Human Genomics

RESEARCH

Open Access

DUCKS4: a comprehensive workflow for Nanopore sequencing analysis of facioscapulohumeral muscular dystrophy (FSHD)

Tamara Löwenstern[1*], Silvia Madritsch[1], David Horner[1], Nadja Brait[1], Naz Güleray Lafci[1], Anna Schachner[1], Maria Gerykova Bujalkova[1], Tadeusz Kałużewski[2], Pawel Szyld[3], Markus Hengstschläger[1], Paul Dremsek[1] and Franco Laccone[1]

Abstract

Motivation Facioscapulohumeral Muscular Dystrophy (FSHD) is an autosomal dominant form of muscular dystrophy caused by genetic or epigenetic changes within the D4Z4-repeat at the DUX4-gene, on chromosome 4q. Genetic analysis is challenging due to a nearly identical region on chromosome 10, multiple haplotypes, long and short repeat subtypes, and complex rearrangements such as translocations and duplications. So far, no single method detects all known causes of FSHD.

Results We have developed an integrated approach combining an optimized wet-lab protocol with an automated bioinformatics workflow, called DUCKS4. It enables read-level resolution of the D4Z4 array for FSHD1 repeat sizing, variant detection for FSHD2, and detection of methylation patterns. Using NCBI BLAST, it assigns reads to chromosomes and haplotypes, supporting robust filtering and analysis. Our approach also includes a novel adaptive sampling workflow for human high molecular-weight DNA. With long-read Nanopore sequencing technology, our tool enables precise determination of D4Z4 array size, individual haplotype assignment, methylation profiling, and complex allele analysis. It also allows for the detection of mosaicism and structural variation like interchromosomal translocations, providing a comprehensive, single-method solution for FSHD analysis.

Availability and implementation DUCKS4 is implemented within a Docker environment. Source code and supplementary materials are accessible at https://github.com/tamara-nano/ducks4, https://github.com/tamara-nano/ducks4_wovar (light version without variant calling) and archived at https://figshare.com/projects/DUCKS4-FSHD/260306

Keywords Long-read sequencing, Nanopore sequencing, FSHD, DUX4, Macrorepeat, Adaptive-sampling

*Correspondence:
Tamara Löwenstern
tamara.loewenstern@meduniwien.ac.at
[1]Institute of Medical Genetics, Medical University of Vienna, Vienna, Austria
[2]Department of Genetics, Polish Mother's Memorial Hospital Research Institute, 93-338 Lodz, Poland
[3]Specjalistyczna Poradnia Genetyczna, Kalisz, Poland

Introduction

Facioscapulohumeral muscular dystrophy (FSHD) is the third most common autosomal dominant muscular dystrophy, affecting ~ 5–10 in 100,000 individuals [1, 2]. It is characterized by progressive, often asymmetric skeletal muscle weakness, typically beginning in the facial and shoulder muscles, and may lead to loss of ambulation in up to 20% of cases [3–6].

Two genetically distinct forms are known: FSHD1 (OMIM: 158900) and FSHD2 (OMIM: 158901). Both involve aberrant expression of the DUX4 gene, leading to muscle cell apoptosis, inflammation, and impaired regeneration [7–9]. DUX4 is embedded in the macrosatellite D4Z4 repeat array at chromosome 4q35. The D4Z4 repeat has a motif size of 3.3 kb, and within the population, 11–150 repeat-units are considered normal. FSHD1 is caused by a contraction of the D4Z4 array to 1–10 units on a permissive 4qA haplotype. In contrast, FSHD2 results from epigenetic hypomethylation of a normal-sized D4Z4 array in combination with a permissive 4qA haplotype, typically due to disease-causing variants in chromatin regulators like *SMCHD1*, *DNMT3B*, or *LRIF1* [10–12]. Importantly, only 4qA haplotypes are pathogenic due to the presence of a polyadenylation signal (PAS) in the pLAM region, 4qB and chr10 homologs, however, lack this feature and are therefore regarded as non-permissive alleles [13, 14].

A major technical challenge in FSHD diagnostics lies in the high sequence identity (~ 99%) between the 4q and 10q D4Z4 regions, impairing haplotype resolution and accurate repeat sizing. Furthermore, over 30 sub-haplotypes have been described across chromosomes 4 and 10, including subtypes and hybrid alleles as well as structurally complex rearrangements such as de novo translocations or in-cis D4Z4 repeat duplications [11, 15–19]. Moreover, mosaicism should be considered in FSHD diagnosis, which is described in cases where the disease occurs despite a negative family history, similarly to de novo pathogenic variants or low penetrance [20, 21].

Current diagnostic tests for FSHD, such as Southern blotting, allow repeat sizing and haplotype assignment but are labor-intensive and have limited resolution [16]. To some extent, optical genome mapping (OGM) offers an alternative, although, importantly, chr4-to-10 translocations cannot be detected [22, 23]. Short-read sequencing fails due to the repetitive structure of the disease-associated region. FSHD2 analysis additionally requires variant calling and methylation profiling. Clinical diagnosis alone is not sufficient, as misdiagnoses of FSHD can occur due to phenotypic overlap with other muscular dystrophies such as Becker (BMD), Duchenne (DMD), or Limb-Girdle Muscular Dystrophies (LGMD), especially in early or atypical cases [24, 25], and blended phenotypes have also been reported [26].

The emergence of Nanopore long-read sequencing technology offers a unique opportunity to unify all relevant analyses in a single assay. This includes D4Z4 repeat sizing, haplotype determination, methylation profiling, and the detection of mosaicism and structural rearrangements. Nevertheless, the D4Z4 locus remains one of the most challenging regions of the human genome. Standard long-read analysis pipelines mostly fail to completely resolve this locus, particularly in structurally complex samples. In recent years, several studies have demonstrated that Nanopore sequencing enables the simultaneous assessment of D4Z4 repeat length, haplotype structure, and CpG methylation using whole-genome sequencing or Cas9-targeted enrichment [27–30]. More recent work extending these strategies to resolve complete end-to-end D4Z4 arrays across FSHD and reference cohorts [31], preprint). Together, these findings underscore both the diagnostic potential of long-read sequencing and the need for specialized bioinformatic workflows to fully exploit its capabilities in FSHD analysis. Following this body of work, we additionally describe an adaptive sampling-based enrichment workflow of high molecular weight genomic (HMW) DNA. This approach allows cost-efficient, flexible and indication-specific enrichment of large genomic regions while preserving long fragment lengths and minimizing wet-lab work.

Here, we present a comprehensive workflow incorporated in a convenient tool, DUCKS4, that allows for fast and simultaneous analysis of FSHD1 and FSHD2, as well as the detection of more complex alleles. This approach offers a comprehensive and scalable alternative to the fragmented and labour-intensive methods traditionally used in FSHD diagnosis. Inspired by previously published BLAST-based strategies for Nanopore-derived data for FSHD-analysis [32, 33], our tool represents a substantially advanced workflow for the complexity of FSHD analysis. Compared to existing methods, DUCKS4 enables the resolution of complex cases that previously remained challenging, including chr4-to-10 translocations, mosaicism, DBED alleles (proximal deletion in the D4Z4-array), and hybrid alleles.

Material and methods

Sample collection

The cell cultures for the human reference genomes HG001 (GM12878), HG002 (GM24385), and HG003 (GM24149) were ordered from coriell.org and cultivated according to the website's recommendations. The benchmark sequences of maternal and paternal alleles of HG002 were downloaded from the HG002 "Q100" project (https://github.com/marbl/HG002?tab=readme-ov-file) of the Telomere-to-Telomere Consortium [34]. The other reference genome sequences were acquired via ncbi.nlm.nih.gov (https://github.com/genome-in-a-bottle

/giab_data_indexes). Patient DNA samples were provided from patient cohorts of the Institute of Medical Genetics, Vienna (Mos1, Neg1, Neg2, Neg3) as well as from Poland (samples Pos1-3, Trans1-4) with patients' informed consent for using the DNA samples for scientific purposes.

Sample preparation and sequencing

High molecular weight (HMW) DNA was extracted from frozen EDTA blood or cell cultures according to the "ultra-long-dna-sequencing-kit-sqk-ulk114-ULK_9177_v114_revK_27Nov2022-promethion" protocol from Oxford Nanopore Technologies (UK), with the following modifications. Blood was washed a minimum of 3 times with RBC-Lysis Solution (Qiagen, Hilden, DE) until the cell pellet appeared near-white and the liquid was clear. Then an additional washing step with 10 min incubation with RBC-Lysis Solution at RT was performed. After discarding the supernatant, 40 µL 1 × PBS was added to the pellets, not resuspended, and frozen at – 20 °C. After thawing, DNA was extracted with the Monarch HMW-DNA Extraction Kit for Tissue (NEB, Ipswich, US). Proteinase K digestion was extended to up to 1 h at 56 °C and 650 rpm on a heater-shaker to ensure complete pellet dissolution. DNA was purified and eluted per protocol, incubated overnight at room temperature and stored for a few days at + 4 °C, then quantified via Qubit. A total of 40–60 µg of DNA was used for library preparation with the SQK-ULK114 Library prep Kit (Oxford Nanopore Technologies, UK). Following library preparation, storing the library at + 4 °C for at least 1–2 weeks prior to sequencing is essential to achieve superior sequencing results. One third of the resulting library was used for subsequent sequencing via a PromethION P2-Solo connected to a GridION machine (Oxford Nanopore Technologies, UK).

