The variability of *SMCHD1* gene in FSHD patients: evidence of new mutations

Introduction

The application of molecular genetics strategies into the clinical practice highlighted the existence of a large gap between the genotype and phenotype in many human disorders [1-2]. This is particularly true for neuromuscular disorders which consist of a heterogenous group of pathologies characterized by progressive weakness and wasting of proximal and/or distal muscles [3-4]. The phenotypic overlap, the limited availability of muscle biopsies and the multi-systemic events occurring in neuromuscular disorders raised the need for a multidisciplinary approach to provide an accurate diagnosis and allow targeted interventions according to the profile of each patient [3-4]. Our group is one of the two Italian Reference Centers for the genetic characterization of Facioscapulohumeral muscular Dystrophy (FSHD, OMIM #158900). FSHD affects approximately 1 in 8300 individuals [5-7]. The disease is characterized by clinical variability and incomplete penetrance, ranging from asymptomatic to wheelchair-dependent individuals [2]. Patients experience a progressive weakness of scapular girdle, facial and humeral muscles in the initial stage of FSHD. Later, the weakness can extend to the muscles of trunk and of lower-extremities, leading thereby to loss of ambulation in 20% of cases [8-9]. Two forms of FSHD are known, namely FSHD1 and FSHD2, which are characterized by identical clinical features but different genetic signatures.

FSHD1 accounts for approximately 95% of cases and it is associated with a contraction of a microsatellite repeat array on 4q35 chromosome [2]. This region is 3.3 kb long and is referred to as *D4Z4* region. In healthy individuals, the repeat consists of 11 to 100 Repeated Units (RU), whereas it is found to be 1-10 RU in FSHD1 patients. The array contraction results in the hypomethylation of *D4Z4* and, consequently, in the expression of *Double Homeobox Protein 4* (*DUX4)* that is toxic for muscle cells [10-11]. However, it is important to remark that as many as 2% of the general population presents 8-10 RU without being affected [2]. These findings suggested that the etiopathogenesis of FSHD might not be due to the *D4Z4* contraction on 4q35 alone, but to a combination of specific genetic and epigenetic signatures, which create a permissive background for the development of disease.

Approximately 5% of patients show clinical symptoms typical of FSHD, without carrying a short allele on *D4Z4* repeat array. This form is clinically identical to FSHD1 but genetically distinct and it is termed FSHD2 (OMIM #158901) [12]. FSHD2 has been associated with mutations in *Structural Maintenance of Chromosomes flexible Hinge Domain containing 1* (*SMCHD1*, 18p11.32, OMIM #614982) gene. *SMCHD1* consists of 48 exons and encodes the homonymous protein belonging to the highly conserved SMC protein family, although it is also regarded as a member of the human

Microrchidia (MORC) family. Both these groups of proteins are involved in the epigenetic regulation of the chromatin status [13-14]. In fact, *SMCHD1* is mainly involved in the regulation of high-order chromosome structures, in the inactivation of X chromosome and generally in the epigenetic regulation of chromatin repression [14-15]. In particular, *SMCHD1* contributes to the somatic repression of *DUX4* by directly binding to *D4Z4* repeat array [16]. SMCHD1 protein harbors a N-terminal *GHKL-ATPase* domain and a non-canonical C-terminal *SMC hinge* domain, both flanked by coiled-coil regions and uncharacterized domains. These functional domains are involved in the homodimerization of the protein, which is regarded as a fundamental mechanism for its activity [15]. The mutational spectrum of *SMCHD1* includes small deletions, splice site mutations and missense mutations [17]. These mutations decrease the binding activity of SMCHD1, resulting in *D4Z4* hypomethylation and incomplete repression of *DUX4*, which is thereby expressed in muscle tissue [16]. In addition, mutations of *SMCHD1* have also been shown to act as disease modifiers in FSHD patients carrying short or borderline *D4Z4* fragments [2,18]. However, the lack of a precise genotype-phenotype correlation in many cases explain the need for a more comprehensive genetic analysis of both *D4Z4* alleles and *SMCHD1*. In our experience, we observed that the majority of patients with a clinical suspect of FSHD carry a borderline (8-10RU) or normal (>11RU) sized D4Z4 fragment, in contrast to patients carrying a short fragment of 1-7RU. In this work, we report the sequence analysis of *SMCHD1* in a cohort of clinically defined FSHD patients in order to assess the distribution of *SMCHD1* variants, considering the *D4Z4* size (short fragment, borderline fragment and normal fragment).

Results

NGS and traditional methodologies proved to be useful to characterize *D4Z4* fragment, 4qA and *SMCHD1* sequence in the patient's cohort. We selected a cohort of patients representative the three categories of patients in terms of fragment size: a number of 30 patients (40%) presenting a normal range (>11RU), 13 subjects (17%) with borderline (8-10RU) and 33 (43%) patients with short fragment (1-7RU) (Fig. 1). All the patients resulted to be 4qA-positive. Successively, the sequence analysis of *SMCHD1* was performed in all patients. The extensive analysis of *SMCHD1* sequence revealed the presence of 88 variants scattered throughout the introns, exons and 3'UTR regions of the gene. The evaluation of frequency distribution and bioinformatics analysis indicated that 8 exonic variants were described as benign or likely benign, whereas 6 exonic variants were classified as Variants of Uncertain Significance (VUS) which need to be further investigated (Suppl. Table 1). In addition, 61 intronic variants were detected, which resulted to have no impact on splicing activity

(Suppl. Table 2). The frequency distribution of the above-described variants are consistent with the frequency distribution observed in the general population and are not correlated with any of the fragment size category.

