



Profiling the Structure and Epigenetics of D4Z4 Arrays in FSHD Using Oxford Nanopore Sequencing

Rebecca Stubbs¹, Bryan Leland¹, Priyesh Rughani¹, Carly Tyer¹, Kenney Tuyen¹, Xiaoguang (David) Dai¹, Tatiana O’Connell-Tilley¹, Heather Jeffery¹, Philipp Rescheneder¹, Phill James¹, Scott Hickey¹, Sissel Juul¹

¹Oxford Nanopore Technologies Inc., New York, NY



Abstract

Molecular analysis of facioscapulohumeral muscular dystrophy (FSHD) is challenging due to the highly repetitive D4Z4 macrosatellite array containing the *DUX4* locus, which exists at two nearly identical subtelomeric regions on chromosomes 4q35 and 10q26. FSHD results from aberrant *DUX4* expression caused either by contraction of the D4Z4 array on a permissive 4qA haplotype or by pathogenic variants in chromatin-regulating genes, each leading to hypomethylation. Current diagnostic approaches require multiple complementary techniques to assess repeat size, haplotype, sequence variation, and epigenetic state. However, in this study, we resolve D4Z4 arrays, quantify number of repeat units, and profile DNA methylation and chromatin accessibility using Oxford Nanopore multiomic sequencing. These methods accurately identified FSHD1 and FSHD2 affected haplotypes. Epigenetic profiling revealed hypomethylation and increased chromatin accessibility at pathogenic 4qA haplotypes in FSHD samples. Together, these results demonstrate that Oxford Nanopore sequencing enables comprehensive, single-platform genetic and epigenetic analysis for FSHD.

Methods

We sequenced seven Coriell cell lines with known FSHD or FSHD-like phenotypes (Fig. 1; Table 1). Three approaches were used to sequence the repeat arrays (Fig. 2). First, the Ultra-Long DNA Sequencing Kit (ULK) was used to unambiguously characterize each haplotype, with individual long reads spanning the entire array. Second, all samples were sequenced with the Ligation Sequencing Kit (LSK) to at least 30x coverage with a read-length N50 of ~30 kb, then assembled into contigs using Hifiasm (ONT)¹ *de novo* assembly. Third, Adaptive Sampling using a digital panel targeting FSHD genes was used to multiplex samples (2-plex) and obtain high read depth in targeted regions. DNA methylation was detected directly from the sequencing signal and analyzed using Modkit². In addition, chromatin stenciling experiments with exogenous 6mA were performed to investigate chromatin accessibility across the *DUX4* loci.

