

# Complete genetic and epigenetic architecture of D4Z4 macrosatellites in FSHD, BAMS, and reference cohorts with D4Z4End2End

Lucinda C. Xiao,<sup>1</sup> Ayush Semwal,<sup>1</sup> Brianna St John,<sup>1</sup> Kathleen Zeglinski,<sup>1</sup> Shian Su,<sup>1,2</sup> James Lancaster,<sup>1</sup> Shifeng Xue,<sup>3</sup> Bruno Reversade,<sup>4</sup> Matthew E. Ritchie,<sup>1,2</sup> Frédérique Magdinier,<sup>5</sup> Marnie E. Blewitt,<sup>1,2</sup> and Quentin Gouil<sup>1,2,6,7</sup>

<sup>1</sup>The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia; <sup>2</sup>Department of Medical Biology, The University of Melbourne, Parkville, VIC 3052, Australia; <sup>3</sup>Department of Biological Sciences, National University of Singapore, 117558, Singapore; <sup>4</sup>Laboratory of Human Genetics and Therapeutics, King Abdullah University of Science and Technology, Thuwal 23955, Kingdom of Saudi Arabia; <sup>5</sup>Marseille Medical Genetics, Aix-Marseille University, INSERM, Marseille 13005, France; <sup>6</sup>Olivia Newton-John Cancer Research Institute, Heidelberg, VIC 3084, Australia; <sup>7</sup>School of Cancer Medicine, La Trobe University, Bundoora, VIC 3086, Australia

The D4Z4 locus is a macrosatellite array on Chromosome 4q normally comprising 8 to >100 3.3-kb repeat units. Its size and repetitiveness render it refractory to most sequencing technologies; consequently, its genetic and epigenetic architectures remain incompletely understood despite their relevance to facioscapulohumeral muscular dystrophy (FSHD). Current FSHD molecular testing relies on complex, multistep and low-resolution assays, which aim to identify contractions on permissive haplotypes (FSHD type 1) or epigenetic reactivation due to pathogenic variants in the epigenetic machinery, most often in *SMCHD1* (FSHD type 2). Recent guideline updates highlight the need for more accurate and comprehensive diagnostic approaches. Here, we leverage ultra-long whole-genome and Cas9-targeted sequencing to develop a fast and accurate workflow, D4Z4End2End, for comprehensive genetic and methylation analysis of D4Z4 alleles. We apply it to samples from two controls, four FSHD1 patients, four FSHD2 patients, and two patients with Bosma arhinia microphthalmia syndrome (BAMS) caused by *SMCHD1* variants, as well as publicly available data from 30 B-lymphoblastoid cell lines from the 1000 Genomes Project and Human Pangenome Reference Consortium. We attain high-depth sequencing of full-length D4Z4 arrays of up to 40 repeat units (~132 kb), accurately capture contracted arrays, genetic mosaicism, and pathogenic *SMCHD1* variants, and generate consensus sequences of all D4Z4 alleles. We identify new allelic variants, analyze complex D4Z4 rearrangements including in-cis duplications, and reveal length- and *SMCHD1*-dependent methylation patterns across the D4Z4 array. Our findings offer insights into D4Z4 genetics and epigenetics, and demonstrate the potential of long-read nanopore sequencing to accelerate FSHD research and diagnostics.

[Supplemental material is available for this article.]

Facioscapulohumeral muscular dystrophy (FSHD) is a genetic disorder that causes progressive weakness of the muscles of the face, scapula and upper arms, as well as the lower legs, hip girdle, and abdomen (Preston and Wang 2025). It is the third most common muscular dystrophy worldwide, with an estimated prevalence of between 1/8000 and 20,000 (Mostacciuolo et al. 2009; Deenen et al. 2014; Wang et al. 2022).

FSHD has a complex etiology involving genetic and epigenetic dysregulation of the D4Z4 macrosatellite array in the subtelomere of Chromosome 4 (4q35). The D4Z4 array comprises 1 to >100 3.3-kb tandem repeats, each of which contains a partial copy of the *DUX4* gene, with a full-length *DUX4* gene being present at the distal end of the array (Fig. 1A). *DUX4* encodes a transcription factor that is normally expressed in a small window of embryonic development, where it plays a key role in zygotic genome activation (De Iaco et al. 2017), but is thereafter silenced

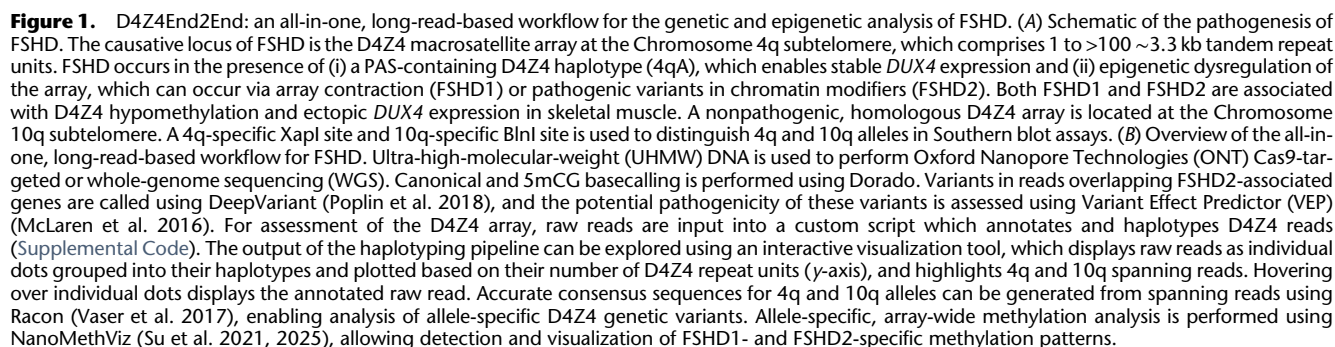
in most somatic tissues (Mocciaro et al. 2021). FSHD arises when loss of silencing at D4Z4 leads to aberrant expression of *DUX4* in skeletal muscle, and subsequent activation of myotoxic signaling pathways (Campbell et al. 2018; Duranti and Villa 2023).

Epigenetic derepression of D4Z4 can occur through two different mechanisms, which define the two main subtypes of FSHD (FSHD1 and FSHD2) (Fig. 1A) (Himeda and Jones 2019; Schätzl et al. 2021; Caputo et al. 2022b). FSHD1 accounts for the majority of cases (~95%), and is caused by contraction of the D4Z4 array to 1–10 repeat units, which can either be inherited in an autosomal dominant manner or arise de novo (10%–30% of cases) (van der Maarel et al. 2000; Lemmers et al. 2004). FSHD2 (~5% of cases) is caused by pathogenic variants in chromatin modifiers including *SMCHD1* (Lemmers et al. 2012), *DNMT3B* (van den Boogaard et al. 2016), and *LRIF1* (Hamanaka et al. 2020) in the presence of an intermediate-length array of commonly 8–20 repeat units (Giardina et al. 2024) or cis-duplicated arrays with short distal

**Corresponding author:** [quentin.gouil@onjcri.org.au](mailto:quentin.gouil@onjcri.org.au)

Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.280907.125>. Freely available online through the *Genome Research* Open Access option.

© 2026 Xiao et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.



arrays (Lemmers et al. 2018a). Noncontracted D4Z4 arrays generally have high levels of DNA methylation, whereas both FSHD1 and FSHD2 are associated with hypomethylation of either the contracted array (FSHD1) or all D4Z4 alleles (FSHD2) (de Greef et al. 2009; De Greef et al. 2010; Gould et al. 2021; Erdmann et al. 2023).

Further factors complicate the genetics of FSHD (Fig. 1A). Two main haplotypes of the 4q D4Z4 allele have been identified, designated 4qA and 4qB, based on differences in the sequence immediately distal to the D4Z4 array (Lemmers et al. 2002). Only the 4qA haplotype contains a polyadenylation signal (PAS) required for stabilization of the *DUX4* mRNA, and thus only 4qA alleles are pathogenic (Lemmers et al. 2010a). Moreover, there is a homologous qA-type D4Z4 array on Chromosome 10q that shares 98% identity with the 4q D4Z4 array (Bakker et al. 1995; Cacurri et al. 1998; Lemmers et al. 2002), yet 10qA alleles contain a single-nucleotide variant (SNV) within the PAS that renders them nonpathogenic (Lemmers et al. 2010a, 2022). Further D4Z4 subhaplotypes have been identified based on other genetic variants within and adjacent to the D4Z4 array (Lemmers et al. 2010b), including two variants of the 4qA allele (4qAS and 4qAL) that differ in the lengths of their distal D4Z4 units, but have been found to be equally pathogenic (Lemmers et al. 2018b). These factors should be accounted for to provide accurate diagnoses for FSHD.

Although previous studies have done much to advance our understanding of FSHD, many aspects of D4Z4 genetics and epigenetics remain poorly understood. It is still unclear the exact mechanisms by which epigenetic factors normally interact with D4Z4 repeats to mediate silencing of the array. Moreover, even amongst patients with the same number of D4Z4 repeats, there is great variability in onset, severity and penetrance of FSHD (Goto et al. 2004; Salort-Campana et al. 2015; Mul et al. 2018b; Ruggiero et al. 2020), suggesting the existence of as-yet unresolved modifiers of the disease. A prominent example of this is the “gray zone” in the repeat number threshold for FSHD1, where 4qA alleles with 8–10 repeat units can give rise to FSHD, but are also found in unaffected control individuals (Giardina et al. 2024). Additionally, Bosma arhinia microphthalmia syndrome (BAMS), another phenotypically distinct human disease caused by pathogenic *SMCHD1* variants, has also been associated with D4Z4 hypomethylation (Shaw et al. 2017), yet most BAMS patients do not have the skeletal muscle manifestations of FSHD (Gordon et al. 2017; Mul et al. 2018a). These outstanding questions all highlight the need for methods that can more comprehensively characterize the genetics and epigenetics of D4Z4 regulation.

Until recently, study of D4Z4 has been limited by the low resolution and/or low read length of prior techniques. Southern blotting has long been the “gold standard” for FSHD molecular testing, and uses two separate sets of restriction enzyme digestions and probe hybridizations to determine (a) the number of D4Z4 units in 4q alleles (distinguished from 10q alleles by haplotype-specific SNVs that create 4q-specific *XapI* and 10q-specific *BlnI* restriction sites), and (b) the A/B haplotype. Minimally, a typical diagnostic pipeline for FSHD, as outlined in the latest guideline recommendations (Giardina et al. 2024), may involve clinical assessment followed by a Southern blot assay to count the D4Z4 repeat number of 4q alleles. If this first assay shows that the patient does not have a 4q allele of 1–7 repeat units (strongly suggestive of FSHD1), additional assays may be required to determine A/B haplotype (additional Southern blotting assay), evaluate FSHD2-associated genes (exome sequencing), or assess for D4Z4 hypomethylation (bisulfite sequencing). Using this diagnostic approach, it can take months to years for a patient to receive a definitive FSHD

diagnosis. Moreover, more complex cases involving mosaicism and complex D4Z4 alleles, including rare 4qA haplotypes, D4Z4 duplications, hybrid arrays, D4Z4 proximal extended deletion (DPED) alleles, and 4q/10q translocations, can give rise to false positives or false negatives, or are often unable to be resolved (Giardina et al. 2024).

Newer technologies such as molecular combing and optical genome mapping have enabled more in-depth characterization of complex D4Z4 alleles that are missed by traditional technologies (Nguyen et al. 2011; Vasale et al. 2015; Nguyen et al. 2017), and are recognised as potential first-line alternatives to Southern blotting for repeat number counting and A/B haplotyping (Giardina et al. 2024). Nevertheless, these assays are still limited by their inability to resolve the base-level D4Z4 sequence, and do not provide methylation information. Several groups have also developed methylation assays that have been shown to accurately diagnose FSHD1 and FSHD2. These methods include a Southern blot-based assay that assesses methylation at the CpG methylation-sensitive *FseI* restriction enzyme site in the proximal D4Z4 repeat (van Overveld et al. 2003; Lemmers et al. 2015); a methylated DNA immunoprecipitation (MeDIP) method targeting three regions (5', mid, 3') within the D4Z4 repeat (Gaillard et al. 2014); and more recent bisulfite-sequencing-based assays that target the *DUX4* 5'/DR1 region (Hartweck et al. 2013; Jones et al. 2014), the final D4Z4 repeat and the region immediately distal to it for 4qA/PAS-positive alleles (Jones et al. 2014; Calandra et al. 2016), or highly informative individual CpG sites within the *DUX4*-PAS and DR1 region (Caputo et al. 2022a; Straffella et al. 2024). However, these assays either average the methylation signal of specific D4Z4 subregions/CpGs across all 4q and 10q repeat units, or specifically target the proximal or distal regions of the array, limiting the study of array-wide D4Z4 methylation patterns.

More recently, several studies have shown the potential of long-read Oxford Nanopore Technologies (ONT) sequencing for the study of the D4Z4 macrosatellite array (Hiramuki et al. 2022; Butterfield et al. 2023; Lemmers et al. 2023; Yeetong et al. 2023; Huang et al. 2024; Li et al. 2024; Wang et al. 2024). Given ONT sequencing is able to simultaneously reveal the genetic sequence and methylation status of the D4Z4 region, the recent guideline updates (Giardina et al. 2024) have identified it as a promising option for future FSHD diagnostics. In this study, we set out to develop an all-in-one, long-read-based workflow for the genetic and epigenetic analysis of FSHD (D4Z4End2End), combining targeted and whole-genome ultra-long sequencing to capture full-length 4q and 10q D4Z4 arrays as well as a panel of FSHD2-associated genes, followed by read-level annotation and haplotyping. We apply D4Z4End2End to the study of several FSHD1, FSHD2, BAMS, and control samples to assess its ability to resolve the genetics and methylation of complete sets of patient 4q and 10q alleles.

## Results

### Haplotyping and annotation of 4q/10q long-read sequencing data to characterize D4Z4 alleles

We extend upon previous studies (Supplemental Table S1) by developing an all-in-one, long-read-based workflow for the genetic and epigenetic analysis of FSHD: D4Z4End2End. D4Z4End2End was designed to process and visualise long-read sequencing data covering the 4q35 and 10q26 regions, including assignment to Chromosome 4q/10q, determination of A/B haplotype, counting

of D4Z4 repeat units, and highlighting of reads that span the entire D4Z4 array (Fig. 1B). Reads that overlap the D4Z4 array were identified by aligning the raw reads against the sequence of a single D4Z4 repeat unit. This avoids potential biases of using a specific genome reference. Alignment of D4Z4-containing reads to the 4q and 10q reference sequences from CHM13v2.0 (containing 4qA and 10qA, Nurk et al. 2022) assigns reads to their chromosome of origin. In addition, XapI and BlnI restriction sites, used to identify each chromosome in current diagnostic approaches, were also assessed within the raw reads, to confirm correct assignment to 4q or 10q. We further assigned the reads to A and B haplotypes based on the presence of pLAM (qA-specific sequence) and qB-specific sequences, respectively. Reads that contained both p13E-11 (a region flanking the proximal end of the D4Z4 array) and either pLAM or qB-specific sequence were classified as spanning reads, and these spanning reads were used to determine the haplotypes and number of D4Z4 units for each allele. To aid in visualization of the results, we created an interactive tool that can be used to explore individual raw 4q and 10q reads that have been processed by the pipeline and annotated with their features (p13E-11, D4Z4 units, pLAM, qB-specific sequence, and XapI and BlnI restriction sites) and haplotypes (4q/10q and A/B) (Figs. 1B, 2).

We first used the pipeline to analyze publicly available raw ONT data and assess assemblies published by the Human Pangenome Reference Consortium (HPRC) (Liao et al. 2023), which have previously been studied in the context of FSHD (Salsi et al. 2023). The pipeline was able to successfully identify and annotate raw reads that spanned complete D4Z4 arrays, enabling accurate D4Z4 counting and resolution of 4q and 10q alleles (Supplemental Table S2). Despite the presence of raw spanning reads in the ONT data, assessment of the associated HPRC assemblies showed many of the D4Z4 arrays to be incomplete or split across multiple scaffolds (Supplemental Table S2), creating difficulties for accurate D4Z4 counting and haplotyping. Thus, based on evidence from the raw reads, we find different D4Z4 repeat numbers to those reported previously (Salsi et al. 2023). Additionally, we used the pipeline to process the publicly available raw ONT data used for the CHM13v2.0 assembly (Nurk et al. 2022), and found the number of repeat units in CHM13v2.0's assembly of the 10q array (33 full 3.3 kb units) to be inconsistent with the raw ONT reads (32 full 3.3 kb units). This shows that there are still outstanding issues in the automated assembly of macrosatellite repeat regions, and demonstrates the need for a tool to visualise the repeat structure of raw reads.

We then used the pipeline to analyze ONT ultra-long whole-genome sequencing (WGS) data from the fibroblasts of two healthy controls, three patients diagnosed with FSHD1, three patients diagnosed with FSHD2, and two patients diagnosed with BAMS (Table 1). The D4Z4 haplotypes of the FSHD1 and FSHD2 patients have previously been studied using Southern blotting and/or molecular combing (Dion et al. 2019; Gaillard et al. 2019; Laberthonnière et al. 2022; Delourme et al. 2023; Gérard et al. 2025), with one of the FSHD1 patients exhibiting somatic mosaicism for a pathogenic 4qA 2RU allele (25% based on Southern blot) and another of the FSHD1 samples harboring a 10qA *cis* triplication allele (Delourme et al. 2023) (Supplemental Table S3). The D4Z4 haplotypes of the controls and BAMS patients have not previously been studied.

