



# Deep characterization of a common D4Z4 variant identifies biallelic *DUX4* expression as a modifier for disease penetrance in FSHD2

Richard JLF Lemmers<sup>1</sup> · Patrick J van der Vliet<sup>1</sup> · Judit Balog<sup>1</sup> · Jelle J Goeman<sup>2</sup> · Wibowo Arindrarto<sup>3</sup> · Yvonne D Krom<sup>4</sup> · Kirsten R Straasheijm<sup>1</sup> · Rashmie D Debipersad<sup>1</sup> · Gizem Özel<sup>1</sup> · Janet Sowden<sup>5</sup> · Lauren Snider<sup>6</sup> · Karlien Mul<sup>7</sup> · Sabrina Sacconi<sup>8</sup> · Baziel van Engelen<sup>7</sup> · Stephen J Tapscott<sup>6</sup> · Rabi Tawil<sup>5</sup> · Silvère M van der Maarel<sup>1</sup>

Received: 24 July 2017 / Revised: 7 September 2017 / Accepted: 9 September 2017 / Published online: 21 November 2017  
© European Society of Human Genetics 2018

## Abstract

Facioscapulohumeral muscular dystrophy is caused by incomplete repression of the transcription factor *DUX4* in skeletal muscle as a consequence of D4Z4 macrosatellite repeat contraction in chromosome 4q35 (FSHD1) or variants in genes encoding D4Z4 chromatin repressors (FSHD2). A clinical hallmark of FSHD is variability in onset and progression suggesting the presence of disease modifiers. A well-known cis modifier is the polymorphic *DUX4* polyadenylation signal (PAS) that defines FSHD permissive alleles: D4Z4 chromatin relaxation on non-permissive alleles which lack the *DUX4*-PAS cannot cause disease in the absence of stable *DUX4* mRNA. We have explored the nature and relevance of a common variant of the major FSHD haplotype 4A161, which is defined by 1.6 kb size difference of the most distal D4Z4 repeat unit. While the short variant (4A161S) has been extensively studied, we demonstrate that the long variant (4A161L) is relatively common in the European population, is capable of expressing *DUX4*, but that *DUX4* mRNA processing differs from 4A161S. While we do not find evidence for a difference in disease severity between FSHD carriers of an 4A161S or 4A161L allele, our study does uncover biallelic *DUX4* expression in FSHD2 patients. Compared to control individuals, we observed an increased frequency of FSHD2 patients homozygous for disease permissive alleles, and who are thus capable of biallelic *DUX4* expression, while *SMCHD1* variant carriers with only one permissive allele were significantly more often asymptomatic. This suggests that biallelic *DUX4* expression lowers the threshold for disease presentation and is a modifier for disease severity in FSHD2.

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1038/s41431-017-0015-0>) contains supplementary material, which is available to authorized users.

✉ Richard JLF Lemmers  
R.J.L.F.Lemmers@lumc.nl

✉ Silvère M van der Maarel  
S.M.van\_der\_Maarel@lumc.nl

<sup>1</sup> Department of Human Genetics, Leiden University Medical Center, Leiden, Netherlands

<sup>2</sup> Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, Netherlands

<sup>3</sup> Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, Netherlands

## Introduction

Facioscapulohumeral dystrophy (FSHD; OMIM 158900 & 158901) is one of the more common hereditary myopathies characterized by a descending pattern of muscle weakness

<sup>4</sup> Department of Neurology, Leiden University Medical Center, Leiden, Netherlands

<sup>5</sup> Neuromuscular Disease Unit, Department of Neurology, University of Rochester Medical Center, Rochester, NY, USA

<sup>6</sup> Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

<sup>7</sup> Neuromuscular Centre Nijmegen, Department of Neurology, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands

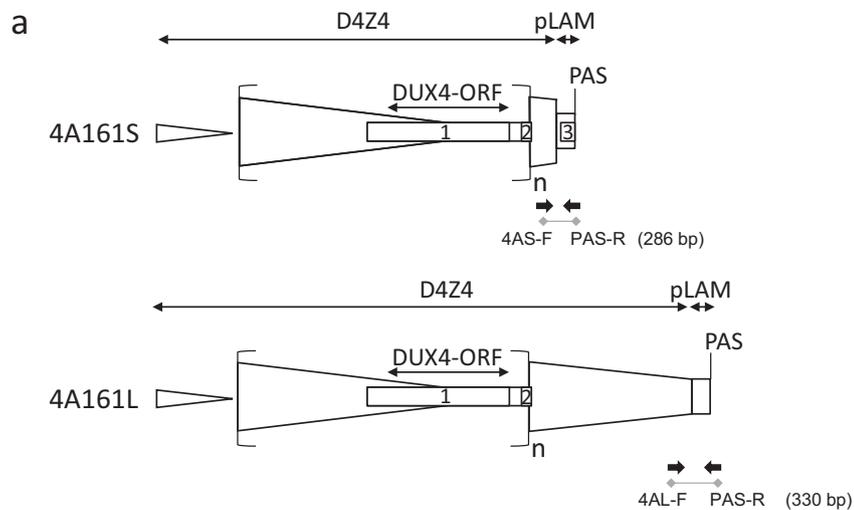
<sup>8</sup> Centre de référence des Maladies neuromusculaires and CNRS UMR6543, Nice University Hospital, Nice, France

starting in the face and progressing into the upper extremity muscles. The disease typically starts in the second decade of life and often the pattern of muscle involvement is asymmetric. With disease progression, also other muscles may become involved. Despite the well recognizable core phenotype, the clinical hallmark of the disease is the large variability in onset and progression, both between and within families. While approximately one fifth of gene carriers will become wheelchair-dependent, an equal proportion of gene carriers will remain asymptomatic or minimally affected throughout life (reviewed in [1]).

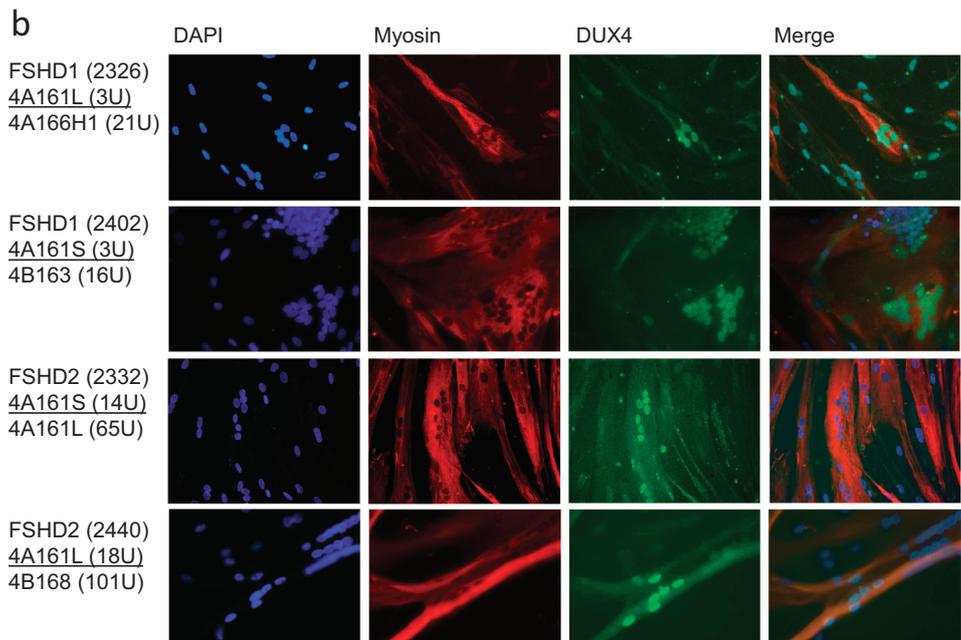
FSHD is caused by incomplete repression of the *DUX4* retrogene (OMIM 606009) in skeletal muscle [2–6]. *DUX4* is a transcription factor normally expressed in the luminal cells of the testis and in cleavage stage embryos but repressed in somatic tissue such as skeletal muscle [7–10].

Incomplete repression, as observed in FSHD muscle, leads to the activation and de-activation of a complex pattern of transcriptional pathways eventually leading to muscle cell death [11–15]. A copy of the *DUX4* retrogene is embedded within each unit of the D4Z4 macrosatellite repeat array on chromosome 4, which is polymorphic in repeat number and normally varies between 8–100 copies [3, 4, 16, 17]. Either by contraction to a size of 1–10 units (FSHD1), or by monoallelic variants in genes encoding somatic D4Z4 chromatin repressors such as SMCHD1 and DNMT3B (FSHD2), somatic repression of *DUX4* is incomplete leading to presence of *DUX4* protein in a proportion of myonuclei [18–20]. Consequently, a prominent difference between FSHD1 and FSHD2 is that while in FSHD1 the D4Z4 chromatin changes are restricted to the contracted D4Z4 array, in FSHD2 these chromatin changes can be

**Fig. 1 a** Graphical representation of the D4Z4 locus on 4A161S and 4A161L alleles. The 4A161 variants vary in size by 1.6 kb in the most distal and partial D4Z4 unit, which is followed by identical pLAM sequences encompassing the *DUX4* polyadenylation signal. The size of the complete D4Z4 unit preceding the most distal partial unit is identical between both variants and all other D4Z4 haplotypes. The length variation of D4Z4 repeat arrays is determined by variability in the number of 3.3 kb internal D4Z4 units (*n*). The *DUX4* gene structure including exons 1–3 and the open reading frame (ORF) is indicated, as well as the D4Z4 and pLAM region and the intron-intron composition in the 4A161S allele. The primers that are used for the genotyping PCR (4A161S and 4A161L) in Fig. 2 are also indicated.



