High resolution breakpoint junction mapping of proximally extended D4Z4 deletions in FSHD1 reveals evidence for a founder effect.

Richard JLF Lemmers[1]*, Patrick J van der Vliet[1], David San Leon Granado[1], Nienke van der Stoep[2], Henk Buermans[1], Robin van Schendel[1], Joost Schimmel[1], Marianne de Visser[3], Rudy van Coster[4], Marc Jeanpierre[5], Pascal Laforet[6], Meena Upadhyaya[7], Baziel van Engelen[8], Sabrina Sacconi[9], Rabi Tawil[10], Nicol C Voermans[8], Mark Rogers[7], Silvère M van der Maarel [1]*

1 Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

2 Department of Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands

3 Academic Medical Center, Department of Neurology, Amsterdam, The Netherlands

4 Department of Pediatrics, Division of Pediatric Neurology, Ghent University Hospital, Ghent, Belgium

5 APHP-Hôpitaux de Paris, Université de Paris, Paris, France

6 Nord-Est/Ile-de-France Neuromuscular Reference Center, FHU PHENIX, Neurology Department, Raymond-Poincaré Hospital, Versailles Saint-Quentin-en-Yvelines - Paris Saclay University, Garches, France.

7 Department of Medical Genetics, Cardiff University, Cardif, UK

8 Department of Neurology, Donders Institute for Brain, Cognition and Behaviour, Radboud University, The Netherlands

9 Centre de référence des Maladies neuromusculaires, Nice University Hospital, Nice, France

10 Department of Neurology, University of Rochester Medical Center, Rochester, NY, USA

*corresponding authors

Richard JLF Lemmers, Silvère M van der Maarel

Department of Human Genetics

Leiden University Medical Center

Leiden, The Netherlands

Tel 0031-71-5269481

Fax 0031-71-5268285

Email: r.j.l.f.lemmers@lumc.nl or maarel@lumc.nl

Abstract

Facioscapulohumeral muscular dystrophy (FSHD) is an inherited myopathy clinically characterized by weakness in the facial, shoulder girdle and upper arm muscles. FSHD is caused by chromatin relaxation of the D4Z4 macrosatellite repeat, mostly by a repeat contraction, facilitating ectopic expression of *DUX4* in skeletal muscle.

Genetic diagnosis for FSHD is generally based on the sizing and haplotyping of the D4Z4 repeat on chromosome 4 by Southern blotting, molecular combing or single-molecule optical mapping, which is usually straight forward but can be complicated by atypical rearrangements of the D4Z4 repeat. One of these rearrangements is a D4Z4 proximally-extended deletion (DPED) allele, where not only the D4Z4 repeat is partially deleted, but also sequences immediately proximal to the repeat are lost, which can impede accurate diagnosis in all genetic methods.

Previously, we identified several DPED alleles in FSHD and estimated the size of the proximal deletions by a complex pulsed-field gel electrophoresis and Southern blot strategy. Here, using next generation sequencing, we have defined the breakpoint junctions of these DPED alleles at the base pair resolution in 12 FSHD families and 4 control individuals facilitating a PCR-based diagnosis of these DPED alleles.

Our results show that half of the DPED alleles are derivates of an ancient founder allele. For some DPED alleles we found that genetic elements are deleted such as *DUX4c, FRG2, DBE-T* and myogenic enhancers necessitating re-evaluation of their role in FSHD pathogenesis.

Introduction

Facioscapulohumeral muscular dystrophy (FSHD; MIM 158900) represents one of the most common inherited myopathies classically defined by an onset of weakness in the facial, shoulder girdle and upper arm muscles in the second decade. With disease progression, which is typically slow, also other muscles may become affected. Other clinical hall marks of the disease are asymmetric muscle involvement and the marked variability in disease onset and severity, even within families (1, 2).

FSHD is caused by a local and partial chromatin relaxation of the D4Z4 macrosatellite repeat, marked by e.g. D4Z4 hypomethylation, in somatic cells facilitating the expression of *DUX4* in skeletal muscle (3-7). The transcription factor DUX4 is normally repressed in this tissue while active in testis and in cleavage stage embryos, where it is involved in zygotic genome activation (8-10). Out of place DUX4 expression in skeletal muscle tissue induces stem cell and germline specific processes, as well as other processes that disturb normal muscle homeostasis eventually resulting in apoptosis (5, 11, 12).

A complete copy of the *DUX4* gene is embedded in the D4Z4 repeat and adjacent sequence located on a specific genetic variant of chromosome 4q called 4qA (13, 14). The D4Z4 repeat is also present on a chromosome 4 variant that is called 4qB, which is equally common to 4qA in the European population and has a different sequence immediately distal to D4Z4, and in the subtelomere of chromosome 10q (15-17). Yet, only derepression of the D4Z4 locus on chromosome 4qA, but not on 4qB or 10q, typically results in the production of stable DUX4 mRNA and DUX4 protein in a subset of myonuclei. This apparent discordance between different chromosomal backgrounds can be explained by the presence of a somatic polyadenylation signal (PAS) for the *DUX4* gene (18). This *DUX4* PAS is only present in sequences immediately distal to the D4Z4 repeat on chromosome 4qA but not on chromosomes 4qB and 10q.

FSHD is genetically heterogeneous with two major genetic mechanisms having been described that converge into the presence of DUX4 in skeletal muscle. In the majority of patients (FSHD1, accounting for ~95% of patients), partial D4Z4 chromatin relaxation is caused by a contraction of the D4Z4 repeat to a size of 1-10U (19). In unaffected European individuals the repeat size varies between 8-100U (20, 21). A minority of patients (FSHD2; ~5%) show a digenic pattern of inheritance having a normal-sized D4Z4 repeat, albeit in the lower normal size range (mostly 8-20U) (22), on a 4qA

chromosome in combination with a heterozygous pathogenic variant in the structural maintenance of chromosomes hinge domain 1 (*SMCHD1*) gene or rarely in the DNA methyltransferase 3B (*DNMT3B*) gene, or bi-allelic mutations in the ligand dependent nuclear receptor interacting factor 1 (*LRIF1*) gene (23-25). All three FSHD2 genes encode chromatin factors that are necessary to establish and/or maintain a repressive D4Z4 chromatin structure in somatic cells. A recent study of a French patient population suggests that FSHD1 and FSHD2 are not separate disease entities but form a continuum in which the combination of D4Z4 repeat size and repressive activity of D4Z4 chromatin factors determines the likelihood of DUX4 expression in skeletal muscle and disease presentation (26).

DNA testing for FSHD is most often based on digestion of genomic DNA isolated from peripheral blood mononuclear cells by specific restriction endonucleases followed by Southern blotting (SB) using a combination of hybridization probes that recognize unique sequences immediately proximal (D4F104S1; p13E-11) and distal (A and B) to the D4Z4 repeat, as well as the repeat itself (D4Z4) (Figure 1) (27). Using this combination of hybridization probes, atypical rearrangements of the D4Z4 repeat have been reported in the past. One of the more common atypical rearrangement found in FSHD1 individuals is a so-called D4F104S1 deletion (or D4Z4 proximally-extended deletion, DPED) allele (28-30). In these individuals, not only the D4Z4 repeat is partially deleted, but also sequences immediately proximal to the repeat are lost. Confirmation of the clinical diagnosis of FSHD can be easily missed in these individuals because of a failure of hybridization of probe p13E-11, which is typically used as a first-tier strategy in FSHD DNA diagnostics to visualize the contracted D4Z4 repeat. FSHD genetic diagnosis can also be performed by emerging technologies such as molecular combing (MC) or single-molecule optical mapping (SMOM) (31, 32). Both methods visualize the size of the D4Z4 repeats and bordering sequences on chromosomes 4q and 10q. FSHD alleles that have a D4F104S1 deletion have recently also been identified using MC(33) and theoretically SMOM should also be able recognize these alleles.

Previously, we mapped the size of the proximal deletion for some DPED alleles by an alternative pulsed-field gel electrophoresis (PFGE) and SB strategy that allows to better characterize the proximal sequences that are lost in these individuals (29). The estimated deletion size in these DPED alleles suggested that several genetic elements previously suggested to contribute to FSHD pathology are

deleted in some DPED alleles. To test if these findings hold true, and to obtain a better estimate of the frequency and origin of these alleles, we decided to analyze DPED alleles in detail. With the advent of next generation sequencing, we have therefore re-explored previous identified cases carrying a DPED allele and included a series of novel cases and controls, and were able to characterize the breakpoint junctions at the base pair resolution.

Results

SB-based identification of individuals with D4F104S1 deletions

We identified in our database 12 FSHD individuals from 11 families and 4 control individuals that have a DPED allele (Table 1). For all, the SB-based genetic analysis with PFGE revealed only 3 of the expected 4 alleles after applying the standard diagnostic probe p13E-11 on EcoRI/HindIII-digested genomic DNA. The DPED-allele only became manifest after hybridization with SB probes D4Z4 and 4qA (Supplementary Figure 1). In 9 FSHD families, we identified a DPED allele with an FSHD1-sized repeat. In previously reported family Rf100, we detected a normal-sized DPED allele (15U) in the father and unaffected son. This allele was *de novo* contracted to 3U in the oldest affected son (30). In family Rf137 we reported a *de novo* partial deletion of the D4Z4 repeat that extended into proximal sequences in the proband (29). Unexpectedly, we also identified two affected individuals who have a repeat size larger than the FSHD1 threshold of 10 D4Z4 units (Rf1161 [15U] and Rf1317 [16U]). The clinical characteristics of these individuals are described in more detail in the supplementary data.

