Aus dem Friedrich-Baur-Institut der Neurologischen Klinik und Poliklinik der
Ludwig-Maximilians-Universität München

Direktor: Prof. Dr. med. Günter Höglinger

The genetics of repeat disorders:

Development of novel long-read sequencing methods and phenotype-genotype studies

Dissertation

zum Erwerb des Doktorgrades der Medizin

an der Medizinischen Fakultät der

Ludwig-Maximilians-Universität München

vorgelegt von

Hannes Erdmann

aus

Havelberg

2023

Mit Genehmigung der Medizinischen Fakultät der

Ludwig-Maximilians-Universität zu München

Erster Gutachter:	*Prof. Dr. med. Angela Abicht*
Zweiter Gutachter:	Prof. Dr. med. Benedikt Schoser
Dritter Gutachter:	Prof. Dr. med. Ortrud Steinlein

Dekan:	Prof. Dr. med. Thomas Gudermann

Tag der mündlichen Prüfung:	19.10.2023

Affidavit

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Promotionsbüro
Medizinische Fakultät

MMRS

Eidesstattliche Versicherung

Erdmann, Hannes

Name, Vorname

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel:

The genetics of repeat disorders:

Development of novel long-read sequencing methods and phenotype-genotype studies

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München, 23. Oktober 2023
Ort, Datum

Hannes Erdmann
Unterschrift Doktorandin bzw. Doktorand

Table of Contents

Abbreviations

ALS	amyotrophic lateral sclerosis
CANVAS	cerebellar ataxia, sensory neuropathy and vestibular areflexia syndrome
Clin-CATS	clinical nanopore Cas9-targeted sequencing
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DNA	desoxyribonucleic acid
dNTP	desoxynucleotides triphosphates
FMRP	fragile X mental retardation protein
FSHD	facioscapulohumeral muscular dystrophy
FSHD-MPA	facioscapulohumeral muscular dystrophy – methylation profile analysis
FRDA	Friedreich's ataxia
FXTAS	fragile-X-associated tremor/ataxia syndrome
FXS	fragile-X syndrome
HD	Huntington's disease
HMW	high-molecular-weight
NGS	next-generation sequencing
ONT	Oxford Nanopore Technology
ORF	open-reading frame
PacBio	Pacific Biosciences
PCR	polymerase chain-reaction
RAN	repeat associated non-AUG
RBP	RNA binding proteins
RNA	ribonucleic acid
RU	repeat unit
SCA	spinocerebellar ataxia
SMRT	single-molecule real-time
STR	short tandem repeat
TR	tandem repeat
UTR	untranslated region
VNTR	variable number tandem repeats
ZMV	zero-mode waveguide

List of Publications

This thesis is based on the following publications, reprinted in chapter 5, 6 and appendix I:

1 Hannes Erdmann, Florian Schöberl, Mădălina Giurgiu, Rafaela Magalhaes Leal Silva, Veronika Scholz, Florentine Scharf, Martin Wendlandt, Stephanie Kleinle, Marcus Deschauer, Georg Nübling, Wolfgang Heide, Sait Seymen Babacan, Christine Schneider, Teresa Neuhann, Katrin Hahn, Benedikt Schoser, Elke Holinski-Feder, Dieter A. Wolf, Angela Abicht.
Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing. *Brain*. 2023;146(5):1831–1843.
https://doi.org/10.1093/brain/awac377

2 Hannes Erdmann[±], Florentine Scharf[±], Stefanie Gehling, Anna Benet-Pagès, Sibylle Jakubiczka, Kerstin Becker, Maria Seipelt, Felix Kleefeld, Karl Christian Knop, Eva-Christina Prott, Miriam Hiebeler, Federica Montagnese, Dieter Gläser, Matthias Vorgerd, Tim Hagenacker, Maggie C. Walter, Peter Reilich, Teresa Neuhann, Martin Zenker, Elke Holinski-Feder, Benedikt Schoser, Angela Abicht.
Methylation of the 4q35 D4Z4 repeat defines disease status in facioscapulohumeral muscular dystrophy. *Brain.* 2023(4);146:1388–1402.
https://doi.org/10.1093/brain/awac336

3 Hannes Erdmann[±], Florentine Scharf[±], Ariane Hallermayr, Hayk Barsegehyan, Maggie C. Walter, Elke Holinski-Feder, Benedikt Schoser, Angela Abicht.
Reply to: An epigenetic basis for genetic anticipation in facioscapulohumeral muscular dystrophy type 1. *Brain.* Published online 22 June 2023.
https://doi.org/10.1093/brain/awad216

Further publications containing contributions by Hannes Erdmann and prepared within the time period of the dissertation in chronological order:

4 Peter Reilich, Beate Schlotter, Federica Montagnese, Berit Jordan, Friedrich Stock, Mario Schäff-Vogelsang, Benjamin Hotter, Katherina Eger, Isabel Diebold, Hannes Erdmann, Kerstin Becker, Ulrike Schön, Angela Abicht.
Location matters – Genotype-phenotype correlation in *LRSAM1* mutations associated with rare Charcot-Marie-Tooth neuropathy CMT2P. *Neuromuscul Disord.* 2021;31(2):123–133.
https://doi.org/10.1016/j.nmd.2020.11.011

5 Florian Schöberl[±], Angela Abicht[±], Clemens Kuepper, Stefanie Voelk, Stefan Sonnenfeld, Matthias Tonon, Annalisa Schaub, Veronika Scholz, Stephanie Kleinle, Hannes Erdmann, Dieter A. Wolf, Peter Reilich.
Sensory neuropathy due to *RFC1* in a patient with ALS: more than a conincidence? *J Neurolog.* 2022;269:2774–2777.
https://doi.org/10.1007/s00415-021-10835-9

6 Hannes Erdmann, Angela Abicht.
Häufige intronische Repeat-Expansionen in *FGF14* – eine weitere lang gesuchte Ursache bei spätmanifester Ataxie. *DGNeurologie.* 2023;6(2):159–160.
https://doi.org/10.1007/s42451-023-00535-1

[±] These authors contributed equally.

1 Contribution to the publications

1.1 Contribution to Paper I: Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing

Paper I is a first authorship of the doctoral candidate.

H.E., F.Schöberl, M.G., R.M.L.S., V.S., E.H.-F., D.A.W. and A.A. conceptualized and designed the study.

H.E., F.Schöberl, M.D., G.N., W.H., S.S.B., C.S., K.H. and B.S. recruited patients of this study.

H.E., R.M.L.S. and M.W. performed the experiments in the laboratory.

M.G., V.S. and F.Scharf implemented the bioinformatics pipeline.

H.E., F.Schöberl, M.G., R.M.L.S., V.S., F.Scharf, M.W., S.K., M.D., B.S., E.H.-F., D.A.W. and A.A. analyzed and interpreted data.

Data acquisition and collection was performed by all authors.

H.E., M.G., V.S. prepared the figures of the manuscript.

H.E., F.Schöberl, M.G., M.D., B.S., E.H.-F., D.A.W. and A.A. wrote the manuscript.

R.M.L.S., V.S., F.Scharf, M.W., S.K., G.N., W.H., S.S.B., C.S., T.N. and K.H. revised the paper and provided feedback.

D.A.W. and A.A. supervised the project and paper writing process.

1.2 Contribution to Paper II: Methylation of the 4q35 D4Z4 repeat defines disease status in facioscapulohumeral muscular dystrophy

Paper II is a joint first authorship of the doctoral candidate and Florentine Scharf. While Florentine Scharf primarily performed the implementation and validation of the method, Hannes Erdmann mainly studied the genotype-phenotype relationship and performed the revalidation of the method by comparing genetic and epigenetic parameters.

H.E., F.S, S.G., A.B.-P., B.S. and A.A. conceptualized and designed the study.

H.E., K.B., M.S., F.K., K.C.K., E.-C.P., M.H., F.M., M.V., T.H., M.C.W., P.R., B.S. and A.A. recruited patients of this study.

H.E., F.S., S.G., A.B.-P., S.J., D.G., T.N., M.Z., E.H.-F., B.S. and A.A. analyzed and interpreted data.

Data acquisition and collection was performed by all authors.

H.E. and F.S. prepared the figures of the manuscript.

H.E., F.S., E.H.-F., B.S. and A.A. wrote the manuscript.

S.G., A.B.-P., S.J., K.B., M.S., F.K, K.C.K., E.-C.P., M.H., F.M., D.G., M.V., T.H., M.C.W., P.R., T.N. and M.Z. revised the paper and provided feedback.

B.S. and A.A. supervised the project and paper writing process.

1.3 Contribution to Paper III: Reply to: An epigenetic basis for genetic anticipation in facioscapulohumeral muscular dystrophy type 1

Paper III is a joint first authorship of the doctoral candidate and Florentine Scharf. While Florentine Scharf focused on performing the bioinformatics analysis, Hannes Erdmann focused on writing the manuscript and interpreting the molecular data in context to the clinical findings. Both authors analyzed, discussed and interpreted data in a joint fashion.

H.E., F.S, B.S. and A.A. conceptualized and designed the study.

H.E., A.A. and M.C.W. recruited patients of this study.

Data acquisition and collection was performed by all authors.

H.E., F.S., A.H, H.B., E.H.-F., B.S. and A.A. analyzed and interpreted data.

H.E. and F.S. prepared the figures of the manuscript.

H.E., F.S., B.S. and A.A. wrote the manuscript.

A.H., H.B., M.C.W., E.H-F. revised the paper and provided feedback.

2 Introduction

2.1 Repetitive elements in the human genome and their role in pathology

It is estimated that half of the human genome consists of repetitive sequences,[7] with tandem repeats being among the most important representatives.[8] Tandem repeats (TR) are arrays of simple nucleotide sequences that are repeated in direct succession. Based on the size of one repeat unit, short tandem repeats (STRs, also known as microsatellites) with sizes of less than 10 base pairs are differentiated from variable number tandem repeats (VNTRs, also known as minisatellites) with 10 to 100 base pairs in size and satellites with more than 100 base pairs.[9] Large satellites consisting of several kilobases are called macrosatellites.[10] Due to slippage events during DNA replication, DNA repair, and nonallelic homologous replication, tandem repeats are highly unstable with respect to their length and show mutation rates of up to five orders of magnitude higher than the average mutation rate of the genome.[11,12] Therefore, TRs are a great source of phenotypic variability but also of heritable human disorders.[13] Currently, more than 50 disorders are known to be caused by the expansion of TRs beyond a locus-specific threshold (mostly STRs) that mainly impair different parts of the nervous system.[14] They include fragile-X syndrome (FXS, *FMR1* locus), Huntington's disease (HD, *HTT* locus), amyotrophic lateral sclerosis (ALS, *C9orf72* locus), Friedreich's ataxia (FRDA, *FXN* locus) and many other conditions.[15] Despite their high abundance and relevance in diseases, tandem repeats are poorly characterized due to difficulties in sequencing and assigning them to the reference genome using current short-read sequencing technologies.[16–18] As such, the genotype-phenotype correlation for known repeat disorders, as well as their pathomechanism is not fully understood. Especially the effect of repeat interruptions or base modifications, such as methylation, on the inheritance and disease severity requires further studies and a standardized analysis in genetic testing. Currently, diagnostics still rely on the laborious determination of the repeat length by Southern blotting or PCR-based methods.[1] Recently, the high potential of novel long-read sequencing technologies for the accurate repeat analysis was shown in the research setting.[19–21] These techniques allow for capturing not only the size of the repeats, but also the entire sequence and methylation status of a locus. This is especially important for the diagnosis of recently discovered complex repeat disorders such as *RFC1* spectrum disorder, SCA31, or SCA37 in which the motive of the altered tandem repeat sequence rather than the repeat size determines pathogenicity, as preliminary studies indicate.[22–24] Additionally, long-read sequencing techniques allow

multiple repeat disorders to be analyzed simultaneously. Thus, differential diagnoses can be evaluated within one analysis without selecting a subset of single analyses by preconceptions based on the clinical phenotype.[1] Long-read sequencing is likely to overcome the current limitation of repeat testing, refine known genotype-phenotype correlations and to identify additional TRs associated with human diseases.[1,25]

Facioscapulohumeral muscular dystrophy (FSHD) is unique among the repeat disorders, as it results from the contraction of a macrosatellite array (D4Z4 repeat array) and not the expansion of a STR.[26] As such, it presents with a unique pathomechanism: the derepression of a somatically silenced gene (*DUX4*) resulting in damage to the skeletal muscle.[27–29] Despite several therapeutic approaches to target FSHD and its high prevalence,[30–34] diagnostics based on genetic features (haplotype, repeat length, pathogenic variants in *SMCHD1,* and other epigenetic suppressor genes) remained imprecise.[2] For FSHD2 patients, in particular, there is a risk of not detecting the disease. Furthermore, predictive testing of family members is limited due to the incomplete penetrance of the disease, as is the prediction of disease severity. To overcome these diagnostic limitations, diagnostic methods based on epigenetic features (methylation) of the D4Z4 macrosatellite array have been proposed but are not implemented in routine diagnostics.[35–37] In particular, the recent debate on whether methylation plays a role in the pathogenesis of FSHD influences a broader investigation and implementation of such methods in diagnostics.[38,39]

2.2 Repeat expansion disorders as origin of adult-onset ataxia

2.2.1 General

The majority of known repeat expansion disorders originate from a microsatellite repeat array that is present in each individual and becomes pathogenic when it exceeds a certain size. Due to the high instability of microsatellite repeat arrays, repeat expansion disorders show some general features that differ from other genetically static disorders.[15] The size of microsatellite repeat arrays varies when inherited from one individual to another. As such, a repeat disorder may manifest in previously unaffected families, especially if paternal individuals carry repeat sizes in the upper normal range. In general, the longer the repeat, the more severe and the earlier the disease manifests. Because pathogenic repeat expansions tend to expand further in the next generation, individuals of this generation often exhibit more severe and earlier manifesting phenotypes, which is referred to as anticipation.