The pipeline was able to successfully highlight and haplotype 4q and 10q reads from the patient fibroblasts (Table 1), in concordance with the results from Southern blot and molecular combing (Supplemental Table S3). Repeat unit counts based on spanning

reads occasionally departed moderately from previous length-based estimates from Southern blot and molecular combing. WGS captured spanning reads for 30 out of 41 alleles, with a maximum of six spanning reads being obtained for a single allele. We captured spanning reads corresponding to the pathogenic, contracted allele for two of the FSHD1 samples, but not for the mosaic FSHD1 sample, perhaps due to the contracted allele having a lower mosaic proportion (25%) (Fig. 2A). The interactive visualization tool also enabled assessment of the length of the distal D4Z4 unit, and we detected both the previously reported 4qAS (~0.3 kb) and 4qAL (~1.9 kb) haplotypes. We additionally identified a new, previously unreported 4qA haplotype with a distal D4Z4 unit of length ~0.6 kb, present in the BAMS1 sample, which we designate “4qAM” (Fig. 2B). All 10qA alleles contained a distal D4Z4 unit of length ~0.3 kb, similar to 4qAS alleles.

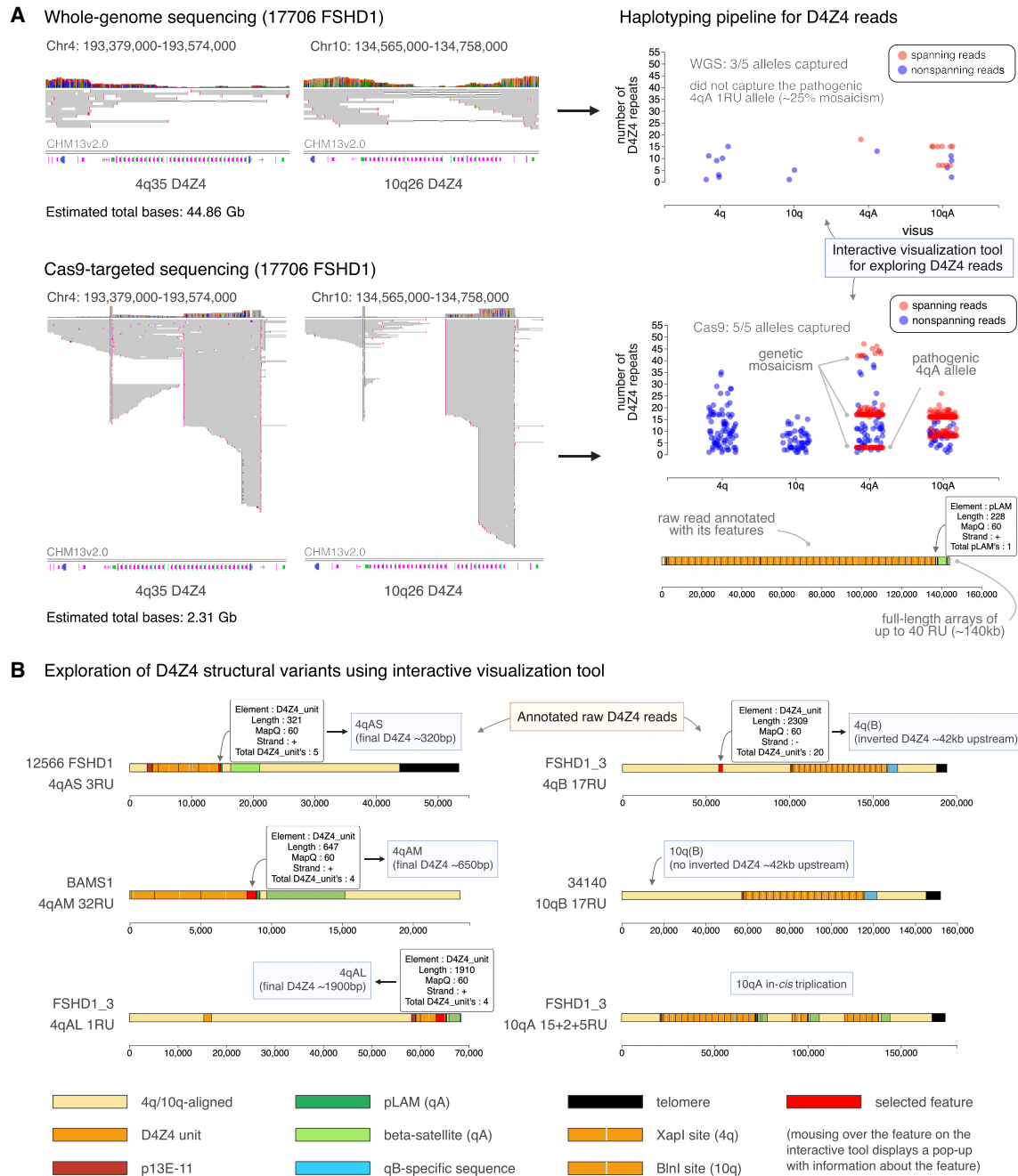
Ultra-long read sequencing also enabled assessment of extended regions of 4q and 10q sequence upstream of and downstream from the D4Z4 array (Fig. 2B). We use the terms “upstream” and “downstream” to refer to genomic regions that are closer to the centromere or telomere relative to the D4Z4 array, respectively. For eight of the 4q alleles, reads were obtained that captured the inverted D4Z4 unit (D4S2463) containing *DUX4c* that is found ~42 kb upstream of the 4q D4Z4 array (Wright et al. 1993; Anseau et al. 2009) (Table 1, Fig. 2B). For three of the samples (12,566, 15,166, and 34,140), 10q alleles were identified that belonged to B haplotypes and contained XapI restriction sites, indicating possible 4q-to-10q translocations upstream of the array. The region of homology between 4q and 10q extends ~42 kb upstream of the D4Z4 array (Deidda et al. 1995; Delourme et al. 2023), and these alleles were able to be confidently assigned to Chromosome 10q as the ultra-long reads spanned >42 kb of upstream sequence (Fig. 2B). Moreover, the pipeline identified reads that spanned all three arrays of the 10qA *in-cis* triplication allele of sample FSHD1\_3 (37-I<sub>1</sub> in Delourme et al. 2023), revealing the allele's complex structure with nucleotide resolution (Fig. 2B). As has been previously suggested, translocated or duplicated alleles with intact *DUX4* and PAS sequences may give rise to *DUX4* expression (Lemmers et al. 2018a, 2022), however these complex alleles cannot always be accurately resolved by Southern blotting and molecular combing (Delourme et al. 2023; Giardina et al. 2024). This demonstrates that our pipeline can comprehensively characterize 4q and 10q D4Z4 alleles, including reliable identification of contracted 4qA alleles to confirm the diagnosis of FSHD1, and identification of other structural variants which may contribute to FSHD pathogenesis.

### Cas9-targeted sequencing of 4q/10q D4Z4 alleles and FSHD2-associated genes

Although ultra-long WGS is able to capture reads that span the entire 4q and 10q D4Z4 arrays with no prior assumption of the repeat composition or structure, a single PromethION flow cell per patient only provides low coverage. This can lead to pathogenic alleles being missed (Table 1, Fig. 2A), difficulty in creating an accurate consensus sequence of each allele, and in exploring cell-to-cell epigenetic variability or indeed genetic mosaicism.

To address this, we used Cas9-targeted sequencing to enrich for the 4q and 10q D4Z4 regions, aiming to capture the whole set of alleles in each patient. We designed guide RNAs targeting regions upstream of and downstream from the D4Z4 array, for both A and B alleles (Supplemental Table S4). Additionally, we complemented this with guide RNAs designed to target a panel of FSHD2-





**Figure 2.** An interactive tool for exploring annotated and haplotyped D4Z4 raw reads from long-read sequencing data. (A) Sequencing data obtained for the 4q35 and 10q26 D4Z4 regions using ultra-long whole-genome sequencing (WGS) and Cas9-targeted sequencing, for sample 17,706 (FSHD1). (Left) Alignment of sequencing data for the D4Z4 region against a single reference genome, such as CHM13v2.0 (32 repeat units for both the 4q35 and 10q26 D4Z4 arrays), leads to large gaps in alignment and potential mis-mapping of reads, making interpretation difficult. (Right) Our pipeline processes the sequencing data to annotate raw reads with their D4Z4 features, enabling accurate repeat unit counting and assignment to 4q/10q and A/B haplotypes. (Top) WGS generates a high total sequencing output (44.86 Gb for sample 17,706) but yields low coverage of the 4q35 and 10q26 regions, which can lead to pathogenic alleles being missed. (Bottom) Cas9-targeted sequencing leads to much higher coverage of the 4q35 and 10q26 regions with a much lower total sequencing output (2.31 Gb for sample 17,706), which enables the capture of low-frequency pathogenic alleles, genetic mosaicism, long full-length arrays, and high numbers of spanning reads. (B) The pipeline is accompanied by an interactive visualization tool that enables exploration of the structure of raw D4Z4 reads, revealing D4Z4 variants such as (left) 4qAS, 4qAL and a newly identified 4qAM allele, as well as (right) 10qB alleles from 4q/10q translocations, and in-cis triplication alleles.

associated genes (*SMCHD1*, *DNMT3B*, and *LRIF1*) (Supplemental Table S4), with a view to developing an all-in-one genetic and epigenetic research and diagnostic tool for FSHD. This enabled us to

obtain high-read-depth sequencing of both Chr 4q35 and Chr 10q26 (Table 1, Fig. 2A) and of *SMCHD1*, *DNMT3B*, and *LRIF1* exons, respectively (Supplemental Fig. S1).

**Table 1.** Haplotyping results and number of spanning reads for D4Z4 alleles from FSHD, BAMS, and control fibroblasts, from whole-genome and Cas9-targeted Nanopore sequencing

| SAMPLE              | 4q                |                      |                           |                            |                               | 10q       |                         |                           |                            |
|---------------------|-------------------|----------------------|---------------------------|----------------------------|-------------------------------|-----------|-------------------------|---------------------------|----------------------------|
|                     | Haplotype         | No. RUs <sup>a</sup> | No. of WGS spanning reads | No. of Cas9 spanning reads | Upstream D4S2463 <sup>b</sup> | Haplotype | No. RUs                 | No. of WGS spanning reads | No. of Cas9 spanning reads |
| Control             |                   |                      |                           |                            |                               |           |                         |                           |                            |
| AG09309             | 4qB <sup>c</sup>  | ≥38 <sup>c</sup>     | 0                         | – <sup>d</sup>             | N                             | 10qA      | 10                      | 2                         | –                          |
|                     | N/A <sup>e</sup>  | N/A                  | 0                         | –                          | N/A                           | 10qA      | 21                      | 1                         | –                          |
| AG10803             | 4qB               | 18                   | 2                         | –                          | Y                             | 10qA      | 10                      | 2                         | –                          |
|                     | 4qB               | ≥35                  | 0                         | –                          | N                             | 10qA      | 13                      | 2                         | –                          |
| FSHD1               |                   |                      |                           |                            |                               |           |                         |                           |                            |
| 12,566              | 4qAS              | 3                    | 4                         | 20                         | Y                             | 10qA      | 9                       | 4                         | 12                         |
|                     | 4qAS              | 42                   | 1                         | 0                          | N                             | 10qB      | 19                      | 2                         | 0                          |
| FSHD1_3             | 4qAL              | 1                    | 3                         | –                          | Y                             | 10qA      | 15 + 2 + 5 <sup>f</sup> | 1 + 1 + 2 <sup>g</sup>    | –                          |
|                     | 4qB               | 17                   | 2                         | –                          | Y                             | 10qA      | 35                      | 1                         | –                          |
| 17,706              | 4qAL              | 1                    | 0                         | 53                         | N                             | 10qA      | 6                       | 4                         | 139                        |
|                     | 4qAL              | 15                   | 1                         | 80                         | Y                             | 10qA      | 14                      | 6                         | 115                        |
|                     | 4qAL              | 40                   | 0                         | 12                         | N                             |           |                         |                           |                            |
| 19,187              | 4qAS              | 2                    | –                         | 118                        | –                             | 10qA      | 15                      | –                         | 33                         |
|                     | 4qAS              | 22                   | –                         | 6                          | –                             | 10qA      | 26 + 1 <sup>f</sup>     | –                         | 12                         |
|                     | 4qAM              | 32                   | –                         | 8                          | –                             |           |                         |                           |                            |
| FSHD2               |                   |                      |                           |                            |                               |           |                         |                           |                            |
| 15,166              | 4qAS              | 12                   | 4                         | –                          | N                             | 10qB      | 13                      | 2                         | –                          |
|                     | 4qB               | 21                   | 2                         | –                          | N                             | N/A       | N/A                     | 0                         | –                          |
| 34,140 <sup>h</sup> | 4qAS              | 25                   | 1                         | –                          | Y                             | 10qB      | 17                      | 3                         | –                          |
|                     | 4qAL <sup>c</sup> | ≥20 <sup>c</sup>     | 0                         | –                          | N                             | 10qA      | 19                      | 3                         | –                          |
| 11,491              | 4qAL              | 22                   | 1                         | –                          | N                             | 10qA      | 6                       | 1                         | –                          |
|                     | N/A               | N/A                  | 0                         | –                          | N/A                           | 10qA      | 7                       | 1                         | –                          |
| 11,440              | 4qAS              | 11                   | –                         | 14                         | –                             | 10qA      | 14                      | –                         | 3                          |
|                     | 4qAS              | 34                   | –                         | 4                          | –                             | 10qA      | 19                      | –                         | 4                          |
| BAMS                |                   |                      |                           |                            |                               |           |                         |                           |                            |
| BAMS-1              | 4qAM              | 32                   | 0                         | –                          | N                             | 10qA      | 12                      | 0                         | –                          |
|                     | 4qAS              | 47                   | 0                         | –                          | N                             | 10qA      | 23                      | 3                         | –                          |
| BAMS-9              | 4qB               | 13                   | 5                         | 100                        | Y                             | 10qA      | 7                       | 1                         | 192                        |
|                     | 4qB               | 26                   | 2                         | 27                         | Y                             | 10qA      | 31                      | 1                         | 9                          |

<sup>a</sup>Refers to the number of complete 3.3 kb KpnI-KpnI D4Z4 units (excluding partial D4Z4 units at the proximal and distal ends).

<sup>b</sup>Refers to whether D4S2463, an inverted and truncated D4Z4 unit found ~42 kb upstream of 4q D4Z4 arrays, is captured in spanning reads from the WGS data.

<sup>c</sup>Alleles for which no spanning reads were captured, but whose haplotypes and minimum numbers of repeats were able to be determined from non-spanning reads covering the distal end of the array.

<sup>d</sup>Dashes (–) indicate that no WGS/Cas9-targeted sequencing was performed for that sample.

<sup>e</sup>“N/A” indicates alleles whose haplotypes and repeat numbers could not be determined due to the absence of spanning reads or unique nonspanning reads.

<sup>f</sup>Alleles with in-cis duplications/triplications of the D4Z4 array.

<sup>g</sup>Number of reads for the triplication allele spanning 1, 2, and all 3 of the arrays, respectively.

<sup>h</sup>Previously diagnosed as FSHD2 based on clinical findings and absence of a contracted D4Z4 array.

For all except one of the samples assayed using Cas9-targeted sequencing (three FSHD1 samples, one FSHD2 sample, and one BAMS sample), we were able to determine the full set of 4q and 10q haplotypes by capturing at least one spanning read from each of the patient's 4q/10q alleles (Table 1). This included the capture of all three 4q alleles from the two mosaic FSHD1 patients (Table 1, Fig. 2A). Although the proportions of spanning 4q reads

for each allele from the mosaic samples did not match the expected mosaic proportions, sample 19,187's 4qA 2RU allele was captured at a much higher frequency than sample 17,706's 4qA 1RU allele, perhaps reflecting its higher mosaic proportion (50% vs. 25%). Overall, there were far more spanning reads for shorter arrays than for longer arrays, likely due to the greater potential for fragmentation of long reads during library preparation,

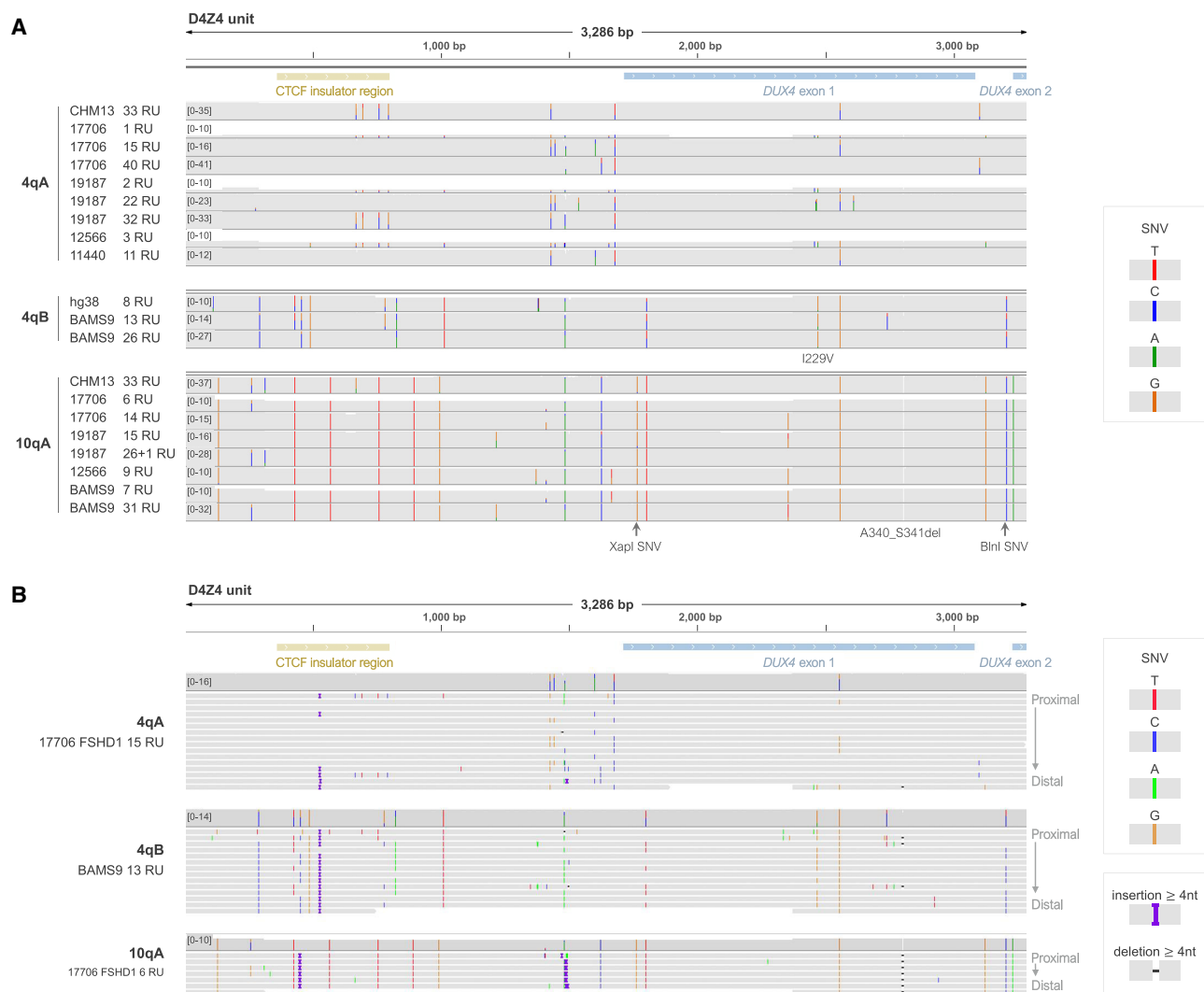
and preference for short reads during sequencing. Nevertheless, we were able to capture multiple spanning reads for alleles of up to 40 repeat units (~142 kb), well beyond the maximum read length for Cas9-targeted sequencing recommended by ONT (30 kb).