Previously, we determined the size of the proximal deletion in four samples by a SB-based strategy (28-30). In this approach, the size of the proximal deletion is estimated by comparison of differently digested DNA fragments in the subtelomere of chromosome 4q after PFGE and therefore not very

accurate. For six suspected DPED-allele carriers we performed MC to confirm the presence of such an allele. With MC the proximal deletion is visualized by the absence of a part of the genetic barcode proximal to the D4Z4 repeat. MC software cannot identify the DPED alleles automatically, rather unusual hybridization patterns are collected separately after which the structure of the allele and the size of the D4Z4 repeat size can be established manually using the software (Figure 2A). Generally, we found a good concordance between the deletion size estimates of the DPED in SB and MC. (Supplementary Figure 2).

Gene expression analysis

To confirm for these cases that FSHD is caused by derepression of *DUX4* from the DPED allele, we performed *DUX4* gene expression analysis and expression analysis of established DUX4 targets. We were able to collect myocytes from patient Rf100.201 (3U) and his unaffected father Rf100.101 (15U) and we collected fibroblasts from patients Rf137.3 (7U) and Rf1161.1 (15U). We used lentivirus-mediated MyoD-forced myogenesis to study *DUX4* and target gene expression in the fibroblast samples. After myogenic differentiation, we observed high *DUX4* and target gene expression in patients Rf100.201, Rf137.3 and Rf1161.1, confirming FSHD (Figure 3). We did not find *DUX4* and target gene expression in cells from the unaffected father Rf100.101.

Mapping of the breakpoint junctions

To precisely define the deletions and to understand the mechanism that creates these proximal deletions we sequenced the breakpoint junctions. For GM18517, the size of the proximal deletion was estimated at 3 kb based on MC (Supplementary Figure 1) allowing the design of several PCR primer combinations that were predicted to span the deletion and enabling us to amplify and sequence the breakpoint junction (designated DPED7). For 11 of the in total 15 DPED occurrences, we developed a method by applying Illumina short read sequencing on genomic DNA enriched for the DEPD-allele (Figure 2B). Genomic DNA was enriched for the D4Z4 repeat by it using restriction endonuclease MseI, which leaves the D4Z4 repeat intact. The efficiency of the enrichment is visualized in the agarose gel picture in Figure 2B. After isolation and verification of the enriched DPED DNA

fragments from PFGE gel, tagmentation was followed by a limited-cycle PCR amplification, and the PCR-amplified DNA was size-separated for 400-600 bp fragments, equimolarly pooled and paired-end sequenced.

On average we generated 7 million paired-end short reads per sample which were aligned to a custom build of the hg19 reference sequence of the 4q subtelomere. After alignment, we obtained images alike as shown in Figure 2C, where reads immediately proximal to the D4Z4 repeat are missing, but they re-occur after a gap close to a MseI restriction site, indicating the size of the proximal deletion. Two more detailed examples are shown in Supplementary Figure 3. For 9/11 analyzed individuals we sequenced the breakpoint junction, while for the remaining two samples NGS-based breakpoint junction determination failed. We identified six different breakpoint junctions; DPED1 - DPED6, as 4 individuals carried the same proximal deletion and breakpoint junction (DPED1) meaning that they have exactly the same breakpoint proximal to the D4Z4 repeat and exactly the same breakpoint within the D4Z4 repeat unit although their D4Z4 repeat sizes differ. Based on the position of all 7 breakpoint junctions (DPED1-DPED6 and DPED 7), we designed PCR primers proximal and distal to the deletion and confirmed the breakpoint junctions by PCR amplification and Sanger sequencing (data not shown). Using these PCR amplicons we analyzed the two samples that failed in the NGS analysis and 3 other samples that were not analyzed by the NGS-based method and identified another 4 samples with DPED1 and one other with DPED7. Thus, DPED1, defined by a proximal deletion of 45 kb, was found in 8 families, all FSHD. These DPED1 families included families Rf100 and Rf132 for which we previously estimated a different proximal deletion size based on the SB approach. For the other deletions, the size ranges from 3 to 67 kb with the breakpoint within D4Z4 being different for the seven identified DPED alleles. The size of all identified deletions and the composition of the breakpoint junctions is shown in Table 1 and Figure 4.

Rearrangement mechanism

To understand the rearrangement mechanism creating these proximal deletions we studied the sequences flanking the breakpoint junctions. Three of the 7 deletions displayed a *de novo* insertion between the donor and acceptor region, which, upon close inspection, seemed to (partly) be derived

from the flanking DNA. For DPED2, we found that the 42 kb proximally extended deletion was accompanied by a 9 nucleotides (GGGGCTGGG) insertion, a sequence that is found in the deleted sequences 17 nucleotides proximal to the distal breakpoint. For DPED4 we observed a 7 nucleotides (GTTGCCG) insertion, which was also seen four nucleotides distal to the distal breakpoint. And for DPED5, we observed a single nucleotide insertion (C) (Supplementary Table 2). This type of repair in DPED2, DPED4 and DPED5, called templated insertions, are indicative for repair by polymerase Theta-Mediated End-Joining (TMEJ) (34). TMEJ is defined as a pathway that repairs breaks via microhomology and the appearance of microhomology. The remaining 4/7 breakpoint junctions seem the result of more simple deletions where the donor and acceptor region were joined without any insertions. Interestingly, all these simple deletions display microhomology at the repair junctions. For two breakpoint junctions the microhomology is only one nucleotide (DPED6 [C] and DPED7 [G]), while the other two show four nucleotides of microhomology (DPED1 [CGTG] and DPED3 [CACA]) (Supplementary Table 2), consistent with a TMEJ repair mechanism.

Evidence for founder effects

Strikingly, 8 of 15 families have a DPED1 allele and most of these unrelated families have a different D4Z4 repeat size on their DPED1 allele (Figure 4). Six of these families originate from the Netherlands, while the other 2 are from Morocco and Algeria (Table 1). This indicates that the breakpoint junction is either a mutation hotspot or that this is a founder mutation. To further explore this we analyzed the DNA sequence up to 4 kb proximal to the deletion. As shown in Supplementary Figure 4, this region encompasses 7 single nucleotide polymorphisms (SNPs) with a minor allele frequency >0.3. We defined a haplotype based on these SNPs and found that all samples carried the identical haplotype. This haplotype was shown to be specific for DPED1 alleles as we could not find this SNP combination in 10 independent, randomly selected 4qA alleles and in 10 independent and randomly selected 4qB alleles (Supplementary Figure 4). Likewise, DPED7 was also identified in two unrelated African control individuals from the 1000 Genome project (Yoruba individuals from Ibadan, Nigeria). SNP-array based studies already confirmed that these individuals are not related (personal communication with IGSR helpdesk, https://www.internationalgenome.org/),(35) which was

corroborated by the observation that they carry a DPED-allele with a different repeat size (48U and 13U), suggestive for a second DPED-founder-allele.

Detailed analysis of DPED patients with D4Z4 repeats beyond the FSHD1 threshold.

We identified two patients who have a DPED allele with a repeat size >10 units (Rf1161 [15U] and Rf1317 [16U]). Because these repeat sizes are commonly found in FSHD2, we determined the D4Z4 methylation in these individuals. For patient Rf1161 we observed relatively low D4Z4 methylation (delta1 -17%), which is just above the FSHD2 threshold of -20% (36). Patient Rf1317 has normal D4Z4 methylation (delta1 9%). To exclude other genetic causes for the disease phenotype in these individuals and to exclude a variant in one of the FSHD2 genes we performed whole exome sequencing (WES). For Rf1317, we did not find a pathogenic variant in other muscular dystrophy genes or in *SMCHD1, DNMT3B* and *LRIF1*. Surprisingly, in patient Rf1161, we identified a variant in intron 10 of *SMCHD1* at position c.1342+3A>G, which was shown to cause an out-of-frame 137 nucleotide deletion in exon 10 due to alternative splicing (Supplementary Figure 5). Despite that the D4Z4 methylation did not reach the FSHD2 threshold, the alternative splicing suggest this is a mild pathogenic SMCHD1 variant. Patient Rf1161 has two permissive alleles, the 15U DPED1 allele and a 14U standard 4A161 allele. Based on a single nucleotide polymorphism (T/C) in *DUX4* between both alleles near the DUX4 transcription start site, we were able to perform allele-specific expression analysis on MyoD-transduced fibroblasts and found evidence for biallelic DUX4 expression (Supplementary Figure 6).

Frequency of DPED alleles in FSHD

Previously, we estimated the frequency of DPED allele at 2% in the FSHD population. This was based on the number of cases identified in our research lab compared to the total number of clinically confirmed FSHD families. We anticipated that this number might be overestimated because of a referral bias to our lab for more detailed genetic analysis after a false negative test by a standard diagnostic procedure. Therefore, we re-calculated the frequency of DPED alleles in FSHD based on samples that were submitted by a few institutes which frequently send in samples without having prior

genetic testing. In this selection, we identified four DPED allele FSHD families in 606 independent families (all with an European background), which brings the frequency to 0,6%. We expect this frequency to be representative for the Caucasian control population. Intriguingly, we identified a DPED allele in three (2xDPED7 and 1xDPED6) out of 60 unrelated Yoruba control individuals from Ibadan (Nigeria), which were included in the 1000 Genomes project. This might indicate that the DPED allele frequency is population-dependent and higher in Sub-Saharan Africa.