Repeat expansion disorders are a frequent cause of hereditary ataxia.[40] Hereditary ataxias are overlapping neurological conditions characterized by progressive stance and gait disorder, ocular motor disturbance, speech difficulties, limb ataxia, and dysdiadochokinesia.[1] The manifestation may result from dysfunction of the cerebellum, the spinal cord or peripheral sensory loss. Depending on their mode of inheritance, autosomal-dominant ataxias such as spinocerebellar ataxias (SCA) 1, 2, 3, 6, 7, 8, 17 and SCA27B (the latter discovered after publishing paper I of this thesis) are differentiated from the X-linked fragile-X-associated tremor/ataxia syndrome (FXTAS) and from autosomal-recessive ataxias such as Friedreich's ataxia (FRDA), *RFC1* spectrum disorder (Figure **1**).[1,14,41–43]

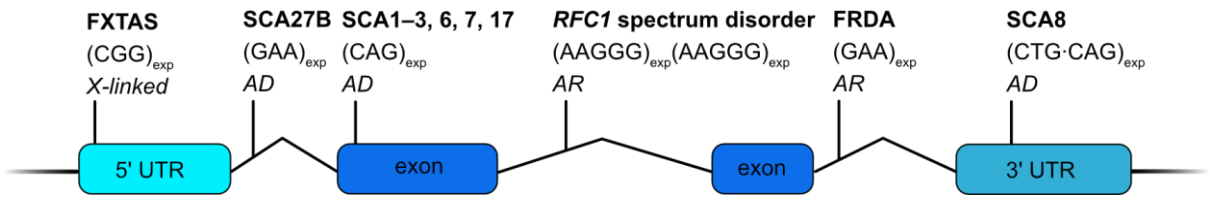

Figure 1. Schematic representation of the most prevalent repeat expansion disorders causing hereditary ataxia in the European population with their (pathogenic) microsatellite motive, mode of inheritance and position in the genome. Own illustration. *Abbreviations:* AD: autosomal-dominant, AR: autosomal-recessive.

2.2.2 Pathomechanism of repeat expansion disorders

Repeat expansions can be part of both coding as well as non-coding regions of the genome.[1] While first (e.g. SCA1, 2, 3, 6, 7, 17) are usually small in size and contain less than one hundred repeat units, repeat expansions in non-coding regions (SCA8, SCA27B, FRDA, *RFC1* spectrum disorder, FXTAS) are significantly larger.[1] They contain several hundred to several thousand repeat units. Associated with their location in the genome, there are different viable pathomechanisms (Figure 2) that can act at the same time:

(A) Transcriptional gene silencing: As in fragile-X syndrome (FXS) full mutations (> 200 CGG repeat units) within the 5'UTR of *FMR1* cause CpG promotor methylation and gene silencing (Figure 2A). Absence of the fragile-X mental retardation protein (FMRP) essential for brain development and neuronal signaling causes serve intellectual disability, developmental retardation and behavioral problems.[44,45] Transcriptional silencing is also the origin of Friedreich's ataxia. In contrast to FXS, expansions of the GAA repeat in intron 1 of the frataxin gene in Friedreich´s ataxia do not cause promotor methylation but the formation of secondary DNA structures hindering transcription by blocking RNA polymerase and heterochromatization.[46,47]

(B) Toxic gain of function of repeat mediated proteins: SCA1–3, 6, 7 and 17 are caused by expansion of the CAG microsatellites in the coding region. Expression results in proteins with large polyglutamine

chains that undergo conformational changes (Figure 2B), aggregate and induce toxicity to the cell.[48] Additionally, normal protein function is altered.

(C) RNA toxicity: RNA that contains a transcribed repeat expansion can form multiple secondary structures that bind RNA-binding proteins (RBPs) with high affinity forming RNA foci (Figure 2C). Sequestration of RBPs leads to their deficiency in physiological processes such as splicing resulting in disruption of cell physiology, e.g. in myotonic dystrophies.[49]

(D) Repeat associated non-AUG (RAN) translation: Translation of repetitive elements without an AUG start codon is the pathomechanism of several repeat expansion disorders such as FXTAS and SCA8 (Figure 2D). The mechanism of RAN translation initiation is only partly understood and likely involves secondary RNA structures formed by the repeat-containing regions.[50] There are multiple reading-frames in sense and antisense direction for each microsatellite that can result in different RAN proteins. Similar to polyglutamine proteins caused by CAG repeat expansions in coding regions, some of these proteins can accumulate and aggregate in specific cell types inducing toxicity by various mechanisms.[51] Protein-mediated toxicity likely overlaps with RNA-mediated toxicity.[52]

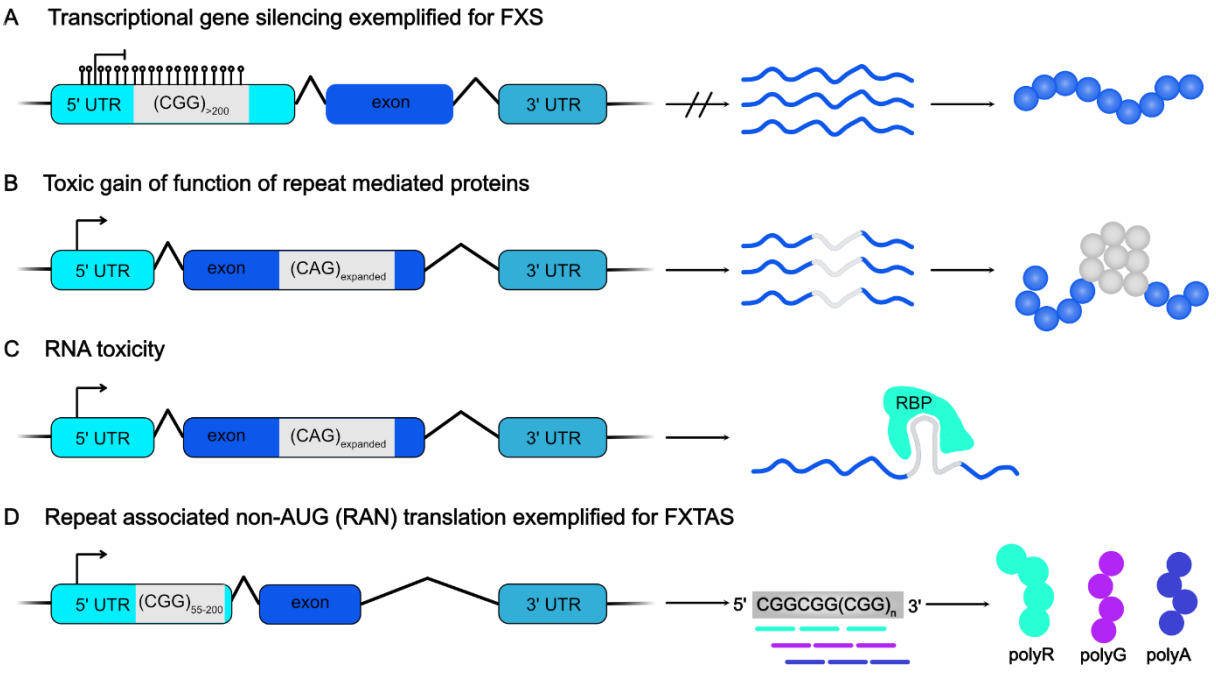

Figure 2. Exemplified pathomechanism of repeat expansion disorders. (A) Gene silencing mechanism as in FXS. A CGG repeat expansion causes promotor methylation and repression of *FMR1* transcription leading to the absence of FMRP. (B) Toxic gain of function of proteins as in SCA1–3, 6, 7 and 17. (C) RNA toxicity due to artificial binding of RBP disturbing physiological cellular processes. (D) RAN translation of the CGG repeat in FXTAS patients. RAN proteins resulting from antisense transcripts of the CGG repeat are not shown. Own illustration adapted from [53].

2.2.3 *RFC1* spectrum disorder – a complex repeat expansion disorder

Recently, the genetic origin of another adult-onset ataxia, *RFC1* spectrum disorder was identified that equals cerebellar ataxia, sensory neuropathy and vestibular areflexia syndrome (CANVAS) in its full presentation (Figure 3A).[24,54–56]

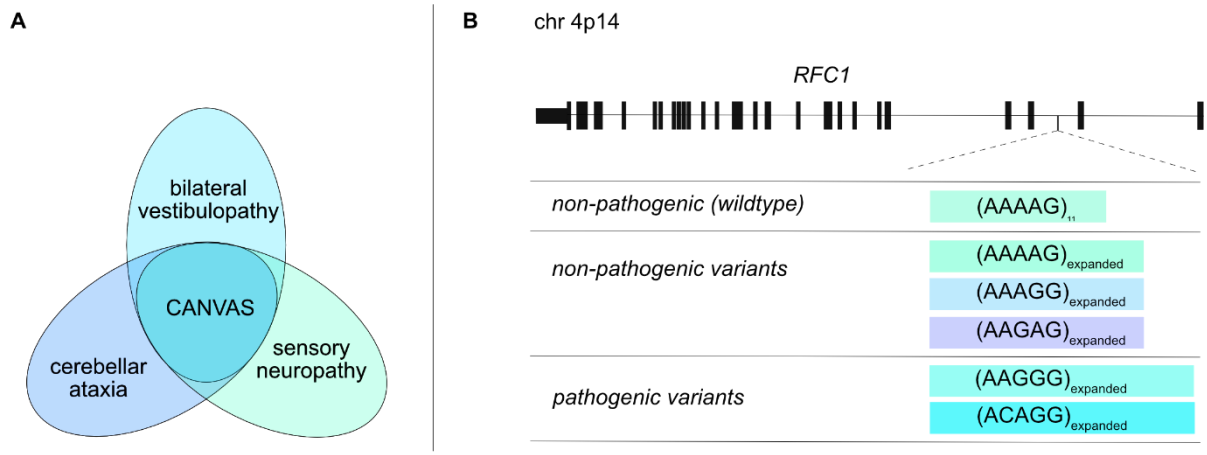

Figure 3. (A) Clinical spectrum of *RFC1* spectrum disorder. (B) Schematic representation of the *RFC1* locus and selected variants of the intronic microsatellite repeat array with their clinical evaluation. Own illustration adapted from [24].

In addition to the symptoms represented by the acronym, presyndromal irritative cough, autonomic dysfunction, motoneuron involvement and other symptoms may occur.[57,58] The disease is inherited autosomal-recessively and originates from a pentameric microsatellite array in the intronic region of the replication factor C1 (*RFC1*) gene. In contrast to other repeat expansion disorders, *RFC1* spectrum disorder does not simply rely on the expansion of a wildtype microsatellite array. Rather, it requires its substitution for a specific alternative microsatellite motive that is expanded (usually AAGGG) (Figure **3**B). Thus, diagnosis of *RFC1* spectrum disorder requires the determination of repeat motive and size. Various methods are currently used for diagnosing *RFC1* spectrum disorder such as Southern blotting or repeat-primed PCR. However, neither method can span the entire region and determine both repeat length and repeat motive.[1,59]

The detailed molecular mechanism by which biallelic AAGGG repeat expansions are causing multisystem neuronal damage involving the cerebrum, the cerebellum, and peripheral and cranial nerves remains elusive.[54,60] Recent studies have shown that patients who are compound-heterozygous for an AAGGG repeat expansion and a truncating pathogenic variant in *RFC1* can also develop a phenotype of *RFC1* spectrum disorder that tends to be more severe.[61,62] Reduced *RFC1* mRNA levels were detected in these patients compared to healthy individuals and patients with biallelic AAGGG repeat expansions in *RFC1*, suggesting a loss-of-function mechanism. A heterozygote carrier frequency of 0.7

to 4% in the European population indicates a high prevalence of *RFC1* spectrum disorder which might to be one of the most common causes of hereditary adult-onset ataxia in Europe.[1,24,55,56,63]

2.2.4 Long-read sequencing methods

Recently developed long-read sequencing techniques are capable to overcome current limitations of next-generation sequencing (NGS) in genetic testing, which are the analysis of structural variants, (large) repeat expansions and genes with corresponding pseudogenes as well as the phasing of variants.[18] Two commercial long-read sequencing platforms are currently available: Pacific Bioscience single molecule real-time sequencing (PacBio SMRT) and Oxford Nanopore Technology (ONT) sequencing (Figure 4).[64]

