A human pan-genomic analysis reconfigures 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*1]

Authors' Affiliations:

1. Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Modena, 41125, Italy. valentina.salsi@unimore.it; pini.saraps@gmail.com; rossella.tupler@unimore.it*.

2. Dipartimento di Bioscienze, Università degli Studi di Milano, Milan, 20133, Italy. matteo.chiara@unimi.it.

3. Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, Consiglio Nazionale delle Ricerche, Bari, 70126, Italy. matteo.chiara@unimi.it; graziano.pesole@uniba.it.

4. Department of Systems Biology and Engineering, Silesian University of Technology, Gliwice, 44-100, Poland. pawel.s.kus@protonmail.com; kimmel@rice.edu.

5. Department of Neurosciences, Reproductive and Odontostomatological Sciences, University Federico II of Naples, Naples, 80131, Italy. ruggilucia@gmail.com.

6. IRCCS Foundation, C. Besta Neurological Institute, Milan, 20133, Italy. Silvia.Bonanno@istituto-besta.it; Lorenzo.Maggi@istituto-besta.it.

7. Department of Clinical and Experimental Medicine, University of Messina, Messina, 98125 Italy. carmelo.rodolico@unime.it.

8. INSPE and Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milan, 20133, Italy. previtali.stefano@hsr.it.

9. Department of Neurorehabilitation, IRCCS Eugenio Medea, Bosisio Parini, 23842, Italy. grazia.dangelo@lanostrafamiglia.it.

10. Department of Medicine, Surgery and Neurosciences, University of Siena, Siena, 53100 Italy. diego.lopergolo@unifi.it.

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

12. Departments of Statistics and Bioengineering, Rice University, Houston, TX, 77005, United States. kimmel@rice.edu.

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

14. Department of Biosciences, Biotechnology and Environment, University of Bari "A. Moro", Bari, 70126, Italy. graziano.pesole@uniba.it.

15. Department of Molecular, Cell, and Cancer Biology, University of Massachusetts Medical School, Worcester, MA 01605, United States. rossella.tupler@unimore.it*.

16. Li Weibo Institute for Rare Diseases Research at the University of Massachusetts Medical School, Worcester, MA 01605, United States. rossella.tupler@unimore.it *.

† Valentina Salsi and Matteo Chiara contributed equally to this study and reserve the right to list themselves as first author.

***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

Facioscapulohumeral muscular dystrophy (FSHD) is the only human disease associated with epigenetic changes at a macrosatellite array. Almost 95% of FSHD cases carry a reduced number (≤10) of tandem 3.3 kilobase repeats, termed D4Z4, on chromosome 4q35; remaining cases bear variants in chromatin remodeling factors, such as SMCHD1, DNMT3B, LRIF1. Reduced CpG methylation is used for the molecular diagnosis of FSHD, but D4Z4-like sequences dispersed in the genome can generate ambiguous results. By analyzing complete haplotype level assemblies from the T2T-CHM13 human genome and 86 haploid genomes from the human pangenome project, we uncovered the extensive number of D4Z4-like elements and their widespread inter- and intra-individual variability. An original analytical approach was developed to elucidate this previously unaccounted wealth of D4Z4-like elements and to analyze CpG methylation at D4Z4 in bisulfite-treated DNA from 29 FSHD index cases and 15 relatives. Integrated analysis of clinical phenotype, D4Z4 CpG methylation level and gene variants showed that low D4Z4 methylation was associated with variants in the *SMCHD1* gene, but not with the patients' clinical phenotypes. This is the first study showing the relevance of the pangenome and T2T-CHM13 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 inconsistent association of phenotypes with methylation profiles advocate for a critical revision of FSHD diagnostic tests based on D4Z4 CpG methylation assays and indicate that molecular investigations must be complemented by family studies for the proper interpretation of results.

Introduction

Facioscapulohumeral muscular dystrophy (FSHD) is the third most common form of hereditary myopathy with an estimated prevalence from 1/8,000 to 1/20,000[1,2]. FSHD is the only hereditary human disease associated with reductions in copy number of a 3.3 kb macrosatellite[3], termed D4Z4. In the general population, the size of the D4Z4 repeat varies between 11 and 150 units, whereas FSHD patients carry fewer than 11 repeats[4]. Each copy of the 3.3 kb D4Z4 repeat contains an open reading frame (ORF) with two homeobox domains, named *DUX4*-Like (*DUXL*)[5], and two different classes of GC-rich repetitive DNA[6,7]: hhspm3, a member of a low copy human repeat family[8], and LSau, a middle repetitive DNA family[8,9]. A tandemly arrayed D4Z4 repeat with 98% identity to the 4q35 array is located at 10q26[10], but only alleles with reduced D4Z4 copy number on chromosome 4q35 have been associated with FSHD (Figure 1).

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)[11] is thought to determine epigenetic alterations and the inappropriate expression of nearby genes leading to disease [12–14]. In FSHD2 [OMIM #158901] affected individuals carry two D4Z4 arrays in the healthy range (>10 RU), but bear damaging heterozygous variants in chromatin remodeling factor genes, such as *SMCHD1* (for structural maintenance of chromosomes flexible hinge domain containing 1)[15], *DNMT3B* (for methyltransferase 3B)[16] and *LRIF1* (for ligand-dependent nuclear receptor interacting factor 1)[17]. It has thus 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 (Figure 1). Recent studies linked the molecular etiology of FSHD with the inappropriate expression of the DUX4 protein from the terminal D4Z4 repeat. Only a specific haplotype labelled 4qA-PAS is considered to be permissive for DUX4 protein expression, due to the

presence of a 260-bp sequence, termed pLAM, which provides a 3'-terminal exon (Exon 3) and a polyadenylation signal (PAS) (AUUAAA) in juxtaposition to the terminal copy of *DUX4* ORF (hereafter ter-*DUX4*) on 4q[18,19] (Figure S1). Although a similar arrangement is observed also on chr 10q26, at this locus the PAS signal is disrupted by a single nucleotide variant (AUCAAA), preventing the polyadenylation of the ter-*DUX4* mRNA. By analogy, 4qB, an alternative (to 4qA-PAS) haplotype at 4q (Figure S1), is considered not permissive for the expression of the DUX4 protein due to the lack of pLam. Based on these observations (D4Z4 reduced copy number or heterozygous variants in chromatin remodelers, associated with 4qA-PAS haplotype) a reduction of CpG methylation of D4Z4 has been proposed as a faithful indicator of anomalous *DUX4* gene expression and it is used as a proxy of disease status in the clinical setting[20].

This model is in contrast with clinical and epidemiological data showing reduced penetrance and clinical variability in FSHD families[21–27], as well as with observations obtained by adopting an articulate neurological assessment for the study of genotype-phenotype correlation in FSHD. This latter approach includes the phenotypic classification of probands and relatives by applying the Comprehensive Clinical Evaluation Form (CCEF)[28], which beside quantifying the degree of motor disability[29], classifies individuals on the basis of detailed phenotypic features which include parameters such as age at onset or site of disease onset beside the detailed description of muscle impairment. Four main descriptive categories have been created. They identify: (1) subjects presenting facial and scapular girdle muscle weakness typical of FSHD (category A, subcategories A1-A3), (2) subjects presenting an incomplete phenotype with muscle weakness limited to scapular girdle or facial muscles (category B, subcategories B1 and B2), (3) asymptomatic or healthy subjects (category C, subcategories C1 and C2), and (4) subjects with myopathic phenotypes presenting clinical features not consistent with FSHD canonical phenotype (category D, subcategories D1 and D2)[28]. The application of this methodology for the

stratification of clinical phenotypes has permitted the quantification of the large phenotypic variability observed in individuals carrying a D4Z4 Reduced Allele (DRA) which is in contrast to the indication that a positive molecular test is the only determining aspect for FSHD diagnosis[30–37]. Studies also showed that the different categories are associated with different disease trajectories; in particular subjects assessed with a classical FSHD phenotype (Category A) display a steeper functional decline, than subjects with limited muscle impairment (Category B) [31,35]. Large genotype-phenotype studies also highlighted differences between probands and relatives, including 30-50% of relatives remaining non-penetrant carriers. Consistent with these findings it has been assessed that DRA with 4-8 RU have the frequency of a common polymorphism (3% in the general population[25,38–40]), which is not compatible with the incidence of FSHD; moreover the DRA 4qA-PAS haplotype has a 2% frequency in the human population[39]. Finally, studies reported a wide variability in D4Z4 CpG methylation levels among FSHD index cases and FSHD families with no clear association between D4Z4 methylation status and disease manifestations or severity[32,41].

Here, we report the results of comparative analyses of 86 haplotype-level assemblies from the Human Pangenome Project dataset (hereafter PGR)[42] and the outcome of high-throughput CpG methylation assay of the D4Z4 repeat based on the T2T-CHM13 human genome assembly (hereafter T2T). Our study displays all the potential of complete, high-quality haplotype level genome assemblies for the analysis of D4Z4 arrays in human diseases, confirms the high frequency of the permissive 4qA-PAS haplotype, which is comparable to a common polymorphism and shows that across the 3.3 kb D4Z4 elements only two sequences are 4q/10q-specific and amenable for the study of methylation patterns in FSHD. This refined CpG methylation assay reveals that 4q/10q-specific reduced methylation correlates with the presence of damaging *SMCHD1* heterozygous variants, but not with the clinical status of FSHD2 cases and advocate revisions of previous findings based on the GRCh38 reference (hereafter hg38).

Subjects and methods

Participants

The study cohort included 26 subjects (8F, 18M) reported to the Miogen Laboratory of the University of Modena and Reggio Emilia for the diagnosis of FSHD[43]. 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[32]) D4Z4 methylation was previously assessed through the BSS assay using primer sets specific for the distal 4qA and 4qA-Long (4qA-L) D4Z4 repeat regions[32].

The clinical phenotype was determined (Table S1) according to the Comprehensive Clinical Evaluation Form (CCEF), which evaluates the distribution and degree of motor impairment, age at onset of motor impairment and site of muscle weakness at onset, the presence of typical and atypical symptoms[24]. The CCEF has been developed by the Italian Clinical Network for FSHD to classify subjects on the basis of these 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). Degree of motor impairment was also established by applying a standardized evaluation that generates the FSHD score [29].

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 primer sets were designed on hg38 genome assembly using MethPrimer tool[44]. 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)[12]. Table S2 and Table S3, Figure S2 summarize the characteristics of the four selected primer sets. 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 primer pairs (Table S2). For every subject we generated two pools of amplicons: one from bisulfite converted (BSC), one from non-converted DNA. Illumina paired-end sequencing was then performed by Eurofins genomics on the amplicon pools for a total of 36 Mb (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[45] were performed to annotate D4Z4-like 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 searches 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-like 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; presence/absence of the pLam sequence and integrity of the AUUAAA polyadenylation signal (PAS). 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/) (Supplemental Material and Methods).

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

D4Z4-like sequences of 3200 or more bp in size (full matches) were extracted from their respective genome assemblies and aligned with Kalign[46]. Conserved alignment blocks were extracted with Gblocks[47], 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[48], using the Tajima-Nei correction for multiple substitutions[49]. Phenetic clustering was performed by using the NJ algorithm, as implemented by hclust()

function from the cluster package in R[50]. Six distinct groups with high sequence identity were defined for the D4Z4 complete elements. D4Z4 from T2T were used to anchor/assign each group to a chromosome.

The same method was applied to cluster ter-*DUX4* coding sequences by sequence similarity. Terminal repeats were identified based on the presence of a qA or qB element *in cis* and within 15 Kb from a complete *DUX4* ORF and tentatively assigned to chr 4 or 10 based on sequence similarity.

In-silico identification of target regions of primer sets in hg38 and T2T genome assemblies

Potential target regions of primer sets in the hg38 and T2T genome assemblies were determined by a custom Perl script (Table S4). Sequence similarity searches of primer sequences on the hg38 and T2T reference assemblies were performed by blastn, with default parameters[45]. All matches of 17 bp or longer and with at most 1 mismatch and all perfect matches of 16 bp 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 BSC space, in this case both the genome and the primer sequences were converted *in-silico*.

Statistical analyses

Distribution of values were compared by applying the non-parametric, two tailed Kolmogorov Smirnoff test, as implemented by the ks.test() function from the stats package of the R programming language.

Analysis of bisulfite converted reads

Quality control and determination of target regions

Reads' quality was assessed by Fastqc[51,52]. Individual reports were merged by MultiQC[53] and quality metrics were visually inspected.

Non-converted amplicons' reads were aligned to T2T and hg38 human genome assemblies using Bowtie2[54]. Sample-level coverage profiles were computed by bedtools genomecov[55]. 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 172 and 57 candidate target regions were identified in the T2T and hg38 assemblies, respectively. 50/57 and 169/172 had at least a high similarity match (17 bp with at most 1 mismatch) to one or more primer sequences.