**b** Immunofluorescence staining of myotubes derived from FSHD1 and FSHD2 patients. The size in units (U) and haplotype background of both D4Z4 repeat arrays on chromosome 4 is indicated and the pathogenic alleles are underlined. *DUX4* protein is expressed from both 4A161S (2402 and 2332) and 4A161L (2332 and 2440) alleles. DAPI staining in blue, myosin in red and *DUX4* in green. Scale bar is indicated



observed on the D4Z4 arrays of both chromosomes 4, and on the highly homologous D4Z4 arrays on chromosome 10 [21, 22]. We and others have reported overlap in repeat sizes between controls and FSHD1 in the range of 8–10 units [16, 23]. This size range is marked by high clinical variability in disease onset, with affected individuals often showing more D4Z4 hypomethylation at their FSHD allele than asymptomatic family members carrying the same disease allele [24–26]. We also identified FSHD2 patients (thus having a variant in *SMCHD1* or *DNMT3B*) who also carry a permissive allele with 8–10 D4Z4 units [19, 27]. These patients typically show an earlier disease onset and more rapid progression in comparison to their family members carrying the D4Z4 disease allele without the FSHD2 variant suggesting that both disease genes can also act as modifier genes in FSHD1.

The *DUX4* retrogene within the D4Z4 unit lacks a stabilizing polyadenylation signal (PAS) but can make use of an additional exon immediately distal to the repeat that contains a stabilizing *DUX4* PAS [2]. The *DUX4*-PAS is polymorphic in nature and this explains the chromosome- and haplotype-specificity of FSHD: a *DUX4*-PAS is present on 4qA chromosomes but absent from 4qB chromosomes and from chromosome 10 [5, 28]. Hence, FSHD patients have at least one 4qA chromosome from which *DUX4* can be expressed in muscle, either because of repeat array contraction, or because of a variant in *SMCHD1* or *DNMT3B*.

The 4qA and 4qB haplotypes can be divided in several subgroups based on a simple sequence length polymorphism (SSLP) proximal to the repeat (Fig. S1) [16, 29]. The sequence of the most distal D4Z4 unit immediately proximal to the pLAM sequence is virtually identical for almost all 4qA haplotypes [5]. The exception to this is the most common haplotype 4A161, accounting for 85% of FSHD patients and defined by an SSLP length of 161 nucleotides and the distal 4qA variant. Earlier studies have shown that the 4A161 haplotype can end in two genetically distinct variants, 4qA and 4qA-L (long), which diverge by a 1.6 kb size difference of the most distal partial D4Z4 unit, followed by identical pLAM sequences in which the *DUX4*-PAS is located (Fig. 1a and Fig. S1) [17].

Although it has been described that D4Z4 repeat contraction results in FSHD1 from the 4A161 haplotype with the canonical 4qA distal end (hereafter called 4A161S) and with an extended 4qA-L distal end (hereafter called 4A161L) [5], the relative frequency and repeat size distribution of 4A161S and 4A161L and the identity of the *DUX4* transcript from the 4A161L haplotype is currently unknown. In this study we determined the frequency and distribution of the two 4A161 haplotypes in the control and FSHD1 and FSHD2 patient population, determined the nature of the *DUX4* transcripts from the 4A161L allele, and explored potential clinical consequences.

## Results

### DUX4 expression in FSHD1 and FSHD2 myotubes with either a 4A161S or 4A161L pathogenic allele

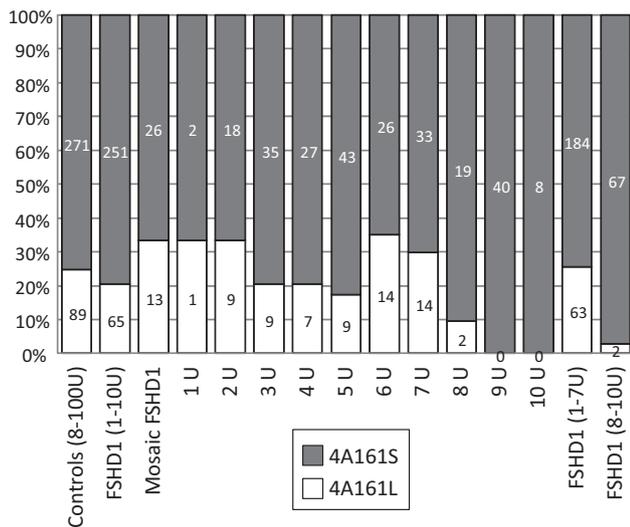
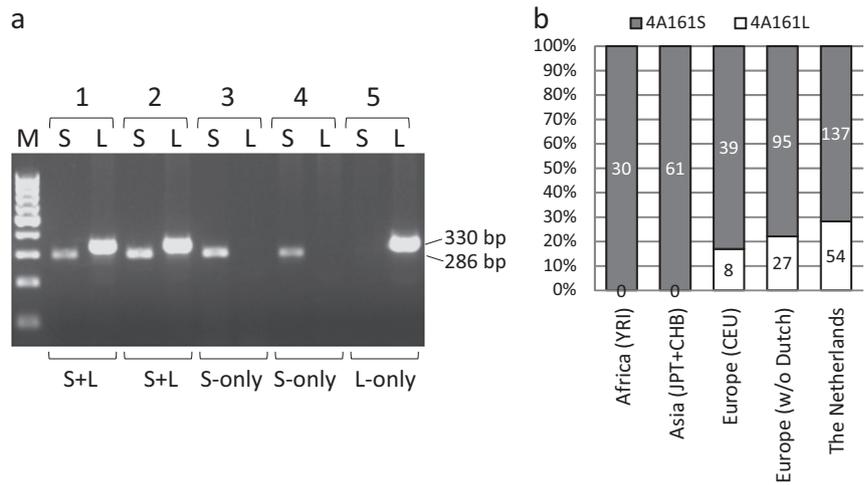
Based on protein prediction, 4A161S and 4A161L alleles both contain an identical *DUX4* open reading frame (ORF) of 424 amino acids since exon 1, which contains the entire ORF, is not affected by the distal variation (Fig. 1a). Previously, we showed that FSHD1 is caused by contracted D4Z4 repeats on both 4A161S and 4A161L alleles [5]. Here, we performed immunofluorescence analysis on myotube cultures derived from FSHD1 and FSHD2 patients, in which the pathogenic repeat is either on a 4A161S or 4A161L background. As shown in representative immunofluorescence images in Fig. 1b, a comparable pattern of *DUX4* positive nuclei can be detected in all situations suggesting that *DUX4* expression is possible from both 4A161 variants and that both variants indeed encode a *DUX4* protein that can be detected with the same antibody.

### 4A161L haplotype is enriched in the European population

To gain insight into the frequency of the 4A161S and 4A161L haplotypes in different populations, we developed 4A161S and 4A161L-specific primer pairs (Fig. 1a) and analysed independent HAPMAP samples from the Nigerian (Yoruban-YRI,  $n = 60$ ; totalling 30 independent 4A161 alleles), Caucasian (CEU,  $n = 58$ ; 50 independent 4A161 alleles), Japanese (JPT,  $n = 44$ ; 31 independent 4A161 alleles), and Han Chinese (CHB,  $n = 45$ ; 30 independent 4A161 alleles) panels for the presence of both variants. Detailed D4Z4 genotyping, including repeat sizing and haplotype analysis, of these panels was already performed in a previous study [29]. The specificity of the primer pairs is shown in Fig. 2a, demonstrating that chromosome 10qA alleles (carrying a 4A161S-like distal end, but with specific sequence variants) and 4qB alleles are not being amplified. Based on available HAPMAP samples carrying one or two 4A161 alleles, we found that the 4A161L haplotype is absent in the African and Asian samples from the HAPMAP panels and can be found at a frequency of 17.0% (8/47) in the Caucasian (of European origin) HAPMAP samples (Fig. 2b).

