A human pan-genomic analysis provides insights into the genetic and epigenetic make up of facioscapulohumeral muscular dystrophy

Authors: Valentina Salsi[1*], Matteo Chiara[2,3*], Sara Pini[1], Paweł Kuś[4], Lucia Ruggiero[5], Silvia Bonanno[6], Carmelo Rodolico[7], Stefano C. Previtali[8], Maria Grazia D'Angelo[9], Lorenzo Maggi[6], Diego Lopergolo[10,11], Marek Kimmel[4,12], Filippo M. Santorelli[13], Graziano Pesole[3,14], Rossella G. Tupler[1,15,16]

Authors' Affiliations:

1. Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Modena, Italy. valentina.salsi@unimore.it; pini.saraps@gmail.com; rossella.tupler@unimore.it.

2. Dipartimento di Bioscienze, Università degli Studi di Milano, Milan, Italy. matteo.chiara@unimi.it.

3. Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, Consiglio Nazionale delle Ricerche, Bari, Italy. matteo.chiara@unimi.it; graziano.pesole@uniba.it.

4. Department of Systems Biology and Engineering, Silesian University of Technology, Gliwice, Poland. pawel.s.kus@protonmail.com; kimmel@rice.edu

5. Department of Neurosciences, Reproductive and Odontostomatological Sciences, University Federico II of Naples, Naples, Italy. ruggilucia@gmail.com.

6. IRCCS Foundation, C. Besta Neurological Institute, Milan, Italy. Silvia.Bonanno@istituto-besta.it; Lorenzo.Maggi@istituto-besta.it.

7. Department of Neurosciences, Policlinico "G. Martino", University of Messina, Messina Italy. carmelo.rodolico@unime.it.

8. INSPE and Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milan, Italy. previtali.stefano@hsr.it.

9. Department of Neurorehabilitation, IRCCS Eugenio Medea, Bosisio Parini, Italy. grazia.dangelo@lanostrafamiglia.it.

10. Department of Medicine, Surgery and Neurosciences, University of Siena, Siena, Italy. diego.lopergolo@unifi.it

11. UOC Neurologia e Malattie Neurometaboliche, Azienda Ospedaliero Universitaria Senese, Policlinico Le Scotte, Viale Bracci, 2, 53100 Siena, Italy. diego.lopergolo@unifi.it

12. Departments of Statistics and Bioengineering, Rice University, Houston, TX, United States. kimmel@rice.edu

13. IRCCS Fondazione Stella Maris, 56128 Pisa, Italy. filippo3364@gmail.com

14. Department of Biosciences, Biotechnology and Environment, University of Bari "A. Moro", Bari, Italy. graziano.pesole@uniba.it.

15. Department of Molecular, Cell, and Cancer Biology, University of Massachusetts Medical School, Worcester. rossella.tupler@unimore.it.

16. Li Weibo Institute for Rare Diseases Research at the University of Massachusetts Medical School, Worcester. rossella.tupler@unimore.it.

*The authors equally contributed to the work

Corresponding Author:

Rossella Tupler, M.D., Ph.D., Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Email: rossella.tupler@unimore.it).

ABSTRACT

Background: Facioscapulohumeral muscular dystrophy (FSHD) is the only human disease associated with epigenetic changes at a macrosatellite array. Alleles with fewer than 10 D4Z4 elements at 4q35 are found in 95% of FSHD cases; remaining cases carrying D4Z4 alleles with 11 or more repeats bear mutations in chromatin remodeling factors. Reduced D4Z4 CpG methylation is used for the molecular diagnosis of FHSD, but D4Z4-like sequences dispersed in the genome can generate ambiguous results.

Results: Using an original analytical approach based on complete haplotype level assemblies from the T2T-CHM13 human genome and the human pangenome project, which includes 86 haploid genomes, we uncovered the extensive number of D4Z4-like elements and their widespread inter- and intra-individual variability, which were not represented in the GRCh38 reference. Bisulfite-treated DNAs from 44 subjects (29 FSHD index cases, 15 relatives) were analyzed through high-depth methylation sequencing assaying 82 CpGs spanning the D4Z4 repeat. Samples stratified in three groups with low, intermediate, and high CpG methylation. Low D4Z4 methylation was associated with variants in the chromatin remodeling factor *SMCHD1,* but not with the patients'clinical phenotypes.

Conclusions: This is the first study showing the relevance of the pangenome and T2T assemblies for investigating the genotype-phenotype correlation in genetic diseases. The extension and the variability of D4Z4-like elements scattered throughout the human genome and the variable phenotypes advocate for a critical revision of FSHD diagnostic tests based

on D4Z4 CpG methylation assays and indicate that they must be complemented by family studies for the proper interpretation of results.

Keywords:

FSHD
D4Z4 macrosatellite
Repetitive DNA elements
T2T genome assembly
Methylation
Comparative genomics
Pangenome assemblies
Genotype-phenotype correlation

Background

Facioscapulohumeral muscular dystrophy (FSHD) is a rare familial disease with an estimated prevalence from 1/8,000 to 1/20,000 (1,2). It is the third most common form of hereditary myopathy and is the only hereditary human disease associated with variations of a 3.3 kb macrosatellite (3), termed D4Z4, located on the chromosome 4 subtelomere, at 4q35. Each copy of the 3.3 kb D4Z4 repeat contains two homeobox sequences, that are part of a putative open reading frame called DUX4 ORF (4), and two different classes of GC-rich repetitive DNA (4,5): hhspm3, a member of a low copy human repeat family (6), and LSau, a middle repetitive DNA family (7,8). Several human chromosomes contain repeat sequences with high similarity to D4Z4 (D4Z4-like) (4,5), including a large tandemly arrayed D4Z4 repeat with 98% identity located at 10q26 (9).

Two genetically distinct FSHD subtypes, FSHD1 and FSHD2, are currently described. In FSHD1 (OMIM #158900), the reduction of 4q35-D4Z4 repeats below a critical threshold (10 or fewer repeat units, RU) (10) is thought to determine epigenetic alterations and the inappropriate expression of nearby genes, leading to disease. Alleles with reduced D4Z4 copy numbers (4-8 RU) however, have the frequency of a common polymorphism being present in the 3% general population (11–14), which is not compatible with the incidence of FSHD. To address this issue, it has been proposed that additional sequence elements, *in cis* to the D4Z4 repeat array are necessary for disease development (15–17). Two distinct 4q

alleles labeled qA and qB, and distal to the D4Z4 array have been described. The qA allele carries pLAM: a 260-bp sequence, which provides a 3'-terminal exon (Exon 3) and a polyadenylation signal (PAS) (AUUAAA) in juxtaposition to the terminal copy of DUX4 ORF on 4q (18,19). A similar arrangement is observed on chr 10, where instead the PAS signal is disrupted by a single nucleotide variant (AUCAAA). The so-called 4qA-PAS "permissive haplotype" is thought to permit the expression of DUX4 ORF from the most distal D4Z4 repeat. Instead, the non-permissive 4qB haplotype lacks the pLAM sequence and DUX4 expression is not gained despite the epigenetic status of the D4Z4 array. Aberrant expression of DUX4 is considered the molecular cause of FSHD1 (Additional file 2: Fig. S1). However, D4Z4 alleles of reduced size (DRA) with 4qA-PAS have a high frequency (2%) in the human population (13).

FSHD2 (OMIM #158901) is contraction-independent with affected individuals carrying two D4Z4 arrays in the healthy range (>10 RU). FSHD2 has been linked with mutations in chromatin remodeling factors genes i.e., *SMCHD1* (for structural maintenance of chromosomes flexible hinge domain containing 1) (20), *DNMT3B* (for methyltransferase 3B) (21) and *LRIF1* (ligand-dependent nuclear receptor interacting factor 1) (22). It has been proposed that, similar to the reduction of D4Z4 RU, the inactivation of these genes alters the epigenetic configuration of the D4Z4 array at 4q35 and causes the inappropriate expression of DUX4 when a 4qA-PAS haplotype is present at 4q.

In the clinical setting, a reduction of CpG methylation of D4Z4 is considered a faithful indicator of anomalous gene expression and it is used as a proxy of disease status (23). However, previous studies reported a wide variability in D4Z4 methylation levels among FSHD index cases and FSHD families with no clear association between D4Z4 methylation status and disease manifestation or severity (24,25).

Until very recently only the euchromatic fraction of the genome was mapped while relevant heterochromatic regions remained uncharted. In the excess of 151 Mbp of sequences distributed throughout the genome are missing in the GRCh38 (hereafter hg38) reference assembly released by the Genome Reference Consortium (GRC) in 2013. These uncharacterized sequences include pericentromeric and subtelomeric regions, ribosomal DNA (rDNA) arrays, short arms of acrocentric chromosomes and centromeric satellite arrays (26). All these sequences are highly repetitive and polymorphic (27,28).

Only the advancement of long-read sequencing technologies has allowed filling these gaps. In 2022, the T2T-CHM13 assembly (hereafter T2T) released by the Telomere-to-Telomere Consortium has provided a complete representation of a human genome sequence, which includes gapless telomere to telomere assemblies for all chromosomes (except chromosome Y) (26). More recently, the human Pangenome Reference Consortium (hereafter PGR), has reported for the first time (17), complete haplotype-level assemblies of human genomes from individuals of different geographic ancestries. These methodological advances are a turning point for the studying size and distribution of tandemly arrayed repetitive elements in the genome and open new possibilities of interpretation for complex diseases such as FSHD.

Here, we report the results of comparative analyses of 86 haplotype-level assemblies from the PGR (29) and the outcome of high-throughput CpG methylation assay of the D4Z4 repeat based on the T2T. Our study displays all the potential of complete, high-quality, haplotype/individual level genome assemblies for the analysis of D4Z4 arrays in human diseases, and advocate revisions of previous findings based on the hg38 reference.

Results

T2T and PGR reveal the extended variability of D4Z4 repeats in the human genome

The extent of D4Z4-like repeat elements was investigated by analyzing T2T and 86 haplotype-level genome assemblies from the PGR (29). D4Z4-like elements were annotated by sequence similarity and classified in five groups based on their size: 300-500 bp; 501-1000 bp; 1001-1500 bp; 1501-2500 bp; 2501-3300 bp. Matches of in-between 3200 and 3300 bp in size were regarded as “complete” D4Z4 repeats.

Our analyses uncovered hundreds of D4Z4-like elements (hereafter D4Z4-l), irrespective of the reported geographic origin of the subjects (Additional file 1: Table S1). Cumulatively D4Z4-l spanned in-between 700 Kb to 1.5 Mb of sequence in the haplotype-level assemblies of the 43 subjects from the PGR and accounted for approximately 1.2 Mb of sequence in T2T. D4Z4-l were scattered through several chromosomes in T2T and accounted for hundreds of Kbs of sequences on chr 1, chr 4, chr 10, chr 15, chr 21 and chr 22 (Fig. 1a). Notably, only 86.5 Kb of D4Z4-l were observed in hg38, representing a more than 10-fold reduction compared with T2T and PGR assemblies (Fig. 1b). In hg38, D4Z4-l were prevalently associated with the long arm of chr 4 and chr 10 (Fig. 1a) where they are

arranged as two large arrays of 8 and 10 complete D4Z4 repeats, respectively. PGR and T2T instead harbored a large number (minimum 285, maximum 626, Additional file 1 Table S1) of "incomplete" D4Z4-l repeats of < 2.5 Kb in size, which are lacking in hg38. Conversely (Fig. 1c), the number of complete D4Z4 elements included in hg38 was within the range of variability observed in T2T and PGR (26 to 121). Interestingly, subjects from AFR (African) ancestry were associated with a higher variability in D4Z4-l copy numbers (Kolmogorov Smirnoff test p-value p-value = 0.0006403), while a narrower range was observed in individuals of EAS (Eastern Asia) and AMR (South America) ancestry (Additional file 2: Fig. S2).

Table 1: Characterization of complete D4Z4 elements clusters.