Highly similar candidate regions were collapsed to provide a non-redundant representation of repetitive sequences and allow non-ambiguous mapping of BSC reads. Different sequence identity thresholds were tested through a titration analysis to identify the ideal sequence similarity threshold. The total number of non-ambiguously mapped BSC reads and the total number of CpGs covered by at least 10 BSC reads in at least 50% of samples (testable CpGs) were recorded (Table 1). Considerations based on the number and cumulative size of genomic regions that were merged, and on the number of testable CpGs, prompted us to select 95% identity as the most appropriate threshold. In conclusion, by comparing hg38 and T2T-based analyses of non-converted reads, we defined a total of 60 and 9 distinct consensus groups (groups of sequences GS) respectively (Table S5 and Table S6). For every target region the consensus sequence was determined by majority rule consensus, by applying the cons EMBOSS program[48]. These consensus sequences were used for the further analysis of methyl-seq data.

Assessment of methylation levels

BSC reads were aligned to consensus target region sequences and methylation levels were determined by Bismark[56] +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*, *DNMT3B*, *LRIF1* variants. We collected the identified *SMCHD1* variants in the cohort and performed a reannotation on GRCh37 genome assembly using wAnnovar[57] (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 splicing variants with a GnomAD exome all MAF < 10^{-4} (GnomAD frequency for PM2 pathogenic moderate rule in Varsome)[58]. 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[59,60] (https://varsome.com/; the analysis was performed on 27.9.22). As *SMCHD1* gene function has not been thoroughly investigated yet, we decided not to filter out the variants predicted as VUS (Variants of Uncertain Significance). Finally, variant validation through Sanger sequencing was performed on probands and on relatives (Table S7).

Results

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

The extent of D4Z4-like repeat elements in the human genome was investigated by analyzing T2T and 86 haplotype-level genome assemblies from the PGR[42].

Our analyses uncovered hundreds of “complete” (>3200 bp) as well as partial (300 – 3200 bp) D4Z4-like sequences (hereafter D4Z4-l), irrespective of the reported geographic origin of the subjects (Table S8). The cumulative size of 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 kilobases of sequences on chr 1, chr 4, chr 10, chr 15, chr 21 and chr 22 (Figure 2A). Notably, in hg38 only 86.5 Kb of D4Z4-l were observed, representing a more than 10-fold reduction compared with T2T and PGR (Figure 2B). In hg38, D4Z4-l were prevalently associated with the long arms of chr 4 and chr 10 (Figure 2A) 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, Table S8) of “incomplete” D4Z4-l repeats of < 2.5 Kb in size, which are lacking in hg38. Conversely (Figure 2B), 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 (Figure S3).

Figure 2C shows the variability and the distribution of complete D4Z4-l elements. Six different groups were defined on the basis of sequence identity levels. D4Z4 from T2T were used to anchor/assign each group to a chromosome. The largest clusters (Table 2) 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 number of complete D4Z4 elements assigned to distinct clusters in the 86 PGR haplotypes is summarized in Figure 2D and Table 2. 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. Substitution rates estimates at haplotype levels suggested a high inter-individual variability, with rates ranging from 0.164 to 0.4312 (average of 0.245) substitutions for 100 bp for chr 4 and from 0.015 to 0.21 (average of 0.13) for chr 10 (Figure S4 and Figure S5). Interestingly, as shown in Table 2, more than 75% of the assemblies carried D4Z4-l assigned to cluster-6 (chr 22); whereas 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).

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

The permissive DRA 4qA-PAS haplotype has been proposed to cause FSHD. However, we previously reported a frequency of 2% for this haplotype[38], which is not compatible with the prevalence of the disease[61]. We thus analyzed the T2T and PGR to investigate the prevalence of this molecular signature in other populations and to study the conservation and the integrity of ter-*DUX4*.

The pLam and telomeric qA and qB sequences (Supplemental Material and Methods) were used to identify and classify terminal D4Z4 repeat in PGR, T2T and hg38. A total of 172 D4Z4 terminal repeats were identified (Table S9) and tentatively assigned to chr 4 or chr 10 based on sequence similarity profiles (as described by the clustering of the D4Z4-l sequences in Figure 2C). 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[62,63]. 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 (Table S9). Patterns of conservation of ter-*DUX4* sequences were investigated. Figure 3A shows that

ter-*DUX4* assigned to the 10q D4Z4 terminal repeat are found to be highly similar and formed a single cluster; while 4q ter-*DUX4* are more heterogeneous and were partitioned in 3 groups (Figure 3A-B). Interestingly, as shown in Figure 3C, in chr 10 and chr4_group 2 ter-*DUX4* 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); Figure 3A also shows that the most variable sequences (chr4_group 1 and chr4_group 3) were preferentially associated with qB alleles. All ter-*DUX4* sequences were found to be intact, and none was associated with frameshifts or premature stop codons. Consistent with previous observations[18,19], all the alleles classified as qB did not carry a pLam and were more proximal to the ter-*DUX4* compared to qA (Figure 3D-G). In the same way, (Figure 3E-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[64]; while in pLam assigned to 10q the PAS sequences 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 (Table 3).

Remarkably, in line with our previous observations[38–40] (Table 3), 4/86 haplotypes (4.6%) carried D4Z4 alleles with 6-8 RU and 4qA-PAS (Figure 3F). None of these haplotypes carried disruptive variants in the ter-*DUX4* or in the pLam sequence element.

No statistically significant difference in the total number of chr 4 D4Z4 repeats was observed when qA/qB alleles pLam+/pLam- alleles were compared (Figure 3H-I), suggesting that alleles that are not permissive for the expression of the ter-*DUX4* are not associated with significant differences in D4Z4 copy numbers compared with the permissive qA-PAS haplotype.

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

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 potential concerns about the reliability of methylation/diagnostic tests based on the hg38. As bisulfite treated DNA sequencing at targeted regions is commonly used to diagnose FSHD[65], the uncovered variability of D4Z4-l sequences prompted us to investigate D4Z4 CpG methylation comparing the hg38 and the T2T.

A set of primers spanning the entire length of the D4Z4 repeat and a total of 126 CpGs (Figure S2 and Table S3) were designed based on hg38 assembly to recapitulate previous findings in the field[41]. This experimental design was used to assess a carefully selected cohort of 26 index cases referred for FSHD2 diagnosis, and 14 relatives all carrying 10 or more RU. All the participants assayed were searched for damaging variants in the chromatin remodelers *SMCHD1, DNMT3B* and *LRIF1*. In addition, 3 FSHD1 index cases and 1 relative carrying 4 RU were studied.

Amplicon targeted regions were sequenced both in non-converted and in BSC DNA samples. This approach was undertaken: i. to test the 4q/10q-specificity of the primer sets commonly used and designed based on hg38 assembly[41], ii. to facilitate the alignment of short BSC derived reads. Both non-converted and BSC NGS data were used for sequence alignment based on hg38 and T2T reference genomes (Table S4).

A total of 77% and 99.43% non-converted short reads mapped to hg38 and T2T respectively. Moreover, the majority of BSC reads (62% for T2T and 54% for hg38) did not align at 4q primer target regions, potentially suggesting “off target” sequences, and/or high levels of heterogeneity in methylation levels. To identify genomic regions targeted by our assay in an unbiased manner, genomic intervals covered by 100 or more non-converted reads in at least 50% of the samples were recorded. A total of 57 and 172 genomic regions

in hg38 and T2T displayed these remarkable levels of coverage. Notably, 50/57 and 169/172 of the target regions delineated by coverage analyses had at least a high similarity match (17 bp with at most 1 mismatch) to one or more primer sequences. Based on these observations sequences above 95% sequence similarity were collapsed to prevent inconsistent mapping of BSC short reads across highly similar D4Z4-l elements. By this approach, we defined a total of 60 and 9 distinct consensus groups (groups of sequences, hereafter GS) in T2T and hg38 respectively (Table S5 and Table S6). Consensus sequences were computed for each GS and used for the further analysis of methyl-seq data.

This *ad-hoc* bioinformatics workflow revealed remarkable differences in the completeness, number, and breadth of testable CpGs when the GSs derived from hg38 and T2T were compared: hg38 125 CpG; T2T 938 CpG. Global methylation patterns were summarized in the form of a heatmap (Figure S6 and Figure S7). In the T2T-based analysis a total of 4 distinct GS including 81 distinct CpGs had no missing data and complete methylation profiles across all samples, as shown in Figure 4 and Table 4: GS1 (primer D6): 28 CpGs; GS2 (primer D1): 32 CpGs; GS5 (chr 13-chr 15): 16 CpGs and GS6 (chr 13-chr 15): 5 CpGs). Notably, only GS1 and GS2 included D4Z4 5' 4q/10q-specific regions exclusively, for a total of 60 CpGs, while GS5 and GS6 mapped in D4Z4-l repeats on chr 13 and 15. The wide-range distributions of CpG methylation levels, as displayed in Figure S8, highlight a high variability at regions GS1 and GS2, while off-target regions GS5 and GS6 have high methylation levels, with a narrower distribution. Figure S9 shows a general overview of the CpG methylation patterns observed in all the 60 distinct GSs defined in our analysis. Taken together these findings indicate that a significant proportion of the BSC reads recovered by our targeted high throughput assay derive from off-target, hyper-methylated D4Z4-l regions. We also observed that when hg38 is considered as reference, the criteria we established for the T2T-based analysis (namely no missing data and

437 complete methylation profiles across all samples) identified only a single group, GS3 (primer
438 D1) including 26 CpGs (Figure 4, Figure 5 and Table 4). In the light of these observations,
439 methylation profiles inferred from T2T were considered more informative and were selected
440 for subsequent analyses. Only primer D1 had complete data both in T2T and hg38 (Table
441 4); however, CpG methylation levels estimated on hg38 were slightly more elevated
442 compared to those observed on T2T at the equivalent target region (Figure S10, median
443 values: 49.13 hg38, 43.75 T2T). An observation that might be consistent with discrepancies
444 in the mapping of BSC reads on different genome assemblies.

445 Further, our analysis revealed that 4q/10q-specific methylation patterns at GS1 and
446 GS2 stratified samples into 3 groups: low methylation (average CpG methylation 24%),
447 intermediate methylation (average CpG methylation 48%), and high methylation (average
448 CpG methylation 71%). The methylation pattern of GS5 and GS6 was highly uniform across
449 all samples, with an average of 89%.

450 Interestingly, the observed methylation profiles did not correlate with the disease
451 status. As shown in Figure 6A, of 16 samples presenting low methylation 10 (62.5%) derived
452 from cases presenting the classical FSHD features (CCEF category A), 6 (37,5%) derived
453 from cases presenting incomplete (2, CCEF clinical Category B) or complex (4, CCEF
454 clinical category D) phenotypes. Of 16 samples, presenting intermediate methylation 9
455 (56.3%) derived from cases assessed as clinical category A. Of 12 samples presenting high
456 methylation 7 were from Category A cases (58.3%). Consistently, the methylation level
457 distributions assessed in DNA from subjects presenting the CCEF clinical categories A, B
458 and D was not statistically different (Figure S11). Instead, the majority (11/16) of participants
459 in the low methylation group carried Pathogenic, Likely Pathogenic variants or Variants of
460 Uncertain Significance according to ACMG classification[60] (P/LP/VUS variants) in *SMCHD1*
461 (Fisher exact test p-value 0.0003). This association was confirmed by the comparison of the
462 methylation profile distributions of *SMCHD1* P/LP/VUS variants carriers and non-carriers

(Figure 4, Kolmogorov-Smirnov test p-value <= 1e-16). Identical patterns were recovered also when methylation levels on the hg38 assembly were assessed with our approach (Figure 5).

Discussion

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

Thirty years ago, reduction below a critical threshold of the number of repeats belonging to the D4Z4 macrosatellite at 4q35 was causally associated with FSHD. D4Z4 is part a family of repetitive elements characteristic of heterochromatin and it was then known that many D4Z4-l were present in the human genome[66]. At that time, the lack of a complete genome assembly hampered the possibility of fully deciphering the significance of molecular and epigenetic findings in FSHD. Very recently, these limitations have been overcome by the availability of complete, high quality haplotype level genome assemblies. 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. We established that the number of 3.3 kb D4Z4 repeats included in each array ranges from 6 RU to 89 RU[40]. We observed that the distance between the qA sequences and the last D4Z4 repeat varies. We also confirmed that qA sequences are more distal compared to qB sequences; the pLam sequence is in association with qA haplotypes[18], and valid PAS is exclusively associated with 4q[64]; chr 10 D4Z4 repeats were all marked by qA, with a single exception: a 10qB allele carrying 1 RU. Notably, PGR haplotype analysis confirmed the high frequency of the permissive 4qA-PAS haplotype, which is comparable to a common polymorphism[61]. In fact, in 4.6% of cases we identified permissive haplotypes consisting of reduced 4q D4Z4 arrays with ≤ 8 RU and the

4qA-PAS. Since all ter-*DUX4* and pLam sequence elements associated with 4qA alleles were found to be conserved and functionally intact, we speculate that disruptive genomic variants at the ter-*DUX4* are unlikely to account for the staggering discrepancy between the frequency of the presumed causative 4qA-PAS allele and the prevalence of FHSD. Furthermore, no skewed association has been detected between 4q D4Z4 arrays copy number and qA/B alleles[19], indicating that non-permissive 4qB alleles and potentially permissive 4qA alleles do not have a different selective pressure for the mainteinance of the physiological number of D4Z4 RU.