Moreover, the analysis of *SMCHD1* sequence pointed out the attention on 7 pathogenic and likely pathogenic variants in 7 FSHD patients carrying a borderline or normal sized *D4Z4* fragment, namely c.182_183dupGT, c.2129dupC, c.3469 G>T, c.5150_5151delAA and c.1131+2_1131+5delTAAG, c.3010A>T, c.853G>C (Table 1).

***SMCHD1*:c.182_183dupGT**

The insertion variant c.182_183dupGT (p.Q62Vfs*48) is localized in the exon 1 of *SMCHD1* and has been found in one individual at the heterozygous state. This variant was not present in none of the annotation databases and has been predicted as disease-causing by Mutation Taster. In fact, the c.182_183dupGT may create a Premature Termination Codon (PTC), causing the termination of the amino acid sequence at the 109th amino acid (instead of the canonical 2006th codon), leading thereby to the activation of Nonsense-mediated mRNA Decay (NMD). The analysis of this variant by SMART prediction tool revealed that the truncated protein may result in the loss of its essential functional domains, namely the *GHKL-ATPase* and the *SMC hinge* domains. This alteration was also visible by the comparison of the wild-type (Fig. 2A) and variant 3D model simulated by Phyre2. Figure 2B illustrates the strong alteration of the protein secondary structures and of the subsequent conformation of the variant protein compared to the wild-type structure. The analysis on HSF indicated that the variant may impact the splicing as well, cause the disruption of a donor splice site, the creation of an Exonic Silencer Site (ESS) or the activation of a cryptic exonic donor site. According to ACMG, the c.182_183dupGT has been classified as a likely-pathogenic variant (Table2). As a matter of fact, the variant is a null variant potentially causing Loss Of Function (LOF) of *SMCHD1* (PVS1) and it is absent in ExAc, GnomAD and 1000Genome Browser (PM2). Concerning *D4Z4* region, the patient showed a 10RU contracted fragment. From a clinical point of view, the patient reported weakness of the scapular girdle muscles with difficulties to rise both the arms, in combination with an increased level of CPK enzyme (3 times higher than expected). Therefore, the *D4Z4* and *SMCHD1* findings (borderline fragment + LOF variant in *SMCHD1*) together with the observed clinical features may confirm the clinical suspect of FSHD.

***SMCHD1*:c.2129dupC**

The insertion variant c.2129dupC (p.A711Cfs*11) has been found in the exon 16 of *SMCHD1* in a single individual at the heterozygous state. This variant has not been annotated in any of the online databases. MutationTaster prediction described the c.2129dupC as a disease-causing variant, since it may create a frameshift and, consequently, a PTC at the 721th aminoacid and NMD. The analysis of

the variant effect by SMART tool suggested that the truncated protein may lose the C-terminal *SMC hinge* domain. Consistently with this data, Phyre2 showed that the 3D model predicted for the variant protein appeared to have a more compressed conformation (Fig. 2C) with respect to the wild-type structure (Fig. 2A). The HSF analysis did not reveal a potential alteration of splicing. According to ACMG, the c.2129dupC can be classified as a pathogenic variant (Table2), since it is a null variant leading to a LOF of *SMCHD1* (PVS1); it is absent in ExAc, GnomAD and 1000Genome Browser (PM2) and there is computational evidence supporting a deleterious effect on the gene product without benign-supporting predictions (PP3). *D4Z4* sizing revealed a borderline fragment (9RU) in this patient, while the clinical assessment reported a mild weakness of pelvic and proximal leg muscles, although the patient was still able to stand up from a chair without support. These data and the presence of a LOF variant in *SMCHD1* support therefore the FSHD-associated clinical symptomatology.

***SMCHD1*:c.3469 G>T**

The variant c.3469 G>T (p.G1157*) is situated within the exon 27 of *SMCHD1* and has been identified in a single patient at the heterozygous state. This variant has not been reported on the online annotation databases and has been predicted to have a damaging effect by Mutation Taster. In fact, the c.3469 G>T creates a frameshift, generating a PTC at the 1157th aminoacid and probably triggering the NMD process. In addition, the analysis performed by SMART and Phyre2 reported that the truncated protein may result in the loss of the *SMC hinge* domain and, consequently, in the disruption of secondary structure and a partial relaxation of the tridimensional conformation of SMCHD1 (Fig. 2D). This variant has also been investigated by HSF, showing that it can affect splicing through the alteration of an Exonic Splicing Enhancer (ESE) site. Following ACMG criteria, the c.3469 G>T can be described as a pathogenic variant (Table2), since it is a null variant (PVS1), it is absent on ExAc, GnomAD and 1000Genome Browser (PM2) and has been predicted to be damaging for the gene or the gene product (PP3). The analysis of *D4Z4* reported a borderline fragment (8RU) in this patient, who experienced a weakness of axial, facial, scapular, lower limbs muscles, although he maintained the ability to walk on tips. These data and the detection a truncating variant in *SMCHD1* may confirm the supposed FSHD pathology.

***SMCHD1*:c.5150_5151delAA**

The c.5150_5151delAA (p.K1717Rfs*16) has been detected in the exon 41, in a single case at the heterozygous state. This variant was predicted to have a pathogenic effect, leading to NMD and causing loss of the C-terminal *SMC hinge* domain. Moreover, the 3D model obtained by Phyre2 highlighted a maintenance of the central coiled-coil domain conformation in the truncated protein (Fig. 2E). This variant has been described as a pathogenic variant (Table2) in our previous work, in

which we described an accurate genotype-phenotype correlation within the proband and his family [19]. However, we decided to include the sample even in the present study because we performed the 3D simulation of the variant protein and we evaluated the 3'UTR region of *SMCHD1*.

