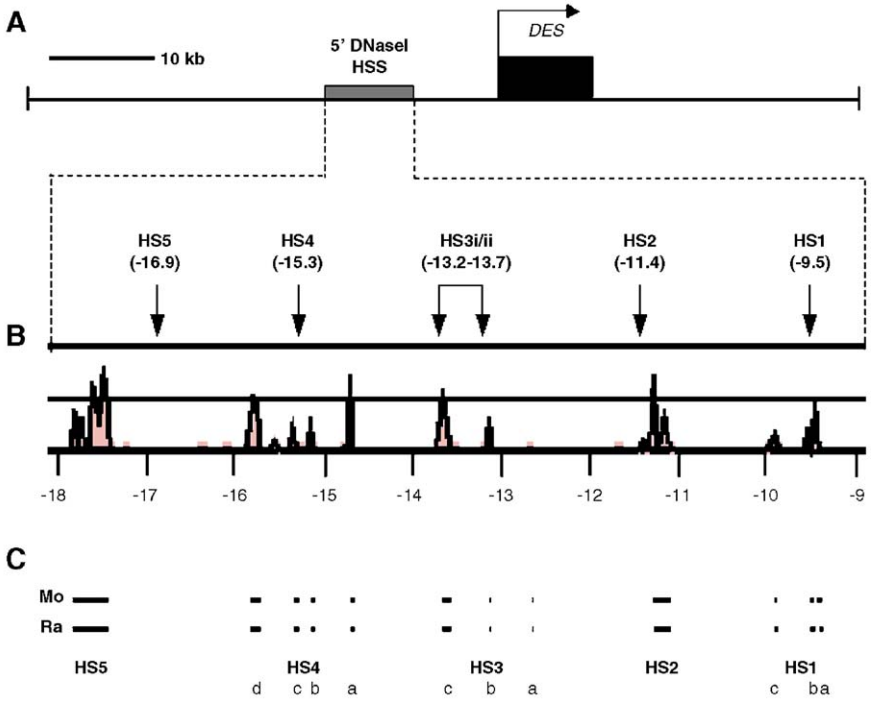

Fig. 5. Bioinformatics analysis of the DNA sequence underlying the muscle-specific DNase I HS sites of 5′ *DES* (Fig. 2) reveals 12 elements that are highly conserved between human and rodent species. (A) Illustration of the relative positions of *DES* and the muscle-specific 5′ DNase I HS sites (HSS) that constitute the *DES* LCR (see Figs. 2 and 3). (B) The region starting from within exon I and extending for 20 kb 5′ of the desmin gene transcriptional start site was obtained from human (AC053503) and mouse (AC114651) BAC clones and subjected to a comparative DNA sequence analysis using the mVISTA program employing a window size set at 50 bp and a high homology threshold of 75%. Peaks highlighting elements with a minimum 75% homology were observed not only as expected within the promoter region (data not shown), but also in the region between −9 and −18 kb from the *DES* transcriptional start site and which corresponded to the positions of the muscle-specific DNase I HS sites (HS1–HS5, vertical arrows; see Fig. 2). Very similar results were obtained with a human–rat comparative analysis (data not shown). (C) Illustration of the positions of the 12 elements of human–mouse (Mo) and human–rat (Ra) homology obtained by BLAST followed by CLUSTALW comparative sequence analysis (Fig. 6) between −9 and −18 kb from the *DES* transcriptional start. The 12 elements are grouped into five locations that correspond to the peaks of homology obtained by mVISTA and the areas occupied by the muscle-specific DNase I HS sites. The elements of human–rodent homology are therefore named by reference to the DNase I HS site with which they are associated. Numbers indicate distances in kilobases from the *DES* start of transcription.

strictly conserved between human and rodent sequences (see Supplementary Fig. 3) and that have been shown to regulate gene expression in skeletal, cardiac, and smooth muscle both during development and in adult life [39].

The 5′ DES DNase I HS site human–rodent homology regions are sufficient for reproducible, physiological levels of muscle-specific expression in transgenic mice

The functional expression data in mice with the 22DES and 4.6DES cosmid clones (Fig. 1) in combination with the DNase I HS site mapping (Fig. 2) and DNA sequence analysis (Figs. 3, 5, and 6) strongly suggest that the region between −9 and −18 kb 5′ of the *DES* Cap site is a potent muscle-specific transcriptional regulatory element that may have an LCR capability. To confirm that this region alone is sufficient to confer LCR function and that no elements are present either within or downstream of *DES*, we asked whether these *DES* 5′ flanking sequences could confer reproducible, high levels of expression upon a heterologous reporter gene. A construct extending from −18.6 to +30 bp from the *DES* transcriptional start site, which includes all the DNase I HS site/human–rodent homology regions, and linked to a human β-globin (*HBB*) reporter gene (18.6DESβ; Fig. 7A), was functionally analyzed in transgenic mice. A total of six transgenic lines harboring 18.6DESβ were obtained and ascertained by Southern blot analysis (data not shown) to range in transgene copy number from 3 to 11. Expression analysis in three mice per transgenic mouse line harboring this construct was conducted using an S1-nuclease protection assay, simultaneously probing for the 3′ end of the *HBB* reporter mRNA [40] and endogenous murine desmin mRNA [20]. *HBB* mRNA levels were calculated as a percentage of murine desmin on a transgene copy number basis as in the case of the 22DES and 4.6DES mice (Fig. 1B).

Unlike in the case of transgenic animals harboring 22DES and 4.6DES (Fig. 1), it is not possible to obtain an absolute measure of 18.6DESβ expression efficiency in relation to the endogenous murine *Des* internal reference since, in all likelihood, the mRNAs derived from these two genes will differ in their relative stabilities in the various muscle tissues analyzed. However, all transgenic lines gave rise to high levels of *HBB* reporter gene expression in all muscle cell types analyzed, with the same degree of reproducibility per transgene copy as that obtained with the 22DES genomic clone (Fig. 7B). Expression in the smooth muscle bladder and uterine compartments was 73–140 and 56–130%, respectively; in cardiac muscle 95–177%; and in skeletal leg muscle 46–136% of endogenous mouse desmin mRNA per transgene copy. As in the