### Genetic analysis of 4q/10q alleles using high-accuracy consensus sequences

For alleles with high read depth from Cas9-targeted sequencing, we used Racon (Vaser et al. 2017) to create full-length consensus sequences for the D4Z4 array. Although many automated assembly tools struggle to accurately reconstruct repetitive alleles (Supplemental Table S2), our consensus-calling approach (see Methods) was able to correct errors while faithfully retaining the

structure of the raw D4Z4 reads, including the number of repeat units and unusual variants such as a truncated internal D4Z4 repeat (Supplemental Fig. S3C). This was able to be readily verified by using our interactive visualization tool (Fig. 2) to compare the repeat structure of the consensus sequences to that of the raw reads.

Using the consensus sequences, we were able to reliably analyze SNVs and small insertions/deletions (indels) from 4qA, 4qB, and 10qA alleles. We identified 4q- and 10q-specific variants that were consistent with those seen in CHM13v2.0, including the two SNVs responsible for the 4q-specific XapI site and 10q-specific BlnI site (Fig. 3A). Additionally, we identified qB-specific variants from the BAMS9 4qB alleles, that were consistent with those found in the 4qB allele from GRCh38 (Fig. 3A). Several haplotype-specific variants were present in the *DUX4* open reading frame



**Figure 3.** Array-wide and base-level analysis of D4Z4 genetic variants using consensus sequences from high-coverage long-read sequencing. (A) Integrative Genomics Viewer (IGV) coverage plots for alignments of D4Z4 units from the consensus sequences for patient 4q and 10q alleles, showing 4qA-, 4qB- and 10qA-specific genetic variants consistent with CHM13v2.0 and GRCh38. (B) IGV alignments of D4Z4 units from the consensus sequences for representative 4qA, 4qB and 10qA alleles from samples 17,706 and BAMS3. Within each alignment track, full-length D4Z4 units are arranged from *top* to *bottom* in order of proximal to distal position within the array. Indels <4 nt which were not consistent between repeat units were hidden to aid in visualization (Supplemental Fig. S3). Consensus sequences were generated using Racon using spanning reads from high-coverage Cas9-targeted ONT sequencing. SNV, single nucleotide variant.

(ORF), including four silent SNVs, a 4qB-specific *A > G* variant leading to an I229V substitution, and a 10qA-specific 6-bp deletion leading to A340\_S341del. Several variants were also identified in the *DUX4* promoter region, but none were found to overlap known promoter elements including the GC box, TACAA box, and initiator sequence (Supplemental Fig. S2).

Comparison of the variants between individual D4Z4 units within each 4q and 10q array showed 4q arrays to be more heterogeneous than 10q arrays (Fig. 3B, Supplemental Fig. S3A–E), consistent with previous observations (Salsi et al. 2023). For both 4qA and 4qB arrays, the proximal repeat unit contained a distinct set of variants compared to most other D4Z4 units in the array, although these variants were occasionally also found in internal repeats (Fig. 3B). Variants in the proximal and internal repeat units have previously been used to define D4Z4 subhaplotypes (Lemmers et al. 2010b), however these have largely been restricted to SNVs affecting restriction enzyme sites. The additional 4qA-, 4qB- and 10qA-specific genetic variants identified here provide further insight into D4Z4 subhaplotypes and the composition of the D4Z4 array.

The composition of the D4Z4 array could also be used to reconstruct full-length haplotypes for alleles without spanning reads, and we demonstrated this for the BAMS1 4qAM allele (Supplemental Fig. S4). We also identified a 4qAM allele from the Cas9-sequencing data for sample 19,187 (FSD1) (Fig. 4), and we found the full-length BAMS1 4qAM allele to be the same as 19,187's 4qAM allele (Supplemental Fig. S3C), suggesting a common ancestral origin.

The consensus sequences also enabled more detailed assessment of the 4q subhaplotype based on sequences proximal and distal to the D4Z4 array. Alignment of the distal D4Z4 sequences of 4qAS, 4qAL and 4qAM alleles showed all three haplotypes to have identical pLAM sequences, containing an intact PAS (Fig. 4, Supplemental Fig. S5). Using the 4qAM consensus sequence, we designed 4qAM-specific primers (Supplemental Table S5, Fig. 4) and validated the presence of the 4qAM allele in sample 19,187 by PCR (Supplemental Fig. S6). To further assess the potential pathogenicity of 4qAM alleles, we also assessed the simple sequence length polymorphism (SSLP) found ~3.5 kb upstream of the D4Z4 array that has been used to further delineate permissive and nonpermissive D4Z4 subhaplotypes (Lemmers et al. 2007). We found that the SSLP for the BAMS1 4qAM allele corresponded to the 4A161 subhaplotype, which is permissive for FSHD. This suggests that 4qAM alleles may have the potential to be pathogenic, and may need to be considered in genetic testing.

These findings demonstrate the utility of our approach for analyzing D4Z4 alleles at array-wide, base-level resolution, including the potential to further characterize known D4Z4 subhaplotypes, discover new subhaplotypes, and resolve the composition of D4Z4 arrays.

### Detection of pathogenic variants in *SMCHD1*, *LRIF1*, and *DNMT3B*

We next applied a variant calling and variant effect prediction pipeline to the whole-genome and Cas9-targeted ONT sequencing data, to assess whether our assay could obviate the need for separate genetic testing. We focused on a panel of FSHD2-associated genes, *SMCHD1*, *LRIF1*, and *DNMT3B*, part of the gene panel for Cas9-mediated enrichment. *SMCHD1* genotyping has previously been performed for the patients included in this study (Supplemental Table S3).

We identified pathogenic *SMCHD1* variants for three of the four FSHD2 samples (11,440, 15,166, and 11,491) and both of

the BAMS samples (BAMS1, BAMS9) that recapitulated the prior genotyping results (Supplemental Table S3). For one of the patients diagnosed as FSHD2 (34,140), no pathogenic variant in *SMCHD1*, *DNMT3B*, or *LRIF1* was found. Visual inspection of ultra-long read alignments showed no evidence of structural variants at these loci either (Supplemental Fig. S7). The presence of a synonymous variant in *SMCHD1* was confirmed (p.L1031), but with no further evidence of pathogenicity, calling into question the diagnosis as FSHD2. This diagnosis was established in early childhood on the basis of a severe muscular dystrophy phenotype; no follow-up information is available for this patient.

The BAMS variants were heterozygous missense variants located in or adjacent to *SMCHD1*'s N-terminal ATPase domain (sample BAMS1: p.E136G, sample BAMS9: p.D420V). Conversely, the FSHD2 variants were heterozygous splice donor (sample 11,440: c.2338+4A>G [intron 18]; sample 11,491: c.5547+5G>A [intron 44]) or frameshift (sample 15,166: c.4608\_4614dup, p.Ala1539Tyrfs\*6 [exon 37]) variants found closer towards the C-terminal end of the protein, predicted to lead to *SMCHD1* haploinsufficiency and consistent with an FSHD2 diagnosis.

In one of the FSHD1 samples (12,566), we also identified a previously unreported p.R1887L missense variant in exon 45 of *SMCHD1* which was predicted to be pathogenic by AlphaMissense (Cheng et al. 2023), with a pathogenicity score of 0.9642. None of the samples had pathogenic variants in *LRIF1* or *DNMT3B*. Notably, whereas some FSHD2-associated genes had only low coverage with whole-genome sequencing, the read depth was significantly increased with Cas9-targeted sequencing, aiding in detection of heterozygous pathogenic variants.

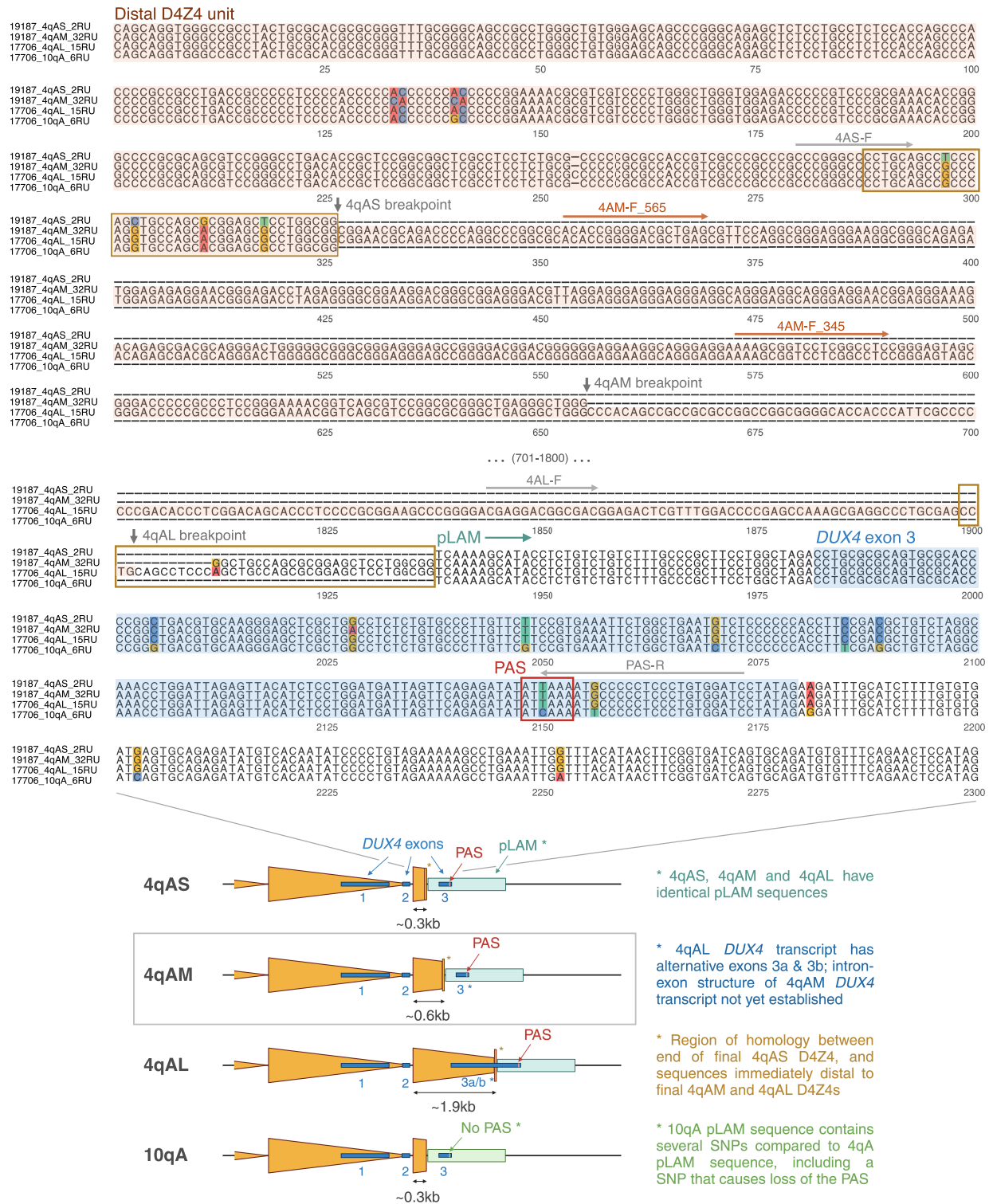
This confirmed the suitability of a single nanopore sequencing assay for 4q haplotyping, 4q structural characterization and variant detection in an FSHD gene panel.

### Allele-specific CpG methylation patterns across the D4Z4 array

#### *Array-size-dependent methylation patterns are seen in FSHD1 and control fibroblasts*

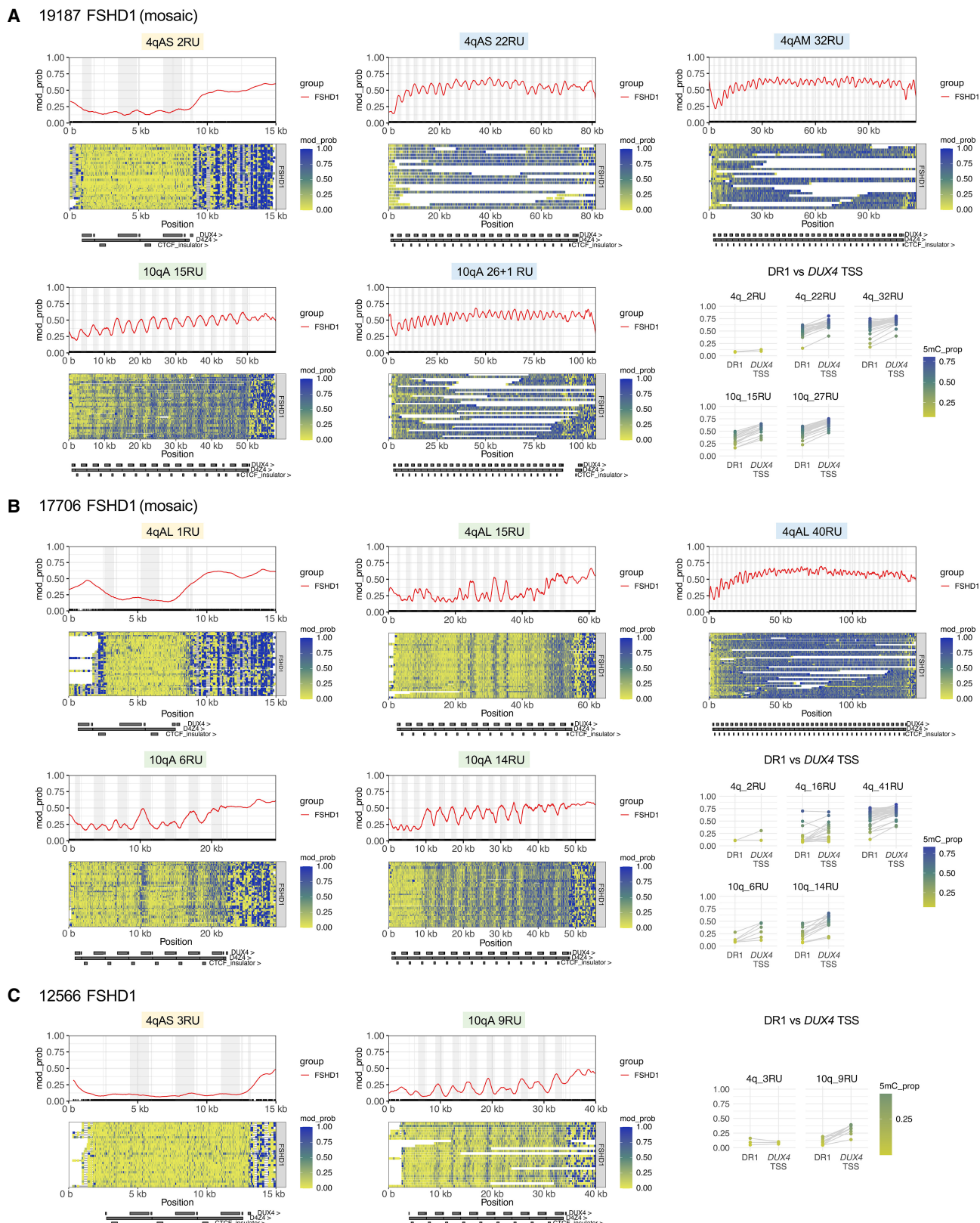
To assess the methylation patterns of D4Z4 arrays of different lengths and *SMCHD1* variant status, we called CpG methylation for the 4q and 10q reads, and used NanoMethViz (Su et al. 2021, 2025) to plot the methylation of individual molecules as well as the smoothed methylation profile across the complete array for each allele.