Cluster	Chromosome	% sequences	Median number of complete repeats (per haplotype)	Average substitution rate per 100 bp	Substitution rates estimates at haplotype level
1	Chr 10	87%	19	0.13	0.015-0.21
2	Chr 4		22	0.245	0.164-0.4312
3	Chr 14	2.3%	0.5	1.73	NA
4	Chr 15, Chr 21	3.1%	1	1.81	NA
5	Chr 22-1	1.5%+6.1%	0	0.798	NA
6	Chr 22-2		2	1.24	NA

Highly conserved regions of complete (> 3.2 Kb in size) D4Z4 elements were aligned and clustered by sequence similarity (Fig. 1c). Six distinct groups were defined. D4Z4 repeats from T2T were used to anchor/assign groups to chromosomes. The largest clusters (Table 1) and the majority of the sequences (87%) were assigned to either chr 4 (cluster-2) or chr 10 (cluster-1); this notwithstanding, complete D4Z4 elements were also observed on chr 14 (cluster-3), chr 22 (cluster-5 and cluster-6), chr 15 and chr 21 (cluster-4). The distribution of the number of different types of D4Z4-l across haplotypes is summarized in Fig. 1d and Table 1. Chr 4 and chr 10 were associated with a median number of 22 and 19 repeats, respectively. The difference was statistically significant according to a Kolmogorov Smirnov test (p-value: 0.02518). Complete D4Z4 repeats on chr 10 displayed lower variability compared with those on chr 4, with an average substitution rate of 0.13 and 0.245 substitutions per 100 bp, respectively (Additional file 2: Fig. S3). Substitution rates estimates at haplotype levels suggested a high inter-individual variability, with rates ranging from 0.164 to 0.4312 substitutions for 100 bp for chr 4 and from 0.015 to 0.21 for chr 10, respectively

(Additional file 2: Fig. S4). Interestingly, less than 50% of the haplotype assemblies in the PGR had one sequence assigned to cluster-3 (chr 14), cluster-4 (chr 15-21) and cluster-5 (chr 22), implicating that, contrary to T2T, the majority of the haplotype-level genome assemblies from the PGR does not incorporate D4Z4 repeat sequences associated with these clusters. Instead, more than 75% of the assemblies carried D4Z4-l assigned to cluster-6 (chr 22) (Table 1).

The pangenome assemblies show the high frequency of the 4qA-PAS haplotype associated with FSHD in the world population

The molecular signature, named 4qA-PAS, has been proposed to define alleles causally related to FSHD. On this basis, healthy subjects carrying reduced D4Z4 alleles would be explained by the absence of the 4qA-PAS. Analyzing 801 healthy subjects from the Italian and Brazilian populations we previously found that the 4qA-PAS associated with a D4Z4 array with 4-8 RU has high frequency 2% (13). We thus exploited the T2T and PGR to investigate the presence of this molecular signature in other populations. The telomeric qA and qB sequences (Additional file 3) were used to identify and classify terminal D4Z4 repeat in PGR, T2T and hg38. A total of 173 D4Z4 terminal repeats were labeled, based on the presence of a qA or qB element *in cis* and within 15 Kb from a complete DUX4 coding sequence (Additional file 1: Table S2). Terminal repeats were tentatively assigned to chr 4 or chr 10 based on sequence similarity profiles (see Fig. 1c). Haplotypes with inconsistent annotations (2 or more qA/qB probes assigned to the same chromosome and none to the other chromosome) were not included in subsequent analysis. By this approach the complete set of alleles at D4Z4 repeat loci at chr 4 and chr 10 was determined for 72 haplotypes from PGR (Additional file 1: Table S2). DUX4 coding sequences from the last repeat were aligned and clustered by sequence similarity. DUX4 sequences assigned to 10q D4Z4 terminal repeat formed a single cluster; while 4q terminal DUX4 were partitioned in 3 groups which included 4, 46 and 21 sequences respectively (Fig. 2a-b). Interestingly (Fig. 2c), chr 10 and chr4_group 2 DUX4 sequences displayed lower levels of variability (average 0.061 subs per 100 bp) compared to both chr4_group 3 (average 0.12 subs per 100 bp) and chr4_group 1 (average 0.142 subs per 100 bp).

The pLam reference sequence was used (Additional file 3) to annotate the presence/absence of pLam element at the terminal repeats and verify the integrity of the AUUAAA polyadenylation signal (PAS). Consistent with previous observations, all the alleles

classified as qB did not carry a pLam; and qB sequences were more proximal to DUX4 compared to qA (Fig. 2d-g). In the same way, (Fig. 2e-f) the vast majority of alleles classified as qA (112/119) were associated with a pLam both on chr 4 (41/47) and chr 10 (71/72). Only terminal 4q pLam had valid PAS; while in pLam assigned to 10q PAS were disrupted by a single nucleotide substitution (AUUAAA→AUCAAA). Perfect conservation of PAS/PAS-like elements was observed both on 4q and 10q terminal D4Z4 repeats. One qB haplotype was assigned to 10q. Remarkably, 4 individuals carried a short (<8 units) D4Z4 qA-PAS haplotype at 4q (Fig. 2f), that is considered the molecular signature of FSHD. No statistically significant difference in the total number of chr 4 D4Z4 repeats was observed when qA/qB alleles pLam+/pLam- alleles were compared (Fig. 2h-i).

D4Z4 hypomethylation correlates with *SMCHD1* pathogenic variants but not with FSHD clinical phenotype

Besides the evaluation of 4q D4Z4 copy number, bisulfite treated DNA sequencing at targeted regions is commonly used to diagnose FSHD (30). Results shown above uncover all the heterogeneity of D4Z4 repeats in the human genome, including the presence of qA D4Z4 alleles with 8 repeats or fewer, and raise important concerns about the reliability of methylation/diagnostic tests based on the hg38 reference assembly. These considerations prompted us to investigate D4Z4 CpG methylation using the T2T. We devised a methylation assay based on hg38 to substantiate previous findings (25). We included 26 FSHD2 index cases and 14 relatives all carrying 10 or more RU. All were searched for *SMCHD1* variants. In addition, 3 FSHD1 index cases and 1 relative carrying 4 RU were studied. For this purpose, D1, D3, D5 and D6 primers (Additional File 2: Fig. S5) were designed. Bioinformatics analyses identified different potential target regions for D1, D3, D5 and D6 primers amplification on the T2T and hg38 assemblies (Additional file 4: Table S3), when bisulfite converted (BSC) and non-converted sequences were considered. Moreover, intrinsic differences in the proportion of aligned reads were observed when non-converted short reads were mapped to hg38 and T2T (hg38 77% reads mapped; T2T 99.43% reads mapped). Notably, the majority of mapped reads (62% for T2T and 54% for hg38) did not align at primer target regions. Short read-coverage profiles were computed, and genomic intervals covered by 100 or more non-converted reads in at least 50% of the samples were recorded. By this approach a total of 57 and 172 high-coverage genomic regions were identified in the

hg38 and T2T assembly, respectively; of these 87% and 98% had at least a high similarity match (17 bp with at most 1 mismatch) to one or more primer sequences.

Table 2: Titration analysis for the identification of the optimal sequence similarity threshold.

Identity-level (%)	# regions	size (Kb)	uniquely mapped BSC reads (%)	# testable CpGs
99	145	67,41	5,43%	237
98	103	59,13	12,13%	411
97	87	54,01	26,74%	873
96	73	46,14	32,47%	997
95	**60**	**40,03**	**49,17%**	**1176**
94	56	38,71	51,01%	1083
93	53	37,55	53,43%	891
92	46	36,49	58,47%	766
91	45	33,02	72,71%	637
90	42	32,75	75,41%	451

Identity-level (%): sequence identity threshold. #regions: number of distinct regions defined by the threshold. size (Kbp): total size of the regions in Kbp. uniquely mapped BSC reads (%): percentage of uniquely mapping BSC reads. # testable CpGs: total number of distinct CpGs for which more than 10 reads were available for more than 25 subjects.

To mitigate these limitations and obtain a non-redundant representation of highly similar D4Z4 elements, repetitive sequences were collapsed. Different sequence identity thresholds were tested, and a titration analysis was performed by evaluating the total number of non-ambiguously mapped BSC reads (Table 2). Considerations based on the number and cumulative size of merged genomic regions, and on the number of CpGs covered by at least 10 BSC reads (Testable CpGs in Table 2), prompted us to select 95% identity as the most appropriate threshold. A total of 60 and 9 distinct consensus groups (groups of sequences GS) were defined on T2T and hg38 (Additional file 5: Table S4 and Table S5). These regions were used for the analysis of methyl-seq data.

Table 3: GS containing CpGs covered by at least 10 BSC reads in at least 50% of samples.

Assembly	Group	Chromosome	N° CpGs
T2T	GS1	Chr4, Chr10	28
	GS2		32
	GS5	Chr 13, Chr 15	16
	GS6		5
hg38	GS3	Chr4, Chr10	26

Remarkable differences were observed in the completeness, number and breadth of testable CpGs when the hg38 and T2T genome references were used: hg38 125 CpG; T2T

938 CpG. Global methylation patterns were summarized in the form of a heatmap (Additional file 2: Fig. S6 and Fig. S7). Only CpGs covered by at least 10 BSC reads in at least 50% of samples were considered. A total of 4 distinct groups and 81 distinct CpGs had complete methylation profiles (no missing data) across all samples (GS1 (primer D6): 28CpGs; GS2 (primer D1): 32CpGs; GS5 (chr13-ch15): 16CpGs; GS6 (chr 13-chr 15): 5CpGs) under these criteria. GS1 and GS2 included 4q/10q-specific regions exclusively. A single group, GS3 (primer D1) and a total of 26 CpGs fulfilled these strict criteria when hg38 was considered (Fig. 3, Fig. 4 and Table 3). In light of these observations, methylation profiles inferred from T2T were considered more accurate/complete and used for subsequent analyses. Only primer D1 had complete data in both T2T and hg38; however, CpGs methylation levels estimated on hg38 were slightly more elevated than those observed on T2T at the equivalent target region (Additional file 2: Fig. S8, median values: 49.13 hg38, 43.75 T2T). An observation that might be consistent with discrepancies in the mapping of BSC reads on different genome assemblies.

Methylation patterns at GS1 (primer D6) and GS2 (primer D1) stratified samples into 3 groups: low methylation (average CpG methylation 24%), intermediate methylation (average CpG methylation 48%), and high methylation (average CpG methylation 71%) (hereafter named). The methylation pattern of GS5 and GS6 was highly uniform across all samples, with an average of 89% CpG methylation.

Interestingly, while all of the non-FSHD affected individuals displayed high levels of CpG methylation at both primer D1 and primer D6 (see C2 in Fig. 3), the observed methylation profiles did not provide a clear stratification of clinical FSHD phenotypes; consistent with this finding no statistically significant differences were recovered when average methylation levels of CCEF clinical categories A, B and D were compared (Additional file 2: Fig. S9). Conversely, the majority (11/16) of participants in the low methylation group carried Pathogenic, Likely Pathogenic variants or Variants of Uncertain Significance according to American College of Medical Genetics and Genomics classification (31) (P/LP/VUS variants) in *SMCHD1* (Fisher exact test p-value 0.0003). This association was confirmed by the comparison of the methylation profile distributions of *SMCHD1* P/LP/VUS variants carriers and non-carriers (Fig. 3, Kolmogorov-Smirnov test p-value <= 1e-16). Identical patterns were recovered also when methylation levels determined on the hg38 assembly were considered (Additional file 2 Fig. S9).

Overall distributions of CpG methylation level, as displayed in Additional file 2: Fig. S10 highlights a high variability at primer targeted regions GS1 (primer_6) and GS2 (primer_1), while off-target regions GS5 and GS6 have high methylation levels, with a more compact distribution. Equivalent patterns are observed when all of the 60 distinct genomic regions defined in our analysis are considered (Additional file 2: Fig. S11).

Discussion

The pangenome assemblies provide novel hints to evaluate the significance of the D4Z4-like sequences in the clinical setting

Thirty years ago, reduction of the number of tandemly arrayed 3.3 D4Z4 repeats at 4q35 was causally associated with FSHD. The D4Z4 macrosatellite belongs to a family of repetitive elements characteristic of heterochromatin and it was then known that many D4Z4-l were present in the human genome (4). However, the lack of a complete genome assembly hampered the possibility of fully deciphering the significance of molecular and epigenetic findings in FSHD. Recent developments in sequencing technologies, and the availability of complete, high quality haplotype level genome assemblies such as the T2T and the PGR have opened unprecedented opportunities to study the molecular function of repetitive DNA elements in humans. By analyzing the haplotypes from the PGR we collected relevant information on the size, the distribution and the composition of the D4Z4 locus, including distal sequence elements that are considered a hallmark of the FSHD molecular signature. As summarized in Table 4 we established that the number of 3.3 kb D4Z4 repeats included in each array ranges from 6 RU to 89 RU (14). PGR analysis (Table 4) confirmed that qA sequences are more distal compared to qB sequences, even though not all qA sequences identified presented the same distance from the last D4Z4 repeat; the pLam sequence was only identified in association with qA haplotypes (18); and valid PAS were exclusively associated with 4q (17); chr 10 D4Z4 repeats were all marked by qA, with a single exception: a 10qB allele carrying 1 RU. Notably, previous data regarding the high frequency of the permissive 4qA-PAS haplotype, comparable to a common polymorphism, were confirmed (13). In fact, in 4.6% of cases we identified permissive haplotypes consisting of reduced 4q D4Z4 arrays with ≤ 8 RU and the 4qA-PAS. No skewed association has been detected between 4q D4Z4 arrays copy number and qA/B alleles (19). Our work (Table 4) showed the consistency between data obtained from the analysis of PGR and previous observations

(4,5,13) confirming the reliability of the PGR and long reads based sequencing technologies to advance our knowledge of the molecular signature used to diagnose FSHD.