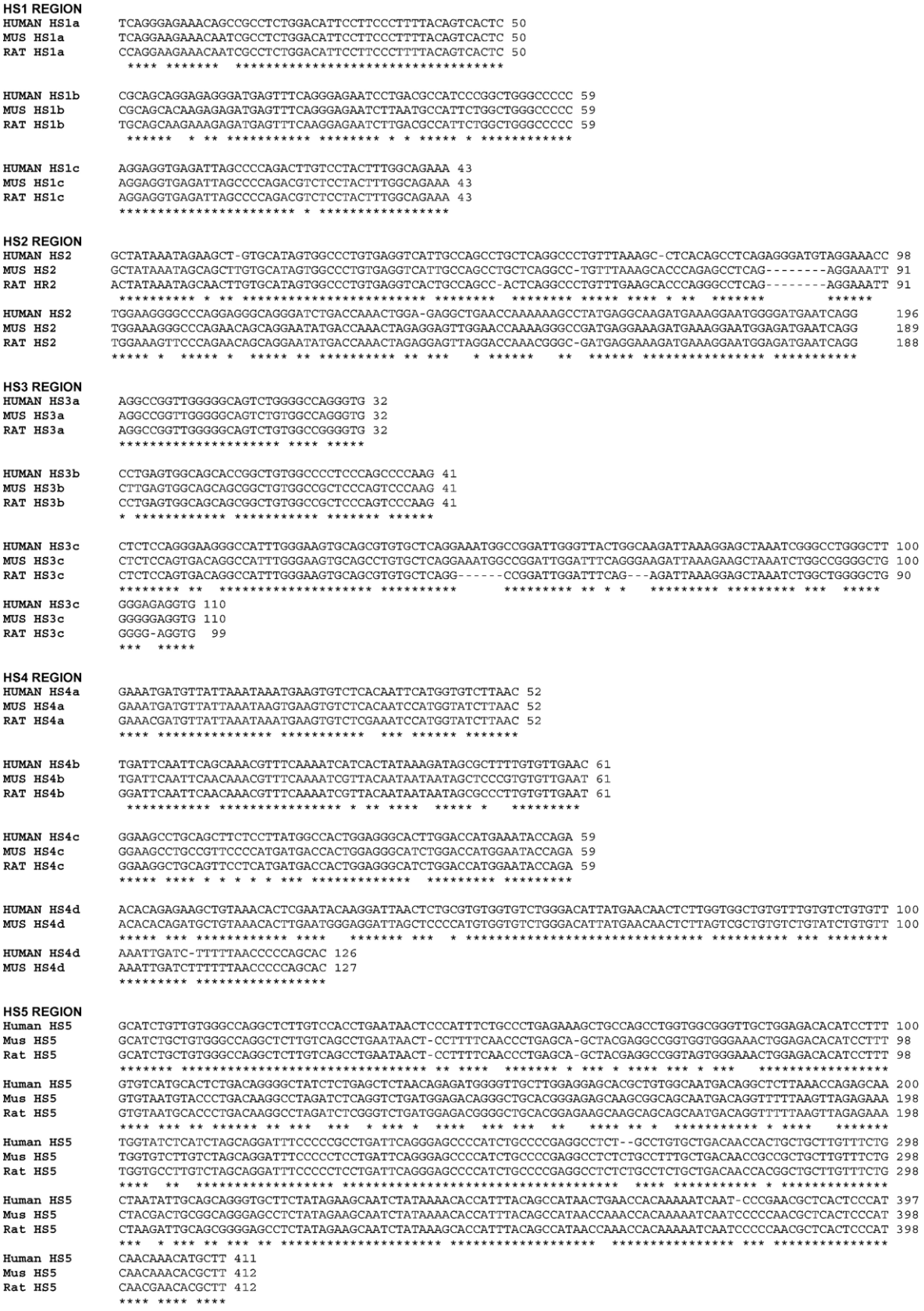

Fig. 6. The sequence of the human–rodent (mouse, rat) regions of homology that underlie the muscle-specific DNase I HS sites. The 20-kb regions 5′ of the *DES* transcriptional start site were obtained from human (AC053503), mouse (AC114651), and rat (AC121633) BAC clones. These were subjected sequentially to a DNA sequence comparison employing BLAST followed by CLUSTALW to obtain a multisequence alignment. Within the area −9 to −18 kb a total of 12 elements that are highly conserved (83–100% homology) between humans and rodents (Mus; Rat) ranging in size from 32 to 411 bp were identified and correspond to the regions of the muscle-specific DNase I HS sites (Figs. 5B and 5C). Stars denote nucleotides that are conserved between all three species.

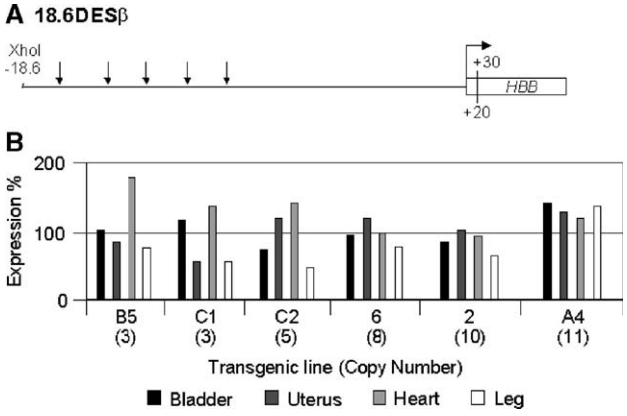

Fig. 7. The 5′ *DES* region is sufficient to confer reproducible, physiological levels of expression in transgenic mice. (A) Illustration of the 18.6DESβ construct that links the 5′ region of *DES* extending from −18.6 to +30 bp from the transcriptional start site (horizontal arrow) to a human β-globin (*HBB*) reporter gene starting at +20 bp. The positions of the muscle-specific DNase I HS sites (Fig. 2) encompassing the elements that are highly conserved between humans and other mammals (Figs. 3, 5, and 6) are indicated (vertical arrows). (B) 18.6DESβ was used to generate six lines of transgenic mice. Transgene copy number and integrity were determined by Southern blot analysis (data not shown). Expression analysis in three mice per transgenic mouse line harboring this construct was conducted using total RNA and an S1-nuclease protection assay in a range of muscle (bladder, uterus, heart, leg muscle) and nonmuscle (liver, kidney, spleen, lung) tissues, simultaneously probing for the 3′ end of the *HBB* reporter mRNA and endogenous murine desmin mRNA (see Materials and methods). Average *HBB* mRNA levels were calculated as a percentage of murine desmin on a transgene copy number basis as in the case of the 22DES and 4.6DES mice (Fig. 1B). Numbers in parentheses denote transgene copy number. The results from the smooth muscle uterine and bladder, skeletal leg muscle, and heart muscle are shown. No expression was detected in nonmuscle tissues (data not shown). Reproducible, physiological levels of expression per transgene copy are seen in all muscle cell types.

case of the 22DES and 4.6DES cosmid clone analysis (Fig. 1), no *HBB* reporter transcripts were detected in nonmuscle tissues in this experiment (data not shown). However, to confirm that the 5′ *DES* region in 18.6DESβ could drive expression in vascular smooth muscle as well as smooth muscle organs (uterus, bladder), the S1-nuclease protection assay was repeated with larger (20–30 μg) amounts of RNA from the low 3-copy lines B5 and C1 and the high 11-copy line A4. Clear signals of both the endogenous murine *Des* and the *HBB* transgene transcripts were observed in the highly vascularized spleen, kidney, liver, and lung tissues that were analyzed, the ratio of which was comparable to that seen in uterus and bladder (data not shown).