The consistency between data obtained from the analysis of PGR and previous observations[6,61,66] shows the validity of the PGR to advance knowledge of the molecular signature used to diagnose FSHD on a world-wide scale.

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 (Table S9). This condition has been reported before[62] and is ascribable to the spreading of repetitive and polymorphic D4Z4-l at different loci of the genome[63,67].

Our comparative analyses of complete D4Z4-l (Figure 2C) highlighted that D4Z4 repeats, as well as the ter-*DUX4*, 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 African 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[20,32,68–74]. In addition, the correlation between D4Z4 methylation level and FSHD clinical status has often been postulated, but different technical settings in the various studies, the lack of clear clinical evaluation of patients and the prominent absence of family studies had hindered the interpretation of results[20,41,73,74]. In the present work, we comprehensively evaluated the D4Z4 methylation status by considering a high number of reads and excluding technical biases introduced by the previous studies based on hg38 and on standard bioinformatics pipelines. We established (Figure S2) that the primer sets BSS-D1 and BSS-D6 identify CpG methylation signals deriving from 4q/10q-specific regions only, whereas the primer sets BSS-D3 and BSS-D5 amplify D4Z4-l distributed not only at 4q and 10q but also on chromosome 1 and on the short arms of acrocentric chromosomes. Thus, the study of D4Z4 CpG methylation can be hazardous if the distribution of D4Z4-l elements and their variability are not correctly taken into account.

Despite all this, the application of short read-based assays for the analysis of repetitive elements of relatively large size has limitations. 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. We can anticipate 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, Butterfield and colleagues (2023)[75] 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 which was irrespective of the

clinical phenotype (FSHD or healthy)[75]. 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 large 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[20], not all DNAs from myopathic patients referred as FSHD2 (26 index cases) 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). Figure 6 depicts the results of several comparisons regarding the following parameters: clinical phenotype, D4Z4 methylation level and *SMCHD1* mutational status. Figure 6A shows that some genotype-phenotype intersections are not in agreement with the current indications for FSHD diagnosis[76,77]. Remarkably, we observed the full range of D4Z4 methylation levels in DNA from the 26 participants presenting a classical FSHD phenotype (CCEF Clinical Category A); of those 16 carriers of a damaging *SMCHD1* variant, 11 presented low 4q/10q-specific D4Z4 CpG methylation, 5 displayed intermediate CpG methylation. We also observed 5 samples with low CpG methylation with no damaging *SMCHD1* or *DNMT3B* or *LRIF1* variants and presenting classical (2), incomplete (1) or complex (2) phenotypes. underlying the genetic heterogeneity of this molecular phenotype.

The variants we identified were distributed all along *SMCHD1* 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 (Figure 6B).

To further explore this aspect, we investigated genotype-phenotype correlation in eight trios/parent-child pairs participating to this study (Figure 6C, Table S7). We observed participants presenting the classical FSHD phenotype (CCEF Category A), healthy subjects (CCEF Category C) or complex phenotypes presenting atypical features (CCEF Category D). According to our data, only in families 1 and 26 subjects presenting a classic FSHD phenotype displayed a reduced CpG methylation profile and carried a damaging heterozygous *SMCHD1* variant. In family 22 the classic clinical phenotype, in presence of D4Z4 reduced methylation, was not associated with variants in known chromatin modifiers. In family 8 the probands presented the classic clinical phenotype associated with intermediate D4Z4 methylation level in presence of a damaging *SMCHD1* variant, as the healthy mother. In family 27 the proband presented the Category A phenotype associated with reduced CpG methylation and a damaging *SMCHD1* variant, whereas intermediate D4Z4 methylation level was detected in the healthy father carrying the same *SMCHD1* allele. In family 30 father and daughter presented a complex phenotype with atypical clinical symptoms (Category D1); both reported clinical onset at pelvic girdle and distal lower limbs. They were heterozygous for *SMCHD1* damaging variant and presented reduced D4Z4 CpG methylation. In families 9 both the probands and her father presented Category A phenotype without displaying reduced D4Z4 CpG methylation and carrying WT *SMCHD1* alleles, as the proband in family 11.

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.

In conclusion, our work uncovered the complexity of the genomic setting of D4Z4-l and showed that a molecular diagnostic test for FSHD2 based on *SMCHD1* inactivation and D4Z4 hypomethylation 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 network leading to muscle wasting in FSHD.

Declarations

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. We acknowledge the Fondazione di Modena for FAR-FOMO 2021 grant funding this work.

Authors' contributions

RGT, VS, MC conceived the study concept and supervised the project; VS, SP conducted the molecular analysis; MC conducted bioinformatics analysis; PK and MK contributed to bioinformatic 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.

Declaration of interests

The authors declare no competing interests.

Web sources

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

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.

Ethics approval and consent to participate

Signed informed consent was obtained from all the subjects prior to the inclusion in the study. The study was approved by the ethics committee of Emilia Romagna Area Vasta Nord 743/2022/OSS/UNIMO SIRER ID 5111.

References

1. Deenen, J.C.W., Arnts, H., Van Der Maarel, S.M., Padberg, G.W., Verschuuren, J.J.G.M., Bakker, E., Weinreich, S.S., Verbeek, A.L.M., and Van Engelen, B.G.M. (2014). Population-based incidence and prevalence of facioscapulohumeral dystrophy. Neurology *83*, 1056–1059. 10.1212/WNL.0000000000000797.

2. Mostacciuolo, M.L., Pastorello, E., Vazza, G., Miorin, M., Angelini, C., Tomelleri, G., Galluzzi, G., and Trevisan, C. Pietro (2009). Facioscapulohumeral muscular dystrophy: Epidemiological and molecular study in a north-east Italian population sample. Clinical Genetics *75*, 550–555. 10.1111/j.1399-0004.2009.01158.x.

3. Dumbovic, G., Forcales, S. V., and Perucho, M. (2017). Emerging roles of macrosatellite repeats in genome organization and disease development. Epigenetics *12*, 515–526. 10.1080/15592294.2017.1318235.

4. Lunt, P.W. (1998). 44th ENMC International Workshop: Facioscapulohumeral muscular dystrophy: Molecular studies, 19-21 July 1996, Naarden, The Netherlands. Neuromuscular Disorders *8*, 126–130. 10.1016/S0960-8966(98)00012-1.

5. Hewitt, J.E., Lyle, R., Clark, L.N., Valleley, E.M., Wright, T.J., Wijmenga, C., Van Deutekom, J.C. t., Francis, F., Sharpe, P.T., Hofker, M., et al. (1994). Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystropothhy. Human Molecular Genetics *3*, 1287–1295. 10.1093/hmg/3.8.1287.

6. Winokur, S.T., Bengtsson, U., Feddersen, J., Mathews, K.D., Weiffenbach, B., Bailey, H., Markovich, R.P., Murray, J.C., Wasmuth, J.J., Altherr, M.R., et al. (1994). 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. Chromosome Research *2*, 225–234. 10.1007/BF01553323.

7. Zhang, X.Y., Loflin, P.T., Gehrke, C.W., Andrews, P.A., and Ehrlich, M. (1987). Hypermethylation of human DNA sequences in embryonal carcinoma cells and somatic tissues but not in sperm. Nucleic Acids Research *15*, 9429–9449. 10.1093/nar/15.22.9429.

8. Agresti, A., Meneveri, R., Siccardi, A.G., Marozzi, A., Corneo, G., Gaudi, S., and Ginelli, E. (1989). Linkage in human heterochromatin between highly divergent Sau3A repeats and a new family of repeated DNA sequences (HaeIII family). Journal of Molecular Biology *205*, 625–631. 10.1016/0022-2836(89)90308-2.

9. Meneveri, R., Agresti, A., Marozzi, A., Saccone, S., Rocchi, M., Archidiacono, N., Corneo, G., Valle, G.D., and Ginelli, E. (1993). Molecular organization and chromosomal location of human GC-rich heterochromatic blocks. Gene *123*, 227–234. 10.1016/0378-1119(93)90128-P.

10. Deidda, G., Cacurri, S., Grisanti, P., Vigneti, E., Piazzo, N., and Felicetti, L. (1995). Physical mapping evidence for a duplicated region on chromosome 10qter showing high homology with the facioscapulohumeral muscular dystrophy locus on chromosome 4qter. European Journal of Human Genetics *3*, 155–167. 10.1159/000472291.

11. Deutekom, J.C.T. van, Wljmenga, C., Tlenhoven, E.A.E. van, Gruter, A.M., Hewitt, J.E., Padberg, G.W., van Ommen, G.J.B., Hofker, M.H., and Fronts, R.R. (1993). FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. Human Molecular Genetics *2*, 2037–2042. 10.1093/hmg/2.12.2037.

12. Gabellini, D., Green, M.R., and Tupler, R. (2002). Inappropriate gene activation in FSHD: A repressor complex binds a chromosomal repeat deleted in dystrophic muscle. Cell *110*, 339–348. 10.1016/S0092-8674(02)00826-7.

13. Cabianca, D.S., Casa, V., Bodega, B., Xynos, A., Ginelli, E., Tanaka, Y., and Gabellini, D. (2012). A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in fshd muscular dystrophy. Cell *149*, 819–831. 10.1016/j.cell.2012.03.035.

14. Laoudj-Chenivesse, D., Carnac, G., Bisbal, C., Hugon, G., Bouillot, S., Desnuelle, C., Vassetzky, Y., and Fernandez, A. (2005). Increased levels of adenine nucleotide translocator 1 protein and response to oxidative stress are early events in facioscapulohumeral muscular dystrophy muscle. Journal of molecular medicine (Berlin, Germany) *83*, 216–224. 10.1007/s00109-004-0583-7.

15. Lemmers, R.J.L.F., Tawil, R., Petek, L.M., Balog, J., Block, G.J., Santen, G.W.E., Amell, A.M., Vliet, P.J. Van Der, Almomani, R., Straasheijm, R., et al. (2012). Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. Nat Genet. *44*, 1370–1374. 10.1038/ng.2454.Digenic.

16. 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. American Journal of Human Genetics *98*, 1020–1029. 10.1016/j.ajhg.2016.03.013.

17. Hamanaka, K., Šikrová, D., Mitsuhashi, S., Masuda, H., Sekiguchi, Y., Sugiyama, A., Shibuya, K., Lemmers, R.J.L.F., Goossens, R., Ogawa, M., et al. (2020). Homozygous nonsense variant in LRIF1 associated with facioscapulohumeral muscular dystrophy. Neurology *94*, E2441–E2447. 10.1212/WNL.0000000000009617.

18. Van Geel, M., Dickson, M.C., Beck, A.F., Bolland, D.J., Frants, R.R., Van der Maarel, S.M., De Jong, P.J., and Hewitt, J.E. (2002). Genomic analysis of human chromosome 10q and 4q telomeres suggests a common origin. Genomics *79*, 210–217. 10.1006/geno.2002.6690.

19. Lemmers, R.J.L.F., de Kievit, P., Sandkuijl, L., Padberg, G.W., van Ommen, G.J.B., Frants, R.R., and van der Maarel, S.M. (2002). Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. Nature Genetics *32*, 235–236. 10.1038/ng999.

20. Gould, T., Jones, T.I., and Jones, P.L. (2021). Precise epigenetic analysis using targeted bisulfite genomic sequencing distinguishes fshd1, fshd2, and healthy subjects. Diagnostics *11*, 1–17. 10.3390/diagnostics11081469.

21. Ricci, G., Scionti, I., Sera, F., Govi, M., D'Amico, R., Frambolli, I., Mele, F., Filosto, M., Vercelli, L., Ruggiero, L., et al. (2013). Large scale genotype-phenotype analyses indicate that novel prognostic tools are required for families with facioscapulohumeral muscular dystrophy. Brain *136*, 3408–3417. 10.1093/brain/awt226.

22. K Goto, I Nishino, Y.K.H. (2004). Very low penetrance in 85 Japanese families with facioscapulohumeral muscular dystrophy 1A K. J Med Genet *41*. 10.4324/9780203994559-14.

23. Wu, Z.Y., Wang, Z.Q., Murong, S.X., and Wang, N. (2004). FSHD in Chinese population: Characteristics of translocation and genotype-phenotype correlation. Neurology *63*, 581–583. 10.1212/01.WNL.0000133210.93075.81.