The PGR analysis also identified 6 haplotypes with peculiar annotations where 2 or more qA/qB probes were assigned to the same chromosome and none to the other chromosome (Additional file 1: Table S2). This condition has been reported before (32) and is ascribable to the spreading of repetitive and polymorphic D4Z4-l at different loci of the genome (33). For example, 21% of analyzed cases show translocations between 4q and 10q repeats (14,34). In addition, in about 6% of cases more complex profiles are identified (32). The high frequency of complex profiles such as mosaicisms, undetectable alleles, duplications, deletions or inversions of D4Z4 repeated elements (35,36) can all be explained by non-allelic homologous recombination events promoted by the recombinogenic nature of this type of repeats. Analyses of large cohorts of cases and controls will be required in the near future to infer general rules to be applied in the clinical setting for the evaluation of these alleles.

Finally, our comparative analyses of complete D4Z4-l (Fig. 1c), highlighted that D4Z4 repeats, as well as the DUX4 ORF in the last repeat, located at chr 4 present a higher variability compared to chr 10. In addition, PGR haplotypes from different ancestries showed different levels of variability in the number of complete D4Z4-l, with a significant difference in D4Z4 copy number between individuals of AFR ancestry, native Americans and far east Asians.

D4Z4 methylation assay must consider the T2T sequence for the identification of 4q/10q specific CpGs

Reduced CpG methylation at D4Z4 has been widely studied and proposed as a biomarker for the presence of FSHD. Previous studies of D4Z4 methylation presented some drawbacks: technical limitations and lack of reproducibility led to discordant results regarding which regions are the most representative for D4Z4 methylation status (24,37–41). In addition, a clear correlation between D4Z4 methylation level and FSHD clinical status has not been demonstrated (25). Our results can explain these discrepancies. In our work, we obtained a comprehensive evaluation of the D4Z4 methylation status considering a high number of reads and excluding technical biases introduced by the previous studies. The alignment on the T2T of the D4Z4 amplicons obtained with the primer sets we designed (Additional file 2:

Figure S5) allowed to establish that the primers sets BSS-D1 and BSS-D6 identify CpG methylation signals deriving from 4q/10q-specific regions and discriminate the primer sets (BSS-D3 and BSS-D5) amplifying D4Z4-l distributed not only at 4q and 10q but also in chromosome 1 and on the short arms of acrocentric chromosomes. Thus, the study of D4Z4 CpG methylation can be hazardous if the existing variability is not correctly taken into account.

Despite all this, our approach presents some technical limitations, due to the application of short reads-based assays for the analysis of repetitive elements of relatively large size. When non-converted reads were aligned to both hg38 and T2T, more than 50% of mapped reads did not align at expected target regions. Nevertheless, the total fraction of mapped reads and the number of testable CpGs were higher when aligned to T2T compared to hg38, confirming that T2T provides a more accurate representation of the human genome. It is easy to envisage that long read sequencing technologies, such as Oxford Nanopore, will provide relevant technical advances and open new possibilities for the study of repetitive elements' organization and function in the genome. In this respect, it is worth noting that Butterfield and colleagues (2023) (42) applied targeted nanopore sequencing to FSHD patients and healthy subjects and showed that an asymmetric methylation gradient forms in a length-dependent manner at the proximal end of the D4Z4 repeat array reaching saturation approximately at the 10th repeat. They also observed a highly similar methylation pattern at 4q and 10q and irrespective of the clinical phenotype (FSHD or healthy), shedding new light on the interpretation of the role of the last D4Z4 repeat (42). This notwithstanding, at present long read sequencing technologies are not mature for systematic large-scale applications in the clinical settings. Instead, the targeted CpG methylation analysis we set up is manageable on a larger scale as it follows a simple protocol, is less expensive and highly reproducible. Remarkably, our data are consistent with the results obtained by Butterfield and colleagues.

D4Z4 CpG methylation level is not a predictor of FSHD disease

Beside the formal demonstration that primer design is a crucial point in the analysis of D4Z4 methylation, our study also revealed that contrary to common knowledge (23), not all DNAs from myopathic patients (29 index cases, 4 relatives) presented a low methylation level. Our analysis identified three clusters with low, intermediate and high D4Z4 CpG methylation (mean at 24%, 48% and 71% respectively). We observed that cases presenting the classical

FSHD features (CCEF category A) are distributed across all clusters and in similar proportions: low methylation 62.5% (10/16); medium methylation 56.3% (9/16); high methylation 58.3% (7/12) (Fig. 5a). Figure 5 depicts the results of several comparisons regarding the parameters we evaluated: clinical phenotype, D4Z4 methylation level and *SMCHD1* variants. Several genotype-phenotype intersections were assessed, some of which are not in complete agreement with the current indications for FSHD diagnosis (16,43): first we observed the full range of D4Z4 methylation levels in the 26 participants presenting a classical FSHD phenotype, of these 11 carried a *SMCHD1* variants; of the 11 participants presenting low 4q/10q D4Z4-l CpG methylation and carrying a *SMCHD1* variant 8 had a classical FSHD phenotype, but three presented incomplete or complex phenotypes; of the 5 carrying a *SMCHD1* variant with medium CpG methylation 3 had a classical FSHD phenotype and two presented incomplete or complex phenotypes. We also observed 5 samples with low CpG methylation with no *SMCHD1* variants and presenting classical (2), incomplete (1) or complex phenotypes (2) (Fig. 5a). In samples from healthy relatives with no *SMCHD1* variants we observed intermediate or high D4Z4 methylation. This picture shows that low D4Z4 CpG methylation level is associated with *SMCHD1* variants, whereas D4Z4 methylation levels do not reflect the clinical status of the subject.

The *SMCHD1* variants we identified were distributed all along the gene body and did not provide information on specific SMCHD1 region/domain that might account for the reduced methylation at 4q/10q-specific D4Z4 sequences (Fig. 5b). Nonetheless, we observed that variants positioned at the 3' end of the gene were associated to slightly higher D4Z4 methylation compared to the variants at 5' end. This could be explained by the presence of the major SMCHD1 functional/structural domains at the 5' end.

Altogether, these results suggest that SMCHD1 modulates, directly or indirectly, D4Z4 methylation at 4q and 10q, at the same time it seems that deleterious variants at *SMCHD1* are not sufficient for the pathogenesis of FSHD and cannot be considered faithful indicators of FSHD clinical status. Indeed, if applied to the trios/parent-child pairs included in this study (Fig. 5c), a molecular diagnostic test for FSHD2 based on *SMCHD1* inactivation and D4Z4 hypomethylation, would produce inconclusive results. According to our data only families 1 and 26 presented concordance between clinical phenotype, 4q/10q D4Z4 methylation profile and *SMCHD1* mutational status; while more complex profiles were identified in families 9, 11,

22 and 30; and none of the considered parameters are useful the genotype-phenotype correlation in family 8 and 27.

Conclusions

Reference genome assemblies provide the scaffold for the analysis and interpretation of large-scale assays based on massive parallel sequencing. Errors and/or inaccurate representation of sequence elements might pose a substantial hindrance for the analysis and interpretation of these data and lead to inaccurate conclusions. Thanks to recent breakthroughs in sequencing technologies, we can now appreciate the extent and variability of large repetitive genomic regions, which until recently were not adequately represented in reference human genome assemblies. In our work we demonstrated for the first time the value of the T2T and of PGR for investigating the genotype-phenotype correlation in a human genetic disease. Our work uncovered the complexity of the genomic setting of D4Z4-l opening to novel insights about the functional and architectural roles of repetitive elements. On this basis, we argue that the use of D4Z4 CpG methylation to diagnose FSHD must follow rigorous protocols to avoid biased interpretation. Molecular analysis should be considered together with precise clinical assessment and complemented by family studies for the proper interpretation of results. Moreover, the variability unveiled by our analysis should warn about the risk of relying on a single ideal reference genome sequence in FSHD. In fact, this simplification could strongly narrow our capacity of observing the variability ascribable to each single human genome producing biased interpretations of sequencing data. Finally, our study indicates that a linear approach to diagnose FSHD is obsolete and novel genomic approaches integrated with the precise phenotypic description of patients and their families are needed for the comprehension of the molecular mechanisms leading to muscle wasting in FSHD.

Methods

Participants

The study cohort included 26 subjects (9F, 17M) reported to the Miogen Laboratory of the University of Modena and Reggio Emilia for the diagnosis of FSHD (44). These cases carried

D4Z4 alleles with 10 or more repeats. In 8 cases the study was extended to one or both parents, for a total of 14 subjects (7F, 7M). Overall, 5 trios and 3 parent-child couples were analyzed. Four subjects carrying a 4U DRA were also included (2F, 2M). In three of these subjects (Family C individuals III-1 and III-3; Family 1252 subject 297/08 (24) D4Z4 methylation was previously assessed through the BSS assay using primers specific for the distal 4qA and 4qA-Long (4qA-L) D4Z4 repeat regions (24).

Clinical status was determined (Additional file 6: Table S6) according to the Comprehensive Clinical Evaluation Form (CCEF), which evaluates the distribution and degree of motor impairment (45). The CCEF has been developed by the Italian Clinical Network for FSHD to classify subjects on the basis of their clinical features. The CCEF classifies: 1) subjects presenting facial and scapular girdle muscle weakness typical of FSHD (category A, subcategories A1-A3), 2) subjects with muscle weakness limited to scapular girdle or facial muscles (category B subcategories B1, B2), 3) asymptomatic/healthy subjects (category C, subcategories C1, C2), 4) subjects with myopathic phenotype presenting clinical features not consistent with FSHD canonical phenotype (category D, subcategories D1, D2). Signed informed consent was obtained from all the subjects prior to the inclusion in the study.

Bisulfite sequencing:

Bisulfite sequence analysis was performed on high molecular weight genomic DNA obtained from peripheral frozen blood through phenol-chloroform extraction. To assess the methylation level at D4Z4 locus specific primers were designed on hg38 genome assembly using MethPrimer tool (46). No CpG sites were allowed in the primer sequence. The amplicon selection was based on following criteria: i) the sequences contain SNPs allowing to discriminate between the 4q and 10q D4Z4 repeats; and ii) include the maximum number of CpG sites possible. Amplicons were representative of the whole D4Z4 array and included functional domains such as the D4Z4 binding element (DBE) (47). Table S7 and S8, Fig. S5 (Additional files 6 and 2) summarize the characteristics of the four selected primers pairs. Specific Illumina adapters were added to each set of primers.

Bisulfite conversion was performed on 1 µg of genomic DNA by using the EpiTec Bisulfite Kit (Cat N°59104 QIAGEN) following the manufacturer's protocol.

PCR amplification of both converted and non-converted DNA was performed using the four selected primers (Additional file 6: Table S7). For every subject we generated two pools

of amplicons: bisulfite converted and non-converted DNA. Illumina paired-end sequencing was then performed by Eurofins genomics on the amplicon pools for a total of 36 Mbp (60K read pairs per amplicon with 2 x 300 bp read mode).

Bioinformatics analyses

Identification and characterization of D4Z4-like sequences and D4Z4 associated elements in human genome assemblies

The complete sequence of a reference D4Z4 element was obtained from the hg38 assembly in UCSC genome browser (https://genome.ucsc.edu/). Sequence similarity searches based on the BLAST program (48) were performed to annotate D4Z4-l elements and functional elements of the D4Z4 array (see below), in the hg38 reference assembly of the human genome [GRCh38.p14 - hg38 - Genome - Assembly - NCBI], the T2T assembly [T2T-CHM13v1.1 - Genome - Assembly - NCBI], and 86 distinct haplotype-level assemblies from the PGR, as available from https://github.com/human-pangenomics/HPP_Year1_Assemblies. Results of BLAST sequences similarity were stored in simple tabular format and processed by custom Perl scripts. A similarity threshold of 85% and an alignment length threshold of 300 aligned or more aligned residues were used to define D4Z4-l elements. Matches were binned in 5 bins according to their size (<500bp; <1000bp; <1500bp; <2500bp; and <3300bp). Matches of size in-between 3200 and 3300 bp were considered to provide a complete representation of a D4Z4 macrosatellite (full match).