As 18.6DESβ gave a reproducible pattern of expression that was very similar to that obtained with the equivalent native *DES* construct (22DES; Fig. 1B, left), these data imply that the 5′ region extending to −18.6 kb from the transcriptional start site is sufficient to drive physiological, pan-muscle-specific expression of *DES* and that no crucial genetic regulatory elements are present either within or downstream of the gene.

Discussion

We present here the first report describing the localization and characterization of a muscle-specific LCR. We originally showed that this transcriptional control element is present on a 240-kb PAC clone (240DES) spanning *DES* [20] and now show that it consists of five regions of muscle-specific DNase I hypersensitivity located between −9 and −18 kb 5′ of the transcriptional start site (Figs. 2–5). The *DES* LCR can reproducibly drive full physiological levels of *DES* expression in all muscle cell types in transgenic mice (Figs. 1 and 7), which makes it functionally unique among muscle-specific transcriptional regulatory elements described to date.

The 22DES cosmid clone gave levels of expression that were greater than 100% per transgene copy, especially in the cardiac compartment (Fig. 1B). This is similar to the results we obtained with the much larger 240DES PAC clone [20]. A feature that the genomic regions contained in 240DES and 22DES have in common is they lack the downstream SPEG/*APEG1* transcription unit (Fig. 3A). If the *DES* LCR regulates the muscle-specific expression pattern of all three genes within this cluster, then the downstream SPEG and *APEG1* gene promoters could compete with *DES* for LCR-mediated activation in striated and smooth muscle cell types, respectively. In support of this possibility is the finding that *DES*-SPEG and *DES-APEG1* are co-expressed within the same muscle cells [30]. This situation would be analogous to the polar, unidirectional competition observed between the LCR and the globin genes during development within the human *HBB* and mouse *Hbb* loci [3]. In the absence of competition from SPEG and *APEG1*, expression of *DES* in the 240DES and 22DES transgenes could therefore be elevated above normal physiological levels (Fig. 1B). This would be equivalent to the increased levels of murine *Hbb2* expression that are observed after deletion of the competing upstream *Hbb1* gene [41] and the embryonic rather than adult stage expression of *HBB* when it is placed in the more LCR-proximal position normally occupied by *HBE1* [4].

In addition, our results show that expression of *DES* from either the 240DES [20] or the 22DES (Fig. 1B, left) transgene is consistently higher per transgene copy in cardiac tissue compared to smooth or skeletal muscle. There are a number of reasons this can be the case. First, this may be an artifact of expressing a human gene in a murine background. Second, it is possible the higher level of *DES* transgene expression in the heart may reflect a true physiological condition; that is, *DES* expression is proportionally higher in cardiac than in skeletal or smooth muscle in humans. Interestingly, *DES* expression from the 4.6DES transgene is also reproducibly higher in the heart than in either skeletal or smooth muscles (Fig. 1B, right), suggesting that a cardiac-specific enhancer element may reside within 4.6 kb of the *DES* transcriptional start site.

Unfortunately, analysis of cells from all of our mouse lines (Figs. 1 and 7) by double-fluorescence in situ hybridization using transgene and γ-satellite repeat-specific probes [5] (see Materials and methods) showed that none possessed a centromeric transgene integration event (data not shown) that would have constituted a clear test for the presence of a dominant chromatin-opening activity [6]. Nevertheless, the reproducible, physiological level of expression per transgene copy, especially at low two to three transgene copy numbers, in all muscle cell types that we observed with 22DES (Fig. 1B,

left) and 18.6DESβ (Fig. 7B) demonstrates that the 5′ *DES* region possesses a dominant transcriptional activating capability and therefore an LCR function. It is also evident that despite the clear trend toward transgene copy number-dependent expression for both the 22DES and the 18.6DESβ transgenes, there is a small (two- to threefold) variation in the level of expression per transgene copy in a given tissue type between different mouse lines, albeit within a physiological range and similar to what is observed with other LCR-type elements [4]. This degree of variability in expression of 22DES and 18.6DESβ is far lower than that observed with transgenes driven only by classical promoter and enhancer elements [42,43]. Transgenes of this type are markedly prone to position effects and typically give low and highly variable levels of expression (varying over several orders of magnitude) that are not proportional to transgene copy number [42,43].

Our data using a cosmid clone encompassing *DES* with 4.57 kb of 5′ flanking sequences (4.6DES; Fig. 1A) also showed that this was sufficient to drive expression in all muscle cell types (Fig. 1B). This is similar to the results obtained using the equivalent 5′ region from murine *Des* to drive a *lacZ* reporter gene in mice [38]. However, in comparison to 22DES (Fig. 1B, left), which possesses full *DES* LCR activity, the level of expression obtained with 4.6DES was markedly lower, being on average approximately 30-fold less in the smooth (bladder/uterine) and 10-fold down in the cardiac and skeletal muscle compartments (Fig. 1B, right).

Although there have been previous reports of regions possessing muscle-specific LCR-like activity linked with the 5′-flanking region of the murine β myosin heavy chain [44] and the muscle promoter of the human [45] and rat [46] aldolase A gene, neither produces full levels of gene expression nor do they function at low transgene copy number, suggesting the lack of complete LCR function. The reproducible, full physiological expression pattern demonstrated by the *DES* 5′ flanking region described in this report, therefore, constitutes the first detailed description of an LCR capable of functioning in all muscle cell types.

