Chromosoma (2012) 121:489–497
DOI 10.1007/s00412-012-0380-y

RESEARCH ARTICLE

Evolution of *DUX* gene macrosatellites in placental mammals

Andreas Leidenroth · Jannine Clapp · Laura M. Mitchell · Daniel Coneyworth · Frances L. Dearden · Leopoldo Iannuzzi · Jane E. Hewitt

Received: 29 June 2012 / Revised: 31 July 2012 / Accepted: 31 July 2012 / Published online: 18 August 2012

Abstract Macrosatellites are large polymorphic tandem arrays. The human subtelomeric macrosatellite D4Z4 has 11–150 repeats, each containing a copy of the intronless *DUX4* gene. *DUX4* is linked to facioscapulohumeral muscular dystrophy, but its normal function is unknown. The *DUX* gene family includes *DUX4,* the intronless *Dux* macrosatellites in rat and mouse, as well as several intron-containing members (*DUXA*, *DUXB*, *Duxbl*, and *DUXC*). Here, we report that the genomic organization (though not the syntenic location) of primate *DUX4* is conserved in the Afrotheria. In primates and Afrotheria, *DUX4* arose by retrotransposition of an ancestral intron-containing *DUXC*, which is itself not found in these species. Surprisingly, we discovered a similar macrosatellite organization for *DUXC* in cow and other Laurasiatheria (dog, alpaca, dolphin, pig, and horse), and in Xenarthra (sloth). Therefore, *DUX4* and *Dux* are not the only *DUX* gene macrosatellites. Our data suggest a new retrotransposition-displacement model for the evolution of intronless *DUX* macrosatellites.

Communicated by David Tremethick

Electronic supplementary material The online version of this article (doi:10.1007/s00412-012-0380-y) contains supplementary material, which is available to authorized users.

A. Leidenroth · J. Clapp · L. M. Mitchell · D. Coneyworth ·
J. E. Hewitt (✉)
Centre for Genetics and Genomics, School of Biology, The
University of Nottingham,
Nottingham, UK
e-mail: jane.hewitt@nottingham.ac.uk

F. L. Dearden
Centre for Veterinary Science, Department of Veterinary Medicine,
University of Cambridge,
Cambridge, UK

L. Iannuzzi
National Research Council (CNR), Institute of Animal Production
Systems in Mediterranean Environments (ISPAAM),
Naples, Italy

Introduction

Macrosatellites are large, repetitive sequences composed of many kilobase-sized tandem repeats, and there are at least 50 such satellites in the human genome (Warburton et al. 2008; Tremblay et al. 2010). However, due to the limitations of sequence assembly, such repetitive loci are still largely unexplored “white space on a map” in most genomes (Alkan et al. 2011). One of the most extensively studied human macrosatellites is D4Z4, which is of particular interest because of its link to the human genetic disorder facioscapulohumeral muscular dystrophy (FSHD).

Nested within each 3.3-kb repeat unit of D4Z4 is one copy of the intronless double-homeobox gene *DUX4* (Hewitt et al. 1994). Aberrant expression of *DUX4* in muscle fibers of patients with FSHD has been causally linked to muscle degeneration and disease (reviewed by van der Maarel et al. 2011). In humans, large tandem arrays of D4Z4 sequences are located at the subtelomeres of both chromosome 4 and 10. These arrays are highly polymorphic in copy number, and each of the four alleles usually contains 11 to >150 D4Z4 units, making *DUX4* the human protein-encoding gene with the highest overall copy number (Alkan et al. 2009). In FSHD, the chromosome 4 array is contracted to fewer than 11 repeats (Wijmenga et al. 1992; van der Maarel et al. 2011). This is thought to “relax” the D4Z4 chromatin and cause the de-repression and transcription of *DUX4* in muscle, where this gene is usually silenced (Snider et al. 2010; Lemmers et al. 2010a).

For human D4Z4, we have extensive knowledge about its internal organization (Lemmers et al. 2004, 2001), epigenetic modifications (de Greef et al. 2009; Zeng et al. 2009), repeat number distribution (Rossi et al. 2007), and recent evolution (Lemmers et al. 2010b), but we know little about the organization and evolution of D4Z4-related sequences in other mammals.

Previous studies showed that both the tandem array organization and the subtelomeric localization of D4Z4 are conserved in chimpanzee, orangutan, and gorilla (Clark et al. 1996; Rudd et al. 2009). D4Z4-like sequences containing intronless *DUX4* open-reading frames were also identified in the genomes of the deeply rooted mammalian clade Afrotheria (elephant, hyrax, and tenrec), while a homolog with a similar structure (*Dux*) was found in mouse and rat (Clapp et al. 2007). Like *DUX4* in primates and Afrotheria, mouse *Dux* is intronless, and many copies of it are embedded within larger repeats (4.9 kb) in tandem array macrosatellites (Clapp et al. 2007).

Other mammals apparently lack intronless *DUX* genes, although their genomes do contain a number of intron-containing *DUX* homologs (*DUXA*, *DUXB*, *Duxbl*, and *DUXC*) (Leidenroth and Hewitt 2010). Of these four genes, *DUXC* is the most closely related to *DUX4* (Leidenroth and Hewitt 2010). *DUXC* is also the only intron-containing *DUX* gene to share a conserved C-terminal domain with *DUX4* and *Dux* (Clapp et al. 2007; Leidenroth and Hewitt 2010). This domain can act as a transcriptional activator: several cases of Ewing-like sarcoma are linked to a fusion of the DUX4-CTD to another DNA binding protein (CIC) by translocations (Kawamura-Saito et al. 2006).

Previously, we reported evidence for *DUXC* homologs in the mammalian groups Laurasiatheria (dog, cow, dolphin, and bat) and Xenarthra (armadillo and sloth) (Leidenroth and Hewitt 2010). Intriguingly, we never identified any intronless (*DUX4*-like) retrogenes in any of these species. Conversely, genomes that contain *DUX4* or *Dux* appeared not to contain *DUXC*. Based on their reciprocal species distribution pattern and close relatedness, *DUX4* and *DUXC* could be functional homologs.