For sixteen 4A161S, eleven 4A161L and 32 other 4qA, 4qB, and 10q haplotypes we cloned and sequenced the distal variable region from independent individuals according to our previous study [5], and confirmed that independent alleles of the same haplotype have identical sequences (Table S1). We did not find the 4qA-L extension in combination with other 4qter haplotypes, as expected

**Fig. 2** 4A161S and 4A161L PCR and frequency of 4A161 haplotypes in different populations. **a** Genotyping PCR for the discrimination between 4A161S (286 bp) and 4A161L (330 bp) alleles. Samples 1 and 2 carry a 4A161S (S) and 4A161L (L) allele, samples 3 and 4 carry one 4A161S and one 4qB allele and sample 5 carries one 4A161L and one 4qB allele. The position of the primers and the amplicon size are shown in Fig. 1a. **b** Bar diagram showing the distribution of 4A161S and 4A161L alleles in three HAPMAP populations and two in-house control populations. The in-house populations were divided in European (without Dutch) and Dutch control individuals



**Fig. 3** Bar diagram showing the distribution of 4A161S and 4A161L alleles in the European controls and FSHD1 patients (non-mosaic and mosaic patients). Only unique control and FSHD1 alleles are included. After binning of 4A161S and 4A161L FSHD1 alleles into 1-7U and 8-10U bins, we observed a significant paucity ( $p = 8.18 \times 10^{-5}$ ) of 8–10 D4Z4 repeat units on a 4A161L allele. On the right side of the graph results are shown for binning to individual D4Z4 repeat array sizes

from the evolution of D4Z4 haplotypes [29]. We expanded the control group (not linked to FSHD) with samples collected in our lab to 667 independent Caucasian samples: 370 from the Netherlands and 297 with non-Dutch Caucasian background. We identified 517 4A161 alleles out of a total of 1334 4q alleles (39.4%), for which we further determined the distal variation in 313 samples. We observed slightly more 4A161L alleles in the Dutch cohort compared

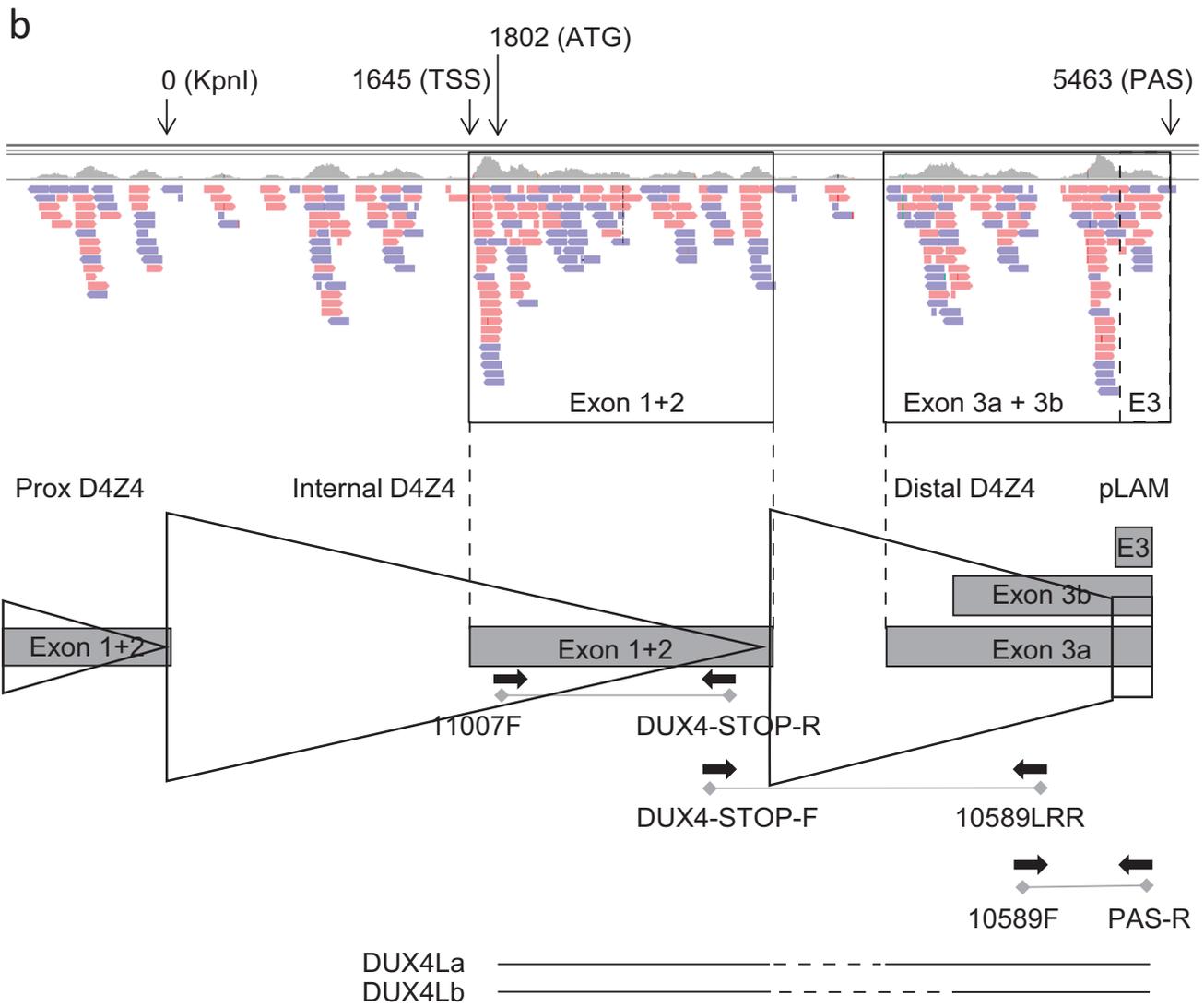
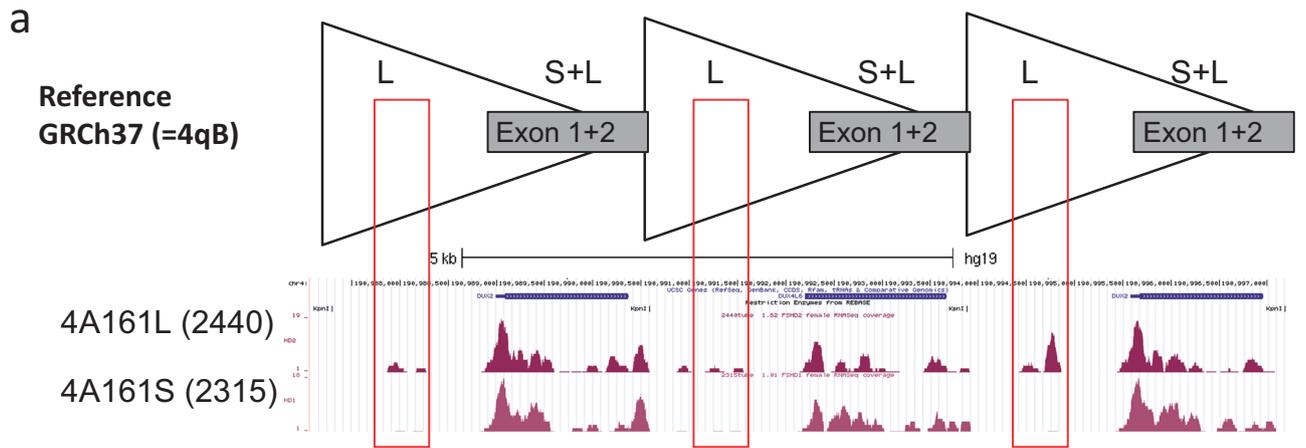
to non-Dutch Caucasian cohort and the CEU HAPMAP samples, but this difference is not significant ( $p = 0.199$ ). Upon merging all Caucasian samples, we identified 89 (24.7%) 4A161L and 271 (75.3%) 4A161S alleles (Fig. 2b).

**Upper-sized FSHD alleles have a different 4A161S and 4A161L distribution**

We next studied the frequency of 4A161S and 4A161L in the Caucasian FSHD population. We analysed 720 FSHD1 patients and 49 unaffected carriers from 432 families and identified a 4A161 FSHD allele in 411 families (95.4%). The other FSHD1 families carried an FSHD allele of a rarer haplotype (4A157, 4A159, 4A168, 4A166H, or D4F104S1 deletion). For 315 families with a 4A161 FSHD allele we determined the distal variation of the disease allele and identified 251 4A161S and 65 4A161L (20.6%) alleles. Within the FSHD1 cohort, we identified 39 mosaic FSHD1 patients with a de novo contraction of D4Z4 and for them we found comparable ratios (33.3% 4A161L), suggesting that both 4A161 variants are equally recombinogenic (Table S2). We then binned the 4A161S and 4A161L FSHD1 alleles, according to the D4Z4 repeat array size. For FSHD1 alleles with a D4Z4 array between 1–7 units, we found ratios between 4A161S and 4A161L comparable to the ratio in the Caucasian control population. However, we found a significantly ( $p = 8.18 \times 10^{-5}$ ) smaller number of FSHD1 patients with an array of 8–10 D4Z4 repeat units on a 4A161L allele (Fig. 3). We also analyzed the FSHD allele in 14 Asian FSHD1 patients and found thirteen 4A161S alleles and one hybrid 4A166H allele, like in the Asian HAPMAP populations (results not shown).