At present the genome database shows a transcription unit of 57 kb in mouse and 23.2 kb in human immediately downstream of the desmin gene designated as aortic preferentially expressed gene 1 (*APEG1* and *Apeg1*, respectively, in human and mouse; e.g., see http://www.ensembl.org/). However, it has been shown that the transcription unit within this region, which possesses an authentic aortic smooth muscle preferential expression pattern, is a small (4.5 kb) gene, giving rise to a 1.8-kb transcript [28]. It was subsequently shown that this 4.5 kb gene was contained within a larger transcription unit from which are produced a number of related serine/threonine protein kinases but with a markedly different tissue expression pattern compared to that of *Apeg1*, namely, striated muscle preferentially expressed proteins α and β (SPEGα and SPEGβ) as well as a brain isoform (BPEG) [30]. In addition, comparative analysis of human and mouse sequences (Fig. 3B; data not shown) indicate that the region corresponding to the 4.5-kb *Apeg1* gene in humans is 5.5 kb in length and again resides within the much larger SPEG-encoding transcription unit (Fig. 3B). These previous studies showed that *Apeg1*, SPEGα/β, and BPEG transcripts have a number of exons in common and may therefore be considered as isoforms from the same overall transcription unit [30]. However, murine SPEGβ is a 10.8-kb transcript that is produced from an independent promoter 5′ of the *Apeg1* Cap site, with the *Apeg1* promoter residing within a SPEGβ intron. In addition, SPEGβ possesses extensive 3′ sequences that are lacking in *Apeg1*. SPEGα appears to share the first four exons with *Apeg1* but possesses the same large 3′ region as that of SPEGβ, giving rise to an 8.4-kb mRNA. Furthermore, little or no *Apeg1* is expressed in striated skeletal/cardiac muscle and very low levels of SPEGα/β are detected in aortic (smooth) muscle [30].

Our sequence analysis demonstrates that humans and rodents as well as other mammalian species are highly homologous with respect to gene organization at the *DES* locus, including the position of *APEG1* (Fig. 3B; data not shown). Northern blotting of muscle-tissue mRNA (Fig. 4) in combination with sequence analysis (Fig. 3B and Supplementary Fig. 1; data not shown) has confirmed that 11- and 5-kb transcripts that include exon sequences lying approximately 8 kb downstream of *DES* are generated and preferentially expressed in striated (skeletal and cardiac) muscle (Accession No. DQ395348; Fig. 4). This contrasts with the current placement of the start of "*APEG1*" in the genome database, which is indicated as 16.9 kb 3′ of *DES* (see http://www.ensembl.org/). This is due to the fact that the first exon of the transcription unit downstream of *DES* predicted by our data (Accession No. DQ395348; see Supplementary Fig. 1 and Fig. 4) is at present not included. Comparison with the equivalent situation in the mouse [30] suggests that the 11-kb mRNA we detect (Fig. 4) would appear to correspond to the transcript encoding SPEGβ [30]. The prominent 5-kb transcript that we also detect (Fig. 4) does not correspond in size to any of those shown to be present in the mouse [30] and may therefore represent a human-specific isoform.

In addition, our sequence analysis (Fig. 3B; data not shown) also suggests that the three-exon, Ensembl-predicted 1.5-kb gene fragment (ENSG00000144585) shown to lie between *APEG1* and *GMPPA* (see http://www.ensembl.org/) is, by homology to the mouse, part of the transcripts encoding SPEGα and SPEGβ [30]. This is supported by our observation that combining the novel first exon and the exon sequences of ENSG00000144585 with the partial SPEG cDNA sequence already present in the database (AY603755) results in a predicted transcript of 10,209 bp (Supplementary Fig. 1), which approximates in size SPEGβ (Fig. 4) [30].

In summary, from the perspective of gene and corresponding protein function it would appear to be inaccurate to designate the entire transcription unit 3′ of *DES*/*Des* as "*APEG1*/*Apeg1*." The data presented in this report will therefore allow a significant update of the genome database with respect to the annotation of *APEG1*/*Apeg1* so that it accurately reflects both gene structure and protein function.

Bioinformatics analysis showed that the 5′ DNase I HS site regions that constitute the *DES* LCR are highly conserved in their location and number between human and other mammalian species (Fig. 3B). A detailed comparative DNA sequence

analysis showed a total of 12 regions ranging in size from 32 to 411 bp and distributed between −9 and −18 kb to be highly (83–100%) conserved between human and rodent (mouse, rat) species and to coincide with the muscle-specific *DES* LCR HS sites (Figs. 5 and 6). This high degree of species conservation of these elements further underscores their functional significance. With the exception of HS3a and HS4c, which lack any predicted factor binding sites, each of the regions of human–rodent homology shows a unique set and pattern of potential muscle-specific and ubiquitous transcription factor binding elements (see Supplementary Fig. 3) [39] that is consistent with their potent ability to drive a dominant pan-muscle-specific gene expression (Figs. 1 and 7). In addition, the fact that the only transcription factor binding sites that are conserved between humans and the opossum are those within HS3b and HS3c (Supplementary Fig. 2) suggests that these two elements may be particularly important for *DES* LCR function. It is also evident that there are a number of regions that reside 3′ of *DES* and 5′ of SPEG that are conserved between human and mammalian species (but not the opossum). The functional significance (if any) of these regions of homology within a transcriptional regulatory context is currently unknown but these elements appear not to be crucial for *DES* LCR activity since reproducible, physiological levels of muscle-specific gene expression can be obtained in their absence (Fig. 7) [20].

The presence of two muscle preferentially expressed genes downstream of *DES* (Fig. 3A) also raises the possibility that the *DES* LCR is an integral and important component of a functional expression module [3] that may regulate and coordinate transcription of not only *DES* but also SPEG and *APEG1*. This multiple cell and, possibly, gene specificity associated with the *DES* LCR is a feature shared by other elements of this type such as those from the human *GH1–CSHL1* cluster, which shows a dual pituitary and placental specificity [11], and the murine *Tcra*/*Tcrd–Dad1* locus, which appears to possess both a T-cell-specific and ubiquitous function [18]. *DES* LCR-mediated gene expression of *DES* and possibly SPEG and *APEG1* may take place via a DNA/chromatin looping mechanism and formation of an ACH as has been found for the equivalent element within the murine *Hbb* and human *HBB* loci [3,47]. Alternatively, a "facilitated tracking" mechanism, as has been proposed to mediate the function of the human hepatocyte nuclear factor-4α gene (*HNF4A*) upstream enhancer [48], may be in operation in this context. In either of these scenarios it would be necessary for different components of the *DES* LCR to function in different cell types to bring about the correct pattern of muscle-specific expression of *DES* and possibly SPEG and *APEG1*. In addition, our data to date do not provide insight as to whether the *DES* LCR can function as an independent tissue-specific transcriptional regulatory element that is capable of fully activating at least some heterologous promoters as has been found with the equivalent elements from the human *HBB* ([49–51]; M. Antoniou and F. Grosveld, unpublished results) and *CD2* [52] genes. Future work is aimed at testing these and other mechanistic models to elucidate the precise nature of *DES* LCR-mediated gene function.