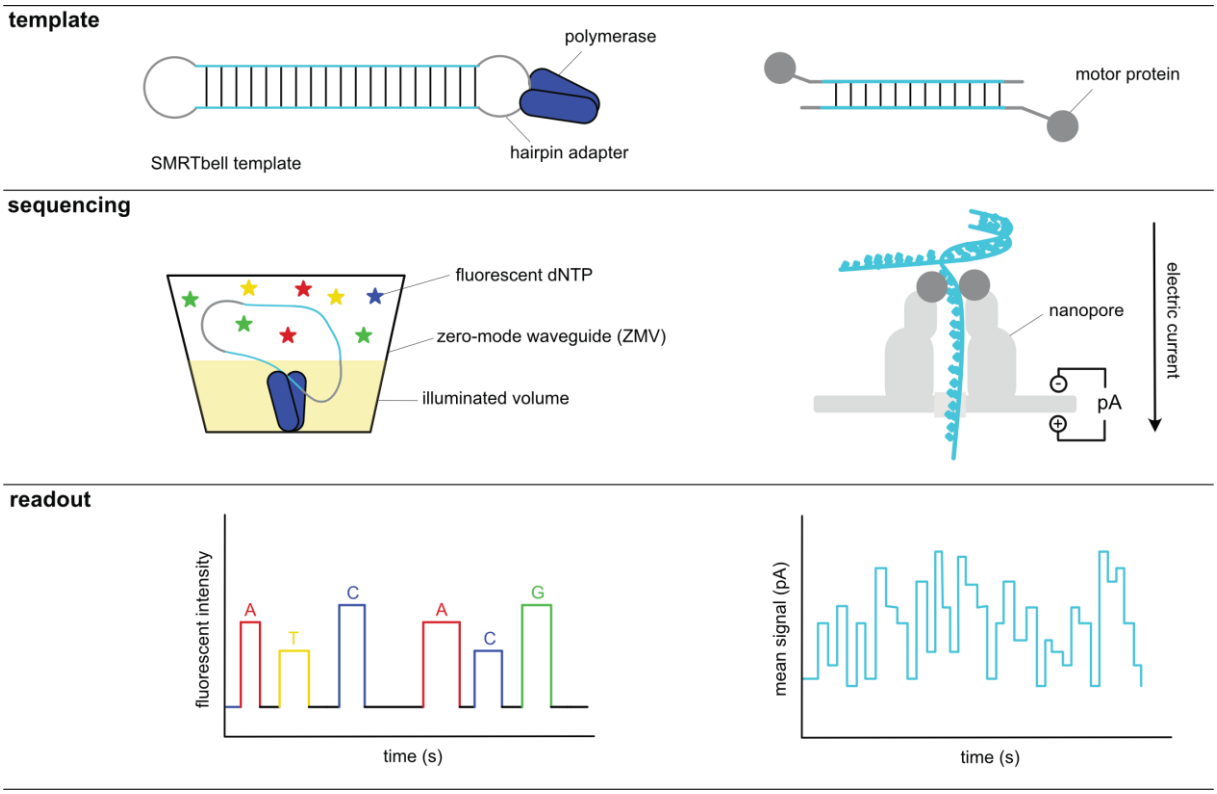

Figure 4. Principle of PacBio SMRT (left) and ONT sequencing (right). Own illustration adapted from [64].

PacBio SMRT is based on nanostructures that provide a small illuminated volume, called zero-mode waveguides (ZMW), sized to accommodate and observe only one DNA molecule at a time.[65] Each ZMW contains an immobilized DNA polymerase that binds the circular DNA template (SMRTbell). It consists of the double-stranded DNA fragment (up to several hundered kilobases in size) to be sequenced and hairpin adapters on both sides. The DNA polymerase turns around the SMRTbell and synthesizes a new DNA strand by incorporating one of the four desoxynucleotides triphosphates (dNTP) bound to a specific fluorescent label. After binding of the dNTP by the polymerase, the fluorophore is excited by a

laser, and the emission is detected by a camera. By native ligation of the dNTP to the existing DNA strand, catalyzed by the polymerase, the fluorophore is cleaved and diffuses out of the illuminated volume. The steps of dNTP incorporation and recording of the emission are repeated thousands of times and performed in parallel in the ZMWs of a flow cell (up to 8 million). While the chronological order of the different light emissions determines the DNA sequence, the kinetics of the polymerase reaction gives information about base modifications.

In contrast, the ONT platform does not rely on sequencing by synthesis and uses linear DNA molecules. A flow cell used for sequencing contains thousands of nanopores within a membrane that is under an electrical voltage.[66–68] A constant current flow passes through the nanopores. To determine the DNA sequence, the double-stranded DNA bound to an adapter is separated into single strands and pulled through the nanopore by a motorprotein, assisted by the polarity of the electrical voltage. When the DNA strand is translocated through the pore, the current changes depending on which nucleotide passes through and what modification it has. Recording and real-time analysis of the resulting current allows to determine the DNA sequence and its base modification (so-called base-calling).

2.2.5 First aim of this thesis – Diagnosis of repeat expansion disorders causing adult-onset ataxia

The first aim of this thesis is to implement a long-read sequencing method for the parallel diagnosis of the most prevalent repeat expansion disorders causing adult-onset ataxia in the European population. In addition to repeat length, which is the most important diagnostic parameter in current analyses, additional parameters such as repeat sequence as well as the methylation pattern in relevant loci will be determined. This allows for the diagnosis of *RFC1* spectrum disorder, the detection of repeat interruptions for the assessment of stability and pathogenicity of intermediate alleles and further characterization of expansions in the *FMR1* gene. For validating the method, repeat lengths determined by long-read sequencing will be compared to those determined by PCR based methods. Additionally, individuals with confirmed repeat expansions will be analyzed as positive controls. After implementing and validating the method, it will be applied to a cohort of patients with adult-onset ataxia. The composition of the *RFC1* repeat array in the whole cohort and the phenotype of patients with *RFC1* spectrum disorder will be characterized in detail.

The results of this study will be summarized in a publication (Paper I: Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing).[1]

2.3 Facioscapulohumeral muscular dystrophy – an epigenetic disease

2.3.1 Clinical and genetic background of facioscapulohumeral muscular dystrophy

Facioscapulohumeral muscular dystrophy (FSHD, OMIM #158900) is the third most common hereditary autosomal-dominant muscular dystrophy with an estimated prevalence of four to ten patients per 100.000 individuals.[33,34] FSHD is clinically characterized by slowly progressive and asymmetric weakness of facial muscles, muscles of the scapula, the upper limb, and the distal lower limb among variable manifestations in other muscles.[69] The clinical phenotype of FSHD varies highly regarding the involvement of muscle groups, the clinical severity or age at disease manifestation and overlaps with the phenotype of other myopathies.[70–72] The disease is caused by the epigenetic derepression of the double homeobox protein 4 (*DUX4*) gene which is silenced in somatic cells of healthy individuals after early embryonic development.[73]

In FSHD patients, stable *DUX4* expression in myocytes impairs various cellular signaling pathways leading to damage and cell death, which in turn results in muscle atrophy.[27–29] The entire *DUX4* open reading frame (ORF) is present in the last repeat unit of a D4Z4 macrosatellite repeat array in the subtelomeric 4q35 region (Figure 5).[74,75] Stable gene expression requires the presence of a specific haplotype (4qA and its variant 4qAL) that provides a polyadenylation signal in the *DUX4* ORF, which is referred to as permissive haplotype.[26]

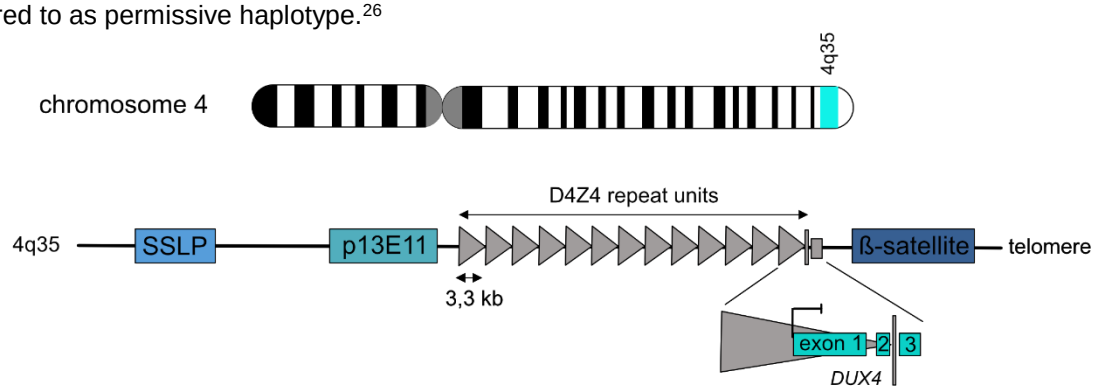

Figure 5. Location and architecture of the 4q35 region containing the D4Z4 repeat array in healthy individuals. *DUX4* present within the most distal repeat unit is silenced and not expressed. Own illustration.

FSHD shows an autosomal-dominant inheritance as it relies on a gain-of-function mechanism. While the majority of cases are associated with contractions of the D4Z4 repeat array to less than 12 repeat units on a permissive allele (referred to as FSHD1), a minority of cases (referred to as FSHD2) are

caused by hypomethylation of the D4Z4 repeat array on a permissive 4q allele due to pathogenic variants in *SMCHD1*, *DNMT3B*, *LRIF1* or other yet unknown factors (Figure 6).[76–78]

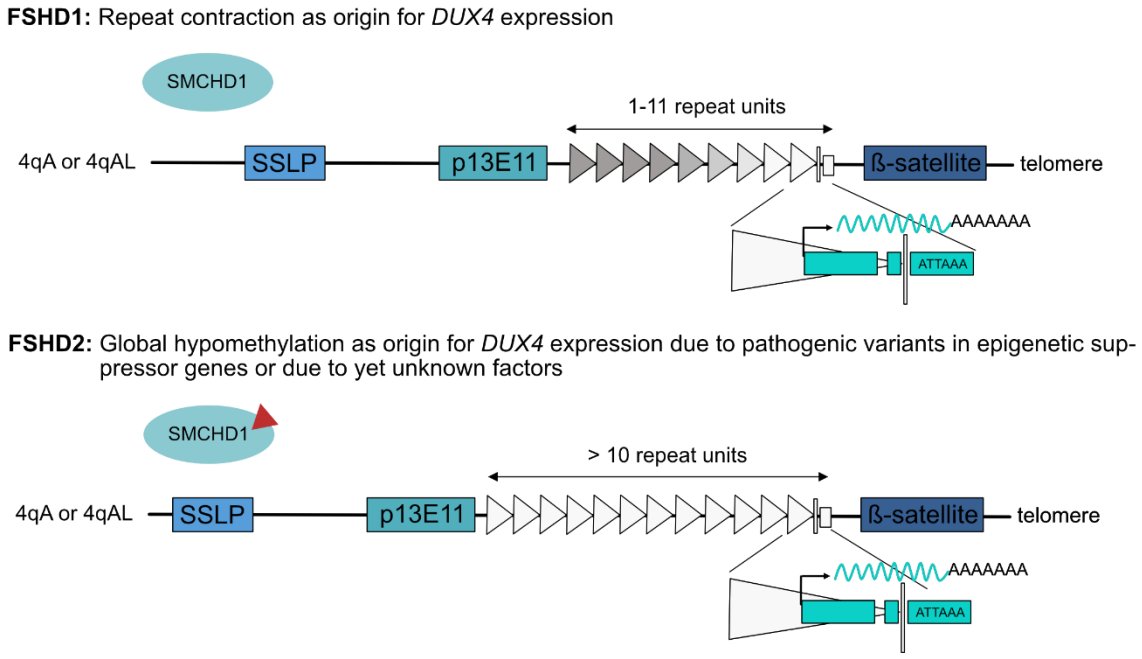

Figure 6. Genetic characteristics of FSHD1 (top) and FSHD2 patients (bottom). 5'-ATTAAA-3' present in exon 3 is the PAS defining the permissive haplotype. White triangles represent hypomethylated D4Z4 repeat arrays. Own illustration.

2.3.2 Diagnosis of FSHD

Diagnosis of FSHD is usually based on determining the genetic parameters associated with FSHD1 (repeat contraction of the D4Z4 repeat array) and FSHD2 (pathogenic variants in epigenetic suppressor genes such as *SMCHD1*) in combination with a haplotype analysis. [76–79] Determining the repeat length is traditionally performed by Southern blotting.[80] Here, high-molecular-weight (HWM) DNA is digested with the combination of different restriction enzymes (*EcoRI*, *XapI* and *BlnI*) to isolate the D4Z4 repeat array and align it to chromosome 4 or 10 after gel electrophoresis, blotting and visualization using a radioactively labeled p13E11 probe (Figure 7).[79,81]

Additional methods for determining repeat size and haplotype at the same time are molecular combing and single-molecule optical mapping.[82–84] Especially the first method allows for deciphering complex rearrangements that might escape Southern blotting.[85] Analysis of epigenetic suppressor genes is usually performed by next-generation sequencing together with genes causing overlapping clinical phenotypes. For determining the haplotype, different assays exist. Most commonly, the assay of Tsumagari *et al.* is used that identifies the 4q161 haplotype as the most prevalent permissive haplotype based on a few single nucleotide polymorphisms within the FSHD locus.[86]

Diagnosis of FSHD based on genetic parameters is limited in precision. Repeat contractions and a permissive haplotype might not be penetrant as they are found in 1-2% of the general population, which do not show symptoms of FSHD.[87,88] Similarly, pathogenic variants in *SMCHD1* and other epigenetic suppressor genes are not fully penetrant. Especially large D4Z4 repeat arrays are likely to prevent derepression of *DUX4* expression.[89] Additionally, some FSHD2 patients might be missed because global hypomethylation of the D4Z4 array has other causes than pathogenic variants in known epigenetic suppressor genes.[76]