DNA obtained using routine clinical laboratory methods (referred to as standard DNA) was extracted from cell cultures using the automated EZ1 Advanced XL robot (Qiagen) with the EZ1 DNA Tissue Kit. Standard DNA from human whole blood (EDTA) was extracted using the Chemagic Revvity robot (Revvity). Left-side size selection for eliminating fragments < 30 kb for ~ 6 µg DNA was performed using a Polyvinylpyrrolidone solution (final concentration: 2% PVP 360 K, 600 mM KCl, 10 mM Tris pH8; [35]). Equal amounts of sample and SRE-Buffer were mixed, incubated for 30 min at 50 °C and centrifuged for 30 min at 10.000 g at RT, and supernatant discarded. The sample was washed twice with 80% EtOH and eluted with ddH2O and incubated for 20 min at RT and 10 min at 37 °C without mixing to carefully elute DNA from the pellet. Afterwards the sample was stored overnight at + 4 °C prior quantification. The Nanopore library was prepared according to the "Ligation sequencing 30 kb human gDNA on P2-Solo for SQK-LSK114, Version GDH_9174_v114, revD, 10.11.2022" protocol for the SQK-LSK114 Kit. The protocol was adapted as follows: the sample was eluted at the end of the end-prep with half of the amount of ddH2O and in the subsequent adapter-ligation step half of the amount of the reagents were used and a library sufficient for one loading on the flow-cell was created. 80% EtOH and LFB were used for the washing steps. Also, where possible the samples were handled with wide-bore-tips to minimize possible fragmentation.

Adaptive sampling targeted chromosomes 4 and 10 (DUX4 regions and D4Z4 arrays) and FSHD2-related genes (supplementary table S4), covering ~ 0.5% of the T2T-CHM13v2.0 genome supplemented with the HG002 4qB allele (Q100 project; see 2.1). Sequence duration was 17 h, achieving 15–40 × coverage in target regions. Basecalling was performed using Dorado [0.7.3 + 6e6c45c] in SUP mode with methylation detection (basecalling model: dna_r10.4.1_e8.2_400bps_sup\@v5.0.0_5mCG_5hmCG\@v1/).

DUCKS4 methodology

The DUCKS4 scripts are written in Python (v3.7.11) and R (v4.3.2; [36]), and are available at https://github.com/tamara-nano/ducks4. The automated workflow for the read-based FSHD statistical analysis of DUCKS4 is outlined below (Fig. 1).

Input file handling and database screening

The DUCKS4 workflow accepts unmapped and mapped BAM files, FASTQ files, or FASTA files. First, the file format is checked and converted to FASTA format if necessary. Input files are searched against a custom-created database for FSHD analysis using BLASTn (v2.14.0; [37]). A visual representation of the target regions and motifs identified in this process can be seen in Fig. 2. The database includes the data for the most common haplotypes 4q35 A (4A161) and B (4B163) and 10q26 (10A164b) [38], D4F104S1 (locus for the p13E-11 probe from Southern Blotting) for the proximal area of the repeat, pLAM for the distal end of the repeat, as well as D4Z4-RU for 4q35 and 10q26. Moreover, two extra regions are defined to test for completeness of the repeat-array: *CLUHP4* (~ 2,3 kb upstream of D4F104S1) and the distal part of *DUX4* (DUX4-end). Sequences for creating the blast database were downloaded from the NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore).

Target regions and off-target markers

BLASTN results are processed via a custom R-based pipeline that determines the exact repeat composition per read by evaluating hit identity, alignment length, and query position. Prior to repeat analysis, off-target hits for 4qA-pLAM/4qB-pLAM and D4F104S1 (p13E-11)

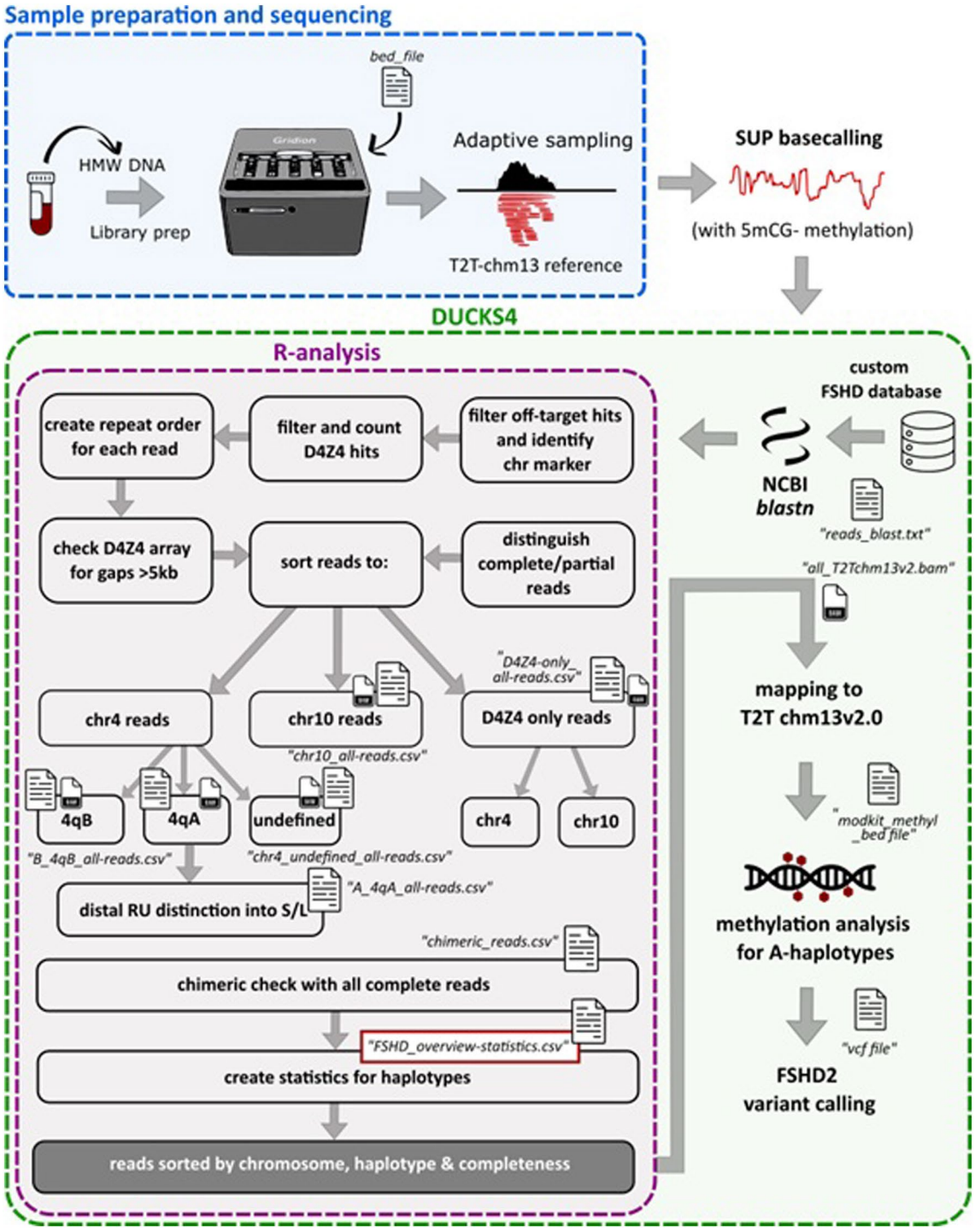

Fig. 1 Overview of sample preparation and DUCKS4 methodology. File outputs are indicated by file icons. An additional BAM output for visualization is marked with a BAM file icon

were identified and filtered. A known off-target hit for 4qB-pLAM (~88% identity) occurs approx. ~58.2 kb upstream of the D4Z4 array on chromosome 4 and is annotated as "pLAM_4qB_lowpid" (Fig. 2, panel 1). This marker also serves as a positive control for full-array coverage and chromosome 4 specificity. Additionally, up to two off-target D4Z4 blast-hits upstream of the D4Z4 array (~42 kb) resulting from an inverted D4Z4 repeat unit on chr4 are annotated as "c4_ctrl" and used to distinguish chr4 from chr10 reads, marking the end of 4q/10q homology [14]. Reads lacking a "c4_ctrl" hit are tested for length and orientation to assess whether spanning this

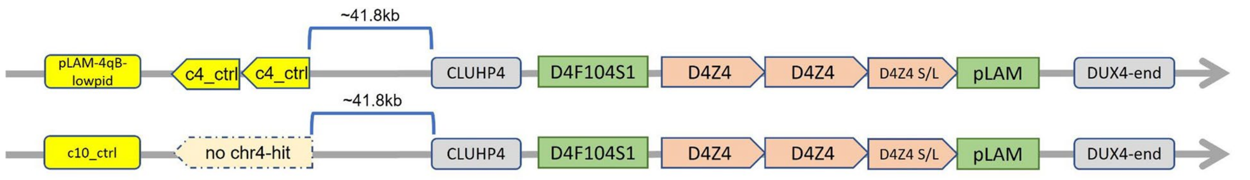

Fig. 2 Scheme of the order of the blast-hits for the D4Z4 repeat on a forward read. The first panel shows chromosome 4, while the second panel shows chromosome 10. The DUCKS4 output represents the order of the output in orientation of the read; therefore, for reverse reads, the repeat-order will be shown in reverse order

region would be possible. If so, and no hit is found, a "no_chr4_hit" tag is added, indicating a possible chr10 origin. A dedicated "c10_ctrl" hit is also included upstream for direct chr10 read identification (Fig. 2, panel 2). After filtering, only D4Z4 hits > 1500 bp are considered. Chr4 and chr10 D4Z4 hits appear in parallel, but only the hit with higher identity is retained, allowing a *qstart* difference of ≤ 100 bp to define parallel hits. Reads are then assigned to chr4 or chr10 based on the dominant hit identity.