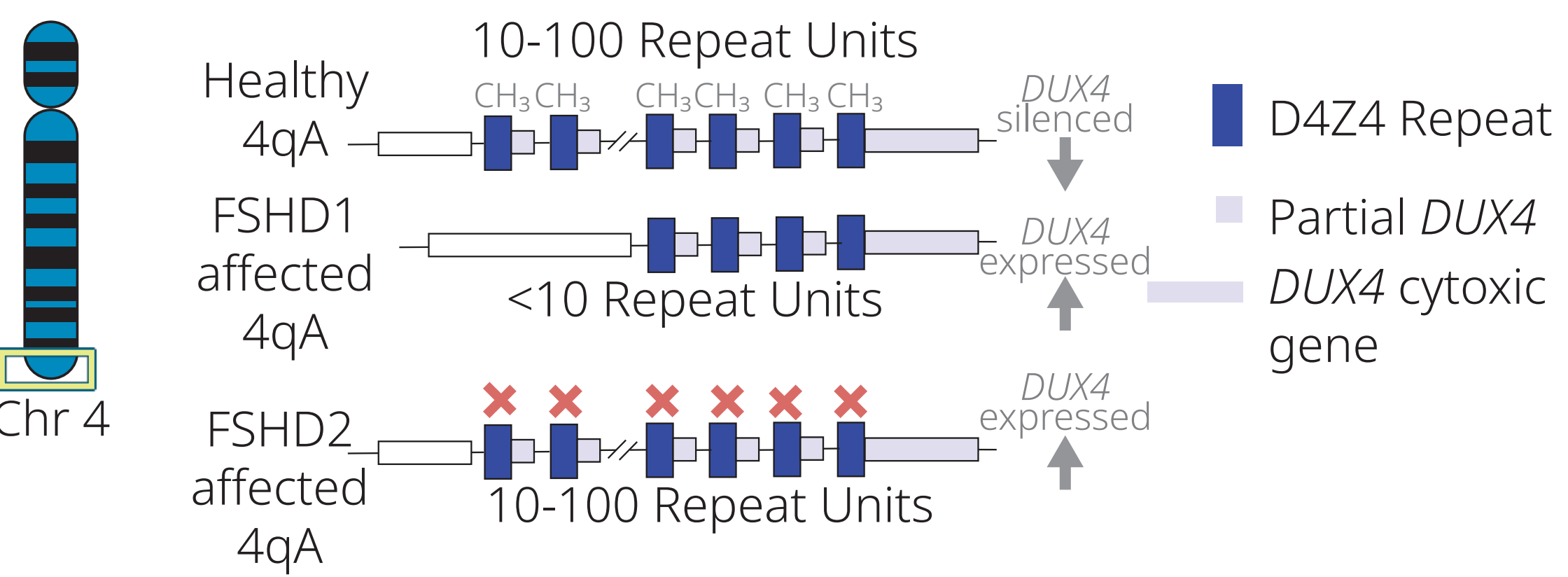


Fig. 1. Overview of FSHD. Schematic of the pathogenesis of FSHD1 and FSHD2. FSHD1 involves contraction of the D4Z4 array (<10 units) on a permissive 4qA haplotype, causing chromatin relaxation and hypomethylation, while FSHD2 results from variants in chromatin-regulating genes producing a hypomethylated state. The D4Z4 macrosatellite array at the chromosome 4q subtelomere, which comprises 1-100 3.3 kb repeat units, is the pathogenic locus of FSHD.

