ARTICLE

Evolutionary Conservation of a Coding Function for D4Z4, the Tandem DNA Repeat Mutated in Facioscapulohumeral Muscular Dystrophy

Jannine Clapp, Laura M. Mitchell, Daniel J. Bolland, Judy Fantes, Anne E. Corcoran, Paul J. Scotting, John A. L. Armour, and Jane E. Hewitt

Facioscapulohumeral muscular dystrophy (FSHD) is caused by deletions within the polymorphic DNA tandem array D4Z4. Each D4Z4 repeat unit has an open reading frame (ORF), termed "*DUX4,*" containing two homeobox sequences. Because there has been no evidence of a transcript from the array, these deletions are thought to cause FSHD by a position effect on other genes. Here, we identify D4Z4 homologues in the genomes of rodents, Afrotheria (superorder of elephants and related species), and other species and show that the *DUX4* ORF is conserved. Phylogenetic analysis suggests that primate and Afrotherian D4Z4 arrays are orthologous and originated from a retrotransposed copy of an intron-containing *DUX* gene, *DUXC*. Reverse-transcriptase polymerase chain reaction and RNA fluorescence and tissue in situ hybridization data indicate transcription of the mouse array. Together with the conservation of the *DUX4* ORF for >100 million years, this strongly supports a coding function for D4Z4 and necessitates re-examination of current models of the FSHD disease mechanism.

Facioscapulohumeral muscular dystrophy (FSHD [MIM 158900]) is the third most common muscular dystrophy in whites, with an autosomal dominant pattern of inheritance and an incidence of ~1 in 20,000.[1] The disease is caused by a unique and enigmatic mechanism; almost all cases of FSHD are associated with contractions within a DNA tandem array (D4Z4) that is located in the subtelomeric region of chromosome 4q35.[2,3] The number of 3.3-kb repeat units within this array is highly polymorphic, varying between 11 and 100 in unaffected individuals. In patients with FSHD, one allele has ⩽10 repeats.[4] However, loss of a complete D4Z4 array on one allele does not result in the disease, suggesting that the repeats themselves play a role in FSHD. A similar tandem array is located on chromosome 10q26 and has 98% nucleotide identity to D4Z4.[5,6] Although this chromosome 10q locus is also polymorphic, short arrays are not associated with FSHD.[4] The 4q telomere has two variants, termed "4qA" and "4qB."[6] However, D4Z4 deletions result in FSHD only when they occur on a 4qA chromosome.[7]

The underlying mechanism whereby these deletions cause FSHD is still unclear.[4] D4Z4 repeats contain two dispersed repeat elements (LSau and hhspm3) that are characteristic of heterochromatic regions of the human genome.[8,9] One widely held view is that D4Z4 has a noncoding, regulatory function and plays a role in the formation or maintenance of heterochromatin at the 4q telomere, repressing the expression of genes within chromosome 4q35. In this epigenetic model, contractions of the array below a threshold number of repeats alters the local chromatin organization, resulting in loss of repression of one or more nearby genes.[4] Because the D4Z4 repeat unit contains a substantial ORF (*DUX4*) with the potential to encode a homeodomain protein,[8,9] an alternative mechanism has been proposed in which the FSHD deletions perturb the expression of this putative homeobox gene.[10] However, no transcript from the array has been identified; consequently, D4Z4 is generally considered to represent an accumulation of pseudogenes and to have only a noncoding function.

Although there is some experimental evidence to support the chromatin hypothesis for FSHD, it remains unproven. This region of 4q is not decorated by histone modifications that are characteristic of heterochromatin.[11] Although chromosome 4q35 is spatially associated with two heterochromatic domains within the nucleus, the nuclear envelope, and the nucleolar region, there is no difference in 4qter nuclear localization between FSHD cells and controls.[12,13] D4Z4 has a high GC content (71%) and is highly methylated.[9] Partial hypomethylation of the D4Z4 array has been reported in patients with FSHD,[14,15] but patients with immunodeficiency, centromeric instability, and facial anomalies (ICF) syndrome (who have mutations in the DNA methyltransferase gene *DNMT3B*) show extensive hypomethylation of D4Z4 but no muscular dystrophy symptoms,[16] arguing against a causal role for the methylation status of the repeat in FSHD.

Gabellini et al.[17] identified a repressor protein complex

From the Institute of Genetics, School of Biology, The University of Nottingham, Queen's Medical Centre, Nottingham, United Kingdom (J.C.; L.M.M.; P.J.S.; J.A.L.A.; J.E.H.); Laboratory of Chromatin and Gene Expression, Babraham Institute, Cambridge, United Kingdom (D.J.B.; A.E.C.); and Medical Research Council Human Genetics Unit, Western General Hospital, Edinburgh (J.F.)

Received February 26, 2007; accepted for publication May 1, 2007; electronically published June 27, 2007.

Address for correspondence and reprints: Prof. Jane Hewitt, Institute of Genetics, School of Biology, The University of Nottingham, Queen's Medical Centre, Nottingham, NG7 2UH, United Kingdom. E-mail: jane.hewitt@nottingham.ac.uk

Am. J. Hum. Genet. 2007;81:264–279. © 2007 by The American Society of Human Genetics. All rights reserved. 0002-9297/2007/8102-0008$15.00
DOI: 10.1086/519311

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics*.

Figure 1. ClustalW alignment of ape D4Z4 repeats. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics*.

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics*.

Figure 2. Clustal alignments of primate D4Z4 sequences. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics*.

bound to D4Z4 and postulated a *cis*-acting model in which loss of repeats from D4Z4 results in a decrease in the amount of repressor complex that is bound and a concomitant loss of transcriptional repression of chromosome 4q35 genes. RT-PCR data in that study showed increased expression of three chromosome 4q35 genes (*FRG1, FRG2,* and *ANT1*) in FSHD muscle samples.[17] However, this gene-expression data has not been reproduced by other studies using microarray and quantitative and allelic RT-PCR approaches.[18–21] Although transgenic mice expressing very high levels of human *FRG1* in skeletal muscle do develop a muscular-dystrophy phenotype,[22] the lack of robust data supporting up-regulation of this gene in patients means that the relationship between *FRG1* and FSHD is unclear.

Little is known about D4Z4 sequences in other organisms. In original comparative studies, D4Z4 homologues were identified only in higher primates, and the DNA sequences of these loci were not determined.[23,24] Physical-mapping data showed that genomes of great apes (chimpanzee, gorilla, and orangutan) all have D4Z4-related arrays at orthologous chromosomal locations.[23,24] In addition, these species also contain many related, dispersed 3.3-kb repeats; as in humans, these are found primarily at heterochromatic locations, such as on acrocentric chromosomes.[23,24] Old and New World monkeys also contain two D4Z4-like arrays, equivalent to the human 4qter and 10qter loci, but appear to lack significant numbers of the dispersed repeats.[23,24]

Here, we have taken advantage of the extensive DNA sequence data from whole-genome projects to re-examine the extent of D4Z4 evolutionary conservation. We have identified D4Z4 homologues in several mammalian species, allowing us to infer the evolutionary history of this locus and to identify a protein-coding function for the repeat.

Material and Methods

Identification and Analysis of Genomic Sequence

To identify sequences with similarity to D4Z4, the human repeat sequence was used to search the National Center for Biotechnology Information Trace Archive by use of discontinuous megablast (BLAST). Trace files for each species were downloaded and then were assembled and manually edited using Sequencher (Genecodes). Reiterative rounds of searching were then used to identify all matching traces within the appropriate archive. Interspersed DNA repeat elements were identified using RepeatMasker. Species for which sequence data were assembled were chimpanzee (*Pan troglodytes*), orangutan (*Pongo pygmaeus pygmaeus*), rhesus macaque (*Macaca mulatta*), white tufted-ear marmoset (*Callithrix jacchus*), tree shrew (*Tupaia belangeri*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), tenrec (*Echinops telfairi*), hyrax (*Procavia capensis*), and African elephant (*Loxodonta africana*). Apes have two D4Z4-related arrays, corresponding to the human 4q35 and 10q26 loci.[21] In our sequence assemblies, lack of linkage information meant that it was not possible to assign individual sequences to a specific paralogue; hence, the consensus sequences are presumably derived from a mixture of the two loci. A search of several Laurasiatherian (dog, cat, and cow) or nonmammalian (zebrafish and chicken) genome sequences failed to identify D4Z4 homologues in these species. Intron-containing *DUX* homologues were identified by blastx (BLAST) searches of genome assemblies in the Ensembl database. Because homeodomains are generally highly conserved, sequences were assigned as *DUX* genes only if they met at least one of the following criteria: conservation of synteny with the human genes, a putative mRNA encoding two DUX-like homeodomains, or exon organization matching known *DUX* genes.

Statistical and Phylogenetic Analysis

Assembled nucleotide sequences from primate D4Z4 repeats were aligned using ClustalW.[25] These alignments are shown in figures 1 and 2. Pairwise divergence between aligned ORF and non-ORF sequences from human, chimpanzee, and orangutan was assessed with a 2×2 χ^2 test for heterogeneity. Neutrality of aligned ORF sequences from human, chimpanzee, orangutan, macaque, and marmoset was tested using the codeml option in PAML, version 3.15,[26] by comparing the likelihoods of model 0 (single ω value) versus model 1 (nearly neutral: two ω values) and model 2 (positive selection: three ω values, allowing one to be >1).

For phylogenetic analysis, the DUX homeodomain amino acid sequences were aligned using ClustalX.[27] Phylogenetic trees were computed with the maximum-likelihood method by use of PROML in the PHYLIP package[28] (version 3.6), with the Jones-Taylor-Thornton model of amino acid substitution. Output trees were drawn using Phylodendron. To evaluate the confidence of the maximum-likelihood tree, PROML analysis (jumble = 10) was performed using a bootstrap set of 100 pseudoalignments generated by SEQBOOT, and the consensus tree was computed with CONSENSE. Bootstrap values for equivalent nodes were placed onto a representative maximum-likelihood tree in figure 3.