Haplotyping, subtyping, and chimaera detection

Chr4 reads are assigned to 4qA or 4qB haplotypes based on the presence of the proximal D4F104S1 and/or the distal pLAM identity. Reads with chromosome 4 markers but lacking definitive haplotype features are labeled as "undefined". Reads containing only D4Z4 units without flanking regions are evaluated based on repeat composition to infer chromosomal origin (chr4 vs. chr10). 4qA-reads are further subclassified based on the length of the most distal D4Z4 unit, distinguishing between the short ("S") and long ("L") sub-haplotypes. Typically, 4qA_S alleles end in a truncated D4Z4 unit (~ 1.8 kb), while 4qA_L alleles retain a full-length unit (~ 3.3 kb). Following repeat annotation, reads are categorized as complete if both proximal (D4F104S1) and distal (pLAM) ends flanking the D4Z4 array are present, or partial if only one side was captured. Additional control targets outside the array, including CLUHP4 (proximal) and DUX4_end (distal), are used to support classification. Finally, complete reads are screened for chimeric features, such as mixed chr4/chr10 repeat patterns. Chimeric reads, e.g., those beginning with chr10-like sequences and ending in 4qA-specific pLAM, may indicate rare structural rearrangements or non-canonical haplotypes involving chromosomes 4 and 10.

DUCKS4 output: aligned haplotypes and output statistics

All sorted and original input reads are aligned to the T2T-CHM13v2.0 reference genome using minimap2 (v2.28-r1209; [39]) and samtools (v1.16.1, [40]) to generate sorted and indexed BAM files. Mean alignment depth is computed using samtools coverage. Detailed repeat composition per haplotype is provided in.csv files, listing read ID, RU (repeat unit) count, distal RU type (S/L), status (complete/partial), and exact repeat order (Table 1). For reverse-oriented reads, repeat composition is reported in reverse. Reads classified as 4qA or 4qB but containing chromosome 10 markers trigger a warning for non-canonical haplotypes. An overview statistics

Table 1 Example of sample HG002 detailed results for 4qA-reads from the DUCKS4 tool

Read.id	RU count	S/L	Status	Repeat-sequence
HG2r1	43	S	complete	DUX4_end–4qA_pLAM–c4S–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–4qA_D4F104S1–CLUHP4-201_exon1
HG2r2	38	S	partial	c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4S–4qA_pLAM–DUX4_end
HG2r3	23		partial	c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–4qA_D4F104S1–CLUHP4-201_exon1
HG2r4	21		partial	c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–4qA_D4F104S1–CLUHP4-201_exon1–c4_ctrl–c4_ctrl–4qB_pLAM_low-pid
HG2r5	19	S	partial	c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4S–4qA_pLAM–DUX4_end
HG2r6	16	S	partial	DUX4_end–4qA_pLAM–c4S–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4
HG2r7	10		partial	4qB_pLAM_low-pid–c4_ctrl–c4_ctrl–CLUHP4-201_exon1–4qA_D4F104S1–c4–c4–c4–c4–c4–c4–c4–c4–c4–c4
HG2r8	9		partial	CLUHP4-201_exon1–4qA_D4F104S1–c4–c4–c4–c4–c4–c4–c4–c4–c4
HG2r9	8	S	partial	c4–c4–c4–c4–c4–c4–c4–c4S–4qA_pLAM–DUX4_end
HG2r10	2		partial	c4–c4–4qA_D4F104S1–CLUHP4-201_exon1
HG2r11	2	S	partial	DUX4_end–4qA_pLAM–c4S–c4

reports read counts per chromosome and haplotype, RU distribution for complete, partial, and D4Z4-only reads, and a summary of chimeric reads with mixed chr4/chr10 features if they occur. A detailed explanation for all additional output files can be found at (https://github.com/tamara-nano/ducks4).

An excerpt of the table, including the information for the read-id, RU count, S/L differentiation, read status and resolved repeat-sequence, is shown.

Methylation analysis

An optional methylation analysis is implemented for 4qA and, if present, chimeric reads. Methylation levels are assessed for the two most distal D4Z4 RU and the *DUX4* gene body [33] using modkit (v0.4.4, Oxford Nanopore Technologies), which quantifies 5mC at CpG sites, creating a bedmethyl file and methylation statistics. A sub workflow within the DUCKS4 package "DUCKS4_ID2bam2meth.py" enables selection, alignment, and methylation calling for user-defined read subsets via read ID. It further enables the selection of an individual read as a custom reference to which a subset of reads can be aligned. Using the corresponding blast-based annotation of this reference read, methylation can be quantified for each annotated segment of the D4Z4 array, enabling reconstruction of methylation gradients across the array (see Supplementary Figure S2). The workflow outputs the custom reference sequence as indexed fasta file, its annotated D4Z4 structure as bed file, per-segment methylation values (bedGraph and summary statistics), and the aligned reads (BAM). While complete reads are recommended as references, partial reads can also be used, in which case methylation analysis is restricted to the covered portion of the array.

FSHD 2 analysis

For optional variant calling the input file is aligned to the GRCh38 reference, then variant calling is performed with Clair3 (v1.0.2; [41]), phasing is done with Whatshap (v2.4 [42]), and structural variant calling is carried out with Sniffles2 (v2.0.7 [43]). Furthermore, variant annotation is carried out by SnpEff (v5.2f [44]) and SnpSift [45] against the human genetic variant databases ClinVar (v20250729 [46]).

Optical genomic mapping

HMW DNA was isolated from frozen EDTA-blood samples with the "Bionano Prep SP G2 Frozen Human Blood DNA Isolation" protocol, and fluorescence labelled with DLE-1 enzyme via the "Bionano DLS-G2" protocol (Bionano Genomics, USA). The labelled HMW-DNA was then processed via a Bionano Saphyr instrument on a Saphyr G3.3 chip and analysed with the Bionano Genomics EnFocus™ FSHD Analysis pipeline for FSHD diagnostics, distinguishing 4qA and 4qB haplotypes and D4Z4-array repeat counts. The analysis was performed with the Bionano Access software (Tools Version 1.8.2, Solve Version Solve3.8.2_main-04-30-24-rel) using the GRCh38 reference database. If needed, a de novo assembly to T2T-chm13v2.0 was performed for genome-wide analysis of structural variants.

Statistics

Determination of sequencing statistics like read length distributions and the median read quality was performed with seqkit (v2.10.0; [47]) and NanoStat (v1.6.0 [48]). Per-read error rates were estimated from primary alignments using CIGAR and MD tags in uniquely mappable regions proximal to the D4Z4 array, calculating the fraction of mismatches, insertions, and deletions relative to aligned bases. This aims to avoid inflation of apparent error rates caused by alignment ambiguity within the highly repetitive D4Z4 macrosatellite itself. We further summarized the proportion of reads ≥ 37 kb, the length of 11 RUs, per sample by DNA input (HMW vs. standard) using boxplots showing the median and interquartile range (IQR; 25th-75th percentiles), with the individual samples indicated as points. For that we used R-studio (version 2025.5.0.496; [49]) with ggplot2 [50].

Results

Comparison of HMW with standard DNA

In total, six whole genome sequencing (WGS) runs and eighteen adaptive sampling (AS) runs were conducted with HMW-DNA, whereas from 7 AS runs only sequencing and alignment parameter are presented in this study (see Supplementary Table S3). Mean WGS coverage was ~ 19,2x (SD: 9,6), and for AS ~ 24,4x (SD: 14,3). Excluding the low-coverage positive controls, AS averaged 30.7x (SD: 14.6). In six cases (Pos1, Trans1, Dup1, Hmw1, Hmw4, Hmw6_L2) one AS run for 17 h on a PromethION flow-cell reached up to 40–50 × in mean depth. For the HMW-DNA an average N50 of ~ 72 kb (SD: 20.5) was reached. Although AS yielded higher coverage, no substantial differences in analytical outcomes were observed between WGS and adaptive datasets for HMW-DNA. We also incorporated the sequencing and alignment parameters for six AS and one WGS run with Standard-DNA where we yielded mean depths averaged 38.4x (SD: 16.8) with mean N50 of ~ 18.3 kb (SD: 4.9; see Supplementary Table S3). Across all samples, HMW-DNA libraries showed comparable median per-read error rates to Standard-DNA libraries, with distributions consistent with expected SUP basecalling performance and indicating overall high read accuracy (Supplementary Table S3).

Storing HMW libraries at + 4 °C prior to sequencing was associated with improved sequencing performance,

whereas libraries sequenced immediately after preparation showed a higher risk of reduced yield and mean depth (Supplementary Table S3). This observed effect was further illustrated by a paired comparison of two aliquots from the same library, where sequencing after 18 days of storage resulted in a substantially higher mean depth (48×) compared to immediate sequencing (5.1x). Notably, the first aliquot was sequenced on a new flow cell with 7459 available pores, while the second aliquot was sequenced on a previously used and washed flow cell with 6613 pores remaining, indicating that the observed improvement was not attributable to pore availability in this case.

To assess the yield of fragments capable of resolving or excluding FSHD1-relevant repeat lengths, we quantified the proportion of reads ≥ 37 kb, corresponding to the minimum size required to span an 11-RU D4Z4 allele (11RU = ~ 36.3 kb). HMW DNA libraries showed a markedly higher fraction of long reads than standard DNA libraries (median 0.358 vs. 0.037; see Fig. 3), representing an approximately tenfold enrichment. Despite higher total read counts in Standard DNA samples, the yield of ≥ 37 kb fragments remained consistently low, indicating that increased sequencing depth alone does not compensate for reduced long-fragment recovery (Supplementary Table S10).

Validation of workflow with reference genomes and positive controls

The accuracy of the DUCKS4 workflow was validated using HMW-DNA from well-characterized reference samples (HG001, HG002, HG003), an additional negative control, and three FSHD1-positive controls, all independently verified by OGM via Bionano Genomics EnFocus™. An overview of mean coverage, read counts, and repeat unit (RU) composition per haplotype is summarized in Supplementary Table S1. Final results, including OGM comparisons, are provided in Supplementary Table S2, and further sequencing metrics in Supplementary Table S3.