***SMCHD1*:c.1131+2_1131+5delTAAG**

The intronic c.1131+2_1131+5delTAAG variant has been found downstream the exon 9 in one patient at the heterozygous state. This variant is novel and has been predicted to affect splicing and lead to NMD. In addition, the c.1131+2_1131+5delTAAG was not found in the annotation databases and has been reported as a disease-causing variant on MutationTaster. However, the prediction of the effect on the protein domains could not be performed because it is not possible to predict how the sequence and the reading frame may be altered following this variant, although it is likely to affect splicing. According to ACMG guidelines, the variant has been classified as likely-pathogenic (Table2), considering that it is a null variant (PVS1) and it is not been reported in ExAc, GnomAD and 1000Genome Browser (PM2). Interestingly, *D4Z4* sizing in the patient revealed a fragment >11RU, which is normally considered as non-pathogenic for FSHD. However, the patient experienced a severe weakness of pelvic muscles, needing thereby a double support to stand up from a chair but retaining the ability to walk independently. In this case, the FSHD clinical phenotype may be explained by the presence of a likely pathogenic variant in *SMCHD1*.

***SMCHD1*:c.3010A>T**

The c.3010A>T (p.K1004*) has been localized in the exon 24 of *SMCHD1* in a single patient at the heterozygous state. This variant has not been reported on the online annotation databases and has been predicted to have a damaging effect by Mutation Taster. In fact, the variant has been predicted to generate a PTC at the 1004th aminoacid, probably triggering the NMD process. The interrogation of HSF indicated that the variant may affect the splicing as well, causing the creation of an ESS or the alteration of ESE site. In addition, the analysis performed by SMART and Phyre2 reported that the truncated protein may result in the loss of the *SMC hinge* domain and, consequently, in the disruption of the secondary structure and the tridimensional conformation of SMCHD1 (Fig.2F). The ACMG classification of c.3010A>T described it as a pathogenic variant (Table2), since it is a null variant (PVS1), it is absent on ExAc, GnomAD and 1000Genome Browser (PM2) and has been predicted to be damaging for the gene or the gene product (PP3). The patients carrying this variant reported a fragment >11RU and experienced a mild weakness of pelvic and proximal leg muscles, although he was still able to stand up from a chair without support. Considering the genotype and clinical picture of the patient, the FSHD symptomology could be explained by the presence of a LOF variant in *SMCHD1*, although he reported a normal size fragment.

***SMCHD1*:c.853G>C**

The variant c.853G>C (p.G285R) has been detected in the exon 7, in a single patient at the heterozygous state. It is a missense variant, which has been described as disease-causing by Mutation Taster. In fact, the variant produces an aminoacid change in the *GHKL-ATPase* protein domain of SMCHD1. The HSF did not reveal a potential alteration of splicing. However, the predictive analysis performed on VarSite reported that the aminoacid substitution may be highly negative in terms of conserved aminoacid properties because of the change from a neutral (G) to a charged residue (R). Supporting this finding, interrogation of Missense3D tool revealed a damaging effect on the protein structure resulting from the steric hindrance, the introduction of a buried charge and the substitution of a buried Glycine residue, which, in turn, impair the bending of the polypeptide chain (Fig.3A-B). According to ACMG guidelines, the c.853G>C could be likely-pathogenic (Table2), considering that it is located in a mutational hotspot within a functional domain of the protein (PM1); it is absent on ExAc, GnomAD and 1000Genome Browser (PM2); it has been found in other affected family members (PP1) and has been predicted to be damaging for the gene or the gene product (PP3). The analysis of *D4Z4* sizing in the patient revealed a borderline fragment (9RU). Moreover, the patient had difficulties in the shoulders abduction and reported muscular hypotrophy proximal to legs. Given these features, the genetic profile of the patient is consistent with FSHD clinical phenotype.

Analysis of the 3'UTR region

The analysis of the 3' UTR region of *SMCHD1* revealed different variants in our patient's cohort. However, our attention was focused on c.*1376A>C (rs7238459); c.*1579G>A (rs559994); c.*1397A>G (rs150573037); c.*1631C>T (rs193227855); c.*1889G>C (rs149259359), considering that the variant alleles may disrupt an existing binding site or create a novel binding site for different miRNAs (Table 3). The rs7238459 reported a MAF= 0.263 in our patient's cohort; which overlaps the frequency observed in the general population (MAF= 0.257). According to PolymiRTS, the variant allele (C) of rs7238459 is able to disrupt a binding site for MIR7850 as well as to create a site for MIR6740.

The rs559994 had a MAF= 0.263 in our cases, resulting to be lower compared to the general population (MAF= 0.438). PolymiRTS interrogation revealed that the variant allele (A) may create a new binding site for MIR548AT.

The rs150573037 has only been detected in two patients (MAF= 0.013). Interestingly, frequency data for this Single Nucleotide Variation (SNV) are only available for the African population (MAF= 0.008) whereas it has not been observed in the European population up to date. Prediction analysis indicated that the variant allele (G) of rs150573037 may generate new binding sites for MIR515, MIR519, MIR519E and MIR5695.

The rs193227855 has been found in two patients (MAF= 0.013) in contrast to the lower frequency (MAF= 0.006) observed only in the control population of American Ancestry. Based on the PolymiRTS prediction analysis, this variant may disrupt the binding site for MIR548E and create new binding sites for MIR495 and MIR548-family members.