For FSHD1 and control samples, an oscillatory pattern of D4Z4 methylation was seen within each repeat unit, with low levels of methylation around the DR1 region, which corresponds to the distal portion of a proposed CTCF binding site (Ottaviani et al. 2009; Hartweck et al. 2013), and high levels of methylation around the *DUX4* transcription start site (TSS) (Fig. 5A–C, Supplemental Fig. S8A–E), consistent with previous studies (Huichalaf et al. 2014; Butterfield et al. 2023). These oscillatory patterns were present in all noncontracted alleles, but absent from the hypomethylated, contracted FSHD1 alleles. Autocorrelation of the methylation status of CpGs across the D4Z4 array indicated strong correlation between sites separated by ~3.3 kbp, corresponding to the size of one D4Z4 unit, as well as between sites separated by ~180 bp (Supplemental Fig. S9), which likely represents nucleosome positioning (Butterfield et al. 2023; Kerr et al. 2023). The high read depth provided by Cas9-targeted sequencing enabled us to see that for each allele, the overall methylation patterns were highly consistent between individual molecules (Fig. 5), but



**Figure 4.** Base-level comparison of distal D4Z4 sequences from 4qAS, 4qAM, 4qAL and 10qA alleles. (Top) Alignment of distal D4Z4 sequences beginning from the proximal KpnI site for the final partial D4Z4 unit, and extending into the proximal portion of the pLAM sequence containing DUX4 exon 3. Sequences are extracted from consensus sequences generated using Racon from high-coverage Cas9-targeted ONT sequencing. Breakpoints for the final partial D4Z4 units (vertical arrows) were determined by alignment to a full-length D4Z4 sequence. The start of pLAM is marked as the first nucleotide following the final partial 4qAS D4Z4 unit. The 4qAM and 4qAL alleles contain an extra length of sequence between the end of their final partial D4Z4 units and the start of pLAM, which corresponds to the end of the 4qAS allele's final partial D4Z4 unit (shown in orange boxes). Locations of primers used for PCR genotyping of 4qAS/4qAM/4qAL (Supplemental Fig. S6, Supplemental Table S5) are shown above their respective sequences; 4AS-F, 4AL-F, and PAS-R (gray) are from Lemmers et al. (2018b), whereas 4AM-F\_565 and 4AM-F\_345 (brown) were newly designed using the 4qAM consensus sequence. (Bottom) Schematic comparing the distal sequences of 4qAS, 4qAM (newly identified), 4qAL and 10qA alleles.





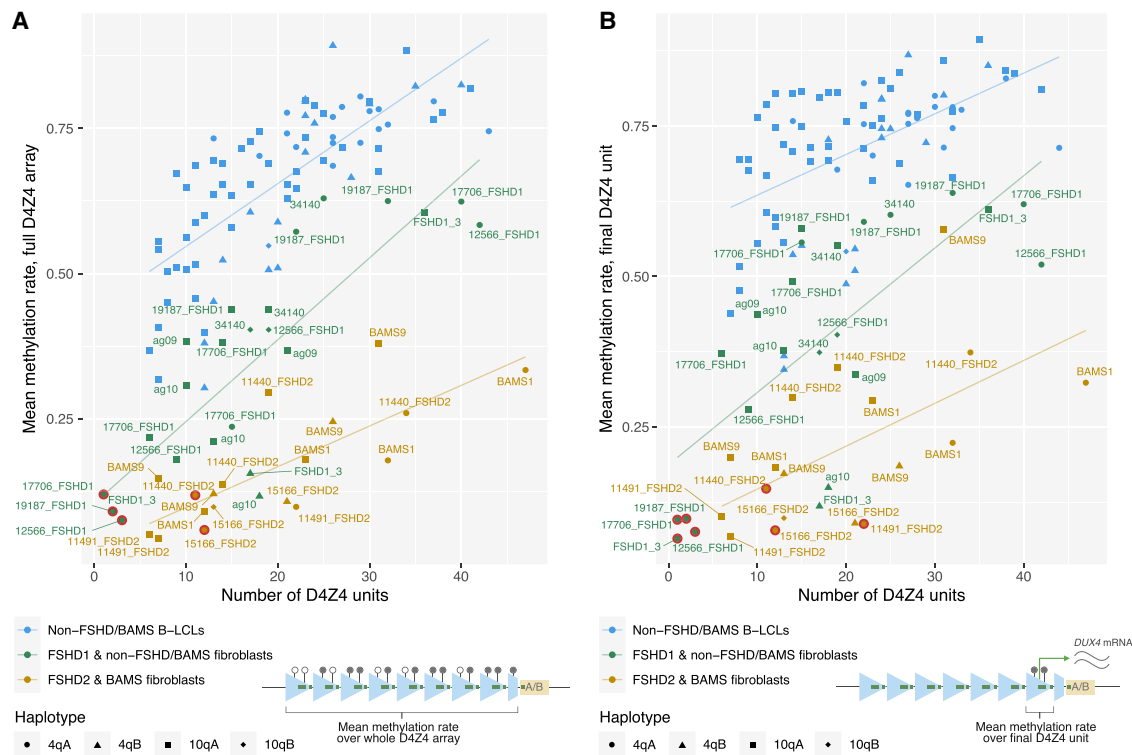
**Figure 5.** Allele-specific, array-wide D4Z4 methylation profiles for FSHD1 fibroblasts from Cas9-targeted sequencing. Single-molecule and smoothed methylation plots for 4q and 10q D4Z4 alleles from the FSHD1 samples (A) 19,187, (B) 17,706, and (C) 12,566, generated using NanoMethViz (Su et al. 2021, 2025). The locations of *DUX4* exons are shaded in gray on the smoothed methylation plot. Annotations for D4Z4 repeat units, CTCF insulator regions, and *DUX4* exons are shown below each plot. Mean methylation rates over the DR1 region (positions 563–814 of the D4Z4 KpnI-KpnI unit) (Hartweck et al. 2013) and *DUX4* transcription start site (TSS) (–200 to +200 with respect to the TSS at position 1688 of the D4Z4 KpnI-KpnI unit) within each D4Z4 unit were also plotted for each allele. Points for DR1 and *DUX4* TSS regions from the same D4Z4 unit are connected by gray lines.

that there was significant intermolecular heterogeneity in methylation state at the level of individual CpG sites (i.e., a given CpG site can be methylated in one read and unmethylated in another; Supplemental Fig. S10).

Differing array-wide methylation patterns were observed for short, intermediate and long D4Z4 arrays (Fig. 5, Supplemental Fig. S8). The short, pathogenic FSHD1 arrays, which ranged from 1 to 3 repeat units, were severely hypomethylated across the full length of the array (Fig. 5A–C and Supplemental Fig. S8C, D, alleles shaded in yellow). Long arrays of >20 repeat units were much more highly methylated, and tended to display a regular, stepwise increase in D4Z4 methylation over the first ten or so repeat units before reaching a constant, high level of methylation (Fig. 5A, B and Supplemental Fig. S8A–E, alleles shaded in blue), similarly to what has been described previously (Butterfield et al. 2023). In contrast, short D4Z4 arrays at the upper end of the pathogenic range for FSHD1 (6–10 repeat units) and intermediate-length arrays (11–20 repeat units) were more variable in their methylation patterns (Fig. 5A–C and Supplemental Fig. S8A–E, alleles shaded in green). Although these arrays were often more highly methylated at their distal ends compared to their proximal ends, this proximal-to-distal increase in methylation did not always occur in a stepwise fashion, as exemplified by the “irregular” methylation patterns seen in several of the alleles from sample 17,706 (Fig. 5B). These observa-

tions may partially explain why array sizes in the 1–4 repeat unit range are seen in more clinically severe and early-onset cases of FSHD, whereas longer arrays are associated with variable penetrance and a broad range of clinical severity (Mul et al. 2018b; Goselink et al. 2019).

The differences in D4Z4 methylation between short, intermediate and long D4Z4 arrays were also reflected in the global D4Z4 methylation level for each allele, as measured by the proportion of all methylated CpGs across the D4Z4 array (henceforth “global D4Z4 5mC rate.”) Overall, we found a positive correlation between the average allelic D4Z4 5mC rate and the number of repeat units for both FSHD1 samples and controls (Pearson correlation coefficient 0.844, Spearman’s correlation coefficient 0.826) (Fig. 6A, Supplemental Fig. S11A). As only the terminal *DUX4* gives rise to stable transcripts in permissive 4qA alleles, we additionally assessed the mean methylation of the final full D4Z4 unit, which contains the promoter and exons 1 and 2 of the terminal *DUX4*. Amongst FSHD1 and control samples, there was a moderate correlation between the number of repeat units and mean methylation rate of the final full D4Z4 unit (“final D4Z4 5mC rate”) (Pearson correlation coefficient 0.701, Spearman’s correlation coefficient 0.732) (Fig. 6B, Supplemental Fig. S11B). As expected, the pathogenic, contracted FSHD1 alleles had very low final D4Z4 5mC rates, ranging from 0.064 to 0.097. However, there were two



**Figure 6.** Correlation of allelic D4Z4 methylation levels with the number of repeat units. Mean methylation rates of individual alleles across (A) the full 4q/10q D4Z4 arrays or (B) the final D4Z4 unit against the number of D4Z4 units, plotted for alleles from *SMCHD1*-wild-type fibroblasts (FSHD1 and non-FSHD/BAMS), pathogenic-*SMCHD1*-variant fibroblasts (FSHD2 and BAMS), and 30 B-lymphoblastoid cell lines (B-LCLs) from the 1000 Genomes Project (Gustafson et al. 2024). Sample 34,140, initially clinically diagnosed as FSHD2, was included in the “non-FSHD/BAMS” group based on the lack of pathogenic *SMCHD1*, *LRIF1*, or *DNMT3B* variants and overall methylation profile. Mean methylation rate was calculated as the total number of methylated CpGs from all reads/the total number of methylated + unmethylated CpGs from all reads, across the defined regions. Regression lines were plotted for each group of samples. Alleles presumed to be the pathogenic alleles for each of the FSHD patients are indicated with red rings. (A) Pearson and Spearman’s correlation coefficients for FSHD1 and non-FSHD/BAMS fibroblasts, FSHD2 and BAMS fibroblasts, and B-LCLs were 0.844 and 0.826, 0.757 and 0.742, and 0.746 and 0.798, respectively. (B) Pearson and Spearman’s correlation coefficients for FSHD1 and non-FSHD fibroblasts, FSHD2 and BAMS fibroblasts, and B-LCLs were 0.701 and 0.732, 0.583 and 0.607, and 0.531 and 0.552, respectively.

intermediate-length *SMCHD1*-wild-type alleles which also had low final D4Z4 5mC rates, the FSHD1\_3 4qB 17RU allele (final D4Z4 5mC rate 0.118) and the AG10803 (ag10) 4qB 18RU allele (final D4Z4 5mC rate 0.149). The AG10803 4qB 18RU allele appeared to be hypomethylated across the full length of the array (global D4Z4 5mC rate 0.116), although there were only two spanning reads to support the methylation signal (Supplemental Fig. S8A). These findings once again suggest that in FSHD1 and control fibroblasts, although contracted D4Z4 arrays are consistently hypomethylated, and long D4Z4 arrays are consistently hypermethylated, methylation patterns for intermediate-length arrays are more variable.

#### *B-lymphoblastoid cell lines show similar D4Z4 methylation patterns to FSHD1 and control fibroblasts, but have higher methylation levels*

To assess whether these length-dependent methylation patterns are also seen in a larger cohort and other cell types, we analyzed D4Z4 methylation in 30 B-lymphoblastoid cell lines (B-LCLs) from the 1000 Genomes Project (1KGP), using publicly available ONT WGS data from the 1KGP ONT Sequencing Consortium (1KGP-ONT) (Gustafson et al. 2024). We found 83 4q/10q alleles with reads that spanned the whole D4Z4 array, ranging in length from 6 to 43 repeat units (Supplemental Table S6). The patterns of DNA methylation across the array in the B-LCLs were similar to those seen in *SMCHD1*-wild-type fibroblasts. All arrays had oscillatory patterns of low methylation around the DR1/CTCF insulator region and high methylation around the *DUX4* TSS (Supplemental Fig. S12). For long arrays (>20 repeats), methylation levels increased over the first ten or so repeat units then plateaued at a constant, high level across the remainder of the array, or were consistently high across the whole array. Short and intermediate-length arrays (6–20 repeat units) tended to display more regular, stepwise increases in methylation from the proximal to distal end of the array, contrasting with the “atypical” methylation patterns seen in several of the alleles from the FSHD1 and control fibroblasts. Similarly to what we found in fibroblasts, there was a positive correlation between the global D4Z4 5mC and final D4Z4 5mC rates, and the number of D4Z4 units (Fig. 6). However, mean D4Z4 methylation rates were generally higher in the B-LCLs compared to the *SMCHD1*-wild-type fibroblasts, suggesting that global D4Z4 methylation levels are cell type specific.

#### *Both FSHD2 and BAMS fibroblasts have global D4Z4 hypomethylation*

We sought to investigate the effect of pathogenic FSHD2 and BAMS *SMCHD1* variants on D4Z4 methylation. In both FSHD2 and BAMS, there was global hypomethylation of D4Z4 arrays compared to similar-length FSHD1 and control samples (Fig. 7A). Despite the differing effects of FSHD2 and BAMS variants on the *SMCHD1* protein (predicted haploinsufficiency and missense mutations in the ATPase domain, respectively), FSHD2 and BAMS samples showed similar overall methylation levels and methylation patterns to each other (Fig. 7, Supplemental Fig. S11). Short and intermediate-length arrays were generally severely hypomethylated across the full length of the array. Longer alleles had relatively higher methylation levels, yet were still markedly hypomethylated compared to similar-length *SMCHD1*-wild-type samples (Fig. 6A), and generally had lower methylation rates for their final D4Z4 units (Fig. 6B) owing to minimal proximal-to-distal increase in array-wide methylation. This was clearly demonstrated by comparison of the 32RU 4qAM alleles from BAMS9 and 19,187\_FSHD1 (Figs. 5A, 7B). However, we also identified sev-

eral reads with very high methylation levels that mapped to the BAMS1 47RU 4qAS allele (Supplemental Fig. S13B), suggesting either the presence of a separate, highly methylated cell population, or an unexpected and sporadic “rescue” of these cells from hypomethylation.

In terms of intra- and inter-repeat methylation patterns, oscillations between low methylation around the DR1/CTCF insulator region and high methylation around the *DUX4* TSS were variably present in hypomethylated alleles, but were prominent in more highly methylated longer alleles (Fig. 7). Autocorrelation of 5mC signals once again indicated methylation patterns of ~3.3 kbp and ~180 bp periodicity (Supplemental Fig. S9), suggesting that the mechanisms that give rise to these patterns are retained in FSHD2 and BAMS.

#### **Complete genetic and epigenetic analysis of in-cis duplicated arrays and upstream inverted D4Z4 units**

Although several studies have used molecular combing to study the structure of in-cis duplication alleles (Vasale et al. 2015; Nguyen et al. 2017; Lemmers et al. 2018a; Delourme et al. 2023; Lemmers et al. 2023), this technique is unable to reveal the underlying genetic sequence or epigenetic status, and many aspects of the genetics and epigenetics of these complex alleles have yet to be resolved. Therefore, we sought to use long-read sequencing to further investigate the composition and methylation of in-cis duplicated arrays.

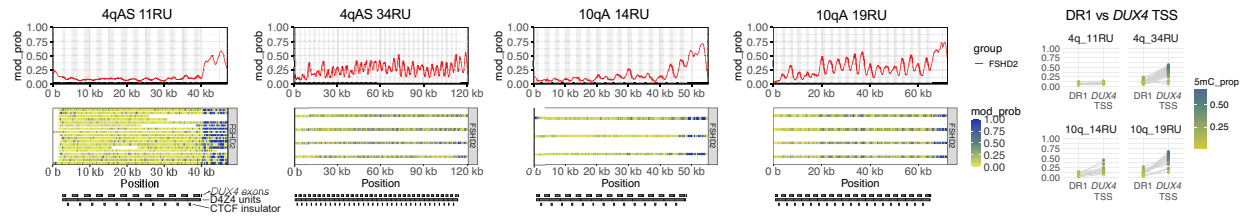
Our pipeline identified in-cis duplications downstream from the main D4Z4 array in two of the FSHD1 alleles (19,187 10qA, FSHD1\_3 10qA). Sample 19,187's complex 10qA allele was composed of a 26RU array followed by a 1RU array, separated by a ~6.5 kb spacer sequence, with the proximal partial D4Z4 of the duplicated array beginning ~0.3 kb after the KpnI restriction site (Fig. 8A). Sample FSHD1\_3's complex 10qA allele contained an in-cis triplication, comprising a 15RU array followed by a 2RU array and a 5RU array, separated by two spacer sequences of ~20 kb, with the proximal partial D4Z4s of the duplicated arrays both beginning ~1.5 kb after the KpnI restriction site (Fig. 8B). Analysis of the genetic sequence revealed all of the arrays to be 10qA-type, as determined by the presence of BlnI sites and 10qA-type pLAM sequences lacking a functional PAS. Moreover, the spacer sequences and the sequences distal to the duplicated array all corresponded to the sub-telomeric sequence that is normally found immediately distal to the 10qA D4Z4 array.

We also identified the first reported case of a 4qB in-cis duplicated array, present in the B-LCL sample HG00675. The allele contained a proximal array of at least 14RU, followed by an in-cis duplicated array of 5RU. The arrays were separated by a ~13.7 kb spacer sequence, with the proximal partial D4Z4 of the duplicated array beginning ~0.8 kb after the KpnI restriction site (Fig. 8C). Both arrays were 4qB-type, and once again, the spacer sequence was identical to the sequence distal to the duplicated array, both corresponding to the 4qB sub-telomeric sequence. These findings mirror those in Lemmers et al. (2023), where the spacer sequence for 4qA-type in-cis duplication alleles was found to correspond to the 4qA sub-telomeric sequence. The shared basic structure of these 4qA, 4qB and 10qA duplication alleles suggests that they emerged via a common underlying mechanism, likely involving intrachromatid gene conversion.