To explain the lack of 4A161L alleles with 8–10 D4Z4 repeat units in the FSHD1 cohort, we analysed the D4Z4 array size distribution for 4A161S and 4A161L alleles ( $n = 360$ ) in the European controls in more detail. Like in the

FSHD1 cohort, in controls we found a paucity of 4A161L alleles with 8–10 D4Z4 repeat units and found that this 4A161L paucity extends to repeat sizes of 14 units. The number of 4A161L alleles ( $n = 1$ ) with array size between 8



and 14 units in controls is significantly smaller ( $p$ -value = 0.0109) compared to the number of 4A161S alleles ( $n = 28$ ) (Figure S2a and S2b). We also determined the allelic variation of 4A161 alleles in FSHD2 patients. We studied 79 FSHD2 families based on D4Z4 hypomethylation and the presence of an *SMCHD1* variant totalling 137 affected and unaffected *SMCHD1* variant carriers. We determined the allelic variation of 4A161 alleles in 72 affected individuals with an ACSS  $\geq 50$  or in probands with a clear FSHD phenotype but without documented ACSS. We found an increased frequency of 4A161 alleles with D4Z4 repeat arrays  $\leq 14$  units in FSHD2 patients, but we did not find a significant paucity of 4A161L alleles in this group compared to  $>15$  units repeat arrays ( $p = 0.301$ , (Fig. S2c)).

### Methylation analysis of 4A161S and 4A161L alleles

D4Z4 has a high GC content (73%) and displays high CpG methylation in somatic cells of control individuals. In FSHD, D4Z4 methylation is reduced. Moreover, CpG methylation analysis of the *FseI* restriction site in D4Z4 has shown D4Z4 repeat size-dependent methylation in controls, and in FSHD1 and FSHD2 individuals. Consequently, in general the shortest D4Z4 repeat sizes have the lowest CpG methylation. Based on the correlation between repeat size and methylation we were able to consistently predict the methylation of D4Z4 given the repeat sizes. The predicted methylation value was almost identical to the observed methylation in controls and FSHD1 patients that carried a D4Z4 repeat between 1–6 units [26]. However, in FSHD1 patients carrying a D4Z4 repeat between 7–10 units we found that the observed methylation is significantly lower than the predicted methylation, as expressed in the Delta1 value. This suggests that for repeat sizes between 7–10 units an additional reduction in CpG methylation contributes to pathogenic chromatin relaxation associated with somatic *DUX4* expression. We confirmed the significant reduced Delta1 value for FSHD1 patients with arrays of 7–10 D4Z4 units in this study (1–6 unit, mean  $-0.8\%$ ; 7–10 unit, mean  $-4.6\%$ ;  $p = 0.0002$ ) (Fig. S3).

A recent bisulfite sequencing methylation analysis studied the methylation of the pLAM region distal to D4Z4 without discriminating between 4A161S and 4A161L [30]. Re-analysis of the most informative CpG from this study (CpG6) showed that D4Z4 repeat size-dependent methylation can be observed in both 4A161S and 4A161L alleles and that the methylation values for both 4qA variants are comparable at identical repeat sizes (Fig. S4).

### Transcript analysis

Based on the DNA sequence, we expected that 4A161L and 4A161S alleles possibly encode for an identical *DUX4* transcript, where the distal D4Z4 extension would represent an extended *DUX4* intron 2. However, RT-PCR analysis using RNA from myotube cultures of FSHD1 patients with a contraction on a 4A161L allele using previously described 4A161S transcript-based primers [5] failed to yield a product. As the reverse PCR primer of this primer pair straddles exons 2 and 3 of the 4A161S *DUX4* transcript, the RT-PCR failure on 4A161L RNA suggested a different *DUX4* intron–exon structure.

To gain more insight in the transcript from 4A161L alleles we aligned *DUX4* 4A161S and 4A161L RNAseq reads from a previous study [13] to the reference genome of chromosome 4. The standard HG19 and HG38 reference genomes do not contain a 4A161 sequence, but the 4qB refseq sequence from hg19 suggested that a portion of the extended D4Z4 region in patients with a pathogenic 4A161L allele is transcribed rather than being intronic (Fig. 4a). Therefore, we aligned the 4A161L RNAseq reads in the integrative genomics viewer (IGV) to the 4A161L reference sequence available in HG38 Patch 7 (KQ983258.1). The new alignment suggested the presence of a *DUX4* transcript with an exon 3 that is proximally extended compared to the *DUX4* transcript from 4A161S alleles (Fig. 4b, alignment in IGV). Based on the IGV alignment RT-PCR primers were designed, which identified the presence of two alternative 4A161L transcripts, which we named *DUX4La* and *DUX4Lb*, where the longest and most prevalent isoform is *DUX4La* (Fig. 4b). The amplification of 4A161L is

◀ **Fig. 4 a** Representative RNAseq coverage tracks in the UCSC genome browser of differentiated myoblast obtained from FSHD patients expressing *DUX4* from a 4A161S, or from a 4A161L allele. The reads are aligned to the 4qB-type D4Z4 locus of reference genome GRCh37 (hg19), as the 4qA sequences are not standard available in GRCh37 and GRCh38. The indicated region includes three D4Z4 units depicted on top of the UCSC track. The boxed region (red) shows the extra reads that are specific for the *DUX4* transcript from 4A161L allele. **b** IGV alignment of *DUX4* reads from 4A161L allele aligned to 4A161L allele available in GRCh38.p7 (KQ983258.1). The 4A161L sequence shown covers a complete internal D4Z4 unit, the most distal partial D4Z4 unit and the pLAM region distal to D4Z4. Forward reads are shown in pink and reverse reads in purple. The exons 1 and 2 region that is similar between 4A161S and 4A161L *DUX4* transcripts is shown. The 4A161L-specific exons 3a and exon 3b region is boxed and shows several specific reads, the 4A161S-specific exon 3 region is shown with a dotted box (E3). An overview of the two major *DUX4* transcripts (*DUX4La* and *DUX4Lb*) produced from a derepressed 4A161L allele is shown below with uninterrupted lines representing exons and interrupted lines representing introns. The primers and amplicons that were used for the identification of the 4A161L transcripts are indicated. The position of the *KpnI* restriction site, *DUX4* transcriptional and translation start sites (TSS and ATG) and polyadenylation signal (PAS) are indicated as well as the intron–exon composition for *DUX4La* and *DUX4Lb*

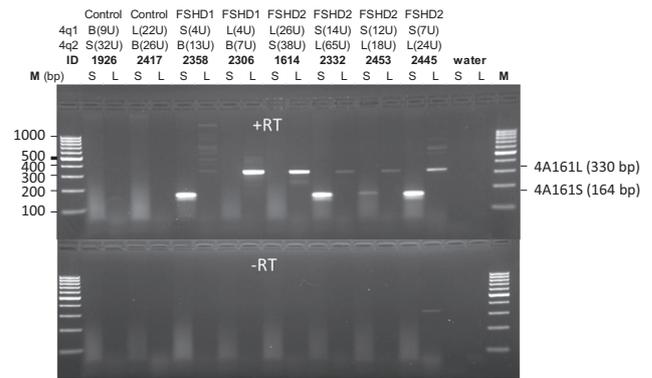
hampered by the high GC-content in D4Z4 and specifically the high Cytosine content (70%) in the region 599–737 bp proximal to the *DUX4* polyadenylation signal. These problems are encountered for PCR and Sanger sequencing of DNA and RNA, and also for 3' RACE reaction. The identified 4A161L *DUX4* transcripts have been submitted to Genbank under accession numbers MF422078 (*DUX4La*) and MF422079 (*DUX4Lb*). 4A161L sequences with intron-exon structure and exon numbering of the *DUX4La* and *DUX4Lb* transcript can be found in Genbank under accession numbers MF693913 and KQ983258.1 (Supplementary Table 1). The gene features of Genbank accession number MF693913 are depicted in detail in Fig. S5.

### Pathogenicity 4A161S and 4A161L alleles in FSHD1

We analyzed the correlation between clinical severity and repeat array size in FSHD1 patients carrying a 4A161S or 4A161L FSHD allele. For this we focussed on FSHD1 alleles that range between 1–7 D4Z4 units, because of the 4A161L paucity in the 8–10 D4Z4 units size range. We selected one patient per family of at least 25 years of age and excluded mosaic FSHD1 patients, and used the age corrected clinical severity score (ACSS) [31, 32]. We did not find a significant difference ( $p = 0.0611$ ) in clinical severity between carriers of a 4A161S or a 4A161L FSHD allele (Fig. S6).