Metaphase FISH

Probe labeling, DNA hybridization, and antibody detection were performed as described elsewhere.[29] Chromosome preparations were made from C5BL/6J mouse spleens by use of standard methods. A cosmid (cosmid 6) and a plasmid insert (Dux_4) were mapped to metaphase spreads, by use of inverted 4′,6-diamidino-2-phenylindole (DAPI) staining, to obtain a G-banded pattern for chromosome identification. The localization was confirmed by

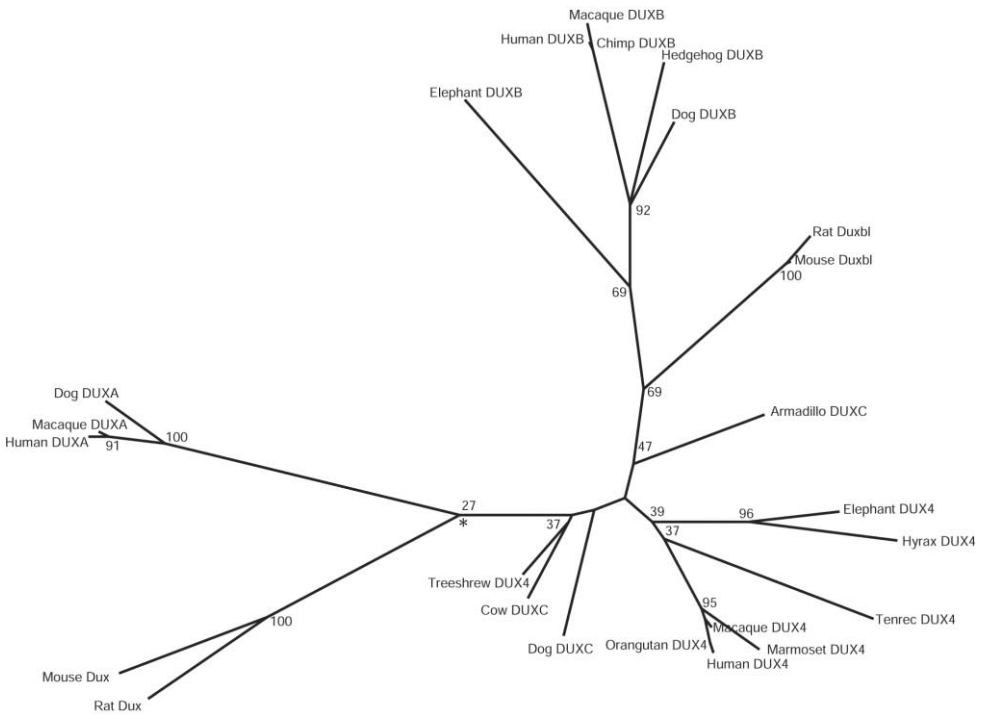

Figure 3. Phylogenetic relationships of DUX proteins. The unrooted maximum-likelihood tree was generated from the aligned, concatenated homeodomain amino acid sequences by use of the PHYLIP package. Branches are scaled according to evolutionary distance. Numbers at nodes represent bootstrap values. The asterisk (*) indicates the location of the rodent Dux node. For both DUXA and DUX4, the chimpanzee and human proteins are identical across the homeodomains; therefore, only the human orthologues are included in the tree. Similar tree topologies were generated using UPGMA and neighbor-joining methods (data not shown).

dual hybridization with a biotin-labeled mouse chromosome 10–specific paint (Cambio).

Physical Mapping

DNA preparation, digestion, electrophoresis, transfer, and hybridization were performed using standard protocols.[30] For Southern analysis, genomic DNA (10 μg) was digested with the restriction enzyme indicated. For pulsed-field gel electrophoresis (PFGE), high–molecular-weight DNA from C57BL/6J or CD1 mice embedded in agarose blocks was digested with *Eco*RV, and fragments were separated by electrophoresis by use of CHEF DRII (BioRad).

RT-PCR

DNase treatment was performed using TURBO DNA-*free* (Ambion), in accordance with the manufacturer's instructions, by use of 10 μg RNA and 2 U or 4 U DNase. A OneStep reaction (Qiagen) was performed in a total volume of 50 μl, containing 1 μg of DNase-treated RNA. After amplification, 15 μl of reaction mix was analyzed on a 2% agarose gel. For a "no reverse transcriptase" control, RNA was added to the reaction mix only after the heating to 95°C had inactivated the reverse-transcriptase enzyme. The exact annealing temperatures and extension times depended on the primers used and the size of the expected product. Primer sequences and reaction details are provided in table 1.

RNA FISH

Appropriate PCR products were cloned into pGEM-T Easy (Promega), for probe generation. For mouse *Dux,* this was 2,078–3,179 bp of EMBL accession number AM398151 (Dux_6). For *Gcc2,* the probe corresponded to 993–2,593 bp of GenBank accession number NM_027375. Probe generation and RNA FISH of mouse splenocytes was performed as described elsewhere.[31] Images of the DAPI stain and the fluorescein isocyanate (FITC) and Texas Red signals of the same region were taken using a black and white camera and were saved as JPEG files before being opened in Adobe Photoshop, where color was added. For FITC (green), the color settings were hue 120, saturation 100, and light −50; for Texas Red (red), the color settings were hue 360, saturation 100, and light −50; and, for DAPI (blue), the color settings were hue 240, saturation 100, and light −50. The files were then layered on top of each other to produce a composite image.

In Situ Hybridization of Tissue Sections

In situ hybridization was performed on frozen and wax sections as described by Rex and Scotting,[32] by use of the same mouse *Dux* probe (Dux_6) as in the RNA-FISH studies. Digoxigenin (DIG)–labeled probes were produced using a T7/Sp6 labeling kit (Roche).

Generation of Epitope-Tagged Proteins

The full-length mouse *Dux* ORF (2,025 bp) was cloned in-frame into the pEGFPN1 vector (Clontech) to give the construct Dux-

Table 1. Primers Used in the Mouse *Dux* RT-PCR Analysis

Primer Name	Primer Sequence (5′→3′) Forward	Reverse	Annealing Temperature (°C)	Product Size (bp)
Dux_2	GCACTCAAGCAGACAGCACA	GTGTCCATTTGCTCCCATGT	57	400
Dux_5	CAGCACATGCAGGAAGATGA	TCAGACCCCCTTCCTTGACT	57	720
Dux_7	ACTTCTAGCCCCAGCGACTC	CCATGCTGCCAGGATTTCTA	57	212
Dux_10	GCCCACAGCTCAAGATCAAG	ATCAAGGAGGGGTTCCAGAG	59	210
Dux_13	CCAACCTTCTGCAGAGAATCC	CACTTGGGAAGGCTTGGACT	57	309
Dux_3	GAATGGGGGTCTCAGATTGC	TGCCTGTACTTCCTGCTTCTTC	57	489
Dux_ATG1	TTTAAGGGGCAGTGGTCACA	CCAGCTCCTTCCTCTCCTTG	59	310
Dux_ATG2	AGTCGATTCTCCCAAGGTGA	GGAGCCTCTGATGGACCTCT	57	273
Dux_TGA1	AACTGCTGACCGAAGTCCAA	CATTTCGGGAAGTCACTGGA	57	278
Dux_TGA2	AACTGCTGACCGAAGTCCAA	CACAGCTCTGCATGAAGCAT	57	626

EGFP. An amino-terminal enhanced green flourescence protein (EGFP)–tagged construct (EGFP-Dux) was made by cloning the ORF in-frame into the pEGFPC3 vector (Clontech). Constructs were also made that encoded either the N-terminus and the double homeodomain region or just the C-terminus. For the N-terminus and homeodomain regions, 1–566 bp of the *Dux* ORF were cloned in-frame into either pEGFPN1, to give the construct DuxHD-EGFP, or into pEGFPC3, to give the construct EGFP-DuxHD. For the C-terminal region, 580–2,025 bp of the ORF were cloned in-frame into pEGFPC1, to give the construct EGFP-CtermDux.

Cell Culture and Transfection

C2C12 myoblasts were grown and maintained in Dulbecco's modified Eagle medium supplemented with 10% fetal calf serum and were transfected using Effectene transfection reagent (Qiagen). To examine EGFP expression, cells were seeded onto glass coverslips. Forty-eight hours after transfection, cells expressing EGFP constructs were washed in PBS, were fixed with 4% paraformaldehyde in PBS, and were washed in PBS before being mounted in VectaShield mounting medium with DAPI.

Results

Codon-Based Analysis of Primate Sequences

We used megablast (BLAST) to identify *DUX4* homologous sequences in shotgun-trace databases for four primates: chimpanzee (*P. troglodytes*), orangutan (*P. pygmaeus pygmaeus*), rhesus macaque (*M. mulatta*), and white tufted-ear marmoset (*C. jacchus*). Although draft genome assemblies are available for several of these species, we chose to search the raw shotgun data directly because of the inherent problems in correctly assembling tandem arrays. For each species, the locations and orientation of mate pairs, the high density of clones, and the identification of nucleotide variants (none of which disrupt the *DUX4* ORF) indicate that the D4Z4 sequences are arranged as tandem arrays. This is consistent with previous genomic mapping data.[23] The consensus sequences have been deposited in EMBL, and a schematic diagram of each repeat unit is shown in figure 4.

D4Z4 orthologues were identified in all these primates, but only the *DUX4* homeobox ORF is conserved in all the species. Although the chimpanzee and orangutan repeats could be aligned to human D4Z4 along the whole 3.3-kb unit (fig. 1), alignment of the Old and New World monkey consensus sequences with those of apes was possible only over three regions: the ORF, a short stretch of DNA immediately upstream of the ORF, and the LSau repeat (fig. 2).

Statistical comparison of these primate nucleotide sequences suggests significant conservation of the *DUX4* ORF. There is significantly less substitutional divergence between aligned human and chimpanzee sequences in the ORF (1.8%) than in non-ORF positions (3.1%; $P \approx .022$). Significantly less ORF divergence is also seen between aligned human and orangutan sequences ($P \approx 4 \times 10^{-8}$). The 3.1% human-chimpanzee non-ORF divergence at D4Z4 is significantly higher than the genome average of 1.23%[34] but is not higher than the 3.5% reported at the Xp/Yp subtelomeric region,[35] which is consistent with the observed high divergence seen in subtelomeric regions.[34] A codon-based analysis of aligned primate *DUX4* ORFs rejected a null hypothesis of neutrality ($P \approx 1.3 \times 10^{-11}$) in a likelihood-ratio test against a maximum-likelihood model in which ~40% of sites are conserved by selection. Therefore, we conclude that *DUX4* has not been preserved by chance but has been conserved by selection.