Finally, the rs149259359 has been reported in a single patient of the cohort and has been observed with a low frequency in the general population (MAF= 0.014). Interestingly, this is the patient carrying the *SMCHD1*_c.5150_5151delAA and already reported in our previous work [19]. The segregation analysis on his family members reported the heterozygous presence of the rs149259359 in both the affected mother and maternal uncle. The analysis performed by PolymiRTS revealed that the variant allele (C) of rs149259359 may disrupt binding sites for MIR3942, MIR4503, MIR4703, MIR6792 and MIR95, whereas it may create novel binding sites for MIR4477B, MIR651 and MIR7856.

Discussion

FSHD is one of the most difficult disease to deal with, because of the complex genetic and epigenetic background underlying its etiopathogenesis. In fact, the variable penetrance and expressivity (observed either in related or unrelated patients) does not allow an accurate diagnosis, which is further complicated by the lack of a precise genotype-phenotype correlation. Although the shortest fragments have been found in severe patients, most of the mild and moderate cases showed borderline (8-10RU) or normal-sized (>11RU) fragments (Fig.4). In these cases, the *D4Z4* analysis was not enough to explain the clinical symptomatology. We therefore decided to extend our study to the analysis of *SMCHD1* sequence, which can be helpful for genotype-phenotype correlation in FSHD patients. On this subject, our previous work described the case of a patient presenting severe FSHD symptoms, in which preliminary genetic analysis did not clarify the phenotype [19]. In fact, a contracted *D4Z4* fragment was detected both in the affected proband and the healthy father, without explaining thereby the severe symptomatology of the proband and highlighting a reduced penetrance of disease within the family. The subsequent analysis of *SMCHD1* revealed the presence of a novel pathogenic variant in the proband, which was also detected in the mother and the maternal uncle who were both affected by mild FSHD symptoms without carrying a short *D4Z4* fragment. The severe phenotype of the proband may therefore be explained by the digenic inheritance of a contracted fragment and a *SMCHD1* variant [19]. In the present study, the analysis of *SMCHD1* sequence reported 88 variants which were localized throughout the introns, exons and 3'UTR regions of the genes. Of them, 69 were classified as polymorphisms with a frequency distribution overlapping those observed in the general population. These variants are probably not related with FSHD neither with the *D4Z4*

fragment size, suggesting that they are not involved in disease etiopathogenesis. Moreover, 5 non-described VUS were also detected, but they need to be re-evaluated as more information and/or literature data will be collected concerning their potential clinical relevance in FSHD. In addition, none of them are correlated with a specific class of *D4Z4* fragment size.

Interestingly, 7 pathogenic and likely pathogenic variants were identified by *SMCHD1* sequencing, namely c.182_183dupGT, c.2129dupC, c.3469 G>T, c.1131+2_1131+5delTAAG, c.5150_5151delAA, c.3010A>T and c.853G>C. All of them were found to strongly impact the protein structure. In fact, these variants were predicted to disrupt the structure and conformation of SMCHD1 and, in most cases, alter splicing or create PTC and truncated protein products. The resulting protein have been predicted to cause the loss of *GHKL-ATPase* and *SMC hinge* domains, which are essential for SMCHD1 to maintain a repressive chromatin structure in muscle cells. These results are in line with the FSHD etiopathogenetic mechanism, which supports a toxic expression of *DUX4* as a consequence of LOF mutations in *SMCHD1*. However, functional assays are necessary to validate the real effect of the identified variants on the protein structure and function. Interestingly, gain-of function mutations localized in *GHKL-ATPase* domain of *SMCHD1* have been shown to cause severe malformations of the human nose, olfactory tract and eyes (namely, Bosma arhinia microphthalmia syndrome; BAMS), whereas LOF or dominant-negative pathogenic *SMCHD1* mutations have been found throughout the sequence of the gene [20]. Why mutations of *SMCHD1* lead to the development of FSHD rather than BAMS is still a matter of investigation. However, these data emphasize the importance of considering the genetic background of patients to clarify the clinical variability of such disorders. The present study showed that the analysis of *D4Z4* fragment and *SMCHD1* sequence were crucial to confirm the suspected clinical phenotype and accomplish a reliable genotype-phenotype correlation. Our data are consistent with Sacconi et al., 2019 who suggested that a borderline *D4Z4* fragment might be considered as a risk factor or a phenotype modifier of FSHD in patients carrying *SMCHD1* causative mutations [21]. On the other hand, patients with borderline *D4Z4* fragment who were negative to *SMCHD1* analysis, could not receive a clear molecular diagnosis, although they appeared phenotypically affected. This data highlights the fact that probably one or more unknown genes contribute to determine the permissive background for FSHD. On this subject, a recent study identified a potentially damaging mutation in the *DNA Methyltransferase 3 Beta* (*DNMT3B*, 20q11.21, #602900) gene, which is a *D4Z4*-chromatin modifier and, therefore, it represent a good candidate gene for FSHD [2,18,23]. In this context, we developed an NGS panel, including a set of candidate genes involved in the epigenetic regulation of the *D4Z4* region and genes targeted by *DUX4 (data in progress)*. Moreover, the analysis of *SMCHD1* sequence revealed the presence of variants in the 3' UTR region of the gene, which may affect the binding of

specific miRNAs or their interaction with target mRNAs. In this perspective, the rs149259359 (G/C) appeared to be the most interesting among the identified 3' UTR variants. In fact, the variant allele (C) of the SNP was predicted to disrupt the binding sites of different miRNAs, including MIR95, which is known to be overexpressed during myogenic differentiation [24]. A disruption of its binding site, may thereby affect the expression of MIR95 and its modulatory effect in myogenic cells, suggesting a potential role as a disease modifier in FSHD. Given the fact that the variant has been identified in the patient carrying the c.5150_5151delAA, we performed segregation analysis on his family members. Interestingly, the rs149259359 was detected in both the affected mother and maternal-uncle, supporting its potential implication in FSHD etiopathogenesis or severity. These findings support the role of epigenetics as hallmark and/or phenotype modifier of disease [22].