### Expression from two permissive alleles in FSHD2

The severity of FSHD2 is in part determined by the size of the permissive allele and in part by the nature of the *SMCHD1* variant [26]. In most FSHD2 cases the size of the pathogenic 4qA allele ranges between 8 and 20 units, while *SMCHD1* variant carriers with 4qA alleles > 20 units often remain asymptomatic. Some FSHD2 patients carry two permissive alleles and in these cases we expect that *DUX4* is exclusively expressed from the shortest array in patient derived myotubes, if the second allele contains >20 D4Z4 units. To study this we used the RT-PCR primers specific for either 4A161S or 4A161L-derived *DUX4* transcripts in an endpoint PCR and analyzed myotube cultures obtained from FSHD2 patients carrying a 4A161S and a 4A161L permissive allele (Fig. 5). Negative control myotube cultures were from unaffected controls carrying normal-sized D4Z4 arrays on a 4A161S or 4A161L allele and PCR-specificity was confirmed on myotube cultures from FSHD1 patients carrying a D4Z4 repeat array contraction on either a 4A161S or a 4A161L FSHD allele. Surprisingly, in 3 out of 4 FSHD2 myotube cultures we observed expression from both the 4A161S as from the 4A161L allele, probably as a consequence of the overall D4Z4 derepression. Although



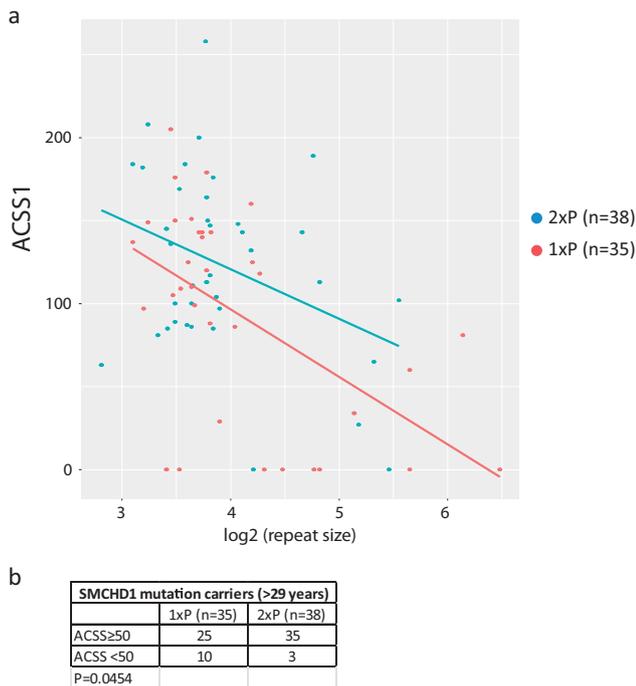
**Fig. 5** Endpoint 4A161S- and 4A161L-specific RT-PCR on differentiated myoblasts obtained from controls and FSHD patients. For both chromosomes 4 (4q1 and 4q2) the haplotype is indicated (4A161S [S], 4A161L [L] or 4qB [B]) and the repeat size is shown between parenthesis. The specificity of the 4A161S and 4A161L primers is shown for the FSHD1 patients carrying a 4A161S or a 4A161L FSHD allele (positive controls) and the control individuals (negative controls). Three of the four FSHD2 patients show expression from both the 4A161S and the 4A161L allele. Sequences of the products were confirmed by Sanger sequencing

**Table 1** Ratio between individuals that carry one (1xP) and two (2x) permissive alleles in controls and FSHD2 patients carrying a permissive allele with  $\leq 30$  D4Z4 units. Compared to control individuals we find in FSHD2 patients a significant skewing ( $p = 0.00142$ ) of individuals carrying two permissive alleles. This difference is significant ( $p = 0.00142$ ) when the shortest permissive allele is 30 units or smaller

	1xP	2xP
Control	151 (60.9%)	97 (39.1%)
FSHD2	38 (40.9%)	55 (59.1%)

these are non-quantitative RT-PCR experiments, the signal intensities of the RT-PCR products suggest that in all cases *DUX4* expression levels are highest from the allele with the shortest D4Z4 repeat array.

This finding suggests that the presence of two permissive 4qA alleles might increase *DUX4* expression levels as a consequence of bi-allelic expression and may result in a more severe phenotype in some FSHD2 cases. This could cause an excess of FSHD2 individuals carrying two permissive alleles because of increased penetrance. To address this issue, we determined the haplotype of the second D4Z4 allele in FSHD2 patients and control individuals carrying a permissive allele with  $\leq 30$  D4Z4 units. As shown in Table 1, 38 FSHD2 patients (40.9%) carried one permissive allele, while 55 (59.1%) carried two permissive alleles. In controls this ratio was 151 (60.9%) against 97 (39.1%), which is a significant difference ( $p = 0.00142$ ). This suggests that there is an excess of FSHD2 patients carrying two permissive alleles.



**Fig. 6** **a** Scatter plot showing the ACSS for *SMCHD1* variant carriers that carry one (1xP, red dots) or two (2xP, blue dots) permissive alleles. All carriers were 30 years or older. The X-axis shows the log<sub>2</sub> value of the D4Z4 repeat size of shortest permissive allele. For both situations we find that the disease severity drops by the size of the repeat ( $p = 0.0008$ ). We did not find a significant difference in the severity between carriers of one or two permissive alleles. Regression lines are shown in blue (2xP) and red (1xP). **b** We found significantly more unaffected or mildly affected (ACSS < 50) carriers with only one permissive allele ( $p = 0.0454$ )

To further study the clinical consequences of having a second permissive allele, we compared the ACSS in FSHD2 patients of 30 years and older [33] carrying one or two permissive alleles. Despite the biallelic expression of *DUX4* in muscle cell cultures of FSHD2 individuals with two permissive alleles, we did not observe a significant difference ( $p = 0.126$ ) in clinical severity between individuals carrying one or two permissive 4qA alleles (Fig. 6a). Interestingly, we observe an opposite ratio of one permissive allele (1xP) vs. two permissive alleles (2xP) between affected *SMCHD1* variant carriers (1xP:2xP = 25:35) and unaffected or mildly affected carriers (1xP:2xP = 10:3). This difference is significant ( $p = 0.0454$ ) and suggests that for FSHD2 the number of permissive alleles may code-terminate the likelihood of developing FSHD (Fig. 6b).

## Discussion

FSHD is clinically characterized by substantial variability in onset and progression of the disease, even within FSHD1 families where individuals at risk carry the same D4Z4

repeat array contraction [34–36]. This suggests that modifiers increase the risk of developing FSHD and determine the progression of the disease. The existence of such modifiers is supported by different lines of evidence. For example, while the diagnostic cut-off for FSHD1 has been defined as maximally 10 D4Z4 repeat units, 1–3% of the European population carries a D4Z4 repeat array of 8–10 units on a FSHD-permissive haplotype [16, 23]. In addition, while variants in the genes encoding the chromatin repressors *SMCHD1* and *DNMT3B* have been shown to cause FSHD2 when combined with an intermediate-sized D4Z4 repeat array (11–20 units) on a FSHD permissive chromosome 4, variants in the same genes have also shown to act as modifier genes for disease severity in some FSHD1 families that carry a permissive D4Z4 repeat array of 8–10 units [18, 19, 27, 37]. Therefore, it is plausible that additional cis (i.e. variants in 4qter) and trans (i.e., variants in D4Z4 chromatin factors) modifiers exist that influence FSHD disease penetrance and progression, in addition to other factors such as perhaps *DUX4* pathway modifiers.

Perhaps the most influential *cis* modifier is the *DUX4*-PAS variant, which is essential for the stabilization of the *DUX4* transcript in muscle cells [5]. Consequently, D4Z4 repeat array contractions on chromosomes that lack the *DUX4*-PAS, such as 4B chromosomes, do not result in the production of *DUX4* protein and disease presentation. Moreover, additional polymorphic sequences immediately downstream of the *DUX4*-PAS were recently identified that are critical for *DUX4* cleavage and polyadenylation [38]. In this study we molecularly characterized and studied the clinical consequence of another major variation in the most common FSHD permissive haplotype 4A161. Previous studies had already demonstrated that the partial most distal D4Z4 repeat unit can differ in size in patients and controls, either being 0.3 kb (4A161S) or 1.9 kb (4A161L) long [5, 17], but the frequency and distribution of both variants over the different haplotypes was unknown. Moreover, while the *DUX4* locus in the 4A161S allele has been extensively studied, the 4A161L has not been studied in this respect.

Our population studies indicate that the 4A161L variant is a rather recent modification of the 4A161S allele that is relatively common in the European population but absent, or very uncommon in the Asian and African populations. This is consistent with our previous study into the evolutionary network of 4qter, in which we predicted 4A161S and 4A161L to be closely related [29]. RNA and protein studies showed that both variants can produce stable *DUX4* transcripts in muscle cell cultures, that both variants display a similar repeat size dependent methylation pattern, and that there is no evidence that either variant affects function more than the other based on the ACSS. Both variants, however, do differ in the usage of the splice acceptor site of exon 3.

4A161L uses two alternative 3' splice sites that are not present on the 4A161S allele. As a consequence two alternative DUX4 transcripts can be produced from 4A161L alleles, DUX4La and DUX4Lb, that carry a 5' extended exon 3 compared to the DUX4 transcript from 4A161S alleles.