24. Nikolic, A., Ricci, G., Sera, F., Bucci, E., Govi, M., Mele, F., Rossi, M., Ruggiero, L., Vercelli, L., Ravaglia, S., et al. (2016). Clinical expression of facioscapulohumeral muscular dystrophy in carriers of 1-3 D4Z4 reduced alleles: Experience of the FSHD Italian National Registry. BMJ Open *6*, 1–10. 10.1136/bmjopen-2015-007798.

25. Nakagawa, M., Matsuzaki, T., Higuchi, I., Fukunaga, H., Inui, T., Nagamitsu, S., Yamada, H., Arimura, K., and Osame, M. (1997). Facioscapulohumeral Muscular Dystrophy: Clinical Diversity and Genetic Abnormalities in Japanese Patients. Internal Medicine *36*, 333–339. 10.2169/internalmedicine.36.333.

26. Sakellariou, P., Kekou, K., Fryssira, H., Sofocleous, C., Manta, P., Panousopoulou, A., Gounaris, K., and Kanavakis, E. (2012). Mutation spectrum and phenotypic manifestation in FSHD Greek patients. Neuromuscular Disorders *22*, 339–349. 10.1016/j.nmd.2011.11.001.

27. Salort-Campana, E., Nguyen, K., Bernard, R., Jouve, E., Solé, G., Nadaj-Pakleza, A., Niederhauser, J., Charles, E., Ollagnon, E., Bouhour, F., et al. (2015). Low penetrance in facioscapulohumeral muscular dystrophy type 1 with large pathological D4Z4 alleles: A cross-sectional multicenter study. Orphanet Journal of Rare Diseases *10*, 4–11. 10.1186/s13023-014-0218-1.

28. 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. Journal of Neurology *263*, 1204–1214. 10.1007/s00415-016-8123-2.

29. 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 and Nerve *42*, 213–217. 10.1002/mus.21671.

30. Ruggiero, L., Mele, F., Manganelli, F., Bruzzese, D., Ricci, G., Vercelli, L., Govi, M., Vallarola, A., Tripodi, S., Villa, L., et al. (2020). Phenotypic Variability Among Patients With D4Z4 Reduced Allele Facioscapulohumeral Muscular Dystrophy. JAMA network open *3*, e204040. 10.1001/jamanetworkopen.2020.4040.

31. Vercelli, L., Mele, F., Ruggiero, L., Sera, F., Tripodi, S., Ricci, G., Vallarola, A., Villa, L., Govi, M., Maranda, L., et al. (2021). A 5-year clinical follow-up study from the Italian National Registry for FSHD. Journal of Neurology *268*, 356–366. 10.1007/s00415-020-10144-7.

32. Nikolic, A., Jones, T.I., Govi, M., Mele, F., Maranda, L., Sera, F., Ricci, G., Ruggiero, L., Vercelli, L., Portaro, S., et al. (2020). Interpretation of the epigenetic signature of facioscapulohumeral muscular dystrophy in light of genotype-phenotype studies. International Journal of Molecular Sciences *21*. 10.3390/ijms21072635.

33. Ricci, G., Cammish, P., Siciliano, G., Tupler, R., Lochmuller, H., and Evangelista, T. (2019). Phenotype may predict the clinical course of facioscapolohumeral muscular dystrophy. Muscle and Nerve *59*, 711–713. 10.1002/mus.26474.

34. Ricci, G., Mele, F., Govi, M., Ruggiero, L., Sera, F., Vercelli, L., Bettio, C., Santoro, L., Mongini, T., Villa, L., et al. (2020). Large genotype–phenotype study in carriers of D4Z4 borderline alleles provides guidance for facioscapulohumeral muscular dystrophy diagnosis. Scientific Reports *10*, 1–12. 10.1038/s41598-020-78578-7.

35. He, J.J., Lin, X.D., Lin, F., Xu, G.R., Xu, L.Q., Hu, W., Wang, D.N., Lin, H.X., Lin, M.T., Wang, N., et al. (2018). Clinical and genetic features of patients with facial-sparing facioscapulohumeral muscular dystrophy. European Journal of Neurology *25*, 356–364. 10.1111/ene.13509.

36. Wang, Z., Qiu, L., Lin, M., Chen, L., Zheng, F., Lin, L., Lin, F., Ye, Z., Lin, X., He, J., et al. (2022). Prevalence and disease progression of genetically-confirmed facioscapulohumeral muscular dystrophy type 1 (FSHD1) in China between 2001 and 2020: a nationwide

772 population-based study. The Lancet regional health. Western Pacific *18*, 100323.
773 10.1016/j.lanwpc.2021.100323.

774 37. Salsia, V., Vattemi, G.N.A., and Tupler, R.G. (2023). The FSHD jigsaw: are we placing the
775 tiles in the right position? Curr Opin Neurol. 10.1097/WCO.0000000000001176.

776 38. Wohlgemuth, M., Lemmers, R.J., Van der Kooi, E.L., Van der Wielen, M.J., Van Overveld, P.G.,
777 Dauwerse, H., Bakker, E., Frants, R.R., Padberg, G.W., and Van der Maarel, S.M. (2003).
778 Possible phenotypic dosage effect in patients compound heterozygous for FSHD-sized
779 4q35 alleles. Neurology *61*, 909–913. 10.1212/WNL.61.7.909.

780 39. Scionti, I., Greco, F., Ricci, G., Govi, M., Arashiro, P., Vercelli, L., Berardinelli, A., Angelini, C.,
781 Antonini, G., Cao, M., et al. (2012). Large-scale population analysis challenges the current
782 criteria for the molecular diagnosis of fascioscapulohumeral muscular dystrophy.
783 American Journal of Human Genetics *90*, 628–635. 10.1016/j.ajhg.2012.02.019.

784 40. Van Overveld, P.G.M., Lemmers, R.J.F.L., Deidda, G., Sandkuijl, L., Padberg, G.W., Frants,
785 R.R., and Van Der Maarel, S.M. (2000). Interchromosomal repeat array interactions
786 between chromosomes 4 and 10: A model for subtelomeric plasticity. Human Molecular
787 Genetics *9*, 2879–2884. 10.1093/hmg/9.19.2879.

788 41. Salsi, V., Magdinier, F., and Tupler, R. (2020). Does DNA methylation matter in FSHD?
789 Genes *11*. 10.3390/genes11030258.

790 42. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H.A., Lucas, J.K., Phillippy, A.M., Popejoy,
791 A.B., Asri, M., Carson, C., Chaisson, M.J.P., et al. (2022). The Human Pangenome Project: a
792 global resource to map genomic diversity. Nature *604*, 437–446. 10.1038/s41586-022-
793 04601-8.

794 43. Bettio, C., Salsi, V., Orsini, M., Calanchi, E., Magnotta, L., Gagliardelli, L., Kinoshita, J.,
795 Bergamaschi, S., and Tupler, R. (2021). The Italian National Registry for FSHD: an
796 enhanced data integration and an analytics framework towards Smart Health Care and
797 Precision Medicine for a rare disease. Orphanet Journal of Rare Diseases *16*, 1–21.
798 10.1186/s13023-021-02100-z.

799 44. Li, L.C., and Dahiya, R. (2002). MethPrimer: Designing primers for methylation PCRs.
800 Bioinformatics *18*, 1427–1431. 10.1093/bioinformatics/18.11.1427.

801 45. Altschup, S.F., Gish, W., Pennsylvania, T., and Park, U. (1990). Basic Local Alignment
802 Search Tool 2Department of Computer Science. 403–410.

803 46. Lassmann, T. (2020). Kalign 3: Multiple sequence alignment of large datasets.
804 Bioinformatics *36*, 1928–1929. 10.1093/bioinformatics/btz795.

805 47. Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use
806 in phylogenetic analysis. Molecular Biology and Evolution *17*, 540–552.
807 10.1093/oxfordjournals.molbev.a026334.

808 48. Rice, P., Longden, L., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology
809 Open Software Suite. Trends in Genetics *16*, 276–277. 10.1016/S0168-9525(00)02024-2.

810 49. Tajima, F., and Nei, M. (1984). Estimation of evolutionary distance between nucleotide
811 sequences. Molecular Biology and Evolution *1*, 269–285.

50. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/.

51. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

52. FastQC (2015). https://qubeshub.org/resources/fastqc.

53. Ewels, P., Lundin, S., and Max, K. (2016). Data and text mining MultiQC : summarize analysis results for multiple tools and samples in a single report. *32*, 3047–3048. 10.1093/bioinformatics/btw354.

54. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods *9*, 357–359. 10.1038/nmeth.1923.

55. Quinlan, A.R., and Hall, I.M. (2010). BEDTools : a flexible suite of utilities for comparing genomic features. *26*, 841–842. 10.1093/bioinformatics/btq033.

56. Krueger, F., and Andrews, S.R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572. 10.1093/bioinformatics/btr167.

57. Chang, X.; Wang, K. (2008). wANNOVAR: annotating genetic variants for personal genomes via the web. Journal of Human Genetics *6*, 2166–2171. 10.1136/jmedgenet-2012-100918.wANNOVAR.

58. Chen, K., Hu, J., Moore, D.L., Liu, R., Kessans, S.A., Breslin, K., Lucet, I.S., Keniry, A., Leong, H.S., Parish, C.L., et al. (2015). Genome-wide binding and mechanistic analyses of Smchd1-mediated epigenetic regulation. Proceedings of the National Academy of Sciences of the United States of America *112*, E3535–E3544. 10.1073/pnas.1504232112.

59. Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C.E., Albarca Aguilera, M., Meyer, R., and Massouras, A. (2019). VarSome: the human genomic variant search engine. Bioinformatics *35*, 1978–1980. 10.1093/bioinformatics/bty897.

60. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, R.H., and Committee., A.L.Q.A. (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. 10.1038/gim.2015.30.

61. Scionti, I., Greco, F., Ricci, G., Govi, M., Arashiro, P., Vercelli, L., Berardinelli, A., Angelini, C., Antonini, G., Cao, M., et al. (2012). Large-scale population analysis challenges the current criteria for the molecular diagnosis of fascioscapulohumeral muscular dystrophy. American Journal of Human Genetics *90*, 628–635. 10.1016/j.ajhg.2012.02.019.

62. Delourme, M., Charlene, C., Gerard, L., Ganne, B., Perrin, P., Vovan, C., Bertaux, K., Nguyen, K., and Magdinier, F. (2023). Complex 4q35 and 10q26 Rearrangements. 10.1212/NXG.0000000000200076.

63. Cacurri, S., Piazzo, N., Deidda, G., Vigneti, E., Galluzzi, G., Colantoni, L., Merico, B., Ricci, E., and Felicetti, L. (1998). Sequence homology between 4qter and 10qter loci facilitates the

instability of subtelomeric KpnI repeat units implicated in facioscapulohumeral muscular dystrophy. American Journal of Human Genetics *63*, 181–190. 10.1086/301906.

64. Lemmers, R.J.L.F., van der Vliet, P.J., van der Gaag, K.J., Zuniga, S., Frants, R.R., de Knijff, P., and van der Maarel, S.M. (2010). Worldwide Population Analysis of the 4q and 10q Subtelomeres Identifies Only Four Discrete Interchromosomal Sequence Transfers in Human Evolution. American Journal of Human Genetics *86*, 364–377. 10.1016/j.ajhg.2010.01.035.

65. de Greef, J.C., Lemmers, R.J.L.F., van Engelen, B.G.M., Sacconi, S., Venance, S.L., Frants, R.R., Tawil, R., and van der Maarel, S.M. (2009). Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. Human Mutation *30*, 1449–1459. 10.1002/humu.21091.

66. Hewitt, J.E., Lyle, R., Clark, L.N., Valleley, E.M., Wright, T.J., Wijmenga, C., Van Deutekom, J.C. t., Francis, F., Sharpe, P.T., Hofker, M., et al. (1994). Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystropothhy. Human Molecular Genetics *3*, 1287–1295. 10.1093/hmg/3.8.1287.

67. Pini, S., Napoli, F.M., Tagliafico, E., La Marca, A., Bertucci, E., Salsi, V., and Tupler, R. (2023). De novo variants and recombination at 4q35: Hints for preimplantation genetic testing in facioscapulohumeral muscular dystrophy. Clinical Genetics *103*, 242–246. 10.1111/cge.14250.

68. Gaillard, M.C., Roche, S., Dion, C., Tasmadjian, A., Bouget, G., Salort-Campana, E., Vovan, C., Chaix, C., Broucqsault, N., Morere, J., et al. (2014). Differential DNA methylation of the D4Z4 repeat in patients with FSHD and asymptomatic carriers. Neurology *83*, 733–742. 10.1212/WNL.0000000000000708.