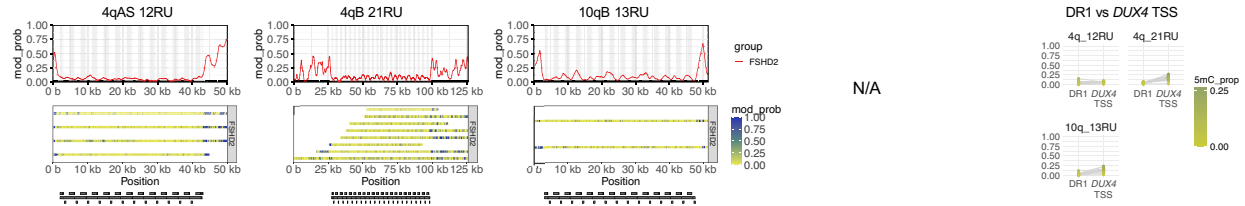
We then assessed the methylation of the duplicated arrays. For all of these alleles, the downstream, contracted arrays were highly methylated, with average methylation levels similar to or

**A** FSHD2 fibroblasts

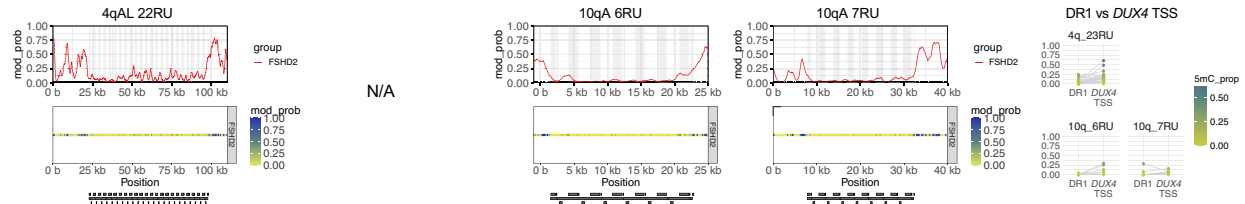
11440\_FSHD2



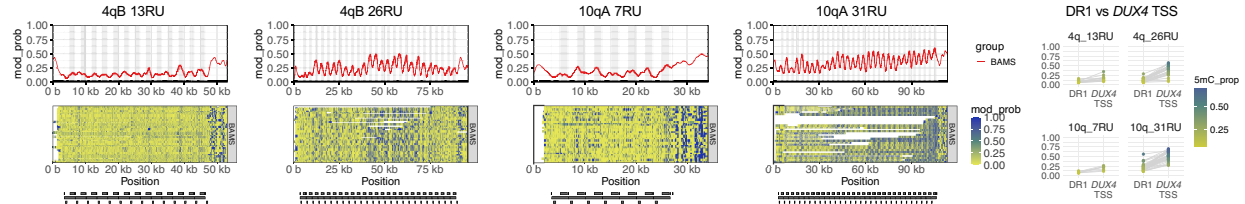
15166\_FSHD2



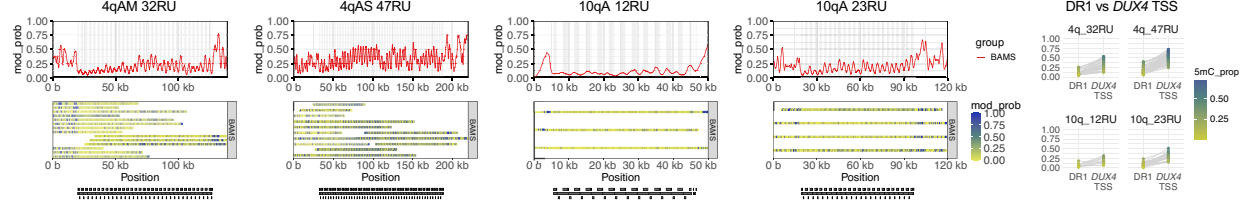
11491\_FSHD2

**B** BAMS fibroblasts

BAMS9



BAMS1

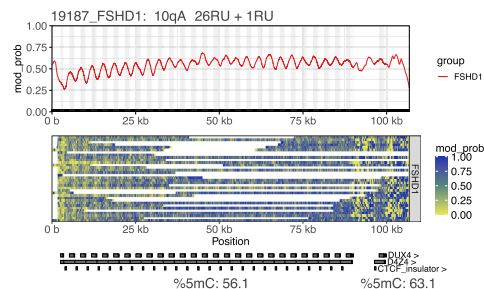
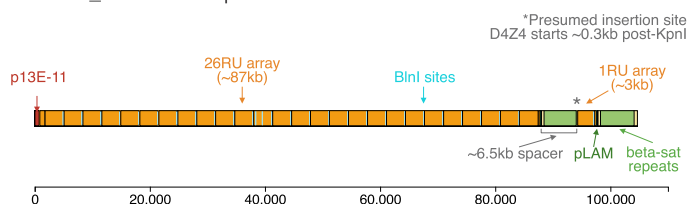
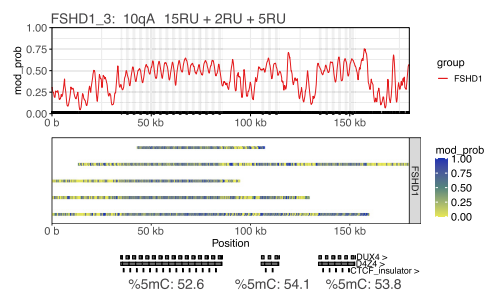
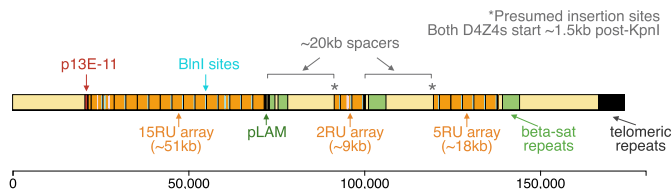
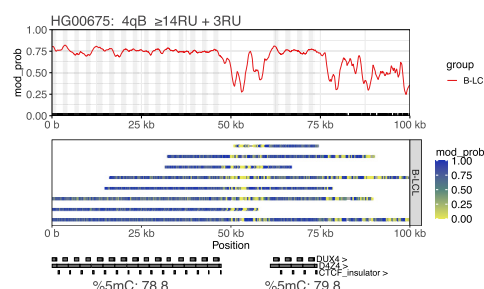
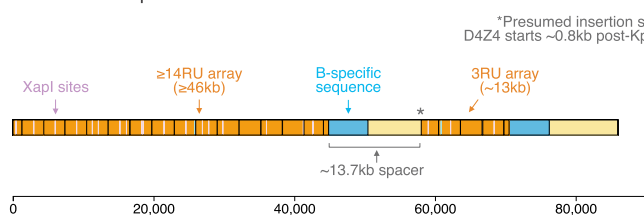


**Figure 7.** Allele-specific, array-wide D4Z4 methylation profiles for FSHD2 and BAMS fibroblasts. Single-molecule and smoothed methylation plots for 4q and 10q D4Z4 alleles from (A) FSHD2 and (B) BAMS samples, generated using NanoMethViz (Su et al. 2021, 2025). The locations of *DUX4* exons are shaded in gray on the smoothed methylation plot. Annotations for D4Z4 repeat units, CTCF insulator regions, and *DUX4* exons are shown below each plot. Mean methylation rates over the DR1 region (positions 563–814 of the D4Z4 KpnI-KpnI unit (Hartweck et al. 2013) and *DUX4* TSS region (–200 to +200 with respect to the TSS at position 1688 of the D4Z4 KpnI-KpnI unit) within each D4Z4 unit were also plotted for each allele. Points for DR1 and *DUX4* TSS regions from the same D4Z4 unit are connected by gray lines. The full-length reference sequences for the BAMS1 4qAM and 4qAS alleles were determined as described in Supplemental Figs. S4 and S13.

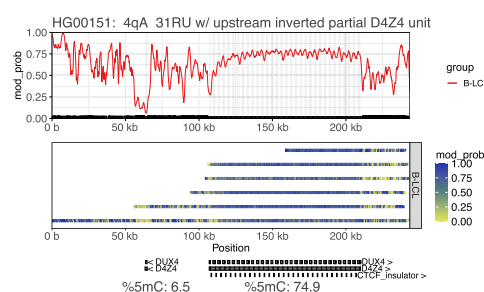
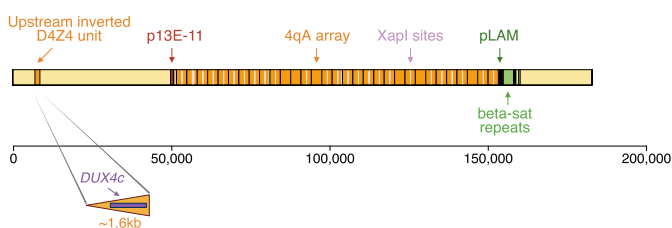
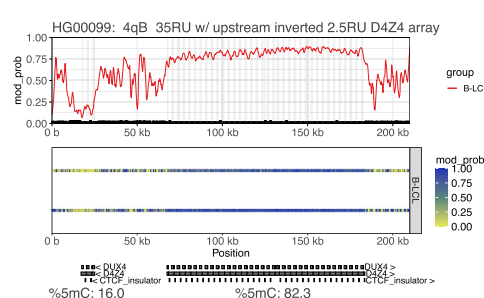
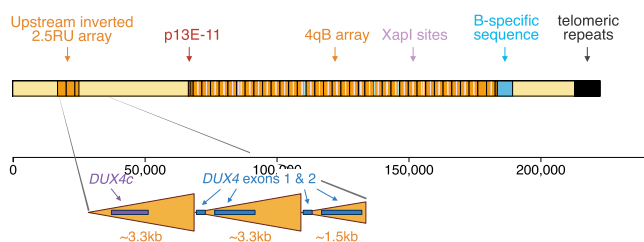
higher than those of the noncontracted upstream arrays (Fig. 8A–C). Additionally, for sample FSHD1\_3's in-cis triplication 10qA allele, the average methylation of the proximal 15RU array was higher than might be expected if it were an isolated 15RU array, based on comparison with the global D4Z4 5mC rate of sample

FSHD1\_3's 4qB 17RU allele (0.155) and the methylation of intermediate-length arrays in other *SMCHD1*-wild-type samples (Fig. 6A). Lemmers et al. (2023) found a similar phenomenon for a mosaic 17+9RU 4qA allele and 17+2RU 4qA allele, where the proximal 17RU allele for the 17+9RU array was more highly methylated



Downstream in-*cis* duplicated D4Z4 arrays**A** 19187\_FSHD1: 10qA 26RU + 1RU**B** FSHD1\_3: 10qA 15RU + 2RU + 5RU**C** HG00675: 4qB ≥14RU + 3RU

## Upstream inverted D4Z4 units and arrays

**D** HG00151: 4qA upstream inverted partial D4Z4 unit**E** HG00099: 4qB upstream inverted 2.5RU D4Z4 array

**Figure 8.** Analysis of the structure and methylation of duplicated D4Z4 arrays and upstream inverted D4Z4 units. (A), (B), (C) Structure of 4q and 10q alleles identified by the pipeline to have in-*cis* duplications downstream from the D4Z4 array, alongside smoothed methylation plots showing that these duplicated arrays are hypermethylated. The overall %5mC for each D4Z4 array is shown *below* the annotation track for the methylation plot. The locations of *DUX4* exons are shaded in gray on the smoothed methylation plot. (D) Representative example of the inverted D4S2463 unit located ~42 kb upstream of the 4q D4Z4 array, which is found to be hypomethylated in FSHD and control fibroblasts and B-LCL samples. (E) The HG00675 cell line from the 1000 Genomes Project (Gustafson et al. 2024) has a 4qB allele with an upstream inverted D4Z4 array of 2.5 repeat units in place of the D4S2463 unit. This upstream inverted array is also hypomethylated.



than for the 17+2RU array. This may indicate the existence of bi-directional methylation spreading mechanisms between proximal and distal arrays, as suggested by Lemmers et al. (2023), or reflect an underlying chromatin state that is dependent upon the additive effect of both duplicated and nonduplicated arrays.

To further characterize the structure and methylation of *in-cis* D4Z4 repeats, we also assessed the inverted D4Z4 unit (D4S2463) that is found ~42 kb upstream of the 4q D4Z4 array. Consistent with previous reports (Wright et al. 1993; Ansseau et al. 2009), we found that D4S2463 is a truncated unit of ~1.5 kb that is homologous to the distal portion of the KpnI-KpnI D4Z4 unit, and contains a homologous *DUX4c* gene. Contrary to the downstream *in-cis* duplicated arrays, the upstream inverted repeat was consistently hypomethylated in the FSHD, BAMS, and control fibroblast samples and in the B-LCL samples (Fig. 8D).

Moreover, we identified one B-LCL allele (HG00099 4qB) that had an inverted D4Z4 array of 2.5 repeat units ~42 kb upstream of the main D4Z4 array, rather than a single D4S2463 unit (Fig. 8E). Several cases of the same allelic variant have also previously been identified using molecular combing (Lemmers et al. 2018a), where the inverted array was suggested to be an expansion of the D4S2463 repeat. Here, we were able to analyze the composition of the inverted array at base-level resolution, and we found that the first inverted D4Z4 unit was ~1.6 kb and corresponded to the distal ~1.6 kb of D4Z4, whereas the second and third inverted D4Z4 units were ~3.3 kb units containing *DUX4* and *DUX4c* sequences, respectively. Once again, Racon (Vaser et al. 2017) was used to create a consensus sequence for the upstream inverted D4Z4 array, and it was found that all three repeat units contained 4qB-type SNVs, whereas the sequence for the last ~1.5 kb of the final repeat corresponded to the normal sequence for the inverted D4S2463 unit (Supplemental Fig. S14). This suggests that the inverted array may have arisen by insertion of a segment of 4qB array at the breakpoint site for the D4S2463 unit, mediated by partial homology between the D4S2463 unit and the normal D4Z4 sequence; or alternatively, that the 4qB-type inverted array is in fact the ancestral allele, and gave rise to the partial D4S2463 repeat by contraction.

We then assessed the methylation of this inverted array. Similarly to the single inverted D4S2463 units, the inverted array was hypomethylated, with a mean 5mC rate of 0.160. Therefore, overall, we found that downstream, *in-cis* duplicated arrays were methylated at similar levels to the first D4Z4 array, whereas upstream inverted arrays were hypomethylated similar to the typical D4S2463 unit. This could suggest that methylation spreading mechanisms at the D4Z4 locus, if present, are orientation-dependent, or may provide further support for an insulator mechanism at the D4Z4 locus that separates upstream inverted and downstream noninverted arrays into separate chromatin compartments (Ottaviani et al. 2009; Gaillard et al. 2019).

## Discussion

The D4Z4 macrosatellite array is one of the most difficult regions in the genome to study. Traditional techniques such as Southern blotting and short-read sequencing are only able to provide fragmented insights into D4Z4 (epi)genetics, hampering our ability to resolve complex D4Z4 alleles, and to unravel the molecular mechanisms acting at this locus in health and disease. Using long-read nanopore sequencing, here we have developed an optimized protocol and all-in-one analysis workflow for FSHD, D4Z4End2End, that is able to identify full sets of patient 4q and

10q alleles, accurately detect pathogenic variants in FSHD2-associated genes (*SMCHD1*, *DNMT3B* and *LRIF1*), and assess D4Z4 methylation status in a single assay.

The interactive annotation and visualization of raw D4Z4 reads outlined issues with current automated macrosatellite assembly. By annotating and exploring individual raw reads, we were able to more reliably resolve the structure, repeat number, and haplotype of D4Z4 alleles, which is critical for accurate downstream analyses and accurate FSHD diagnoses.

We applied our workflow to the study of primary fibroblast samples from two control individuals, four FSHD1 patients, four FSHD2 patients, and two BAMS patients. Although this represents a smaller cohort compared to several previous ONT studies (Hiramuki et al. 2022; Butterfield et al. 2023; Yeetong et al. 2023; Huang et al. 2024), we were able to attain high coverage and capture longer arrays (up to 42 repeat units (~140 kb)), cases of mosaicism, and several complex alleles (including entire *in-cis* duplication and triplication alleles (~173 kb)), while also providing comparisons between full-length FSHD and BAMS alleles. By generating accurate consensus sequences from high-coverage, full-length patient alleles, we were able to perform detailed analysis of D4Z4 genetic variation, including the identification of haplotype-specific SNVs, and characterization of inter-repeat and inter-array variation. Moreover, we discovered a previously unreported 4qA variant, “4qAM,” which differs from the previously identified 4qAS and 4qAL variants by the length of its distal repeat unit, but has an identical pLAM sequence and intact polyadenylation signal. Further studies could elucidate the prevalence and distribution of 4qAM alleles in the general population, whether they can serve as the causative allele for FSHD, and whether they give rise to different *DUX4* transcripts, as is the case for 4qAS and 4qAL alleles (Lemmers et al. 2018b). If 4qAM alleles are indeed permissive for *DUX4* expression, this would need to be considered in FSHD genetic testing, including updated conversion tables for calculation of repeat unit size using Southern blotting, and the use of 4qAM-specific primers such as the ones we designed in addition to the 4qAS- and 4qAL-specific primers already used for (bisulfite)-PCR assays (Jones et al. 2014; Lemmers et al. 2018b; Giardina et al. 2024).