A noteworthy observation is that some FSHD2 patients can express DUX4 from both alleles. Biallelic DUX4 expression is surprising since earlier studies have shown that intermediate-sized D4Z4 alleles ranging from 11–20 units are particularly sensitive to the consequences of heterozygous SMCHD1 variants, while only few FSHD2 patients carry a permissive allele that exceed 20 units [26]. Perhaps other mechanisms are at play in these patients which make both alleles susceptible to DUX4 expression. We did not find evidence that biallelic DUX4 expression leads to a more severe disease presentation. Rather, our study in the FSHD2 population does suggest that the presence of two permissive alleles, and therefore being at risk for biallelic DUX4 expression, increases the penetrance of FSHD. This suggests that the threshold for DUX4 expression in vivo is higher in FSHD2 individuals consistent with the relatively milder disease presentation in FSHD2 [33].

Another surprising observation is the significant lack of 4A161L alleles with D4Z4 repeat sizes of 8–10 units in the FSHD1 population, but not in the 1–7 units range. This may suggest that 4A161L alleles are less susceptible to DUX4 expression and FSHD than 4A161S alleles in this size range. However, we did not find a difference in the ACSS between FSHD1 patients carrying either a 4A161S or a 4A161L allele with 1–7 D4Z4 units. We also did not find a difference between the frequency of 4A161S and 4A161L alleles for de novo mosaic D4Z4 contractions leading to FSHD1, suggesting that both haplotypes are equally recombinogenic. Our population studies, however, revealed 10.3% 4A161S alleles with repeat sizes of 8–14 D4Z4 units and only 1.1% 4A161L alleles with this size range. This parallel between controls and FSHD1 therefore suggests that in this size range there is no selection pressure but it is rather a consequence of the recent evolutionary origin of this haplotype.

FSHD2 patients show a different size distribution of permissive (mostly 4A161) alleles. While in controls 8–14 units compared to >14 unit repeat sizes on 4A161S and 4A161L alleles are less common to rare, they are very common in FSHD2 patients (Figure S2) as this size range is highly sensitive to D4Z4 chromatin relaxation and DUX4 derepression as a consequence of SMCHD1 variants [26]. Strikingly, for unknown reasons, we do not observe the paucity of 8–14 units 4A161L alleles in FSHD2 patients.

Previously, we and others recognized a difference in the predictive value of a genetic diagnosis for FSHD1, where 1–7 D4Z4 unit arrays typically result in an FSHD

phenotype, while the clinical consequences of arrays 8–10 units are highly unpredictable and can also be found in the control population [16, 23]. Recently, several groups showed that for carriers of 7–10 unit D4Z4 repeat the degree of CpG methylation to some extent correlates with disease penetrance [24–27]. Apparently, in this size range, FSHD mostly develops in a complex fashion with additional requirement factors that further reduce the methylation of D4Z4 beyond that what can be expected from the contraction itself. The current finding of a shortage of 4A161L compared to 4A161S alleles in the 8–10 units size range, corroborates this finding. D4Z4 array sizes between 8–10 units on 4A161S alleles can be found in 1–3% of the control population and in combination with D4Z4 hypomethylation result in FSHD. However, on 4A161L alleles these repeat sizes are extremely rare and therefore will unlikely appear in the hypomethylation dependent 8–10 unit FSHD population, because of the digenic nature.

Overall, recent developments in our understanding of the genetic and epigenetic factors that govern DUX4 repression in skeletal muscle may ask for a revision of the definition of FSHD1 and FSHD2. We propose that FSHD1 and FSHD2 belong to the same disease continuum in which FSHD1 represents a spectrum of the disease caused by D4Z4 contractions on a permissive D4Z4 array to relatively short D4Z4 repeat sizes (1–7U) with limited additive effect of variation in epigenetic modifiers of the D4Z4 chromatin structure. FSHD2 is typically more often found in individuals with longer D4Z4 array sizes (8–20U) in combination with the contribution of additional factors that determine the likelihood of DUX4 expression in skeletal muscle, such as genetic variants that influence the activity or levels of the D4Z4 chromatin repressors SMCHD1 and DNMT3B.

## Materials and methods

### Subjects

This study was approved by the relevant Medical Ethical Committees from participating institutions. For controls, we analyzed HAPMAP samples from the Nigerian (Yoruban-YRI,  $n = 60$ ), Caucasian (CEU,  $n = 58$ ), Japanese (JPT,  $n = 44$ ) and Han Chinese (CHB,  $n = 45$ ) panels. The Caucasian control (not linked to FSHD) samples were expanded with 667 independent Caucasian samples collected at the LUMC, 370 from the Netherlands and 297 with other European background. For FSHD1, we analyzed 720 patients and 48 unaffected carriers (family members from FSHD1 patients) from 431 families. Thirty-nine of the 720 patients carried a mosaic FSHD allele. For FSHD2, we analyzed 79 families with one or more FSHD2 patients. In total we studied 139 carriers of an SMCHD1 variant, of

which 110 were affected and 29 unaffected. Eleven of these 29 unaffected cases carried two non-permissive alleles. For most individuals the D4Z4 repeat sizes and haplotypes have been described previously [26, 29]. Clinical evaluation was performed by an experienced neurologist after informed consent. For the clinical severity we used the age corrected severity score (ACSS), based on the 10-scale Ricci score [ACSS = (Ricci score/age at examination) × 1000] [31, 32]. The standardized clinical form is available at the website of the Fields Center for FSHD Research ([www.urmc.rochester.edu/fieldscenter/](http://www.urmc.rochester.edu/fieldscenter/)).

### Myoblasts

For immunofluorescence and *DUX4* transcription analysis we used 9 primary myoblast cell cultures. For all myoblasts we analyzed the D4Z4 repeat sizes and haplotype as described for blood. For the different experiments we used two control lines (1926 and 2417), 4 FSHD1 lines (2326, 2358, 2402, and 2417) and 5 FSHD2 lines (1614, 2332, 2440, 2445, and 2453). Detailed genotype and methylation information can be found in Table S3.

### D4Z4 repeat sizing, haplotype analysis and methylation analysis

All cohorts were analyzed for D4Z4 repeat size, genetic background (haplotype) and CpG methylation at D4Z4. Genomic DNA (gDNA) was isolated from PBMCs or from lymphoblastoid cell lines for HAPMAP samples. The sizing of the D4Z4 repeats was done by pulsed field gel electrophoresis, as described previously [26, 29]. Haplotypes were determined by Southern blot hybridization of HindIII digested genomic DNA with probes A and B in combination with PCR-based SLP analysis according to previously described protocols [29]. Methylation at the D4Z4 repeat arrays was determined on blood-derived genomic DNA by Southern blot and the methylation sensitive restriction enzyme FseI, after which the delta1 value was calculated as described previously [26]. Detailed protocols are freely available from the Fields Center website ([www.urmc.rochester.edu/fields-center/](http://www.urmc.rochester.edu/fields-center/)).

### Muscle cell culture and immunofluorescence staining

Human primary myoblast cell lines originated from the University of Rochester bio repository. Primary myoblasts were cultured in DMEM/F-10 media (#31550, Gibco/Life Technologies, Bleiswijk, The Netherlands) supplemented with 20% heat inactivated fetal bovine serum (FBS #10270, Gibco/Life Technologies), 1% penicillin/streptomycin (#15140, Gibco/Life Technologies), 10ng/ml rhFGF (#G5071, Promega, Leiden, The Netherlands) and 1 μM

dexamethasone (#D2915, Sigma-Aldrich, Zwijndrecht, The Netherlands). For differentiation the myoblasts were cultured for 2–4 days in DMEM/F-12 Glutamax media (#31331, Gibco/Life Technologies) containing 1% penicillin and streptomycin and 2% KnockOut serum replacement formulation (#10828, Gibco/Life Technologies). Immunofluorescence staining on differentiated myoblasts was performed as described previously [39]. The following antibodies were used; E5-5 (1:100), primary rabbit-*DUX4* antibody directed against the C-terminal region of human *DUX4* [40] with Alexa Fluor 488-conjugated donkey anti-rabbit (1:400; Life Technologies) and MF-20 anti-myosin (1:100, Dako North America, Carpinteria CA, USA) with Alexa-594-conjugated goat anti-mouse (1:400; Life Technologies). The cells were mounted on microscope slides using Aqua Poly/Mount (PolySciences, Warrington PA, U.S.A.) containing 500 ng/ml DAPI.