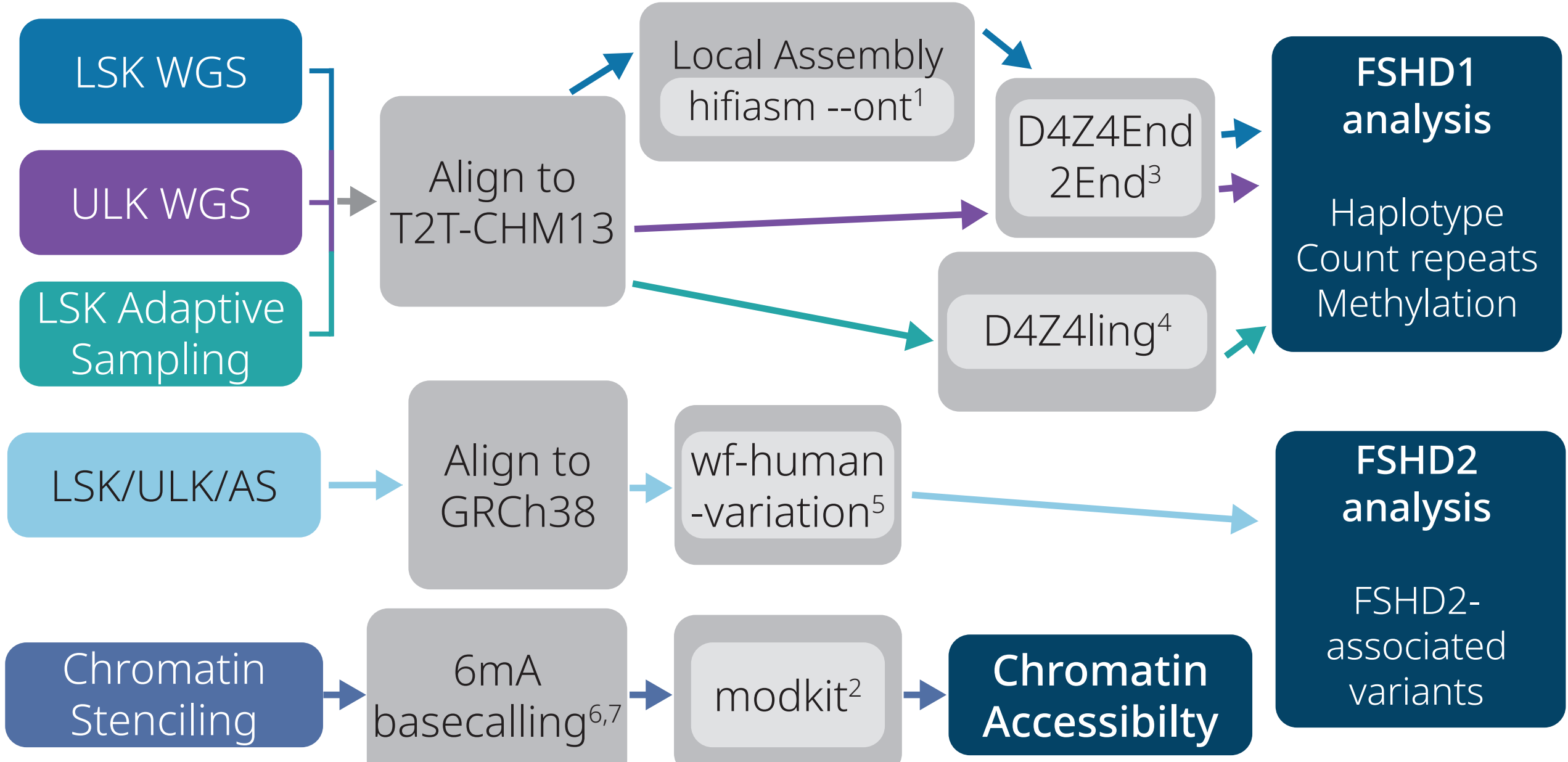


Fig. 2. Analysis workflow highlighting Oxford Nanopore Technologies sequencing strategies for characterizing the genetic and epigenetic features of FSHD. Open-source bioinformatic tools shown in light gray boxes. LSK = Ligation Sequencing Kit, ULK=Ultra-Long DNA Sequencing Kit, WGS=Whole Genome Sequencing, AS=Adaptive Sampling.

References

1. Cheng, H. et al. Efficient near-telomere-to-telomere assembly of nanopore simplex reads. *Nature* (2026).
2. Modkit. Oxford Nanopore Technologies (2026).
3. Xiao, L. C. et al. D4Z4End2End: complete genetic and epigenetic architecture of D4Z4 macrosatellites in FSHD, BAMS and reference cohorts.
4. Yew, D. et al. Targeted long-read sequencing enables comprehensive analysis of the genetic and epigenetic landscape of inherited myopathies.
5. EPI2ME Labs. wf-human-variation. Oxford Nanopore Technologies.
6. Stergachis, A. B., Debo, B. M., Haugen, E., Churchman, L. S. & Stamatoyannopoulos, J. A. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* 368, 1449–1454 (2020).
7. Chromatin accessibility know-how document. Oxford Nanopore Technologies (2026).

Accurate assemblies of D4Z4 arrays resolve repeat number and haplotype

Using both raw ULK array-spanning reads and LSK-assembled contigs, we successfully identified and counted the repeat units across the four target haplotypes in all cell lines (Fig. 3). For all samples, we obtained at least one read in the ULK dataset that spanned the entire repeat array, with the longest individual spanning read being 565 kb. To evaluate assembly accuracy, we compared D4Z4 repeat unit counts from the ULK reads with those from the LSK-assembled contigs (Fig. 3A), and the counts were highly concordant ($R^2 = 0.998$; Fig. 3B).