Repeat number calculations occasionally differed from previously reported numbers in draft pangenome assemblies and Southern blot/molecular combing analyses. Discrepancies with draft pangenome assemblies can be explained by the current assemblers’ difficulties with macrosatellites. Often those regions still require manual scaffolding and curation (Tørresen et al. 2019). Both the assemblies and our analysis use the same raw data, so the direct annotation of multiple spanning reads via D4Z4End2End is more likely to reflect the truth. Discrepancies with Southern blot and molecular combing can arise from their limited resolution on longer alleles (e.g., 37 vs. 40 RU), the specific probes, scanners and software version used, and the caveats of extrapolating RU numbers from array length rather than analyzing the read structure at base resolution with nanopore sequencing. For example, for patient sample 19,187, our sequencing data showed that one internal repeat of a 10qA allele is truncated (~1.2 kb deletion), which was counted by D4Z4End2End in the RU count (15 RU), but could account for length-based Southern blot/molecular combing estimating one fewer repeat unit (14 RU). The method of counting may also account for differences of +/-1 units with reported lengths, as we exclude partial repeats at either end of the array in accordance with the updated guidelines. For FSHD1\_3, all of our RU measurements are one fewer than the

reported counts based on Southern blot/molecular combing, which may have been scored using slightly different criteria.

Accurate haplotyping and the capture of full-length spanning reads also allowed us to analyze the allele-specific and array-wide methylation profiles of FSHD and BAMS D4Z4 alleles, in aggregate and at single-molecule resolution. We were able to observe broad patterns in D4Z4 methylation that correlate with D4Z4 array length and *SMCHD1* status, including clear hypomethylation of contracted arrays in FSHD1, and of all 4q and 10q arrays in FSHD2. Notably, one of the samples that had been clinically diagnosed as FSHD2 (34,140) displayed comparatively high levels of methylation for all alleles, consistent with the absence of contracted arrays or pathogenic *SMCHD1*, *DNMT3B*, or *LRIF1* variants. This illustrates that our workflow can be used to confirm a diagnosis of FSHD in patients with clinical manifestations of the disease, or conversely, to suggest an alternative diagnosis.

High-coverage and full-length methylation profiles also yielded several interesting findings. Many arrays showed a step-wise proximal-to-distal increase in methylation, however we observed some arrays with “atypical” methylation patterns, especially in intermediate-length fibroblast alleles. Of note, overall DNA methylation patterns were highly consistent between individual molecules from the same allele, even for arrays with atypical methylation patterns, although not at the level of individual CpG sites. This aligns with observations from previous studies, which have shown that overall DNA methylation patterns can be stably inherited in clonal populations while being heterogeneous at the per-base level, which may represent differences in nucleosome phasing between molecules, stochasticity in the de novo and maintenance activities of epigenetic modifiers, and the influence of the methylation states of neighboring CpGs (Busto-Moner et al. 2020; Wang et al. 2020; Kerr et al. 2023). Future work should reveal whether those full-length D4Z4 methylation characteristics hold in other relevant cell types, such as myocytes to investigate disease pathogenesis and primary peripheral blood cells which could be used for diagnostic testing.

Previous studies have also indicated that BAMS is associated with D4Z4 hypomethylation in a similar manner to in FSHD2 (Shaw et al. 2017), and here we have performed the first full-length analysis of methylation at the D4Z4 array in BAMS patients. Our results confirm that, at least in primary fibroblasts, D4Z4 methylation patterns are highly similar between BAMS and FSHD2, with retention of oscillatory patterns at the ~3.3 kbp and ~180 bp levels, but with overall marked hypomethylation. So far, there has only been one reported case of a patient with overlapping symptoms of BAMS and FSHD, on the background of a FSHD2-permissive D4Z4 genotype (Shaw et al. 2017). Shaw et al. (2017) and Mohassel et al. (2022) identified several other BAMS patients without clinical FSHD who had potentially permissive D4Z4 genotypes, however array size information was not available for all patients, D4Z4 methylation values were aggregate results obtained via bisulfite sequencing, and for young patients FSHD onset might not have been reached (as it typically manifests in young adults). Further genotyping and allele-specific methylation analysis could be used to more comprehensively characterize FSHD-permissive alleles in BAMS patients, and explore the possibility of as-yet undiscovered mechanisms that protect these patients from FSHD.

Finally, we also characterized several complex D4Z4 alleles containing duplicated D4Z4 arrays. Our results suggest that 4qA, 4qB and 10qA duplication alleles likely arise via the same mechanism, which results in in-cis duplications of the distal end of the subtelomere in the noninverted orientation. We found that down-

stream in-cis duplicated arrays are highly methylated, whereas upstream inverted D4Z4 repeats are hypomethylated, which may suggest orientation-dependent long-range DNA methylation spreading mechanisms, inheritance of methylation patterns from the ancestral array, or insulator mechanisms that generate local differences in chromatin organization. Lemmers et al. (2023) found that in-cis duplication alleles may be able to give rise to FSHD-level *DUX4* expression and cause clinical FSHD in the absence of D4Z4 hypomethylation, demonstrating the need to further characterize the epigenetic mechanisms mediating *DUX4* repression.

Although D4Z4End2End was developed and tested as a research/diagnostic-enabling pipeline only, and roll out to larger cohorts in a clinical laboratory environment will be crucial, we anticipate that our approach has the potential to greatly simplify and streamline FSHD diagnostics. The field of long-read sequencing diagnostics is still emerging and there are no commonly accepted guidelines yet regarding data acquisition or analysis (Eisfeldt et al. 2025). For FSHD specifically, we propose the following principles as a starting point. For FSHD diagnosis, only the short and demethylated 4qA allele is clinically relevant. In principle, as few as two independent full-length reads spanning this allele could be sufficient for a confident diagnosis, because the large structural elements (pLAM, D4Z4 repeats) and pronounced hypomethylation do not demand base-level precision. Thus as little as 15× coverage with read lengths exceeding 70 kb (longer than the pathogenic threshold for both FSHD1 and 2) should be adequate for most samples. FSHD2 could potentially still be diagnosed with nonspanning reads if key indicators such as low methylation, 4qA haplotype, and a pathogenic *SMCHD1* or related variant are evident, therefore 40 kb reads might generally be sufficient. In mosaic cases, the required coverage scales with the variant allele frequency; for example, to detect a mosaic variant at 10% frequency, total coverage of the 4q35 region should increase to approximately 75× (although the enrichment bias for shorter fragments in Cas9-targeted sequencing would decrease this requirement in practice). In addition, 15×–20× coverage is generally sufficient to detect small and large germline variants in FSHD-associated genes (Kolesnikov et al. 2024; Smolka et al. 2024). Live analysis could allow stopping of data acquisition as soon as a confident diagnosis is reached. Although the Cas9 protocol is no longer officially supported by ONT for the current R10.4.1 flow cell chemistry, optimization efforts from other groups ensure that this approach remains viable (Scholz et al. 2024; Benarroch et al. 2025). Several adaptations to the protocol could optimize it even further for use in clinical testing. First, in this study we have sequenced DNA from primary fibroblasts derived from skin biopsies, however for diagnostic testing, peripheral blood samples would be more suitable in terms of decreased cost and invasiveness, and increased acceptability to patients. Previous studies that have performed nanopore sequencing of peripheral blood have also shown reduced D4Z4 methylation of pathogenic alleles compared to controls (Hiramuki et al. 2022; Butterfield et al. 2023) as well as in-cis duplication alleles and somatic mosaicism (Lemmers et al. 2023). Second, in our study we use a single flow cell per patient sample, however we envisage that real-time analysis could facilitate dynamic processing of the sequencing data using the haplotyping pipeline, so that runs can be stopped once enough spanning reads are reached. This could allow multiple samples to be run on a single flow cell after sequential washes and reloads, decreasing sequencing costs. Using one flow cell per patient brings the reagent cost to about USD1500; as batching is not required,

turnaround can be very quick, 1–3 days, and the single assay can detect both FSHD type 1 and type 2. By comparison, FSHD testing via Southern blot is around a third of the price but turnaround is typically 6–8 weeks, although batching requirements can draw this out to 6 months; if the result is negative for FSHD type 1, further tests for FSHD type 2 increase the time and cost. Third, adaptive sampling has recently emerged as an alternative to Cas9 for targeted sequencing (Loose et al. 2016), and may be able to further simplify library preparation while also capturing sequences further upstream of and downstream from the D4Z4 array, obviating the need for separate whole-genome sequencing. Adaptive sampling targets regions of interest based on an easily editable browser extensible data (BED) file; therefore, another benefit of this approach is that it could allow additional genetic loci to be added with the discovery of new causative variants for FSHD, and enable optional inclusion of loci responsible for other neuromuscular diseases (NMDs), with no added cost or complexity in library preparation. Indeed, previous studies have demonstrated the efficacy of adaptive sampling for diagnosing short-tandem repeat (STR) expansion-associated NMDs (Miyatake et al. 2022; Stevanovski et al. 2022), and we imagine that STRs, the D4Z4 region, FSHD2-associated genes, and causative loci for other genetic NMDs could be combined into a single neuromuscular gene panel for targeted sequencing.

Finally, there are still many unknowns surrounding D4Z4 regulation, and our workflow could help to answer fundamental questions in macrosatellite biology, such as further investigating the dynamic changes in D4Z4 methylation across different cell types, stages of development, and disease states. Indeed, in addition to embryonic development and FSHD, *DUX4* expression is also seen physiologically in the human testis and thymus, and pathologically in several solid cancers, inflammatory disorders, and herpesvirus infection (Mocciaro et al. 2021). Elucidating the epigenetic changes in these contexts may help to reveal the underlying mechanisms of *DUX4* silencing and reactivation, paving the way for the identification of new therapeutic targets. Additionally, with the recent development of simultaneous profiling techniques such as DiMeLo-seq (Altemose et al. 2022), nanoNOMe (Lee et al. 2020) and nanoHiMe-seq (Yue et al. 2022), the long-read, single-molecule approach presented here has the potential to be extended to the exploration of even more dimensions of D4Z4 regulation, such as *SMCHD1* and CTCF binding, nucleosome occupancy, and histone modifications, alongside genetic sequence and methylation. Our workflow for investigating macrosatellite (epi)genetic landscapes therefore provides a means to explore alterations in the molecular machinery in functional studies, investigate the effects of emerging gene therapies, and map these dynamic and intriguing repetitive regions of the genome with more depth, detail and precision than previously possible.

## Methods

### Patient sample collection and cell culture

All participants provided signed informed consent for research and publication at the time of recruitment. The study complied with the Declaration of Helsinki. This study received ethics approval from the Walter Eliza Hall Institute (Ethics ID: 20/16B) and the French Agency of Biomedicine (CRBAP-HM OCP05P01E003; Étude D-2023-FSHD). FSHD and BAMS patients were clinically diagnosed and have previously received clinical genetic testing for D4Z4 array size (FSHD and controls), D4Z4 methylation (FSHD)

and *SMCHD1* genotype (FSHD, BAMS, and controls), as described in Supplemental Table S3. Controls were previously confirmed to have D4Z4 array size >10 units and to have no pathogenic *SMCHD1* variant. Primary fibroblasts from FSHD and BAMS patients and controls were obtained via skin biopsies. Fibroblasts were cultured in DMEM with 4.5 g/L D-glucose, FBS South American, Penicillin-Streptomycin 5000 U/mL, and GlutaMAX (Thermo Fisher Scientific), and incubated at 37°C, 5% CO<sub>2</sub>/9% air in a humidified incubator.

### DNA extraction

DNA was extracted from primary fibroblasts using the Monarch high-molecular weight (HMW) DNA extraction kit for cells and blood as per the protocol (NEB T3050) using 300 rpm agitation, and eluted in elution buffer (EB) from the relevant ONT kit used for library preparation (SQK-ULK001, SQK-ULK114 or SQK-CS9109).

### ONT library preparation and sequencing

For ultra-long whole-genome sequencing, libraries were prepared using the ONT ultra-long DNA sequencing kit (either SQK-ULK001 or SQK-ULK114) according to the relevant ONT protocol. For Cas9-targeted sequencing, Cas9 crRNAs targeting 4q/10q, *SMCHD1*, *DNMT3B*, and *LRIF1* (Supplemental Table S3) were designed using CHOPCHOP (v3; Labun et al. 2019) (GRCh38 or CHM13 T2T v1.1, CRISPR-Cas9, nanopore enrichment, with efficiency score parameter “Doench et al. 2014 – only for NGG PAM”). Cas9 RNP formation and DNA library preparation for Cas9-targeted sequencing were performed using the ONT Cas9 Sequencing Kit (SQK-CS9109) as per the ONT protocol, using the Long Fragment Buffer in the last wash to retain fragments of 3 kb and longer. Libraries were loaded on R9.4.1 PromethION flow cells (FLO-PRO002) if prepared using SQK-ULK001 or SQK-CS9109, and on R10.4.1 (FLO-PRO114M) if prepared using SQK-ULK114 (Supplemental Table S7), and sequenced to exhaustion on a PromethION P24 (72 h).

### Basecalling

Reads from FAST5 files were basecalled using the latest Guppy version at the time of sequencing (v6.1.1 or v6.3.4) using the modified basecalling model `dna_r9.4.1_450bps_modbases_5mc_cg_sup_prom.cfg`. Reads from pod5 files from later sequencing runs were basecalled using the latest Dorado version at the time of sequencing (v0.3.2 or v0.3.3), using model `dna_r10.4.1_e8.2_400bps_sup@v4.2.0` with the argument `--modified-bases 5mCG_5hmCG` for concurrent modified basecalling.

### Haplotyping pipeline

Processing of FASTQ files to identify, annotate and haplotype 4q and 10q reads was performed using a custom pipeline, D4Z4End2End. Briefly, reads were mapped against a D4Z4 reference sequence (the first full KpnI-KpnI D4Z4 unit from the 4q array of CHM13v1.0) using minimap2 (v2.24, Li 2018), to identify D4Z4-containing reads and count the number of D4Z4 units. These reads were then mapped against the 4q and 10q regions of CHM13v2.0, and secondary and supplementary alignments were filtered out using `samtools view -F 2304` (Li 2011), to identify 4q- and 10q-specific D4Z4 reads and assign them to 4q or 10q haplotypes. D4Z4-containing reads were also mapped against reference sequences for p13E-11 (from Hewitt et al. 1994), pLAM (sequence from end of last D4Z4 to end of *DUX4* exon 3 in 4q allele of CHM13v2.0), and the 4qB distal region (from GRCh38) for read

annotation. Reads containing pLAM and qB-sequence were assigned to haplotype A and B, respectively. Reads containing both p13E-11 and a D4Z4-distal feature (pLAM or qB-sequence) were classified as spanning reads, and all other reads were classified as nonspanning reads. Reads were also searched for exact matches for XapI sites, BlnI sites, and 4qA/10qA-specific poly(A)-signal sequences using regular expression matching in Python. Noise-Cancelling Repeat Finder (v1.01.00, Harris et al. 2019) was used to search reads for beta-satellite repeats (from Greig and Willard 1992) and telomeric repeats. The interactive visualization tool for exploration of haplotyped reads was created using *D3.js* and is publicly available alongside the D4Z4End2End workflow. The haplotyping pipeline was used to process the ultra-long whole-genome sequencing and Cas9-targeted sequencing data from all patient samples, and publicly available ONT sequencing data for CHM13v2.0 from the Telomere to Telomere (T2T) Consortium (accessed from [s3://human-pangenomics/T2T/CHM13/nanopore/rel8-guppy-5.0.7/reads.fastq.gz](https://s3://human-pangenomics/T2T/CHM13/nanopore/rel8-guppy-5.0.7/reads.fastq.gz)), 1KGP-ONT (Gustafson et al. 2024) data available on <https://s3.amazonaws.com/1000g-ont/index.html>, and Human Pangenome Consortium draft assemblies and sequence data available at GitHub ([https://github.com/human-pangenomics/HPP\\_Year1\\_Assemblies](https://github.com/human-pangenomics/HPP_Year1_Assemblies)) (Liao et al. 2023).

### D4Z4 genetic analysis

For comparison of the haplotyping pipeline results for CHM13 and the HPRC cell lines to their corresponding assemblies (CHM13v2.0 and HPRC assemblies), D4Z4 arrays in the assemblies were annotated by mapping against the D4Z4, p13E-11, pLAM and 4qB reference sequences described above, using minimap2 (Li 2018).

Consensus sequences for patient 4q and 10q alleles were generated from spanning reads identified by the haplotyping pipeline, using a two-step process: (1) minimap2 (Li 2018) for overlap detection, followed by (2) Racon (v1.4.20) (Vaser et al. 2017) for consensus calling, using the `--no-trimming` argument and using the longest spanning read as the target sequence for correction. minimap2 (Li 2018) was used to extract 4q/10q D4Z4 units from the consensus sequences and from the CHM13v2.0 and GRCh38 assemblies (GRCh38 accessed from <http://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/>), and to map them against a 4qA-type D4Z4 reference sequence. The 4qA-type D4Z4 reference sequence was generated by performing a MUSCLE (v3.8) (Edgar 2004) multiple sequence alignment of all full-length D4Z4 units from the CHM13v2.0 4q allele, and taking the most frequent value at each position (nucleotide or gap). D4Z4 alignments were visualized using Integrative Genomics Viewer (v2.17.4) (Robinson et al. 2011). MUSCLE was also used for multiple sequence alignments of distal D4Z4 sequences, which were visualized using the *R* package *ggmsa* (Zhou et al. 2022).

For the BAMS1 sample, WhatsHap (v2.1) (Martin et al. 2016) was used to phase the nonspanning 4qA reads that overlapped the proximal end of the D4Z4 array.

### FSHD2 gene panel variant calling

BAM files for patient samples were run through DeepVariant (v1.5.0) (Poplin et al. 2018) using model type ONT\_R104, to generate variant call format (VCF) files targeting *SMCHD1*, *DNMT3B*, and *LRIF1*. VCF files were filtered for “PASS” calls using BCFtools (Danecek et al. 2021). VCF files were input into Variant Effect Predictor (VEP) (v107.0) (McLaren et al. 2016) using the T2T-CHM13v2.0 VEP cache file, to annotate variant calls for potential functional impact. Variant calls overlapping *SMCHD1*, *DNMT3B*, or *LRIF1* with IMPACT=HIGH or IMPACT=MODERATE as output by VEP were further assessed by comparison to previously reported

pathogenic *SMCHD1* variants (<https://databases.lovd.nl/shared/genes/SMCHD1>) and variants in healthy controls, as well as variant pathogenicity predictions from AlphaMissense (Cheng et al. 2023).