### Determination of the distal variation at 4A161

D4Z4-pLAM sequences for 4A161S and 4A161L haplotypes were obtained in individuals that carry one 4qA and one 4qB chromosome. We used 4qA specific primers on BlnI-digested genomic DNA to eliminate the 10qA chromosomes and to specifically amplify the 4qA chromosome. To amplify the 3.5 kb D4Z4-pLAM sequence of 4A161S haplotype we used forward primer 9406LRF 5'-AGC GTT CCA GGC GGG AGG GAA G-3' and reverse primer PAS-LRR 5'-CAG GGG ATA TTG TGA CAT ATC TCT GCA CTC ATC-3' and for 3.6 kb D4Z4-pLAM sequence of 4A161L haplotype we used the same reverse primer in combination with forward primer 11007LRF 5'-AGC CCA GGG TCC AGA TTT GGT TTC AG-3' [5]. When individuals carried a 4A161S and 4A161L haplotype, we were not able to generate a 3.5 kb 4A161L PCR fragment because of preferential amplification of the shorter 4A161S PCR fragment. In these cases, we generated a shorter 1.9 kb 4A161L PCR fragment using forward primer 9406LRF in combination with reverse primer PAS-LRR. These PCR reactions were performed on 125–140 ng of genomic DNA, in a solution containing 3.5 μM each primer, 0.2 mM dATP, 0.4 mM dCTP, 0.2 mM dTTP, 0.2 mM dGTP and 0.2 mM 7-deaza-dGTP, 2.5 U of LA-Taq DNA polymerase and supplemented with 2xGC buffer I (TAKARA), in a total volume of 20 μl. The PCR conditions consisted of an initial denaturation step at 94 °C for 5 min., followed by 34 cycles of denaturation at 94 °C for 30 s, annealing at 68 °C for 30 s, and extension at 72 °C for 3 min and 30 s. The final extension time was 10 min. at 72 °C.

For genotyping of the distal variation (4A161S or 4A161L) of 4A161 alleles we developed 4A161S- and 4A161L-specific primer pairs and for both PCR reactions we used 20 ng of genomic DNA in total a volume of 10 μL.

For 4A161S amplicon (286 bp) we used forward primer 4AS-F 5'-CCC GCC CGG GCC CCT GCA-3' and reverse primer PAS-R 5'-GAT CCA CAG GGA GGG GGC ATT TTA-3'. For the PCR reaction of 4A161S, final concentrations are as follows: 1× Accuprime buffer II (Invitrogen), 10% DMSO, 0.25 μM each primer, 0.15 μl AccuPrime Taq Polymerase. The PCR conditions consisted of an initial denaturation step at 94 °C for 2 min. and 30 sec., followed a touchdown PCR: 98 °C, 30 s; 73 > 63 °C, 30 s and 68 °C, 30 s for 11 cycles, followed by 26 cycles at 94 °C, 30 s; 64 °C, 30 s; 68 °C, 30 s. The final extension time was 10 min. at 72 °C. For the 4A161L amplicon (330 bp) we used forward primer 4AL-F 5'-CGA GGA CGG CGA CGG AGA C -3' in combination with reverse primer PAS-R. The final concentrations of the 4A161L PCR are as follows: 1× buffer Dreamtaq, 0.2 mM dNTPs, 0.25 μM each primer, 0.05 units Dreamtaq Polymerase. The PCR conditions consisted of an initial denaturation step at 95 °C for 5 min., followed by 35 cycles of denaturation at 95 °C for 10 s, annealing at 64 °C for 20 s, and extension at 72 °C for 15 s. The final extension time was 5 min. at 72 °C.

### In silico alignment DUX4 reads

Reads were first analysed using FastQC (version 0.10.1). If any known adapters were detected to be overrepresented by FastQC, the full adapter sequence was then used for clipping using the cutadapt tool (version 1.5), setting the minimum read length to 20. In any case, low quality bases in all reads were trimmed using the sickle tool (version 1.33) with default options. If cutadapt was performed before running sickle, the paired reads are processed using a custom script beforehand to ensure the read pairs were still in-sync. Mapping was done to the respective 4A161 allele sequence using the GSNAP tool (version 2014-12-23) with the following flags: -batch 4, -nthreads 8, -novelsplicing 1, and -format sam. The resulting alignments were then compressed, sorted, and indexed, before viewed in the IGV tool.

### Transcription analysis DUX4

Total RNA was extracted using the Qiagen RNeasy isolation kit with DnaseI treatment. The RNA concentration was determined on a ND-1000 spectrophotometer (Thermo Scientific, Wilmington, USA). cDNA was synthesized from 2 μg of total RNA using poly dT primers (Fermentas RevertAid Transcripase kit). After 5 min at 70 °C, the reverse transcription buffer was added, followed by 1 h incubation at 42 °C, 10 min at 70 °C and put on ice. After the reaction, cDNA was treated with RNaseH and 75 μL water was added to an end volume of 100 μL.

For identification of the 4A161L transcript, we used three overlapping amplicons. For the 5' amplicon we used forward primer 11007LRF 5'-AGC CCA GGG TCC AGA TTT GGT TTC AG-3' and reverse primer DUX4-STOP-R 5'-CTA AAG CTC CTC CAG CAG AGC CCG GTA TTC-3', for the central amplicon we used DUX4-STOP-F 5'-GAA TAC CGG GCT CTG CTG GAG GAG CTT TAG-3' and reverse primer 10589LRR 5'-CGG AAG GGA CCC AGG GCG TCG AG-3' and for the 3' amplicon we used 10589F 5'-CGC CCT GGG TCC CTT CCG-3' and reverse primer PAS-R 5'-GAT CCA CAG GGA GGG GGC ATT TTA-3'. For all PCRs we used the following conditions in a 25 μl PCR reaction; input cDNA 2 μl cDNA, 1x Accuprime buffer II (Invitrogen), supplemented with 0.2 mM 7-Deaza-2'-deoxyguanosine, 0.2 μM of each primer and 0.5 μl AccuPrime Taq Polymerase. The PCR conditions consisted of an initial denaturation step at 95 °C for 2 min., followed by 40 cycles of denaturation at 94 °C for 30 s, annealing at 58 °C for 30 s, and extension at 68 °C for 2 min. 30 s. The final extension time was 10 min. at 72 °C. All the primers used were designed using Primer 3 software. To study bi-allelic expression in FSHD2 we used the following PCR conditions. For the 4A161S amplicon (164 bp) we used forward primer DUX4RT-F2 5'- CCC AGG TAC CAG CAG ACC-3' and reverse primer pLAMR4 5'- TCC AGG AGA TGT AAC TCT AAT CCA-3'. In a 15 μl PCR reaction we used 5 microliter cDNA, 1× Sybrgreen PCR mix (buffer and dNTP and Polymerase) and 0.33 μM of each primer. The PCR conditions consisted of an initial denaturation step at 95 °C for 3 min., followed by 35 cycles of denaturation at 95 °C for 10 s and annealing/extension for 30 s at 60 °C. For the 4A161L transcription analysis we used the same conditions and amplicon (330 bp) as we used to determine the 4A161L variation on DNA (primers 4AL-F and PAS-R in a PCR volume of 15 uL with 5 uL cDNA).

### Statistical analysis

To compare the ratio of 4A161S vs. 4A161L alleles between the groups in Figs. 2, 3, 6 and Figure S2, we used a Pearson chi-square test. Yates' continuity correction was used in chi-square tests in 2 × 2 tables.

To model the relationship between the Delta1 value and the repeat length in Figure S3, we fitted a linear mixed model with Delta1 as the response and repeat length as a fixed effect (dichotomized at a threshold of 7.5). The family identifier was included as a random effect in the model.

For FSHD1 patients, severity was modeled as a function of repeat size and 4A161S vs. 4A161L haplotype using a linear mixed model with repeat length and haplotype and their interaction as fixed effects, and with family identifier as random effect. In this model we first tested for the presence of an interaction effect. Since the interaction was not

significant, we left it out of the final model when testing whether haplotype had a significant influence on severity, assuming a common slope between haplotypes.

To test the effect of the number of permissive alleles on the severity in FSHD2 patients a similar linear mixed model was used, now containing the number of permissive alleles, the logarithm of the length of the shortest permissive allele and their interaction as fixed effects and family as random effect. As above, we first tested the interaction effect, leaving it out when it was found to be non-significant. The effect of the number of permissive alleles was subsequently assessed in a simpler model with common slope between groups.

All mixed models were fitted in R 3.3.2 [41] in the MASS library [42] and Wald *t*-tests were used for assessing significance.

## Web Resources

The URLs for data presented herein are as follows:

Detailed information on genotyping protocols, <http://www.urmc.rochester.edu/fields-center/>

GenBank, <http://www.ncbi.nlm.nih.gov/Genbank/>

Primer3, [http://frodo.wi.mit.edu/cgi-bin/primer3/primer3\\_www.cgi](http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi).

## Genbank accession numbers

*DUX4* transcript accession numbers MF422078 (*DUX4La*) and MF422079 (*DUX4Lb*) and for *D4Z4* genomic accession numbers KQ983258.1, MF693913, MF503833, MF503834, MF503835, MF503836, MF741677, MF741678 (4A161L); MF776830, MF776831, MF776832, MF776833, MF776834, MF776835 (4A161S); MF776826, MF776827, MF776828, MF776829 (4A166) and MF776824, MF776825 (4A166H1).