Conservation of DUX4 *ORF in Nonprimate Mammals*

We next identified D4Z4 homologues in several other mammalian species: the tree shrew (*T. belangeri*), mouse, rat, tenrec (*E. telfairi*), hyrax (*P. capensis*), and African elephant (*L. africana*). The presence of D4Z4-related arrays in hyrax, tenrec, and elephant indicates an ancient origin for this repeat, since all these species are members of the Afrotheria, a deeply rooted mammalian order that diverged from other eutherian mammals >100 million years ago.[36]

Comparison of the organization of all these mammalian loci identifies two very striking properties of these D4Z4-related sequences (fig. 4). First, there is maintenance of the *DUX4* ORF in all the species. In contrast, apart from the ape sequences, there is little nucleotide similarity outside this ORF, even between relatively closely related species such as humans and Old or New World monkeys.

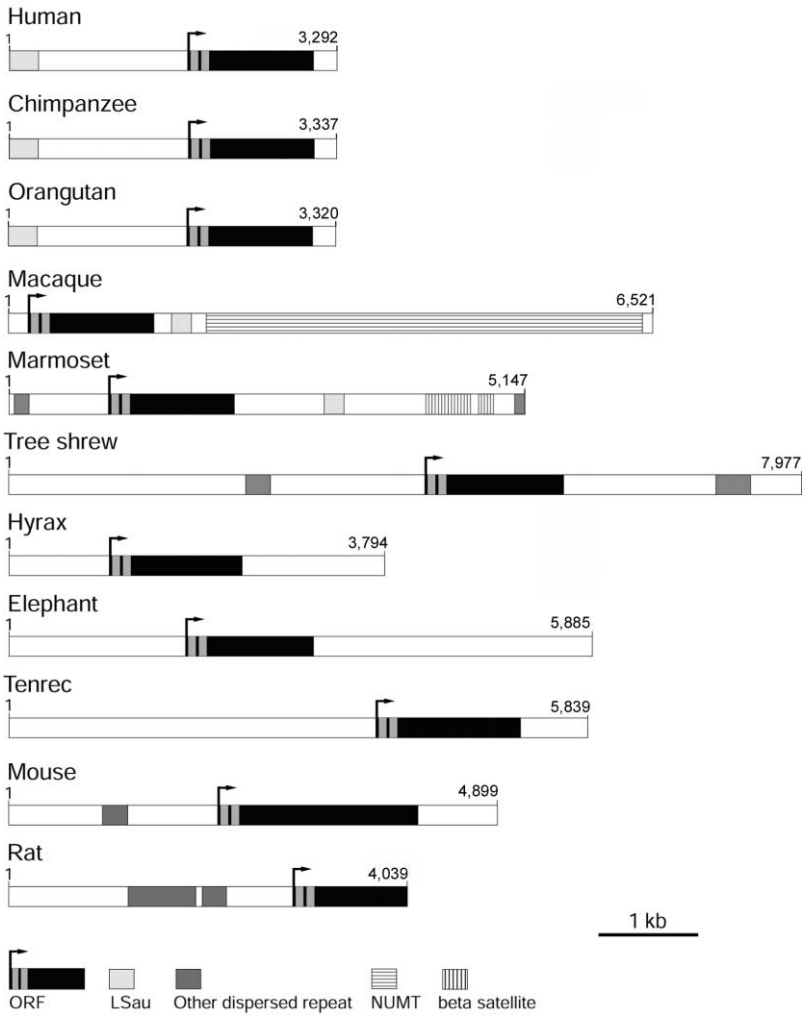

Figure 4. Schematic diagram of mammalian D4Z4–related repeat units. Repeats and sequence elements are drawn to scale. DNA repeat elements were identified using RepeatMasker. Sequence elements are shaded according to the key and are defined in the EMBL database entries. The macaque D4Z4 repeat has an insertion of 4.4 kb of mtDNA sequence that has not interrupted the *DUX4* ORF. Such nuclear DNA sequences of mitochondrial origin (NUMTs) are found in many eukaryotes.[33] Human sequence is from EMBL accession number AF117653, and rat sequence is from BAC clone CH230-14H6 (EMBL accession number AC135091). Other sequences are for chimpanzee (*P. troglodytes* [EMBL accession number BN000980]), orangutan (*P. pygmaeus* [BN000981]), rhesus macaque (*M. mulatta* [BN000983]), white tufted-ear marmoset (*C. jacchus* [BN000982]), tree shrew (*T. belangeri* [BN000984]), mouse (*M. musculus* [AM398147]), tenrec (*E. telfairi* [BN000990]), hyrax (*P. capensis* [BN000988]), and African elephant (*L. africana* [BN000989]).

Second, all the homologues are organized as multiple copies in a head-to-tail arrangement, where the repeat units in an array are almost identical within a species but differ between species. Examination of genome scaffold assemblies and physical-mapping data[23] indicates that these mammalian arrays typically contain at least 10 repeat units.

Thus, D4Z4 arrays exist in distantly related mammalian species and are not confined to primates, as suggested elsewhere.[24] The *DUX4* ORF is conserved in all these arrays, indicative of a protein-coding function. The high level of nucleotide divergence between the human and mouse homeodomain sequences presumably accounts for the failure of previous DNA-hybridization approaches to identify this locus. We did not identify any homologues outside the taxon Eutheria, so D4Z4 sequences may be confined to placental mammals.

Evolutionary Relationships of D4Z4 Arrays and Mammalian Intron-Containing DUX *Genes*

Recently, two human intron-containing *DUX* genes (*DUXA* and *DUXB*) were reported by Booth and Holland,[37] who proposed that D4Z4 may have originated from a retrotransposed copy of *DUXA*. We examined the relationship of the intronless *DUX4* ORFs in the D4Z4-related sequences with *DUXA* and *DUXB*. First, mammalian intron-containing *DUX* genes were identified by database

searching, and we then used synteny information (when available), gene organization, and phylogenetic analysis to assign these genes into four paralogous groups: *DUXA, DUXB, DUXC,* and *DUXB-like* (*Duxbl*) (table 2). No intron-containing *DUX* genes were identified in nonmammalian organisms.

Orthologues of both *DUXA* and *DUXB* are present in a wide range of mammalian species, including members of the superorders Laurasiatheria, Euarchontoglires, and Afrotheria. The dog genome contains a third intron-containing *DUX* gene (*DUXC*) in addition to *DUXA* and *DUXB*; potential *DUXC* orthologues were also found in cow and armadillo. Mouse and rat appear to lack orthologues of *DUXA, DUXB,* and *DUXC*. However, an intron-containing *DUX* gene was identified, and, on the basis of genomic location and phylogenetic analysis, we have assigned this gene as *Duxbl* (table 2).

In phylogenetic trees of >200 homeodomains, all the intron-containing and intronless DUX sequences clustered together, indicating that they represent a distinct class of homeodomain proteins (data not shown). We next performed more-extensive phylogenetic analysis of the DUX-type homeodomains. The predicted amino acid sequences of the homeodomains were aligned using ClustalX (fig. 5*a*); only those proteins for which both homeodomains were identified with certainty were included in this alignment. For each protein, the two homeodomains were then concatenated and aligned; this alignment was then used to generate phylogenetic trees. Figure 3 shows an unrooted maximum-likelihood tree, with the branch lengths scaled to indicate relative evolutionary distances. From the genomic data and this phylogenetic analysis, we infer that there has been a complex pattern of *DUX* gene duplication and loss during mammalian evolution.

The DUXA proteins form a distinct cluster within the tree, with a bootstrap value of 100%. The bootstrap support for the grouping of the mammalian DUXB proteins is also strong (69%), although the primate order is poorly resolved in this branch. Together with the synteny data, this indicates that both *DUXA* and *DUXB* genes existed in the common ancestor of all mammals. Because we could not identify either *DUXA* or *DUXB* in mouse or rat genomes, even though these have been sequenced to great depth, it is likely that both genes have been lost in the murine lineage. Although the relationship of the *Duxbl* genes to the other *DUX* loci is less clear, the murine Duxbl branch roots to the base of the DUXB tree with good bootstrap support (69%), indicating that this gene probably arose from duplication of an intron-containing *DUX* gene in a common mammalian ancestor but has subsequently been lost except in the rodent lineages.

Implications for the Evolutionary Origins of D4Z4

Our data have allowed us, for the first time, to postulate the evolutionary history of D4Z4. The primate and Afrotherian DUX4 and DUXC homeodomains cluster together in the maximum-likelihood tree (fig. 3). Apart from tree shrew, the topology of the DUX4 tree reflects that of mammalian phylogeny, indicating that the primate and Afrotherian *DUX4* sequences are orthologues that arose from amplification of a retrotransposed copy of *DUXC*. In addition to the homeodomain regions, there is a short stretch of C-terminal residues that is conserved only in DUXC, DUX4, and the murine Dux proteins (fig. 5*b*), supporting a common ancestral relationship between these genes. We found no evidence of the existence of intron-containing *DUXC* genes in primates or Afrotheria, nor could we identify any *DUX4* arrays in Laurasiatheria. Thus, we infer that the ancestral *DUX4* array and *DUXC* were both present in the common ancestor of all mammalian orders ~105 million years ago, with *DUX4* giving rise to the human D4Z4 array. The *DUX4* array has been lost from Laurasiatheria, and *DUXC* is lost in both the Afrotherian and primate lineages. Thus, D4Z4 represents a tandem array of homeodomain ORFs that has been conserved for >100 million years of evolution.