Altogether, the present study highlight how NGS platforms can be helpful to disclose *SMCHD1* as well as other candidate genes effect in FSHD pathogenesis. However, NGS still needs to be always combined with labor-intensive, outdated genetic methodologies (such as southern blotting) to better characterize the complex etiopathogenetic background of FSHD. On this subject, the recent development of alternative molecular approaches, including molecular combing and optical mapping platforms, proved to be the most feasible alternatives for FSHD molecular diagnosis and investigation [2]. Moreover, a deeper characterization of the leading mechanisms underlying the disease can be critical for undertaking the most suitable molecular assays and enable an accurate genotype-phenotype correlation [25-26]. In this perspective, integrating molecular findings and clinical data is essential to develop precision medicine protocols for FSHD patients.

Methods

Description of patients' cohort

The study involved 76 Italian individuals with a clinical suspect of FSHD enrolled at the Gemelli University Hospital Foundation of Rome and the University of Campania "Luigi Vanvitelli". Recruited patients had an average age of 50 years and a 47:53 male/female ratio. The clinical evaluation of patients was performed by specialized physicians following the dedicated guidelines [26-27]. All participants provided signed informed consent for research and publication at the time of recruitment. The study was approved by the ethics committee of Santa Lucia Foundation and complied with Declaration of Helsinki.

DNA extraction and *D4Z4* analysis

The DNA was initially extracted from lymphocytes according to standard procedures. Successively, the extracted DNA was digested on agarose plugs by EcoRI, EcoRI/BlnI and XapI restriction

enzymes and subsequently separated by Pulsed-Field Gel Electrophoresis (PFGE) as previously described [19]. The *D4Z4* RU were evaluated by southern blotting and hybridization with p13E-11 probe according to standard procedures. Linear Gel Electrophoresis (LGE) was utilized to confirm the results. In addition, 4qA and 4qB alleles were subjected to digestion with HindIII and EcoRI, PFGE and southern blot hybridization with radio-labeled 4qB and 4qA probes, according to standard procedures [19].

SMCHD1 sequence analysis

SMCHD1 gene was extensively investigated by Next Generation Sequencing (NGS) and direct sequencing, searching for putative variants located within the intronic, exonic and 3'UTR regions. To this purpose, the DNA was re-extracted from 400 µl of peripheral blood using MagPurix Blood DNA Extraction Kit and MagPurix Automatic Extraction System (Resnova) according to the manufacturer's instructions. *SMCHD1* gene was sequenced using Ion Torrent S5 and Ion Ampliseq Customized Panel, designed by Ion Ampliseq Designer (Thermo Fisher Scientific). The panel is expected to screen approximately 99.72% of target sequences, considering a minimum coverage of 20X. The construction of the library was performed by Ion AmpliSeq™ Library Kits Plus and utilizing approximately 10 ng/µl of starting DNA for multiplex PCR reactions. Two purification steps (using AMPure XP, Beckman Coulter) were performed to remove unwanted contaminants, followed by a final PCR according to manufacturer's instruction. The quality of library was evaluated by Qubit R 2.0 Fluorometer (Thermo Fisher Scientific). The enrichment procedures were performed by Ion Chef System (Thermo Fisher Scientific). Ion 510™ & Ion 520™ & Ion 530™ Kit-Chef (Thermo Fisher Scientific) were utilized for template amplification, enrichment and sequencing. Samples were run on Ion 520™ Chip (850 flows required) and Ion Torrent S5 (Thermo Fisher Scientific). The results were analyzed using Ion Reporter 5.6 (Thermo Fisher Scientific), Integrated Genome Viewer (IGV), taking the hg19 as reference genome building and NM_015295.2 as reference sequence for *SMCHD1*. The putative variants and *SMCHD1* sequence regions uncovered by NGS were analyzed by direct sequencing. To this purpose, 100 ng/µl of genomic DNA was amplified using the AmpliTaq Gold DNA Polymerase (Applied Biosystems) and PCR reagents in a total volume of 25 µL, following manufacturer's instructions. The amplified samples were sequenced using Big Dye Terminator v3.1 Cycle Sequencing Kit (ThermoFisher Scientific) and run on ABI3130xl (Applied Biosystems). Electropherograms were finally analyzed with Sequencing Analysis Software v.6 (Applied Biosystems).

Interpretation of variants

The identified variants were firstly investigated by looking at frequencies and data reported on publicly available database (1000Genome browser, ExAC, Clinvar, HGMD, GnomAD). UniProt annotation database [29] was used to obtain the aminoacid sequence and the protein domains of wild-type SMCHD1. The functional effect of the detected variants was evaluated by bioinformatic predictive tools, including Mutation Taster, Varsome, SMART, Human Splicing Finder (HSF), Phyre2, VarSite, Missense3D. In particular, MutationTaster evaluates the potential pathogenic effect of DNA sequence alterations by predicting the functional consequences of amino acid substitutions, intronic and synonymous alterations, short insertions and/or deletions (indels) and variants spanning intron-exon borders affecting splicing activity [30]. Varsome is a powerful annotation tool and search engine for human genomic variants, allowing the classification of variants according to ACMG criteria [31]. SMART, VarSite, Missense3D and Phyre2 enable the prediction of the effect of the variants on the protein structure (Kelley et al., 2015; Letunic et al., 2015; Letunic et al., 2018) [32-36]. In particular, SMART performs the analysis of the architecture of protein domains whereas Phyre2, VarSite and Missense3D are able to analyze the effect of aminoacid changes on protein structure, providing a 3D model of the predicted results. HSF predicts the effects of variants on the splicing mechanisms [37]. PolymiRTs Database 3.0 was used to analyze the variants detected within the 3' untraslated region (3'UTR) of *SMCHD1*. It allows the evaluation of the functional impact of genetic variants located in microRNA (miRNAs) seed regions and miRNAs target sites, predicting the effect on the miRNA-mRNA binding (Bhattacharya et al., 2014) [38].