Overall, the *DES* LCR would appear to be a highly complex element, which in all probability functions by cooperation between its various components [2–4] to fulfill the need to regulate *DES* expression both spatially and temporally in all muscle cell types during development and adult life [20]. In addition, the structure and function of the *DES* LCR may represent a class of transcriptional regulatory elements that generally direct a developmentally multifaceted spatiotemporal muscle-specific gene expression program, further exemplified by the −48- to −58-kb region within the murine *Mrf4–Myf5* locus [53].

Finally, given that the defining characteristic of LCRs is their ability to overcome chromatin position effects and to provide reproducible, high levels of expression, the inclusion of these elements in integrating (retroviral and lentiviral) vectors can greatly assist in overcoming these problems within a gene therapy context as has been found with the use of the human *HBB* [54] and *CD2* [55] elements. In addition, we have also shown using human *HBB* as a model system that LCR elements are required to maintain gene expression from replicating episomal vectors [23]. Our data show that the *DES* LCR is capable of driving high, reproducible levels of muscle-specific expression of a heterologous gene (Fig. 6). This raises the possibility that efficient *DES* LCR-based vectors could be developed for targeting gene expression to any desired muscle cell type for a wide range of applications, including gene therapy for the muscular dystrophies, the systemic release of a therapeutic protein, and DNA vaccines.

Materials and methods

Cosmid genomic clones and plasmid test gene constructs

Genomic clones spanning the human desmin gene (*DES*) were isolated using standard procedures from a genomic library (provided by Professor Frank Grosveld) in the cosmid vector pTCF [56]. Nitrocellulose filters, on which the library was spread, were hybridized with an 810-bp *Xho*I–*Kpn*I fragment extending from −1700 to −890 bp from the *DES* transcriptional start site. The extent of the two clones isolated was precisely determined by sequencing and comparison to the human genome database (see Bioinformatics analysis section below). These two clones contained an intact *DES* gene and were found to be 35.76 and 39.92 kb in size. The 39.92-kb clone possesses 22.33-kb 5′ and 9.25-kb 3′ flanking regions and was therefore designated as 22DES. The 35.76-kb clone was found to have 4.57 kb of 5′ flanking and 22.85 kb 3′ sequences and was designated as 4.6DES.

The 18.6DESβ construct, which consisted of the *DES* 5′ region extending from −18.6 kb to +30 bp from the transcriptional start site linked to a human β-globin (*HBB*) reporter gene was constructed as follows: a 2449-bp *Sac*I fragment spanning the *DES* Cap site between −2275 and +174 bp was subcloned into the *Sac*I site of pBluescript (Stratagene Europe, Amsterdam, The Netherlands). This was then linearized with *Sac*II, which cuts at the 3′ end of the *DES* promoter region insert, and subjected to partial digestion with *Bal*31. Digested ends were then blunted through reaction with T4 DNA polymerase, and desmin gene inserts extending to −1.7 kb from the Cap site were excised from the pBluescript vector by digestion with *Xho*I, isolated by gel electrophoresis, and recloned between the *Xho*I and the *Eco*RV sites of pBluescript. A clone that had been deleted to a position +30 bp from the *DES* transcriptional start site (and therefore missing the ATG translation start codon, which is 50 bp downstream from this point) was then selected and confirmed by DNA sequencing and designated 1.7DES. A 3265-bp *HBB* subclone extending from a linker-generated *Xba*I site at +20 to a native *Xba*I site at +3285 bp (1679 bp after the polyadenylation site) was then linked to 1.7DES at its +30 position by insertion into the *Xba*I site

within the pBluescript polylinker to produce the construct 1.7DESβ. The 5′ flanking *DES* sequences of 1.7DESβ were then extended by insertion of a 16.9-kb *Xho*I fragment from −1.7 to −18.6 kb 5′ of *DES*, by insertion into the pBluescript polylinker *Sal*I site, to give the construct 18.6DESβ.

Generation of transgenic mice

DNA fragments for microinjection to generate transgenic mice from the 22DES and 4.6DES cosmid genomic clones were excised by digestion with *Nru*I, which cuts within the pTCF vector backbone either side of the *Bam*HI genomic DNA insert cloning site [56]. The 18.6DESβ construct was separated from pBluescript by digestion with *Xho*I and *Not*I, which cut within the vector polylinker at the 5′ and 3′ ends of the insert, respectively. Genomic insert was isolated from vector sequences by ultracentrifugation fractionation through NaCl gradients and used to generate transgenic mice by standard procedures as previously described [57]. Transgene copy number and integrity was determined by Southern blot analysis as previously described [20].

Determining the site of transgene integration in mice

Fibroblast cells from transgenic mice were propagated from ear biopsies in medium consisting of DMEM (Invitrogen Ltd., Paisley, UK) supplemented with 4500 mg/L L-ananyl-L-glutamine, 4.5 μg/μl glucose, 10 μg/ml each penicillin and streptomycin, and 20% (v/v) heat-inactivated fetal calf serum (PAA Laboratories, Linz, Austria). These cells were then subjected to double-fluorescence in situ hybridization using transgene and γ-satellite repeat-specific probes as previously described [5], to assess if any integration events had taken place within a centromeric region of a mouse chromosome.

DNase I hypersensitive site analysis

Pure skeletal myoblasts from transgenic mice carrying a single copy of a fully functional 240-kb PAC clone (240DES) spanning the *DES* locus [20] were used as the starting material for this analysis and were produced as follows. The 240DES transgenic mouse line DN28 [20] was bred to homozygosity and crossed to homozygous ImmortoMouse animals [21]. Clones of skeletal muscle myoblasts from the resultant double transgenic line were then isolated, maintained, and induced to undergo terminal differentiation to fused myotubes as previously described [22]. Nuclei isolated from fused myotube cultures were then isolated and subjected to DNase I HS site mapping as before [23,24]. Genomic DNA from the DNase I-treated nuclei was digested with *Xho*I and analyzed by Southern blotting. The presence of DNase I HS sites was assessed within a 16.9-kb region extending from −1.7 to −18.6 kb from the *DES* transcriptional start site. Blots were sequentially probed with a 5′ 626-bp *Xho*I–*Sac*I fragment (position −18.6 to −17.974 kb) and a 3′ 940-bp *Pst*I fragment (position −2163 to −3103 bp). The same procedure was applied to nuclei isolated from the human myelogenous leukemia cell line K562 [26], which acted as a nonmuscle negative control.