With a WGS run for HG002, two chr4 alleles were identified: 4qA with 43 RU and 4qB with 27 RU, along with two chr10 alleles (18 RU and 33 RU). These findings were confirmed by aligning filtered reads to the phased HG002 reference genome from the Q100 project [34], and by processing the reference genome through DUCKS4. OGM results closely matched, with only one RU difference (42 RU for 4qA, 26 RU for 4qB, Table 2). An additional AS run for HG002 resolved chr10 and 4qB completely; 4qA was partially resolved (> 41 RU). For HG003 (father of HG002), a 4qA_S allele with 28 RU and a 4qB allele with 27 RU were detected. HG001 (not related to HG002) carried a 4qA_S allele with 48

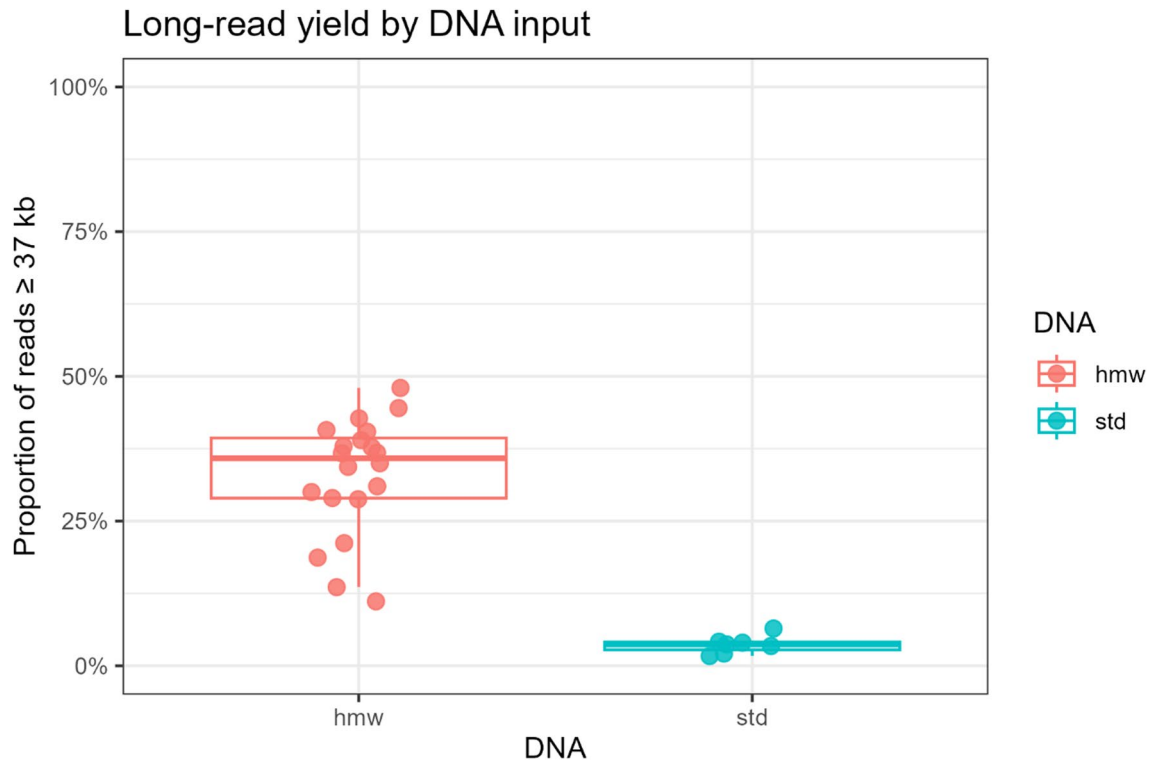

Fig. 3 Long read-yield by DNA input. Proportion of reads ≥ 37 kb was calculated from primary alignments overlapping chr4:192,406,429-193,419,672 (on-target read set, located proximal of the D4Z4 array). Proportion was defined as: number of on-target primary reads with read length ≥ 37 kb)/total number of on-target primary reads. Each point represents one sample, the samples are grouped by DNA input (HMW vs. Standard (std))

Table 2 Overview of Nanopore DUCKS4 and OGM analyses

OGM and Nanopore results of the haplotypes with RU-count on chromosome 4										
Samples			Nanopore - Results					OGM-Results		
				Chr 4				Chr 4		
Sample	Type	method	mean depth (x)	HP1	D4Z4-RU	HP2	D4Z4-RU	HP 1	D4Z4-RU (+/- 1)	HP 2	D4Z4-RU (+/- 1)
HG001	reference	WGS	31,4	4qA_S	48	4qB	20	4qA	48	4qB	19
HG001	reference	AS	18,4	4qA_S	48	4qB	20				
HG001	reference	AS	31,1	4qA_S	48	4qB	20				
HG002	reference	WGS	19,4	4qA_S	43	4qB	27	4qA	42	4qB	26
HG002	reference	AS	16,9	4qA_S	>41	4qB	27				
HG003	reference	WGS	20,4	4qA_S	28	4qB	27	4qA	27	4qB	28
Pos1	pos. control	AS	50,7	4qA_S	5	4qB	15	4qA	4	4qB	14
Pos2	pos. control	AS	4,7	4qA_S	7	4qB	>23	4qA	6	4qB	34
Pos3	pos. control	AS	6,8	4qA_S	8	4qB	22	4qA	7	4qB	21
Neg1	neg. control	WGS	17,4	4qA_L	42	4qA_S	64	4qA	42	4qA	63

RU and a 4qB allele with 20 RU. All results were confirmed with OGM. Two AS experiments were also performed for HG001 with the same results. Three positive FSHD1 cases (Pos1–3) served as test samples during development. Despite low mean coverage in Pos2 (4.7 ×) and Pos3 (6.8 ×), both chr4 alleles could be resolved and validated with OGM. In Pos2, only one chr10 allele was detected by DUCKS4 (32 RU), but heterozygous SNVs indicated a second allele. Pos1 achieved a mean depth of ~ 50 × using adaptive sampling. For Pos1 the extended methylation profiling approach via ID2bam2meth was used demonstratively at the chr4 and chr10 alleles: an increasing methylation gradient can be observed on the contracted and hypomethylated 4qA-allele (5RU) and the contracted and hypomethylated 10qA-allele (7RU; see Supplementary Figure S2). Three negative controls (Neg1, Neg2, Neg3) were also analysed, with Neg1 additionally validated by OGM.

Minor discrepancies of ± 1 RU between Nanopore and OGM results are attributed to the S/L structure of the distal repeat. For 4qA_S haplotypes, which terminate in a shortened RU (~ 1.8 kb), OGM typically undercounts by one RU. In contrast, for 4qA_L (e.g. 4A161L), counts were identical across platforms (Table 2). For 4qB, which is not permissive and therefore not stratified by S/L, a ± 1 RU difference was also observed, though clinically irrelevant.

The results for reference samples, positive and negative control are shown. WGS = whole genome sequencing, AS = adaptive sampling. The complete table with all samples is in Supplementary Table S2.

Analysis of atypical and clinically relevant cases

Clinical diagnostic samples may harbor structurally complex variants, including somatic mosaicism, and structural variants like chromosomal translocations. Given their potential to complicate molecular diagnosis, representative cases were selected to evaluate the performance of the DUCKS4 workflow in resolving such complex genomic rearrangements.

Detection of mosaicism

Additionally, to a FSHD1-positive case, we investigated the asymptomatic mother (MOS1), who previously showed only mildly elevated creatine kinase levels, an unspecific but potentially supportive finding for FSHD [20]. Southern blotting was not available, prompting long-read analysis for molecular characterization.

Overview statistics (Supplementary Table S1) revealed exclusively 4qA-reads and three distinct D4Z4 arrays: two complete arrays with 2 and 16 RU, and one partial array exceeding 26 RU. Based on DUCKS4 allele-specific statistics (Fig. 4A) and the size of the most distal RU (S/L status), the reads could be assigned to two 4qA_L alleles (2 RU and > 26 RU) and one 4qA_S allele (16 RU). The latter lacked the proximal DBET (D4F104S1) and CLUHP4 regions and showed a ~ 37.9 kb gap within the repeat array, consistent with a D4Z4-proximal extended deletion (DBED or p13E-11 deletion allele). This was independently confirmed using Sniffles2 detecting a ~ 8 kb deletion including CLUHP4 and DBET (Fig. 4A, B). Such permissive 4qA-DBED alleles occur at low frequency (~ 3%) within the population [16, 51–53]. Methylation profiling (Fig. 4C) showed the 2 RU allele to be hypomethylated (41.2%, ~ 25% read depth), indicating a permissive allele in mosaic, whereas the > 26 RU and 16 RU alleles were hypermethylated (80.7% and 73.5%, respectively) each with ~ 25 and ~ 50% read depth.

The 2 RU and the 16 RU DBED allele were fully resolved. The > 26 RU allele could be identified from both sides of the D4Z4 array via SNV and S/L distinction, but not fully spanned. Repeat composition of all reads is provided in Supplementary Data (MOS1). OGM was not performed on this sample.