Until now, *DUXC* was presumed to be a single-copy gene. We had therefore previously hypothesized that the intronless *DUX4* and *Dux* macrosatellites arose in the common ancestor of placental mammals through the reverse-transcription and retrotransposition of a spliced ancestral *DUXC* mRNA into a new genomic location. It was thought that local copy-number expansion of these new retrogenes then created the D4Z4 macrosatellites (Clapp et al. 2007; Leidenroth and Hewitt 2010). This model implied that the primate, murine, and Afrotheria lineages had lost *DUXC* but retained the intronless *DUX* macrosatellites, while the Laurasiatheria and Xenarthra retained *DUXC* but lost *DUX4*.

Here, we present new data that suggest an alternative evolutionary model. The genome of the most recent common ancestor of all placental mammals contained a *DUXC* macrosatellite but no intronless tandem arrays. The intronless *DUX4* and *Dux* genes then arose independently several times by separate retrotransposition events that displaced the ancestral *DUXC* macrosatellite in primates, murines, and the Afrotheria.

Materials and methods

Sample acquisition and tissue culture

Illumina reads were downloaded from the DNA database of Japan (http://trace.ddbj.nig.ac.jp). The cow fibroblast cell line GM06034 was purchased from the Coriell Institute for Medical Research. The cow embryonic fibroblast cell line (BFF3) was donated by Ramiro Alberio. Cells were grown under standard tissue culture conditions. For the *DUXC* FISH, blood samples from *Bos taurus* and *Bubalus bubalis* were used. Tenrec (*Microgale cowani*), hyrax (*Procavia capensis*), and elephant (*Loxodonta africana*) cell lines were provided by Willem Rens, Department of Veterinary Medicine, University of Cambridge.

Copy-number analysis with mrsFAST and mrCaNaVaR

We used mrsFAST (Hach et al. 2010) (2.3.0.2) and mrCaNaVaR (0.32) (Alkan et al. 2009). We had previously assembled the *B. taurus DUXC* locus from trace archive data (Clapp et al. 2007). Using this assembly, we built a small reference genome including *DUXC* (5.9 kb), using *DUXA* (5.5 kb) and *ZAR1* (6.3 kb) as controls and cow chromosome 24 (66 Mb) to estimate background read-depth.

The reference genome was processed with RepeatMasker (www.repeatmasker.org). Tandem Repeat Finder (Benson 1999) was run with parameters “2 7 7 80 10 50 500 -m”. Genome assembly gaps were downloaded from UCSC (http://genome.ucsc.edu). The reference sequence was indexed with mrsFAST–index, and copy windows were defined with mrCaNaVaR–prep. Reads were mapped using default parameters and 5 % hamming. Aligned *.sam files were processed with mrCaNaVaR–read and –call modes, and copy-numbers calls were calculated as an average across windows spanning *DUXA* (two windows) and *DUXC* (three windows); outer windows were excluded to avoid edge effects.

DNA preparation and restriction digests

Genomic DNA was extracted using standard methods (Miller et al. 1988). For pulsed-field gel electrophoresis (PFGE), cells were embedded in low gelling agarose type VII (SIGMA-Aldrich) and equilibrated in the appropriate restriction enzyme buffer. Enzymes were purchased from Roche (*Bln*I, *Bam*HI, and *Eco*RV), NEB (*Pvu*II), Fermentas (*Bgl*II), and Promega (*Hin*dIII).

Gel electrophoresis and Southern blot

Digested plugs were equilibrated in 0.5×TBE buffer and run in a Bio-Rad Chef Mapper tank in a 0.5×TBE 1 % gel. Cow DNA was separated at 15 °C for 48 h with 5–120 s ramp at 4.5 V/cm to resolve the smaller fragments, or at 16 °C for 26 h with 8–120 s ramp at 6 V/cm to resolve the larger fragments. Linear 1 % agarose gels (20×20 cm) were run for 24 h at 40 V in 1 × TAE buffer. DNA markers for sizing were: Lambda ladder (NEB), MidRange PFG marker I (NEB), or digoxigenin-labeled high-molecular weight marker II (Roche). DNA was transferred to a positively charged nylon membrane (Roche) using standard protocols.

Probes were amplified using BioMix^{TM} Red (Bioline) and cloned into pGEM T-Easy vector (Promega). Primer pairs were 5′ CTATACAGCACTCATCAAATCTAGC 3′+ 5′ CCCAAAAGCAATGCCAAACTAGTC 3′ (p13E-11) and 5′ TGGTTTCAAAACCGAAGAGC 3′+5′ AGGA GAGGACCCTGGAGAAG 3′ (cow *DUXC*). Digoxigenin-labeled probes were synthesized with the PCR DIG probe synthesis kit (Roche) according to manufacturer's instructions. To probe the lambda ladder, 0.5 μg of *Bst*EII cut lambda DNA (NEB) was labeled with Fluorescein High-Prime (Roche) according to manufacturer's instructions.

Membranes were pre-hybridized with DIG EasyHyb (Roche), and probes were hybridized overnight at 42 °C (p13E-11) or 59 °C (DUXC) in a roller oven. For signal detection, the Roche DIG wash and block buffer set was used. For linear gels, anti-digoxigenin-AP antibody (Roche) was diluted 1:20,000 in 20 ml fresh 1×DIG blocking solution. For pulsed-field gels, antibody detection was instead performed in 5 ml blocking solution supplemented with 2 μl anti-fluorescein NEF709 antibody (Perkin-Elmer). Signal was detected with CPD-star (Roche).

DUX4/DUXC metaphase FISH analysis

The Afrotheria *DUX4* probes were amplified from genomic DNA by PCR and cloned into T-Vector. Tenrec (*M. cowani*): 5′ GTGGCCAGGAAGATGACAAA 3′+5′ TGACGCT TTCAGAGGCTTGT 3′. Hyrax (*P. capensis*): 5′ GCTTTGCCCTCGTTTACCTG 3′+5′ GGAGGC ATTTCCTTTCGCAAC 3′. Elephant (*L. africana*): 5′ GAACTCCTCCCTGCCATCAC 3′+5′ TCTCTCCCC ACAGTGCTTGA 3′. Probes were between 2.1 and 2.4 kb in size and labeled using biotin. The hyrax and tenrec probes span part of the *DUX4* ORF. FISH was performed as described previously (Rens et al. 2006). The *DUXC* probe in pGEM T-Easy (see above) was labeled and hybridized to metaphase chromosomes using fluorescence in situ hybridization and cow RBPI-banding methods as described elsewhere (Iannuzzi and Di Berardino 2008).