D4Z4 arrays were annotated based on the terminal repeat for the presence of qA or qB sequence variants and presence/absence of the pLam sequence. qA/B has been annotated based on the sequence of the respective probes used for the FSHD diagnostic protocol; pLam has been annotated based on the sequence reported in UCSC genome browser (https://genome.ucsc.edu/) (Additional file 3).

Similarity-based clustering of D4Z4 like elements and terminal DUX4 coding sequences

D4Z4-l sequences of 3200 or more bp in size (full matches) were extracted from their respective genome assemblies and aligned with Kalign (49). Conserved alignment blocks

were extracted with Gblocks (50), using the following parameter: Minimum Length Of A Block: 3, Allowed Gap Positions: none. A sequence similarity matrix was computed by means of the EMBOSS distmat program (51), using the Tajima-Nei correction for multiple substitutions (52). Phenetic clustering was performed by using the NJ algorithm, as implemented by hclust() function from the cluster package in R (53). The same method was applied to cluster terminal DUX4 coding sequences by sequence similarity.

In-silico identification of primers target regions

Potential primers target regions in the hg38 and T2T genome assembly were determined by a custom Perl script. Sequence similarity searches of primer sequences on the hg38 and T2T reference assemblies were performed by blastn, with default parameters (48). All matches of 17bp or longer and with at most 1 mismatch and all perfect matches of 16bp or longer were recorded. Genomic coordinates were cross-referenced and potential target genomic regions were identified as those spanned by a 5' and 3' primer pairs, in the correct orientation, and within a distance >50 bp and <600 bp. The same procedure was applied for the identification of potential primer target regions in the BSS space, in this case both the genome and the primer sequences were converted *in-silico*.

Analysis of BSC reads

Quality control and determination of target regions

Reads' quality was assessed by Fastqc (54,55). Individual reports were merged by MultiQC (56) and quality metrics were visually inspected.

Non-converted amplicons' reads were aligned to T2T and hg38 human genome assemblies using Bowtie2 (57). Sample-level coverage profiles were computed by bedtools genomecov (58). Genomic regions covered by 100 or more reads in at least 50% of samples were considered for subsequent analyses. By this approach a total of 173 and 57 candidate target regions were identified in the T2T and hg38 assemblies, respectively.

Highly similar candidate regions were collapsed to provide a non-redundant representation of repetitive sequences and allow non-ambiguous mapping of BSC reads. A titration analysis was undertaken to identify the ideal sequence similarity threshold, which

was set at 95%. Target regions were collapsed into 60 and 9 distinct 95% sequence identity respectively for the hg38 and the T2T assembly. For every target region the consensus sequence was determined by majority rule consensus, by applying the cons EMBOSS program (51).

Assessment of methylation levels

BSC reads were aligned to consensus target region sequences and methylation levels were determined by Bismark (59) +Bowtie2. The bismark_methylation_extractor tool was applied to compute CpG methylation levels. Only CpGs covered by more than 10 reads in at least 25 patients were considered for the delineation of methylation profiles.

SMCHD1 variants analysis

FSHD2 probands are routinely tested for the presence of *SMCHD1* variants. We collected the identified *SMCHD1* variants in the cohort and performed areannotation on GRCh37 genome assembly using wAnnovar (60) (https://wannovar.wglab.org/; the analysis was performed on the 27.9.22) and filtered. We excluded intron variants and considered only exonic (missense, stop or frameshift variants) and canonical splicing variants with a GnomAD exome all MAF > 10^{-4} (GnomAD frequency for PM2 pathogenic moderate rule in Varsome) (61). Among the filtered variants we then excluded the ones which were benign and likely benign according to Varsome ACMG (American College of Medical Genetics and Genomics) prediction (31,62) (https://varsome.com/; the analysis was performed on 27.9.22). Finally, variant validation through Sanger sequencing was performed on probands and on relatives where present (Additional file 7: Table S9).

Declarations

Ethics approval and consent to participate: signed informed consent was obtained from all the subjects prior to the inclusion in the study. Ethics committee of Emilia Romagna Area Vasta Nord gave ethical approval for this work (743/2022/OSS/UNIMO SIRER ID 5111).
Consent for publication: Not applicable.

Availability of data and materials: data generated and/or analyzed during the current study and not included in this published article are available from the corresponding author on reasonable request.

The PGR dataset analyzed during the current study is available at https://github.com/human-pangenomics/HPP_Year1_Assemblies.

Competing interests: The authors declare that they have no competing interests.

Funding: FAR-FOMO 2021.

Authors' contributions: RGT, VS, MC conceived the study concept and supervised the project; VS, SP conducted the molecular analysis; MC conducted bioinformatics analysis; LR, SCP, MGD, CR, SB, LM, DL contributed to the patients clinical analysis and sample collection; FMS, GP contributed to results discussion; RGT, VS, MC, SP wrote the paper (original draft); all authors read and approved the final manuscript.

Acknowledgements: We are grateful to Francesca Brocco for contributing to the molecular analysis. In addition, we are indebted to all FSHD patients and their families for participating in this study.

Authors' information: Valentina Salsi and Matteo Chiara contributed equally to this study and reserve the right to list themselves as first author.

References

1. Deenen JCW, Arnts H, Van Der Maarel SM, Padberg GW, Verschuuren JJGM, Bakker E, et al. EPopulation-based incidence and prevalence of facioscapulohumeral dystrophy. Neurology. 2014;83(12):1056–9.
2. Mostacciuolo ML, Pastorello E, Vazza G, Miorin M, Angelini C, Tomelleri G, et al. Facioscapulohumeral muscular dystrophy: Epidemiological and molecular study in a north-east Italian population sample. Clin Genet. 2009;75(6):550–5.
3. Dumbovic G, Forcales S V., Perucho M. Emerging roles of macrosatellite repeats in genome organization and disease development. Epigenetics [Internet]. 2017;12(7):515–26. Available from: https://doi.org/10.1080/15592294.2017.1318235
4. Hewitt JE, Lyle R, Clark LN, Valleley EM, Wright TJ, Wijmenga C, et al. Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystropothhy. Hum Mol Genet. 1994;3(8):1287–95.
5. Winokur ST, Bengtsson U, Feddersen J, Mathews KD, Weiffenbach B, Bailey H, et al. The DNA rearrangement associated with facioscapulohumeral muscular dystrophy involves a heterochromatin-associated repetitive element: Implications for a role of chromatin structure in the pathogenesis of the disease. Chromosom Res. 1994;2(3):225–34.
6. Zhang XY, Loflin PT, Gehrke CW, Andrews PA, Ehrlich M. Hypermethylation of human DNA sequences in embryonal carcinoma cells and somatic tissues but not in sperm. Nucleic Acids Res. 1987;15(22):9429–49.

7. Agresti A, Meneveri R, Siccardi AG, Marozzi A, Corneo G, Gaudi S, et al. Linkage in human heterochromatin between highly divergent Sau3A repeats and a new family of repeated DNA sequences (HaeIII family). J Mol Biol. 1989;205(4):625–31.
8. Meneveri R, Agresti A, Marozzi A, Saccone S, Rocchi M, Archidiacono N, et al. Molecular organization and chromosomal location of human GC-rich heterochromatic blocks. Gene. 1993;123(2):227–34.
9. Deidda G, Cacurri S, Grisanti P, Vigneti E, Piazzo N, Felicetti L. Physical mapping evidence for a duplicated region on chromosome 10qter showing high homology with the facioscapulohumeral muscular dystrophy locus on chromosome 4qter. Eur J Hum Genet. 1995;3(3):155–67.
10. Deutekom JCT va., Wljmenga C, Tlenhoven EAE va., Gruter AM, Hewitt JE, Padberg GW, et al. FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. Hum Mol Genet. 1993;2(12):2037–42.
11. Nakagawa M, Matsuzaki T, Higuchi I, Fukunaga H, Inui T, Nagamitsu S, et al. Facioscapulohumeral Muscular Dystrophy: Clinical Diversity and Genetic Abnormalities in Japanese Patients. Intern Med. 1997;36(5):333–9.
12. Wohlgemuth M, Lemmers RJ, Van der Kooi EL, Van der Wielen MJ, Van Overveld PG, Dauwerse H, et al. Possible phenotypic dosage effect in patients compound heterozygous for FSHD-sized 4q35 alleles. Neurology. 2003;61(7):909–13.
13. Scionti I, Greco F, Ricci G, Govi M, Arashiro P, Vercelli L, et al. Large-scale population analysis challenges the current criteria for the molecular diagnosis of fascioscapulohumeral muscular dystrophy. Am J Hum Genet [Internet]. 2012;90(4):628–35. Available from: http://dx.doi.org/10.1016/j.ajhg.2012.02.019
14. Van Overveld PGM, Lemmers RJFL, Deidda G, Sandkuijl L, Padberg GW, Frants RR, et al. Interchromosomal repeat array interactions between chromosomes 4 and 10: A model for subtelomeric plasticity. Hum Mol Genet. 2000;9(19):2879–84.
15. Lemmers RJFL, Wohlgemuth M, Frants RR, Padberg GW, Morava E, Van Der Maarel SM. Contractions of D4Z4 on 4qB subtelomeres do not cause facioscapulohumeral muscular dystrophy. Am J Hum Genet. 2004;75(6):1124–30.
16. Lemmers RJLF, Wohlgemuth M, Van Der Gaag KJ, Van Der Vliet PJ, Van Teijlingen CMM, De Knijff P, et al. Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. Am J Hum Genet. 2007;81(5):884–94.
17. Lemmers RJLF, van der Vliet PJ, van der Gaag KJ, Zuniga S, Frants RR, de Knijff P, et al. Worldwide Population Analysis of the 4q and 10q Subtelomeres Identifies Only Four Discrete Interchromosomal Sequence Transfers in Human Evolution. Am J Hum Genet. 2010;86(3):364–77.
18. Van Geel M, Dickson MC, Beck AF, Bolland DJ, Frants RR, Van der Maarel SM, et al. Genomic analysis of human chromosome 10q and 4q telomeres suggests a common origin. Genomics. 2002;79(2):210–7.
19. Lemmers RJLF, de Kievit P, Sandkuijl L, Padberg GW, van Ommen GJB, Frants RR, et al. Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. Nat Genet. 2002;32(2):235–6.
20. Lemmers RJLF, Tawil R, Petek LM, Balog J, Block GJ, Santen GWE, et al. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. Nat Genet. 2012;44(12):1370–4.
21. Van Den Boogaard ML, Lemmers RJLF, Balog J, Wohlgemuth M, Auranen M, Mitsuhashi S, et al. Mutations in DNMT3B Modify Epigenetic Repression of the D4Z4