Analysis of transgene expression in mice

Total RNA from muscle and nonmuscle mouse tissues was prepared using TRIzol Reagent (Invitrogen Ltd.) in accordance with the manufacturer's instructions. Analysis of 22DES and 4.6DES transgene expression by an S1-nuclease protection assay using an end-labeled DNA probe was as previously described [20]. Expression analysis of the 18.6DESβ transgene was also by an S1-nuclease protection assay using a 3′ *Eco*RI–*Pst*I *HBB* probe [40]. All samples were double probed for endogenous murine desmin mRNA and results expressed as a percentage of murine desmin on a transgene copy number basis [20]. Results were quantified by phosphorimager analysis using a Typhoon 9200 instrument (Amersham Biosciences UK Ltd., Chalfont St. Giles, Buckinghamshire, UK).

Bioinformatics analysis of the DES locus

BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) searches using human desmin cDNA sequences (BC32116) identified a 183,625-bp BAC clone, AC053503, which contained the complete *DES* gene. The organization of the *DES* locus, that is, gene order and direction of transcription, was ascertained using the NIX program of the former UK Medical Research Council HGMP facilities, the Human Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway), NCBI Map Viewer (http://www.ncbi.nlm.nih.gov/mapview/), Ensembl Genome Browser (http://www.ensembl.org/), and former Celera public databases. BLAST analysis using appropriate cDNA sequences of mouse (BC031760) and rat (BC061872) desmin identified BAC clones AC114651 (172,694 bp) and AC121633 (266,167 bp) that spanned the mouse and rat *Des* loci, respectively. NCBI Map Viewer, Human Genome Browser, and Ensembl Genome Browser were used to confirm gene order and direction of transcription at these rodent *Des* loci. ECR Browser [32] (http://ecrbrowser.dcode.org/) was used to conduct a multiple vertebrate comparative genome analysis between human and mouse, rat, canine, opossum, chicken, and various fish (tetraodon, *Fugu*, zebrafish) species. Sliding windows of 30 bp throughout the alignment with a minimum identity of 80% were the parameters employed.

Identification of regions of homology between the human *DES* 5′ flanking region and the equivalent area from mouse and rat was determined using a number of sequence comparison programs. The region starting from within exon I and extending for 20 kb 5′ of the desmin gene transcriptional start site was obtained from the human (AC053503), mouse (AC114651), and rat (AC121633) BAC clones and initially scrutinized for areas of homology by dot-plot analysis using the Jemboss software package, which was available from the former HGMP. This was followed by analysis employing the mVISTA program (http://www-gsd.lbl.gov/vista/) with a window size set at 50 bp and a high homology threshold of 75%. The precise location of regions of human–rodent homology was identified by DNA sequence comparison of the same 20-kb upstream regions employing BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) followed by CLUSTALW (http://www.ebi.ac.uk/clustalw/) to obtain a multisequence alignment.

Putative transcription factor binding sites within the 5′ *DES* regions of human–rodent homology were identified using a range of prediction programs. The option allowing analysis of orthologous pairs of sequences within ConSite (http://www.phylofoot.org/; http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite) was used to scrutinize aligned human and mouse regions of homology. This was chosen so that only those potential transcription factor binding sites that are conserved between human and rodent species would be highlighted. The presence of AP-1 and GATA-4 binding sites was assessed using PATCH (http://www.gene-regulation.com/cgi-bin/pub/programs/patch/bin/patch.cgi). All transcription factor binding sites indicated were confirmed against the databases held in ConSite, PATCH, TFSEARCH (http://www.cbrc.jp/research/db/TFSEARCH.html), and rVista (http://www-gsd.lbl.gov/vista/rDetails.html).

Human muscle multiple-tissue Northern blot analysis

A human multiple-tissue Northern blot containing poly(A)$^+$ RNA from skeletal muscle, uterus, colon, small intestine, bladder, heart, stomach, and prostate tissues was hybridized using standard conditions in accordance with the manufacturer's instructions (BD Biosciences Clontech, Palo Alto, CA, USA). The hybridization probe was generated by PCR amplification using forward primer 5′-CCAGCTCTCAGTGGCCATGCAGA-3′, reverse primer 5′-ACCTCCAGCACTGTGAGCACCG-3′, and cosmid clone 4.6DES as the DNA template. Standard PCR conditions with an annealing temperature of 60.5°C were employed. The amplification product is 400 bp in size and corresponds to a region extending from −47 bp to the end of the predicted first exon of SPEGβ mRNA (see Supplementary Fig. 1), which resides 8241–8641 bp 3′ of *DES* (Fig. 3A) and is located at position 92,766–93,166 on BAC clone AC053503.

Acknowledgments

We are grateful to Samantha Routledge (Department of Medical and Molecular Genetics, King's College London) for constructive comments on the manuscript. This work was funded by the Muscular Dystrophy Campaign, UK, and in part by the NWO, The Netherlands.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ygeno.2006.01.009.

References

[1] L.D. Hurst, C. Pal, M.J. Lercher, The evolutionary dynamics of eukaryotic gene order, Nat. Rev. Genet. 5 (2004) 299–310.

[2] N. Dillon, P. Sabbattini, Functional gene expression domains: defining the functional unit of eukaryotic gene regulation, Bioessays 22 (2000) 657–665.

[3] W. de Laat, F. Grosveld, Spatial organization of gene expression: the active chromatin hub, Chromosome Res. 11 (2003) 447–459.

[4] Q. Li, K.R. Peterson, X. Fang, G. Stamatoyannopoulos, Locus control regions, Blood 100 (2002) 3077–3086.

[5] M. Antoniou, L. Harland, T. Mustoe, S. Williams, J. Holdstock, E. Yague, T. Mulcahy, M. Griffiths, S. Edwards, P.A. Ioannou, A. Mountain, R. Crombie, Transgenes encompassing dual-promoter CpG islands from the human TBP and HNRPA2B1 loci are resistant to heterochromatin-mediated silencing, Genomics 82 (2003) 269–279.

[6] D. Kioussis, R. Festenstein, Locus control regions: overcoming heterochromatin-induced gene inactivation in mammals, Curr. Opin. Genet. Dev. 7 (1997) 614–619.