Taking into account frequency and predictive results, the variants of *SMCHD1* have been classified according to the American College of Medical Genetics (ACMG) Standards and Guidelines, which help providing clinical interpretation of variants, by discriminating among benign, likely benign, uncertain significance, likely pathogenic and pathogenic variants (Richards 2015) [39].

Acknowledgements: this work is supported by 5X 2016 National Health Ministry 2.

Conflict of interest statement: the authors declare that they have no conflict of interest.

References

1. Cascella, R., Strafella, C., Caputo, V., Errichiello, V., Zampatti, S., Milano, F., Potenza, S., Mauriello, S., Novelli., G., et al. (2018) Towards the application of precision medicine in Age-Related Macular Degeneration. *Prog. Retin. Eye Res.* , **63**, 132-46.
2. Zampatti, S., Colantoni, L., Strafella, C., Galota, R.M., Caputo, V., Campoli, G., Pagliaroli, G., Carboni, S., Mela, J., Peconi, C., et al. (2019) Facioscapulohumeral muscular dystrophy

(FSHD) molecular diagnosis: from traditional technology to the NGS era. *Neurogenetics* , **20**, 57-64.

3. Turakhia, P., Barrick, B., Berman, J. (2013) Patients with neuromuscular disorder. *Med. Clin. North. Am. ,* **97**, 1015-1032.
4. von der Hagen, M., Schallner, J., Kaindl, A.M., Koehler, K., Mitzscherling, P., Abicht, A., Grieben, U., Korinthenberg, R., Kress, W., von Moers, A., et al. (2006) Facing the genetic heterogeneity in neuromuscular disorders: linkage analysis as an economic diagnostic approach towards the molecular diagnosis. *Neuromuscul. Disord.* , **16**, 4-13.
5. Gaillard, M.C., Puppo, F., Roche, S., Dion, C., Campana, E.S., Mariot, V., Chaix, C., Vovan, C., Mazaleyrat, K., Tasmadjian, A., et al. (2016) Segregation between SMCHD1 mutation, D4Z4 hypomethylation and Facio-Scapulo-Humeral Dystrophy: a case report. *BMC Med Genet.* , **17**, 66.
6. Mason, A.G., Slieker, R.C., Balog, J., Lemmers, R.J.L.F., Wong, C.J., Yao, Z., Lim, J.W., Filippova, G.N., Ne, E., Tawil, R., et al. (2017) SMCHD1 regulates a limited set of gene clusters on autosomal chromosomes. *Skelet Muscle.* , **7**, 12.
7. Alavi, A., Esmaeili, S., Nafissi, S., Kahrizi, K., Najmabadi, H. (2018) Genotype and phenotype analysis of 43 Iranian facioscapulohumeral muscular dystrophy patients; Evidence for anticipation. *Neuromuscul Disord.* , **28**, 303-314.
8. Statland, J.M., Tawil, R. (2016) Facioscapulohumeral Muscular Dystrophy. *Continuum (MinneapMinn).* , **22**, 1916-1931.
9. Mul, K., Lassche, S., Voermans, N.C., Padberg, G.W., Horlings, C.G., van Engelen, B.G. (2016) What's in a name? The clinical features of facioscapulohumeral muscular dystrophy. *Pract Neurol.* , **16**, 201-207.
10. Sharma, V., Harafuji, N., Belayew, A., Chen, Y.W. (2013) DUX4 differentially regulates transcriptomes of human rhabdomyosarcoma and mouse C2C12 cells. *PLoS One.* , **8**, 1-9.
11. Feng, Q., Snider, L., Jagannathan, L., Tawil, R., van der Maarel, S.M., Tapscott, S.J., Bradley, R.K. (2015) A feedback loop between nonsensemediated decay and the retrogene DUX4 in facioscapulohumeral muscular dystrophy. *eLife.* , **4**, 1-13.
12. De Greef, J.C., Lemmers, R.J., Camano, P., Day, J.W., Sacconi, S., Dunand, M, van Engelen, B.G., Kiuru-Enari, S., Padberg, G.W., Rosa, A.L., et al. (2010) Clinical features of facioscapulohumeral muscular dystrophy 2. *Neurology.* , **75**, 1548-1554.
13. Koch, A., Kang, H.G., Steinbrenner, J., Dempsey, D.A., Klessig, D.F., Kogel, K.H. (2017) MORC Proteins: Novel Players in Plant and Animal Health. *Front. Plant Sci.* , **8**, 1720.