The location of the rodent Dux node (indicated by an asterisk in fig. 3) outside this DUX4/DUXC group indicates that, despite the conserved C-terminal region, the mouse and rat *Dux* arrays probably originated from an independent amplification of a retrotransposed *DUX*-type gene before mammalian radiation. Therefore, we have designated the rodent array sequences as *Dux,* rather than *DUX4,* to indicate that their relationship is likely to be paralogous, rather than orthologous. This model implies that, within the genomes examined in this study, this *Dux* array was retained only in rodent lineages. The long branch lengths indicate that the murine *Dux* sequences are evolving rapidly. Thus, whatever their evolutionary origin, these mouse and rat loci are the only *DUX* arrays in these species, they are the most closely related rodent sequences to human *DUX4,* and they share a conserved C-terminal region with DUX4 proteins.

There are no naturally occurring animal models of FSHD, and this has severely limited mechanistic studies of the disease. The identification of a potential homologue of D4Z4 in the mouse raises the possibility of engineering a mutation that might mimic the FSHD deletions. Therefore, we examined the organization of this locus in more detail and examined whether the mouse *Dux* array is transcribed.

Physical Mapping of the Mouse Dux *Array*

We identified several cosmid and BAC clones containing mouse *Dux* sequences, and mapping of these clones confirmed the tandem-array organization. Five independent

Table 2. Mammalian Intron-Containing *DUX* Genes

The table is available in its entirety in the online edition of *The American Journal of Human Genetics*.

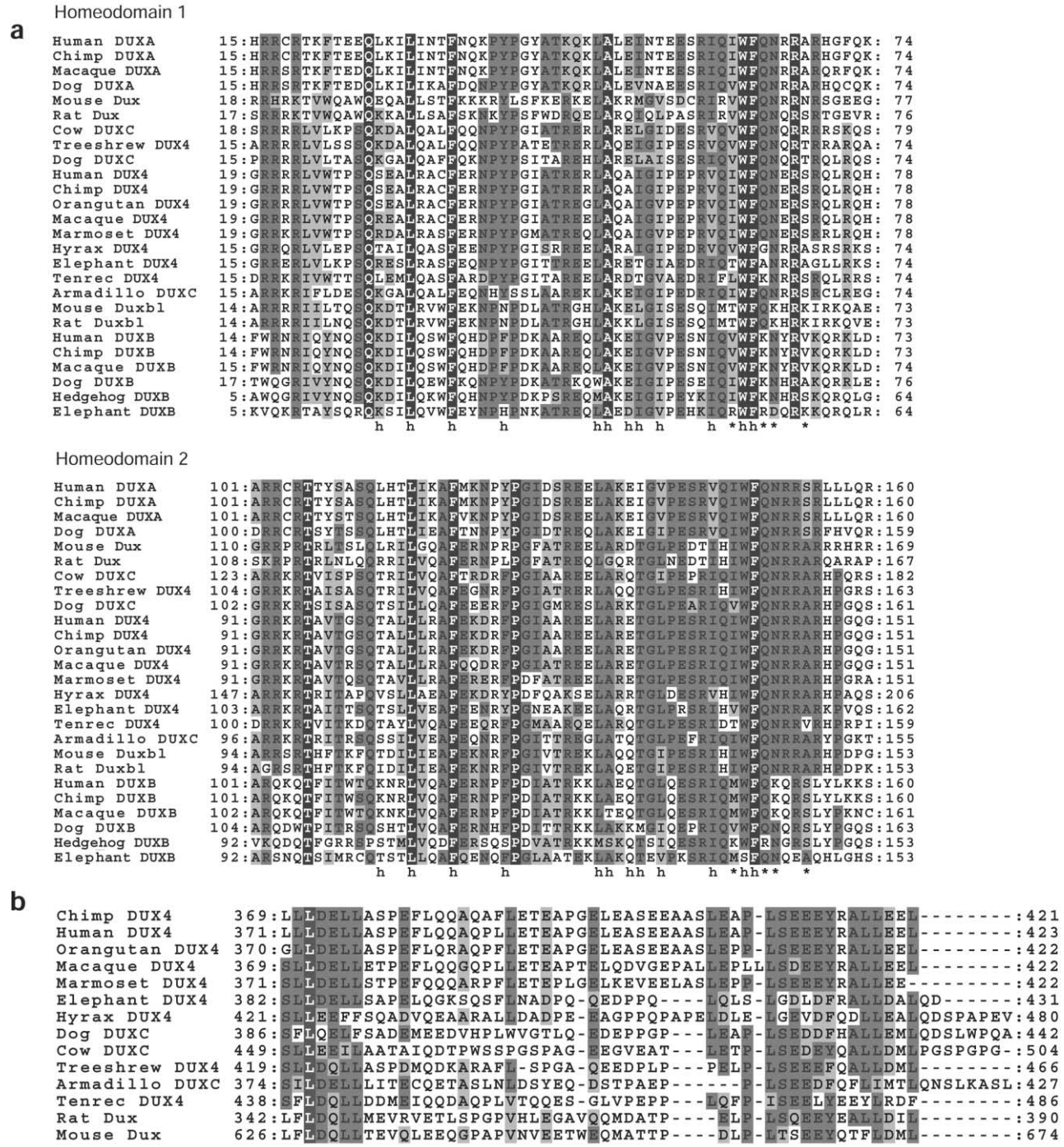

Figure 5. Clustal alignments of mammalian DUX proteins. *a,* Alignment of the two homeodomain regions. The alignment shows several invariant or highly conserved amino acids; comparison with a homeodomain consensus sequence[38] indicates that the majority of these either are hydrophobic residues associated with protein packing or are involved in DNA binding. The number of amino acid residues in the linker region between the two homeodomains varies between both paralogues and orthologues; indeed, the hyrax DUX4 protein has a repeated GQ motif that varies in copy number between D4Z4 repeats. An asterisk (*) indicates amino acid residues predicted to be involved in DNA binding, and "h" indicates residues that are involved in packing of the structure and are usually hydrophobic in homeodomains. *b,* Alignment of the C-terminal regions of DUXC, DUX4, and the mouse and rat Dux proteins. Residues that are invariant in all species are highlighted in black. Residues that are conserved in at least 60% of sequences are highlighted in dark gray, with conservative substitutions highlighted in light gray. Numbering relates to the sequence deposited in the EMBL database or transcript information from Ensembl (table 2).

repeats from one cosmid clone (cosmid 6) were subcloned and sequenced (EMBL accession numbers AM398147–AM398151). Each mouse *Dux* repeat unit is 5 kb and contains an ORF of 2 kb (fig. 6*a*). On Southern-blot analysis of genomic DNA, *Dux* probes produce strongly hybridizing bands of the expected sizes, with only a few minor bands, indicating that the majority of the *Dux* repeats match the consensus sequence (fig. 6*b*). To confirm that there is a single *Dux* locus in mice, we used FISH on metaphase chromosomes with either cosmid 6 (which contains at least five repeat units) or a plasmid probe to part of the repeat unit (Dux_4). Both probes produced a single hybridization signal on chromosome 10 (fig. 6*c*). Therefore, unlike human and higher primate genomes, the mouse genome does not contain significant numbers of divergent, dispersed *Dux* sequences. There is no synteny homology between this region of mouse chromosome 10 and the FSHD region on human chromosome 4q35; the mouse *Frg1* gene maps to chromosome 8.[39]

To examine the long-range organization of the *Dux* array, a probe to the repeat was hybridized to mouse genomic DNA digested with *Eco*RV, which does not cut within the repeat unit. Surprisingly, in C57BL/6J mice, this identified two strongly hybridizing bands of 50 kb and 90 kb and a much fainter band of 19 kb (fig. 6*d*). Whereas several inbred strains of mice gave an identical hybridization pattern (data not shown), outbred CD1 mice showed a more complex pattern, with additional bands of ~70 kb and ~170 kb, suggestive of polymorphisms in the *Dux* array size and/or number.