### Methylation analysis

To generate allele-specific methylation plots, haplotyped reads were extracted from the modBAM files and re-aligned using minimap2 to either the allele's Racon consensus sequence, or a raw read from the allele if no consensus sequence was generated. For samples with high coverage Cas9-targeted sequencing (17,706, 19,187, BAMS9), all 4q and 10q reads were mapped against the full set of consensus sequences for the sample's alleles, to haplotype both spanning and nonspanning reads. For samples with lower coverage, reads included for methylation analysis were either spanning reads, or reads that could be confidently assigned to an allele based on allele-specific differences in SNVs (XapI or BlnI sites for 4q vs. 10q), distal sequences (AS/AM/AL/B haplotype) and/or number of repeat units. Allele-specific methylation plots were generated using NanoMethViz (Su et al. 2021, 2025), using the `plot_region()` function with `smoothing_window=2000`. Annotations shown below the methylation plots were generated using minimap2, as described above.

To perform allele-specific autocorrelation of 5mC signals, bedmethyl files were first produced from the re-aligned modBAM files using modkit (v0.2.5) (<https://github.com/nanoporetech/modkit>) with the `--cpg` and `--combine-strands` arguments. A vector of length  $n$  where  $n$  is the length in nucleotides of the reference sequence for the allele was populated with the %5mC (“fraction modified” from modkit) for all CpG sites with coverage  $\geq 5$ , and “NA” otherwise. The `acf()` function within *R* was then used to calculate the autocorrelation between %5mC values across the length of the vector, and the results were plotted for lag values (corresponding to distance in nucleotides between CpG sites) of up to 10,000 nucleotides (approximately three full D4Z4 units) and up to 500 nucleotides. Allele-specific %5mC plots were generated from the bedmethyl files, once again filtering for CpG sites with coverage  $\geq 5$ , using ggplot2 (v3.5.1) (Wickham 2011). Smoothed lines were generated using the `geom_smooth()` function using the LOESS method with `span = 0.2`.

Mean methylation rate was calculated using the `region_methy_stats()` function within NanoMethViz (Su et al. 2021, 2025), using a threshold of 0.5 to call methylated and unmethylated sites. Regression lines for the scatterplots of mean methylation prevalence versus number of repeats were produced using `geom_line(stat="smooth", method="lm")`, and Pearson and Spearman's correlations were calculated using the `cor()` function within *R*. For comparison of the mean methylation prevalence of the DR1 site and *DUX4* TSSs within each repeat unit, the DR1 site was defined as nucleotides 563–814 of the D4Z4 unit with respect to the start of the KpnI site (Hartweck et al. 2013), and the *DUX4* TSS region was defined as –200nt to +200nt with respect to the *DUX4* TSS (nucleotide 10,732 in GenBank AF117653.3).

### PCR genotyping of the distal 4qAM variant

We designed two forward primers for the novel 4qAM variant, to be compatible with the reverse primer from (Lemmers et al. 2018b) and resulting in 345 and 565 bp amplicons (sequences in Supplemental Table S5). 12.5  $\mu$ L PCRs were set up for 4qAS, 4qAM and 4qAL alleles, each with 20 ng genomic DNA, 2 $\times$  GoTaq Green MasterMix (Promega) and 0.25  $\mu$ M concentration for each primer pair. Following (Lemmers et al. 2018b), the 4qAS amplification used a touchdown protocol whereas the two 4qAM



and the 4qAL amplifications used a regular PCR protocol. Touchdown protocol: initial denaturation of 2' at 95°C, followed by 10 cycles of touchdown PCR: 95°C 30", 70>60°C 30", 72°C 30"; then 26 cycles (95°C 30", 60°C 30", 72°C 30"), and a 5' final extension at 72°C. Regular protocol: 95°C for 2', 35 cycles of (95°C 30", 60°C 30", 72°C 30"), and a 5' final extension at 72°C. PCR products were run on a 1.8% agarose gel.

### Code availability

The D4Z4End2End pipeline inclusive of the interactive visualization tool is freely available at GitHub (<https://github.com/lucindaxiao/D4Z4End2End>) under an MIT licence, and archived as Supplemental Code.

### Data access

All patient sequencing data generated in this study have been submitted to the European Genome-Phenome Archive (EGA; <https://ega-archive.org/>) under accession number EGAD50000001551.

### Competing interest statement

Q.G. and L.C.X. received travel support from ONT to attend a conference.

### Acknowledgments

We thank Stephen Wilcox and Sarah MacRaidl from the WEHI Genomics Platform for their assistance with nanopore sequencing; and Kelsey Breslin and Hannah Vanyai for their help with cell culture. M.E.R., M.E.B., and Q.G. are supported by Australian National Health and Medical Research Council (NHMRC) Investigator Grants (GNT2017257, GNT1194345 plus GNT2041117, and GNT2007996, respectively). This work was supported by an early career research grant from the Brockhoff foundation and the Marian and E.H. Flack Trust to Q.G. The Walter and Eliza Hall Institute receives support from the Victorian State Government through its Operational Infrastructure Support Program. Additional support was provided by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS) program and an Australian National Health and Medical Research Council Independent Research Institutes Infrastructure Support Scheme (IRISS) grant (9000719).

**Author contributions:** Study design: F.M., M.E.B., and Q.G.; data collection: L.C.X., B.S.J., J.L., F.M., and Q.G.; analysis pipeline development: L.C.X., A.S., K.Z., and S.S.; analysis and interpretation of the results: L.C.X., A.S., S.X., B.R., M.E.R., F.M., M.E.B., and Q.G.; supervision: M.E.R., M.E.B., and Q.G.; draft manuscript preparation: L.C.X., M.E.R., M.E.B., and Q.G.

### References

Altemose N, Maslan A, Smith OK, Sundararajan K, Brown RR, Mishra R, Detweiler AM, Neff N, Miga KH, Straight AF, et al. 2022. DiMeLo-seq: a long-read, single-molecule method for mapping protein–DNA interactions genome wide. *Nat Methods* **19**: 711–723. doi:10.1038/s41592-022-01475-6

Anseau E, Laoudj-Chenivess D, Marcowycz A, Tassin A, Vanderplanck C, Sauvage S, Barro M, Mahieu I, Leroy A, Leclercq I, et al. 2009. Dux4c is up-regulated in FSHD. It induces the MYF5 protein and human myoblast proliferation. *PLoS One* **4**: e7482. doi:10.1371/journal.pone.0007482

Bakker E, Wijmenga C, Vossen RHAM, Padberg GW, Hewitt J, van Der Wielen M, Rasmussen K, Frants RR. 1995. The FSHD-linked locus D4F104S1 (p13E-11) on 4q35 has a homologue on 10qter. *Muscle Nerve* **18**: S39–S44. doi:10.1002/(ISSN)1097-4598

Benarroch L, Boëlle PY, Madry H, Mohand Oumoussa B, Eura N, Nishino I, Labrèche K, Gorbunova V, Bassez G, Stojkovic T, et al. 2025. Comparative analysis of CRISPR/Cas9-targeted nanopore sequencing approaches in repeat expansion disorders. *Genomics Proteomics Bioinformatics* qzaf094. doi:10.1093/gpbjnl/qzaf094

Busto-Moner L, Morival J, Ren H, Fahim A, Reitz Z, Downing TL, Read EL. 2020. Stochastic modeling reveals kinetic heterogeneity in post-replication DNA methylation. *PLoS Comput Biol* **16**: e1007195. doi:10.1371/journal.pcbi.1007195

Butterfield RJ, Dunn DM, Duval B, Moldt S, Weiss RB. 2023. Deciphering D4Z4 CpG methylation gradients in facioscapulohumeral muscular dystrophy using nanopore sequencing. *Genome Res* **33**: 1439–1454. doi:10.1101/gr.277871.123

Cacurri S, Piazzi N, Deidda G, Vigneti E, Galluzzi G, Colantoni L, Merico B, Ricci E, Felicetti L. 1998. Sequence homology between 4qter and 10qter loci facilitates the instability of subtelomeric KpnI repeat units implicated in facioscapulohumeral muscular dystrophy. *Am J Hum Genet* **63**: 181–190. doi:10.1086/301906

Calandra P, Cascino I, Lemmers RJLF, Galluzzi G, Teveroni E, Monforte M, Tasca G, Ricci E, Moretti F, van der Maarel SM, et al. 2016. Allele-specific DNA hypomethylation characterises FSHD1 and FSHD2. *J Med Genet* **53**: 348–355. doi:10.1136/jmedgenet-2015-103436

Campbell AE, Belleville AE, Resnick R, Shadle SC, Tapscott SJ. 2018. Facioscapulohumeral dystrophy: activating an early embryonic transcriptional program in human skeletal muscle. *Hum Mol Genet* **27**: R153–R162. doi:10.1093/hmg/ddy162

Caputo V, Megalizzi D, Fabrizio C, Termine A, Colantoni L, Bax C, Gimenez J, Monforte M, Tasca G, Ricci E, et al. 2022a. D4Z4 methylation levels combined with a machine learning pipeline highlight single CpG sites as discriminating biomarkers for FSHD patients. *Cells* **11**: 4114. doi:10.3390/cells11244114

Caputo V, Megalizzi D, Fabrizio C, Termine A, Colantoni L, Caltagirone C, Giardina E, Cascella R, Straffella C. 2022b. Update on the molecular aspects and methods underlying the complex architecture of FSHD. *Cells* **11**: 2687. doi:10.3390/cells11172687

Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Seargent T, et al. 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**: eadg7492. doi:10.1126/science.adg7492

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of samtools and bcftools. *GigaScience* **10**: giab008. doi:10.1093/gigascience/giab008

de Greef JC, Lemmers RJ, van Engelen BG, Sacconi S, Venance SL, Frants RR, Tawil R, van der Maarel SM. 2009. Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent fshd. *Hum Mutat* **30**: 1449–1459. doi:10.1002/humu.v30:10

De Greef J, Lemmers R, Camano P, Day J, Sacconi S, Dunand M, Van Engelen B, Kiuru-Enari S, Padberg G, Rosa AL, et al. 2010. Clinical features of facioscapulohumeral muscular dystrophy 2. *Neurology* **75**: 1548–1554. doi:10.1212/WNL.0b013e3181f96175

De Iaco A, Planet E, Coluccio A, Verp S, Duc J, Trono D. 2017. DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat Genet* **49**: 941–945. doi:10.1038/ng.3858

Deenen JC, Arnts H, Van Der Maarel SM, Padberg GW, Verschuuren JJ, Bakker E, Weinreich SS, Verbeek AL, Van Engelen BG. 2014. Population-based incidence and prevalence of facioscapulohumeral dystrophy. *Neurology* **83**: 1056–1059. doi:10.1212/WNL.0000000000000797

Deidda G, Cacurri S, Grisanti P, Vigneti E, Piazzi N, 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. *Eur J Hum Genet* **3**: 155–167. doi:10.1159/000472291

Delourme M, Charlene C, Gerard L, Ganne B, Perrin P, Vovan C, Bertaux K, Nguyen K, Bernard R, Magdiner F. 2023. Complex 4q35 and 10q26 rearrangements: a challenge for molecular diagnosis of patients with facioscapulohumeral dystrophy. *Neurol Genet* **9**: e200076. doi:10.1212/NXG.00000000000020076

Dion C, Roche S, Laberthonnière C, Brouqsault N, Mariot V, Xue S, Gurzaul AD, Nowak A, Gordon CT, Gaillard MC, et al. 2019. SMCHD1 is involved in de novo methylation of the DUX4-encoding D4Z4 macrosatellite. *Nucleic Acids Res* **47**: 2822–2839. doi:10.1093/nar/gkz005

Duranti E, Villa C. 2023. Influence of DUX4 expression in facioscapulohumeral muscular dystrophy and possible treatments. *Int J Mol Sci* **24**: 9503. doi:10.3390/ijms24119503