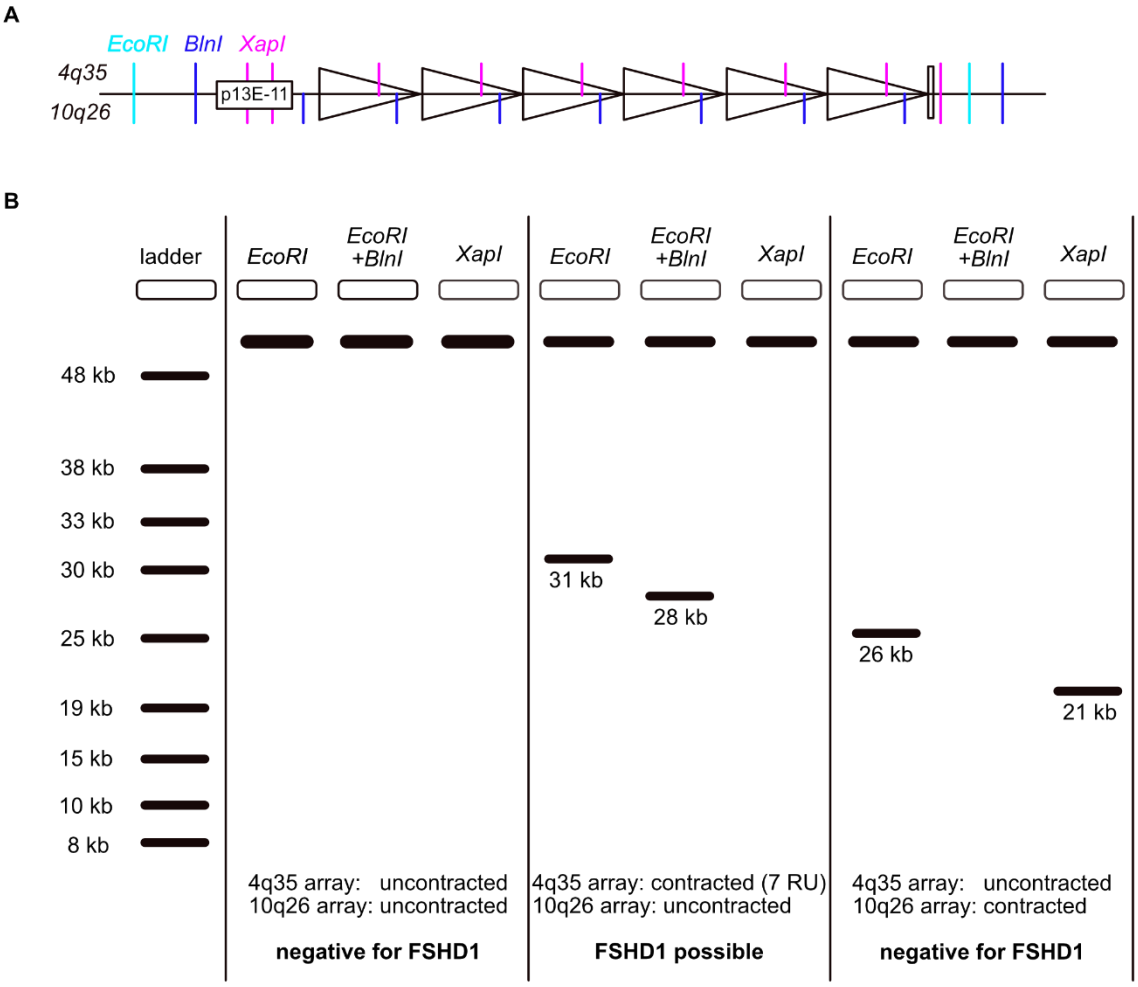

Figure 7. (A) D4Z4 repeat arrays on chromosome 4q35 and 10q26 and their restriction sides. (B) Examples of Southern blot results in FSHD testing. (Left) Uncontracted 4q35 and 10q26 arrays giving only fragments larger than 48 kb in each digest. (Middle) Repeat contraction of one 4q35 D4Z4 repeat array giving a fragment of 31 kb when digested with *EcoRI* that is further reduced in size by 3 kb after co-digest with *EcoRI* and *BlnI* and not visible after digest with *XapI*. Result is compatible with FSHD1 when the contraction is *in cis* to a permissive haplotype. (Right) Repeat contraction of one 10q26 D4Z4 repeat array giving a fragment of 26 kb when digested with *EcoRI* that is reduced in size by 5 kb when digested with *XapI* and not detectable after co-digest with *EcoRI* and *BlnI*. The individual is negative for FSHD1 as the D4Z4 repeat array on chromosome 4q35 is uncontracted. Own illustration.

Although it is consensus that FSHD is caused by epigenetic changes leading to the expression of *DUX4*, it is under debate whether, and if so to what extent, methylation profiles represent these changes.[39,90,91] It is discussed whether methylation represents disease status and directly correlates with *DUX4*

expression or whether changes are unspecific in FSHD patients. The evaluation of methylation is complicated by contrary results from a few epigenetic tests that utilize different techniques and analyze varying regions of the FSHD locus. Especially the amplification of homologous regions not originating from the 4q35 macrosatellite array can falsify results.[38] Several epigenetic assays have been developed for FSHD testing, but still need to be established in diagnostics.[35–37]

Technically, methylation patterns can be assayed by bisulfite sequencing reactions (Figure 8).[92,93] The method is based on hydrolytic deamination of unmethylated cytosines to uracil catalyzed by bisulfite. Because of the higher electrophilicity of unmethylated cytosines, the reaction occurs almost exclusively for unmethylated and not for methylated cytosines. After PCR amplification of the converted fragments, uracil will be replaced for thymine. Comparison of the initial sequence with the sequence after bisulfite conversion and PCR amplification allows to determine the methylation state of the cytosines in the native DNA.

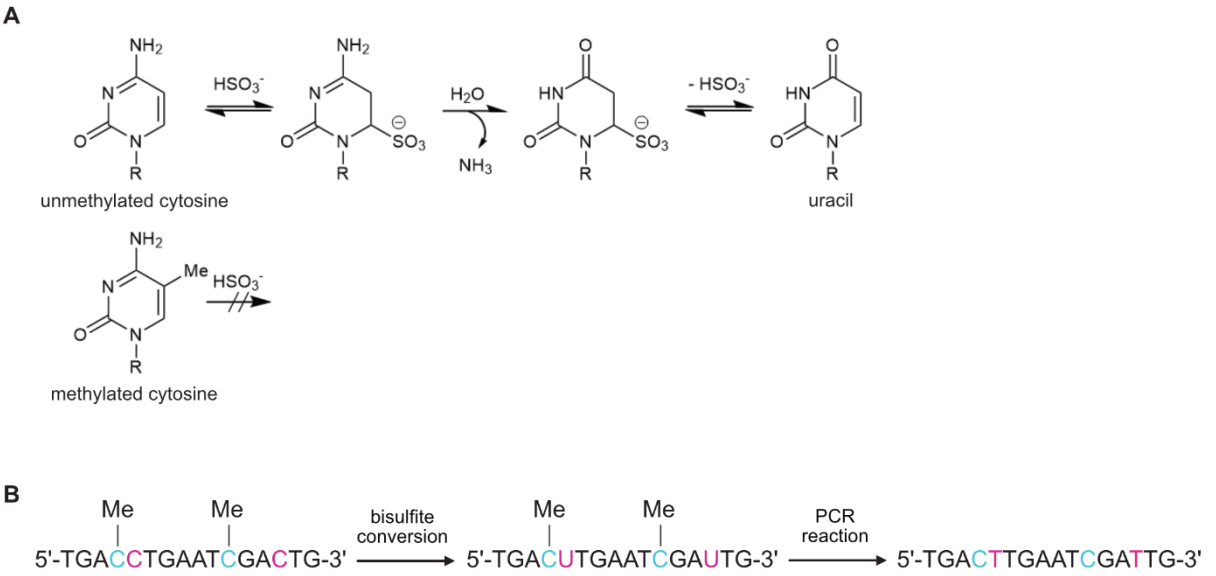

Figure 8. (A) Mechanism of the bisulfite catalyzed hydrolytic deamination of unmethylated cytosine to give uracil: (Top) The reaction is initiated by reversible nucleophilic addition of hydrogen sulfite to cytosine leading to a dearomatized sulphonate intermediate. After addition of water, ammonia is irreversibly eliminated and gives the sulfonated uracil. Rearomatization by elimination and regeneration of hydrogen sulfite gives uracil. (Bottom) Due to the decreased electrophilicity of methylated cytosine, addition of hydrogen sulfite is not favored preventing deamination. (B) Sequence of methylation analysis by bisulfite conversion. Denatured fragments of interest are treated with bisulfite to convert unmethylated cytosines into uracils that are replaced by thymines after PCR amplification. Own illustration.

2.3.3 Second aim of this thesis – Diagnosis and phenotype-genotype studies of FSHD patients

A comprehensive evaluation of epigenetic methods for the diagnosis of FSHD is urgently needed, to gain further insight into the FSHD pathomechanism, establish new biomarkers for the disease, and develop new therapeutics. Therefore, the second aim of this thesis is to evaluate methylation profiles of the 4q35 D4Z4 repeat array first as a reliable qualitative biomarker for diagnosing FSHD and second as a quantitative parameter representing disease severity. Based on the implemented methylation profile analysis (FSHD-MPA), the results of this epigenetic test are compared with genetic parameters from Southern blotting and NGS sequencing and discussed in the context of the patients' phenotypes. Additionally, the clinical severity of affected patients will be assessed and scored to correlate it with D4Z4 repeat length and methylation level within the most distal repeat unit, respectively. Consequences for the genetic testing in patients with FSHD phenotype will be derived.

Implementation and validation of the diagnostic method (previous work) and subsequent revalidation of the method and phenotype-genotype studies (subject of this work) will be summarized in a joined publication (Publication 2**:** Methylation of the 4q35 D4Z4 repeat defines disease status in facioscapulohumeral muscular dystrophy).[2] In addition, potential advancements in FSHD diagnostics through ONT long-read sequencing will be explored.[3]

3 Zusammenfassung

Die Analyse und Sequenzierung von komplexen genomischen Regionen, die aus repetitiven Elementen aufgebaut sind, stellt eine Herausforderung in der klinischen Genetik dar. Die molekulare Diagnose entsprechender Erkrankungen wie Repeatexpansionserkrankungen oder der Fazioskapulohumeralen Muskeldystrophie (FSHD) ist daher aufgrund technischer Limitationen der bisher angewandten Methoden eingeschränkt. Weiterhin ist der Zusammenhang zwischen Genotyp und Phänotyp für viele Erkrankungen bisher nicht vollständig verstanden.

Im Rahmen der vorgelegten Doktorarbeit wird zum einen (1) eine neue diagnostische Methode zur parallelen Erfassung von Repeat-Expansionserkrankungen implementiert und validiert: Eine long-read Sequenziermethode zur gezielten und parallelen Repeatanalyse (clinical nanopore Cas9-targeted sequencing, Clin-CATS) von Patienten mit im Erwachsenenalter manifestierender Ataxie. Zum anderen (2) wird im Rahmen der Arbeit die Relevanz von Methylierungsprofilen in der Diagnostik und klinischen Bewertung der FSHD untersucht.

(1) Für eine umfassende Repeatanalyse von Patienten mit einer Ataxie des Erwachsenenalters wurde Clin-CATS entwickelt, das die zehn in Deutschland am häufigsten ursächlichen Repeaterkrankungen (Stand: Zeitpunkt der Publikation des zugehörigen Papers[1]) erfasst: Die spinocerebellären Ataxien (SCA) 1-3, 6-8, 17, *RFC1*-Spektrumserkrankung, Friedreich-Ataxie (FRDA) und Fragiles-X-assoziiertes Tremor/Ataxie-Syndrom (FXTAS). Assoziierte Repeatregionen werden mit CRIPSR/Cas9 angereichert und anschließend mit Oxford Nanopore Technology long-read Sequenzierung sequenziert. Aus den Sequenzdaten werden die Repeatlängen, die Repeatsequenzen zur Bestimmung von Repeatunterbrechungen und der Zusammensetzung des *RFC1* Repeatarrays, sowie die *FMR1*-Promotor-Methylierung abgeleitet. Der Vergleich der mittels Clin-CATS bestimmten Repeatlängen zeigt eine hohe Übereinstimmung mit denen der konventionellen PCR-basierten Repeatanalysen. Pathogene Repeatexpansionen werden zuverlässig erkannt. Weitere Parameter, die im Rahmen der Analyse bestimmt werden, verbessern zusätzlich die diagnostische Präzision. Die Analyse von 100 Patienten mit einer im Erwachsenenalter manifestierenden Ataxie mittels Clin-CATS identifizierte ursächliche Repeatexpansionen bei 28 Patienten. Darunter sind seltene Erkrankungen wie eine sehr spät manifestierende FRDA oder ein männlicher FXTAS Patient, der trotz eines vollständig expandierten *FMR1*-Repeatarrays passend zum Phänotypen einen nicht-methylierten *FMR1*-

Promotor aufwies. Clin-CATs verdeutlicht die hohe Variabilität des *RFC1*-Repeatarrays und zeigt, dass die *RFC1*-Spektrumserkrankung in Deutschland eine häufige Ursache für erbliche Ataxien des Erwachsenenalters ist.