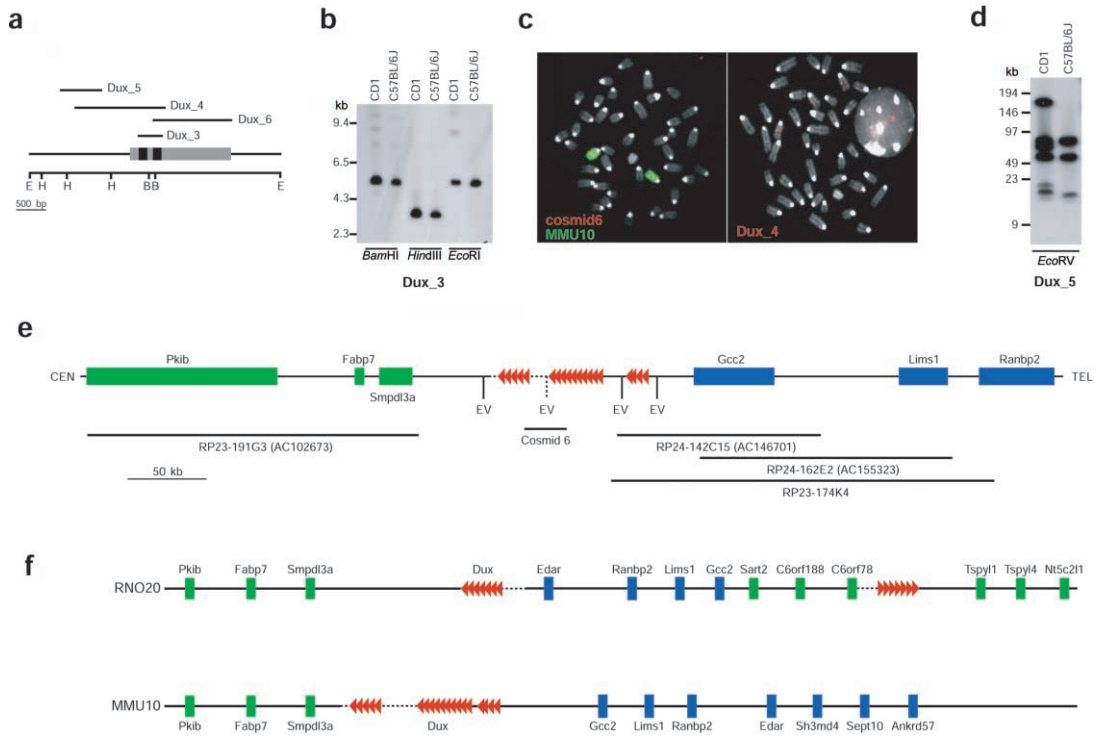

Figure 6. Physical mapping of the mouse *Dux* locus. *a,* Schematic diagram of a mouse *Dux* repeat, indicating key restriction-enzyme sites and the locations of probes used in this study. E = *Eco*RI; B = *Bam*HI; H = *Hin*dIII. The ORF is shown as a shaded box, with the homeoboxes in black. *b,* Southern blot of genomic DNA from C57BL/6J or CD1 mice probed with ^{32}P-labeled Dux_3. The filter was washed under high-stringency conditions and was exposed for 4 h. *c,* FISH analysis of mouse chromosomes. For both the cosmid 6 and the plasmid Dux_4 probes, a single signal was seen on mouse chromosome 10, identified by the chromosome paint in the cosmid 6 panel. *d,* PFGE analysis of *Eco*RV-digested genomic DNA. The filter was hybridized with ^{32}P-labeled Dux_5, was washed under high-stringency conditions, and was exposed for 6 h. *e,* Map of the *Dux* array region, indicating the locations of BAC and cosmid clones. *Dux* sequences are shown as red arrowheads. CEN = centromere; TEL = telomere; EV = *Eco*RV. There are three separate *Dux* clusters that are not joined in the current mouse genome assembly. The dotted line indicates the unsequenced region. The mouse *Dux* locus maps to an evolutionary chromosomal break point; genes that lie telomeric to the arrays have orthologues on human chromosome 2 (*blue*); genes that lie centromeric have orthologues on human chromosome 6 (*green*). We could not find any *DUX*-like sequences in either of these human regions. *f,* Schematic map of the rat and mouse *Dux* loci (not to scale), following the Ensembl assembly (release 42). The rat genome sequence is incomplete for this locus. In the current assembly, the rat *Dux* sequences are located in two arrays, with an intervening region of ~1 Mb. Comparison of gene order between mouse and rat indicates that there has probably been at least one inversion and additional rearrangements in this region during recent murine evolution. Color coding of genes is as in panel e. MMU10 = mouse chromosome 10; RNO20 = rat chromosome 20.

Because the Ensembl Genome Browser showed that this region of chromosome 10 contains a large gap within the assembly, we produced a map of this locus, using sequence information from shotgun assembly contigs combined with physical mapping of our BAC and cosmid clones. From these data, the mouse locus appears to contain three separate but tightly linked *Dux* arrays. One array of two complete and four partial units is present within BAC clone RP24-142C15; this cluster corresponds to the 19-kb *Eco*RV band seen by PFGE (fig. 6*d*). The 50-kb and 90-kb fragments presumably represent two arrays of 10–18 repeats, only one of which could be anchored within our genomic map. The intervening region contains several *Eco*RV sites, accounting for the two fragments seen by PFGE. Southern-blot analysis by use of rare-cutter restriction enzymes is in agreement with this map (data not shown). The gap in the assembly is likely to be the result of inherent problems in the cloning and assembly of tandem repeat sequences. Closure of this gap will require identification and sequencing of additional BACs. Therefore, we have identified a mouse homologue of D4Z4 that, like the human locus, is organized as a large tandem array, with each repeat containing a *Dux* ORF.

Expression of Mouse Dux *Array and Bidirectional Transcription*

Although the human D4Z4 array has long been known to have the potential to encode a homeodomain protein,[8] no transcript from this locus has been demonstrated. There are hundreds of dispersed copies of D4Z4-like sequences in the human genome, some of which are apparently transcribed,[8,10,40,41] although many of these are clearly pseudogenes.[8,40] This background of highly related sequences has complicated human RT-PCR analysis. The apparent lack of significant numbers of dispersed copies of the mouse *Dux* repeat suggested that expression studies would be more straightforward for this species. Therefore, we investigated whether the mouse *Dux* locus is transcribed.

By RT-PCR, although we were unable to successfully amplify across the entire ORF, we used overlapping sets of PCR primers to confirm that the whole ORF is transcribed (fig. 7). *Dux* transcripts were amplified from a range of tissues and in a range of embryonic stages (fig. 8*a*). Although this PCR was not quantitative, the most robust expression was detected in the CNS. In both muscle tissue and in the C2C12 myoblast cell line (in both myoblasts and differentiated myotubes), amplification was weak but consistent. Because the *Dux* transcription unit lacks introns, all RNA samples were treated rigorously with DNase before amplification, and control reactions without reverse transcriptase were always performed.

We also used the independent method of RNA FISH to test whether the array is transcribed. We used mouse splenocytes because they are readily amenable to this procedure and because we had shown that they express mouse *Dux* by RT-PCR. An antisense probe for the nearby *Gcc2* gene was used as a control (fig. 8*b*), and a sense *Gcc2* probe gave no signal (data not shown). A small but significant number of cells showed colocalization of signals from *Gcc2* and *Dux* (fig. 8*c*), indicating generation of primary RNA transcripts from the *Dux* array. Surprisingly, a sense *Dux* probe also gave signals that colocalized with *Gcc2,* indicating that antisense *Dux* transcripts were also being produced (fig. 8*c*). The existence of this *Dux* antisense mRNA was confirmed by RT-PCR (fig. 8*d*).

Since RT-PCR showed that the adult brain was a major site of *Dux* transcription, we analyzed this in more detail by in situ hybridization (fig. 8*e*). Signals were predominantly restricted to three areas: the inner granule layer of the cerebellum, the dentate gyrus of the hippocampus, and the olfactory bulb. Weak expression was also seen in the cornu ammonis (CA) regions of the hippocampus (indicated by arrows in fig. 8*e*). Since all these major sites of expression are the regions where granule neurons are found, we examined earlier stages of development, to determine whether expression was restricted to this neuronal subgroup. This verified that *Dux* transcripts were present in immature precursors of granule cells in the external granule layer of the cerebellum (see expression in postnatal day 8 [P8] cerebellum in fig. 8*e*) and in the immature dentate gyrus and olfactory bulb (data not shown). In addition, in younger brains, we saw broader expression in the cortex, the region around the lateral ventricles, and scattered cells in other regions. Thus, the mouse *Dux* array is transcribed in a broad range of immature cell types in the brain, but strong CNS expression is maintained only in granule cell populations in the adult. In all these regions, the staining was strong and was indistinguishable between sense and antisense probes, confirming the observations by RT-PCR and RNA-FISH that transcription is approximately equal from both strands wherever this gene is expressed.

Localization of Epitope-Tagged Mouse Dux Protein to the Nucleus

If a homeodomain protein were produced from the mouse *Dux* array, it would be expected to localize to the nucleus, as shown elsewhere for epitope-tagged proteins corresponding to the hypothetical translation product of a human *DUX4* mRNA.[42] Therefore, we examined the location of exogenously expressed epitope-tagged mouse Dux protein. Both N- and C-terminal EGFP-tagged Dux proteins were restricted to the nuclei of transfected cells (fig. 9), and V5-tagged constructs showed the same localization pattern (data not shown). Constructs encoding only the homeodomain regions also produced nuclear localization, although the EGFP-tagged C-terminal domain localized throughout the cytoplasm and nucleus (fig. 9). Therefore, as for the human protein, the nuclear-localization signals for the mouse Dux protein have been conserved and reside within the homeodomain regions.

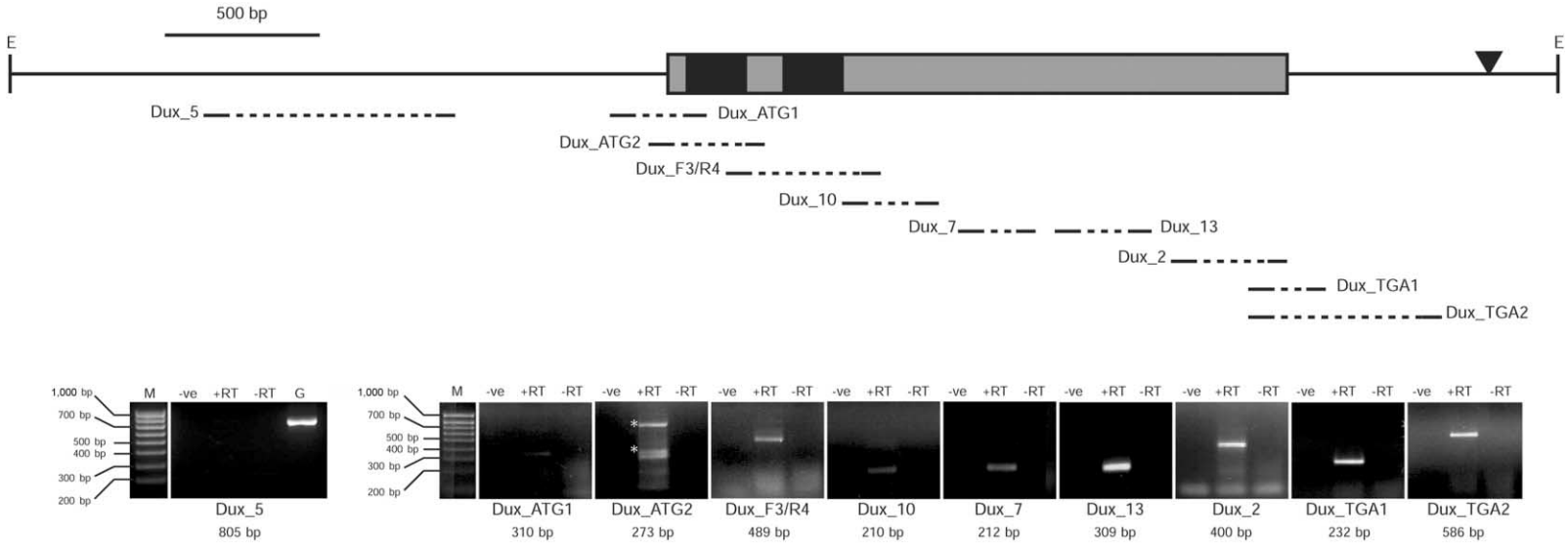

Figure 7. RT-PCR analysis of the mouse *Dux* repeat. Representative agarose gels of RT-PCR and genomic PCR products. M = molecular-weight ladder; -ve = no template; -RT = RNA added after inactivation of reverse transcriptase; +RT = RNA present throughout the OneStep reaction; G = genomic DNA template. The ORF is indicated by the gray rectangle, and the homeobox sequences by the black boxes. The putative polyA addition site is indicated by the black triangle. Primer sequences and reaction conditions are provided in table 1.