PCR-based testing for DPED alleles

In our study we found that the majority of DPED-FSHD-alleles carry DPED1. For this founder breakpoint junction, but also for the other 6 breakpoint junctions we designed a PCR detection strategy. These PCR amplicons can also be used to identify DPED carriers among patients with a 'false' negative genetic test, or in family members from DPED carriers (PCR primers in Supplementary Table 1). To test this, we performed the PCR-analysis for DPED2 on family members from patient Rf1067 for which multiple members show FSHD features. As shown, the PCR seamlessly identifies the DPED allele in all family members with FSHD-like features, demonstrating the utility of this test (Supplementary Figure 7).

Discussion

Partial deletions of the D4Z4 repeat to a size of 1-10U on a *DUX4* permissive chromosome 4qA is considered diagnostic for FSHD1. Traditionally, DNA diagnosis of FSHD is SB-based using the probe p13E-11 as first-tier DNA diagnostic test to visualize the contracted repeat in EcoRI and EcoRI/BlnI-digested genomic DNA. Best practice standards have been developed for SB-based DNA diagnosis to specify the haplotype of the alleles and to identify more complex D4Z4 rearrangements such as DPED, hybrid, or duplication alleles (18, 22). This more detailed analysis requires high quality DNA in agarose plugs, PFGE analysis and a series of digestions and hybridizations with multiple probes (p13E-11, D4Z4, A and B) recognizing sequences proximal, distal and within the D4Z4 repeat, which makes it labor-intensive. Therefore, in many diagnostic laboratories only probe p13E-11 and linear gel electrophoresis are primarily being used for the routine DNA testing for FSHD. Not using the

extended analysis including probe D4Z4 and PFGE runs the risk that DPED alleles, such as those characterized in this study, are missed. This therefore likely represents a group of underdiagnosed FSHD individuals (29, 37).

Recently, new technologies have been developed for the genetic diagnosis of FSHD like MC and SMOM. Both methods use high quality DNA in agarose plugs and are suitable to identify the FSHD1 mutation, the A/B haplotype and repeat size of the other D4Z4 repeats on chromosomes 4 and 10 in a diagnostic setting. In addition to the repeat size and haplotype, both methods also visualize the region proximal to the D4Z4 repeat, by either a fluorescence barcode over a region of 60 kb (MC), or by >100 kb of fluorescence labels for specific DNA recognition sites in the FSHD locus (SMOM). (31, 32). This theoretically enables the identification of DPED-alleles in FSHD and an estimate of the proximal deletion size. Indeed, DPED-alleles have already been identified by MC(33), and it will be interesting to test DPED cases by SMOM as now more diagnostic laboratories are implementing this technology in their routine service. However, like for SB, MC and SMOM cannot provide information at the nucleotide resolution.

Here we have applied massive parallel sequencing to map the breakpoint junctions in a series of 8 unrelated FSHD families and 3 unrelated control individuals with DPED alleles. For 9/11 samples studied by massive parallel sequencing on D4Z4 repeat-enriched genomic DNA, we were able to identify the breakpoint junction, suggesting that this sequencing approach was effective. In addition, one DPED allele was analyzed by a direct PCR approach. Two samples that failed in the NGS analysis and three other DPED samples were analyzed by a PCR strategy specific for each breakpoint junction and showed to be identical to one of the previously identified breakpoint junctions. In total, we identified 7 different breakpoint junctions in 15 unrelated samples. The size of the proximal-extended deletions in 4 of the FSHD families was previously estimated by PFGE and SB. The precise sequence-based mapping shows that the SB-based mapping of the deletion size was not accurate, as two deletions that were previously differently sized by SB (55kb; Rf132 and 45 kb; Rf100) turn out to be identical (both DPED1, 45 kb). Indeed, estimation of the size of the proximal-extended deletion by PFGE is complicated by the presence of the telomere repeat within the fragments analyzed which varies in size between 2-20kb (29).

To elucidate the possible rearrangement mechanism, we analyzed the sequence of the individual breakpoint junctions that we identified. Based on templated insertions and on microhomology that we found in the regions flanking the deletions, we suggest that the deletion products are the result of double strand break repair by TMEJ (34). Templated insertions were observed for DPED2 (CGGGCTGGG), DPED4 (GTTGCCG) and DPED5 (C), probably due to template switching during repair. To find evidence of TMEJ derived human variants, Schimmel and coauthors analyzed reported variants in the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/). They identified 5% (n=3699) indels of which 3% showed the typical TMEJ associated templated insertions (34). Although for most of the ClinVar indel variants the deletions are smaller than observed for the DPED alleles, some are larger than 1 kb up to 200 kb (not shown).

Previously, a long noncoding transcript (DBE-T) was identified in the FSHD locus, which transcription starts proximal to the D4Z4 repeat, within D4F104S1 and p13E-11 (38). This DBE-T was shown to be specifically upregulated in FSHD and demonstrated to recruit the Trithorax group protein ASH1L to the D4Z4 repeat to de-repress *DUX4*. Sequence analysis of the DPED-alleles identified in controls and FSHD individuals showed that the 726 nucleotide 5' end of the DBE-T transcript was deleted in all cases. In addition, two myogenic enhancers for *DUX4* have been identified in at 4,8 kb (DME1) and 18,0 kb (DME2) proximal to D4Z4, respectively (39). Consequently, the DME1 and DME2 enhancers are deleted in DPED1, DPED2 and DPED4 and DME1 in DPED3. Because all carriers of a DPED-FSHD1 allele (DPED1-DPED4) present a classical FSHD phenotype, despite the lack of the complete DBE-T transcript or one or two of the myogenic enhancers, this observation challenges the need for these elements in the development of FSHD.

Similarly, in some affected DPED carriers the proximal genes *FRG2* and *DUX4c* (located within the inverted copy of the D4Z4 repeat) are lost. Both genes were previously reported to be transcriptionally dysregulated in FSHD myotubes and to possibly contribute to FSHD pathogenesis. DUX4c was suggested to promote DUX4 toxicity by facilitating nuclear clustering (40) while the function of FRG2 is unknown (41). The transcriptional dysregulation of DUX4c is believed to occur *in cis* as an indirect result of the D4Z4 repeat contraction and is, in light of this study, therefore unlikely to be

essential for FSHD pathogenesis. *FRG2* is, however, a DUX4 target gene (42) and therefore its dysregulation is still possible from *FRG2* copies other than the one deleted from the DPED allele.

We showed that 8 (DPED1) out of 11 unrelated FSHD individuals and 2 (DPED7) of the 3 unrelated African controls from the 1000 Genome project have an identical breakpoint junction, while their D4Z4 repeat sizes are different. Since the presence of an identical breakpoint junction suggests a founder effect, we analyzed SNPs closely proximal to DPED1 and were able to confirm that all these patients carry the same haplotype, which was not commonly found among standard European 4qA and 4qB haplotypes, strongly supporting the hypothesis that this represents a founder allele. The DPED1 allele was also found in two FSHD patients with a North-African genetic background suggesting that this allele originates from Africa and might be a relatively old founder allele. Indeed, the *de novo* contraction observed in family Rf100 suggests that the DPED1 allele is a pre-existing condition in the population that upon contraction to a FSHD-sized D4Z4 repeat results in DUX4 expression and disease presentation.

Strikingly, 2/8 FSHD patients with a DEPD1-allele carry a repeat size that exceeds the FSHD1 range. Clinical re-evaluation confirmed that both have classical FSHD (supplementary information). For patient Rf1161, we found D4Z4 methylation almost reaching the FSHD2 threshold which can be explained by an intronic *SMCHD1* variant that results in out-of-frame alternative splicing of intron 10. Possibly, the methylation reduction in Rf1161 did not reach the FSHD2 threshold as the intronic variant is at the less conserved third position of the splice site consensus and therefore might only moderately affect the splicing. He carries two permissive alleles with a medium-sized D4Z4 arrays and we showed biallelic DUX4 expression, as we have reported before in affected *SMCHD1* mutation carriers with two permissive alleles and further supporting the clinical diagnosis of FSHD.(43) For patient Rf1317, we did not find reduced methylation and did not find a pathogenic variant by WES analysis. Although mildly affected, we currently have no conclusive genetic explanation for this patient and unfortunately DUX4 expression studies are not possible in the absence of additional patient material.

To conclude, we estimate the frequency of DEPD-alleles in FSHD at approximately 0.6%, which is comparable to the number found by others (33). The identification of DPED alleles using the standard

genetic diagnosis can be challenging, but awareness and applying methods specific for long DNA molecules such as MC, SMOM, or PFGE with SB will improve this. For segregation analysis of a characterized familial DEPD allele, we advise to use the specific PCR amplicons that we designed. This PCR-based method is cost effective and quick and does not require high quality DNA. In family Rf1067 with DPED2, we showed that the PCR amplicon can identify DPED allele carriers even for low quality and low concentration DNA samples. Especially, the DPED1 amplicon seems suitable as this breakpoint junction seems to cover nearly 70% of all FSHD DEPD-alleles.