(2) Als Voraussetzung für die Implementierung von FSHD-MPA als präzise diagnostische Methode zum Nachweis von FSHD basierend auf epigenetischen Parametern, wurde zunächst gezeigt, dass sich gesunde Personen, sowie FSHD1- und FSHD2-Patienten in den Methylierungsmustern des D4Z4-Repeatarrays auf Chromosom 4q35 signifikant unterscheiden. Mit Hilfe der Bisulfit-Konvertierung wird mittels FSHD-MPA der Methylierungsgrad einer Region innerhalb des distalen D4Z4-Repeat-Arrays von permissiven 4q35-Allelen, sowie der durchschnittliche Methylierungsgrad einer zweiten Region, die in jedem D4Z4-Repeat auf Chromosom 4q35 vorhanden ist, bestimmt. Es wird gezeigt, dass der gesamte Repeatarray gesunder Personen hypermethyliert ist, während bei FSHD1-Patienten lediglich der distale Repeat hypomethyliert ist und FSHD2-Patienten eine globale und distale Hypomethylierung des Repeatarrays aufweisen. In einer Kohorte von 148 Patienten mit einem klinischen FSHD Phänotyp oder einer positiven Familienanamnese für FSHD verdeutlicht der Vergleich epigenetischer und genetischer Parameter, dass Methylierungsprofile präzise diagnostische Parameter darstellen. Darüber hinaus zeigen FSHD1- und FSHD2-Patienten eine epigenetische Überschneidung, erkennbar an einigen Patienten mit globaler und distaler Hypomethylierung, die keine pathogenen Varianten in bekannten epigenetischen Suppressorgenen, aber eine Repeatkontraktion aufweisen. Methylierungsprofile ermöglichen die Einschätzung der Penetranz genetischer Parameter, weshalb sie als prädiktive Marker fungieren können. Der Methylierungsgrad innerhalb des distalen Repeatarrays korreliert stark mit der alterskorrigierten klinischen Schwere und weist einen stärkeren Zusammenhang mit dieser auf als es die Repeatlänge tut. Somit ist die distale Methylierung in der vorliegenden Studie ein präziserer und universellerer Biomarker für den Schweregrad der Erkrankung. Der Krankheitsstatus der FSHD wird somit besser durch epigenetische als durch genetische Parameter repräsentiert. Repeatkontraktionen und pathogene Varianten in epigenetischen Suppressorgenen sind eher als Risikofaktoren der Krankheit anzusehen, als als direkte Krankheitsursachen. Zur Weiterentwicklung der aktuellen FSHD Diagnostik mittels FSHD-MPA wurden erste Analysen des D4Z4 Repeatarrays mittels ONT long-read Sequenzierung durchgeführt.[3] Die Ergebnisse zeigen, dass aktuelle Limitierungen der FSHD Diagnostik mittels ONT Sequenzierung überwunden werden können, da die Methode die Bestimmung aller relevanter Parameter (Methylierungsprofil, Haplotyp,

Repeatlänge) spezifisch für jedes Allel innerhalb einer Analyse ermöglicht. Zusätzlich ist das Methylierungsprofil des gesamten FSHD Locus und nicht nur das Methylierungslevel spezifisch amplifizierbarer Bereiche zugänglich. Zusammen mit anderen Arbeiten verdeutlichen diese Ergebnisse, dass die Epigenetik des FSHD Locus die fehlende Verknüpfung zwischen Phänotyp und genetischen Merkmalen darstellt.[94–96]

Zusammenfassend zeigen beide Projekte, dass neue diagnostische Methoden ein Schlüssel sind, um die große Komplexität von Erkrankungen, die in Zusammenhang mit der Veränderung von repetitiven Elementen unseres Genoms stehen, präziser zu erfassen. Dies ist insbesondere mit Hinblick auf eine Vielzahl von bisher nicht bekannten Repeat-assoziierten Erkrankungen oder Risikofaktoren in unserem Genom von hoher Relevanz.

4 Abstract

The analysis of repetitive elements in the human genome remains a challenge in clinical genetics. As next-generation sequencing is limited in analyzing repeat disorders and complex regions within the human genome, specific diagnostic methods are required. This thesis describes (1) the implementation and validation of a long-read sequencing method for parallel repeat analysis of patients with adult-onset ataxia as well as (2) the analysis of the relevance of methylation profiles in the diagnosis and clinical evaluation of FSHD.

For a comprehensive repeat analysis of patients with adult-onset ataxia, clinical nanopore Cas9-targeted sequencing (Clin-CATS) was designed to cover the ten repeat disorders most frequently causing adult-onset ataxia in Germany (status when publishing manuscript 1): spinocerebellar ataxias (SCA) 1–3, 6–8, 17, *RFC1* spectrum disorder, Friedreich's ataxia (FRDA) and fragile-X-associated tremor/ataxia syndrome (FXTAS). Associated repeat loci are enriched using CRIPSR/Cas9 and subsequently sequenced using Oxford Nanopore Technology long-read sequencing. Sequencing data are used to derive repeat length, repeat sequence to identify repeat interruptions and the repeat composition of the *RFC1* repeat array, as well as *FMR1* promoter methylation. Repeat lengths obtained by Clin-CATS show a high concordance to those determined by conventional PCR-based repeat analysis. Pathogenic repeat expansions were reliably detected and the comprehensive set of parameters determined improved diagnostic precision of Clin-CATS over conventional repeat testing. The analysis of 100 patients with an adult-onset ataxia phenotype by Clin-CATS revealed causative repeat expansions in 28 patients, including rare conditions such as a very-late onset FRDA or a high-function FXTAS male carrying a non-methylated *FMR1* promotor despite a fully expanded *FMR1* repeat array. Clin-CATs highlights the high polymorphism of the *RFC1* repeat array and reveals *RFC1* spectrum disorder to be a frequent cause of hereditary adult-onset ataxia in Germany.

After verifying FSHD1 and FSHD2 patients as well as healthy individuals to significantly differ in the methylation patterns of their D4Z4 repeat arrays on chromosome 4q35, FSHD-MPA was established as a diagnostic method for diagnosing FSHD. Utilizing bisulfite conversion FSHD-MPA determines the methylation level of a region within the most distal D4Z4 repeat array of 4q35 alleles carrying the permissive haplotype (4qA or 4qAL, distal methylation) and the average methylation level of a second region present within each D4Z4 repeat unit of chromosome 4q35 (global methylation). Healthy individuals show global and distal hypermethylation, while FSHD1 patients show isolated distal

hypomethylation and FSHD2 patients global and distal hypomethylation. Within a cohort of 148 patients with a clinical phenotype of FSHD or a positive family history of FSHD, methylation profiles are proven as precise diagnostic parameters for diagnosing FSHD by comparing the results from our epigenetic test with the results of Southern blotting and NGS sequencing of the epigenetic suppressor genes *SMCHD1*, *DMNT3B* and *LRIF1* as well as the clinical phenotype. Furthermore FSHD1 and FSHD2 patients show an epigenetic overlap as some patients with global and distal hypomethylation have repeat contractions in the absence of pathogenic variants in known epigenetic suppressor genes. Methylation profiles allow to access the penetrance of genetic parameters indicating their potential in predictive testing. Distal methylation level and age-corrected clinical severity show high correlation level that are stronger than those of repeat length and age-corrected clinical severity in the cohort studied. As such distal methylation is a more precise and universal biomarker for disease severity in the present study accounting for FSHD1 as well as for FSHD2. Thus, the disease status of FSHD is better represented by epigenetic than by genetic parameters. Repeat contractions and pathogenic variants in epigenetic suppressor genes should be considered more as risk factors of the disease than as direct causes of the disease. Further refinements of FSHD diagnostics can be achieved by ONT long-read sequencing which yields all relevant diagnostic parameters within one analysis and specific for each allele including the methylation profile of the whole D4Z4 repeat locus.

5 Paper I: Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing

Hannes Erdmann, Florian Schöberl, Mădălina Giurgiu, Rafaela Magalhaes Leal Silva, Veronika Scholz, Florentine Scharf, Martin Wendlandt, Stephanie Kleinle, Marcus Deschauer, Georg Nübling, Wolfgang Heide, Sait Seymen Babacan, Christine Schneider, Teresa Neuhann, Katrin Hahn, Benedikt Schoser, Elke Holinski-Feder, Dieter A. Wolf, Angela Abicht. *Brain.* 2023;146(5):1831–1843.

DOI:

10.1093/brain/awac377

Impact factor of Brain (2021):

15.255

Rank by Journal Citation Indicator Category Clinical Neurology (2021):

6/267 (JCI Percentile 97.94%)

https://doi.org/10.1093/brain/awac377 BRAIN 2023: 146; 1831–1843 | 1831

BRAIN
ORIGINAL ARTICLE

Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing

Hannes Erdmann,[1,2] Florian Schöberl,[3] Mădălina Giurgiu,[1,4] Rafaela Magalhaes Leal Silva,[1] Veronika Scholz,[1] Florentine Scharf,[1] Martin Wendlandt,[1] Stephanie Kleinle,[1] Marcus Deschauer,[5] Georg Nübling,[3] Wolfgang Heide,[6] Sait Seymen Babacan,[7] Christine Schneider,[8] Teresa Neuhann,[1] Katrin Hahn,[9] Benedikt Schoser,[2] Elke Holinski-Feder,[1,10] Dieter A. Wolf[1,11] and Angela Abicht[1,2]

Instability of simple DNA repeats has been known as a common cause of hereditary ataxias for over 20 years. Routine genetic diagnostics of these phenotypically similar diseases still rely on an iterative workflow for quantification of repeat units by PCR-based methods of limited precision.

We established and validated clinical nanopore Cas9-targeted sequencing, an amplification-free method for simultaneous analysis of 10 repeat loci associated with clinically overlapping hereditary ataxias. The method combines target enrichment by CRISPR–Cas9, Oxford Nanopore long-read sequencing and a bioinformatics pipeline using the tools STRique and Megalodon for parallel detection of length, sequence, methylation and composition of the repeat loci. Clinical nanopore Cas9-targeted sequencing allowed for the precise and parallel analysis of 10 repeat loci associated with adult-onset ataxia and revealed additional parameter such as *FMR1* promotor methylation and repeat sequence required for diagnosis at the same time. Using clinical nanopore Cas9-targeted sequencing we analysed 100 clinical samples of undiagnosed ataxia patients and identified causative repeat expansions in 28 patients. Parallel repeat analysis enabled a molecular diagnosis of ataxias independent of preconceptions on the basis of clinical presentation. Biallelic expansions within *RFC1* were identified as the most frequent cause of ataxia. We characterized the *RFC1* repeat composition of all patients and identified a novel repeat motif, AGGGG.

Our results highlight the power of clinical nanopore Cas9-targeted sequencing as a readily expandable workflow for the in-depth analysis and diagnosis of phenotypically overlapping repeat expansion disorders.

1 Medical Genetics Center (MGZ), Munich, 80335, Germany
2 Friedrich-Baur-Institute, Department of Neurology, Klinikum der Universität, Ludwig-Maximilians-Universität, Munich, 80336, Germany
3 Department of Neurology, Klinikum der Universität, Ludwig-Maximilians-Universität, Munich, 81377, Germany
4 Experimental and Clinical Research Center of the Max Delbrück Center for Molecular Medicine and Charité-Universitätsmedizin Berlin, Berlin, 13125, Deutschland
5 Department of Neurology, Klinikum Rechts der Isar, Technical University of Munich, Munich, 81675, Germany
6 Department of Neurology, General Hospital Celle, Celle, 29223, Germany
7 Department of Neurology and Neurological Intensive Care, Klinikum Darmstadt, Darmstadt, 64283, Germany
8 Department of Neurology, University Medical Center Augsburg, Augsburg, 86156, Germany
9 Department of Neurology and Experimental Neurology, Charité Berlin, Berlin, 10117, Germany
10 Department of Medicine IV, Klinikum der Universität, Ludwig-Maximilians-Universität, Munich, 80336, Germany
11 Department of Medicine II, Klinikum rechts der Isar, Technical University Munich, Munich, 81675, Germany

Received March 31, 2022. Revised August 14, 2022. Accepted September 19, 2022. Advance access publication October 13, 2022

1832 | BRAIN 2023: 146; 1831–1843 H. Erdmann et al.

Correspondence to: Angela Abicht, Medical Genetics Center, Bayerstr. 3–5, 80335 Munich, Germany
E-mail: angela.abicht@mgz-muenchen.de

Keywords: repeat analysis; Oxford nanopore sequencing; hereditary ataxia; RFC1; spinocerebellar ataxia

Introduction

The efficiency of medical diagnostics relies on high-performance methods such as next-generation sequencing that enable the analysis of the genome for pathogenic alterations in a high-throughput and resource-efficient manner.[1] However, due to its small read size, next-generation sequencing has known limitations in analysing structural variants such as tandem repeats despite recent approaches aimed at estimating the repeat lengths using statistical models.[2–6] To date, >40 Mendelian disorders are known to be caused by tandem repeat expansions with many of them resulting in severe neurological impairment.[7,8] An important group among repeat disorders are the hereditary ataxias. These conditions comprise overlapping cerebellar phenotypes mainly manifesting by progressive stance and gait disorders, cerebellar ocular motor disturbances, speech difficulties as well as limb ataxia and dysdiadochokinesia. Additionally, other neurological symptoms such as ophtalmoparesis, neuropathy, extrapyramidal movement disorders as well as cognitive and behavioural changes might occur to various extents.[9,10] According to their mode of inheritance, hereditary ataxias are classified into autosomal dominant ataxias with spinocerebellar ataxias (SCA) 1, 2, 3 and 6 being the most common adult-onset subtypes, autosomal-recessive ataxias such as Friedreich ataxia (FRDA) and X-linked fragile-X-associated tremor/ataxia syndrome (FXTAS).[11,12]

Hereditary ataxias can be caused by the expansion of repeats in both coding and non-coding regions. Fully penetrant alleles vary in repeat size among these disorders and tend to be longer when occurring in non-coding regions. Examples for short tandem repeat (STR) expansions in coding regions are the polyglutamine disorders, SCA1, 2, 3, 6, 7 and 17, where exonic CAG repeat size may vary from 2 to usually <100 repeat units (Supplementary Table 1). In contrast, full penetrance alleles of tandem repeats in non-coding regions such as regulatory 5′ or 3′ UTRs or introns may reach between 100 to several thousand repeat units. Examples of non-coding repeat expansion disorders include SCA8, FRDA and FXTAS, which harbour 3′ UTR CTG·CAG, intronic GAA or 5′ UTR CGG repeats, respectively (Fig. 1A and Supplementary Table 1). Another example is the recently described repeat in the intronic region of the replication factor C1 (*RFC1*) gene. Biallelic expansions were identified as an autosomal-recessively inherited cause of the adult-onset ataxia syndrome, CANVAS (cerebellar ataxia, neuropathy vestibular areflexia syndrome),[13–16] marked, in its full presentation, by the clinical triad of cerebellar ataxia, sensory neuro(no)pathy and bilateral vestibular failure.[17–23] Clinical complete and incomplete presentations are summarized as *RFC1* spectrum disorder. Pathogenicity requires the presence of a specific alternative pentanucleotide repeat sequence and its expansion beyond a certain range.[24] To date, nine different *RFC1* repeat sequence motifs have been identified and classified as either benign, pathogenic or of unknown significance (Supplementary Table 2).[13,14,25–29]

Diagnosis of hereditary ataxia due to repeat expansions is routinely performed by determining the number of repeat units using Southern blot or PCR without capturing sequence information or base modifications.[10] Access to repeat sequence information, including interruptions and base modifications is crucial, however, for accurate prediction of penetrance, inheritance and clinical phenotype. Indeed, this type of information is essential for the diagnosis of complex pathogenic repeat expansions such as in *RFC1*-associated adult-onset ataxia[13,16] or for assessing the clinical relevance of expansions in *FMR1*.[13,16,30,31] As such, parallel, amplification-free and sequence aware analysis of the clinically most relevant hereditary ataxias related to tandem repeats would be highly desirable for an efficient and clinically valid diagnostic workflow.