**Acknowledgements** We thank all FSHD families for participating in our studies. This study was supported by funds from the National Institute of Neurological Disorders and Stroke award number P01 NS069539 to SJT, RT and SMvdM, the Prinses Beatrix Spierfonds (award numbers W.OP14-01 and W.OB17-01) to SMvdM, Spieren voor Spieren to BvE and SMvdM, and the European Union Framework Programme 7 (agreement 2012-305121, NEUROMICS) to SMvdM.

## Compliance with ethical standards

**Conflict of interests** The scientific significance and the materials of our report have not been previously published and will not be submitted for publication elsewhere. The authors declare that they have no competing interests.

## References

- Statland J, Tawil R. Facioscapulohumeral muscular dystrophy. *Neurol Clin*. 2014;32:721–8, ix.
- Dixit M, Anseau E, Tassin A, et al. *DUX4*, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of *PITX1*. *Proc Natl Acad Sci USA*. 2007;104:18157–18162.
- Gabriels J, Beckers MC, Ding H, et al. Nucleotide sequence of the partially deleted *D4Z4* locus in a patient with FSHD identifies a putative gene within each 3.3 kb element. *Gene*. 1999;236:25–32.
- Hewitt JE, Lyle R, Clark LN, et al. Analysis of the tandem repeat locus *D4Z4* associated with facioscapulohumeral muscular dystrophy. *Hum Mol Genet*. 1994;3:1287–1295.
- Lemmers RJ, van der Vliet PJ, Klooster R, et al. A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science*. 2010;329:1650–1653.
- Snider L, Asawachaicharn A, Tyler AE, et al. RNA transcripts, miRNA-sized fragments and proteins produced from *D4Z4* units: new candidates for the pathophysiology of facioscapulohumeral dystrophy. *Hum Mol Genet*. 2009;18:2414–2430.
- De, IA, Planet, E, Coluccio, A et al. *DUX*-family transcription factors regulate zygotic genome activation in placental mammals. *Nat Genet*. 2017;49:941–945.
- Geng LN, Yao Z, Snider L, et al. *DUX4* activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Dev Cell*. 2012;22:38–51.
- Hendrickson, PG, Dorais, JA, Grow, EJ et al. Conserved roles of mouse *DUX* and human *DUX4* in activating cleavage-stage genes and *MERV1/HERV1* retrotransposons. *Nat Genet*. 2017;49:925–940.
- Whiddon, JL, Langford, AT, Wong, CJ et al. Conservation and innovation in the *DUX4*-family gene network. *Nat Genet*. 2017;22:4661–4672.
- Block GJ, Narayanan D, Amell AM, et al. *Wnt*/beta-catenin signaling suppresses *DUX4* expression and prevents apoptosis of FSHD muscle cells. *Hum Mol Genet*. 2013;22:4661–4672.
- Bosnakovski D, Xu Z, Gang EJ, et al. An isogenetic myoblast expression screen identifies *DUX4*-mediated FSHD-associated molecular pathologies. *EMBO J*. 2008;27:2766–2779.
- Snider L, Geng LN, Lemmers RJL, et al. Facioscapulohumeral dystrophy: incomplete suppression of a retrotransposed gene. *PLoS Genet*. 2010;6:e1001181.
- Wallace, LM, Garwick, SE, Mei, W et al. *DUX4*, a candidate gene for facioscapulohumeral muscular dystrophy, causes p53-dependent myopathy in vivo. *Ann Neurol*. 2010;69:540–552.
- Jagannathan, S, Shadle, SC, Resnick, R et al. Model systems of *DUX4* expression recapitulate the transcriptional profile of FSHD cells. *Hum Mol Genet*. 2016;25:4419–4431.
- Lemmers RJ, Wohlgemuth M, van der Gaag KJ, et al. Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. *Am J Hum Genet*. 2007;81:884–894.
- van Deutekom JC, Wijmenga C, van Tienhoven EA, et al. FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. *Hum Mol Genet*. 1993;2:2037–2042.
- Lemmers RJ, Tawil R, Petek LM, et al. Digenic inheritance of an *SMCHD1* mutation and an FSHD-permissive *D4Z4* allele causes facioscapulohumeral muscular dystrophy type 2. *Nat Genet*. 2012;44:1370–1374.
- van den Boogaard ML, Lemmers RJ, Balog J, et al. Mutations in *DNMT3B* Modify Epigenetic Repression of the *D4Z4* Repeat and the Penetrance of Facioscapulohumeral Dystrophy. *Am J Hum Genet*. 2016;98:1020–1029.
- Wijmenga C, Hewitt JE, Sandkuijl LA, et al. Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nat Genet*. 1992;2:26–30.

21. de Greef JC, Lemmers RJ, van Engelen BG, et al. Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. *Hum Mutat.* 2009;30:1449–1459.
22. van Overveld PG, Lemmers RJ, Sandkuijl LA, et al. Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. *Nat Genet.* 2003;35:315–317.
23. Scionti I, Fabbri G, Fiorillo C, et al. Facioscapulohumeral muscular dystrophy: new insights from compound heterozygotes and implication for prenatal genetic counselling. *J Med Genet.* 2012;49:171–178.
24. Gaillard, MC, Roche, S, Dion, C et al. Differential DNA methylation of the D4Z4 repeat in patients with FSHD and asymptomatic carriers. *Neurology.* 2014;83:733–742.
25. Jones TI, King OD, Himes CL, et al. Individual epigenetic status of the pathogenic D4Z4 macrosatellite correlates with disease in facioscapulohumeral muscular dystrophy. *Clin Epigenet.* 2015;7:37.
26. Lemmers RJ, Goeman JJ, van der Vliet PJ, et al. Inter-individual differences in CpG methylation at D4Z4 correlate with clinical variability in FSHD1 and FSHD2. *Hum Mol Genet.* 2015;24:659–669.
27. Sacconi S, Lemmers RJ, Balog J, et al. The FSHD2 gene SMCHD1 is a modifier of disease severity in families affected by FSHD1. *Am J Hum Genet.* 2013;93:744–751.
28. Lemmers RJ, de Kievit P, Sandkuijl L, et al. Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. *Nat Genet.* 2002;32:235–236.
29. Lemmers RJLF, van der Vliet PJ, van der Gaag KJ, et al. Worldwide population analysis of the 4q and 10q subtelomeres identifies only four discrete duplication events in human evolution. *Am J Hum Genet.* 2010;86:364–377.
30. Calandra, P, Cascino, I, Lemmers, RJ et al. Allele-specific DNA hypomethylation characterises FSHD1 and FSHD2. *J Med Genet.* 2016;53:348–355.
31. Ricci E, Galluzzi G, Deidda G, et al. Progress in the molecular diagnosis of facioscapulohumeral muscular dystrophy and correlation between the number of KpnI repeats at the 4q35 locus and clinical phenotype. *Ann Neurol.* 1999;45:751–757.
32. van Overveld PG, Enthoven L, Ricci E, et al. Variable hypomethylation of D4Z4 in facioscapulohumeral muscular dystrophy. *Ann Neurol.* 2005;58:569–576.
33. de Greef JC, Lemmers RJ, Camano P, et al. Clinical features of facioscapulohumeral muscular dystrophy 2. *Neurology.* 2010;75:1548–1554.
34. Goto K, Nishino I, Hayashi YK. Very low penetrance in 85 Japanese families with facioscapulohumeral muscular dystrophy 1A. *J Med Genet.* 2004;41:e12.
35. Ricci G, Scionti I, Sera F, et al. Large scale genotype-phenotype analyses indicate that novel prognostic tools are required for families with facioscapulohumeral muscular dystrophy. *Brain.* 2013;136:3408–3417.
36. Salort-Campana E, Nguyen K, Bernard R, et al. Low penetrance in facioscapulohumeral muscular dystrophy type 1 with large pathological D4Z4 alleles: a cross-sectional multicenter study. *Orphanet J Rare Dis.* 2015;10:2.
37. Larsen, M, Rost, S, El, HN et al. Diagnostic approach for FSHD revisited: SMCHD1 mutations cause FSHD2 and act as modifiers of disease severity in FSHD1. *Eur J Hum Genet.* 2014;23:808–816.
38. Peart, N and Wagner, EJ. A distal auxiliary element facilitates cleavage and polyadenylation of Dux4 mRNA in the pathogenic haplotype of FSHD. *Hum Genet.* 2017; e-pub ahead of print 24 May 2017; <https://doi.org/10.1007/s00439-017-1813-8>.
39. Krom YD, Dumonceaux J, Mamchaoui K, et al. Generation of isogenic D4Z4 contracted and noncontracted immortal muscle cell clones from a mosaic patient: a cellular model for FSHD. *Am J Pathol.* 2012;181:1387–1401.
40. Geng LN, Tyler AE, Tapscott SJ. Immunodetection of human double homeobox 4. *Hybridoma.* 2011;30:125–130.
41. R Development Core Team. R: A Language and Environment for Statistical Computing. 2017. Ref Type: Online Source
42. Venables, WN and Ripley, BD. *Modern Applied Statistics with S.* 4th edition ed. Springer: New York, NY, USA, 2002.