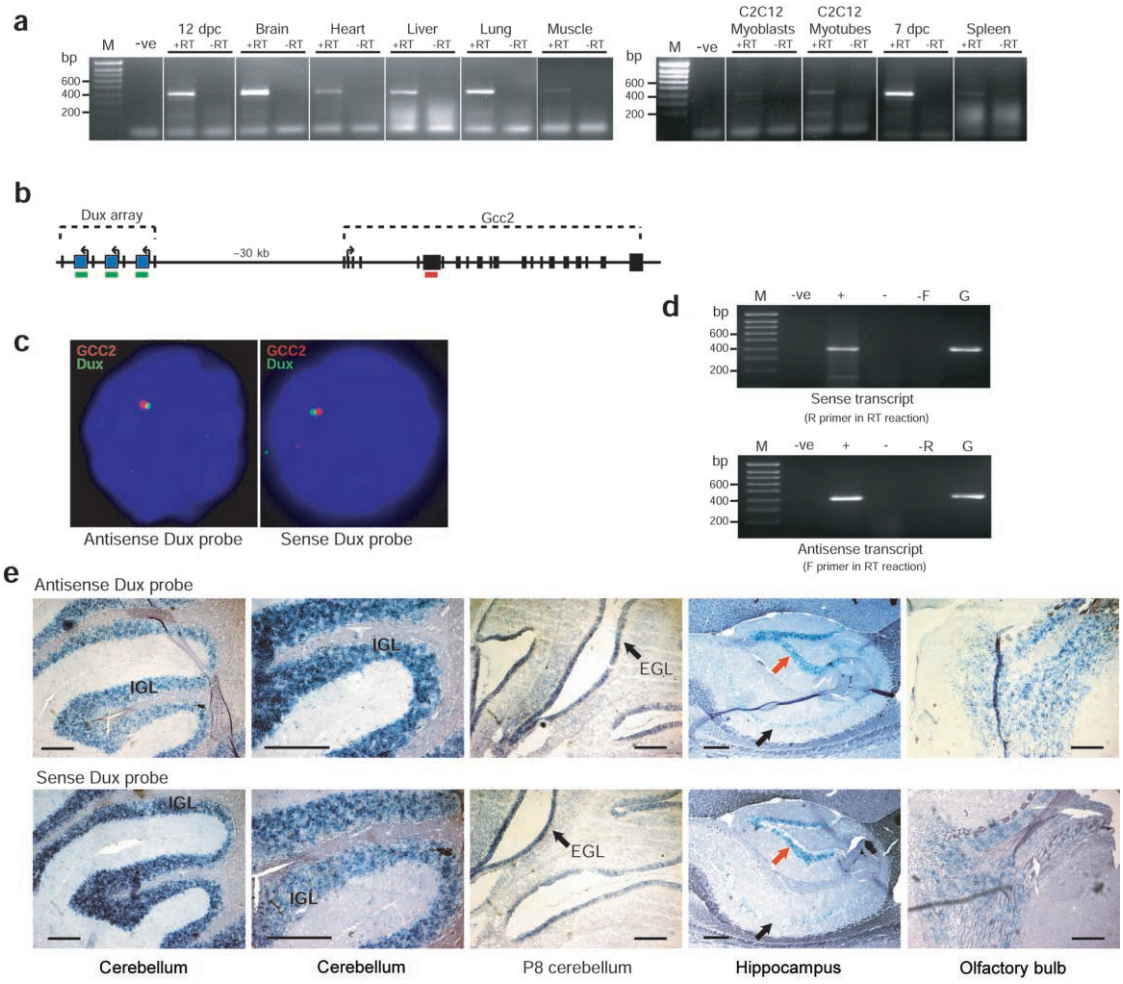

Figure 8. Evidence of transcription from the mouse *Dux* locus. *a,* RT-PCR of mouse tissues by use of primers Dux_2f and Dux_2r, which should give a product of 400 bp (table 1). RT indicates a control reaction, where RNA was added after inactivation of the reverse transcriptase. dpc = days post coitum; C2C12 = mouse myoblast cell line. *Dux* transcripts were amplified from a range of tissues and embryonic stages. Detection of amplification was robust in the brain. In muscle cells (both in vivo and in vitro), amplification was weak but consistent. Sequencing of the products confirmed that they originated from the array. -ve = no template; -RT = control reaction where RNA was added after inactivation of the reverse transcriptase. +RT = RNA present throughout the OneStep reaction. *b,* Schematic representation of the mouse *Dux* and *Gcc2* loci, indicating the direction of transcription and the positions of the probes used in RNA FISH experiments; both probes are 1.6 kb. *Dux* probes (Dux_6 in fig. 6*a*) were labeled with DIG, and *Gcc2* probes with dinitrophenol. *c,* Representative mouse splenocyte nuclei from RNA FISH experiments. Hybridized probes were detected with secondary antibodies conjugated to FITC (for detection of DIG-labeled *Dux* probes) or Texas Red (for detection of DNP-labeled *Gcc2* probes). In both cases, an antisense probe was used to identify sense *Gcc2* transcripts, and a sense *Gcc2* probe gave no signal (data not shown). However, for *Dux,* both antisense and sense probes gave signals that colocalized with *Gcc2* signal, indicating that both sense and antisense transcripts are generated from the array. *d,* RT-PCRs were performed, as for panel a, by use of 1 μg of DNase-treated 7-dpc RNA, except that only one primer (as indicated) was included in the reverse-transcriptase step; the second primer was added after the reverse-transcriptase enzyme had been inactivated. M = molecular-weight ladder; -ve = no RNA added; + = RNA present through the OneStep reaction; − = RNA added after inactivation of reverse transcriptase; -F = control reaction in which the forward primer was not included; -R = control reaction in which the reverse primer was not included; G = genomic DNA positive control. *e,* Nonradioactive in situ hybridization analysis of mouse adult brain by use of the probe Dux_6. All sections are from adult mouse brain, except for that from P8 cerebellum. In the hippocampus, the black arrow indicates the weakly stained CA region; a red arrow indicates the dentate gyrus. EGL = external granule layer; IGL = inner granule layer. Scale bar = 100 μm, except for hippocampus, for which scale bar = 200 μm.

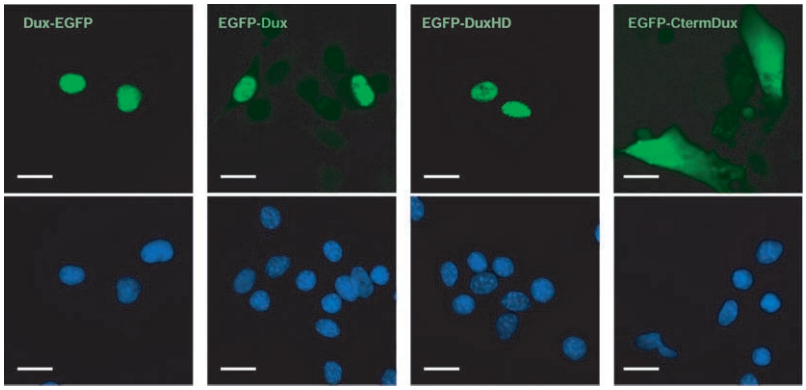

Figure 9. Localization of epitope-tagged mouse Dux protein to the nucleus. Fluorescence images are of EGFP-tagged Dux protein constructs transfected into C2C12 myoblast cells. Nuclei are counterstained with DAPI. Both the full-length protein and the homeodomain regions show nuclear localization. The tagged C-terminal region alone is distributed throughout the cell. Scale bar = 20 μm.

Discussion

Conservation of the D4Z4 Array and Coding Function

Until now, the possibility that D4Z4 is a gene has been largely discounted, primarily because of the lack of evidence of an mRNA.[8,10,40,43] However, we present here data that strongly indicate a protein-coding function for the array. The *DUX4* ORF has been conserved for >100 million years, and analysis of primate sequences by use of a codon model indicates selection for maintenance of a protein-coding capacity. Importantly, in no species did we identify any DUX4 homologues that contained stop codons. Although the absence of stop codons from 420-aa DUX4 ORFs in 11 species across a total of >480 million years of evolution is significantly different from expectation under neutrality ($P \approx 2.4 \times 10^{-25}$), a selective constraint at the RNA level (such as antisense function) could reduce the ORF mutation rate to below the rate for neutral sequences and thereby produce fewer stop codons than expected. We therefore examined the observed mutation rate across the whole DUX4 tree in 120 aligned homeobox codons. Under the null hypothesis that there is no coding function requiring an ORF, more than eight mutations creating stop codons are expected in the entire tree, and the complete absence of stop codons from the alignment is significant (Poisson $P \approx 1.47 \times 10^{-4}$). Overall, we conclude that DUX4 has not been preserved by chance but has been conserved by selection and that the most likely explanation for this selection is a protein-coding function.

In agreement with the proposed coding function, we have demonstrated transcription from the mouse *Dux* array and show that, like human DUX4,[42] exogenous expression of mouse Dux protein results in a nuclear-localization pattern that is consistent with its expected function as a transcription factor. Thus, contrary to the long-held assumption that D4Z4 is simply a heterochromatic DNA repeat, our data indicate a coding function for this locus.

There are several explanations that might account for the failure thus far to identify human D4Z4 transcripts. For example, the mRNA may be rare, expression might be spatially or developmentally restricted (only a limited number of human tissues and cell lines have been tested), or the mRNA may be refractory to RT-PCR amplification and/or cloning. Indeed, even though our in situ data clearly demonstrate mouse *Dux* transcription, RT-PCR analysis of this locus was not straightforward, since we were unable to amplify the full-length mRNA. The evolutionary selection at the codon level to maintain the *DUX4* ORF indicates that a D4Z4 protein is expressed. Therefore, the case for a more thorough and meticulous search for human D4Z4 transcripts by use of a range of approaches and tissues is now indisputable.