Sequence analysis and alignments

BLAST analysis, trace archive searches, and phylogenetic analysis were performed as previously described (Leidenroth and Hewitt 2010).

Results

Telomeric *DUX4* macrosatellites are also present in the Afrotheria

The telomeric location of *DUX4* on chromosomes 4 and 10 is conserved in primates (Clark et al. 1996; Rudd et al. 2009). In human and ape genomes, additional D4Z4-related arrays are preferentially found in pericentromeric regions and the heterochromatin of acrocentric chromosomes (Lyle et al. 1995; Clark et al. 1996). To identify *DUX4* in the Afrotheria genomes, we hybridized elephant, hyrax, and tenrec chromosomes with species-specific *DUX4* probes (Fig. 1 and Online Resource 1). In all three species, most signals mapped to telomeric or pericentromeric regions of acrocentric chromosomes. In elephant, there are multiple telomeric signals of varying intensity, where stronger signals could represent higher sequence identity or copy number. In hyrax, there is a single telomeric signal on an acrocentric chromosome. Tenrec shows signals near centromeres on two chromosomes. The signals on the two chromosomes have different intensities, probably indicating two arrays of different sizes. Further support for *DUX4* tandem arrays in the Afrotheria comes from the reference assemblies of elephant and hyrax, which include contigs with four and five repeats (Fig. 1d and e).

We could not determine which Afrotheria chromosomes carried the *DUX4* signals, but we would expect no conservation of the chromosomal location of *DUX4* between Afrotheria and primates: in the primate lineage, *DUX4* maps distal to *FRG1* at the end of chromosome 4q. However, in other mammals, *FRG1* is not located at the end of a chromosome but internally; it is located next to the gene *ASAH*, which has an ortholog on human chromosome 8p (Grewal et al. 1997). There is conservation of synteny between *FRG1* and *ASAH* in all mammalian groups except primates. In the ancestral Eutherian genome, *FRG1* and *ASAH* were neighboring loci, but in primates, a chromosomal fission event distal to *FRG1* separated them, generating 4qter (Ferguson-Smith and Trifonov 2007). In non-primates, there are no *DUX* genes located near *FRG1* homologs. This strongly suggests that in primates, *DUX4* was transferred distal to *FRG1* on 4qter after the chromosomal fission. Therefore, although Afrotheria genomes do contain *DUX4* arrays, these are not found at the orthologous location to that of the primate arrays. Inspection of the elephant genome (loxAfr3) confirms

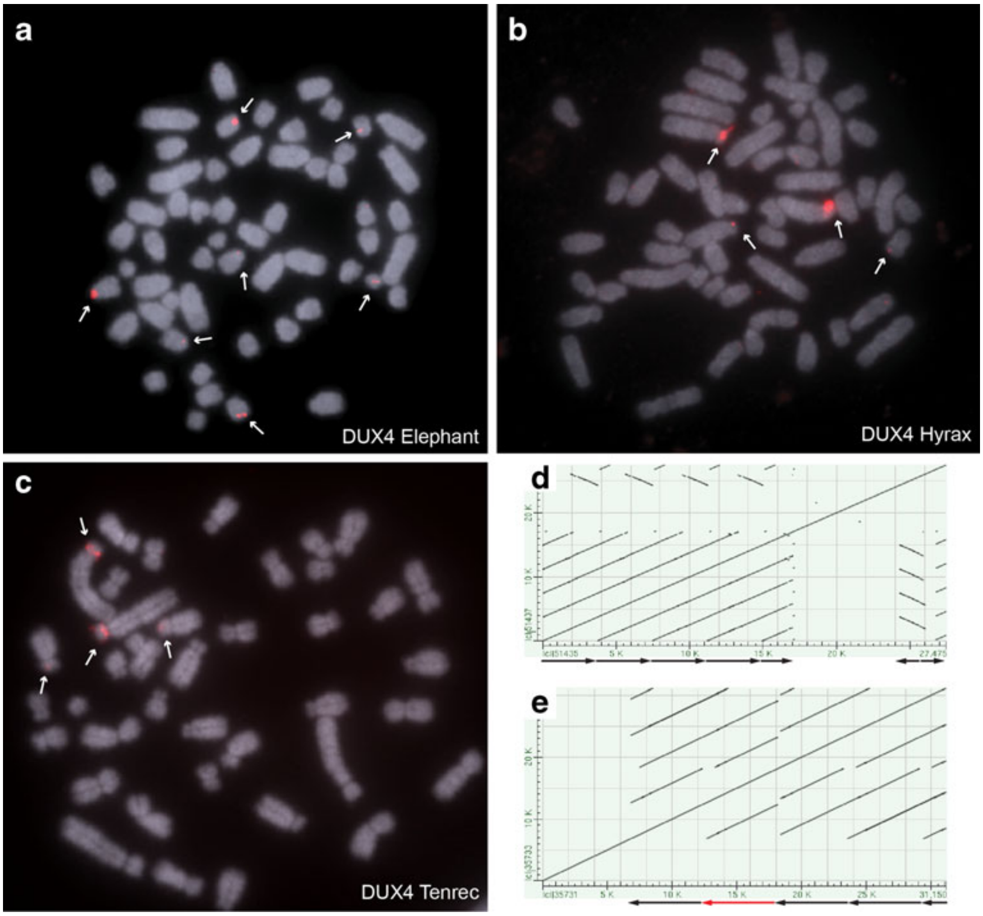

Fig. 1 Telomeric *DUX4* satellites are also present in the Afrotheria. **a**–**c** Representative metaphase FISH images, with chromosomes mounted in DAPI and *DUX4* probes labeled with Cy3. The species-specific *DUX4* probes produce strong telomeric signals in elephant, hyrax, and tenrec (*white arrows*). **d** Dot plot of hyrax contig 209751 aligned against itself using BLAST. This contig contains four tandem copies of ~3.9 kb each and a partially inverted repeat. *Arrow direction* indicates orientation of the *DUX4* open-reading frame within the repeat. **e** The elephant contig 85902/85903 (scaffold 112) contains four copies of about ~5.5 kb each. The *red arrow* marks a repeat with a *DUX4* ORF that has an internal stop codon

conservation of this ancestral *FRG1*/*ASAH* linkage group, with no evidence for *DUX4* at this location.