Repeat and the Penetrance of Facioscapulohumeral Dystrophy. Am J Hum Genet. 2016;98(5):1020–9.
22. Hamanaka K, Šikrová D, Mitsuhashi S, Masuda H, Sekiguchi Y, Sugiyama A, et al. Homozygous nonsense variant in LRIF1 associated with facioscapulohumeral muscular dystrophy. Neurology. 2020;94(23):E2441–7.
23. Gould T, Jones TI, Jones PL. Precise epigenetic analysis using targeted bisulfite genomic sequencing distinguishes fshd1, fshd2, and healthy subjects. Diagnostics. 2021;11(8):1–17.
24. Nikolic A, Jones TI, Govi M, Mele F, Maranda L, Sera F, et al. Interpretation of the epigenetic signature of facioscapulohumeral muscular dystrophy in light of genotype-phenotype studies. Int J Mol Sci. 2020;21(7).
25. Salsi V, Magdinier F, Tupler R. Does DNA methylation matter in FSHD? Genes (Basel). 2020;11(3).
26. Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze AM, Mitchell R. Vollger, Nicolas Altemose5, Lev Uralsky, Ariel Gershman SA, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge SEA, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas N-CC, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima PCD, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti RSF, et al. The complete sequence of a human genome. Science (80-). 2022;
27. Altemose N, Logsdon GA, Bzikadze A V., Sidhwani P, Langley SA, Caldas G V., et al. Complete genomic and epigenetic maps of human centromeres. Science (80-). 2022;376(6588).
28. Hori Y, Shimamoto A, Kobayashi T. The human ribosomal DNA array is composed of highly homogenized tandem clusters. 2021;1971–82.
29. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. Nature. 2022;604(7906):437–46.
30. de Greef JC, Lemmers RJLF, van Engelen BGM, Sacconi S, Venance SL, Frants RR, et al. Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. Hum Mutat. 2009;30(10):1449–59.
31. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K RH, Committee. ALQA. 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. 2015;17(5):405–24.
32. Delourme M, Charlene C, Gerard L, Ganne B, Perrin P, Vovan C, et al. Complex 4q35 and 10q26 Rearrangements. 2023;
33. Cacurri S, Piazzo N, Deidda G, Vigneti E, Galluzzi G, Colantoni L, et al. Sequence homology between 4qter and 10qter loci facilitates the instability of subtelomeric KpnI repeat units implicated in facioscapulohumeral muscular dystrophy. Am J Hum Genet. 1998;63(1):181–90.
34. Van Deutekom JCT, Bakker E, Lemmers RJLF, Van Der Wielen MJR, Bik E, Hofker MH, et al. Evidence for subtelomeric exchange of 3.3 kb tandemly repeated units between chromosomes 4q35 and 10q26: Implications for genetic counselling and etiology of FSHD1. Hum Mol Genet. 1996;5(12):1997–2003.
35. Rossella Tupler, Angela Berardinelli, Laura Barbierato, Rune Frants JEH, Giovanni Lanzi, Paola Maraschio LT. Monosomy of distal 4q does not cause

facioscapulohumeral muscular dystrophy. Med Genet. 1996;
36. Bakker E, Wijmenga C, Vossen RHAM, Padberg GW, Hewitt J, Wielen M Van Der, et al. The FSHD-linked locus D4F104S1 (p13E-11) ON 4q35 has a homologue on 10qter. 1995;11:39–44.
37. Gaillard MC, Roche S, Dion C, Tasmadjian A, Bouget G, Salort-Campana E, et al. Differential DNA methylation of the D4Z4 repeat in patients with FSHD and asymptomatic carriers. Neurology. 2014;83(8):733–42.
38. Lemmers RJLF, Goeman JJ, Van der Vliet PJ, Van Nieuwenhuizen MP, Balog J, Vos-Versteeg M, et al. Inter-individual differences in CpG methylation at D4Z4 correlate with clinical variability in FSHD1 and FSHD2. Hum Mol Genet. 2015;24(3):659–69.
39. Jones TI, King OD, Himeda CL, Homma S, Chen JCJ, Beermann M Lou, et al. Individual epigenetic status of the pathogenic D4Z4 macrosatellite correlates with disease in facioscapulohumeral muscular dystrophy. Clin Epigenetics [Internet]. 2015;7(1):1–22. Available from: ???
40. Calandra P, Cascino I, Lemmers RJLF, Galluzzi G, Teveroni E, Monforte M, et al. Allele-specific DNA hypomethylation characterises FSHD1 and FSHD2. J Med Genet. 2016;53(5):348–55.
41. Dion C, Roche S, Laberthonnière C, Broucqsault N, Mariot V, Xue S, et al. SMCHD1 is involved in de novo methylation of the DUX4-encoding D4Z4 macrosatellite. Nucleic Acids Res. 2019;47(6):2822–39.
42. Butterfield R. fascioscapulohumeral muscular dystrophy. 2023;
43. Tawil R, Kissel JT, Heatwole C, Pandya S, Gronseth G, Benatar M. Evidence-based guideline summary: Evaluation, diagnosis, and management of facioscapulohumeral muscular dystrophy. Neurology. 2015;85(4):357–64.
44. Bettio C, Salsi V, Orsini M, Calanchi E, Magnotta L, Gagliardelli L, et al. The Italian National Registry for FSHD: an enhanced data integration and an analytics framework towards Smart Health Care and Precision Medicine for a rare disease. Orphanet J Rare Dis. 2021;16(1):1–21.
45. Nikolic A, Ricci G, Sera F, Bucci E, Govi M, Mele F, et al. Clinical expression of facioscapulohumeral muscular dystrophy in carriers of 1-3 D4Z4 reduced alleles: Experience of the FSHD Italian National Registry. BMJ Open. 2016;6(1):1–10.
46. Li LC, Dahiya R. MethPrimer: Designing primers for methylation PCRs. Bioinformatics. 2002;18(11):1427–31.
47. Gabellini D, Green MR, Tupler R. Inappropriate gene activation in FSHD: A repressor complex binds a chromosomal repeat deleted in dystrophic muscle. Cell. 2002;110(3):339–48.
48. Altschup SF, Gish W, Pennsylvania T, Park U. Basic Local Alignment Search Tool 2Department of Computer Science. 1990;403–10.
49. Lassmann T. Kalign 3: Multiple sequence alignment of large datasets. Bioinformatics. 2020;36(6):1928–9.
50. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17(4):540–52.
51. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 2000;16(6):276–7.
52. Tajima F, Nei M. Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol. 1984;1(3):269–85.
53. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Internet]. Available from:

https://www.r-project.org/.
54. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Internet]. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
55. FastQC [Internet]. 2015. Available from: https://qubeshub.org/resources/fastqc
56. Ewels P, Lundin S, Max K. Data and text mining MultiQC : summarize analysis results for multiple tools and samples in a single report. 2016;32(June):3047–8.
57. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
58. Quinlan AR, Hall IM. BEDTools : a flexible suite of utilities for comparing genomic features. 2010;26(6):841–2.
59. Krueger F, Andrews SR. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27(11):1571–2.
60. Chang, X.; Wang K. wANNOVAR: annotating genetic variants for personal genomes via the web. J Hum Genet. 2008;6(9):2166–71.
61. Chen K, Hu J, Moore DL, Liu R, Kessans SA, Breslin K, et al. Genome-wide binding and mechanistic analyses of Smchd1-mediated epigenetic regulation. Proc Natl Acad Sci U S A. 2015;112(27):E3535–44.
62. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35(11):1978–80.

Figure legends:

Fig. 1 Stratification and variability of D4Z4-l elements**. a**. D4Z4-l elements in hg38 and T2T. Cumulative size per chromosome of D4Z4-l elements. Only chromosomes for which at least 1 Kbp of D4Z4-l sequence was identified, in any of the 2 assemblies, are reported. Size (in Kbp) is shown on the Y axis. Different colors are used for T2T and hg38 and to mark complete D4Z4-l (see legend). **b** Cumulative size (in Kbp) of D4Z4-l elements in the hg38, T2T and PGR haplotype level assemblies. left: cumulative size of full D4Z4-l elements. right: cumulative size of D4Z4-l elements. Data is displayed in the form of a barplot. Values are the Y axis, labels on the X axis. For the human pangenome haplotype level assemblies, 3 distinct bars are used to indicate the minimum, maximum and average values. Color codes according to the legend. **c** Dendrogram of complete D4Z4-l elements. Red rectangles and color codes are used to delineate the 5 main groups of D4Z4-l elements, as identified by sequence similarity-based clustering. **d** Total estimated number of complete D4Z4 elements per sequence similarity group. Distributions of values are displayed in the form of a boxplot, with groups indicated on the X axis and copy numbers on the Y axis. The top-right panel provides zoom on groups chr14, chr15-21, chr22-1 and chr22-2.

Fig. 2 Characterization of D4Z4 alleles. **a** Heatmap of terminal DUX4 ORFs. Scaled identity levels are displayed by the color gradient on the right. The dendrogram on the left shows clustering of terminal DUX4 ORFs based on sequence identity. Colored vertical bars are used to indicate: presence/absence of pLam; classification as qA or qB; sequence similarity cluster (see top). Color codes are explained directly under each bar. **b** Numerosity of sequence similarity clusters. The barplot shows (y-axis) the total number of DUX4 in each cluster (x-axis). **c** Boxplot of substitution rates. Substitution rates (substitutions by 100 bp) by cluster. Distribution of values are shown as a boxplot. Clusters are on the y-axis, values on the x-axis. **d** Distance of qA and qB elements from terminal DUX4 ORFs. Data are represented in the form of a histogram. Distances (in Kbp) are indicated on the x-axis. qA and qB elements are marked with different colors (see legend). **e-f** Annotation of terminal D4Z4 repeats. Two barplots are displayed for the 3 (out of 4) clusters to which 5 or more sequences have been assigned. Each barplot reports, for qA and qB alleles respectively: the total count (tot); the number pLam-like elements (pLam); the number pLam-like elements with a valid PAS (plamPolyA); the total number of haplotypes with 8 or less D4Z4 repeats. Values

are indicated on the y-axis, labels on the x-axis. Distinct sequence similarity clusters are indicated by a corresponding label directly under each plot. **g-h** total number of complete D4Z4 like elements in g) qA and qB haplotypes; h) haplotypes with (pLam+) and without (pLam-). Distribution of values are displayed in the form of a boxplot, with values on the X-axis and groups on the Y axis. Color codes are consistent throughout Fig. 2.

Fig. 3 CpG Methylation profiles of cohort subjects inferred on T2T. A total of 82 CpGs for which complete data are available for all the study subjects are shown. **a** heatmap of % CpG methylation: individuals are reported on the rows and CpGs on the columns. The dendrogram shows a clustering of the subjects based on the observed methylation profiles. Colored vertical bars are used to display CCEF grades, and the presence/absence of deleterious genomic variants in *SMCHD1*. Color codes are illustrated directly under each bar. The colored-bar at the top demarcates each of the 4 distinct 95% sequence identity genomic region clusters with complete data for all the subjects. **b** boxplots of methylation levels distributions by subject. The color-scale used for the heatmap and the boxplot is shown at the bottom. **c** violin-plot of average methylation levels in subjects with/without deleterious genetic variants in *SMCHD1*. Only CpGs in primer-targeted regions are considered.

Fig. 4 Methylation profiles of cohort subjects inferred on hg38. A total of 32 distinct CpGs, for which complete data are available for all the study subjects, are displayed. **a** heatmap of % CpG methylation: individuals are reported on the rows and CpGs on the columns. The dendrogram shows a clustering of the subjects based on the observed methylation profiles. Colored vertical bars are used to display CCEF grades, and the presence/absence of deleterious genomic variants in *SMCHD1*. Color codes are illustrated directly under each bar. The colored-bar at the top demarcates each of the 4 distinct 95% sequence identity genomic regions clusters with complete data for all the subjects. **b** boxplots of methylation levels distributions by subject. The color-scale used for the heatmap and the boxplot is shown at the bottom. **c** violin-plot of average methylation levels in subjects with/without deleterious genetic variants in *SMCHD1*. Only CpGs in primer-targeted regions are considered.

Fig. 5 Genotype-phenotype correlation. **a** Venn diagrams reporting the number of subjects presenting different combinations of 4q/10q specific D4Z4 methylation level, *SMCHD1* mutational status and FSHD clinical category. **b** ACMG classification, median 4q/10q-specific D4Z4 methylation of variant carriers, variant type and location are reported for each *SMCHD1*

variant. When the variant is carried by more than one subject, median D4Z4 methylation status of all the subjects is reported. **c** Distribution of 4q/10q specific D4Z4 methylation, *SMCHD1* mutational status and FSHD clinical category in the 8 trios/parent-child couples included in the study. Circles/squares are colored according to the median 4q/10q-specific D4Z4 methylation level; the color is fading as it represents the variability of CpG methylation pattern in each individual. WT= subject carrying no *SMCHD1* mutation, VAR= subject carrying a *SMCHD1* mutation.