Long-read DNA sequencing techniques such as provided by Oxford Nanopore Technologies (ONT) are opening up new opportunities for analysing repeat expansions.[2] CRISPR–Cas9-based target enrichment combined with ONT sequencing and bioinformatics identification of repeat sizes from raw sequencing signals ('squiggles') was recently shown to enable accurate sequencing and analysis of genomic regions of interest in the research setting.[32–34] Another approach for target enrichment in long-read sequencing is the 'ReadUntil' functionality of ONT devices, which allows the selective recognition and sequencing of regions of inteterest within genomic DNA libraries.[35,36] Its use for repeat analysis has been demonstrated, but average on-target coverage appears modest compared to CRISPR–Cas9 enrichment. As such, it is less viable for DNA samples in clinical genetic testing at the current point.[37]

In a clinical setting, a long-read sequencing workflow for parallel diagnosis of phenotypically similar repeat expansion diseases remains to be established. In the clinic, special challenges apply, including variable sample quality combined with high demands on accuracy, timeliness and cost effectiveness.

Here, we implemented clinical nanopore Cas9-targeted sequencing (Clin-CATS) as an efficient platform for parallel repeat expansion analysis of hereditary ataxias based on ONT long-read sequencing. Demonstrating the versatility of the approach, we applied Clin-CATS to the parallel diagnosis of 10 of the most common ataxias resulting from expansions and modifications of either short coding repeats (SCA1, SCA2, SCA3, SCA6, SCA7, SCA17) or long non-coding repeats (SCA8, FXTAS, FRDA and *RFC1* spectrum disorder) (Fig. 1A and Supplementary Table 1).[11,13,38,39] Using Clin-CATS we analysed 100 patients with symptoms of adult-onset ataxia and studied the prevalence, genetics and genotype–phenotype relationships of the different hereditary ataxias.

Materials and methods

Patient DNA samples and study approval

To implement and validate Clin-CATS, a cohort of 14 patients with known clinical status and repeat lengths determined by PCR was assembled (validation cohort, Supplementary Table 5). The cohort consisted of nine patients with normal-size alleles of various short coding or long non-coding repeat loci as well as of one patient each with confirmed repeat expansion in *ATXN1* causing SCA1, *ATXN3* causing SCA3, biallelic pathogenic repeat expansions in *RFC1* causing CANVAS, a premutation within *FMR1* causing FXTAS and a fully expanded *FMR1* allele causing Fragile-X Syndrome (FXS). Subsequently, a cohort of 100 patients with symptoms of

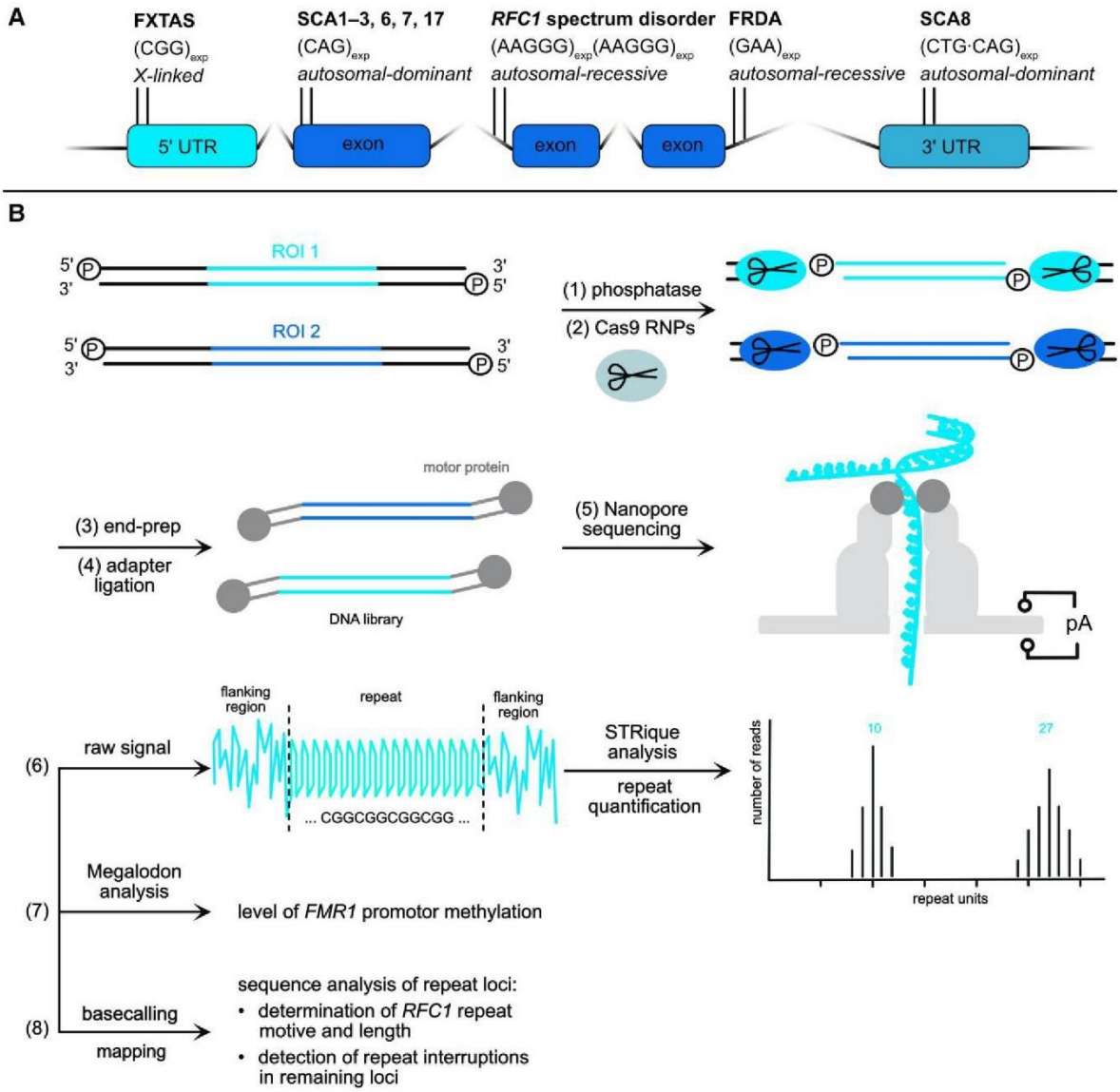

Figure 1 Method and repeat loci analysed in patients with adult-onset ataxia. (**A**) Schematic representation of the loci included in our parallel repeat analysis with their repeat motifs, their genomic regions and their mode of inheritance. (**B**) Clin-CATS approach for repeat analysis by long-read ONT sequencing and bioinformatics analysis. Library preparation steps 1–4: (1) dephosphorylation of DNA extracted from EDTA blood to exclude ligation and sequencing of off-target fragments. (2) Excision of repeat-containing regions of interest by CRISPR–Cas9 RNP particles with formation of 5′ phosphates for adapter ligation. (3) End-prep (dA-tailing of the 3′ end) of generated fragments with subsequent (4) adapter ligation to enable sequencing of targets. (5) Sequencing of the generated fragments by the ONT platform. (6) Repeat analysis using the bioinformatics tool STRIque based on the raw signal to give the repeat size of the alleles. (7) Quantification of the methylation of the *FMR1* promotor using the bioinformatics tool Megalodon. (8) Base calling and mapping of the sequenced reads to determine repeat motif and repeat length in *RFC1* as well as repeat interruptions in other loci from genome browser view.

adult-onset ataxia (diagnostic cohort, Table 2 and Supplementary Table 7) was analysed by Clin-CATS. All 114 patients were analysed as diagnostic samples within our institute. Informed consent to participate in this study was obtained from all patients and was approved by local institutions (Bayerische Landesärztekammer, 2019–210). All genetic analyses and investigations were performed in accordance with the guidelines of the Declaration of Helsinki.

Extraction of genomic DNA

Genomic DNA (gDNA) was obtained from total peripheral EDTA blood samples by extraction of white blood cells with a Biomek FX system (Beckman Coulter) using the NucleoMag® Blood 3 ml Kit (Machery-Nagel, no. 744502.1) as per the manufacturer's instructions. All DNA samples showed high purity as determined by optical density measurements (A260/A280 > 1.9 and A260/A230 > 2.0).

Library preparation and Oxford Nanopore Cas9-targeted sequencing

Library preparation was performed following the protocols of Gilpatrick *et al.*,[34] Giesselmann *et al.*[33] and the Oxford Nanopore Cas9-targeted sequencing protocol as described in the following. To enrich repeat-containing regions of interest associated with different forms of hereditary ataxia, guiding RNAs (gRNAs) were

composed by assembling a set of 22 synthetic CRISPR RNAs (crRNAs) and *trans*-activating crRNAs (tracrRNAs). crRNAs were individually designed using CHOPCHOP 8.[40] Sequences are given in Supplementary Table 3. Both, crRNAs and tracrRNAs, were purchased from IDT. An equimolar pool (100 µM) of crRNAs was prepared to enrich the regions of interest. For each target, one pair (two pairs for *RFC1*) of crRNAs up- and downstream to the regions of interest were used. Equimolar amounts (100 µM) of the crRNA pool and the tracrRNA (IDT) were mixed and diluted to 10 µM with nuclease-free duplex buffer (IDT). Incubation at 95°C for 5 min and subsequent cool down at room temperature for 10 min yielded the desired gRNA complex. Ribonucleoprotein (RNP) complex was formed by incubating a mixture of gRNA complex (10 µM), Hifi Cas9 Nuclease V3 (IDT, 64 µmol) and Cut Smart buffer (New England Biolabs) for 30 min at room temperature.

In parallel, 5 µg of input gDNA were dephosphorylated by incubation with Quick calf intestinal alkaline phosphatase (New England Biolabs) in 1× Cut Smart buffer (New England Biolabs) at 37°C for 10 min followed by enzyme deactivation at 80°C for 2 min. To the dephosphorylated DNA, the preformed RNP complex, dATP (New England Biolabs) and Taq polymerase (New England Biolabs) were added. The reaction mixture was incubated at 37°C for 15 min to enable the Cas9 reaction, before dA-tailing of the DNA ends was performed by incubation at 72°C for 5 min.

Adapter ligation was performed by combining ligation buffer (Oxford Nanopore SQK-LSK109 kit), nuclease-free water, NEBNext Quick T4 DNA ligase (New England Biolabs) and adapter mix (Oxford Nanopore SQK-LSK109 kit). The resulting mixture was incubated at room temperature for 10 min. DNA was purified by AMPure XP beads (Beckman Coulter, 0.3×) according to the manufacturer's protocol in TE buffer (IDT, pH 8.0) and eluted in elution buffer (Oxford Nanopore SQK-LSK109 kit). The library was mixed with sequencing buffer (Oxford Nanopore SQK-LSK109 kit) and loading beads (Oxford Nanopore SQK-LSK109 kit) before applying it on a primed flow cell (Oxford Nanopore FLO-MIN106D R9) for sequencing on the GridION X5 sequencer.