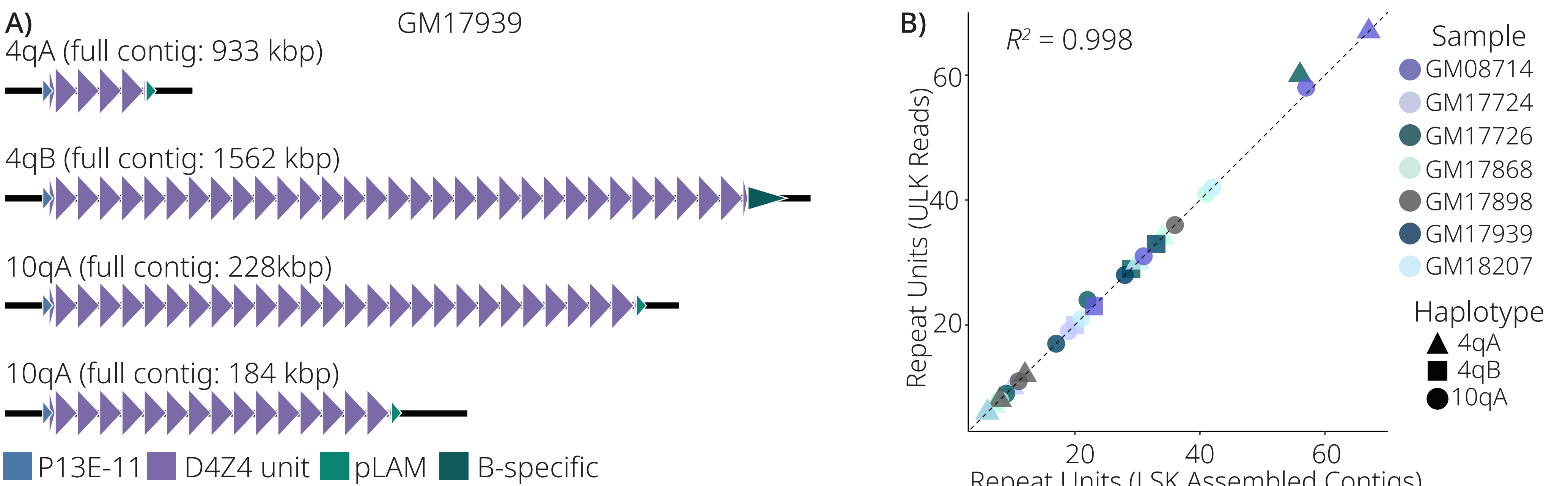


Fig. 3. Assembly of LSK reads accurately represents the number of D4Z4 repeats. A) Schematic of assembled contigs for the four haplotypes with D4Z4 repeats. Assemblies created with hifiasm (ONT) and LSK reads (>30kb read-length N50). Each schematic includes key features of the repeat and the total length of the contig. B) Comparison of number of D4Z4 repeat units in the LSK-assembled contigs and the ULK reads for every sample in the study.

Table 1. Samples, sequencing coverage, and read-length N50.

Sample	Type	Adaptive Sampling*		LSK WGS†		ULK WGS‡	
		On-Target Read N50	On-Target Coverage	Read N50	Coverage	Read N50	Coverage
GM08714	ICF	17 kb	102.3x	31 kb	35x	83 kb	34x
GM17724	FSHD1	22 kb	124.7x	42 kb	37x	99 kb	20x
GM17726	LGMD2A	26 kb	65.4x	33 kb	52x	86 kb	19x
GM17868	FSHD1	23 kb	115.0x	38 kb	50x	122 kb	16x
GM17898	FSHD1	25 kb	102.6x	33 kb	34x	108 kb	20x
GM17939	FSHD1	25 kb	107.5x	33 kb	41x	84 kb	34x
GM18207	FSHD1	20 kb	82.1x	40 kb	39x	100 kb	17x