Materials and Methods

Subjects

This study was approved by the Medical Ethical Committees from the participating institutions. All individuals carrying a D4Z4 proximally extended deletion (DPED) allele have been participating in our ongoing effort to genotype in detail individuals suspected of having FSHD over the past 30 years. FSHD individuals were mostly included after testing false negative by the standard SB-based genetic analysis using linear gel electrophoresis of liquid genomic DNA. Three of the control samples are from the Yoruba population in the 1000 Genomes Project. Clinical evaluation of all FSHD cases 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] (44, 45). The clinical characteristics of three FSHD individuals who carry a DPED allele with D4Z4 repeat >10U and for four affected members of family Rf1064 (9U DPED-FSHD allele) is provided in more detail in the supplementary information.

Gene expression analysis

For gene expression analysis we generated myocytes or fibroblast cultures from the indicated individuals according to previously described methods (detailed protocols at the Fields Center website (www.urmc.rochester.edu/fields-center/). Primary myocytes and fibroblast cell cultures from genetically confirmed FSHD1 individuals and non-affected individuals served as controls and originated from the University of Rochester Medical Center bio repository. Fibroblast were transduced into myocytes using MyoD as described previously (46). Differentiation into myotubes was induced by serum reduction at 80-100% confluency. Expression analysis for *DUX4*, *GUSB*, DUX4 target genes (*MBD3L2*, *ZSCAN4* and *TRIM43*) and myogenic differentiation genes was performed in triplicate by previously described PCR conditions and primer pairs (47). Detailed genotype and D4Z4 methylation information for all cell cultures can be found in Supplementary Table 3. The primers and size of the amplicons are shown in Supplementary Table 1.

Biallelic *DUX4* expression analysis for Rf1161 was performed after a full length DUX4 RT-PCR with forward primer 5'-TGG CTG GCT GTC CGG GCA GGC-3' and reverse primer 5'-GAT CCA CAG GGA GGG GGC ATT TTA-3'. cDNA from patient Rf100.201 served as a control for the DPED1 allele DUX4 sequence. In a 20 µl PCR reaction we used 5 microliter cDNA in a solution containing 0.2 µM of 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 and 2.5 U of LA-Taq DNA polymerase (Takara) supplemented with 2xGC buffer. The PCR conditions consisted of an initial denaturation step at 94 °C for 5 min., followed by 39 cycles of denaturation at 94 °C for 30 s, annealing at 60 °C for 30 s, and extension at 72 °C for 2 min. The final extension time was 10 min. at 72 °C. For Sanger sequencing of the PCR fragment, primer 5'-GAG GGT GCT GTC CGA GGG TG-3' was used.

Genetic analysis of the D4Z4 repeats and methylation analysis

These studies were done as described previously (47, 48). Briefly, genomic DNA embedded in agarose plugs were digested with EcoRI/HindIII, EcoRI/BlnI and XapI for D4Z4 repeat sizing and HindIII for determining the haplotype and separated by pulsed field gel electrophoresis. After separation, DNA was transferred to charged nylon membranes (Hybond-XL, Amersham) and serially

probed with radioactively labeled probes p13E-11, D4Z4, 4qA and 4qB. Hybridizing fragments were visualized by phosphor imaging on a Typhoon scanner (Amersham). Key individuals having a DPED allele were also analyzed by molecular combing as described previously (22). Briefly, DNA was combed on a glass slide, which was hybridized with antibody-labeled FSHD-specific probes and scanned by the Fibervision HeliXScan. D4Z4 alleles were selected and counted using the general procedure by Fiberstudio 0.9.12 software. SB-based methylation analysis was done using the methylation-sensitive restriction endonuclease *Fse*I as described and delta1 methylation values were calculated as before.

Fragment isolation, library preparation and sequencing

Short read sequencing of the breakpoint junctions in the DPED alleles first required an enrichment for the locus of interest. For this, genomic DNA was digested with the restriction endonuclease MseI (recognition site TTAA), which is absent in the D4Z4 repeat but is very common in the rest of the human genome and generates fragments that are generally <500 base pairs. The D4Z4 repeats can be visualized after PFGE and SB of MseI-digested genomic DNA and hybridization with probe D4Z4. Based on the position of the D4Z4 hybridization signals on the MseI digested SB, we determined the position of the DPED-alleles in the agarose gel for each sample (Figure 2). A new agarose gel with MseI-digested genomic DNA was re-run under the same conditions, after which the agarose fragment at the expected position of the DPED-allele was cut from gel and DNA was extracted. The DNA was simultaneously fragmented and tagged by Nextera XT, followed by a limited-cycle PCR amplification. The PCR-amplified DNA of the different DPED carriers was run on an agarose gel and fragments ranging from 400 to 600 bp (insert size 250-450 bp) were extracted from the gel. Finally, tagged fragments from 11 samples were equimolar pooled and paired-end sequenced (2x125bp, ~7 million reads per sample) on an Illumina HiSeq 2500.

DNA-seq analysis

Quality assessment of the raw reads was done using FastQC v 0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Quality filtering was done with Trim

Galore v 0.6.5 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore) using default parameters for paired-end Illumina reads. The remaining reads were mapped to the human genome (build hg19) using BWA-MEM with the following parameters "—no-mixed —very-sensitive –X 15000" (49). The duplicates were removed using Picard Tools v2.25 (http://broadinstitute.github.io/picard/). The reads were filtered based on their insert size to identify only the pairs with an insert size greater than 500 bps in the genomic regions close to D4Z4. This filtering was done using SAMtools (50). The remaining reads were used for the identification of the breakpoint junctions, this identification was done using Pindel v 0.2.5b9 and with a manually search in the D4Z4 locus on chromosome 4 using Integrative Genomics Viewer (IGV) v 2.9.2 (51, 52). The genome browser tracks were generated with bamCoverage, a tool included in deepTools package version 3.5.1 (53).

PCR confirmation of breakpoint junctions

For each of the different DPED alleles a PCR amplicon was designed that spans the deletion. PCR amplifications were performed on 125 ng of genomic DNA in a solution containing 3.5 µM of each primer, 0.2 mM dATP, 0.2 mM dCTP, 0.2 mM dTTP, 0.2 mM dGTP and 0.2 mM 7-deaza-dGTP, 2.5 U of Accuprime HiFi DNA polymerase and Accuprime HiFi buffer II (both Thermo Fisher Scientific), in a total volume of 20 µl. The PCR conditions consisted of an initial denaturation step at 94 °C for 2 min, followed by 34 cycles of denaturation at 94 °C for 30 s, annealing at 60 °C for 30 s, and extension at 68 °C for 1 min. The final extension time was 10 min at 68 °C. The primers and size of the amplicons are shown in Supplementary Table 1.

SNP analysis DPED1 alleles and common 4q haplotypes

For the analysis of SNPs (rs1882893; rs1882894; rs6820491; rs11944561; rs4863031; rs4863307 and rs10005853) proximal to DPED1 in carriers of this DPED allele, we used a forward primer that was located 5.5 kb proximal to the deletion 5'-GCT TTA TTC AGC TGG GAT CAT CCG CAG ACT CAT G-3' and a reverse primer in D4Z4 5'-GAG TCT CTC ACC GGG CCT AGA CCT AG-3'. For this long range PCR we used 150 ng genomic DNA with 2 uL dNTPs (2 mM), 0.2 U of PrimeSTAR

GXL DNA Polymerase and GXL buffer (both Takara) in a total volume of 20 uL. The PCR conditions consisted of an initial denaturation step at 98°C for 3 min, followed by a touchdown PCR with initial 10 cycles of 98°C for 30 s, 74°C for 30 s (−1°C every cycle) and 68°C for 8 min, followed by 30 cycles of 94°C denaturation for 30 s, 64°C annealing for 30 s, and 68°C extension for 8 min. Followed by a final extension step at 68°C extension for 10 min. For the same variants on standard 4qA and 4qB alleles we used two shorter overlapping PCR amplicons. The forward primer 5'-GGG ATC ATC CGC AGA CTC ATG-3' for the most proximal amplicon (1009 nt) overlaps with the long range primer and was used in combination with reverse primer 5'-CGT GCG GAA AAG TGG GAG TA-3' and for the distal amplicon (578 nt) we used primers 5'-TAC TCC CAC TTT TCC GCA CG-3' and 5'-ATT TTG GAT TCC TCG CCG CC-3' (Supplementary Table 1). These PCR reactions were performed on 150 ng of genomic DNA with 2 uL dNTPs (2 mM), 0.2 U of Phusion DNA polymerase and GC buffer in a total volume of 20 µl. The PCR conditions consisted of an initial denaturation step at 98°C for 3 min., followed by 35 cycles of denaturation at 98°C for 30 s, annealing at 60°C for 30 s, and extension at 72°C for 1 min. Followed by a final extension step at 72°C extension for 5 min.