Analysis of Oxford Nanopore sequencing data

Base calling from electrical data was performed using Guppy (v.5.0.16).[41,42] The generated FASTQ files were aligned to the human reference genome (GRCh38/hg38) using Minimap2 (v.2.17)[43] to identify the reads spanning the targets of interest. For quality control of the aligned reads, we used NanoPlot (v.1.29.1).[44] The bioinformatics tool STRique (v.0.2.1)[33] was used to determine the number of repeat units for all reads assigned to the targets of interest. Repeat size distributions obtained by STRique were visualized as violin plots and used to determine the repeat size for each allele by computing local maxima using FindPeaks (v.2.1.1)[45] and visual assessment of the plots. For extended repeat expansions (>100 repeat units) showing multiple local maxima, a density function was used to determine the repeat size range representing the reads of the expanded allele. Only fragments spanning the entire repeat were considered for the repeat length quantification. For determining the repeat length of the *ATXN8OS/ATXN8* locus, the size of the non-pathogenic $(CTA{\cdot}TAG)_n$ and pathogenic $(CTG{\cdot}CAG)_n$ repeat were determined separately and summed up to be comparable with the PCR results. Due to its polymorphism and complexity, the *RFC1* locus was analysed manually by visual inspection of all reads assigned to that region in the IGV viewer. Repeat lengths distribution as well as average repeat sizes were extracted for each allele separately whenever possible. Methylation of cytosines was calculated from raw sequencing signals using Megalodon v.2.3.5[46] and Guppy v.5.0.16[42] within the two *FMR1* promotor regions according to the Eukaryotic Promotor Database EPD (https://epd.epfl.ch/index.php): chrX: 147911902–147911961 and chrX: 147912021–147912080. The percentage of methylation was calculated as average over all reads and cytosines.

Repeat length analysis by PCR-based methods is described in the Supplementary material.

Statistical analysis

Agreement of repeat sizes determined by PCR-based repeat analysis and Clin-CATS was determined using the method of Bland–Altman providing mean differences and 95% limits of agreement. Reference values for *FMR1* promotor methylation level of males and females with normal-size repeat lengths in the validation cohort were determined as average value with 95% confidence interval. Coverage of loci of interest within our validation cohort as well as repeat size distribution of alleles carrying the different motifs in *RFC1* of all patients within our validation cohort are represented as a box plot.

Data availability

Anonymized data from this study are available from the corresponding author on reasonable request.

Results

To establish Clin-CATS for the diagnosis of repeat expansion-associated hereditary ataxias being most common in Germany,[47] we assembled a panel of 10 repeat loci (Fig. 1A and Supplementary Table 1), including the SCA subtypes SCA1, SCA2, SCA3, SCA6, SCA7, SCA8, as well as SCA17, FXTAS, FRDA and *RFC1*-associated ataxia (*RFC1* spectrum disorder). The repeat regions of interest were enriched by *in vitro* CRISPR–Cas9 cleavage

Table 1 Regions of interest included in the analysis

Gene (disease)	Chromosomal position	Fragment size (wild-type) (kb)
ATXN1 (SCA1)	chr6: 16324749–16332104	7.355
ATXN2 (SCA2)	chr12: 111597404–111600286	2.882
ATXN3 (SCA3)	chr14: 92067918–92080268	12.350
CACNA1A (SCA6)	chr19: 13205681–13212882	7.201
ATXN7 (SCA7)	chr3: 63911775–63913479	1.704
ATXN8OS/ATXN8 (SCA8)	chr13: 70135508–70144322	8.814
TBP (SCA17)	chr6: 170556995–170568148	11.153
FMR1 (FXS/FXTAS)	chrX: 147909228–147915682	6.454
FXN (FRDA)	chr9: 69035458–69045114	9.656
RFC1 (CANVAS)	chr4: 39345760–39351593	5.833

Target regions of interest, chromosomal positions (reference genome GRch38/hg38.p11) and sizes of the fragments generated by CRISPR–Cas9.

Table 2 Patients with intermediate or pathogenic repeat expansions identified in the diagnostic cohort of 100 patients with ataxia symptoms

Patient	Result Clin-CATS	PCR repeat analysis	*RFC1* motif (allele 1)	Repeat units of *RFC1* (allele 1)	*RFC1* motif (allele 2)	Repeat units of *RFC1* (allele 2)	Age at analysis
3	SCA17: 36/52	35/51	AAAAG	11	AAGAG	45 (35–50)	70
4	*RFC1* spectrum disorder	–	AAGGG	680 (450–775)	AAGGG	790 (750–900)	74
5	*RFC1* spectrum disorder	–	AAGGG	610 (565–640)	ACAGG	645 (615–695)	65
6	SCA3: 22/68	23/67	AAAAG	11	AAAGG	630 (570–660)	69
8	SCA6: 11/21; carrier for *RFC1* spectrum disorder	11/22	AAAAG	105 (100–110)	AAGGG	430 (405–455)	74
10	Negative; carrier for *RFC1* spectrum disorder	–	AAAAG	110 (90–120)	AAGGG	305 (260–330)	84
14	FXTAS (high function male): 97–359; *FMR1* promotor methylation: 3.2%	>200	AAAAG	11	AAAAG	11	61
21	Negative; carrier for *RFC1* spectrum disorder	–	AAAAG	11	AAGGG	500 (440–525)	83
34	*RFC1* spectrum disorder	–	AAGGG	>600	AAGGG	>400	51
40	*RFC1* spectrum disorder	–	AAGGG	850 (810–915)	AAGGG	600 (545–640)	59
42	*RFC1* spectrum disorder	–	AAGGG	620 (565–670)	AAGGG	620 (565–670)	80
45	SCA2: 22/37	22/36	AAAAG	11	AAAGG	505 (465–540)	78
47	*RFC1* spectrum disorder	–	AAGGG	725	AAGGG	835	73
49	FXTAS: 90–105; *FMR1* promotor methylation: 0.1%	92	AAAGG/ AAAGGG	90 (75–105)	AAAAG	120 (105–125)	71
52	*RFC1* spectrum disorder	–	AAGGG	>270	AAGGG	>250	66
56	*RFC1* spectrum disorder; carrier for FRDA: 9/69	10/71	AAGGG	610	AAGGG	755	63
61	Intermediate *ATXN1* allele: 33/36 with two CAT interruptions	31/37	AAAAG	11	AAAAG	120 (100–130)	32
62	*RFC1* spectrum disorder	–	AAGGG	890 (845–925)	AAGGG	685 (575–750)	74
65	SCA3: 13/66	14/64	AAAAG	11	AAAAG	80 (70–90)	68
68	*RFC1* spectrum disorder	–	AAGGG	745 (640–815)	AAGGG	925 (850–990)	70
69	*RFC1* spectrum disorder	–	AAGGG	715	AAGGG	865	80
70	Intermediate *ATXN1* allele: 33/37 with two CAT interruptions	32/36	AAAAG	85 (70–90)	AGGGG	125 (105–135)	64
71	SCA8: 26/120–214	23/134	AAAAG	11	AAAGG/ AAAGGG	85 (75–90)	46
72	SCA8: 87–240	90/133	AAAAG	11	AAAAG	11	48
74	carrier for FRDA: 9/530–810	10/670	AAAAG	11	AAAAG	115 (95–125)	88
80	*RFC1* spectrum disorder	–	AAGGG	630 (555–700)	AAGGG	790 (750–815)	75
82	*RFC1* spectrum disorder; intermediate *ATXN8/ATXN8OS* allele: 29/59	29/60	AAGGG	715 (665–765)	AAGGG	915 (910–915)	63
85	SCA6: 12/22	12/22	AAAAG	11	AAAAG	11	76
88	*RFC1* spectrum disorder	–	AAGGG	720 (670–765)	AAGGG	815 (790–855)	59
90	SCA7: 10/43	9/47	AAAAG	11	AAAAG	120 (115–130)	37
91	*RFC1* spectrum disorder	–	AAGGG	750	AAGGG	815	71
94	Negative; carrier for *RFC1* spectrum disorder	–	AAAAG	11	AAGGG	420 (430–440)	22
95	*RFC1* spectrum disorder	–	AAGGG	700 (650–745)	AAGGG	860 (840–880)	82
98	FRDA: 145-315/790-1280	260/1070	AAAAG	100 (95–110)	AAAGGG/ AAGGG	55 (50–60)	53
100	Negative; carrier for *RFC1* spectrum disorder	–	AAAAG	11	AAGGG	680 (630–670)	70

Identified intermediate and pathogenic repeat expansions in all regions of interest, confirmatory PCR results for expanded alleles and identified *RFC1* motives on both alleles and their sizes are given. Results of all patients analysed are given in Supplementary Table 7.

of gDNA (Fig. 1B) using a set of 22 gRNAs to give fragments ranging in size from 1.7 to 12.3 kb (Table 1 and Supplementary Table 3).

Clin-CATS: establishment and validation

To validate Clin-CATS, we analysed 14 fully characterized individuals, including nine patients with normal-sized alleles, one patient with a confirmed expansion in *ATXN1*, one with an expansion in *ATXN3* and one with a biallelic AAGGG expansion in the *RFC1* gene causing *RFC1* spectrum disorder (CANVAS), as well as a patient affected by FXTAS and FXS, respectively.

After long-read sequencing, bioinformatic analysis revealed an average target coverage ranging from 51× to 330× (Fig. 2). *ATXN1*, *ATXN2*, *ATXN3*, *ATXN7*, *ATXN8OS/ATXN8*, *RFC1*, *FMR1*, *FXN* and

TBP loci showed excellent coverage with a mean sequencing depth greater or than 140×. *CACNA1A* showed a lower coverage resulting in a mean coverage of 51×. Within our validation cohort, we observed some degree of variability in the coverage of loci across samples. This was attributed to variable degrees of DNA fragmentation depending on the delay between blood sampling and DNA extraction (data not shown).

Based on the raw sequencing data, the STRique bioinformatics tool determined the repeat sizes of all reads assigned to the regions of interest, resulting in a distribution of repeat sizes for all repeat loci (Fig. 1B).[33] Allele sizes were defined as the local maxima of the obtained repeat size distributions. For sequence analysis and evaluation of the *RFC1* locus raw data were further processed and mapped to the reference genome after base calling.

To validate the performance of Clin-CATS, we benchmarked it against the current standard method for repeat length analysis, which is PCR-based fragment analysis. For each locus, the mean difference of both methods as well as the 95% limits of agreement using the Bland–Altman method (Fig. 3, Supplementary Table 6) were determined considering all normal-sized alleles and alleles with short repeat expansions (<100 repeat units). With both methods similar results were obtained for *ATXN1*, *ATXN2*, *ATXN3*, *CACNA1A*, *FXN*, *ATXN7* and *TBP* (mean difference $\leq \pm 0.5$ repeat units), while repeat sizes of *ATXN8OS/ATXN8* (mean difference 1.1 repeat units) and *FMR1* repeats (mean difference 1.4 repeat units) tended to be slightly overestimated by Clin-CATS (Fig. 3, central vertical lines). For *ATXN2*, *CACNA1A* and *FXN*, repeat sizes showed a small variance between the two test methods (95% limits of agreement $< \pm 2$ repeat units), while slightly higher variance of repeat sizes was observed for *ATXN1*, *ATXN3*, *ATXN7* and *TBP* (95% limits of agreement $< \pm 3$ repeat units) as well as for *ATXN8OS/ATXN8* and *FMR1*, for which we determined 95% limits of agreement of −1.0–3.2 repeat units and −2.0 −4.9 repeat units, respectively (Fig. 3, outer vertical lines). PCR-based repeat determinations are known to have limited precision with errors typically ranging from ±1 repeat unit for unexpanded alleles to ±3 units for short allele expansions (<100 repeat units).[48,49] Considering these limitations, the concordance between Clin-CATS and standard PCR measurements for normal-sized alleles is deemed excellent. Also, the pre- and full mutation *FMR1* alleles (Fig. 4A, Supplementary Table 5), as well as the expansion within the *ATXN1* and *ATXN3* locus were precisely detected.

In addition, Clin-CATS adds unique diagnostic information not provided by PCR testing:

(i) Length distribution: for large expansions (>100 repeat units) in *FMR1* (Fig. 4A and Supplementary Table 5), we obtained extensive length distributions that do not allow determination of local maxima and probably reflects actual biological heterogeneity of repeat sizes present in the patients' cells that escapes routine PCR analysis.

(ii) Repeat composition: by determining the composition of the *RFC1* locus, the repeat motif and repeat length can be derived simultaneously. Analysis of a patient with a full clinical presentation of CANVAS and a genetically confirmed biallelic AAGGG expansions in *RFC1* confirmed that Clin-CATS can accurately assess the composition of the *RFC1* locus and diagnose pathological alterations in this locus (Fig. 4B). Two alleles were detected with one carrying 950 and the other carrying 795 repeat units on average. As observed for other long repeat expansions, repeat sizes showed some degree of length-heterogeneity, here within the range of ~150 repeat units for both alleles.

(iii) Additionally, the availability of long-read sequencing information over the entire repeat range permits the detection of repeat interruptions to further assess penetrance and stability of repeat expansions.[50,51]

(iv) Methylation: *FMR1* promotor methylation analysis allows to further delineate the pathogenic potential of *FMR1* expansions. We quantified methylation of the *FMR1* promotor region upstream of the CGG tandem repeat from raw sequencing signals (Supplementary Table 8) using the tool Megalodon.[46] Full expansion of the CGG triplet (>200 repeat units) is known to cause *FMR1* promotor hypermethylation, resulting in loss of *FMR1* mRNA and protein thus causing FXS.[30] Premutation carriers (55–200 repeat units), predominantly males, may develop FXTAS in later adulthood.[30,31,52] In contrast to FXS, the *FMR1* promotor region is typically non-methylated in FXTAS, and *FMR1* mRNA levels are elevated thus leading to repeat-associated non-AUG (RAN) translation and the production of neurotoxic homopolymeric peptides.[53,54] Analysis of methylation levels confirmed unmethylated *FMR1* promotors in males with normal-size *FMR1* alleles [average 0.1% (95%–CI: 0.0–0.1%)] as well as a patient affected by FXTAS with an expanded allele in the premutation range (0.6%) and methylation of ~50% in females with normal-size *FMR1* alleles 43.2% (95%–CI: 42.0–44.5%) due to X-inactivation (Supplementary Table 8). In contrast, a male patient carrying an *FMR1* full expansion leading to FXS had a hypermethylated *FMR1* allele (69.4%).