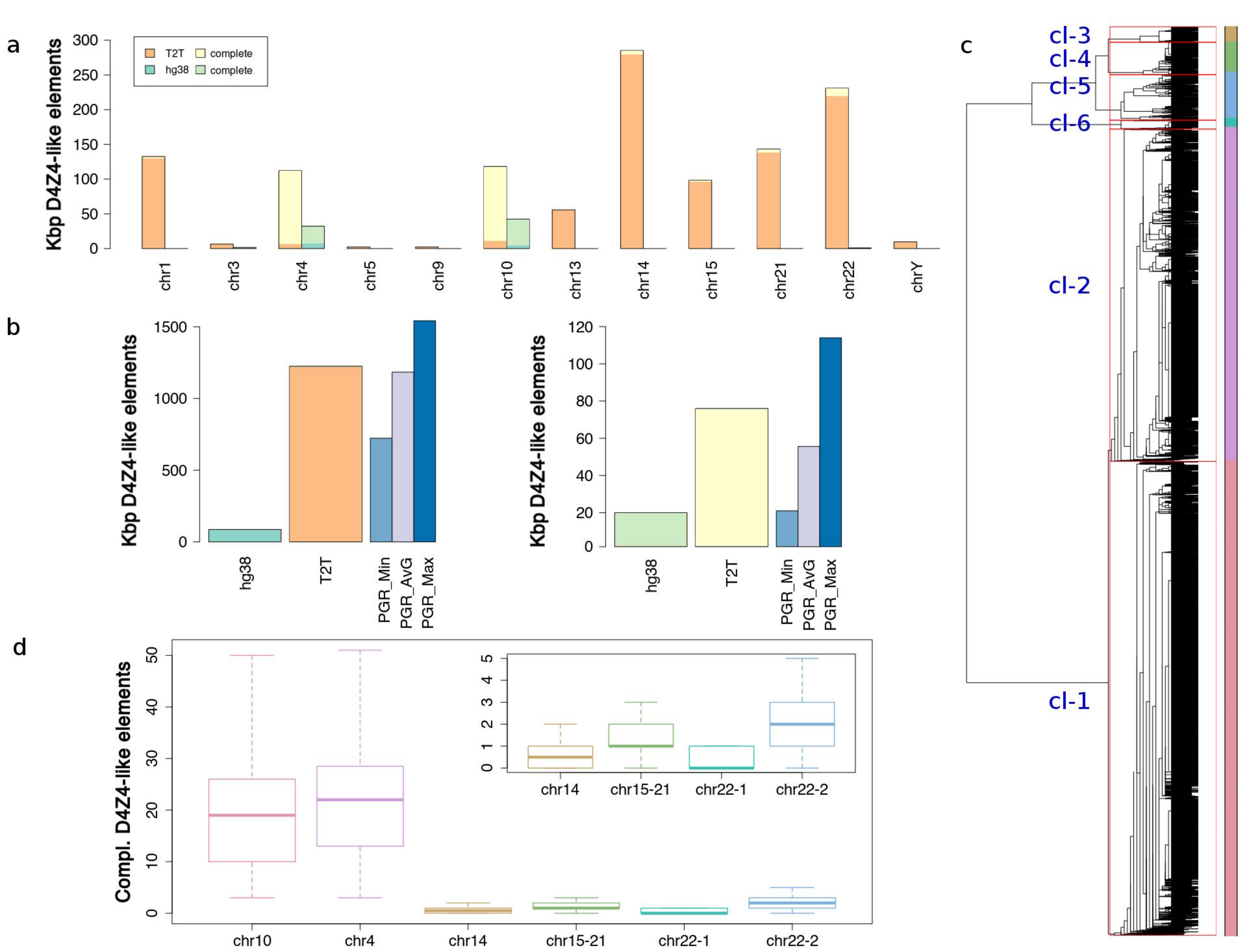
Additional files table of content				
File name		**File format**	**Title**	**Description**
Additional file 1	Table S1	.xlsx	**Number of D4Z4-like elements.**	Column A. assembly: genome assembly/id according to the human pangenome project. Column B-G. D4Z4-l elements were classified according to their size (bp): 350-500, 501-1000, 1001-1500, 1501-2500, 2501-3300, complete D4Z4 (3200-3300 bp). Column H. Total size (bp): total size of D4Z4-l elements. Column I. sex: sex of the subject. Column J-K. HapMap population: human geographic population according to HapMap; HapMap macro-population: human geographic macro-population according to HapMap
	Table S2		**Haplotype level classification of D4Z4 alleles.**	Column A. ID: unique identifier of the assembly. Column B-C. qType-qA/qB: classification and number of 4q35 subtelomeric haplotype. Column D. complete: complete set of alleles at D4Z4 repeat loci at chr 4 and chr 10. Column E. groups: group/cluster assigned to the terminal DUX4 sequence. Column F. Nrepeat: number of D4Z4 repeats at 4q35. Haplotypes with less than 8 repeats on chr4 are in bold. Column G. pLam: presence(1)/absence(0) of pLam. Column H. validPAS: number of valid PAS sequences. Column I. Scaffold/chromosome: scaffold/chromosome of the terminal repeat. Column J. strand: genome strand +=forward, -=reverse. Column K-L. dux4S: start genomic coordinate of the terminal dux CDS; dux4E: end genomic coordinate of the terminal DUX4 CDS. Column M-N. plamS: start genomic coordinate of plam; plamE: end genomic coordinate of plam. Column O-P. qS: start genomic coordinate of qA/qB; qE: end genomic coordinate of qA/qB. Column Q. DistQ(qA/qB): distance (bpl of qA/qB elements from the terminal DUX4 CDS.

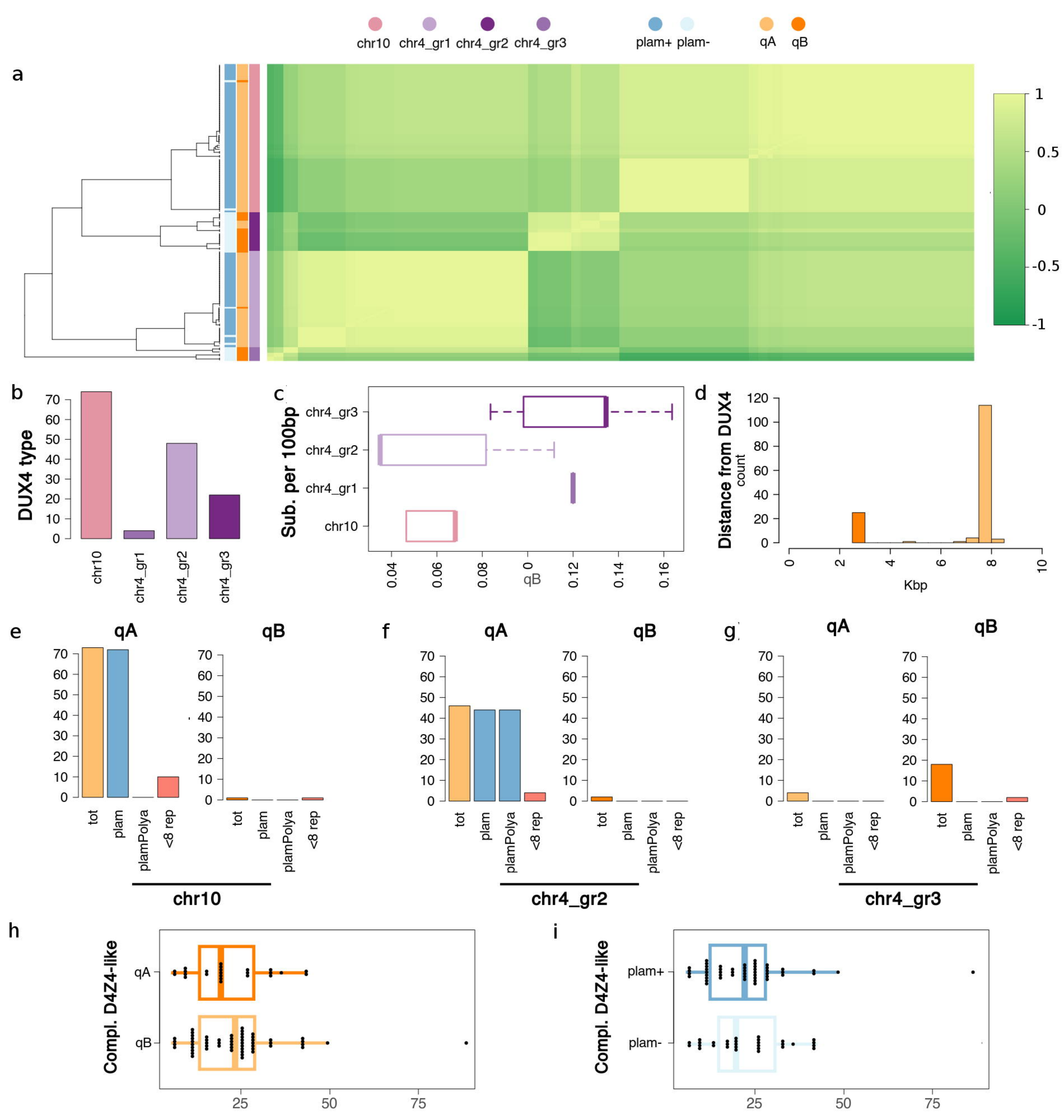
Additional file 2	Fig.S1	.pdf	**4q35 haplotypes**	Common D4Z4 haplotypes include a variable number of D4Z4 tandemly arrayed elements including a DUX4 ORF (exons 1 and 2) containing a double homeobox and the heterochromatic elements Lsau and hhsmp3. 4q alleles are usually associated to either qA or qB haplotype while 10q alleles are only associated to qB haplotype. qA haplotype includes the pLam sequence carrying an AUUAAA functional polyadenylation signal on chromosome 4q and an AUCAAA non-functional polyadenylation signal on chromosome 10q.
	Fig. S2		**Estimated copy number of D4Z4 repeats human geographic macro-populations.**	Distributions of values are represented in the form of a boxplot. Human geographic macro-populations are indicated on the Y axis.
	Fig.S3		**Estimated nucleotide substitution rates in D4Z4 repeats.**	Substitutions rates (substitutions per 100bp) are shown, cumulatively, for repeats assigned to chr10 and chr4. Boxplots are used to display the distribution of values. Chromosomes are indicated on the Y axis.
	Fig.S4		**Estimated nucleotide substitution rates in D4Z4 repeats in haplotypes level assemblies.**	Substitutions rates (substitutions per 100bp) are shown, for repeats assigned to chr10 and chr4, for every haplotype level assembly. Boxplots are used to display the distribution of values. Individuals are reported on the x-axis, substitution rates on the y-axis. **a** chr4. **b** chr10.