Whole exome sequencing

Whole exome sequencing was performed by GenomeScan BV (Leiden, The Netherlands). Briefly, 200 ng genomic DNA was fragmented into 200 to 500 bp fragments. Exomes were captured using the Agilent SureSelectXT human all exon v7 capture library (5191-4006) accompanied by Illumina paired end Sequencing on the NovaSeq6000 (Illumina, San Diego, USA) according to manufacturer's protocols. NovaSeq control software NCS v1.7 was used and image analysis, base calling, and quality check was performed with the Illumina data analysis pipeline RTA3.4.4 and Bcl2fastq v2.20. The exome sequencing protocol was validated for clinical use according to ISO 15189. Variant calling and filtering was essentially done as described previously (54). LOVDplus (Leiden Genome Technology Center, Leiden, The Netherlands) was used for interpretation of variants. First a Gene panel for Muscle Disorders (https://www.lumc.nl/sub/4080/att/1768852) was analyzed as well as the FSHD2 genes *SMCHD1*, *DNMT3B* and *LRIF1*. If these genes did not reveal any likely pathogenic variant, a full exome analysis was performed.

Acknowledgements

We thank all FSHD families for participating in our studies. We thank the IGSR helpdesk and prof. dr. P. de Knijff from the LUMC the Netherlands for advice on possible interfamilial relationships in the 1000 Genomes samples. We thank prof. dr. M. Tijstermans from the LUMC the Netherlands for advice on possible rearrangement mechanisms. Several authors are members of the European Reference Network for Rare Neuromuscular Diseases [ERN EURO-NMD] and/or members of the Netherlands Neuromuscular Center (NL-NMD). This work was supported by funds from the National Institute of Neurological Disorders and Stroke grant number P01NS069539, the Prinses Beatrix Spierfonds grant numbers W.OP14-01 and W.OB17-01, and Spieren voor Spieren.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

References

1 Mul, K., Lassche, S., Voermans, N.C., Padberg, G.W., Horlings, C.G. and van Engelen, B.G. (2016) What's in a name? The clinical features of facioscapulohumeral muscular dystrophy. *Pract Neurol*, **16**, 201-207.

2 Padberg, G.W. (1982). Facioscapulohumeral disease. Leiden University. *PhD thesis*

3 Snider, L., Geng, L.N., Lemmers, R.J., Kyba, M., Ware, C.B., Nelson, A.M., Tawil, R., Filippova, G.N., van der Maarel, S.M., Tapscott, S.J. *et al.* (2010) Facioscapulohumeral dystrophy: incomplete suppression of a retrotransposed gene. *PLoS Genet*, **6**, e1001181.

4 Wang, L.H., Friedman, S.D., Shaw, D., Snider, L., Wong, C.J., Budech, C.B., Poliachik, S.L., Gove, N.E., Lewis, L.M., Campbell, A.E. *et al.* (2019) MRI-informed muscle biopsies correlate MRI with pathology and DUX4 target gene expression in FSHD. *Hum Mol Genet*, **28**, 476-486.

5 Rickard, A.M., Petek, L.M. and Miller, D.G. (2015) Endogenous DUX4 expression in FSHD myotubes is sufficient to cause cell death and disrupts RNA splicing and cell migration pathways. *Hum Mol Genet*, **24**, 5901-5914.

6 Balog, J., Thijssen, P.E., Shadle, S., Straasheijm, K.R., van der Vliet, P.J., Krom, Y.D., van den Boogaard, M.L., de Jong, A., RJ, F.L., Tawil, R. *et al.* (2015) Increased DUX4 expression during muscle differentiation correlates with decreased SMCHD1 protein levels at D4Z4. *Epigenetics*, **10**, 1133-1142.

7 Yao, Z., Snider, L., Balog, J., Lemmers, R.J., Van Der Maarel, S.M., Tawil, R. and Tapscott, S.J. (2014) DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. *Hum Mol Genet*, **23**, 5342-5352.

8 Hendrickson, P.G., Dorais, J.A., Grow, E.J., Whiddon, J.L., Lim, J.W., Wike, C.L., Weaver, B.D., Pflueger, C., Emery, B.R., Wilcox, A.L. *et al.* (2017) Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat Genet*, **49**, 925-934.

9 Whiddon, J.L., Langford, A.T., Wong, C.J., Zhong, J.W. and Tapscott, S.J. (2017) Conservation and innovation in the DUX4-family gene network. *Nat Genet*, **49**, 935-940.

10 De Iaco, A., Planet, E., Coluccio, A., Verp, S., Duc, J. and Trono, D. (2017) DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat Genet*, **49**, 941-945.

11 Geng, L.N., Yao, Z., Snider, L., Fong, A.P., Cech, J.N., Young, J.M., van der Maarel, S.M., Ruzzo, W.L., Gentleman, R.C., Tawil, R. *et al.* (2012) DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Dev Cell*, **22**, 38-51.

12 Kowaljow, V., Marcowycz, A., Ansseau, E., Conde, C.B., Sauvage, S., Matteotti, C., Arias, C., Corona, E.D., Nunez, N.G., Leo, O. *et al.* (2007) The DUX4 gene at the FSHD1A locus encodes a pro-apoptotic protein. *Neuromuscul. Disord*, **17**, 611-623.

13 Dixit, M., Ansseau, E., Tassin, A., Winokur, S., Shi, R., Qian, H., Sauvage, S., Matteotti, C., van Acker, A.M., Leo, O. *et al.* (2007) DUX4, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of PITX1. *Proc Natl Acad Sci U S A*, **104**, 18157-18162.

14 Gabriels, J., Beckers, M.C., Ding, H., De Vriese, A., Plaisance, S., van der Maarel, S.M., Padberg, G.W., Frants, R.R., Hewitt, J.E., Collen, D. *et al.* (1999) Nucleotide sequence of the partially deleted D4Z4 locus in a patient with FSHD identifies a putative gene within each 3.3 kb element. *Gene*, **236**, 25-32.

15 Bakker, E., Van der Wielen, M.J., Voorhoeve, E., Ippel, P.F., Padberg, G.W., Frants, R.R. and Wijmenga, C. (1996) Diagnostic, predictive, and prenatal testing for facioscapulohumeral muscular dystrophy: diagnostic approach for sporadic and familial cases. *J Med Genet*, **33**, 29-35.

16 Deidda, G., Cacurri, S., Piazzo, N. and Felicetti, L. (1996) Direct detection of 4q35 rearrangements implicated in facioscapulohumeral muscular dystrophy (FSHD). *J Med Genet*, **33**, 361-365.

17 Lemmers, R.J.L.F., de Kievit, P., Sandkuijl, L., Padberg, G.W., van Ommen, G.J.B., Frants, R.R. and van der Maarel, S.M. (2002) Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. *Nature Genetics*, **32**, 235-236.

18 Lemmers, R.J., van der Vliet, P.J., Klooster, R., Sacconi, S., Camano, P., Dauwerse, J.G., Snider, L., Straasheijm, K.R., van Ommen, G.J., Padberg, G.W. *et al.* (2010) A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science*, **329**, 1650-1653.

19 Wijmenga, C., Hewitt, J.E., Sandkuijl, L.A., Clark, L.N., Wright, T.J., Dauwerse, H.G., Gruter, A.M., Hofker, M.H., Moerer, P., Williamson, R. *et al.* (1992) Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nat Genet*, **2**, 26-30.

20 Lemmers, R.J.L.F., Wohlgemuth, M., van der Gaag, K.J., van der Vliet, P.J., van Teijlingen, C.M.M., de Knijff, P., Padberg, G.W., Frants, R.R. and van der Maarel, S.M. (2007) Specific sequence variations within the 4q35 region are associated with Facioscapulohumeral muscular dystrophy. *American Journal of Human Genetics*, **81**, 884-894.

21 Scionti, I., Fabbri, G., Fiorillo, C., Ricci, G., Greco, F., D'Amico, R., Termanini, A., Vercelli, L., Tomelleri, G., Cao, M. *et al.* (2012) Facioscapulohumeral muscular dystrophy: new insights from compound heterozygotes and implication for prenatal genetic counselling. *J Med Genet*, **49**, 171-178.

22 Lemmers, R., van der Vliet, P.J., Vreijling, J.P., Henderson, D., van der Stoep, N., Voermans, N., van Engelen, B., Baas, F., Sacconi, S., Tawil, R. *et al.* (2018) Cis D4Z4 repeat duplications associated with facioscapulohumeral muscular dystrophy type 2. *Hum Mol Genet*, **27**, 3488-3497.

23 Hamanaka, K., Sikrova, D., Mitsuhashi, S., Masuda, H., Sekiguchi, Y., Sugiyama, A., Shibuya, K., Lemmers, R., Goossens, R., Ogawa, M. *et al.* (2020) Homozygous nonsense variant in LRIF1 associated with facioscapulohumeral muscular dystrophy. *Neurology*, **94**, e2441-e2447.

24 Lemmers, R.J., Tawil, R., Petek, L.M., Balog, J., Block, G.J., Santen, G.W., Amell, A.M., van der Vliet, P.J., Almomani, R., Straasheijm, K.R. *et al.* (2012) Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat Genet*, **44**, 1370-1374.

25 van den Boogaard, M.L., Lemmers, R., Balog, J., Wohlgemuth, M., Auranen, M., Mitsuhashi, S., van der Vliet, P.J., Straasheijm, K.R., van den Akker, R.F.P., Kriek, M. *et al.* (2016) Mutations in DNMT3B Modify Epigenetic Repression of the D4Z4 Repeat and the Penetrance of Facioscapulohumeral Dystrophy. *Am J Hum Genet*, **98**, 1020-1029.