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797. doi:10.1093/nar/gkh340

- Eisfeldt J, Ek M, Nordenskjöld M, Lindstrand A. 2025. Toward clinical long-read genome sequencing for rare diseases. *Nat Genet* **57**: 1334–1343. doi:10.1038/s41588-025-02160-y
- Erdmann H, Scharf F, Gehling S, Benet-Pagès A, Jakubiczka S, Becker K, Seipelt M, Kleefeld F, Knop KC, Prott EC, et al. 2023. Methylation of the 4q35 D4Z4 repeat defines disease status in facioscapulohumeral muscular dystrophy. *Brain* **146**: 1388–1402. doi:10.1093/brain/awac336
- Gaillard MC, Roche S, Dion C, Tasmdjian A, Bouget G, Salort-Campana E, Vovan C, Chaix C, Brouqsault N, Morere J, et al. 2014. Differential DNA methylation of the D4Z4 repeat in patients with FSHD and asymptomatic carriers. *Neurology* **83**: 733–742. doi:10.1212/WNL.0000000000000708
- Gaillard MC, Brouqsault N, Morere J, Laberthonnère C, Dion C, Badja C, Roche S, Nguyen K, Magdinier F, Robin JD. 2019. Analysis of the 4q35 chromatin organization reveals distinct long-range interactions in patients affected with Facio-Scapulo-Humeral Dystrophy. *Sci Rep* **9**: 10327. doi:10.1038/s41598-019-46861-x
- Gérard L, Delourme M, Tardy C, Ganne B, Perrin P, Chaix C, Trani JP, Eudes N, Laberthonnère C, Bertaux K, et al. 2025. SMCHD1 genetic variants in type 2 facioscapulohumeral dystrophy and challenges in predicting pathogenicity and disease penetrance. *Eur J Hum Genet* **33**: 784–792. doi:10.1038/s41431-024-01781-x
- Giardina E, Camaño P, Burton-Jones S, Ravenscroft G, Henning F, Magdinier F, Van Der Stoep N, Van Der Vliet PJ, Bernard R, Tomaselli PJ, et al. 2024. Best practice guidelines on genetic diagnostics of facioscapulohumeral muscular dystrophy: update of the 2012 guidelines. *Clin Genet* **106**: 13–26. doi:10.1111/cge.v106.1
- Gordon CT, Xue S, Yigit K, Filali H, Chen K, Rosin N, Yoshiura Ki, Oufadem M, Beck TJ, McGowan R, et al. 2017. De novo mutations in SMCHD1 cause Bosma arhinia microphthalmia syndrome and abrogate nasal development. *Nat Genet* **49**: 249–255. doi:10.1038/ng.3765
- Goselink RJ, Mul K, van Kernebeek CR, Lemmers RJ, van der Maarel SM, Schreuder TH, Erasmus CE, Padberg GW, Statland JM, Voermans NC, et al. 2019. Early onset as a marker for disease severity in facioscapulohumeral muscular dystrophy. *Neurology* **92**: e378–e385. doi:10.1212/WNL.00000000000006819
- Goto K, Nishino I, Hayashi YK. 2004. Very low penetrance in 85 Japanese families with facioscapulohumeral muscular dystrophy 1A. *J Med Genet* **41**: e12. doi:10.1136/jmg.2003.008755
- Gould T, Jones TI, Jones PL. 2021. Precise epigenetic analysis using targeted bisulfite genomic sequencing distinguishes FSHD1, FSHD2, and healthy subjects. *Diagnostics* **11**: 1469. doi:10.3390/diagnostics.11081469
- Greig GM, Willard HF. 1992.  $\beta$  satellite DNA: characterization and localization of two subfamilies from the distal and proximal short arms of the human acrocentric chromosomes. *Genomics* **12**: 573–580. doi:10.1016/0888-7543(92)90450-7
- Gustafson JA, Gibson SB, Damaraju N, Zalusky MP, Hoekzema K, Twesigomwe D, Yang L, Snead AA, Richmond PA, De Coster W, et al. 2024. High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res* **34**: 2061–2073. doi:10.1101/gr.279273.124
- Hamanaka K, Sikrová D, Mitsuhashi S, Masuda H, Sekiguchi Y, Sugiyama A, Shibuya K, Lemmers RJ, Goossens R, Ogawa M, et al. 2020. Homozygous nonsense variant in LRIF1 associated with facioscapulohumeral muscular dystrophy. *Neurology* **94**: e2441–e2447. doi:10.1212/WNL.00000000000009617
- Harris RS, Cechova M, Makova KD. 2019. Noise-cancelling repeat finder: uncovering tandem repeats in error-prone long-read sequencing data. *Bioinformatics* **35**: 4809–4811. doi:10.1093/bioinformatics/btz484
- Hartweck LM, Anderson LJ, Lemmers RJ, Dandapat A, Toso EA, Dalton JC, Tawil R, Day JW, van der Maarel SM, Kyba M. 2013. A focal domain of extreme demethylation within D4Z4 in FSHD2. *Neurology* **80**: 392–399. doi:10.1212/WNL.0b013e3182f075c
- Hewitt JE, Lyle R, Clark LN, Valleley EM, Wright TJ, Wijmenga C, van Deutekom JC, Francis F, Sharpe PT, Hofker M, et al. 1994. Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystrophy. *Hum Mol Genet* **3**: 1287–1295. doi:10.1093/hmg/3.8.1287
- Himeda CL, Jones PL. 2019. The genetics and epigenetics of facioscapulohumeral muscular dystrophy. *Annu Rev Genomics Hum Genet* **20**: 265–291. doi:10.1146/genom.2019.20.issue-1
- Hiramuki Y, Kure Y, Saito Y, Ogawa M, Ishikawa K, Mori-Yoshimura M, Oya Y, Takahashi Y, Kim DS, 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. *J Transl Med* **20**: 517. doi:10.1186/s12967-022-03743-7
- Huang M, Zhang Q, Jiao J, Shi J, Xu Y, Zhang C, Zhou R, Liu W, Liang Y, Chen H, et al. 2024. Comprehensive genetic analysis of facioscapulohumeral muscular dystrophy by Nanopore long-read whole-genome sequencing. *J Transl Med* **22**: 451. doi:10.1186/s12967-024-05259-8
- Huichalaf C, Micheloni S, Ferri G, Caccia R, Gabellini D. 2014. DNA methylation analysis of the macrosatellite repeat associated with FSHD muscular dystrophy at single nucleotide level. *PLoS One* **9**: e115278. doi:10.1371/journal.pone.0115278
- Jones TI, Yan C, Sapp PC, McKenna-Yasek D, Kang PB, Quinn C, Salameh JS, King OD, Jones PL. 2014. Identifying diagnostic DNA methylation profiles for facioscapulohumeral muscular dystrophy in blood and saliva using bisulfite sequencing. *Clin Epigenetics* **6**: 23. doi:10.1186/1868-7083-6-23
- Kerr L, Kafetzopoulos I, Grima R, Sproul D. 2023. Genome-wide single-molecule analysis of long-read DNA methylation reveals heterogeneous patterns at heterochromatin that reflect nucleosome organisation. *PLoS Genet* **19**: e1010958. doi:10.1371/journal.pgen.1010958
- Kolesnikov A, Cook D, Nattestad M, Brambrink L, McNulty B, Gorzynski J, Goenka S, Ashley EA, Jain M, Miga KH, et al. 2024. Local read haplotagging enables accurate long-read small variant calling. *Nat Commun* **15**: 5907. doi:10.1038/s41467-024-50079-5
- Laberthonnère C, Novoa-del Toro EM, Delourme M, Chevalier R, Brouqsault N, Mazaleyra K, Streichenberger N, Manel V, Bernard R, Salort Campana E, et al. 2022. Facioscapulohumeral dystrophy weakened sarcomeric contractility is mimicked in induced pluripotent stem cells-derived innervated muscle fibres. *J Cachexia Sarcopenia Muscle* **13**: 621–635. doi:10.1002/jcsm.v13.1
- Labun K, Montague TG, Krause M, Torres Cleuren YN, Tjeldnes H, Valen E. 2019. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res* **47**(W1): W171–W174. doi:10.1093/nar/gkz365
- Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, Sadowski N, Sedlazeck FJ, Hansen KD, Simpson JT, Timp W. 2020. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat Methods* **17**: 1191–1199. doi:10.1038/s41592-020-01000-7
- Lemmers RJ, de Kievit P, Sandkuijl L, Padberg GW, van Ommen GJB, Frants RR, van der Maarel SM. 2002. Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. *Nat Genet* **32**: 235–236. doi:10.1038/ng999
- Lemmers RJLF, Overveld PGMv, Sandkuijl LA, Vrieling H, Padberg GW, Frants RR, Maarel SMvd. 2004. Mechanism and timing of mitotic rearrangements in the subtelomeric D4Z4 repeat involved in facioscapulohumeral muscular dystrophy. *Am J Hum Genet* **75**: 44–53. doi:10.1086/422175
- Lemmers RJ, Wohlgemuth M, van der Gaag KJ, van der Vliet PJ, van Teijlingen CM, de Knijff P, Padberg GW, Frants RR, van der Maarel SM. 2007. Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. *The American Journal of Human Genetics* **81**: 884–894. doi:10.1086/521986
- Lemmers RJLF, van der Vliet PJ, Klooster R, Sacconi S, Camano P, Dauwerse JG, Snider L, Straasheijm KR, Jan van Ommen G, Padberg GW, et al. 2010a. A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**: 1650–1653. doi:10.1126/science.1189044
- Lemmers RJLF, van der Vliet PJ, van der Gaag KJ, Zuniga S, Frants RR, de Knijff P, van der Maarel SM. 2010b. Worldwide population analysis of the 4q and 10q subtelomeres identifies only four discrete interchromosomal sequence transfers in human evolution. *Am J Hum Genet* **86**: 364–377. doi:10.1016/j.ajhg.2010.01.035
- Lemmers RJ, Tawil R, Petek LM, Balog J, Block GJ, Santen GW, Amell AM, van Der Vliet PJ, Almomani R, Straasheijm KR, 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. doi:10.1038/ng.2454
- Lemmers RJ, Goeman JJ, van der Vliet PJ, van Nieuwenhuizen MP, Balog J, Vos-Versteeg M, Camano P, Ramos Arroyo MA, Jerico I, Rogers MT, et al. 2015. Inter-individual differences in CpG methylation at D4Z4 correlate with clinical variability in FSHD1 and FSHD2. *Hum Mol Genet* **24**: 659–669. doi:10.1093/hmg/ddu486
- Lemmers RJLF, van der Vliet PJ, Vreijling JP, Henderson D, van der Stoep N, Voermans N, van Engelen B, Baas F, Sacconi S, Tawil R, et al. 2018a. Cis D4Z4 repeat duplications associated with facioscapulohumeral muscular dystrophy type 2. *Hum Mol Genet* **27**: 3488–3497. doi:10.1093/hmg/ddy236
- Lemmers RJ, van der Vliet PJ, Balog J, Goeman JJ, Arindarto W, Krom YD, Straasheijm KR, Debipersad RD, Özel G, Sowden J, et al. 2018b. Deep characterization of a common D4Z4 variant identifies biallelic DUX4 expression as a modifier for disease penetrance in FSHD2. *Eur J Hum Genet* **26**: 94–106. doi:10.1038/s41431-017-0015-0
- Lemmers RJLF, van der Vliet PJ, Blatnik A, Balog J, Zidar J, Henderson D, Goselink R, Tapscott SJ, Voermans NC, Tawil R, et al. 2022. Chromosome 10q-linked FSHD identifies DUX4 as principal disease gene. *J Med Genet* **59**: 180–188. doi:10.1136/jmedgenet-2020-107041
- Lemmers RJLF, Butterfield R, van der Vliet PJ, de Bleecker JL, van der Pol L, Dunn DM, Erasmus CE, D'Hooghe M, Verhoeven K, Balog J, et al. 2023.

- Autosomal dominant *in cis* D4Z4 repeat array duplication alleles in facioscapulohumeral dystrophy. *Brain* **147**: 414–426. doi:10.1093/brain/awad312
- Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987–2993. doi:10.1093/bioinformatics/btr509
- Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100. doi:10.1093/bioinformatics/bty191
- Li K, Quiat D, She F, Liu Y, He R, Haghighi A, Liu F, Zhang R, DePalma SR, Yang Y, et al. 2024. Genetic diagnosis of facioscapulohumeral muscular dystrophy type 1 using rare-variant linkage analysis and long-read genome sequencing. *Genet Med Open* **2**: 101817. doi:10.1016/j.gimo.2024.101817
- Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617**: 312–324. doi:10.1038/s41586-023-05896-x
- Loose M, Malla S, Stout M. 2016. Real-time selective sequencing using nanopore technology. *Nat Methods* **13**: 751–754. doi:10.1038/nmeth.3930
- Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, Schöenhuth A, Marschall T. 2016. Whatshap: fast and accurate read-based phasing. *bioRxiv* doi:10.1101/085050
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* **17**: 122. doi:10.1186/s13059-016-0974-4
- Miyatake S, Koshimizu E, Fujita A, Doi H, Okubo M, Wada T, Hamanaka K, Ueda N, Kishida H, Minase G, et al. 2022. Rapid and comprehensive diagnostic method for repeat expansion diseases using nanopore sequencing. *NPJ Genom Med* **7**: 62. doi:10.1038/s41525-022-00331-y
- Mocciaro E, Runfola V, Ghezzi P, Pannese M, Gabellini D. 2021. DUX4 role in normal physiology and in fshd muscular dystrophy. *Cells* **10**: 3322. doi:10.3390/cells10123322
- Mohassel P, Chang N, Inoue K, Delaney A, Hu Y, Donkervoort S, Saade D, Billioux BJ, Meader B, Volochayev R, et al. 2022. Cross-sectional neuromuscular phenotyping study of patients with arhinia with SMCHD1 variants. *Neurology* **98**: e1384–e1396. doi:10.1212/WNL.0000000000200032
- Mostacciolo M, Pastorello E, Vazza G, Miorin M, Angelini C, Tomelleri G, Galluzzi G, Trevisan C. 2009. Facioscapulohumeral muscular dystrophy: epidemiological and molecular study in a north-east Italian population sample. *Clin Genet* **75**: 550–555. doi:10.1111/cge.2009.75.issue-6
- Mul K, Lemmers RJ, Kriek M, van der Vliet PJ, van den Boogaard ML, Badrising UA, Graham JM, Lin AE, Brand H, Moore SA, et al. 2018a. FSHD type 2 and Bosma arhinia microphthalmia syndrome. *Neurology* **91**: e562–e570. doi:10.1212/WNL.0000000000005958
- Mul K, Voermans NC, Lemmers RJ, Jonker MA, van der Vliet PJ, Padberg GW, van Engelen BG, van der Maarel SM, Horlings CG. 2018b. Phenotype-genotype relations in facioscapulohumeral muscular dystrophy type 1. *Clin Genet* **94**: 521–527. doi:10.1111/cge.2018.94.issue-6
- Nguyen K, Walrafen P, Bernard R, Attarian S, Chaix C, Vovan C, Renard E, Dufrane N, Pouget J, Vannier A, et al. 2011. Molecular combing reveals allelic combinations in facioscapulohumeral dystrophy. *Ann Neurol* **70**: 627–633. doi:10.1002/ana.v70.4
- Nguyen K, Puppo F, Roche S, Gaillard MC, Chaix C, Lagarde A, Pierret M, Vovan C, Olschwang S, Salort-Campana E, et al. 2017. Molecular combing reveals complex 4q35 rearrangements in Facioscapulohumeral dystrophy. *Hum Mutat* **38**: 1432–1441. doi:10.1002/humu.2017.38.issue-10
- Nurk S, Koren S, Rhie A, Rautiainen M, Bizikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376**: 44–53. doi:10.1126/science.abj6987
- Ottaviani A, Rival-Gervier S, Boussouar A, Foerster AM, Rondier D, Sacconi S, Desnuelle C, Gilson E, Magdinier F. 2009. The D4Z4 macrosatellite repeat acts as a ctcf and A-type lamins-dependent insulator in facio-scapulo-humeral dystrophy. *PLoS Genet* **5**: e1000394. doi:10.1371/journal.pgen.1000394
- Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**: 983–987. doi:10.1038/nbt.4235
- Preston MK, Wang LH. 2025. Facioscapulohumeral muscular dystrophy. In *GeneReviews*. University of Washington, Seattle.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26. doi:10.1038/nbt.1754
- 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. doi:10.1001/jamanetworkopen.2020.4040
- 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 J Rare Dis* **10**: 2. doi:10.1186/s13023-014-0218-1
- Salsi V, Chiara M, Pini S, Kuš P, Ruggiero L, Bonanno S, Rodolico C, Previtali SC, D'Angelo MG, Maggi L, et al. 2023. A human pan-genomic analysis reconfigures the genetic and epigenetic make up of facioscapulohumeral muscular dystrophy. medRxiv doi:10.1101/2023.06.13.23291337
- Schätzl T, Kaiser L, Deigner HP. 2021. Facioscapulohumeral muscular dystrophy: genetics, gene activation and downstream signalling with regard to recent therapeutic approaches: an update. *Orphanet J Rare Dis* **16**: 129. doi:10.1186/s13023-021-01760-1
- Scholz V, Schönrock V, Erdmann H, Prokosh V, Schoedel M, Almus M, Sauer M, Mayer V, Breithausen E, Van Buren I, et al. 2024. Parallel in-depth analysis of repeat expansions: an updated Clin-CATS workflow for nanopore R10 flow cells. *bioRxiv* doi:10.1101/2024.11.05.622099
- Shaw ND, Brand H, Kupchinsky ZA, Bengani H, Plummer L, Jones TI, Erdin S, Williamson KA, Rainger J, Stortchevoi A, et al. 2017. SMCHD1 mutations associated with a rare muscular dystrophy can also cause isolated arhinia and Bosma arhinia microphthalmia syndrome. *Nat Genet* **49**: 238–248. doi:10.1038/ng.3743
- Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, et al. 2024. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* **42**: 1571–1580. doi:10.1038/s41587-023-02024-y
- Stevanovski I, Chintalapathi SR, Gamaarachchi H, Ferguson JM, Pineda SS, Scriba CK, Tchan M, Fung V, Ng K, Cortese A, et al. 2022. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci Adv* **8**: eabm5386. doi:10.1126/sciadv.abm5386
- Strafella C, Megalizzi D, Trastulli G, Proietti Piorgio E, Colantoni L, Tasca G, Monforte M, Zampatti S, Primiano G, Sancrica C, et al. 2024. Integrating D4Z4 methylation analysis into clinical practice: improvement of FSHD molecular diagnosis through distinct thresholds for 4qA/4qA and 4qA/4qB patients. *Clin Epigenetics* **16**: 148. doi:10.1186/s13148-024-01747-2
- Su S, Gouil Q, Blewitt ME, Cook D, Hickey PF, Ritchie ME. 2021. Nanomethviz: An R/Bioconductor package for visualizing long-read methylation data. *PLoS Comput Biol* **17**: e1009524. doi:10.1371/journal.pcbi.1009524
- Su S, Xiao L, Lancaster J, Cameron T, Breslin K, Hickey PF, Blewitt ME, Gouil Q, Ritchie ME. 2025. A streamlined workflow for long-read DNA methylation analysis with NanoMethViz and Bioconductor. *F1000Research* **13**: 1243. doi:10.12688/f1000research.155204.2
- Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ, et al. 2019. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res* **47**: 10994–11006. doi:10.1093/nar/gkz841
- van den Boogaard ML, Lemmers RJ, Balog J, Wohlgemuth M, Auranen M, Mitsuhashi S, van der Vliet PJ, Straasheijm KR, van den Akker RFP, Kriek M, et al. 2016. Mutations in DNMT3B modify epigenetic repression of the D4Z4 repeat and the penetrance of facioscapulohumeral dystrophy. *Am J Hum Genet* **98**: 1020–1029. doi:10.1016/j.ajhg.2016.03.013
- van der Maarel SM, Deidda G, Lemmers RJL, van Overveld PGM, van der Wielen M, Hewitt JE, Sandkuijl L, Bakker B, van Ommen GJB, Padberg GW, et al. 2000. De novo facioscapulohumeral muscular dystrophy: frequent somatic mosaicism, sex-dependent phenotype, and the role of mitotic transchromosomal repeat interaction between chromosomes 4 and 10. *Am J Hum Genet* **66**: 26–35. doi:10.1086/302730
- van Overveld PGM, Lemmers RJL, Sandkuijl LA, Enthoven L, Winokur ST, Bakels F, Padberg GW, van Ommen GJB, Frants RR, van der Maarel SM. 2003. Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. *Nat Genet* **35**: 315–317. doi:10.1038/ng1262
- Vasale J, Boyar F, Jocsom M, Sulcova V, Chan P, Liaquat K, Hoffman C, Meservey M, Chang I, Tsao D, et al. 2015. Molecular combing compared to Southern blot for measuring D4Z4 contractions in FSHD. *Neuromuscul Disord* **25**: 945–951. doi:10.1016/j.nmd.2015.08.008
- Vaser R, Sović I, Nagarajan N, Sikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**: 737–746. doi:10.1101/gr.214270.116
- Wang Q, Yu G, Ming X, Xia W, Xu X, Zhang Y, Zhang W, Li Y, Huang C, Xie H, et al. 2020. Imprecise DNMT1 activity coupled with neighbor-guided correction enables robust yet flexible epigenetic inheritance. *Nat Genet* **52**: 828–839. doi:10.1038/s41588-020-0661-y
- 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 population-based study. *Lancet Reg Health West Pac* **18**: 100323. doi:10.1016/j.lanwpc.2021.100323
- Wang Y, Zhao Z, Meng F, Kong X. 2024. Accurate prenatal diagnosis of facioscapulohumeral muscular dystrophy 1 using nanopore sequencing. *J Med Genet* **61**: 1096–1102. doi:10.1136/jmg-2023-109832
- Wickham H. 2011. ggplot2. *Wiley interdiscip Rev Comput Stat* **3**: 180–185. doi:10.1002/wics.v3.2
- Wright TJ, Wijmenga C, Clark LN, Frants RR, Williamson R, Hewitt JE. 1993. Fine mapping of the fshd gene region orientates the rearranged fragment detected by the probe p13e-11. *Hum Mol Genet* **2**: 1673–1678. doi:10.1093/hmg/2.10.1673
- Yeetong P, Kulsirichawaroj P, Kumutpongpanich T, Srichomthong C, Od-ek P, Rakwongkhachon S, Thamcharoenvipas T, Sanmaneechai O, Pongpanich M, Shotelersuk V. 2023. Long-read Nanopore sequencing identified D4Z4 contractions in patients with facioscapulohumeral muscular dystrophy. *Neuromuscul Disord* **33**: 551–556. doi:10.1016/j.nmd.2023.05.004
- Yue X, Xie Z, Li M, Wang K, Li X, Zhang X, Yan J, Yin Y. 2022. Simultaneous profiling of histone modifications and DNA methylation via nanopore sequencing. *Nat Commun* **13**: 7939. doi:10.1038/s41467-022-35650-2
- Zhou L, Feng T, Xu S, Gao F, Lam TT, Wang Q, Wu T, Huang H, Zhan L, Li L, et al. 2022. ggmsa: a visual exploration tool for multiple sequence alignment and associated data. *Brief Bioinform* **23**: bbac222. doi:10.1093/bib/bbac222

Received May 7, 2025; accepted in revised form January 1, 2026.