Translocation chr4/chr10

Translocations involving the 4qA subtelomeric region and proximal 10q D4Z4 arrays have been described as rare causes of FSHD1 [17]. Such events may generate chimeric reads, which are reads that span sequences from both chromosomes, visible as split or supplementary

A

read.id	RU_count	S_or_L	status	repeat_sequence
MOS1_4qA_r2	17	S	complete	4qB_pLAM_low-pid - c10 - gap_37857bp - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4S - 4qA_pLAM - DUX4_end
MOS1_4qA_r3	17	S	complete	4qB_pLAM_low-pid - c10 - gap_37971bp - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4S - 4qA_pLAM - DUX4_end
MOS1_4qA_r4	17	S	complete	4qA_pLAM - c4S - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - gap_39609bp - c10 - 4qB_pLAM_low-pid
MOS1_4qA_r30	2	L	complete	4qA_D4F104S1 - c4 - c4L - 4qA_pLAM - DUX4_end
MOS1_4qA_r31	2	L	complete	CLUHP4-201_exon1 - 4qA_D4F104S1 - c4 - c4L - 4qA_pLAM
MOS1_4qA_r32	2	L	complete	CLUHP4-201_exon1 - 4qA_D4F104S1 - c4 - c4L - 4qA_pLAM
MOS1_4qA_r35	2	L	complete	CLUHP4-201_exon1 - 4qA_D4F104S1 - c4 - c4L - 4qA_pLAM - DUX4_end
MOS1_4qA_r1	26	L	partial	DUX4_end - 4qA_pLAM - c4L - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4
MOS1_4qA_r5	16	S	partial	c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4S - 4qA_pLAM - DUX4_end
MOS1_4qA_r6	15		partial	CLUHP4-201_exon1 - 4qA_D4F104S1 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4
MOS1_4qA_r7	11	S	partial	c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4S - 4qA_pLAM - DUX4_end
MOS1_4qA_r8	10		partial	c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - 4qA_D4F104S1 - CLUHP4-201_exon1 - c4_ctrl - c4_ctrl - 4qB_pLAM_low-pid
MOS1_4qA_r9	10		partial	c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - 4qA_D4F104S1 - CLUHP4-201_exon1
MOS1_4qA_r10	10		partial	c4_ctrl - c4_ctrl - CLUHP4-201_exon1 - 4qA_D4F104S1 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4
MOS1_4qA_r11	10	S	partial	DUX4_end - 4qA_pLAM - c4S - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4
MOS1_4qA_r12	10	L	partial	4qA_pLAM - c4L - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4
MOS1_4qA_r13	9	L	partial	c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4L - 4qA_pLAM
MOS1_4qA_r14	9	S	partial	c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4S - 4qA_pLAM
MOS1_4qA_r15	8		partial	c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - 4qA_D4F104S1 - CLUHP4-201_exon1
S1_4qA_r16	7	L	partial	4qA_pLAM - c4L - c4 - c4 - c4 - c4 - c4 - c4
S1_4qA_r17	6	S	partial	4qA_pLAM - c4S - c4 - c4 - c4 - c4 - c4

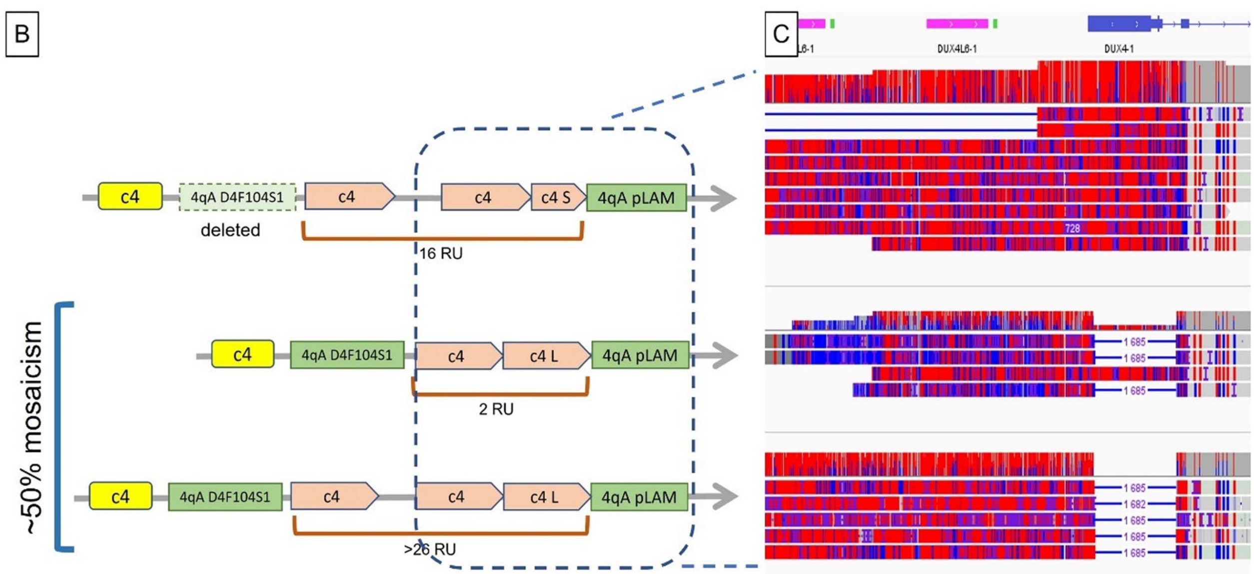

Fig. 4 Somatic mosaicism of MOS1, read composition, and genotype. A. Excerpt from the DUCKS4 output for the 4qA reads with the repeat composition featuring S/L distal RUs. B. Genotype schemes of the three 4qA alleles: The ~ 8 kb deletion of the DBET-region with D4F104S1 (so-called DBED) is shown. C. Alignment at the distal end of the D4Z4-array, example reads of the three 4qA alleles with methylation data are shown separately: blue–hypomethylated, red–hypermethylated. The 4qA_S allele had around 50% of the read depth and was methylated. The other reads were split into 2 more alleles: the contracted and hypomethylated 4qA_L-allele with 2 RU and the methylated 4qA_L allele with > 26 RU. Data visualized with IGV-viewer

alignments. However, structural rearrangements within the highly homologous D4Z4 array, especially with breakpoints inside the array, are not detectable by OGM due to resolution limitations [22] and can also pose challenges for ONT-based variant calling and phasing. Within a family of four cases (mother with three sons, Trans1-Trans4) chimeric reads and the complete chr 10 reads could not be properly phased and remained in the unphased alignment pile-up (example Trans1, Supplementary Figure S1). Furthermore, minimap2's supplementary alignment feature labeled only a subset of chimeric reads: 33% (Trans1), 14% (Trans2), 17% (Trans3), and 18% (Trans4) (Supplementary Table T1–T4 analysis). Such misalignment may compromise the detection of translocations.

The DUCKS4 workflow includes a dedicated chimerism-detection module that filters and annotates chimeric reads, identifying proximal and distal haplotypes as well as full repeat composition (Table 2). In a family with four affected individuals (mother and three children, Trans1–Trans4), a translocation was consistently identified (t(4;10)(q35;q26)). For example, in Trans1, two 4qA alleles (20 and 41 RU), one chr10 allele (24 RU), and one chimeric allele (5 RU) were detected (Supplementary Table S1). Of 18 chimeric reads, 9 included a chr10 marker (c10_ctrl), indicating that these reads extended beyond the D4Z4 region into the unique chr10

sequences. All chimeric reads had a consistent composition: 4 chr10-type RUs followed by 1 chr4-type RU ending in a 4qA_S haplotype, suggesting a translocation breakpoint between the 4th and 5th RU of chr10 (see Fig. 5 upper panel). The chimeric reads were further aligned against a selected read via the ID2bam2meth workflow and the methylation analysis further showed an increasing gradient of methylation over the D4Z4-array (see Fig. 5 lower panel).

To confirm the D4Z4 unit types in chimeric reads, restriction enzyme recognition sites were inspected manually. Based on XapI and BlnI digestion motifs from Southern Blotting, RU types were assigned as follows: chr4-type (B−/X+), chr10-type (B+/X−), and hybrid-type (B−/X−, typical for non-permissive 4qA 4A166 alleles). This confirmed the composition inferred by DUCKS4. Additional SNV and SSLP analyses further resolved the proximal segment as 10A166b and the distal segment as either 4A161S or 4A168, with intact 4qA pLAM PAS sequences, confirming a de novo translocation with a permissive haplotype.

Haplotyping-sub-haplotypes and non-standard chromosomes

The analysis of D4Z4 arrays is complicated by a wide diversity of non-canonical alleles and sub-haplotypes on both chromosomes 4 and 10. Over 30 sub-haplotypes have been described to date, including at least 9 for 4qA, 8 for 4qB, and 13 for chr10 [7, 11, 16, 17, 38]. This complexity likely reflects the high sequence homology between chr4 and chr10, the subtelomeric position of D4Z4, and its repetitive nature.

Chimeric reads identified by DUCKS4 can point to rare sub-haplotypes or structural variants. In cases where chromosome-specific markers are absent or alleles are incomplete, manual inspection of the SSLP region and proximal restriction enzyme sites (BlnI/XapI) can aid in haplotype resolution. To support this, we provide supportive material like a haplotype identification key (Supplementary Table S8) and a BED file (Haplotypes_identification_regions.bed, Supplementary Table S7) marking key regions (SSLP, enzyme motifs, informative SNPs), which can be cross-referenced with SNP data for D4F104S1 and pLAM (Supplementary Tables S5, S6).