69. Lemmers, R.J.L.F., Goeman, J.J., Van der Vliet, P.J., Van Nieuwenhuizen, M.P., Balog, J., Vos-Versteeg, M., Camano, P., Ramos Arroyo, M.A., Jerico, I., Rogers, M.T., et al. (2015). Inter-individual differences in CpG methylation at D4Z4 correlate with clinical variability in FSHD1 and FSHD2. Human Molecular Genetics *24*, 659–669. 10.1093/hmg/ddu486.

70. Jones, T.I., King, O.D., Himeda, C.L., Homma, S., Chen, J.C.J., Beermann, M. Lou, Yan, C., Emerson, C.P., Miller, J.B., Wagner, K.R., et al. (2015). Individual epigenetic status of the pathogenic D4Z4 macrosatellite correlates with disease in facioscapulohumeral muscular dystrophy. Clinical Epigenetics *7*, 1–22. 10.1186/s13148-015-0072-6.

71. Calandra, P., Cascino, I., Lemmers, R.J.L.F., Galluzzi, G., Teveroni, E., Monforte, M., Tasca, G., Ricci, E., Moretti, F., Van Der Maarel, S.M., et al. (2016). Allele-specific DNA hypomethylation characterises FSHD1 and FSHD2. Journal of Medical Genetics *53*, 348–355. 10.1136/jmedgenet-2015-103436.

72. Dion, C., Roche, S., Laberthonnière, C., Broucqsault, N., Mariot, V., Xue, S., Gurzau, A.D., Nowak, A., Gordon, C.T., Gaillard, M.C., et al. (2019). SMCHD1 is involved in de novo methylation of the DUX4-encoding D4Z4 macrosatellite. Nucleic Acids Research *47*, 2822–2839. 10.1093/nar/gkz005.

73. Hiramuki, Y., Kure, Y., Saito, Y., Ogawa, M., Ishikawa, K., Yoshimura, M.M., Oya, Y., Takahashi, Y., Kim, D.S., Arai, N., et al. (2022). Simultaneous measurement of the size and methylation of chromosome 4qA - D4Z4 repeats in facioscapulohumeral muscular

dystrophy by long - read sequencing. Journal of Translational Medicine, 1–12. 10.1186/s12967-022-03743-7.

74. Erdmann, H., Scharf, F., Gehling, S., Benet-Pagès, A., Jakubiczka, S., Becker, K., Seipelt, M., Kleefeld, F., Knop, K.C., Prott, E.-C., et al. (2023). Methylation of the 4q35 D4Z4 repeat defines disease status in facioscapulohumeral muscular dystrophy. Brain *146*, 1388–1402. 10.1093/brain/awac336.

75. Butterfield, R.J., Dunn, D.M., Duvall, B., Moldt, S., and Weiss, R.B. (2023). Deciphering D4Z4 CpG methylation gradients in fascioscapulohumeral muscular dystrophy using nanopore sequencing. bioRxiv : the preprint server for biology. 10.1101/2023.02.17.528868.

76. Lemmers, R.J.L.F., Wohlgemuth, M., Van Der Gaag, K.J., Van Der Vliet, P.J., Van Teijlingen, C.M.M., De Knijff, P., Padberg, G.W., Frants, R.R., and Van Der Maarel, S.M. (2007). Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. American Journal of Human Genetics *81*, 884–894. 10.1086/521986.

77. Tawil, R., Kissel, J.T., Heatwole, C., Pandya, S., Gronseth, G., and Benatar, M. (2015). Evidence-based guideline summary: Evaluation, diagnosis, and management of facioscapulohumeral muscular dystrophy. Neurology *85*, 357–364. 10.1212/WNL.0000000000001783.

Figure titles and legends

Figure 1. Scheme of the proposed models for FSHD1 and FSHD2 pathogenesis

Tandemly arrayed D4Z4 repeats (triangles) and the 4q35 haplotype elements are represented along with the molecular signatures proposed for FSHD1 and FSHD2 pathogenesis.

Figure 2. 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 Kb of D4Z4-l sequence was identified, in any of the 2 assemblies, are reported. Size (in Kb) 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 Kb) of D4Z4-l elements in the hg38, T2T and PGR haplotype level assemblies. Left: cumulative size of D4Z4-l elements. Right: cumulative size of full D4Z4-l elements. Data is displayed in

the form of a barplot. Values are on 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 chr 14, chr 15-21, chr 22-1 and chr 22-2.

Figure 3. Characterization of D4Z4 alleles

(A) Heatmap of ter-*DUX4*. Scaled identity levels are displayed by the green color gradient on the right. The dendrogram on the left shows clustering of ter-*DUX4* 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 ter-*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 ter-*DUX4*. Data are represented in the form of a histogram. Distances (in Kb) are indicated on the x-axis. qA and qB elements are marked with different colors (see legend). (E-G) 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-l elements in H) qA and qB haplotypes; I) 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 Figure 2.

Figure 4. 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: each row represents the CpG methylation pattern of a single individual. Columns represent the methylation pattern of each analyzed CpG. The dendrogram shows the clustering of the subjects based on the observed methylation profiles. Colored vertical bars are used to display CCEF clinical category status, 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, restricted to CpGs in regions GS1 and GS6, in subjects with/without deleterious genetic variants in *SMCHD1*.

Figure 5. 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.

Figure 6. 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* variants, VAR= subject carrying a damaging *SMCHD1* variants.

Table titles and legends

Table 1. 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 (Kb): total size of the regions in Kb. 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. Italic characters indicate the selected sequence similarity threshold.

Table 2. 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

Table 3- Comparison between PGR analysis results and previous knowledge

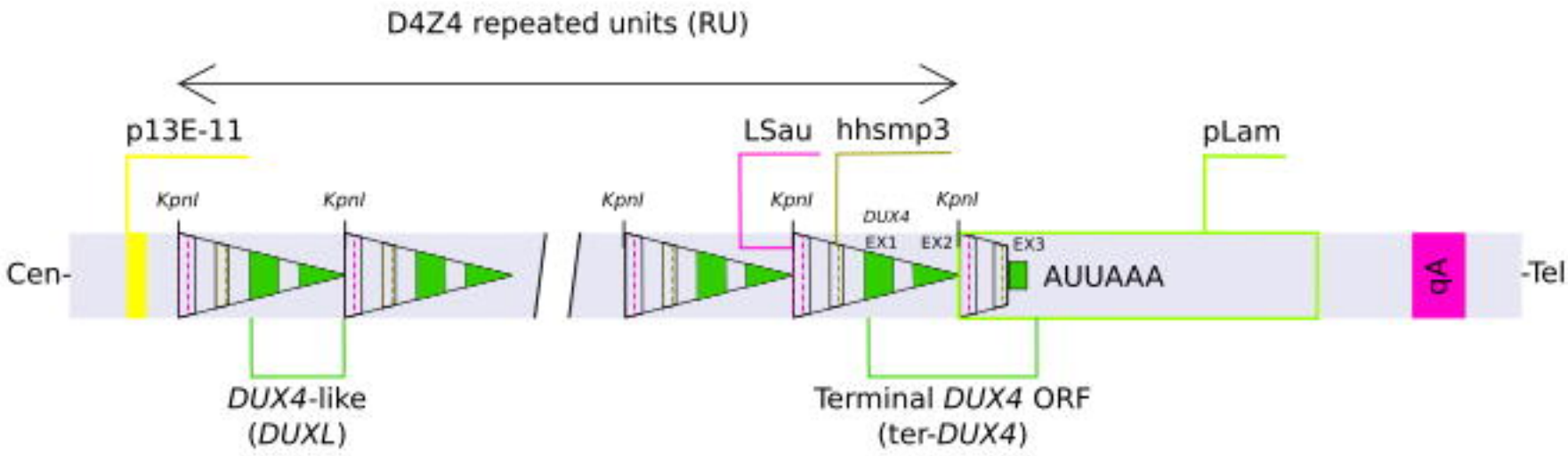
Haplotype	PGR analysis	Previous knowledge
	- 6 to 89 RU alleles; - The majority of alleles classified as qA are associated with a pLam both on 4q and 10q; - AUUAAA PAS at 4q; - 4.6% individuals carry a 4qA-PAS alleles with <8RU; - qA sequence is not always at the same distance from the last D4Z4 repeat (from 4.7 Kb to 8.3 Kb).	- 6 to ~90 RU alleles[40]; - qA alleles are associated with a pLam both on 4q and 10q[18]; - AUUAAA PAS at 4q[64]; - 3% healthy population carry a DRA and 1.3% a DRA with 4qA-PAS haplotype[39].
	- 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[64]; - 10q is associated predominantly to qA haplotype[19]; - sporadic cases of 10qB alleles[62].

Table 4: GS containing CpGs covered by at least 10 BSC reads in at least 50% of samples and no missing data across all samples

Assembly	Group	Chromosome	N° CpGs
T2T	GS1	Chr 4, Chr 10	28
	GS2		32
	GS5	Chr 13, Chr 15	16
	GS6		5
hg38	GS3	Chr 4, Chr 10	26

FSHD

Prevalence 1:8.000 - 1:20.000



FSHD 1
95% cases

FSHD 2
5% cases

D4Z4 reduced allele (DRA)
≤10 RU
+
4qA-PAS haplotype

>10 RU
+
4qA-PAS haplotype
+
Variants in chromatin remodeling genes
(SMCHD1/DNMT3B/LRIF-1)