DUXC copy-number analysis by read-depth analysis

Our surveys of sequence archives indicated that *DUXC* might also be present at high copy number, with the number of sequence traces far exceeding the average fold-coverage of genomes. This was surprising, as we expected *DUXC* to be a single-copy gene like *DUXA, DUXB*, or *Duxbl.* Given the close relatedness of *DUXC* to the high copy-number genes *DUX4* and *Dux*, we decided to investigate this further. We chose to study *DUXC* in cow, as we had access to next-generation sequencing data for different cattle breeds, and a cell line for experimental confirmation.

Gene copy numbers can be estimated by counting next-generation sequencing reads. While bacterial shotgun cloning bias makes inferences from Sanger traces difficult, short-read technologies largely avoid this issue.

Several algorithms can estimate genomic copy numbers from the "relative enrichment" of reads over the average background of diploid loci. MrsFAST and mrCaNaVaR have been developed to study genome-wide copy numbers (Alkan et al. 2009; Hach et al. 2010). In three human genomes, these algorithms identified *DUX4* as the gene with the highest overall number (Alkan et al. 2009). In HapMap individual Yoruba NA18507, mrCaNaVaR counted 97 total diploid copies for *DUX4* (Alkan et al. 2009). This agrees reasonably well with a *DUX4*-copy estimate for NA18507 (82 copies) that was based on PFGE (Online Resource 2a).

We used these tools to test *DUXC* copy number in *B. taurus* by remapping publicly available Illumina data to a custom-built reference (see "Materials and methods"). We included *DUXA* and *ZAR1*, a known single-copy locus in cow (Uzbekova et al. 2006), as controls. The analysis takes into account bias from over- and underrepresentation of sequences of extreme GC contents (Online Resource 2b). Online Resource 3 summarizes our copy-number estimates for datasets from four different breeds. While both *ZAR1* and *DUXA* were assigned copy-number calls of around two, the high copy-number predictions for *DUXC* ranged from 174 (Sahiwal breed) to more than 400 (N'Dama breed). We interpreted these results as order-of-magnitude and evidence for *DUXC* amplification rather than precise estimates of copy number. Nonetheless, this means that *DUXC* is present at a copy number similar to that of *DUX4* in humans.

The *DUXC* copies are arranged in large tandem arrays

The bioinformatics analysis does not reveal whether the hundreds of *DUXC* copies are dispersed as single sequences across the genome, or whether they are clustered in a single locus such as a tandem-array macrosatellite. We tested the BFF3 cell line with PCR primers facing in opposite orientations in *DUXC* (Fig. 2a). We observed a product of 576 bp (data not shown), as would be predicted by a tandem-array organization of adjacent repeats (confirmed by Sanger sequencing, data not shown). PCR and trace archive data allowed the repeat unit size to be defined as 5.89 kb, and showed "inverted" mate pairs, where the end sequences of clones were orientated in opposite directions, indicating a tandem-array arrangement.

PFGE and Southern blotting using enzymes that have no restriction site within *DUXC* (*Bgl*II, *Hin*dIII and *Eco*RV) excises large fragments of 1–2 Mb (Fig. 2b and c). *Hin*dIII and *Eco*RV generate additional fragments of around 85 and 242 kb, and a size decrease in one of the large fragments. Thus, this individual animal carries large arrays of 150–350 repeats, which is consistent with our bioinformatics analysis of other individuals. There is one predicted *Bln*I site per *DUXC* repeat. Accordingly, linear gel electrophoresis followed by Southern blot shows a single strong band, migrating at the size of one repeat (Fig. 2d). There are two predicted cut sites each for *Bam*HI and *Pvu*II which generate bands of 3.9 or 4.3 kb, consistent with a head-to-tail orientation (Fig. 2a, d).