In our dataset of 14 samples, six contained at least one non-canonical allele. For example, HG001 showed two chimeric D4Z4 alleles composed of a proximal chr10-type D4F104S1 and a distal chr4-type pLAM: one with 48 RUs (4qA) and one with 20 RUs (4qB) (Fig. 6). Two canonical chr10 alleles (16 and 21 RUs) were also present. Chromosome 4-specific markers were detected for both chimeric alleles. Based on the DUCKS4 haplotype key (Supplementary Table S8), the 48 RUs allele, which showed two proximal c10-D4Z4 hits followed by c4-D4Z4 hits, was fully assigned to the permissive 4qA sub-haplotype 4A166H, whereas the 20 RU allele was classified as a 4qB haplotype. For verification, the alignment of the 48 RU chimeric repeat confirmed a characteristic restriction pattern across the first three repeat units (B^+X^-, B^+X^-, B^-X^+), reflecting a chr10–chr10–chr4 origin matching the DUCKS4 results (Fig. 6 upper and

read.id	RU_count	S_or_L	status	repeat_sequence	D4F104S1_type	pLAM_type
T1_chim_r1	5	S	complete	c10_ctrl - no_chr4_hit - CLUHP4-201_exon1 - 10qA_D4F104S1 - c10 - c10 - c10 - c10 - c4S - 4qA_pLAM - DUX4	10qA	4qA
T1_chim_r2	5	S	complete	CLUHP4-201_exon1 - 10qA_D4F104S1 - c10 - c10 - c10 - c10 - c4S - 4qA_pLAM - DUX4_end	10qA	4qA
T1_chim_r3	5	S	complete	CLUHP4-201_exon1 - 10qA_D4F104S1 - c10 - c10 - c10 - c10 - c4S - 4qA_pLAM - DUX4_end	10qA	4qA
T1_chim_r4	5	S	complete	no_chr4_hit - CLUHP4-201_exon1 - 10qA_D4F104S1 - c10 - c10 - c10 - c10 - c4S - 4qA_pLAM - DUX4_end	10qA	4qA
T1_chim_r5	5	S	complete	CLUHP4-201_exon1 - 10qA_D4F104S1 - c10 - c10 - c10 - c10 - c4S - 4qA_pLAM - DUX4_end	10qA	4qA
T1_chim_r6	5	S	complete	c10_ctrl - no_chr4_hit - CLUHP4-201_exon1 - 10qA_D4F104S1 - c10 - c10 - c10 - c10 - c4S - 4qA_pLAM - DUX4	10qA	4qA
T1_chim_r7	5	S	complete	c10_ctrl - no_chr4_hit - CLUHP4-201_exon1 - 10qA_D4F104S1 - c10 - c10 - c10 - c10 - c4S - 4qA_pLAM - DUX4	10qA	4qA

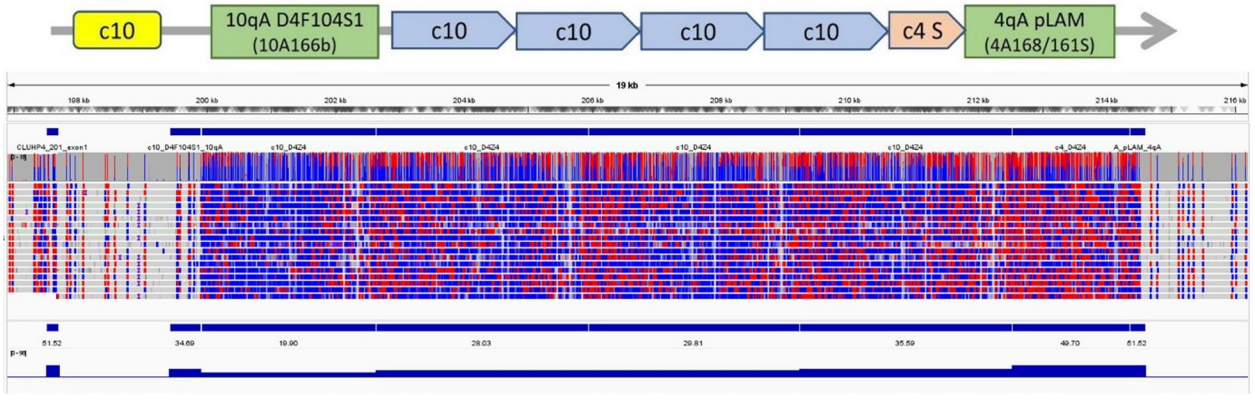

Fig. 5 Resolved haplotype of the translocation of Trans1. Upper panel shows excerpt from the DUCKS4 detailed output for chimeric reads. The middle panel shows the resolved genotype of the translocated allele. The lower panel shows the methylation for each segment of the D4Z4-array and depicting a methylation gradient; reads aligned against a custom reference (DUCKS4_ID2bam2meth workflow, Tool: IGV-desktop-viewer, red–hypermethylated, blue–hypomethylated)

read.id	RU count	S/L	status	repeat_sequence	D4F104S1 type	pLAM type	count c4	count c10
HG1_chim_r1	48	S	complete	4qB_pLAM_low-pid - c4_ctrl - c4_ctrl - CLUHP4-201_exon1 - 10qA_D4F104S1 - c10 - c10 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4S - 4qA_pLAM - DUX4_end	10qA	4qA	46	2
HG1_chim_r2	48	S	complete	4qB_pLAM_low-pid - c4_ctrl - c4_ctrl - CLUHP4-201_exon1 - 10qA_D4F104S1 - c4 - c10 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4S - 4qA_pLAM - DUX4_end	10qA	4qA	47	1
HG1_chim_r3	48	S	complete	4qA_pLAM - c4S - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c10 - c10 - 10qA_D4F104S1 - CLUHP4-201_exon1 - c4_ctrl - c4_ctrl - 4qB_pLAM_low-pid	10qA	4qA	46	2
HG1_chim_r4	48	S	complete	4qA_pLAM - c4S - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c10 - c10 - 10qA_D4F104S1 - CLUHP4-201_exon1	10qA	4qA	46	2
HG1_chim_r5	20	NA	complete	CLUHP4-201_exon1 - 10qA_D4F104S1 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - 4qB_pLAM	10qA	4qB	20	0
HG1_chim_r6	20	NA	complete	c4_ctrl - CLUHP4-201_exon1 - 10qA_D4F104S1 - c4 - c4 - c10 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - c4 - 4qB_pLAM	10qA	4qB	19	1

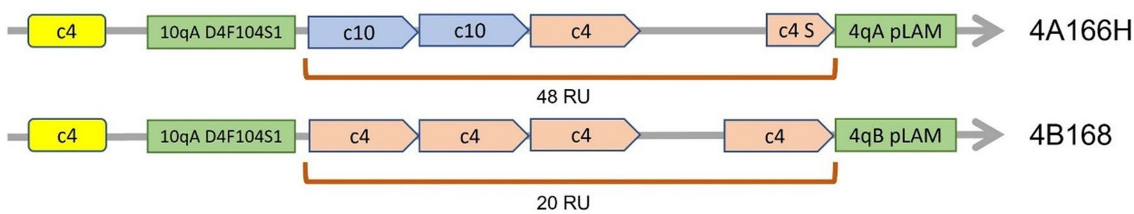

Fig. 6 HG001 chimeric-reads sequence composition. Upper panel shows DUCKS4's detailed chimeric-reads output. A warning is given when it indicates a sub-haplotype or non-standard chromosome rather than a translocation. Lower panel shows the scheme of resolved haplotypes. No FSHD2 associated variants were detected

lower panel), which is characteristic for the 4A166H sub-haplotype [17]. SSLP-region inspection further verified both alleles and identified the 20 RU allele as the 4qB subtype 4B168.

Discussion

In this study, we present DUCKS4, an automated read-based pipeline to resolve the D4Z4 array and detect haplotypes, repeat number, and composition per read. A key advantage is its initial BLAST-based approach, which avoids issues with alignment and phasing against a fixed reference due to the high regional homology between the chromosomes 4 and 10 and the extreme heterogeneity within the D4Z4 locus itself. Therefore, the blast-based pre-selection approach allows for precise annotation of the repeat-sequence and its specific chromosome/haplotype motifs for the individual reads. This is especially beneficial for complex alleles, non-standard chromosomes, and sub-haplotypes, where read alignments often hinder interpretation. By sorting annotated reads by haplotype and completeness, aligning filtered subsets, and providing fully annotated D4Z4-array tables per read, DUCKS4 simplifies sub-haplotype assignment and structural characterization.

DUCKS4 reliably distinguishes canonical 4qA, 4qB, and 10qA alleles, as well as detects chimeric reads, indicating sub-haplotypes or non-standard chromosomal configurations. This is critical, as complex D4Z4 alleles can lead to false positive/negative genetic FSHD1 diagnosis with other methods like Southern blotting and/or OGM. For example, the 4qA subtype 4A166 is non-permissive, despite an intact PAS, and occurs in ~4% of Europeans, undetectable by OGM, and only by extended Southern blotting [16, 17, 38]. Hybrid 4qA alleles containing both chr4- and chr10-type repeat units may mimic chr10 alleles in Southern blotting, leading to false-negative results, whereas OGM detection is unaffected [16, 22, 23]. In contrast, non-canonical chromosomes (e.g. ancient translocated founder alleles, 10qA/BT) are non-permissive, but may be mistaken for 4qA alleles in Southern blotting [16, 54, 55].

Within the main workflow of DUCKS4 methylation is quantified at the distal end of the 4qA D4Z4 array, following established approaches described previously [32]. An extended methylation profiling workflow is introduced in the sub workflow ID2bam2meth of DUCKS4 allowing for methylation profiling over the whole D4Z4-array, if needed. As demonstrated by the sample Pos1 and Trans1, an increasing methylation gradient can be observed on both contracted and hypomethylated 4qA and 10qA alleles of Pos1 and over the translocated contracted 10qA/4qA allele of Trans1, which is also described in Butterfield [27] for contracted D4Z4-arrays (see Fig. 5 and Supplementary Figure S2).