The tandem-repeat organization of D4Z4 is unusual for a gene but is not without precedent. Although tandemly arrayed protein-encoding genes are rare in humans, two verified examples are *TSPY* (an intron-containing, placenta-specific, protein-encoding gene that is contained within the 20-kb unit of the DYZ5 tandem array on the Y chromosome[44]) and the deubiquitinating enzyme, USP17, encoded by the RS447 tandem repeats on chromosomes 4p and 8p.[45] Indeed, USP17 has several properties in common with DUX4; the RS447 repeat unit is a similar size (4.7 kb), the copy number is polymorphic (range 34–94), and the transcription unit is intronless. In addition, like the mouse *Dux* array, both sense and antisense RS447 transcripts have been reported. Antisense transcription of RS447 is thought to provide a regulatory mechanism for USP17[45]; thus, we speculate that antisense *Dux* transcripts might have a similar function.

Although it was not unexpected to find homologues of *DUX4* in other mammals, it was surprising to find evolutionary conservation of the tandem-array organization and high repeat copy number. Such striking conservation implies a functional requirement for a minimum copy number of *DUX4* units, and this might be relevant to FSHD, in which a reduction in D4Z4 repeats below the

threshold of 11 results in the disease. The tandem-array organization of *TSPY* is conserved in other species and, here, the high repeat copy may be a mechanism to protect this Y-chromosome gene from deleterious mutations and to maintain a critical dose of TSPY during human spermatogenesis.[46,47] Although the cytogenetic locations of the Afrotherian arrays are unknown, all the primate D4Z4 repeats are found at telomeric locations.[23,24] Human subtelomeres are dynamic regions that are prone to DNA rearrangements, making them potentially hazardous locations for genes to reside.[48,49] Therefore, it is possible that the array structure of the *DUX4* loci is also a mechanism for protection against mutation. However, the mouse and rat *Dux* arrays are not located at telomeres, so this appears to be an unlikely explanation for the rodent gene organization.

Evolutionary Origins of D4Z4

The intronless *DUX4* arrays presumably originated from amplification of a retrotransposed pseudogene. The intron-containing *DUX* genes, which also encode double-homeodomain proteins, are the most likely candidates for a progenitor gene.[37] Our phylogenetic and synteny data indicate that all four intron-containing DUX paralogues existed before the mammalian radiation, with subsequent loss of some paralogues in different lineages. Because the sequence cover of most mammalian genomes is low and because we used stringent criteria to assign putative *DUX* genes, it is likely that we have underestimated the complement of these genes for some species. As additional genome sequences become available, it should be possible to infer the complex evolutionary history of this mammalian gene family with more certainty.

The phylogenetic grouping and the conserved C-terminal domain indicate that *DUX4* most likely originated from a retrotransposed copy of a *DUXC* gene, rather than from *DUXA* as suggested elsewhere.[37] Recent phylogenetic analysis strongly supports the placement of Afrotheria at the base of the Placentalia,[36] indicating that *DUX4* was present in the common ancestor of all mammals. The array appears to have been lost in Laurasiatheria and Xenarthra but retained in Afrotheria and in primates, where it gave rise to D4Z4. Alternatively, *DUX4* may have evolved twice, although the topology of the DUX4 branch resembles that of the mammalian phylogenetic tree, arguing against this. The rodent *DUX* arrays do appear to have originated from a separate retrotransposon amplification event. However, since the rodent Dux proteins share a conserved C-terminal domain with DUX4 and DUXC that is not present in other DUX proteins, we speculate that the progenitor for the rodent array might also have been a *DUXC* gene.

Implications for the FSHD Disease Mechanism

Although first identified as potential homeobox gene,[2,8] D4Z4 has long been regarded to be a pseudogene, and current epigenetic models of the FSHD mechanism do not take into account any possibility of a coding function for D4Z4.[4] Since our data do indicate a coding function for D4Z4, the possibility that *DUX4* is directly involved in the disease must now be reconsidered. A disease model in which *DUX4* is overexpressed would explain the dominant inheritance pattern and the requirement that at least one copy of D4Z4 be present at 4qter for development of the disease (complete loss of one 4q D4Z4 array does not result in FSHD[4]). We note that cellular toxicity of overexpressed exogenous human DUX4 has been reported,[10] and we have also observed reduced cell viability in cells expressing epitope-tagged mouse Dux protein (data not shown), suggesting that cells are likely to be sensitive to DUX protein levels. Evidence of a role for DUX4 in transcriptional regulation comes from a study that identified DUX4 fusions in two individuals with Ewing-like sarcoma carrying the recurrent translocation t(4;19)(q35;q13).[50] In each case, the C-terminal region of DUX4 was fused to the HMG box protein CIC, the human homologue of the *Drosophila* transcriptional repressor *Capicua,*[51] such that the fusion protein—and, indeed, the C-terminal region of DUX4 alone—now had the ability to activate CIC target genes.[50] These data show that the DUX4 coding sequence has retained the ability to encode a transcriptional activator.

Identification of a mouse homologue of D4Z4 also raises the possibility of development of a mouse model of the FSHD mutation. Although it is uncertain whether the mouse *Dux* array and human D4Z4 have equivalent functions, we are now in a position to design strategies to test this hypothesis. Currently, however, we can only speculate how our mouse *Dux* data might relate to the FSHD disease mechanism. The lack of evidence of D4Z4 transcription suggests that, if the human repeat is expressed, it is at very low levels. In contrast, by RT-PCR, the mouse *Dux* gene is widely expressed, with high levels in the CNS. FSHD typically has no CNS symptoms, except in a few very severe cases.[52] However, given that the FSHD mechanism is still unclear, it is difficult to make predictions about tissue-specific consequences of the D4Z4 deletions. For example, if DUX4 were overexpressed in FSHD, only some tissues (perhaps those that normally express little or no protein) might be affected.

Epigenetic models of FSHD propose that D4Z4 deletions perturb chromatin structure and affect expression of chromosome 4q genes.[4] Proposed mechanisms include changes in nuclear localization and in DNA methylation. There is evidence that the shortened D4Z4 arrays in FSHD are partially hypomethylated.[14,15] However, rather than a model of changes altering the expression of distant genes (which have proved difficult to identify), our data support a model in which such changes might lead to misexpression of *DUX4*.

D4Z4 deletions must also occur within the correct genomic context to cause FSHD. Deletions within the chromosome 10q array do not result in the disease.[4] Further-

more, only deletions that occur on a chromosome bearing a 4qA-type telomere are pathogenic.[7] A model in which FSHD is a result of misregulation of DUX4 does not immediately explain these findings but, like the proposed epigenetic mechanisms, implies that functionally important sequences must reside outside the array. In our model, such sequences would play a role in regulating *DUX4* transcription.

In conclusion, this study has added significantly to our understanding of the evolution of the D4Z4 tandem array and highlights the possibility that it plays a more direct role in the FSHD disease mechanism than was previously believed. For example, the simple prediction that FSHD is caused by inappropriate expression of DUX4 itself rather than that of unrelated, linked genes needs to be disproved before a mechanistic role for the coding function of D4Z4 in the disease can be disregarded.

Acknowledgments

Previous members of the laboratory of J.E.H. are thanked for their contributions to the study of D4Z4 evolution. Chris Moore is thanked, for assistance in the phylogenetic analysis, and Robert Lyle, for helpful comments on the manuscript. J.C. was funded by a Medical Research Council PhD studentship. J.E.H. holds a Biotechnology and Biological Sciences Research Council (BBSRC) Research Development Fellowship. Funding is gratefully acknowledged from the Muscular Dystrophy Association (MDA) (United States), the FSH Society (FSH Society Delta Railroad Construction Company Research Fellowship Grant FSHS-DR-008), The Wellcome Trust (Equipment Grant 066536/Z/01/Z), and the University of Nottingham.

Web Resources

Accession numbers and URLs for data presented herein are as follows:

BLAST, http://www.ncbi.nlm.nih.gov/blast/
ClustalW, http://www.ebi.ac.uk/clustalw/
EMBL, http://www.ebi.ac.uk/embl/ (for accession numbers AF117653, AC135091, AM398147–AM398151, BN000980–BN000984, and BN000988–BN000990)
Ensembl, http://www.ensembl.org/index.html
GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for accession number NM_027375)
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for FSHD)
Phylodendron, http://iubio.bio.indiana.edu/treeapp/
RepeatMasker, http://www.repeatmasker.org/