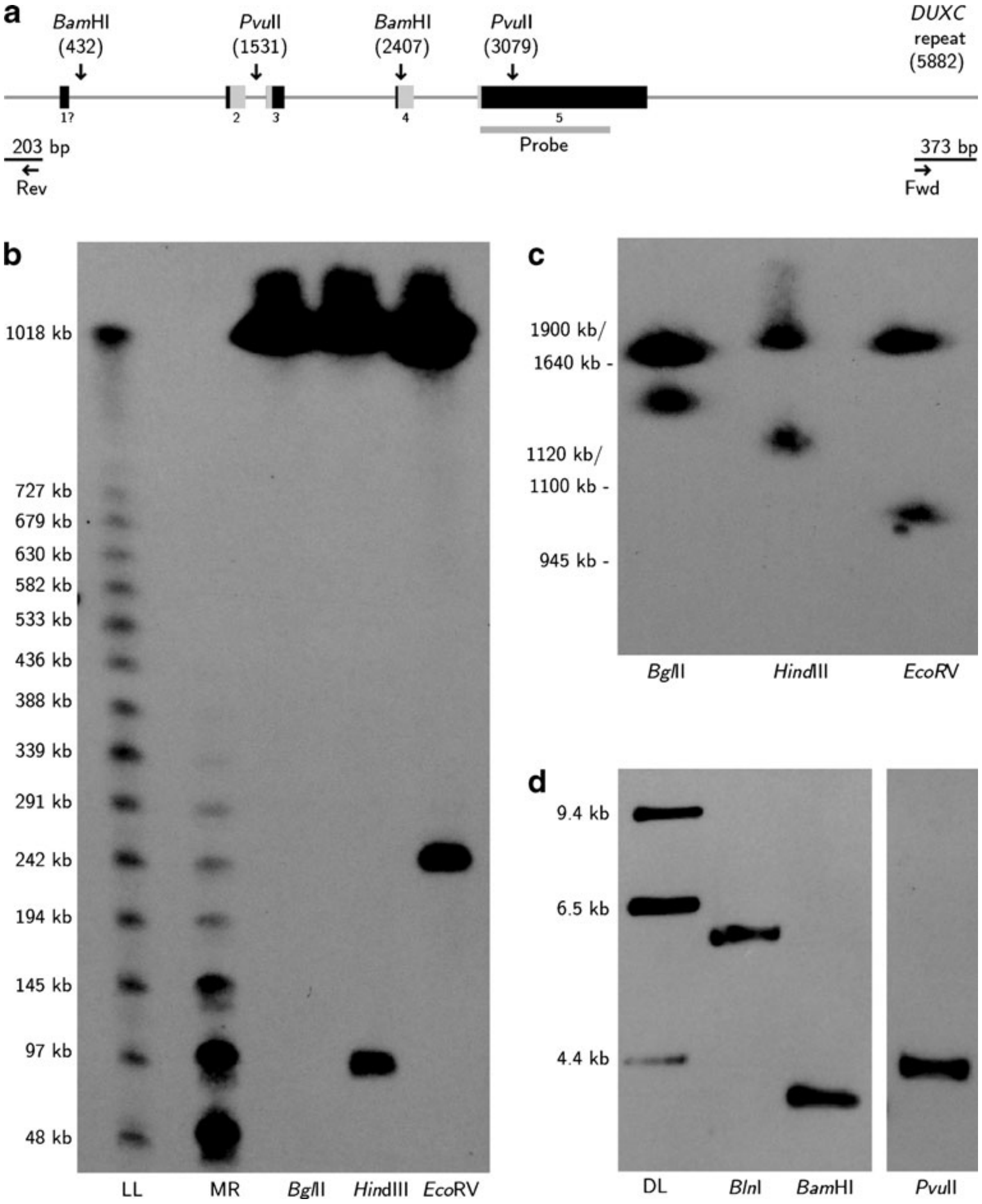

Fig. 2 *DUXC* in cow is present in large tandem arrays. All gels were subjected to Southern blotting using the cow *DUXC* probe. **a** Diagram of one repeat unit of the *DUXC* sequence. Positions of restriction enzyme sites for Southern blots in parentheses. Exons are represented by *boxes*, homeoboxes are highlighted in *blue*. If repeats are arranged in tandem, a digest would yield positive fragments of 3907 bp with *Bam*HI. 4334 and 1548 bp bands are produced by *Pvu*II, but only the large fragment is detected by this probe. Primer positions for the repeat-junction spanning PCR are shown (total product size 576 bp). **b** Southern blot using enzymes predicted not cut within the *DUXC* array (PFGE parameters 15 °C for 48 h with 5–120 s ramp at 4.5 V/cm). Large positive fragments >1 Mb are unresolved under these conditions. *Hin*dIII and *Eco*RV digests also yield smaller fragments of around 85 kb and 242 kb. **c** PFGE using parameters to improve resolution of the large fragments (16 °C for 26 h with 8–120 s ramp at 4.5 V/cm). **d** Linear gel electrophoresis shows two digests with a single-cutter (*Bln*I) and two double-cutters (*Bam*HI and *Pvu*II) that yield bands of the expected sizes of 5.9, 3.9, and 4.3 kb, respectively. *LL* lambda ladder, *MR* mid range marker, *DL* DIG ladder

The *DUXC* macrosatellite in cow is located in a pericentromeric region

We hybridized metaphase chromosomes of two species of cattle with our *DUXC* probe. In *B. taurus*, we observed a single strong signal in the chromosome BTA7q12 pericentromeric region (Fig. 3a). This also confirmed that the signals on our pulsed-field gels represent a single *DUXC* locus. In river buffalo (*B. bubalis*), there was a strong signal at the homoeologous locus in the pericentromeric region of BBU9q12 (Fig. 3b).

Other Laurasiatheria also have *DUXC* tandem arrays

To see if the genomic organization of *DUXC* was conserved in other species, we analyzed reference genomes. In the *B. taurus* assembly, BLASTn detects a single *DUXC* sequence at the tip of chromosome 7, which agrees with the FISH signal (Fig. 4). It is not surprising that the reference contains only a single repeat compared to the hundreds we observed by PFGE, as reads from identical repeats either "collapse" onto one single sequence upon assembly or are excluded from the assembly. Therefore, we independently assembled cow *DUXC* Sanger sequences from the NCBI trace archive. This also indicated a high copy number and a tandem-array organization (Online Resource 4).

Using a similar analysis, we found numerous examples of *DUXC* tandem arrays in other members of the Laurasiatheria. The dolphin assembly has two separate contigs, each containing four tandem *DUXC* repeats, and we also observed inverted mate-pair traces for this species. The alpaca genome also contains a contig with tandemly arrayed *DUXC* copies (Fig. 4 and Online Resource 4). Both the pig and dog reference genomes contain a single copy of their respective *DUXC* ortholog in the subtelomeric regions on chromosome 17 (Fig. 4), but these are probably also collapsed tandem arrays, which is supported by trace archive data with multiple inverted mate pairs in both of these species. Although the horse reference genome contains a single *DUXC* copy on chromosome 1, our trace archive interrogation identified several others, for which we also observed inverted mate pairs (Online Resource 4). Interestingly, we also found a tandem array *DUXC* locus in sloth, a member of the Xenarthra (Fig. 4). Together, this suggests that the tandem-array organization of *DUXC* is conserved throughout the Laurasiatheria and Xenarthra.

Additionally, there is a preferential localization of both *DUX4* and *DUXC* to telomeric or pericentromeric regions. However, for most Afrotheria *DUX4* and Laurasiatheria *DUXC* homologs, the short sequence contigs preclude any analysis of synteny. Although pig, dog, and cow *DUXC* loci have been assigned to chromosomes in Ensembl (17ter, 17ter, and 7ter, respectively), there was no conservation of synteny between these regions.

The recent arrival of *DUX4* on 4qter in primates means that for primate *DUX4* and Laurasiatheria *DUXC*, the observed subtelomeric localization is not due to chromosomal synteny. Therefore, this preferential localization could be due to other mechanisms such as convergent evolution, or a mechanistic pressure.