Complementing the workflow, we also present a sample preparation protocol using HMW-DNA, either enriched via Nanopore adaptive sampling or from WGS data. We successfully implemented adaptive sampling for human HMW-DNA from whole blood samples, enabling targeted sequencing of all relevant regions required for comprehensive FSHD analysis. An important practical advantage of adaptive sampling compared to conventional WGS is its substantially improved cost and resource efficiency. While a single WGS experiment typically consumes the full capacity of a PromethION flow cell, adaptive sampling requires only a fraction of the available sequencing output. In our hands, this enabled up to three, and in rare cases four, independent adaptive

sampling runs per flow cell, allowing multiple samples to be enriched sequentially on a single flow cell and making the approach more economical for targeted FSHD analysis.

For improved enrichment performance, we recommend selecting relatively large target regions and/or multiple target loci, as enrichment efficiency improves when a sufficient fraction of the reference genome is covered. This consideration motivated our use of extended target regions spanning approximately 0.5% of the human genome, which yielded robust enrichment in our experiments. In addition, we recommend extended flanking regions (also often referred as buffer regions) around the primary targets when working with HMW DNA, as adaptive sampling decisions rely on early sequence information from long DNA fragments. Consistent with Nanopore recommendations to use the N10 fragment length as a guideline, and given that we observed fragment lengths up to ~ 500 kb–1.5 Mb, we opted for larger flanking buffers than would be required for Standard DNA preparations. Beyond improving enrichment efficiency, these extended target regions can facilitate the detection of complex structural variation, including large deletions, rearrangements, and regions of loss of heterozygosity.

Taken together, these results show that DNA input quality, and especially the preservation of long DNA fragments, is critical for obtaining reads that reliably span the D4Z4 array, which is essential for confidently excluding FSHD1 (> 10–11 RU). While adaptive sampling is also feasible with Standard-DNA on the D4Z4 locus [56], such libraries typically exhibit N50 read lengths of only ~ 20–30 kb which limits the recovery of reads spanning longer alleles. In contrast, HMW DNA consistently increased the fraction of reads exceeding 11 RU (> ~ 37–40 kb).

Recently, D4Z4End2End was introduced as a long-read workflow aimed at reconstructing complete end-to-end D4Z4 arrays together with methylation profiles using a consensus-based assembly approach applied to Cas9-targeted sequencing data, which requires high sequencing depth [31]. In contrast, DUCKS4 is designed to accept long-read data from diverse sequencing strategies and focuses on read-level resolution using a blast-based annotation and preselection approach. This enables detailed haplotype and repeat characterization even in cases of incomplete array coverage or lower sequencing depth. In addition, DUCKS4 provides reference-guided methylation profiling across the D4Z4 array, facilitating comparison of array-wide methylation patterns. Taken together, the two approaches are therefore complementary, addressing different experimental designs and analytical priorities.

With HMW-DNA, full-length DNA repeat-expansions up to 30–50 RU can be fully resolved, whereas repeat-expansions larger than that are spanned at least partially. Performing an adaptive sampling run for 17 h on a PromethION flow cell typically yields coverage of 20–30 ×, and in some cases even up to 40–50 ×. Even for lower coverage samples (like Pos2 and Pos3), read lengths were sufficient to differentiate haplotypes and detect contracted FSHD1 repeat arrays. Overall, the DUCKS4 tool is flexible and can process any long-read sequence data provided in BAM/UBAM or FASTQ format and works independently from the here described wet-lab and sequencing approach.

In certain scenarios, such as hybrid-alleles (e.g. 4A166), we recommend to verify DUCKS4 analysis by manual confirmation. The current BLAST-based workflow distinguishes chr4 and chr10 type D4Z4 RUs, but cannot resolve hybrid RUs lacking BInI and XapI restriction sites (B-/X-). Instead, chr4 repeat units are called (data not shown), requiring manual inspection. Furthermore, the workflow does not make an automated overall evaluation, as the pathobiomechanism for FSHD is still not completely understood. Therefore, the results always need to be genetically interpreted. While low-coverage FSHD1 controls were successfully resolved, we advise against using low-coverage datasets.

For reliable performance, we recommend ≥ 15 × read depth, ideally 20–25 ×, which is recommended for SV and SNV-calling. The optional SNV annotation pipeline provides preliminary annotation of variants based on SnpSift [45]-annotation against the ClinVar [46] database. Identified variants, together with the methylation status of the D4Z4-array should be further reviewed by clinical geneticists following established guidelines for pathogenicity assessment and integrating patient-specific clinical data to confirm clinical relevance if used for diagnostic purposes. In future versions of DUCKS4, an extended SNV analysis pipeline could be implemented to improve haplotype resolution and support FSHD2 diagnostics. Integration of automated SSLP sizing is also a promising addition under consideration.

Methodologically, DUCKS4 demonstrates that BLAST-based pre-selection with per-read annotations and selective alignment is an effective strategy for robust analysis in highly repetitive and variable, segmentally duplicated regions minimizing manual curation.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40246-026-00921-2.

Additional file 1.

Additional file 2.

Additional file 3.

Acknowledgements
We would like to thank Mateja Smogavec, Jürgen Neesen, Jeannette Ackermann, Jean Elbers, Alina Malashka, and the diagnostic teams from the Institute of Medical Genetics of the Medical University of Vienna and Isabel Abellan Schneyder and her team from Oxford Nanopore Technologies for the technical support. We would also like to thank Agnieszka Pawelak from the Department of Genetics of the Wroclaw Medical University.

Author contributions
F.L. S.M. and P.D. conceived and supervised the project; M. G.-B., T. K., and P.S. supported sampling acquisition; T.L., D.H., and N.B. performed data analysis; T.L., D.H., and N.G. collected data; P.D., A.S., and N.G. performed OGM; S.M., N.B., A.S., M.G.B. and M.H. advised the project; M.H. acquired funding; T.L. wrote the manuscript; All authors read and edited the manuscript.

Funding
The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability
Due to the sensitive nature of the patient information and the potential for identifying individual participants, a portion of the data underlying this study cannot be made publicly available. The data may be obtained from the corresponding author upon reasonable request. The other raw data underlying this article are available on https://figshare.com/articles/dataset/HG001-HG002-HG003_high_molecular_weight_HMW_DNA_D4Z4-locus_on_chr4_chr10/29930690

Code availability
Source code is accessible at https://github.com/tamara-nano/ducks4, https://github.com/tamara-nano/ducks4_wovar (light version without variant calling) and archived at https://figshare.com/projects/DUCKS4-FSHD/260306.

Declarations

Ethics approval and consent to participate
This study was approved by the ethics committee of the Medical University of Vienna under EK Nr. 1078/2024. Informed consent was obtained from all the patients included in this study.

Competing interests
The authors declare no competing interests.

Received: 17 November 2025 / Accepted: 20 January 2026
Published online: 06 February 2026