26 Sacconi, S., Briand-Suleau, A., Gros, M., Baudoin, C., Lemmers, R., Rondeau, S., Lagha, N., Nigumann, P., Cambieri, C., Puma, A. *et al.* (2019) FSHD1 and FSHD2 form a disease continuum. *Neurology*, **92**, e2273-e2285.

27 Lemmers, R.J. (2017) Analyzing Copy Number Variation Using Pulsed-Field Gel Electrophoresis: Providing a Genetic Diagnosis for FSHD1. *Methods Mol. Biol*, **1492**, 107-125.

28 Deak, K.L., Lemmers, R.J., Stajich, J.M., Klooster, R., Tawil, R., Frants, R.R., Speer, M.C., van der Maarel, S.M. and Gilbert, J.R. (2007) Genotype-phenotype study in an FSHD family with a proximal deletion encompassing p13E-11 and D4Z4. *Neurology*, **68**, 578-582.

29 Lemmers, R.J., Osborn, M., Haaf, T., Rogers, M., Frants, R.R., Padberg, G.W., Cooper, D.N., van der Maarel, S.M. and Upadhyaya, M. (2003) D4F104S1 deletion in facioscapulohumeral muscular dystrophy: phenotype, size, and detection. *Neurology*, **61**, 178-183.

30 Lemmers, R.J.L.F., van der Maarel, S.M., van Deutekom, J.C.T., van der Wielen, M.J.R., Deidda, G., Dauwerse, H.G., Hewitt, J., Hofker, M., Bakker, E., Padberg, G.W. *et al.* (1998) Inter- and intrachromosomal subtelomeric rearrangements on 4q35: implications for facioscapulohumeral muscular dystrophy (FSHD) aetiology and diagnosis. *Hum Mol Genet*, **7**, 1207-1214.

31 Nguyen, K., Walrafen, P., Bernard, R., Attarian, S., Chaix, C., Vovan, C., Renard, E., Dufrane, N., Pouget, J., Vannier, A. *et al.* (2011) Molecular combing reveals allelic combinations in facioscapulohumeral dystrophy. *Ann Neurol*, **70**, 627-633.

32 Zhang, Q., Xu, X., Ding, L., Li, H., Xu, C., Gong, Y., Liu, Y., Mu, T., Leigh, D., Cram, D.S. *et al.* (2019) Clinical application of single-molecule optical mapping to a multigeneration FSHD1 pedigree. *Mol Genet Genomic Med*, **7**, e565.

33 Nguyen, K., Broucqsault, N., Chaix, C., Roche, S., Robin, J.D., Vovan, C., Gerard, L., Megarbane, A., Urtizberea, J.A., Bellance, R. *et al.* (2019) Deciphering the complexity of the 4q and 10q subtelomeres by molecular combing in healthy individuals and patients with facioscapulohumeral dystrophy. *J Med Genet*, **56**, 590-601.

34 Schimmel, J., van Schendel, R., den Dunnen, J.T. and Tijsterman, M. (2019) Templated Insertions: A Smoking Gun for Polymerase Theta-Mediated End Joining. *Trends Genet*, **35**, 632-644.

35 Roslin NM, L.W., Paterson AD, Strug LJ. (2016) Quality control analysis of the 1000 Genomes Project Omni2.5 genotypes. *bioRxiv*.

36 Lemmers, R., van der Stoep, N., Vliet, P.J.V., Moore, S.A., San Leon Granado, D., Johnson, K., Topf, A., Straub, V., Evangelista, T., Mozaffar, T. *et al.* (2019) SMCHD1 mutation spectrum for

facioscapulohumeral muscular dystrophy type 2 (FSHD2) and Bosma arhinia microphthalmia syndrome (BAMS) reveals disease-specific localisation of variants in the ATPase domain. *J Med Genet*, **56**, 693-700.

37 Ehrlich, M., Jackson, K., Tsumagari, K., Camano, P. and Lemmers, R.J. (2007) Hybridization analysis of D4Z4 repeat arrays linked to FSHD. *Chromosoma*, **116**, 107-116.

38 Cabianca, D.S., Casa, V., Bodega, B., Xynos, A., Ginelli, E., Tanaka, Y. and Gabellini, D. (2012) A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. *Cell*, **149**, 819-831.

39 Himeda, C.L., Debarnot, C., Homma, S., Beermann, M.L., Miller, J.B., Jones, P.L. and Jones, T.I. (2014) Myogenic enhancers regulate expression of the facioscapulohumeral muscular dystrophy-associated DUX4 gene. *Mol Cell Biol*, **34**, 1942-1955.

40 Vanderplanck, C., Tassin, A., Ansseau, E., Charron, S., Wauters, A., Lancelot, C., Vancutsem, K., Laoudj-Chenivesse, D., Belayew, A. and Coppee, F. (2018) Overexpression of the double homeodomain protein DUX4c interferes with myofibrillogenesis and induces clustering of myonuclei. *Skelet Muscle*, **8**, 2.

41 Rijkers, T., Deidda, G., van Koningsbruggen, S., van Geel, M., Lemmers, R.J.L.F., van Deutekom, J.C.T., Figlewicz, D., Hewitt, J.E., Padberg, G.W., Frants, R.R. *et al.* (2004) FRG2, an FSHD candidate gene, is transcriptionally upregulated in differentiating primary myoblast cultures of FSHD patients. *Journal of Medical Genetics*, **41**, 826-836.

42 Thijssen, P.E., Balog, J., Yao, Z., Pham, T.P., Tawil, R., Tapscott, S.J. and Van der Maarel, S.M. (2014) DUX4 promotes transcription of FRG2 by directly activating its promoter in facioscapulohumeral muscular dystrophy. *Skelet Muscle*, **4**, 19.

43 Lemmers, R.J.L.F., van der Vliet, P.J., Balog, J., Goeman, J.J., Arindrarto, W., Krom, Y.D., Straasheijm, K.R., Debipersad, R.D., Ozel, G., Sowden, J. *et al.* (2018) Deep characterization of a common D4Z4 variant identifies biallelic DUX4 expression as a modifier for disease penetrance in FSHD2. *European Journal of Human Genetics*, **26**, 94-106.

44 Ricci, E., Galluzzi, G., Deidda, G., Cacurri, S., Colantoni, L., Merico, B., Piazzo, N., Servidei, S., Vigneti, E., Pasceri, V. *et al.* (1999) Progress in the molecular diagnosis of facioscapulohumeral

muscular dystrophy and correlation between the number of KpnI repeats at the 4q35 locus and clinical phenotype. *Annals of Neurology*, **45**, 751-757.

45 van Overveld, P.G., Enthoven, L., Ricci, E., Rossi, M., Felicetti, L., Jeanpierre, M., Winokur, S.T., Frants, R.R., Padberg, G.W. and van der Maarel, S.M. (2005) Variable hypomethylation of D4Z4 in facioscapulohumeral muscular dystrophy. *Ann Neurol*, **58**, 569-576.

46 Yao, Z., Fong, A.P., Cao, Y., Ruzzo, W.L., Gentleman, R.C. and Tapscott, S.J. (2013) Comparison of endogenous and overexpressed MyoD shows enhanced binding of physiologically bound sites. *Skelet Muscle*, **3**, 8.

47 Lemmers, R., van der Vliet, P.J., Blatnik, A., Balog, J., Zidar, J., Henderson, D., Goselink, R., Tapscott, S.J., Voermans, N.C., Tawil, R. *et al.* (2021) Chromosome 10q-linked FSHD identifies DUX4 as principal disease gene. *J Med Genet*, in press.

48 Lemmers, R.J.L.F., Goeman, J.J., van der Vliet, P.J., van Nieuwenhuizen, M.P., Balog, J., Vos-Versteeg, M., Camano, P., Arroyo, M.A.R., Jerico, I., Rogers, M.T. *et al.* (2015) Inter-individual differences in CpG methylation at D4Z4 correlate with clinical variability in FSHD1 and FSHD2. *Human Molecular Genetics*, **24**, 659-669.

49 Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM *ArXiv*.

50 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

51 Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat Biotechnol*, **29**, 24-26.

52 Ye, K., Schulz, M.H., Long, Q., Apweiler, R. and Ning, Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865-2871.

53 Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F. and Manke, T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*, **44**, W160-165.

54 de Koning, M.A., Haak, M.C., Adama van Scheltema, P.N., Peeters-Scholte, C., Koopmann, T.T., Nibbeling, E.A.R., Aten, E., den Hollander, N.S., Ruivenkamp, C.A.L., Hoffer, M.J.V. *et al.* (2019) From diagnostic yield to clinical impact: a pilot study on the implementation of prenatal exome sequencing in routine care. *Genet Med*, **21**, 2303-2310.