Discussion

Conservation of the genomic organization between *DUX4* and *Dux*

We have shown that the unusual genomic organization of *DUX4* is found not only in primates but also in the Afrotheria. The murine *Dux* locus shares both sequence and genomic organization with *DUX4* (Clapp et al. 2007), but does not share any synteny with primate *DUX4*. Although the *Dux* tandem arrays are not in a telomeric or pericentromeric location, they lie adjacent to a murine-specific chromosomal fusion point (Clapp et al. 2007). According to the NCBI m37 assembly, the genomic sequence adjacent to *Dux*

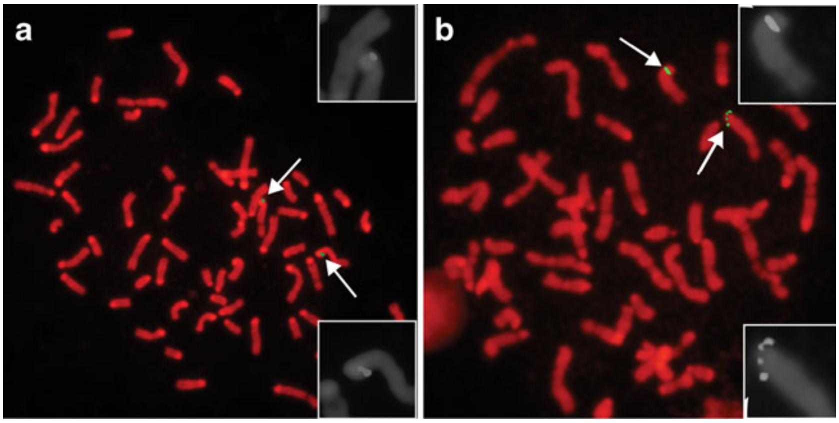

Fig. 3 The *DUXC* array in cow is a single locus near pericentromeres. **a** RPBI banding and overlaid FITC signals with a *DUXC* probe show that the *DUXC* array in cow localizes to a single locus in the pericentromeric region on 7q12. **b** In River buffalo, the probe hybridizes to the homoeologous locus on 9q12. There is only one hybridizing locus in each species, indicating that the *DUXC* macrosatellite is a single locus containing many copies

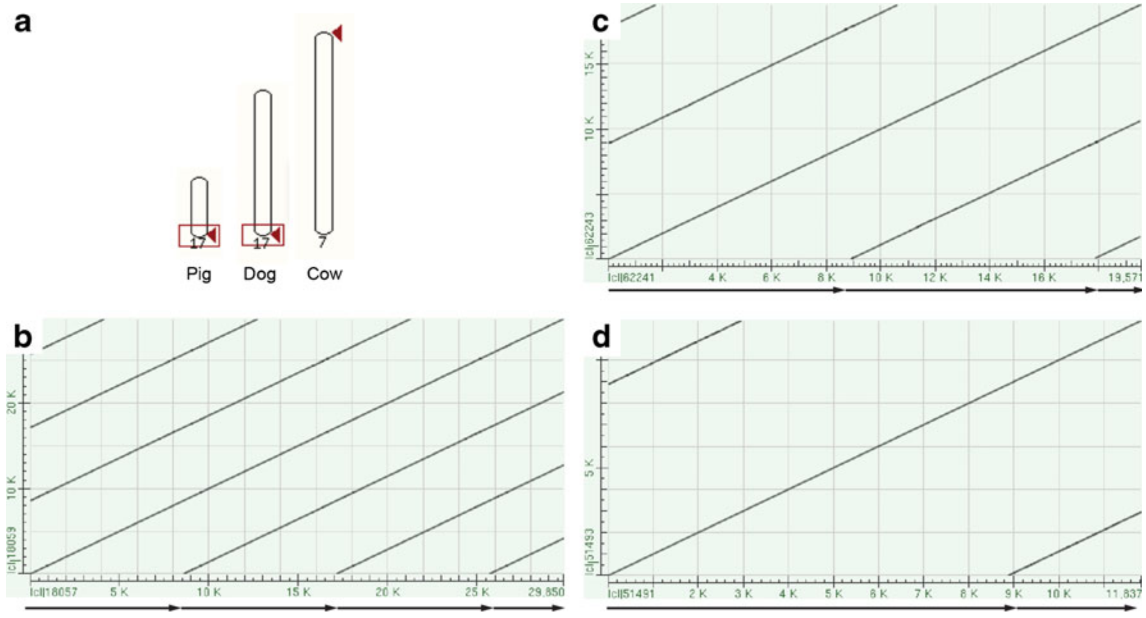

Fig. 4 Telomeric *DUXC* tandem arrays are present in other Laurasiatheria. **a** Ensembl BLAST analysis of the *DUXC* orthologs from cow, pig, and dog all show loci at telomeric regions. **b**–**d** Dot plots of three *DUXC* loci. The sequences were aligned against themselves using BLAST with default parameters. *Arrow* indicates orientation of transcription. For example, **b** shows that the dolphin contig 166 contains four *DUXC* copies in tandem, with a single repeat unit size of ~8.5 kb. **c** shows the Alpaca *DUXC* with three 8.5 kb repeats in tandem. **d** shows a dot plot of two tandem repeats of *DUXC* in sloth, a member of the Xenarthra

contains approximately 500 bp of a degenerate (TTAGGG)n array (10:57582100-57582613), indicative of a recent subtelomeric origin (Flint et al. 1997). This illustrates both the high mobility and plasticity of *DUX*-containing macrosatellites, and the difficulty of assigning orthology.

Origins of intronless *DUX* macrosatellites

Genomes that contain *DUX4* or *Dux* arrays appear to lack *DUXC*, and vice versa. As *DUX4* and *Dux* lack introns but share highly similar homeodomain sequences as well as the C-terminal transcriptional activation domain with *DUXC*, these intronless genes probably arose by retrotransposition from an ancestral *DUXC* gene. Unexpectedly, we have found that *DUXC* is also a telomeric/pericentromeric high-copy macrosatellite, which suggests a simple model for the observed distribution of these genes in mammals (Fig. 5). In this model, *DUX4* (and *Dux*) arose multiple times independently in different mammalian lineages, with the *DUXC* retrotranspositions occuring not at random genomic sites but at the parental *DUXC* macrosatellite. Although in most cases, processed retrogenes insert into random genomic locations and are often “dead on arrival” (Vinckenbosch et al. 2006), reverse-transcripts can also displace their parental gene (Derr and Strathern 1993) by repair of double strand breaks through homologous recombination (Hu 2006). The high *DUXC* copy number would have provided many homologous targets for recombination; additionally, high-copy genes may be highly expressed in the germline and provide many cDNA template molecules, with ample opportunity for retrogenes to arise (Vinckenbosch et al. 2006; Fink 1987). Although little is known about the expression of *DUX* genes in humans, robust *DUX4* expression has