References

1. Padberg GW (2004) Facioscapulohumeral muscular dystrophy. In: Upadhyaya M, Cooper DN (eds) Facioscapulohumeral muscular dystrophy (FSHD): clinical medicine and molecular cell biology. BIOS Scientific Publishers, Oxford, United Kindgom, pp 41–54
2. Wijmenga C, Hewitt JE, Sandkuijl LA, Clark LN, Wright TJ, Dauwerse HG, Gruter A-M, Hofker MH, Moerer P, Williamson R, et al (1992) Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. Nat Genet 2:26–30
3. van Deutekom JCT, Wijmenga C, van Tienhoven EAE, Gruter A-M, Hewitt JE, Padberg GW, van Ommen G-JB, Hofker MH, Frants RR (1993) FSHD associated rearrangements are due to deletion of integral copies of a 3.2 kb tandemly repeated unit. Hum Mol Genet 2:2037–2042
4. Tawil R, van der Maarel SM (2006) Facioscapulohumeral muscular dystrophy. Muscle Nerve 34:1–15
5. Deidda G, Cacurri S, Grisanti P, Vigneti E, Piazzo N, Felicetti L (1995) Physical mapping evidence for a duplicated region on chromosome 10qter showing high homology with the FSHD locus on chromosome 4qter. Eur J Hum Genet 3:155–167
6. van Geel M, Dickson MC, Beck AF, Bolland DJ, Frants RR, van der Maarel SM, de Jong PJ, Hewitt JE (2002) Genomic analysis of human chromosome 10q and 4q telomeres suggests a common origin. Genomics 79:210–217
7. Lemmers RJ, de Kievit P, Sandkuijl L, Padberg GW, van Ommen GJ, Frants RR, van der Maarel SM (2002) Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. Nat Genet 32:235–236
8. Hewitt JE, Lyle R, Clark LN, Valleley EM, Wright TJ, Wijmenga C, van Deutekom JCT, Francis F, Sharpe PT, Hofker M, et al (1994) Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystrophy. Hum Mol Genet 3:1287–1295
9. Winokur ST, Bengtsson U, Feddersen J, Mathews KD, Weiffenbach B, Bailey H, Markovich RP, Murray JC, Wasmuth JJ, Altherr MR, et al (1994) The DNA rearrangement associated with facioscapulohumeral muscular dystrophy involves a heterochromatin-associated repetitive element: implications for a role of chromatin structure in the pathogenesis of the disease. Chromosome Res 2:225–234
10. Gabriels J, Beckers M, Ding H, De Vriese A, Plaisance S, van der Maarel S, Padberg GW, Frants RR, Hewitt JE, Collen D, et al (1999) Nucleotide sequencing of the partially deleted D4Z4 locus in a patient with FSHD identifies a putative gene within each 3.3 kb element. Gene 236:25–32
11. Yang F, Shao CB, Vedanarayanan V, Ehrlich M (2004) Cytogenetic and immuno-FISH analysis of the 4q subtelomeric region, which is associated with facioscapulohumeral muscular dystrophy. Chromosoma 112:350–359
12. Masny PS, Bengtsson U, Chung SA, Martin JH, van Engelen B, van der Maarel SM, Winokur ST (2004) Localization of 4q35.2 to the nuclear periphery: is FSHD a nuclear envelope disease? Hum Mol Genet 13:1857–1871
13. Tam R, Smith KP, Lawrence JB (2004) The 4q subtelomere harboring the FSHD locus is specifically anchored with peripheral heterochromatin unlike most human telomeres. J Cell Biol 167:269–279
14. van Overveld PGM, Lemmers RJ, Sandkuijl LA, Enthoven L, Winokur ST, Bakels F, Padberg GW, van Ommen GJ, Frants RR, van der Maarel SM (2003) Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. Nat Genet 35:315–317
15. van Overveld PGM, Enthoven L, Ricci E, Rossi M, Felicetti L, Jeanpierre M, Winokur ST, Frants RR, Padberg GW, van der Maarel SM (2005) Variable hypomethylation of D4Z4 in facioscapulohumeral muscular dystrophy. Ann Neurol 58:569–576

16. Kondo T, Bobek MP, Kuick R, Lamb B, Zhu X, Narayan A, Bourc'his D, Viegas-Pequignot E, Ehrlich M, Hanash SM (2000) Whole-genome methylation scan in ICF syndrome: hypomethylation of non-satellite DNA repeats D4Z4 and NBL2. Hum Mol Genet 9:597–604
17. Gabellini D, Green MR, Tupler R (2002) Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle. Cell 110:339–348
18. van Deutekom JCT, Lemmers RJLF, Grewal PK, van Geel M, Romberg S, Dauwerse HG, Wright TJ, Padberg GW, Hofker MH, Hewitt JE, et al (1996) Identification of the first gene (*FRG1*) from the FSHD region on human chromosome 4q35. Hum Mol Genet 5:581–589
19. Jiang GF, Yang PG, van Overveld PGM, Vedanarayanan V, van der Maarel SM, Ehrlich M (2003) Testing the position-effect variegation hypothesis for facioscapulohumeral muscular dystrophy by analysis of histone modification and gene expression in subtelomeric 4q. Hum Mol Genet 12:2909–2921
20. Winokur ST, Chen YW, Masny PS, Martin JH, Ehmsen JT, Tapscott SJ, van der Maarel SM, Hayashi Y, Flanigan KM (2003) Expression profiling of FSHD muscle supports a defect in specific stages of myogenic differentiation. Hum Mol Genet 12:2895–2907
21. Osborne RJ, Welle S, Venance SL, Thornton CA, Tawil R (2007) Expression profile of FSHD supports a link between retinal vasculopathy and muscular dystrophy. Neurology 68:569–577
22. Gabellini D, D'Antona G, Moggio M, Prelle A, Zecca C, Adami R, Angeletti B, Ciscato P, Pellegrino MA, Bottinelli R, et al (2006) Facioscapulohumeral muscular dystrophy in mice overexpressing FRG1. Nature 439:973–977
23. Clark LNC, Koehler U, Ward DC, Wienberg J, Hewitt JE (1996) Analysis of the organisation and localisation of the FSHD-associated tandem array in primates: implications for the origin and evolution of the 3.3 kb repeat family. Chromosoma 105:180–189
24. Winokur ST, Bengtsson U, Vargas JC, Wasmuth JJ, Altherr MR (1996) The evolutionary distribution and structural organisation of the homeobox-containing repeat D4Z4 indicates a functional role for the ancestral copy in the FSHD region. Hum Mol Genet 5:1567–1577
25. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680
26. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comp Appl Biosci 13:555–556
27. Jeanmougin F, Thompson JD, Gibson TJ, Gouy M, Higgins DG (1998) Multiple sequence alignment with ClustalX. Trends Biochem Sci 23:403–405
28. Felsenstein J (1989) PHYLIP—phylogeny inference package, version 3.2. Cladistics 5:164–166
29. Chong SS, Pack SD, Roschke AV, Tanigami A, Carrozzo R, Smith ACM, Dobyns WB, Ledbetter DH (1997) A revision of the lissencephaly and Miller-Dieker syndrome critical regions in chromosome 17p13.3. Hum Mol Genet 6:147–155
30. Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
31. Bolland DJ, Wood AL, Johnston CM, Bunting SF, Morgan G, Chakalova L, Fraser PJ, Corcoran AE (2004) Antisense intergenic transcription in V(D)J recombination. Nat Immunol 5: 630–637
32. Rex M, Scotting SJ (1999) In situ hybridisation to sections (non-radioactive). In: Sharpe PT, Mason I (eds) Molecular embryology: methods and protocols. Humana Press, Totowa, NJ, pp 645–654
33. Ricchetti M, Tekaia F, Dujon B (2004) Continued colonization of the human genome by mitochondrial DNA. PLoS Biol 2: 1313–1324
34. The Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87
35. Baird D, Royle N (1997) Sequences from higher primates orthologous to the human Xp/Yp telomere junction region reveal gross rearrangements and high levels of divergence. Hum Mol Genet 6:2291–2299
36. Nikolaev S, Montoya-Burgos JI, Margulies EH, NISC Comparative Sequencing Program, Rougemont J, Nyffeler B, Antonarakis SE (2007) Early history of mammals is elucidated with the ENCODE multiple species sequencing data. PLoS Genet 3:3–7
37. Booth HAF, Holland PWH (2006) Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. Gene 387:7–14
38. Bürglin TR (1994) A comprehensive classification of homeobox genes. In: Duboule D (ed) Guidebook to the homeobox genes. Oxford University Press, Oxford, United Kingdom, pp 25–72
39. Grewal PK, Bolland DJ, Carim Todd L, Hewitt JE (1998) High resolution mapping of mouse chromosome 8 identifies an evolutionary chromosomal breakpoint. Mamm Genome 9: 603–607
40. Lyle R, Wright TJ, Clark LN, Hewitt JE (1995) The FSHD-associated repeat, D4Z4, is a member of a dispersed family of homeobox-containing repeats, subsets of which are clustered on the short arms of the acrocentric chromosomes. Genomics 28:389–397
41. Ding H, Beckers M-C, Plaisance S, Marynen P, Collen D, Belayew A (1998) Characterization of a double homeodomain protein (DUX1) encoded by a cDNA homologous to 3.3 kb dispersed repeat elements. Hum Mol Genet 7:1681–1694
42. Ostlund C, Garcia-Carrasquillo RM, Belayew A, Worman HJ (2005) Intracellular trafficking and dynamics of double homeodomain proteins. Biochemistry 44:2378–2384
43. Alexiadis V, Ballestas ME, Sanchez C, Winokur ST, Vedanarayanan V, Warren M, Ehrlich M (2007) RNAPol-ChIP analysis of transcription from FSHD-linked tandem repeats and satellite DNA. Biochim Biophys Acta 1769:29–40
44. Manz E, Schneiders F, Brechlin AM, Schmidtke J (1993) TSPY-related sequences represent a microheterogeneous gene family organised as constitutive elements in DYZ5 tandem repeat units on the human Y chromosome. Genomics 17:726–731
45. Saitoh Y (2000) The RS447 human megasatellite tandem repetitive sequence encodes a novel deubiquitinating enzyme with a functional promoter. Genomics 67:291–300
46. Graves JAM, Disteche CM, Toder R (1998) Gene dosage in the evolution and function of mammalian sex chromosomes. Cytogenet Cell Genet 80:94–103
47. Vogel T, Schmidtke J (1998) Structure and function of TSPY, the Y-chromosome gene coding for the "testis-specific protein." Cytogenet Cell Genet 80:209–213

48. Mefford HC, Trask BJ (2002) The complex structure and dynamic evolution of human subtelomeres. Nat Rev Genet 3: 91–102
49. Linardopoulou E, Williams EM, Fan C, Friedman C, Young LM, Trask BJ (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature 437:94–100
50. Kawamura-Saito M, Yamazaki Y, Kaneko K, Kawaguchi N, Kanda H, Mukai H, Gotoh T, Motoi T, Fukayama M, Aburatani H, et al (2006) Fusion between CIC and DUX4 up-regulates PEA3 family genes in Ewing-like sarcomas with t(4;19)(q35;q13) translocation. Hum Mol Genet 15:2125–2137
51. Lee CJ, Chan WI, Cheung M, Cheng YC, Appleby VJ, Orme AT, Scotting PJ (2002) CIC, a member of a novel subfamily of the HMG-box superfamily, is transiently expressed in developing granule neurons. Mol Brain Res 106:151–156
52. Padberg GW, Lunt PW, Koch M, Fardeau M (1991) Diagnostic criteria for facioscapulohumeral muscular dystrophy. Neuromusc Dis 1:231–234