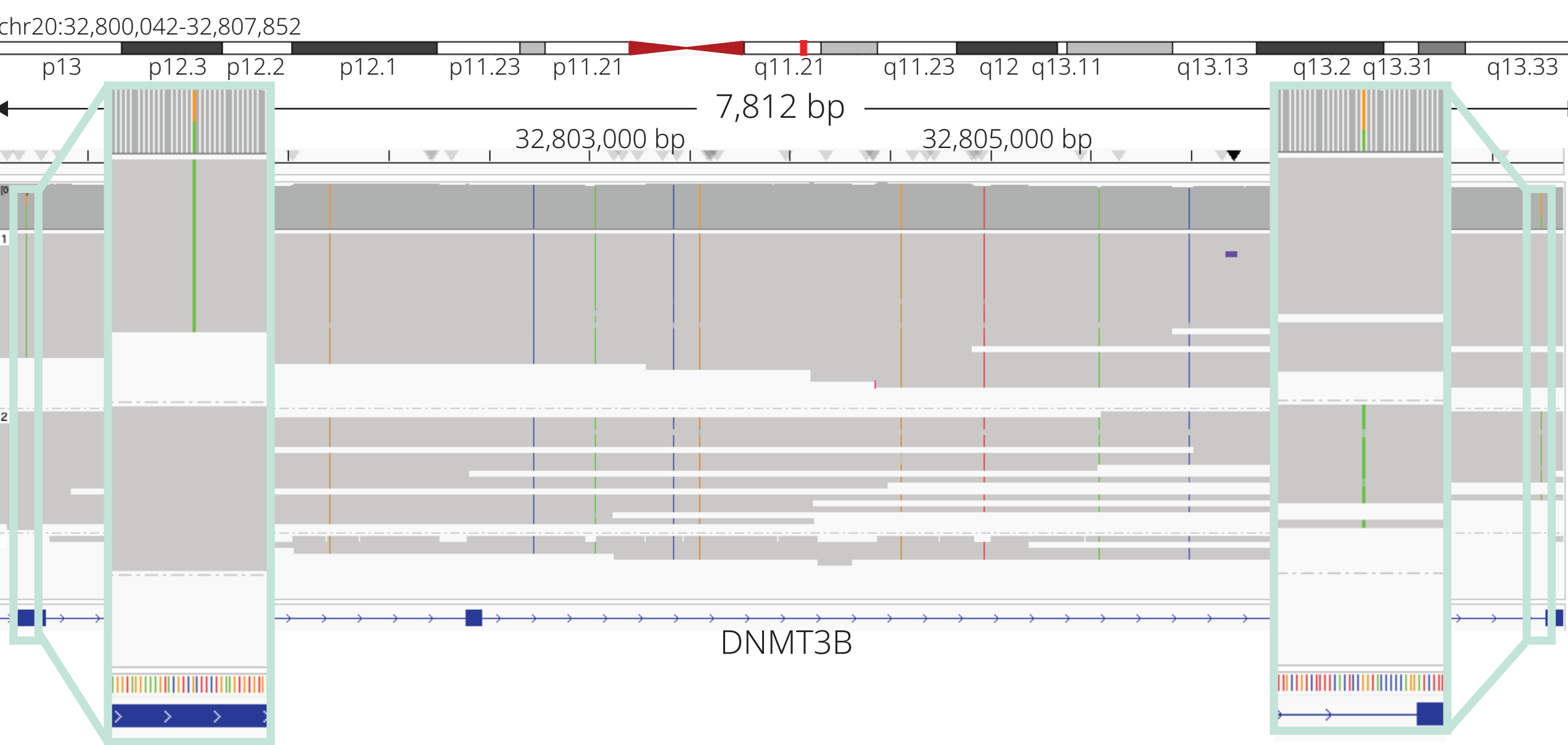
*Adaptive Sampling sequencing metrics are per sample, samples sequenced two per flowcell.

†LSK sequencing metrics are per flowcell, one flowcell per sample.

‡ULK sequencing metrics are per flowcell, two flowcells per sample.

LSK = Ligation Sequencing Kit, ULK = Ultra-Long DNA Sequencing Kit.

Phased analysis resolves compound heterozygous variants in *DNMT3B*



Variants in *DNMT3B* impair the establishment of DNA methylation patterns. Our analyses identified two *DNMT3B* variants in GM08714 and assigned them to separate haplotypes through parental-free phasing, supporting compound heterozygosity in this FSHD2-associated gene (Fig. 4).

Fig. 4. Compound heterozygous *DNMT3B* variants are associated with widespread hypomethylation in GM08714. Analysis using wf-human-variation identified two *DNMT3B* variants. The read pileups confirm haplotype-specific support for each variant. This sample's D4Z4 hypomethylation is shown in Fig. 5Bt.

Adaptive Sampling identifies hypomethylated, contracted 4qA alleles

Adaptive Sampling generated 65–115× on-target coverage across seven samples, with two samples sequenced per flowcell and read N50 values ranging from 17.1 to 26.0 kb (Table 1). For analysis, reads that aligned to the *DUX4* regions of the reference genome were further assigned to haplotypes, classified as spanning or non-spanning, and evaluated for native DNA methylation (Fig. 2). The number of D4Z4 repeats was then calculated for each read and assessed for haplotype and methylation status⁴. Unlike ULK reads and LSK-derived assemblies, Adaptive Sampling reads do not span longer D4Z4 arrays. However, individual Adaptive Sampling reads could span the short, contracted 4qA alleles. This integrated read-level analysis correctly identified the presence or absence of hypomethylated, contracted 4qA molecules across all seven samples directly from Adaptive Sampling reads (Fig. 5B).