Legends to Figures

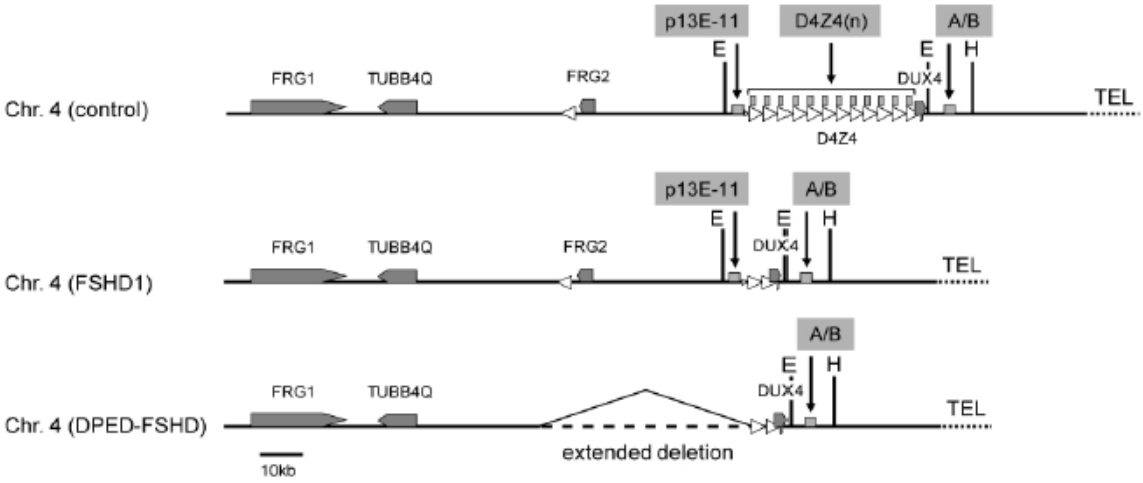

Figure 1. Graphical representation of the FSHD locus on chromosome 4, showing the D4Z4 repeat, the *DUX4* gene, proximal genes (*FRG1, TUBB4Q* and *FRG2*) and the hybridization probes (p13E-11, A/B and D4Z4) that are used for SB-based genetic diagnosis. The top figure shows the non-affected situation and the middle the FSHD1 situation with a contracted D4Z4 repeat. The bottom figure shows an FSHD1 allele, in which the region proximal to D4Z4 is partially deleted including hybridization probe p13E-11.

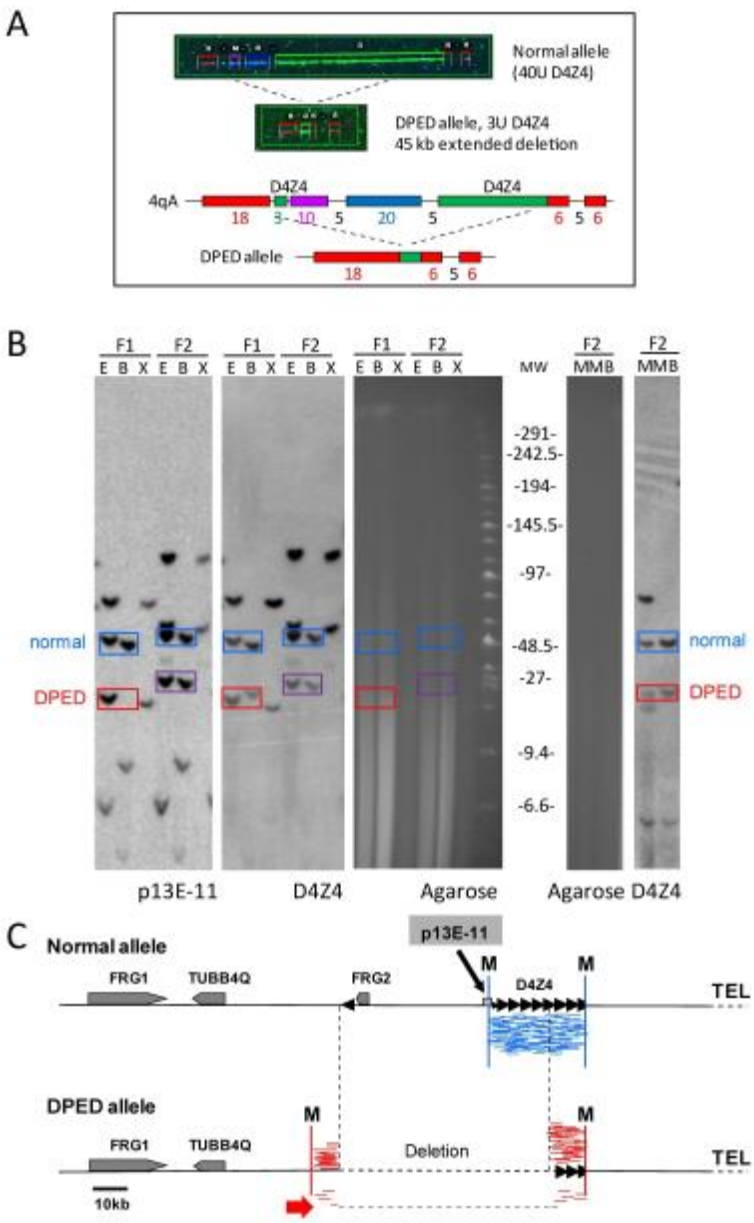

Figure 2. A) Top: Example of MC on wild type 4qA and a DPED-FSHD allele. Bottom: Schematic overview of the position and color of the fluorescent MC probes for both alleles in the boxed figure. The D4Z4 repeat and the highly homologous inverted unit on chromosome 4 are indicated in green. The DPED allele shows a deletion of approximately 45 kb and has 3 D4Z4 units. B) Schematic overview of the DNA enrichment and short read sequencing procedure for DPED alleles. Left from molecular weight (MW) marker: SB and agarose gel after PFGE with EcoRI/HindIII (E), EcoRI/BlnI (B) and XapI (X)-digested genomic DNA from FSHD individuals with a DPED (F1) and a standard (F2) FSHD allele. The p13E-11 hybridized SB for individual F1 shows a normal-sized chromosome 4 band (blue rectangular box) and an FSHD allele (purple). For individual F1, only one chromosome 4

allele (blue) is visible, the FSHD-sized DPED allele is not visible (marked with a dashed red rectangular box) as the p13E-11 recognition site is deleted. Subsequent hybridization of the same blot with probe D4Z4 reveals the DPED allele (marked with a red rectangular box) for F2. Right from MW marker: agarose gel and SB upon digestion with MseI (M) and MseI/BlnI (MB) with indicated DPED allele (red) and the other chromosome 4 allele (blue). MseI removes most of the genomic background as shown in the agarose gel. Guided by the SB results, the DEPD allele was isolated from gel at the expected position and prepared for short read sequencing. C) The isolated DPED allele DNA was subjected to short read sequencing and the sequence reads were aligned to the reference D4Z4 sequence. The blue lines in the D4Z4 locus mark the reads for a normal chromosome 4 allele (top) and red lines mark the reads in the D4Z4 locus mark the reads for DPED allele (bottom). The gap uncovers the deletion and reads that straddle the deletion indicate the exact proximal breakpoint (marked with bold red arrow). The position of the restriction endonuclease MseI (M) site flanking the deletion are specified as well as the position of the p13E-11 recognition site. Indicated are the D4Z4 repeat (D4Z4), the inverted D4Z4 unit, genes proximal to D4Z4 (*FRG1, TUBB4Q* and *FRG2*), the telomere (TEL) and a scale bar.