↓

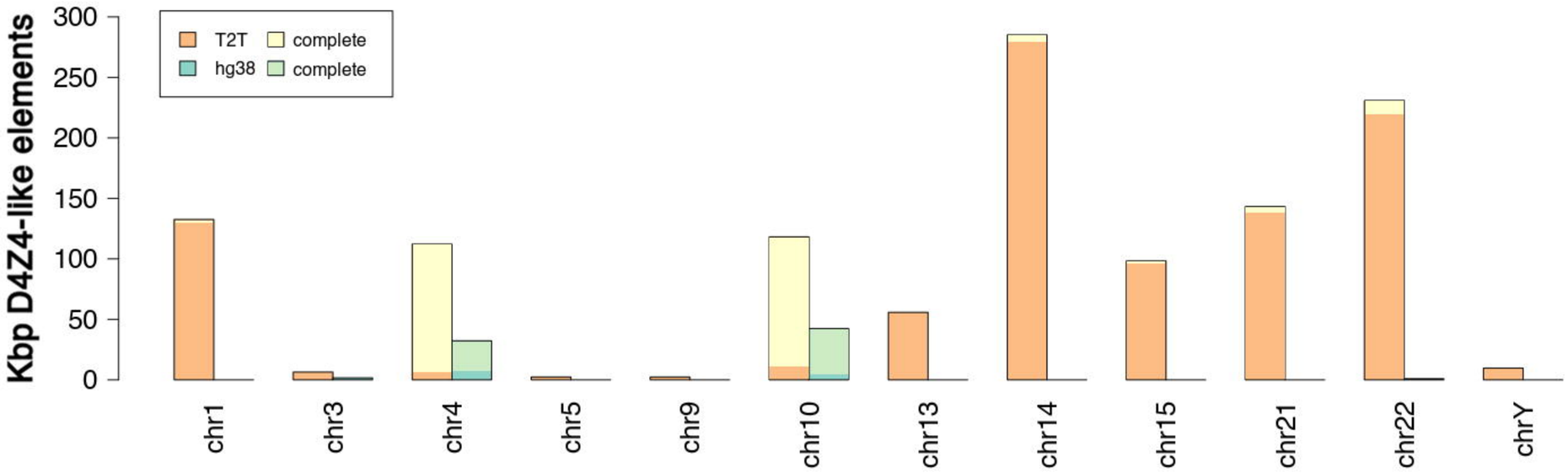
Terminal D4Z4
reduced methylation

D4Z4 repeats
reduced methylation

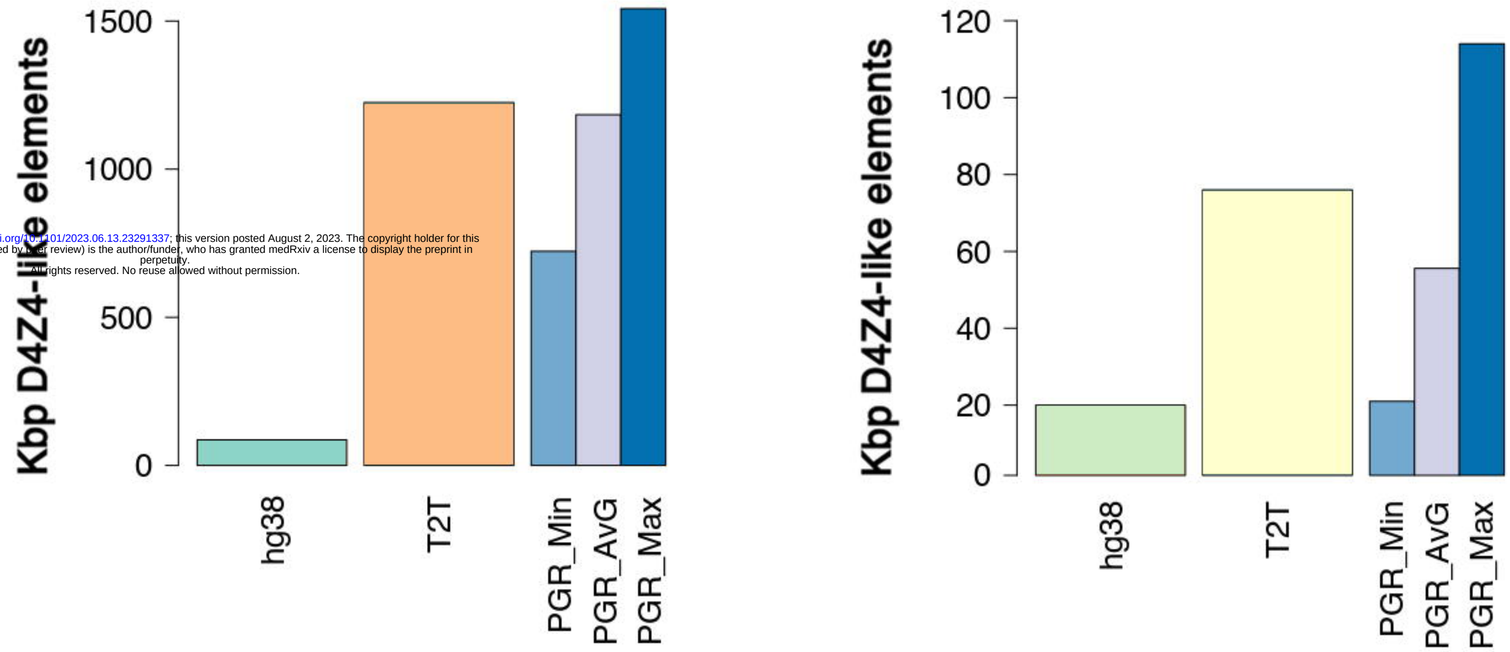
↓

Terminal DUX4 ORF (ter-DUX4) toxic expression

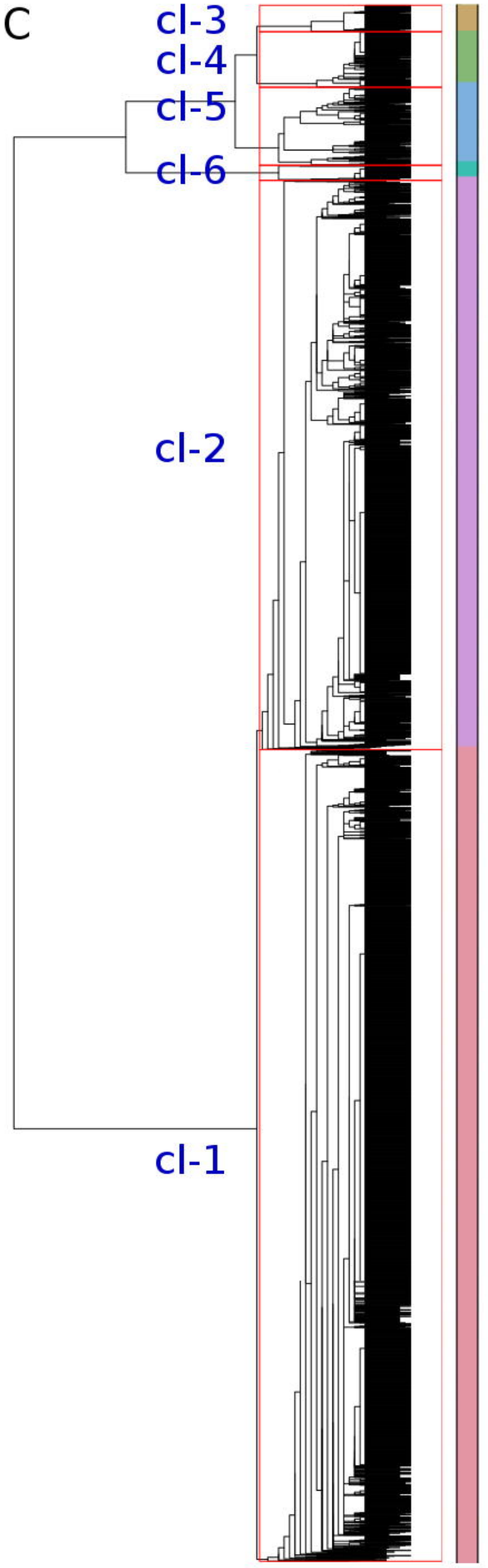
A



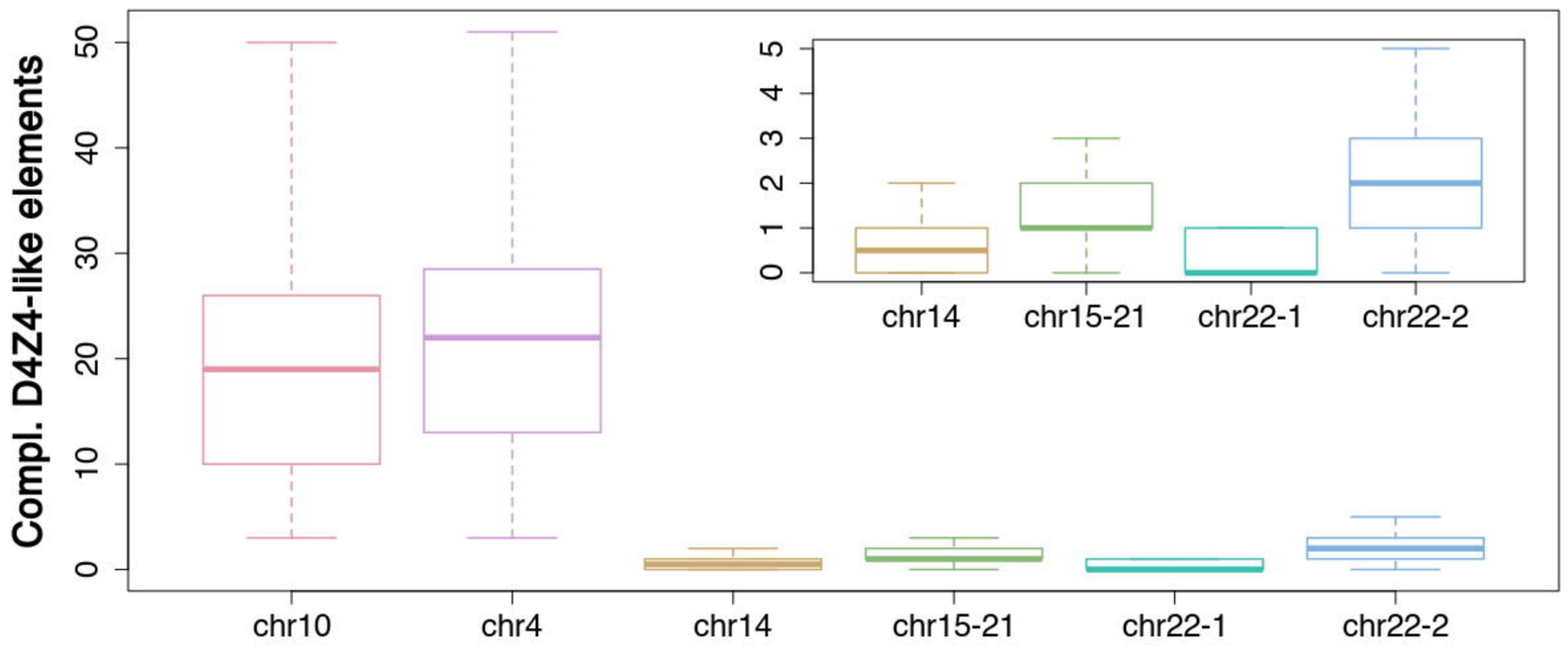
B

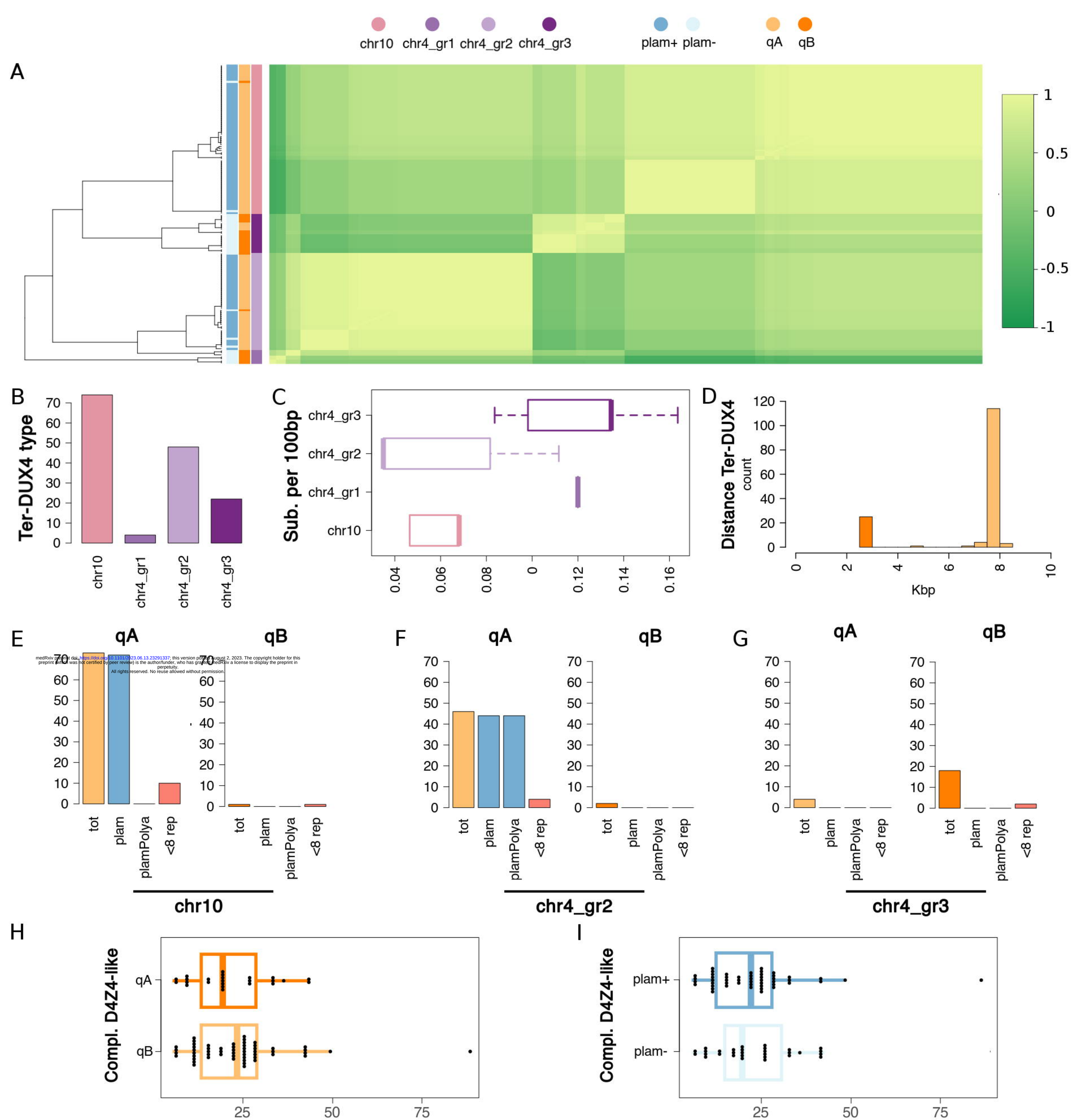


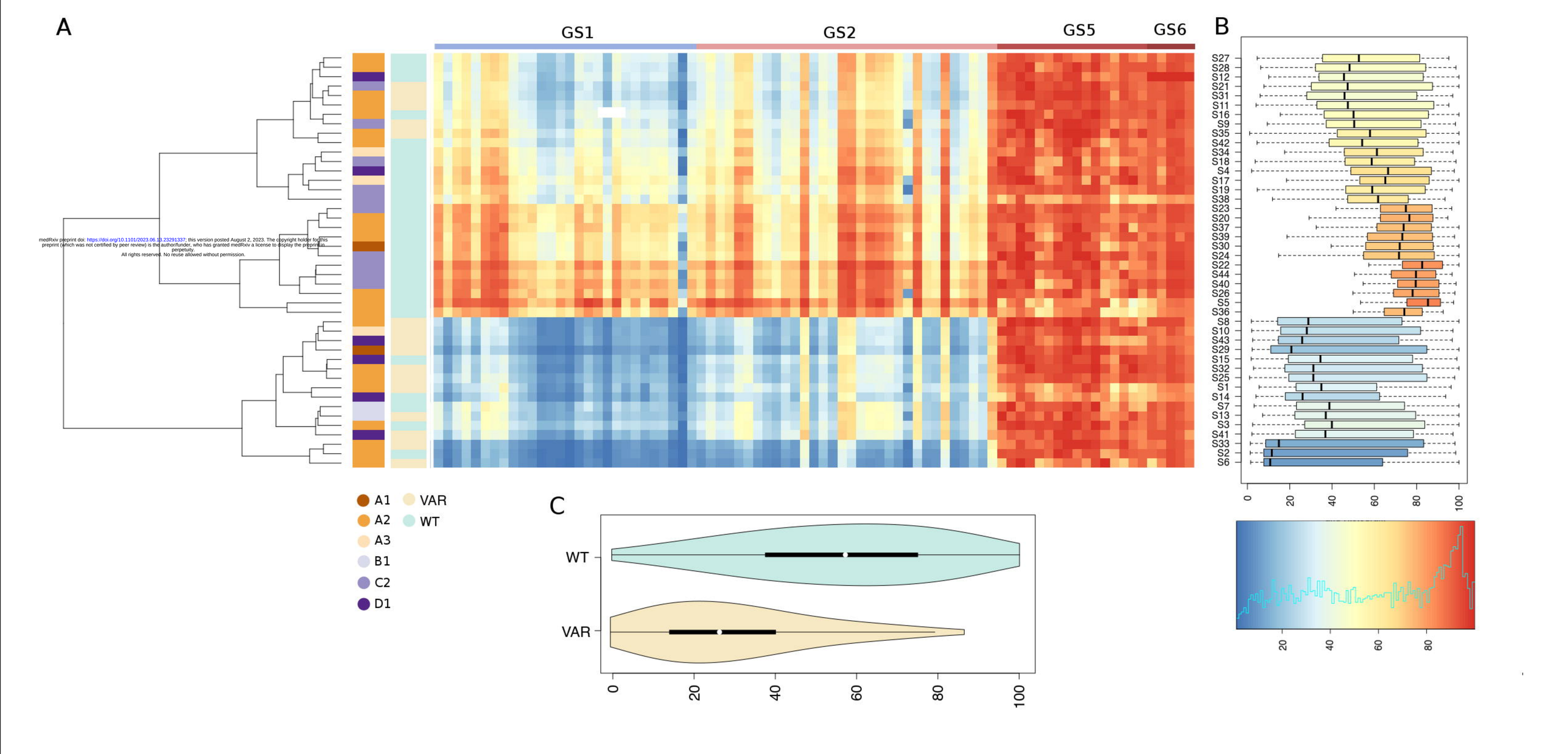
C

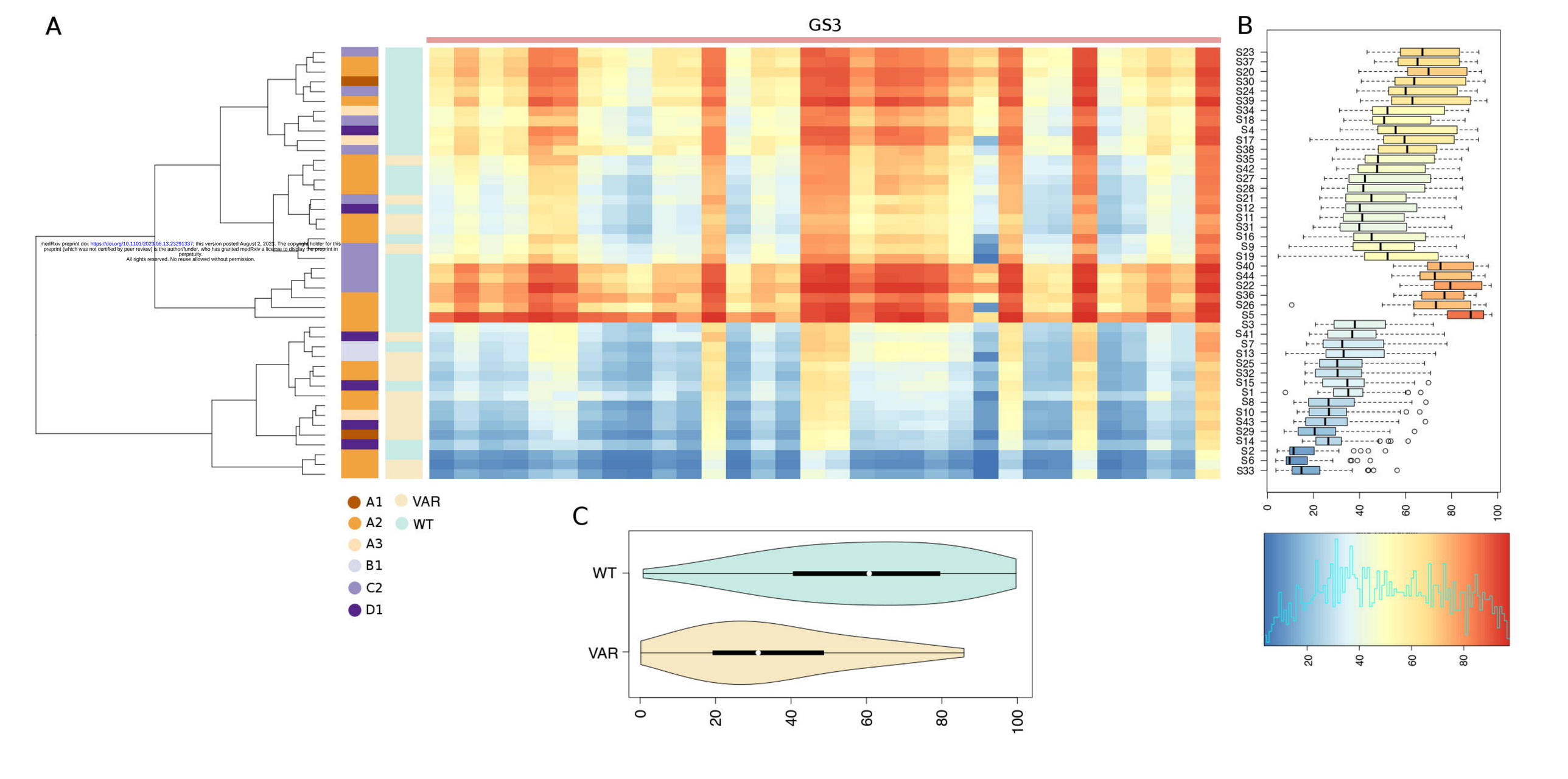


D

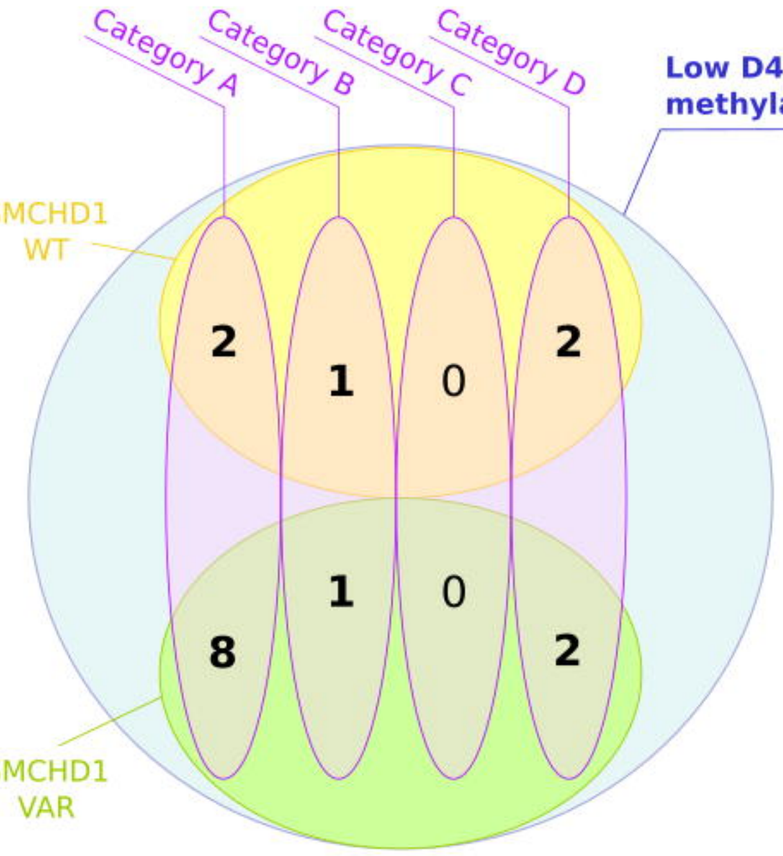