14. Blewitt, M.E., Gendrel, A.V., Pang, Z., Sparrow, D.B., Whitelaw, N., Craig, J.M., Apedaile, A., Hilton, D.J., Dunwoodie, S.L., Brockdorff, N. et al. (2008) SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation. *Nature Genet.* , **40**, 663-669.
15. Jansz, N., Chen, K., Murphy, J.M., Blewitt, M.E. (2017) The epigenetic regulator SMCHD1 in development and disease. *Trends Genet.* , **33**, 233-243.
16. Lemmers, R.J., O'Shea, S., Padberg, G.W., Lunt, P.W., van der Maarel, S.M. (2012) Best practice guidelines on genetic diagnostics of Facioscapulohumeral muscular dystrophy: workshop 9th June 2010, LUMC, Leiden, The Netherlands. *Neuromuscul Disord.* , **22**, 463-470.
17. Larsen, M., Rost., S., El Hajj, N., Ferbert, A., Deschauer, M., Walter, M.C., Schoser , B., Tacik, P., Kress, W., Müller, C.R. (2015) Diagnostic approach for FSHD revisited: SMCHD1 mutations cause FSHD2 and act as modifiers of disease severity in FSHD1. *Eur J Hum Genet.* , **23**, 808-816.
18. DeSimone, A.M., Pakula, A., Lek, A., Emerson, C.P. Jr. (2017) Facioscapulohumeral muscular dystrophy. *Compr. Physiol.* , **7**, 1229-1279.
19. Cascella, R., Strafella, C., Caputo, V., Galota, R.M., Errichiello, V., Scutifero, M., Petillo, R., Marella, G.L., Arcangeli, M., Colantoni, L., et al. (2018) Digenic Inheritance of Shortened Repeat Units of the D4Z4 Region and a Loss-of-Function Variant in SMCHD1 in a Family With FSHD. *Front Neurol. ,* **9**, 1027.
20. Wilkie, A.O. (2017) Many faces of SMCHD1. *Nat Genet.* , **49**,176-178.
21. Sacconi, S., Briand-Suleau, A,. Gros, M., Baudoin, C., Lemmers, R.J.L.F., Rondeau, S., Lagha, N., Nigumann, P., Cambieri, C., Puma, A., et al. (2019) FSHD1 and FSHD2 form a disease continuum. *Neurology* , 92, e2273-e2285.
22. Lemmers, R.J.L.F., 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.* , pii: jmedgenet-2019-106168.
23. van den Boogaard, M.L., Lemmers, R.J.L.F., 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.

24. Colangelo, V., François, S., Soldà, G., Picco, R., Roma, F., Ginelli. E., Meneveri, R. (2014) Next-generation sequencing analysis of miRNA expression in control and FSHD myogenesis. *PLoS One.* , **9**, e108411.
25. Strafella, C., Caputo, V., Galota, M.R., Zampatti, S., Marella, G., Mauriello, S., Cascella, R., Giardina, E. (2018) Application of precision medicine in neurodegenerative diseases. *Front. Neurol.* , **9**, 701.
26. Cascella, R., Strafella, C., Longo, G., Manzo, L., Ragazzo, M., De Felici, C., Gambardella, S., Marsella, L.T., Novelli, G., Borgiani, P., et al. (2017) Assessing individual risk for AMD with genetic counseling, family history, and genetic testing. *Eye (Lond)* , **32**, 446-450.
27. Ricci, G., Ruggiero, L., Vercelli, L., Sera, F., Nikolic, A., Govi, M., Mele, F., Daolio, J., Angelini, C., Antonini, G., et al. (2016) A novel clinical tool to classify facioscapulohumeral muscular dystrophy phenotypes. *J. Neurol.* , **263**, 1204-1214.
28. Lamperti, C., Fabbri, G., Vercelli, L., D'Amico, R., Frusciante, R., Bonifazi, E., Fiorillo, C., Borsato, C., Cao, M., Servida, M., et al. (2010) A standardized clinical evaluation of patients affected by facioscapulohumeral muscular dystrophy: the FSHD clinical score. *Muscle Nerve* , **42**, 213-217.
29. UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* , **8**, D506-D515.
30. Schwarz, J.M., Cooper, D.N., Schuelke, M., Seelow, D. (2014) MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods.* , **11**, 361-362.
31. Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C.E., Albarca Aguilera, M., Meyer, R., et al. (2018) VarSome: The Human Genomic Variant Search Engine. *Bioinformatics*. , **35**, 1978-1980.
32. Kelley, L.A., Mezulis ,S., Yates, C.M., Wass, M.N., Sternberg, M.J. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* , **10**, 845-858.
33. Letunic, I., Doerks, T., Bork, P. (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.* , **43**, D257–D260.
34. Letunic, I., Bork, P. (2018) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* , **46**, D493–496.
35. https://www.ebi.ac.uk/thornton-srv/databases/cgi bin/DisaStr/GetPage.pl?uniprot_acc=n/a&template=home.html.
36. Ittisoponpisan, S., Islam, S.A., Khanna, T., Alhuzimi, E., David, A. & Sternberg, M.J.E. (2019) Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *J. Mol. Biol.* , **431**, 2197-2212.

37. Desmet, F.O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., Béroud, C. (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* , **37**, e67.
38. Bhattacharya, A., Ziebarth, J.D., Cui, Y. (2014) PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. *Nucleic Acids Res.* , **42**, D86-D91.
39. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* **17**, 405-424.

Tables

Table 1: Genetic characterization of the *SMCHD1* mutations identified in 7 FSHD patients.

Patient ID	*D4Z4* size	4qA	*SMCHD1*_variant position	*SMCHD1*_HGVS nomenclature
I	10RU	+	18:2656257_2656258	c.182_183dupGT
II	9RU	+	18:2707627_2707628	c.2129dupC
III	8RU	+	18:2739473	c.3469 G>T
IV	>11RU	+	18:2697122_2697125	c.1131+2_1131+5delTAAG
V	8RU	+	18:2772345_2772346	c.5150_5151delAA
VI	>11RU	+	18:2729369	c.3010A>T
VII	9RU	+	18:2688725	c.853G>C

Table 2: Prediction analysis and ACMG classification of the 7 *SMCHD1*mutations.
ESS: Exonic Silencer Site; ESE: Exonic Splicing Enhancer; WT: wild-type.