[7] C. Bonifer, U. Jagle, M.C. Huber, The chicken lysozyme locus as a paradigm for the complex developmental regulation of eukaryotic gene loci, J. Biol. Chem. 272 (1997) 26075–26078.

[8] Y. Wang, J.P. Macke, S.L. Merbs, D.J. Zack, B. Klaunberg, J. Bennett, J. Gearhart, J. Nathans, A locus control region adjacent to the human red and green visual pigment genes, Neuron 9 (1992) 429–440.

[9] S. Carson, M.V. Wiles, Far upstream regions of class II MHC Ea are necessary for position-independent, copy-dependent expression of Ea transgene, Nucleic Acids Res. 21 (1993) 2065–2072.

[10] L. Montoliu, T. Umland, G. Schutz, A locus control region at −12 kb of the tyrosinase gene, EMBO J. 15 (1996) 6026–6034.

[11] Y. Su, S.A. Liebhaber, N.E. Cooke, The human growth hormone gene cluster locus control region supports position-independent pituitary- and placenta-specific expression in the transgenic mouse, J. Biol. Chem. 275 (2000) 7902–7909.

[12] E.W. Baxter, W.J. Cummings, R.E.K. Fournier, Formation of a large, complex domain of histone hyperacetylation at human 14q32.1 requires the serpin locus control region, Nucleic Acids Res. 33 (2005) 3313–3322.

[13] G.R. Lee, P.E. Fields, I. Griffin, J. Thomas, R.A. Flavell, Regulation of the Th2 cytokine locus by a locus control region, Immunity 19 (2003) 145–153.

[14] D.R. Greaves, F. Wilson, G. Lang, D. Kioussis, Human CD2 3′-flanking sequences confer high-level, T cell-specific, position-independent gene expression in transgenic mice, Cell 56 (1989) 979–986.

[15] C. Chauveau, E.A. Jansson, S. Muller, M. Cogne, S. Pettersson, Cutting edge: Ig heavy chain 3′ HS1-4 directs correct spatial position-independent expression of a linked transgene to B lineage cells, J. Immunol. 163 (1999) 4637–4641.

[16] L. Madisen, M. Groudine, Identification of a locus control region in the immunoglobulin heavy-chain locus that deregulates c-myc expression in plasmacytoma and Burkitt's lymphoma cells, Genes Dev. 8 (1994) 2212–2226.

[17] Q. Dang, D. Walker, S. Taylor, C. Allan, P. Chin, J. Fan, J. Taylor, Structure of the hepatic control region of the human apolipoprotein E/C-I gene locus, J. Biol. Chem. 270 (1995) 22577–22585.

[18] B.D. Ortiz, D. Cado, V. Chen, P.W. Diaz, A. Winoto, Adjacent DNA elements dominantly restrict the ubiquitous activity of a novel chromatin-opening region to specific tissues, EMBO J. 16 (1997) 5037–5045.

[19] P. Sabbattini, A. Georgiou, C. Sinclair, N. Dillon, Analysis of mice with single and multiple copies of transgenes reveals a novel arrangement for the λ5-V_{preB1} locus control region, Mol. Cell. Biol. 19 (1999) 671–679.

[20] S. Raguz, C. Hobbs, E. Yague, P.A. Ioannou, F.S. Walsh, M. Antoniou, Muscle-specific locus control region activity associated with the human desmin gene, Dev. Biol. 201 (1998) 26–42.

[21] P.S. Jat, M. Noble, P. Ataliotis, Y. Tanaka, N. Yannoutsos, L. Larsen, D. Kioussis, Direct derivation of conditionally immortal cell lines from an H-2Kb- tsA58 transgenic mouse, Proc. Natl. Acad. Sci. USA 88 (1991) 5096–5100.

[22] J.E. Morgan, J.R. Beauchamp, C.N. Pagel, M. Peckham, P. Ataliotis, P.S. Jat, M.D. Noble, K. Farmer, T.A. Partridge, Myogenic cell lines derived from transgenic mice carrying a thermolabile T antigen: a model system for the derivation of tissue-specific and mutation-specific cell lines, Dev. Biol. 162 (1994) 486–498.

[23] C.-M. Chow, A. Athanassiadou, S. Raguz, L. Psiouri, L. Harland, M. Malik, M.A. Aitken, F. Grosveld, M. Antoniou, LCR-mediated, long-term tissue-specific gene expression within replicating episomal plasmid and cosmid vectors, Gene Ther. 9 (2002) 327–336.

[24] L. Harland, R. Crombie, S. Anson, J. deBoer, P.A. Ioannou, M. Antoniou, Transcriptional regulation of the human TATA binding protein gene, Genomics 79 (2002) 479–482.

[25] Z. Li, D. Paulin, Different factors interact with myoblast-specific and myotube-specific enhancer regions of the human desmin gene, J. Biol. Chem. 268 (1993) 10403–10415.

[26] C.G. Gahmberg, L.C. Andersson, K562—A human leukemia cell line with erythroid features, Semin. Hematol. 18 (1983) 72.

[27] S. Wilk, E. Wilk, R.P. Magnusson, Purification, characterization, and cloning of a cytosolic aspartyl aminopeptidase, J. Biol. Chem. 273 (1998) 15961–15970.

[28] C.-M. Hsieh, S.-F. Yet, M.D. Layne, M. Watanabe, A.M. Hong, M.A. Perrella, M.-E. Lee, Genomic cloning and promoter analysis of aortic preferentially expressed gene-1: identification of a vascular smooth muscle-specific promoter mediated by an E box motif, J. Biol. Chem. 274 (1999) 14344–14351.

[29] R. Kikuno, T. Nagase, M. Nakayama, H. Koga, N. Okazaki, D. Nakajima, O. Ohara, HUGE: a database for human KIAA proteins, a 2004 update integrating HUGEppi and ROUGE, Nucleic Acids Res. 32 (2004) D502–D504.

[30] C.-M. Hsieh, S. Fukumoto, M.D. Layne, K. Maemura, H. Charles, A. Patel, M.A. Perrella, M.-E. Lee, Striated muscle preferentially expressed genes α and β are two serine/threonine protein kinases derived from the same gene as the aortic preferentially expressed gene-1, J. Biol. Chem. 275 (2000) 36966–36973.

[31] S.B. Sutter, M.O. Raeker, A.B. Borisov, M.W. Russell, Orthologous relationship of obscurin and Unc-89: phylogeny of a novel family of tandem myosin light chain kinases, Dev. Genes Evol. 214 (2004) 352–359.