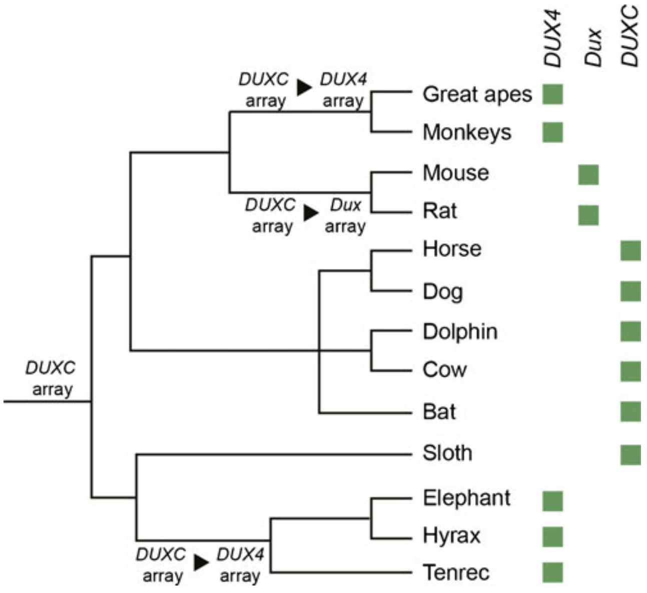

Fig. 5 A new model for the evolution of *DUX* gene macrosatellites. The genome of the common ancestor of placental mammals contained a *DUXC* macrosatellite. In some mammalian lineages, spliced *DUXC* sequences were retrotransposed back into the array, resulting in intron loss by gene conversion followed by array homogenization. This was the origin of *DUX4* macrosatellites, which thus displaced *DUXC* macrosatellites in primates and Afrotheria, while Laurasiatheria and Xenarthra maintained the *DUXC* array

been shown in human testis by Snider et al. (2010), who also reported *DUXC* transcripts in dog testis. Similarly, the mouse *Duxbl* homolog is expressed in testis and ovary (Wu et al. 2010).

In our model, *DUXC* already existed as a high-copy tandem array in the common ancestor of placental mammals. A retrocopy integration in the germline of an intronless, spliced *DUXC* sequence into a single *DUXC* repeat unit could be followed by the local spreading and homogenization of the intronless variant through the rest of the tandem repeats. Such array homogenization ("concerted evolution") is known to occur in tandem arrays like the rDNA clusters (Ganley and Kobayashi 2007).

According to this model, *DUX4*, *Dux*, and *DUXC* did not need to acquire their tandem array structures independently. Replacement of the parental *DUXC* gene in this manner would also explain why mammals have either *DUX4/Dux* or *DUXC*, but never both (this work; Leidenroth and Hewitt 2010). Thus, it may be more appropriate to think of the *DUX4/Dux* retropositions as gene conversions leading to intron loss of *DUXC* (Hu 2006), which makes *DUXC* and *DUX4* effectively "retro-orthologs".

We found no evidence for conservation of synteny for the *DUXC* orthologs and Afrotheria *DUX4*, but telomeric and subtelomeric regions are well known for their plasticity (Mefford and Trask 2002) and are therefore often poorly integrated in genome assemblies. This also means that lower-resolution techniques such as chromosome painting are unlikely to answer the question of synteny between these *DUX* genes. However, the impending arrival of long-read sequencing could soon offer a useful alternative tool.

The unusual genomic arrangement of *DUXC*, *DUX4*, and *Dux* appears to be conserved throughout the placental mammals. There must be mechanistic or selective pressures maintaining all of these unusual tandem arrays in mammals, but these are currently unknown.

Acknowledgments Thanks to Phillipa Muston (University of Nottingham, UK) for donating the Fluorescein High-Prime lambda probe and Ramiro Alberio (University of Nottingham, UK) for donating the BFF3 cow cell line. We are grateful to Bill Wickstead (University of Nottingham, UK) and Willem Rens for critical reading of the manuscript and Can Alkan (University of Washington, USA) for help with the mrsFAST and mrCaNaVaR analysis. We gratefully acknowledge funding for a PhD studentship for AL from the Muscular Dystrophy Campaign (UK), grants from the Muscular Dystrophy Association (USA), and FSH Society Inc (USA), and grant 077121 from the Wellcome Trust.

References

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41 (10):1061–1067. doi:10.1038/ng.437

Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Methods 8(1):61–65. doi:10.1038/nmeth.1527

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27(2):573–580

Clapp J, Mitchell LM, Bolland DJ, Fantes J, Corcoran AE, Scotting PJ, Armour JAL, Hewitt JE (2007) Evolutionary conservation of a coding function for D4Z4, the tandem DNA repeat mutated in facioscapulohumeral muscular dystrophy. Am J Hum Genet 81 (2):264–279. doi:10.1086/519311

Clark LN, 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(3):180–189

de Greef JC, Lemmers RJL, van Engelen BGM, Sacconi S, Venance SL, Frants RR, Tawil R, van der Maarel SM (2009) Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. Hum Mutat 30(10):1449–1459

Derr LK, Strathern JN (1993) A role for reverse transcripts in gene conversion. Nature 361(6408):170–173. doi:10.1038/361170a0

Ferguson-Smith MA, Trifonov V (2007) Mammalian karyotype evolution. Nat Rev Genet 8(12):950–962. doi:10.1038/nrg2199

Fink GR (1987) Pseudogenes in yeast? Cell 49(1):5–6

Flint J, Bates GP, Clark K, Dorman A, Willingham D, Roe BA, Micklem G, Higgs DR, Louis EJ (1997) Sequence comparison of human and yeast telomeres identifies structurally distinct subtelomeric domains. Hum Mol Genet 6(8):1305–1313

Ganley AR, Kobayashi T (2007) Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. Genome Res 17(2):184–191. doi:10.1101/gr.5457707

Grewal PK, van Deutekom JC, Mills KA, Lemmers RJ, Mathews KD, Frants RR, Hewitt JE (1997) The mouse homolog of FRG1, a candidate gene for FSHD, maps proximal to the myodystrophy mutation on chromosome 8. Mamm Genome 8(6):394–398