***SMCHD1* mutations**	**MutationTaster**	**SMART**	**Human Splicing Finder**	**ACMG**
c.182_183dupGT	Disease causing	Loss of *GHKL-ATPase* domain and *SMC hinge* domain	Disruption of a donor splice site; activation of an exonic cryptic donor site or creation of an ESS	Likely pathogenic
c.2129dupC	Disease causing	Loss of *SMC hinge* domain	No significant splicing motif alteration detected	Pathogenic
c.3469 G>T	Disease causing	Loss of *SMC hinge* domain	Alteration of an ESE	Pathogenic
c.5150_5151delAA	Disease causing	Loss of *SMC hinge* domain	Creation of an ESS or alteration of an ESE	Pathogenic
c.1131+2_1131+5delTAAG	Disease causing	NA	Alteration of the WT donor site	Likely Pathogenic
c.853G>C	Disease causing	No significant alteration of domain organization	No significant splicing motif alteration detected	Likely Pathogenic
c.3010A>T	Disease causing	Loss of *SMC hinge* domain	Creation of an ESS or alteration of an ESE	Pathogenic

Table 3: Bioinformatic prediction of 3'UTR variants altering the match to the miRNA seed region.
MAF: Minor Allele Frequency. *Calculated on 76 patients, # Referred to 1000Genomes allele frequencies

Genomic position	**SNP**	**MAF FSHD***	**MAF EUR#**	**Effect of the variant allele**	**Targeted miRNA**
18:2803926	rs7238459 (A/C)	C: 0.263	C: 0.257	Disruption of a conserved miRNA site	MIR-7850
				Creation of a new miRNA site	MIR-6740
18:2804129	rs559994 (G/A)	A: 0.263	A: 0.438	Creation of a new miRNA site	MIR-548AT
18:2804439	rs149259359 (G/C)	C: 0.006	C: 0.014	Disruption of a conserved miRNA site	MIR-3942
					MIR-4503
					MIR-4703
					MIR-6792
					MIR-95
				Creation of a new miRNA site	MIR-4477B
					MIR-651
					MIR-7856
18:2803947	rs150573037 (A/G)	G: 0.013	G: 0.000	Creation of a new miRNA site	MIR-515
					MIR-519D
					MIR-519E
					MIR-5695
18:2804181	rs193227855 (C/T)	T: 0.013	T:0.000	Disruption of a conserved miRNA site	MIR-548E
				Creation of a new miRNA site	MIR-495
					MIR-548AC
					MIR-548AE
					MIR-548AH

					MIR-548AJ
					MIR-548AM
					MIR-548AQ
					MIR-548D
					MIR-548H
					MIR-548J
					MIR-548X
					MIR-548Z
					MIR-5688

Figure Legends

Fig. 1: distribution of the *D4Z4* fragment size which have been subdivided in order to distinguish short (1-7RU), borderline (8-10RU) and normal sized (>11RU) fragments. RU: Repeated Units.

Fig. 2: A. Predicted conformation of the three wild-type domains of SMCHD1, based on the domain organization released by Uniprot (Entry: A6NHR9). In particular, the N-terminal region (1-702 AA) harboring the *GHKL-ATPase* domain (111-702 AA) is based on the template c5ix1A (PDB header: transcription; PDB Molecule: morc family cw-type zinc finger protein 3; PDBTitle: crystal structure of mouse morc3 atpase-cw cassette in complex with2 amppnp and h3k4me3 peptide). The central coiled-coil domain (703-1719 AA) is based on the template c4e9lA (PDB header: cell adhesion; PDB Molecule: attaching and effacing protein, pathogenesis factor; PDBTitle: fdec, a novel broadly conserved escherichia coli adhesin eliciting 2 protection against urinary tract infections). The C-terminal region (1720-2005 AA) harboring the *SMC hinge* domain (1720-1847AA) is based on c2wd5A (PDB header: cell cycle Chain: A: PDB Molecule: structural maintenance of chromosomes protein 1a; PDBTitle: smc hinge heterodimer (mouse). **B-F**: 3D model predicted by Phyre 2 tool. The structure resulting from the presence of the c.182_183dupGT (**B**) is based on the template d1e9ya1 (Fold: beta-clip Superfamily: Urease, beta-subunit). The structures resulting from the presence of the c.2129dupC and c.3469G>T (**C** and **D**, respectively) are based on the template c5ix1A (PDB header: transcription. PDB Molecule: morc family cw-type zinc finger protein 3). The structure resulting from the c.5150_5151delAA (**E**) is based on the template c4e9IA (PDB header: cell adhesion; PDB Molecule: attaching and effacing protein, pathogenesis factor). The structure referred to the

c.3010A>T (**F**) is based on the template c5ix1A (PDB header: transcription. PDB Molecule: morc family cw-type zinc finger protein). The 3D model simulation of the *SMCHD1*_c.1131+2_1131+5delTAAG is not available because the aminoacid sequence alteration following this variant cannot be predicted.

Fig. 3: predicted structure of the N-terminal region structure of SMCHD1 showing the aminoacid change resulting from the c.853G>C. The predicted models are based on the on the template c5ix1A (PDB header: transcription; PDB Molecule: morc family cw-type zinc finger protein 3; PDBTitle: crystal structure of mouse morc3 atpase-cw cassette in complex with2 amppnp and h3k4me3 peptide). **A.** SMCHD1 structure showing the wild-type residue (G). **B.** SMCHD1 structure showing the variant residue (R).

Fig. 4: histogram reporting the non-linear correlation between the *D4Z4* fragment size and the severity of disease in the patient's cohort analyzed in the present study. RU: Repeated Units.