[32] I. Ovcharenko, M.A. Nobrega, G.G. Loots, L. Stubbs, ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes, Nucleic Acids Res. 32 (2004) W280–W286.

[33] J. Flint, C. Tufarelli, J. Peden, K. Clark, R.J. Daniels, R. Hardison, W. Miller, S. Philipsen, K.C. Tan-Un, T. McMorrow, J. Frampton, B.P. Alter, A.-M. Frischauf, D.R. Higgs, Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the α-globin cluster, Hum. Mol. Genet. 10 (2001) 371–382.

[34] R.C. Hardison, J. Oeltjen, W. Miller, Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome, Genome Res. 7 (1997) 959–966.

[35] G.G. Loots, R.M. Locksley, C.M. Blankespoor, Z.E. Wang, W. Miller, E. M. Rubin, K.A. Frazer, Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons, Science 288 (2000) 136–140.

[36] J.C. Oeltjen, T.M. Malley, D.M. Muzny, W. Miller, R.A. Gibbs, J.W. Belmont, Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains, Genome Res. 7 (1997) 315–329.

[37] H. Li, Y. Capetanaki, An E box in the desmin promoter cooperates with the E box and MEF-2 sites of a distal enhancer to direct muscle-specific transcription, EMBO J. 13 (1994) 3580–3589.

[38] M. Mericskay, A. Parlakian, A. Porteu, F. Dandre, J. Bonnet, D. Paulin, Z. Li, An overlapping CArG/octamer element is required for regulation of

desmin gene transcription in arterial smooth muscle cells, Dev. Biol. 226 (2000) 192–208.
[39] W.W. Wasserman, J.W. Fickett, Identification of regulatory regions which confer muscle-specific gene expression, J. Mol. Biol. 278 (1998) 167–181.
[40] M. Antoniou, F. Geraghty, J. Hurst, F. Grosveld, Efficient 3′-end formation of human β-globin mRNA in vivo requires sequences within the last intron but occurs independently of the splicing reaction, Nucleic Acids Res. 26 (1998) 721–729.
[41] W.R. Shehee, P. Oliver, O. Smithies, Lethal thalassemia after insertional disruption of the mouse major adult β-globin gene, Proc. Natl. Acad. Sci. USA 90 (1993) 3177–3181.
[42] A. Auerbach, Production of functional transgenic mice by DNA pronuclear microinjection, Acta Biochim. Pol. 51 (2004) 9–31.
[43] R.D. Palmiter, R.L. Brinster, Germ-line transformation of mice, Annu. Rev. Genet. 20 (1986) 465–499.
[44] S. Knotts, H. Rindt, J. Robbins, Position independent expression and developmental regulation is directed by the β myosin heavy chain gene's 5′ upstream region in transgenic mice: position independent expression and developmental regulation is directed by the b myosin heavy chain gene's 5′ upstream region in transgenic mice, Nucleic Acids Res. 23 (1995) 3301–3309.
[45] M. Salminen, P. Maire, J.-P. Concordet, C. Moch, A. Porteu, A. Kahn, D. Daegelen, Fast-muscle-specific expression of human aldolase A transgenes, Mol. Cell. Biol. 14 (1994) 6797–6808.
[46] K. Hashido, Y. Arai, S. Kajihara, K. Joh, H. Yatsuki, K. Hori, J. Miyazaki, K. Yamamura, T. Mukai, Copy-dependent and position-independent expression of rat aldolase A gene, J. Biochem. 118 (1995) 601–606.
[47] G.P. Patrinos, M. de Krom, E. de Boer, A. Langeveld, A.M.A. Imam, J. Strouboulis, W. de Laat, F.G. Grosveld, Multiple interactions between regulatory regions are required to stabilize an active chromatin hub, Genes Dev. 18 (2004) 1495–1509.
[48] P. Hatzis, I. Talianidis, Dynamics of enhancer–promoter communication during differentiation-induced gene activation, Mol. Cell 10 (2002) 1467–1477.
[49] G. Blom van Assendelft, O. Hanscombe, F. Grosveld, D.R. Greaves, The β-globin dominant control region activates homologous and heterologous promoters in a tissue-specific manner, Cell 56 (1989) 969–977.
[50] P. Collis, M. Antoniou, F. Grosveld, Definition of the minimal requirements within the human β-globin gene and the dominant control region for high level expression, EMBO J. 9 (1990) 233–240.
[51] M. Antoniou, F. Grosveld, β-Globin dominant control region interacts differently with distal and proximal promoter elements, Genes Dev. 4 (1990) 1007–1013.
[52] D.R. Greaves, F.D. Wilson, G. Lang, D. Kioussis, Human CD2 3′-flanking sequences confer high-level, T cell-specific, position-independent gene expression in transgenic mice, Cell 56 (1989) 979–986.
[53] J. Hadchouel, J.J. Carvajal, P. Daubas, L. Bajard, T. Chang, D. Rocancourt, D. Cox, D. Summerbell, S. Tajbakhsh, P.W. Rigby, M. Buckingham, Analysis of a key regulatory region upstream of the Myf5 gene reveals multiple phases of myogenesis, orchestrated at each site by a combination of elements dispersed throughout the locus, Development 130 (2003) 3415–3426.
[54] C. May, S. Rivella, J. Callegari, G. Heller, K.M. Gaensler, L. Luzzatto, M. Sadelain, Therapeutic haemoglobin synthesis in β-thalassaemic mice expressing lentivirus-encoded human β-globin, Nature 406 (2000) 82–86.
[55] S. Indraccolo, S. Minuzzo, F. Roccaforte, R. Zamarchi, W. Habeler, L. Stievano, V. Tosello, D. Klein, W.H. Gunzburg, G. Basso, L. Chieco-Bianchi, A. Amadori, Effects of CD2 locus control region sequences on gene expression by retroviral and lentiviral vectors, Blood 98 (2001) 3607–3617.
[56] F.G. Grosveld, T. Lund, E.J. Murray, A.L. Mellor, H.H. Dahl, R.A. Flavell, The construction of cosmid libraries which can be used to transform eukaryotic cells, Nucleic Acids Res. 10 (1982) 6715–6732.
[57] N. Dillon, F. Grosveld, Transcriptional analysis using transgenic animals, in: B.D. Hames, S.J. Higgins (Eds.), Gene transcription: a practical approach, IRL Press, Oxford, UK, 1993, pp. 153–188.