References

1. Deenen JC, Arnts H, van der Maarel SM, et al. Population-based incidence and prevalence of facioscapulohumeral dystrophy. Neurology. 2014;83(12):1056–9. https://doi.org/10.1212/WNL.0000000000000797. (**PMID: 25122204; PMCID: PMC4166358**).
2. Vercelli L, Mele F, Ruggiero L, et al. A 5-year clinical follow-up study from the Italian National Registry for FSHD. J Neurol. 2021;268(1):356–66. https://doi.org/10.1007/s00415-020-10144-7.
3. Statland JM, McDermott MP, Heatwole C, et al. Reevaluating measures of disease progression in facioscapulohumeral muscular dystrophy. Neuromuscul Disord. 2013;23(4):306–12. https://doi.org/10.1016/j.nmd.2013.01.008.
4. Tawil R, Figlewicz DA, Griggs RC, et al. Facioscapulohumeral dystrophy: a distinct regional myopathy with a novel molecular pathogenesis. Ann Neurol. 1998;43:279–82. https://doi.org/10.1002/ana.410430303.
5. Tawil R, Van Der Maarel SM. Facioscapulohumeral muscular dystrophy. Muscle Nerve. 2006;34:1–15. https://doi.org/10.1002/mus.20522.
6. Tawil R, van der Maarel SM, Tapscott SJ. Facioscapulohumeral dystrophy: the path to consensus on pathophysiology. Skelet Muscle. 2014;4:12. https://doi.org/10.1186/2044-5040-4-12.
7. Lemmers RJ, van der Vliet P, Klooster R, et al. A unifying genetic model for facioscapulohumeral muscular dystrophy. Science. 2010;329(5999):1650–3. https://doi.org/10.1126/science.1189044. (**PMID:20724583;PMCID:PMC4677822**).
8. Lemmers RJ, Tawil R, Petek L, et al. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. Nat Genet. 2012;44(12):1370–4. https://doi.org/10.1038/ng.2454.
9. Lim KRQ, Nguyen Q, Yokota T. DUX4 signalling in the pathogenesis of Facioscapulohumeral muscular dystrophy. Int J Mol Sci. 2020;21:729. https://doi.org/10.3390/ijms21030729.
10. Hamanaka K, Šikrová D, Mitsuhashi S, et al. Homozygous nonsense variant in *LRIF1* associated with facioscapulohumeral muscular dystrophy. Neurology. 2020;94(23):e2441–7. https://doi.org/10.1212/WNL.0000000000009617.
11. Lemmers RJ, van der Vliet P, van der Gaag K, et al. Worldwide Population Analysis of the 4q and 10q Subtelomeres Identifies Only Four Discrete Interchromosomal Sequence Transfers in Human Evolution. Am J Human Genet. 2010;86(3):364–377. ISSN 0002–9297. https://doi.org/10.1016/j.ajhg.2010.01.035.
12. van den Boogaard ML, Lemmers RJ, Balog J, et al. Mutations in DNMT3B Modify Epigenetic Repression of the D4Z4 Repeat and the Penetrance of Facioscapulohumeral Dystrophy. Am J Hum Genet. 2016;98(5):1020–9. https://doi.org/10.1016/j.ajhg.2016.03.013. (**PMID:27153398;PMCID:PMC4863565**).
13. Lemmers RJ, de Kievit P, Sandkuijl L, et al. Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. Nat Genet. 2002;32:235–6. https://doi.org/10.1038/ng999.
14. Van Geel M, Dickson MC, Beck AF, et al. Genomic analysis of human chromosome 10q and 4q telomeres suggests a common origin. Genomics. 2002;79(2):210–7. https://doi.org/10.1006/geno.2002.6690.
15. Delourme M, Charlene C, Gerard L, et al. Complex 4q35 and 10q26 rearrangements: a challenge for molecular diagnosis of patients with facioscapulohumeral dystrophy. Neurol Genet. 2023;9(3):e200076.
16. Giardina E, Camaño P, Burton-Jones S, et al. Best practice guidelines on genetic diagnostics of facioscapulohumeral muscular dystrophy: update of the 2012 guidelines. Clin Genet. 2024;106(1):13–26. https://doi.org/10.1111/cge.14533.
17. Lemmers, RJ, van der Vliet; P, Blatnik; A, et al. Chromosome 10q-linked FSHD identifies *DUX4* as principal disease gene. J Med Genet. 2022;59(2):180–188. https://doi.org/10.1136/jmedgenet-2020-107041. PMID: 33436523; PMCID: PMC8273184.
18. Lemmers RJLF, Butterfield RJ, van der Vliet PJ, de Bleecker JL, van der Pol L, Dunn DM, et al. Autosomal dominant in cis D4Z4 repeat array duplication alleles in facioscapulohumeral dystrophy. Brain. 2024;147(2):414–26. https://doi.org/10.1093/brain/awad312. (**PMID:37703328**).
19. Nguyen K, Puppo F, Roche S, et al. Molecular combing reveals complex 4q35 rearrangements in Facioscapulohumeral dystrophy. Hum Mutat. 2017;38(10):1432–41.
20. Preston, MK, Wang, L. Facioscapulohumeral muscular dystrophy. In: Adam MP, Feldman J, Mirzaa GM, et al., eds. *GeneReviews®* [Internet]. Seattle (WA): University of Washington, Seattle; 1993–2025. Updated 2025 Jul 10. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1443/.
21. Qiu L, Ye Z, Lin L, et al. Clinical and genetic features of somatic mosaicism in facioscapulohumeral dystrophy. J Med Genet. 2020. https://doi.org/10.1136/jmedgenet-2019-106638.
22. Guruju NM, Jump V, Lemmers R, et al. Molecular diagnosis of facioscapulohumeral muscular dystrophy in patients clinically suspected of FSHD using optical genome mapping. Neurol Genet. 2023;9(6):e200107. https://doi.org/10.1212/NXG.0000000000200107.
23. Stence AA, Thomason JG, Pruessner JA, et al. Validation of optical genome mapping for the molecular diagnosis of facioscapulohumeral muscular dystrophy. J Mol Diagn. 2021. https://doi.org/10.1016/j.jmoldx.2021.07.021.
24. Angelini C. LGMD phenotype due to a new gene and dysferlinopathy investigated by next-generation sequencing. Neurol Genet. 2015;1(4):e39. https://doi.org/10.1212/NXG.0000000000000039.
25. Luglio A, Maggi E, Riviello F, et al. Hereditary neuromuscular disorders in reproductive medicine. Genes. 2024;15:1409. https://doi.org/10.3390/genes15111409.
26. Tan M, Huo H, Feng J, et al. Facioscapulohumeral muscular dystrophy type 1 combined with becker muscular dystrophy: a family case report. Front Genet. 2025;15:1522203. https://doi.org/10.3389/fgene.2024.1522203.
27. Butterfield RJ, Dunn DM, Duval B, Moldt S, Weiss RB. Deciphering D4Z4 CpG methylation gradients in fascioscapulohumeral muscular dystrophy using nanopore sequencing. Genome Res. 2023;33(9):1439–54. https://doi.org/10.1101/gr.277871.123.

28. Hiramuki Y, Kure Y, Saito Y, Ogawa M, Ishikawa K, Mori-Yoshimura M, et al. Simultaneous measurement of the size and methylation of chromosome 4qA-D4Z4 repeats in facioscapulohumeral muscular dystrophy by long-read sequencing. J Transl Med. 2022;20(1):517. https://doi.org/10.1186/s12967-022-03743-7.
29. Li K, Quiat D, She F, Liu Y, He R, Haghighi A, et al. Genetic diagnosis of facioscapulohumeral muscular dystrophy type 1 using rare-variant linkage analysis and long-read genome sequencing. Genet Med Open. 2024;2:101817. https://doi.org/10.1016/j.gimo.2024.101817.
30. Wang Y, Zhao Z, Meng F, Kong X. Accurate prenatal diagnosis of facioscapulohumeral muscular dystrophy 1 using nanopore sequencing. J Med Genet. 2024;61(12):1096–102. https://doi.org/10.1136/jmg-2023-109832.
31. Xiao LC, Semwal A, St John B, Zeglinski K, Su S, Lancaster J, et al. D4Z4End2End: complete genetic and epigenetic architecture of D4Z4 macrosatellites in FSHD, BAMS and reference cohorts. medRxiv. 2025. 2025.04.24.25326320. https://doi.org/10.1101/2025.04.24.25326320. PMID: not available (preprint).
32. Huang M, Zhang Q, Jiao J, et al. Comprehensive genetic analysis of facioscapulohumeral muscular dystrophy by Nanopore long-read whole-genome sequencing. J Transl Med. 2024;22:451. https://doi.org/10.1186/s12967-024-05259-8.
33. Yeetong P, Kulsirichawaroj P, Kumutpongpanich T, et al. Long-read Nanopore sequencing identified D4Z4 contractions in patients with facioscapulohumeral muscular dystrophy. Neuromuscul Disord. 2023;7(7):551–6. https://doi.org/10.1016/j.nmd.2023.05.004.
34. Rautiainen M, Nurk S, Walenz BP, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol. 2023;10(10):1474–82. https://doi.org/10.1038/s41587-023-01662-6.
35. Jones A, Torkel C, Stanley D, et al. High-molecular weight DNA extraction, clean-up and size selection for long-read sequencing. PLoS ONE. 2021;16:e0253830. https://doi.org/10.1371/journal.pone.0253830.
36. R Core Team. R: A language and environment for statistical ## computing. R Foundation for Statistical Computing, Vienna, Austria; 2021. https://www.R-project.org/
37. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.
38. Lemmers RJ, Wohlgemuth M, van der Gaag K, et al. Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. Am J Hum Genet. 2007;81(5):884–94. https://doi.org/10.1086/521986.
39. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021;37:4572–4. https://doi.org/10.1093/bioinformatics/btab705.
40. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):giab008. https://doi.org/10.1093/gigascience/giab008.
41. Zheng Z, Li S, Su J, et al. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. Nat Comput Sci. 2022;2:797–803. https://doi.org/10.1038/s43588-022-00387-x.
42. Martin M, Ebert P, Marschall T. Read-Based Phasing and Analysis of Phased Variants with WhatsHap. Methods Mol Biol. 2023;2590:127–38. https://doi.org/10.1007/978-1-0716-2819-5_8. (**PMID: 36335496**).
43. Smolka M, Paulin LF, Grochowski CM, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024;42:1571–80. https://doi.org/10.1038/s41587-023-02024-y.
44. Cingolani P, Platts A, le Wang, L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;2:80–92 (**PMID: 22728672**).
45. Cingolani P, Patel VM, Coon M, et al. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front Genet. 2012;3:35.
46. Landrum M, Lee J, Riley G, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(1):D980–5. https://doi.org/10.1093/nar/gkt1113. (**PMID: 24234437**).
47. Shen W, Sipos B, Zhao L. Seqkit2: a swiss army knife for sequence and alignment processing. iMeta. 2024. https://doi.org/10.1002/imt2.191.
48. De Coster W, D'Hert S, Schultz D, et al. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34(15):2666–9. https://doi.org/10.1093/bioinformatics/bty149.
49. Posit team. RStudio: integrated Development Environment for R. Boston MA: Posit Software, PBC; 2025. Available from http://www.posit.co/.
50. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. 2016. ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.
51. Lemmers RJ, Osborn M, Haaf T, et al. D4F104S1 deletion in facioscapulohumeral muscular dystrophy: phenotype, size, and detection. Neurology. 2003;61(2):178–83.
52. Lemmers RJ, van der Vliet P, Granado D, et al. High-resolution breakpoint junction mapping of proximally extended D4Z4 deletions in FSHD1 reveals evidence for a founder effect. Hum Mol Genet. 2022;31(5):748–60.
53. Nguyen K, Broucqsault N, Chaix C, et al. Deciphering the complexity of the 4q and 10q subtelomeres by molecular combing in healthy individuals and patients with facioscapulohumeral dystrophy. J Med Genet. 2019;56(9):590–601.
54. Lemmers RJ, O'Shea S, Padberg G, et al. Best practice guidelines on genetic diagnostics of Facioscapulohumeral muscular dystrophy: Workshop 9th June 2010, LUMC, Leiden, The Netherlands. Neuromusc Disord. 2012;22(5):463–470, ISSN 0960–8966: https://doi.org/10.1016/j.nmd.2011.09.004.
55. Zernov NV, Guskova AA, Skoblov MY. FSHD1 diagnosis in a Russian population using a qpcr-based approach. Diagnostics. 2021;11(6):982. https://doi.org/10.3390/diagnostics11060982.
56. Huang M, Zhang Q, Wu S, et al. Accurate detection of D4Z4 repeats, methylation and allele haplotype in facioscapulohumeral muscular dystrophy 1 using nanopore long-read adaptive sampling sequencing: a pilot study. J Med Genet. 2025. https://doi.org/10.1136/jmg-2025-110827.

Publisher's Note