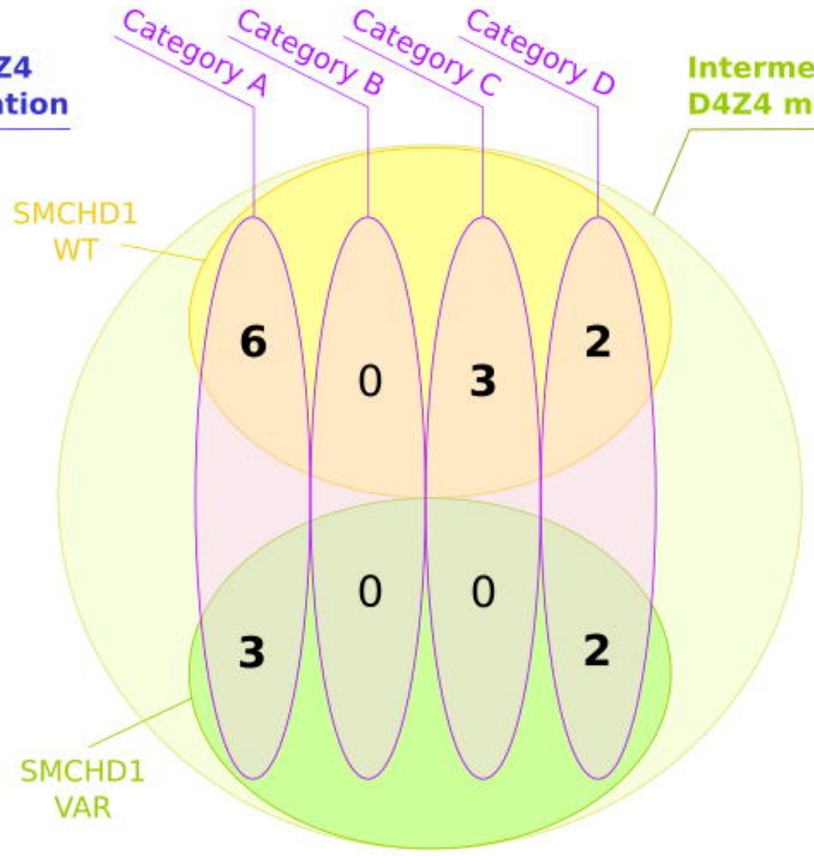


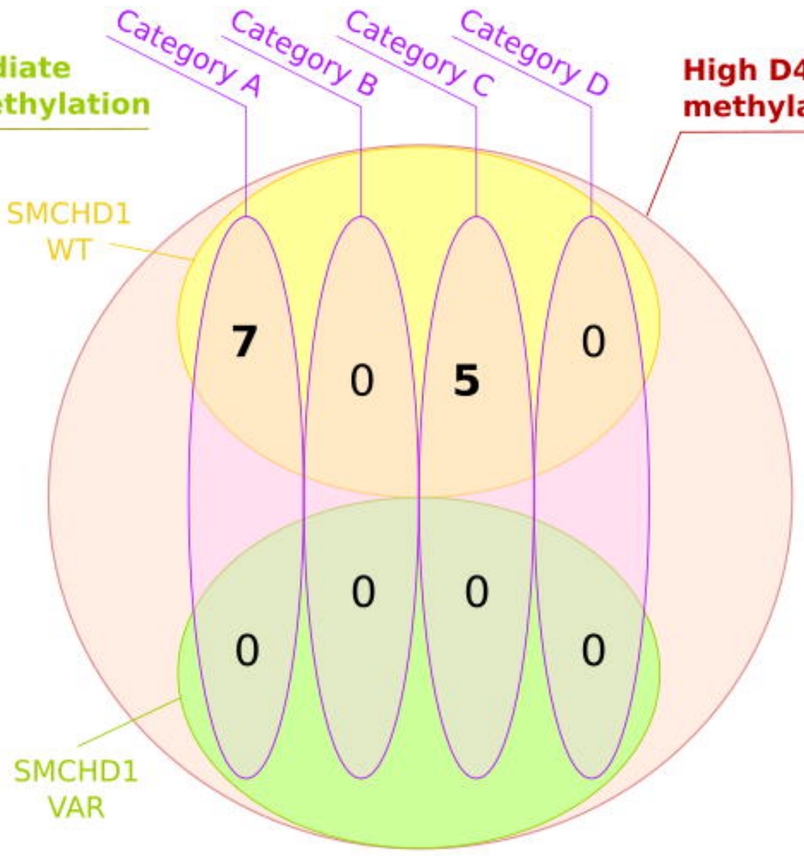




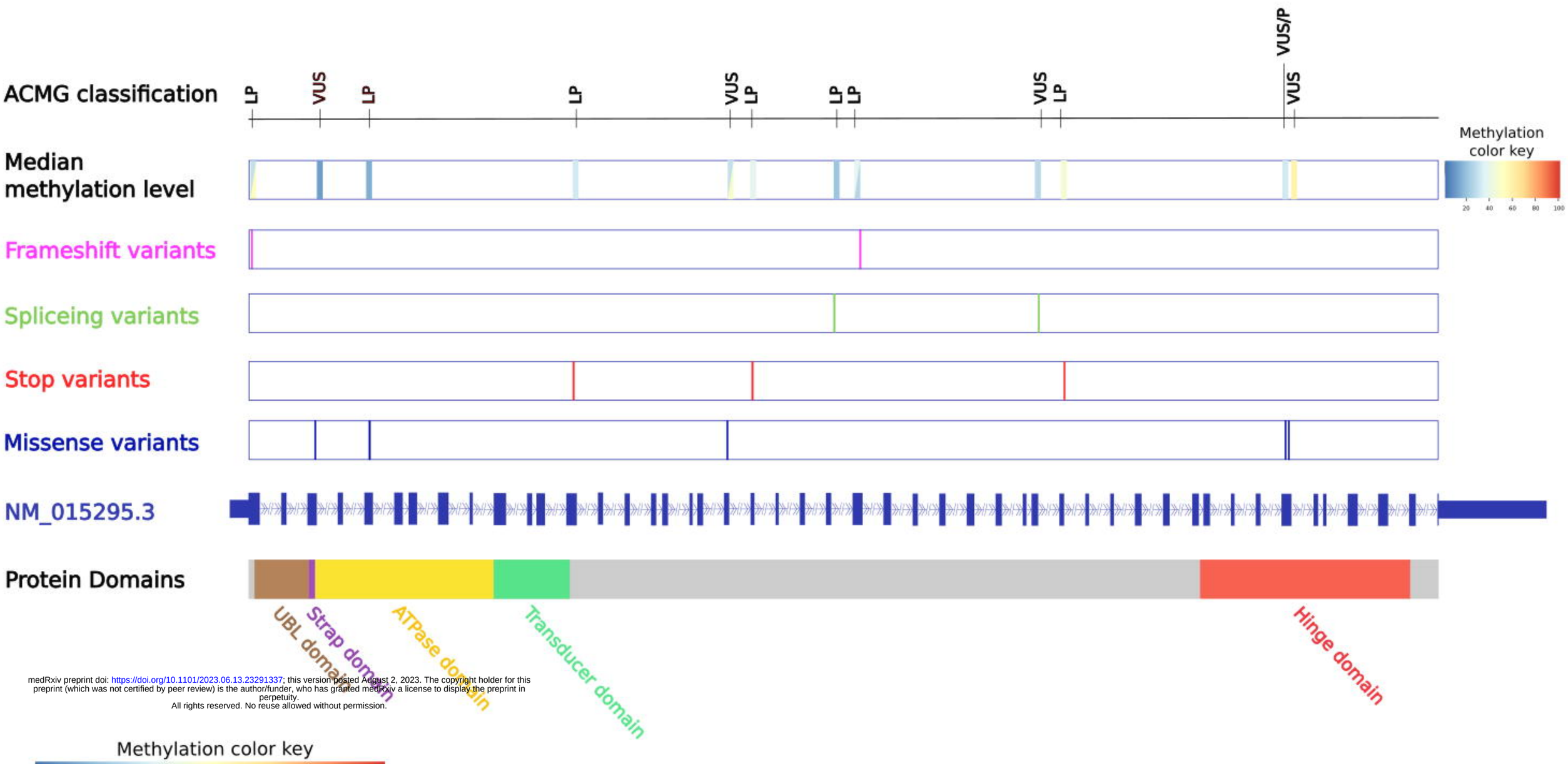
A



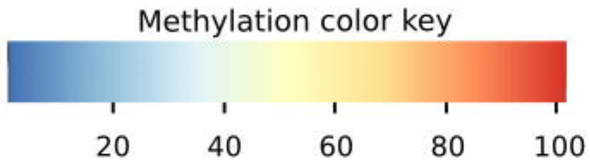


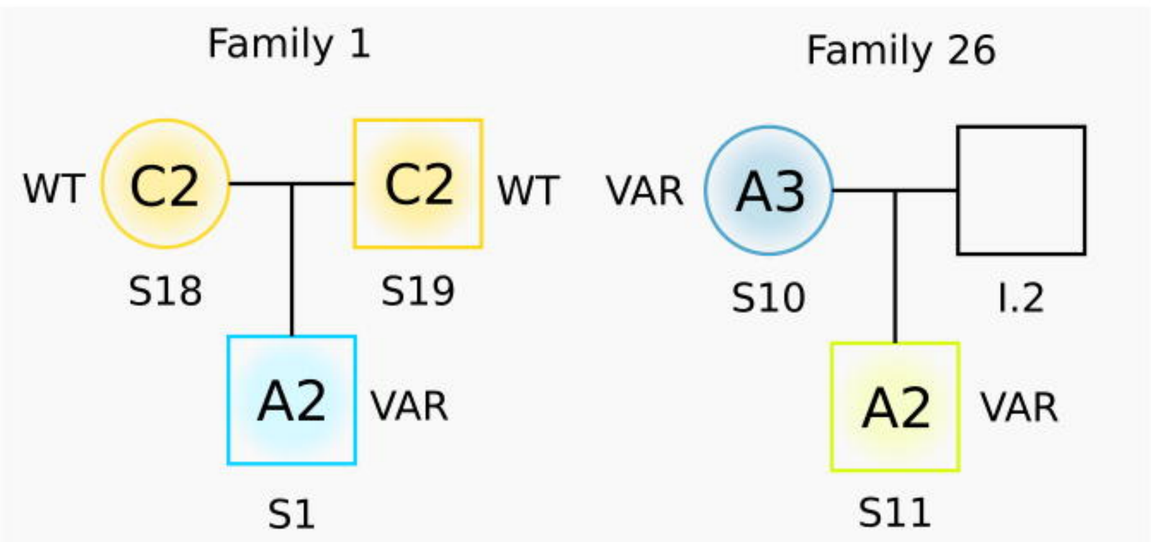


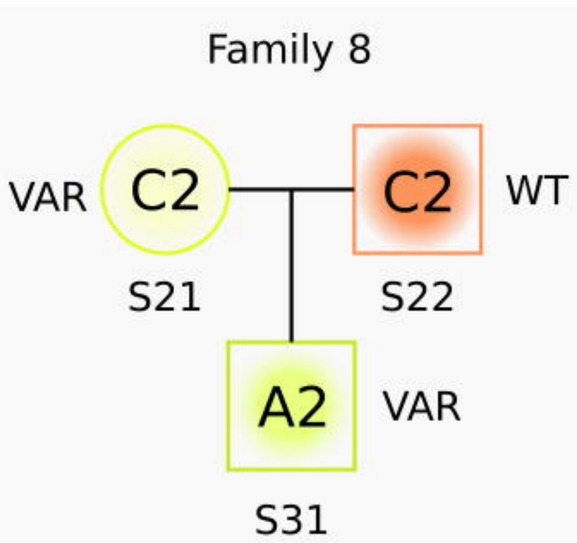
B

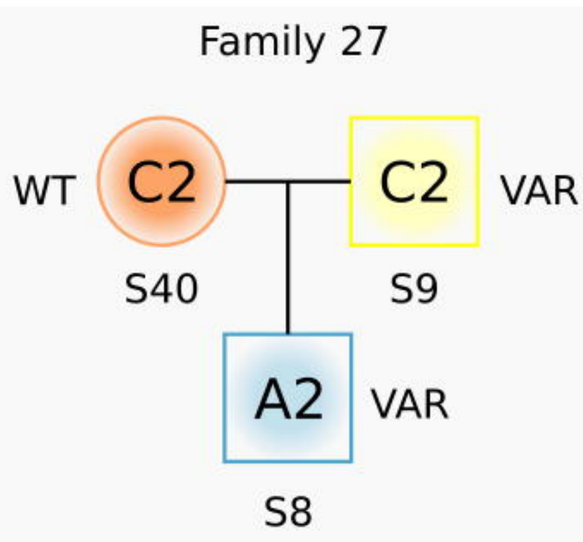


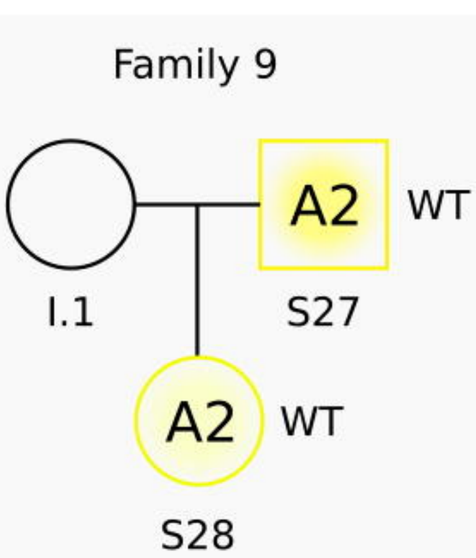
C

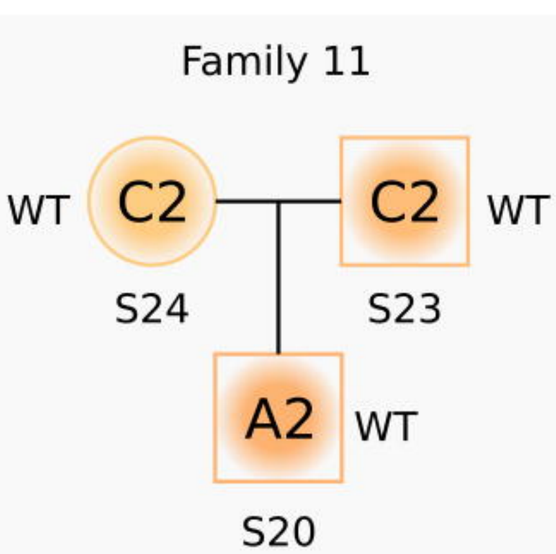


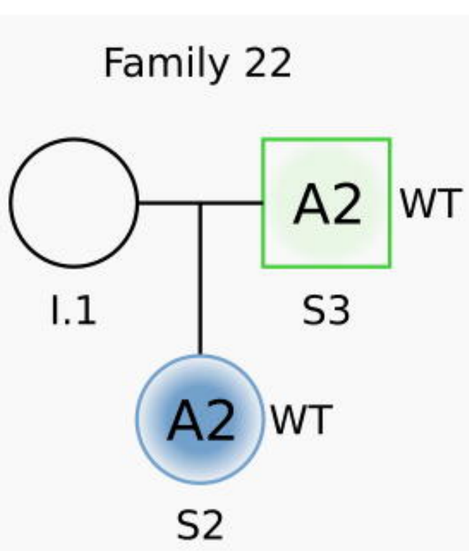


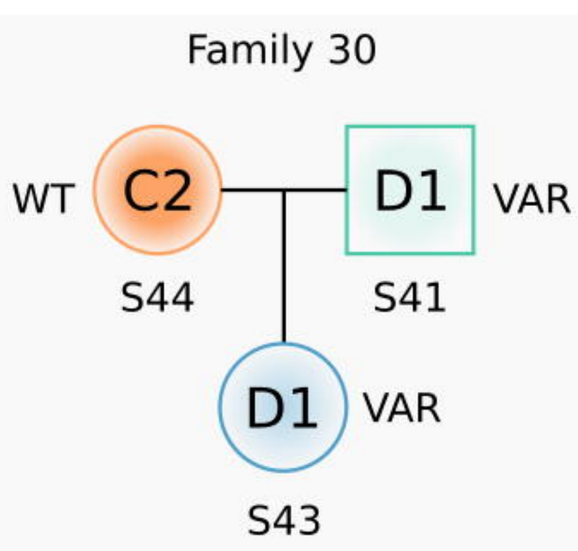












GRCh38 — T2T-CHM13 — Human Pangenome Project

Genetic variability

3.3 Kb

86.5 Kb
D4Z4-like

10

4

DRA

No DRA

+4qA

+*SMCHD1*
variants

Reduced methylation

Transcriptional derepression

FSHD diagnosys
(Molecular)

?

Phenotypic variability

300 bp → 3.3 Kb

0.7 - 1.5 Mb
D4Z4-like

1

10

13

14

15

21

22

4

DRA

No DRA

4.6%

FSHD diagnosys
(Clnical)

Other factors

+4qA

+/-*SMCHD1*
variants

Transcriptional derepression

Variable methylation