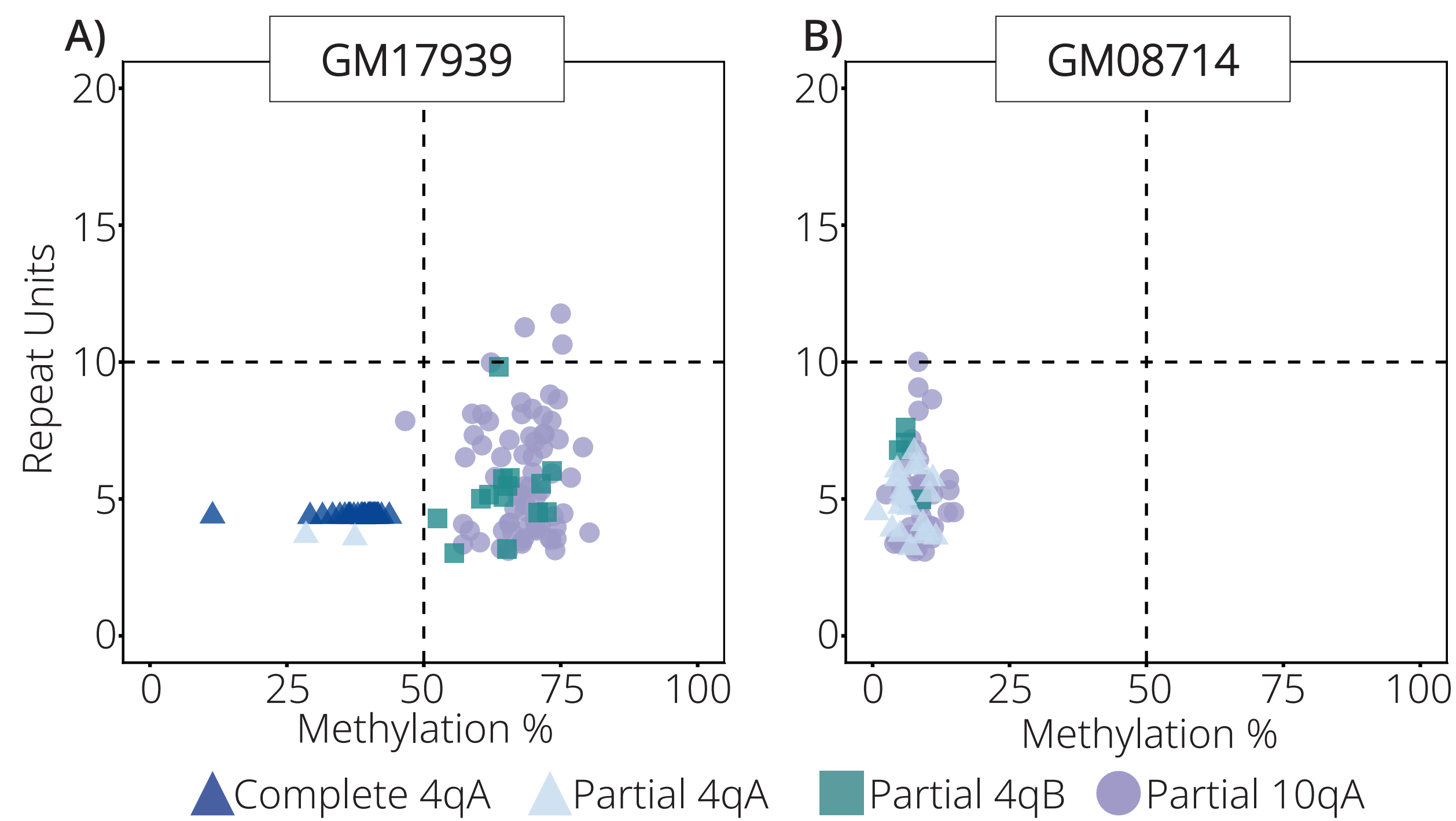


Fig. 5. Plots of individual Adaptive Sampling reads showing D4Z4 copy number and distal-most D4Z4 unit methylation by haplotype. Colors and shapes indicate allele haplotype (4qA, 4qB, 10qA) and read type. Complete reads span the entire D4Z4 array, whereas partial reads contain a haplotype-defining terminal sequence and only the adjacent D4Z4 repeats. Horizontal dashed lines represent the 10-copy threshold for FSHD1 pathogenic contraction, and vertical dashed lines mark 50% methylation. Representative examples are shown for A) FSHD1 with a hypomethylated 5-copy D4Z4 array on a 4qA haplotype and normal (>50%) distal methylation of the non-contracted 4qB and B) FSHD2-like sample with no D4Z4 contraction but widespread hypomethylation across all alleles.

Epigenetic profiles distinguish pathogenic from nonpathogenic alleles

Contracted D4Z4 arrays in FSHD samples are characterized by hypomethylation and increased chromatin accessibility. In contrast, nonpathogenic arrays remain hypermethylated and relatively inaccessible. Integrated analysis of native DNA methylation and 6mA-based chromatin stenciling captured these contrasting epigenetic states at the read level (Fig. 6).