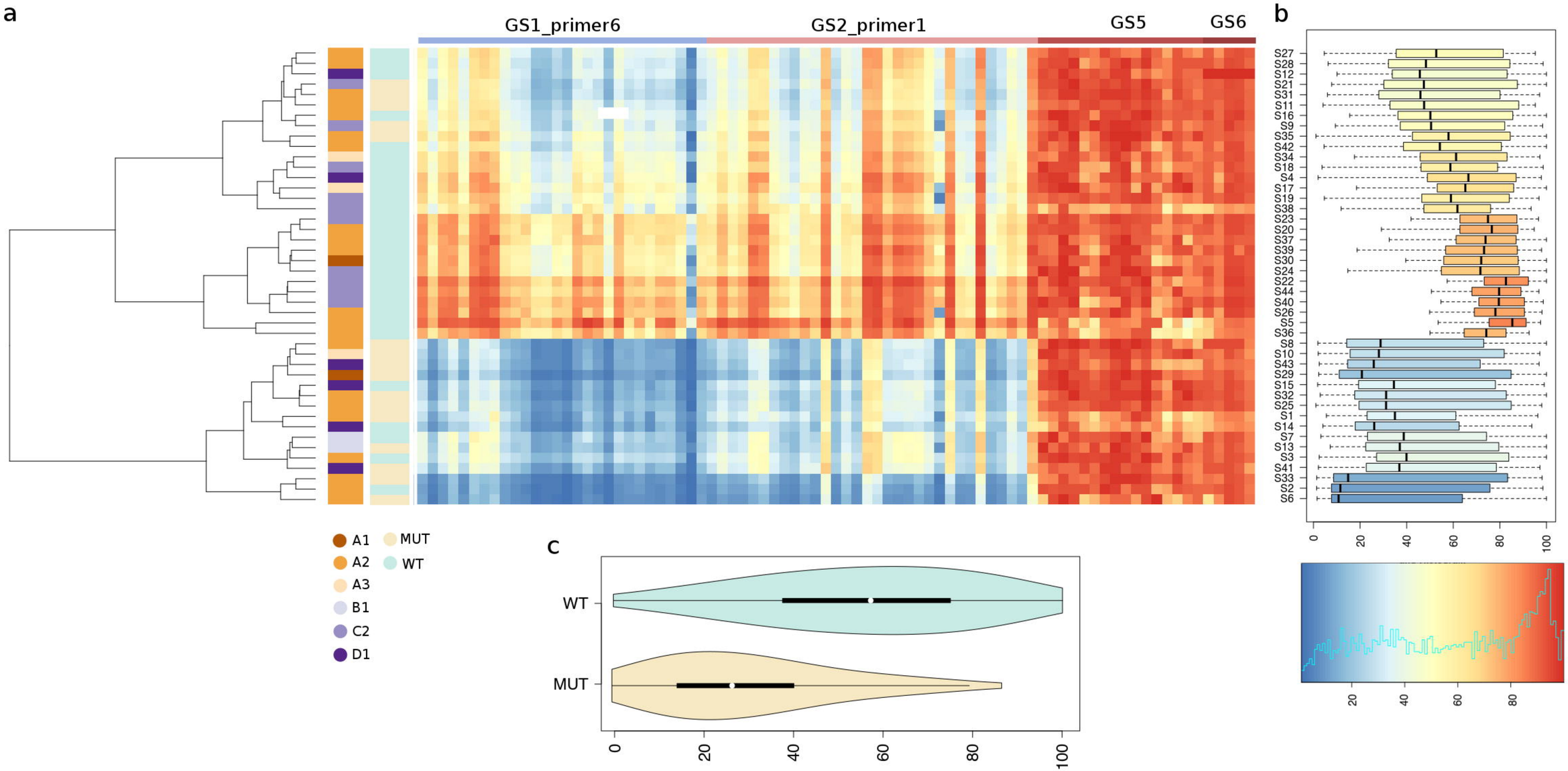
	Fig.S5		**Amplicon locations at D4Z4 repeats.**	The D4Z4 array at 4q35 is reported: arrows represent the 3.3 Kb repeats, the p13E-11 probe annealing site is shown in yellow, the pLAM sequence is indicated with the green box. An expanded representation on the last repeat along with the pLam is represented at the bottom. Typical heterochromatic domains LSau and hhspm3 are reported in magenta and brown; DUX4 exons are reported in green. The position of the amplicons generated from the designed primer sets are reported (coordinates are relative to the 3.3 kb KpnI to KpnI repeat). Primer 1 and 6 were designed on the reverse strand, while primer 3 and 5 on the forward strand.
	Fig.S6		**Cohort subjects' methylation profiles, T2T.**	**a.** Heatmap, with individuals reported on the rows, and CpGs on the columns. A total of 938 distinct CpGs are shown. Missing data are indicated in white. The dendrogram shows a clustering of the subjects based on the observed methylation profile. Colored vertical bars are used to display the stratification of the individuals according to CCEF grades, and the presence/absence of deleterious genomic variants in *SMCHD1*. Color code legends are displayed directly under each bar. The colored-bar at the top demarcates distinct 95% identity regions. **b.** Boxplots display average methylation levels of the 938 CpGs for every subject. **c.** Violin-plot of average methylation levels, in subjects with/without deleterious genetic variants in *SMCHD1*.
	Fig.S7		**Cohort subjects' methylation profiles, hg38.**	**a.** Heatmap, with individuals reported on the rows, and CpGs on the columns. A total of 125 distinct CpGs are shown. Missing data are indicated in white. The dendrogram shows a clustering of the subjects based on the observed methylation profile. Colored vertical bars are used to display the stratification of the individuals according to CCEF grades, and the presence/absence of deleterious genomic variants in *SMCHD1*. Color code legends are displayed directly under each bar. The colored-bar at the top demarcates distinct 95% identity regions. **b.** Boxplots display average methylation levels of the 125 CpGs for every subject. **c.** Violin-plot of average methylation levels, in subjects with/without deleterious genetic variants in *SMCHD*.
	Fig. S8		**Average methylation level at D1 primer on T2T and hg38.**	The distributions of values are displayed in the form of a boxplot. Genome assemblies are indicated on the Y axis, average methylation levels on the X-axis.

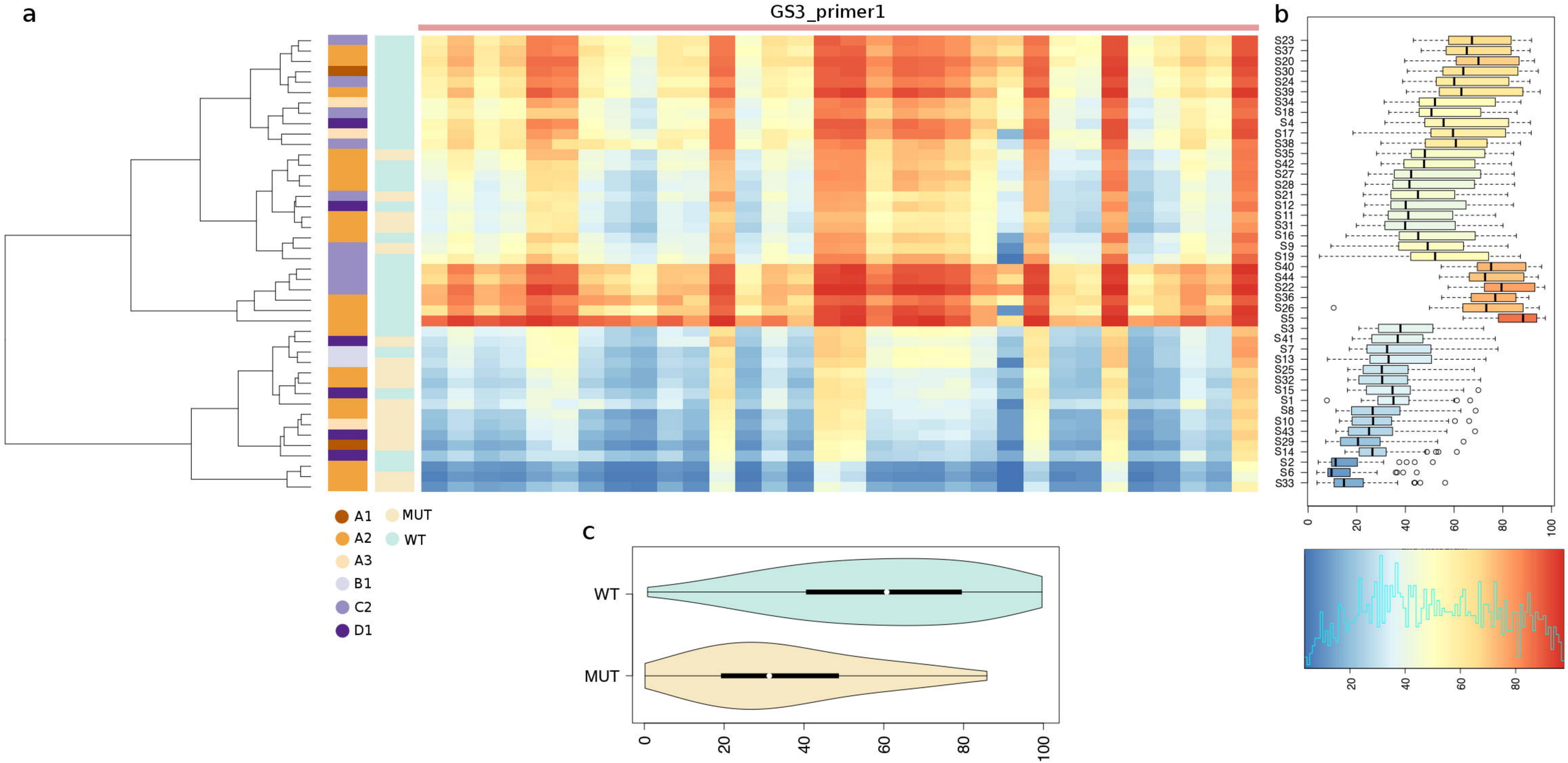
	Fig.S9		**Average methylation level at primers targeted CpGs, according to CCEF grade.**	**a**. T2T reference assembly. **b**. hg38 reference assembly. Boxplots are used to display the distribution of values. CCEF grades are reported on the Y axis, average methylation levels on the X-axis.
	Fig.S10		**Raincloud plot of methylation distribution per group of sequences (GS).**	Boxplots and histograms show D4Z4 CpGs methylation distribution for GS1 (primer D6), GS2 (primer D1), GS5 and GS6 clusters. Dots on the left of each plot represent cytosines from every single subject and are colored according to the *SMCHD1* mutational status.
	Fig.S11		**Raincloud plot of CpG methylation at all group of sequences (GS).**	Boxplots and probability density plots show D4Z4 CpGs methylation distribution for all GSs.
Additional file 3		.docx	**Reference sequences**	FASTA sequences used for characterization of D4Z4 repeat array haplotypes.
Additional file 4	Table S3	.docx	**Primers target region.**	chr**:** chromosome. T2T_notConverted: total number of genomic regions targeted by non-converted primers in the T2T genome assembly. T2T_converted: total number of genomic regions targeted by BSC primers in the T2T genome assembly. hg38_notConverted: total number of genomic regions targeted by non-converted primers in the hg38 genome assembly. hg38_converted: total number of regions by BSC primers in the hg38 genome assembly
Additional file 5	Table S4	.xlsx	**List of genomic regions, T2T.**	Regions covered by at least 100 non-converted reads in at least 50 patients in the T2T reference genome assembly. Column A. genomic coordinates: genomic coordinates in chr:start-end format. Column B. macro-region: assignment to 95% sequence identity clusters.
	Table S5		**List of genomic**	Regions covered by at least 100 non-converted reads in at least 50 patients in the hg38 reference genome

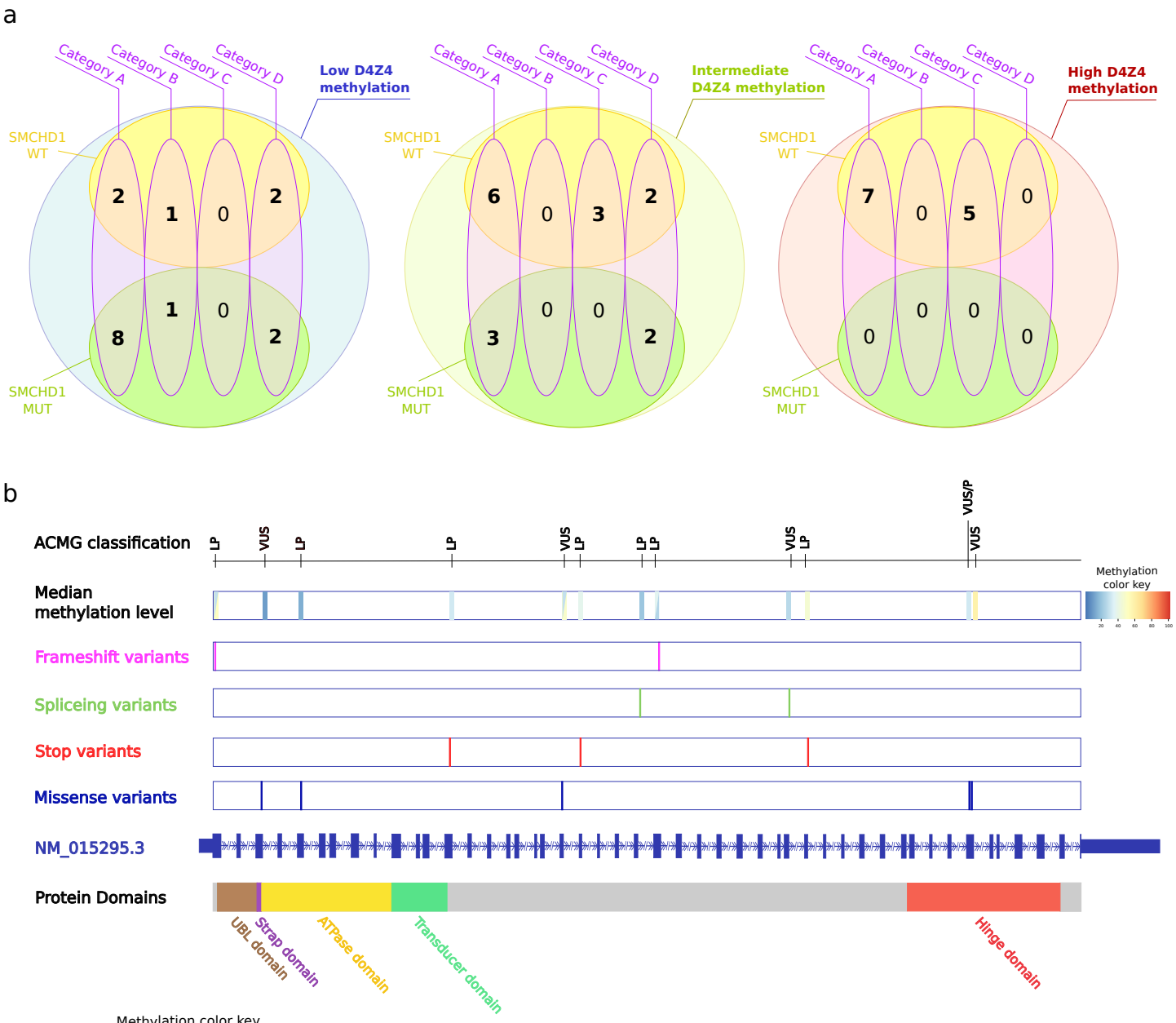
			regions, hg38.	assembly. Column A. genomic coordinates: genomic coordinates in chr:start-end format. Column B. macro-region: assignment to 95% sequence identity clusters.
Additional file 6	Table S6	.docx	**Distribution of subjects according to sex and FSHD category.**	
	Table S7		**Forward (FW) and reverse (RW) sequence of primers.**	D= non-converted sequence; BSS-D= converted sequence.
	Table S8		**Amplicon size and number of analyzed CpGs.**	
Additional file 7	Table S9	.xlsx	**Summary of clinical and molecular features of the cohort.**	*A1-3= classic FSHD phenotype; B1= scapular girdle weakness; D1= complex phenotype; C2= healthy. **LP=likely pathogenic; VUS= variants of uncertain significance; VUS/P= variant of uncertain significance with some pathogenic evidence; FS= frameshift variant; MISS= missense variant; SPL= splicing variant.