Hach F, Hormozdiari F, Alkan C, Birol I, Eichler EE, Sahinalp SC (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Meth 7(8):576–577. doi:10.1038/nmeth0810-576

Hewitt JE, Lyle R, Clark LN, Valleley EM, Wright TJ, Wijmenga C, van Deutekom JCT, Francis F, Sharpe PT, Hofker M, Frants RR, Williamson R (1994) Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular-dystrophy. Hum Mol Genet 3(8):1287–1295

Hu K (2006) Intron exclusion and the mystery of intron loss. FEBS Lett 580(27):6361–6365. doi:10.1016/j.febslet.2006.10.048

Iannuzzi L, Di Berardino D (2008) Tools of the trade: diagnostics and research in domestic animal cytogenetics. J Appl Genet 49 (4):357–366. doi:10.1007/BF03195634

Kawamura-Saito M, Yamazaki Y, Kaneko K, Kawaguchi N, Kanda H, Mukai H, Gotoh T, Motoi T, Fukayama M, Aburatani H, Takizawa T, Nakamura T (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(13):2125–2137. doi:10.1093/hmg/ddl136

Leidenroth A, Hewitt JE (2010) A family history of DUX4: phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. BMC Evol Biol 10:364. doi:10.1186/1471-2148-10-364

Lemmers RJL, de Kievit P, van Geel M, van der Wielen MJ, Bakker E, Padberg GW, Frants RR, van der Maarel SM (2001) Complete allele information in the diagnosis of facioscapulohumeral muscular dystrophy by triple DNA analysis. Ann Neurol 50(6):816–819

Lemmers RJ, van Overveld PG, Sandkuijl LA, Vrieling H, Padberg GW, Frants RR, van der Maarel SM (2004) Mechanism and timing of mitotic rearrangements in the subtelomeric D4Z4 repeat involved in facioscapulohumeral muscular dystrophy. Am J Hum Genet 75(1):44–53. doi:10.1086/422175

Lemmers R, van der Vliet PJ, Klooster R, Sacconi S, Camano P, Dauwerse JG, Snider L, Straasheijm KR, van Ommen GJ, Padberg GW, Miller DG, Tapscott SJ, Tawil R, Frants RR, van der Maarel SM (2010a) A unifying genetic model for facioscapulohumeral muscular dystrophy. Science 329(5999):1650–1653. doi:10.1126/science.1189044

Lemmers R, van der Vliet PJ, van der Gaag KJ, Zuniga S, Frants RR, de Knijff P, van der Maarel SM (2010b) Worldwide population analysis of the 4q and 10q subtelomeres identifies only four discrete interchromosomal sequence transfers in human evolution. Am J Hum Genet 86(3):364–377. doi:10.1016/j.ajhg.2010.01.035

Lyle R, Wright TJ, Clark LN, Hewitt JE (1995) 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(3):389–397

Mefford HC, Trask BJ (2002) The complex structure and dynamic evolution of human subtelomeres. Nat Rev Genet 3(2):91–102. doi:10.1038/nrg727

Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16(3):1215–1215

Rens W, Fu B, O'Brien PC, Ferguson-Smith M (2006) Cross-species chromosome painting. Nat Protoc 1(2):783–790. doi:10.1038/nprot.2006.91

Rossi M, Ricci E, Colantoni L, Galluzzi G, Frusciante R, Tonali PA, Felicetti L (2007) The facioscapulohumeral muscular dystrophy region on 4qter and the homologous locus on 10qter evolved independently under different evolutionary pressure. BMC Med Genet 8. doi:10.1186/1471-2350-8-8

Rudd MK, Endicott RM, Friedman C, Walker M, Young JM, Osoegawa K, de Jong PJ, Green ED, Trask BJ, Progra NCS (2009) Comparative sequence analysis of primate subtelomeres originating from a chromosome fission event. Genome Res 19(1):33–41. doi:10.1101/gr.083170.108

Snider L, Geng LN, Lemmers RJLF, Kyba M, Ware CB, Nelson AM, Tawil R, Filippova GN, van der Maarel SM, Tapscott SJ, Miller DG (2010) Facioscapulohumeral dystrophy: incomplete suppression of a retrotransposed gene. PLoS Genet 6(10):e1001181

Tremblay DC, Alexander G, Moseley S, Chadwick BP (2010) Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. BMC Genomics 11:632. doi:Doi10.1186/1471-2164-11-632

Uzbekova S, Roy-Sabau M, Dalbies-Tran R, Perreau C, Papillier P, Mompart F, Thelie A, Pennetier S, Cognie J, Cadoret V, Royere D, Monget P, Mermillod P (2006) Zygote arrest 1 gene in pig, cattle and human: evidence of different transcript variants in male and female germ cells. Reprod Biol Endocrinol 4:12. doi:10.1186/1477-7827-4-12

van der Maarel SM, Tawil R, Tapscott SJ (2011) Facioscapulohumeral muscular dystrophy and DUX4: breaking the silence. Trends Mol Med 17(5):252–258. doi:10.1016/j.molmed.2011.01.001

Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A 103(9):3220–3225. doi:10.1073/pnas.0511307103

Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G (2008) Analysis of the largest tandemly repeated DNA families in the human genome. BMC Genomics 9:533. doi:10.1186/1471-2164-9-533

Wijmenga C, Hewitt JE, Sandkuijl LA, Clark LN, Wright TJ, Dauwerse HG, Gruter AM, Hofker MH, Moerer P, Williamson R, van Ommen GJB, Padberg GW, Frants RR (1992) Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. Nat Genet 2(1):26–30

Wu SL, Tsai MS, Wong SH, Hsieh-Li HM, Tsai TS, Chang WT, Huang SL, Chiu CC, Wang SH (2010) Characterization of genomic structures and expression profiles of three tandem repeats of a mouse double homeobox gene: Duxbl. Dev Dyn 239(3):927–940

Zeng W, de Greef JC, Chen Y-Y, Chien R, Kong X, Gregson HC, Winokur ST, Pyle A, Robertson KD, Schmiesing JA, Kimonis VE, Balog J, Frants RR, Ball AR, Lock LF, Donovan PJ, van der Maarel SM, Yokomori K (2009) Specific loss of histone H3 lysine 9 trimethylation and HP1gamma/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD). PLoS Genet 5(7):e1000559