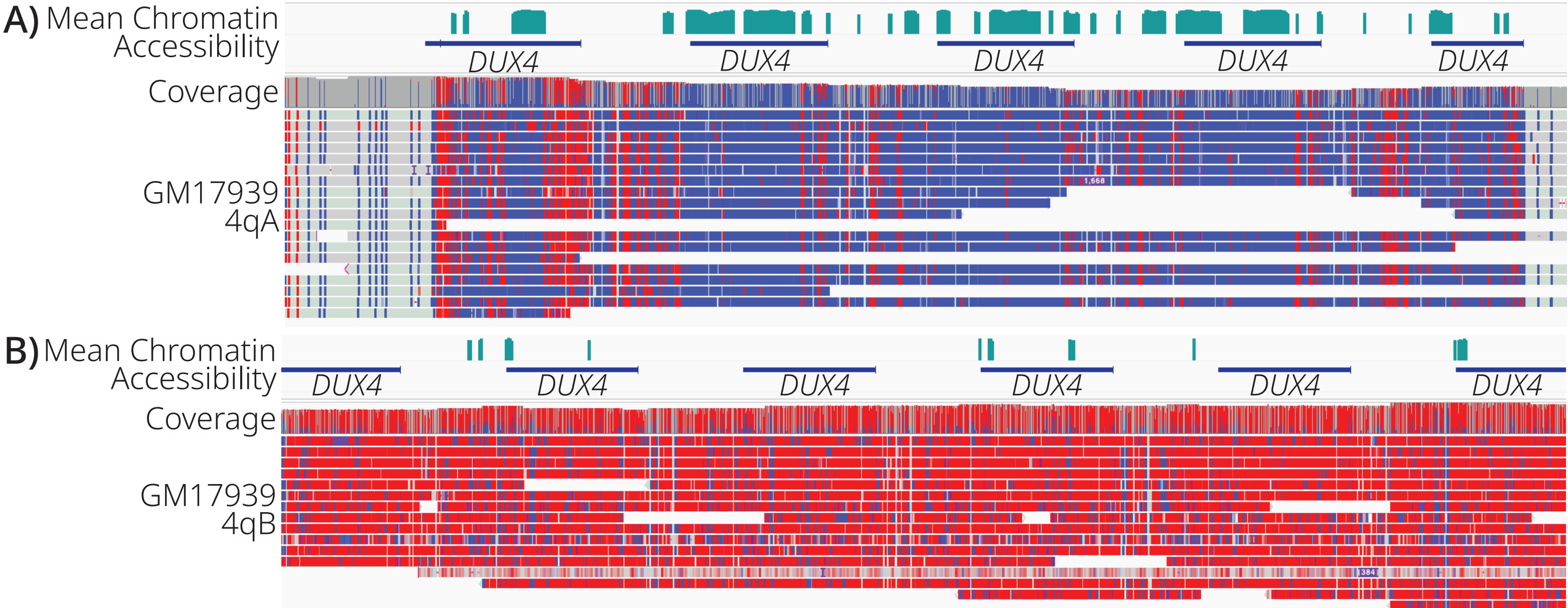


Fig. 6. Integrated epigenetic profiling distinguishes D4Z4 haplotypes. Representative LSK read pileups for A) a pathogenic 4qA allele containing five D4Z4 repeats and B) a six-repeat region of a non-pathogenic 4qB allele containing 36 repeats total. Reads are colored by methylation state, with red indicating hypermethylation and blue indicating hypomethylation; dark green bars indicate mean chromatin accessibility.

Conclusion

Oxford Nanopore sequencing enables single-platform FSHD testing by resolving and counting D4Z4 repeat units across all four *DUX4* alleles. The ULK (array-spanning reads), LSK (*de novo* assembly with hifiasm), and AS (high-coverage, individual reads evaluated for native DNA methylation) workflows, allowed for accurate downstream haplotyping and variant analysis. Direct detection of native DNA methylation revealed the expected hypomethylation of the pathogenic array, and chromatin accessibility analysis using chromatin stenciling identified open chromatin at the pathogenic contracted allele. Taken together, our findings demonstrate a single-platform, comprehensive approach for FSHD analysis using Oxford Nanopore multiomic sequencing.