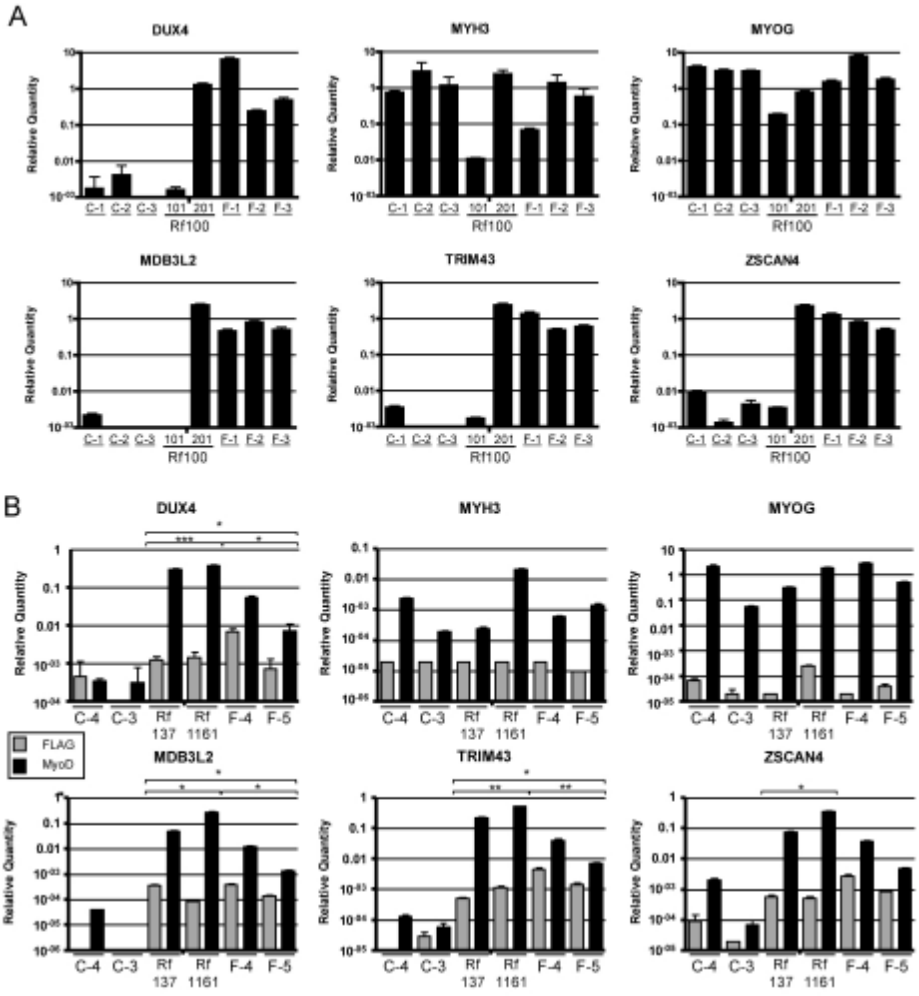

Figure 3. A) Quantitative RT-PCR on myotubes from patient Rf100.201 (3U DPED1 allele) and his unaffected father Rf100.101 (15U DPED1 allele), genetically standard FSHD patients (F-1, F-2 and F-3) and control individuals (C-1, C-2 and C-3). Relative expression levels to *GUS* of *DUX4* and the early and late myogenic differentiation markers (*MYOG* and *MYH3*, respectively) are shown in the top panel and that of the DUX4 target genes *ZSCAN4*, *TRIM43* and *MBD3L2* in the bottom panel. Analyses were done in triplicate and the graph represents means ± SEM. Rf100.201 shows an FSHD-like expression profile and Rf100.101 an expression profile comparable to the negative controls. B) Quantitative RT-PCR on MyoD-transduced fibroblast of Rf137.3 (3U DPED3 allele), Rf1161.1 (13U DPED1 allele) and genetically standard FSHD patients (F-4 and F-5) and control individuals (C-3 and C-4). Fibroblasts were transduced with MyoD (black bars) to induce transdifferentiation towards myogenic lineage or with 3xFlag as a negative control (grey bars). Relative expression levels are shown identical to myotubes analysis in panel A. Rf137.3 and Rf1161.1 show an expression profile

comparable to the standard FSHD patients. Analyses were done in triplicate and the graph represents mean ± SEM. Comparisons between controls vs. DPED-FSHD patients, controls vs. standard FSHD patients and controls vs. patients in both groups were performed using unpaired two-tailed t-tests. Significant differences are indicated by an asterisk on top of the bracket assembling the specific FSHD groups (where $p<0.05$ is indicated by *, $p<0.005$ by ** and $p<0.0005$ by ***).

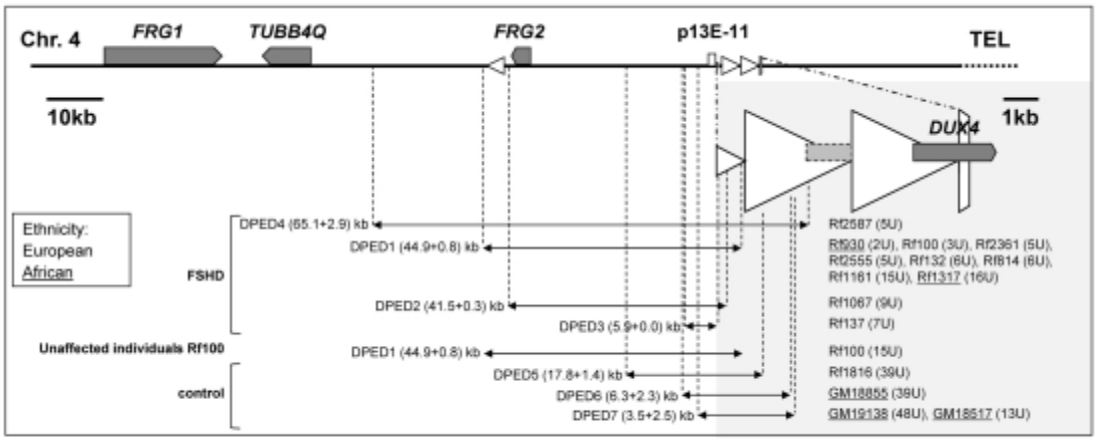

Figure 4. Detailed overview of the deleted region (in kb) proximal and within D4Z4 for all identified breakpoint junctions in 11 FSHD families (including two unaffected family members in family Rf100 with normal sized repeats on a DPED1 allele) and in 4 control individuals. The breakpoint junctions are sorted by the size of the deleted region proximal to D4Z4 and if they are identified in FSHD or control individuals. The family ID or 1000 Genomes ID and the D4Z4 repeat size (in units) is indicated on the right and individuals with an African genetic background are underlined. Indicated are the D4Z4 repeat (D4Z4, triangles), the complete *DUX4* gene at the distal end of the repeat and the partial *DUX4* copy (dashed light gray in the first D4Z4 unit). We also show the inverted D4Z4 unit, the telomere (TEL) as well as the genes proximal to D4Z4 (*FRG1*, *TUBB4Q* and *FRG2*). To visualize the different breakpoints within D4Z4, this part of the sequence is magnified 5 times in the light grey area. The size marker for both magnifications is indicated.

Table 1. Genotype and deletion information for all DPED allele carrier index cases and family members. Columns 1-4 show the information about the breakpoint junctions (DPED allele name and the size of the deletion proximal and in D4Z4 and the total deletion size). Columns 5-9 show the family ID, the personal number of the individual (Nr) and parents (father, mother) and the sex (M/F). Columns 10-14 show the clinical information with age at onset (AaO), clinical severity score (CSS)(44), age at examination (AAE), the age corrected severity score (ACSS)(45) and status. The D4Z4 repeat size in units and haplotype on both chromosome 4 alleles is given in columns 15-16 (the DPED allele in grey) and the DNA methylation (Delta1)(48) in column 17. Columns 18 and 19 show the country and continent of origin.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		16		17	18	19
BP	Deletion proximal to D4Z4 (kb)	Deletion in D4Z4 (kb)	Total size deletion (kb)	Family ID	Nr	Father	Mother	M/F	AaO	CSS	AAE	ACSS	Status	Chr. 4q-1		Chr. 4q-2		Delta1	Country	Continent
	44.9	0.8	45.7	Rf930	1	0	0	F	4	7	18	389	FSHD	2U	4AdelS	19U	4A161S	-5	Morocco	Africa
	44.9	0.8	45.7		101	0	0	M	NA	0	41	0	control	15U	4AdelS	44U	4B168	9	The Netherlands	Europe
	44.9	0.8	45.7		201	101	102	M	NA	10	37	270	FSHD	3U	4AdelS	18U	4B163	12	The Netherlands	Europe
	44.9	0.8	45.7	Rf100	202	101	102	M	NA	0	9	0	control	15U	4AdelS	18U	4B163	NA	The Netherlands	Europe
	no deletion				203	101	102	M	NA	0	7	0	control	18U	4B163	44U	4B168	NA	The Netherlands	Europe
	no deletion				102	0	0	F	NA	0	38	0	control	18U	4B163	18U	4B163	-2	The Netherlands	Europe
BP1	44.9	0.8	45.7	Rf2555	1	0	0	M	20	6	67	90	FSHD	5U	4AdelS	58U	4B168	-5	The Netherlands	Europe
	44.9	0.8	45.7	Rf2361	1	0	0	M	67	6	68	88	FSHD	5U	4AdelS	16U	4B163	-1	The Netherlands	Europe
	44.9	0.8	45.7	Rf132	37	0	0	M	NA	4	28	143	FSHD	6U	4AdelS	15U	4A1611L	NA	The Netherlands	Europe
	44.9	0.8	45.7		41	0	0	M	NA	6	54	111	FSHD	6U	4AdelS	21U	4B162	NA	The Netherlands	Europe
	44.9	0.8	45.7	Rf814	1	0	0	F	NA	AF	57	AF	FSHD	6U	4AdelS	59U	4A161S	NA	The Netherlands	Europe
	44.9	0.8	45.7	Rf1161	1	0	0	M	10	5	62	81	FSHD	14U	4A161S	15U	4AdelS	-17	The Netherlands	Europe
	44.9	0.8	45.7	Rf1317	1	0	0	M	NA	2	47	43	FSHD	16U	4AdelS	52U	4A161S	9	Algeria	Africa
BP2	41.5	0.3	41.8	Rf1067	306	101	102	F	12	7	55	127	FSHD	9U	4AdelL	26U	4B163	1	UK	Europe
	no deletion				1	0	0	M	NA	0	NA	0	control	15U	4B163	20U	4B163	NA	UK	Europe
BP3	5.9	0.0	5.9	Rf137	3	1	2	M	10	8	42	190	FSHD	7U	4AdelL	14U	4B163	NA	UK	Europe
	no deletion				2	0	0	F	NA	0	NA	0	control	28U	4A161L	29U	4A161S	NA	UK	Europe
BP4	65.1	2.9	68.1	Rf2587	1	0	0	M	NA	10	63	159	FSHD	5U	4AdelS	56U	4A161	NA	USA	Europe
BP5	17.8	1.4	19.2	Rf1816	2	0	0	F	NA	0	43	0	control	39U	4AdelS	21U	4A161S	NA	The Netherlands	Europe
BP6	6.3	2.3	8.6	GM18855	2	0	0	F	NA	0	NA	0	control	39U	4AdelS	6U	4A166H1	NA	Nigeria	Africa
BP7	3.5	2.5	6.0	GM19138	1	0	0	M	NA	0	NA	0	control	48U	4AdelS	18U	4A161S	NA	Nigeria	Africa
	3.5	2.5	6.0	GM18517	2	0	0	F	NA	0	NA	0	control	13U	4AdelS	71U	4B163	NA	Nigeria	Africa