a
Kbp D4Z4-like elements
T2T
complete
hg38
complete
chr1
chr3
chr4
chr5
chr9
chr10
chr13
chr14
chr15
chr21
chr22
chrY
b
Kbp D4Z4-like elements
hg38
T2T
PGR_Min
PGR_AvG
PGR_Max
c
cl-3
cl-4
cl-5
cl-6
cl-2
cl-1
d
Compl. D4Z4-like elements
chr10
chr4
chr14
chr15-21
chr22-1
chr22-2


chr10
chr4_gr1
chr4_gr2
chr4_gr3
plam+
plam-
qA
qB
a
1
0.5
0
-0.5
-1
b
DUX4 type
70
60
50
40
30
20
10
0
chr10
chr4_gr1
chr4_gr2
chr4_gr3
c
Sub. per 100bp
chr4_gr3
chr4_gr2
chr4_gr1
chr10
0.04
0.06
0.08
0
0.12
0.14
0.16
qB
d
Distance from DUX4
count
120
100
80
60
40
20
0
0
2
4
6
8
10
Kbp
e
qA
qB
tot
plam
plamPolya
<8 rep
chr10
f
qA
qB
chr4_gr2
g
qA
qB
chr4_gr3
h
Compl. D4Z4-like
qA
qB
25
50
75
i
Compl. D4Z4-like
plam+
plam-
25
50
75


a
GS1_primer6
GS2_primer1
GS5
GS6
A1
A2
A3
B1
C2
D1
MUT
WT
b
S27
S28
S12
S21
S31
S11
S16
S9
S35
S42
S34
S18
S4
S17
S19
S38
S23
S20
S37
S39
S30
S24
S22
S44
S40
S26
S5
S36
S8
S10
S43
S29
S15
S32
S25
S1
S14
S7
S13
S3
S41
S33
S2
S6
0
20
40
60
80
100
c
WT
MUT
0
20
40
60
80
100


a
GS3_primer1
A1
A2
A3
B1
C2
D1
MUT
WT
b
S23
S37
S20
S30
S24
S39
S34
S18
S4
S17
S38
S35
S42
S27
S28
S21
S12
S11
S31
S16
S9
S19
S40
S44
S22
S36
S26
S5
S3
S41
S7
S13
S25
S32
S15
S1
S8
S10
S43
S29
S14
S2
S6
S33
0
20
40
60
80
100
c
WT
MUT
0
20
40
60
80
100


a
Category A
Category B
Category C
Category D
Low D4Z4 methylation
Intermediate D4Z4 methylation
High D4Z4 methylation
SMCHD1 WT
SMCHD1 MUT
b
ACMG classification
LP
VUS
VUS/P
Median methylation level
Methylation color key
Frameshift variants
Spliceing variants
Stop variants
Missense variants
NM_015295.3
Protein Domains
UBL domain
Strap domain
ATPase domain
Transducer domain
Hinge domain

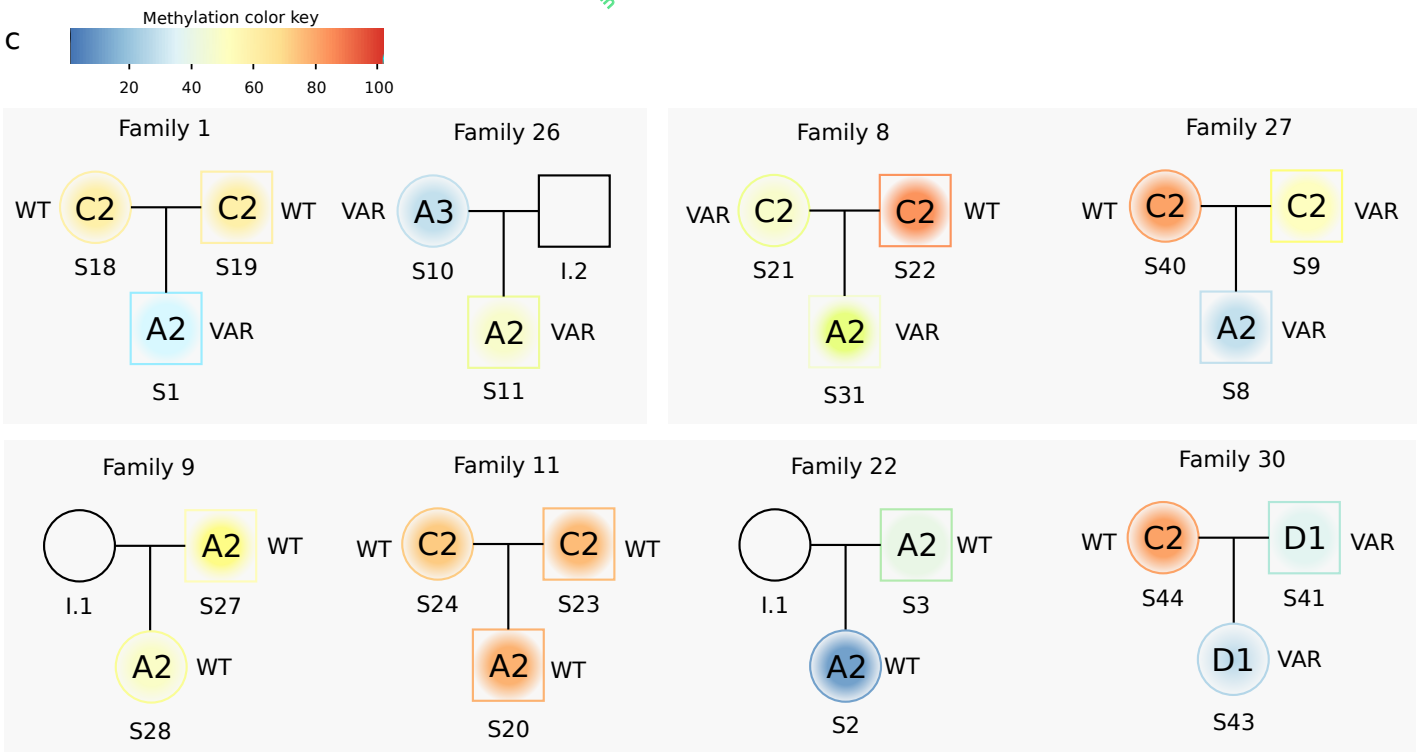
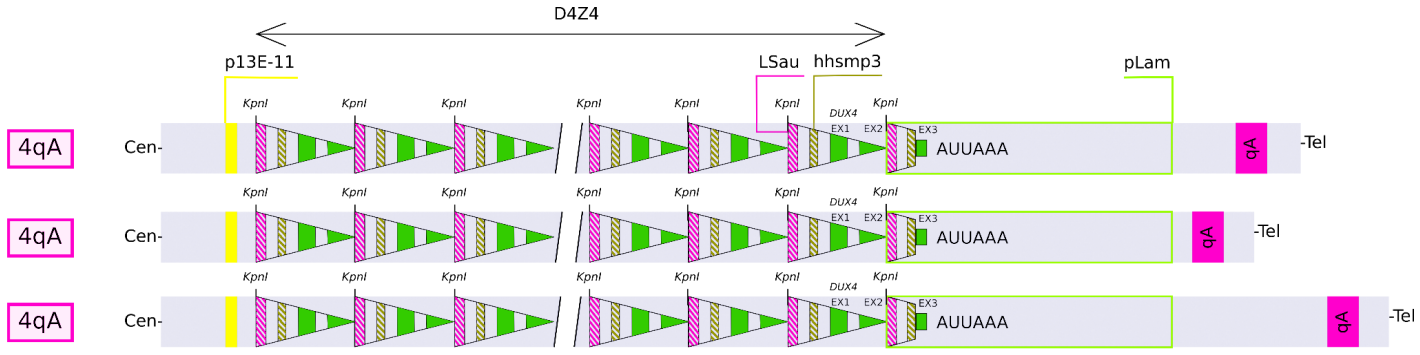
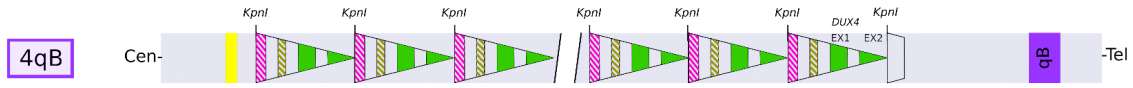
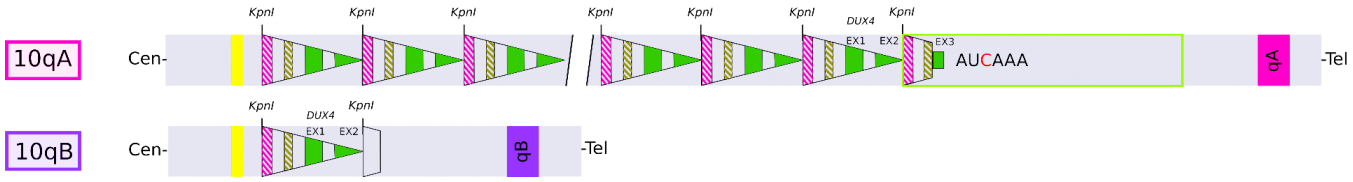

c
Methylation color key
20
40
60
80
100
Family 1
Family 26
Family 8
Family 27
Family 9
Family 11
Family 22
Family 30
WT
VAR
C2
A3
A2
D1
S18
S19
S1
S10
I.2
S11
S21
S22
S31
S40
S9
S8
I.1
S27
S28
S24
S23
S20
S3
S2
S44
S41
S43

Table 4- Comparison between PGR analysis results and previous knowledge.

Haplotype	PGR analysis	Previous knowledge
	- 6 to 89 RU alleles; - The majority of alleles classified as qA are associated with a pLam both on 4q and 10q; - AUUAAA PAS at 4q - 4.6% 4qA-PAS alleles with <8RU; - qA sequence is not always at the same distance from the last D4Z4 repeat (from 4.7 Kb to 8.3 Kb).	- 6 to ~90 RU alleles (14); - qA alleles are associated with a pLam both on 4q and 10q (18); - AUUAAA PAS at 4q (17); - 3% healthy population carry a DRA and 1.3% a DRA with permissive haplotype (13).
	- qB alleles do not carry a pLam sequence; - qB sequence is more proximal than qA (2.5 Kb and 7.3 Kb respectively); - No differences in D4Z4 repeat numbers between 4qA and 4qB alleles.	- qB alleles do not carry a pLam sequence (18); - qB sequence is more proximal than qA (19); - Observed equal frequency of 4qA and 4qB alleles (19).
	- AUCAAA PAS at 10q; - One 10qB allele was identified.	- AUCAAA PAS at 10q (17); - 10q is associated predominantly to qA haplotype (19); - sporadic cases f 10qB alleles (32).

Haplotype: graphical representation of 4q/10q D4Z4 elements; PGR analysis: results obtained by the annotation of D4Z4 elements on PGR; Previous knowledge: D4Z4 elements characterization according to literature.