(v) Differentiating the variable length of neighbouring repeat structures: by standard PCR methods, the potentially pathogenic $(CTG{\cdot}CAG)_n$ repeat present in *ATXN8OS/ATXN8* can only be quantified in combination with the proximal non-pathogenic $(CTA{\cdot}TAG)_n$ repeat, which is also variable in size. Clin-CATS can distinguish both repeats and therefore improves the assessment of pathogenicity and penetrance of repeat expansions in *ATXN8OS/ATXN8*.

Clin-CATS: analysis of 100 patients with adult-onset ataxia

Having validated Clin-CATS, we analysed 100 patients with adult-onset ataxia (Table 2 and Supplementary Table 7). For all patients, genetic testing for repeat expansions was requested by the attending physicians based on clinical suspicion or after exclusion of other causes of ataxia. In 28 patients (28%), Clin-CATS detected a pathogenic repeat expansion causative for the clinical symptoms (Fig. 5A). Expansions in *ATXN2*, *ATXN7*, *FXN* and *TBP* were detected in one patient each; expansions in *ATXN3*, *CACNA1A*, *ATXN8OS/ATXN8* and *FMR1* were diagnosed in two patients each; and 16 patients carried biallelic pathogenic expansions of the *RFC1* locus. Additionally, one patient affected by *RFC1*-spectrum disorder and one patient without causative findings within our panel were found to be additionally heterozygous for repeat expansions in FRDA. One patient affected by SCA6 and four patients without causative findings were heterozygous for the pathogenic $(AAGGG)_{exp}$ repeat composition in *RFC1*. In addition, Clin-CATS identified two patients with intermediately sized *ATXN1* alleles (36 and 37 repeat units), which were considered non-pathogenic but are likely to expand within the next generation if uninterrupted.[55] Assessment of the repeat sequence revealed two CAT interruptions within each intermediate allele, thus stabile transmission can be assumed. All repeat expansions and intermediate alleles were confirmed by PCR repeat size analysis.

Clin-CATS-based methylation analysis provided a valuable contribution in diagnostics: one male patient affected by FXTAS carried a fully expanded *FMR1* allele and showed clinical symptoms of cerebellar ataxia without intellectual disability. Methylation analysis of the *FMR1* promotor revealed a non-methylated allele (3.2%) that apparently explains the FXTAS phenotype (Table 2 and Supplementary Table 8) and identifies the patient as a high-functioning FXTAS male despite a fully expanded *FMR1* allele. A second male patient affected by FXTAS showed a premutated *FMR1* allele (90–105 repeat units) and a non-methylated *FMR1* promotor (0.1%). The Clin-CATS result of an unmethylated promotor in the high-functioning FXTAS male was confirmed by *FMR1* MLPA methylation analysis.

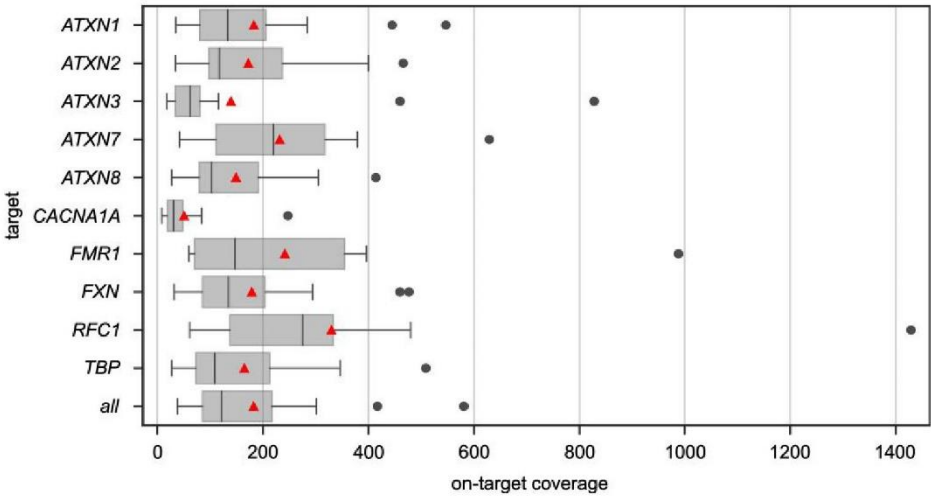

Figure 2 Coverage of the regions of interest. Box plot of the on-target coverage of each locus of interest and overall targets of the 14 samples of the validation cohort. Triangles symbolize mean coverage.

By analysis of *RFC1*'s repeat composition (repeat motif and repeat sizes), 16 of 100 ataxia patients were diagnosed with *RFC1* spectrum disorder (Table 2). Of 16 patients, 15 with biallelic pathogenic *RFC1* alleles showed homozygous alleles carrying AAGGG expansions. Exact repeat sizes could be determined in 14 patients ranging in size from ~600 to 925 repeat units. For large pathogenic expansions, coverage was often lower than for benign expansions. Thus, two patients (Table 2) with vastly expanded alleles carrying the pathogenic AAGGG motifs (>200 repeat units) were readily identified although the exact repeat length remained undetermined. For three patients with *RFC1* spectrum disorder (Table 2), we quantified the repeat expansion based on a few reads per allele without giving a length distribution. In one European patient with the common pathogenic $(AAGGG)_{exp}$ repeat composition on one allele, the expansion of the rare pathogenic ACAGG motif was detected with an average of 645 repeat units on the other allele, which was previously reported only in Asian-Pacific families.[25] We determined frequencies of the repeat motifs and number of repeat units of *RFC1* for all 100 patients analysed (Fig. 5B and C and Supplementary Table 7). As non-pathogenic repeat compositions we identified the configuration in the reference genome $(AAAAG)_{11}$ in 40.5% of alleles, $(AAAAG)_{exp}$ with expansion sizes between 30 and 160 repeat units (average of 105 repeat units) in 34.0% of alleles, and homogeneous $(AAAGG)_{exp}$ with expansion sizes from 245 to 630 repeat units (average of 475 repeat units) in 3.0% of alleles. The likely non-pathogenic repeat composition $(AAGAG)_{exp}$ was seen in two alleles (1.0%) with 45 repeat units each. As pathogenic repeat compositions we identified $(AAGGG)_{exp}$ ranging in size from 305 to 925 repeat units (average of 705 repeat units) in 18.0% of alleles and $(ACAGG)_{exp}$ with 645 repeat units in one patient (0.5%). Clin-CATS also allowed for identifying the yet unknown $(AGGGG)_{exp}$ repeat composition in the heterozygous state in one patient (0.5%) containing 125 repeat units. In addition, five alleles with heterogeneous repeat composition carrying two different repeat motifs in substantial amounts were identified: four alleles (2%) carried heterogeneous $(AAAGG/AAAGGG)_{exp}$ repeat compositions with expansion sizes from 85 to 90 repeat units (average 89 repeat units) and one allele carried the heterogeneous $(AAAGGG/AAGGG)_{exp}$ composition with 55 repeat units. To further characterize the phenotype of *RFC1* spectrum disorder, we analysed the clinical presentation of affected patients (Supplementary Table 9). All 16 patients with biallelic pathogenic expansion in *RFC1* showed sensory axonal polyneuropathy and cerebellar ataxia. Of the 16 patients, 12 were additionally affected by bilateral vestibulopathy, representing complete CANVAS. In addition, autonomic dysfunction was reported for four out of 15 patients for whom relevant information was available. Chronic irritable cough was a common symptom, occurring several years before other symptoms in nine out of 15 patients. Average age at disease manifestation was 59 years, ranging from 37 to 70 years. In patients closely monitored for signs of progression, the latency between onset of polyneuropathic impairment and ataxia was typically <3 years.

Discussion

Continued developments in sequencing technology and bioinformatics have been opening up opportunities for new procedures in genetic testing. Here, we adopted Nanopore Cas9 targeted long-read sequencing combined with a bioinformatics pipeline based on the tools STRique and Megalodon to implement Clin-CATS, a clinical grade diagnostics workflow for parallel analysis of 10 repeat loci associated with the most common causes of adult-onset ataxia in the European population (Fig. 1).

Although we observed some variability in coverage, all regions of interest were sufficiently covered to allow precise repeat analysis (Fig. 2). While differences in target coverage of individual loci may be due to varying on-target activity of individual CRIPSR–Cas9 RNP complexes, we have observed that DNA fragmentation contributes to the variability between samples. In preanalytics, it is therefore essential to minimize time to analysis and sample deterioration during shipping and storage. Throughout the study, blood collection with PAXgene® tubes was implemented, which was found to stabilize the DNA during shipping and storage. In particular, the fragmentation of pathogenic AAGGG repeat expansions in *RFC1* was prevented.

Clin-CATS enables parallel repeat length quantification of ataxia-related loci without elaborate sample preparation and is thus feasible in routine diagnostics. While the 'ReadUntil' functionality as an alternative approach for target enrichment enables

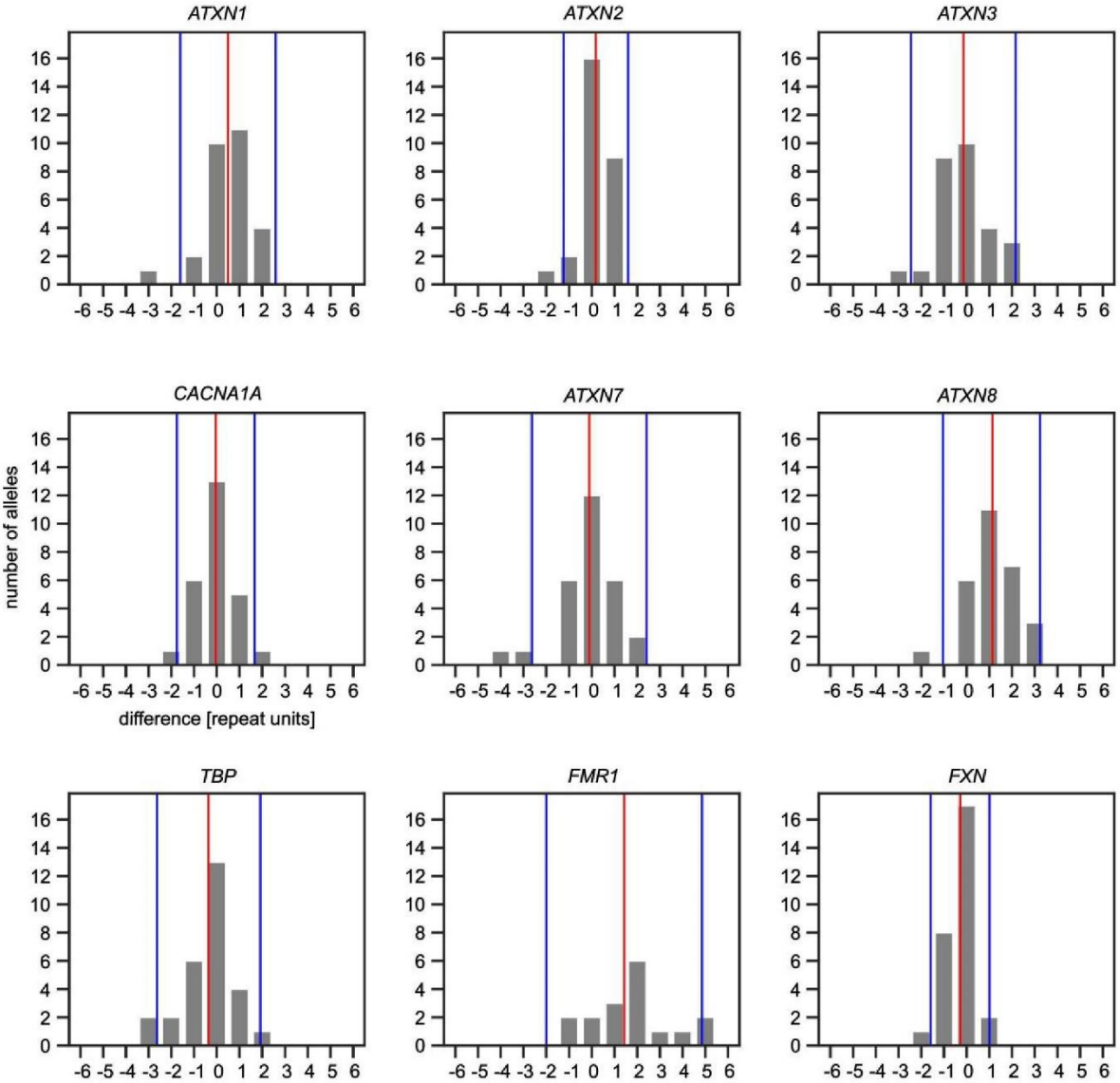

Figure 3 Precision of repeat size analysis. Concordance of repeat lengths of unexpanded alleles and short repeat expansions (<100 repeat units) determined by Clin-CATS and standard PCR-based repeat quantification. The *x*-axes represent the relative difference in repeat sizes. The *y*-axes show the number of alleles showing the respective difference. The central vertical line indicates the mean difference, while the outer vertical lines represent the 95% limit of agreement.

median on-target coverage of 14–23× for the loci included in our panel, it was 31–275× (average coverage 51–330×) with CRISPR–Cas9 enrichment.[37] In the clinical setting, this is especially important to compensate for varying DNA quality and to sufficiently cover long repeat expansions with a broad range of repeat lengths as observed for patients with FRDA, FXTAS or SCA8. Although 'ReadUntil' is less laborious in library preparation and more flexible in selecting regions of interest than CRISPR–Cas9-target enrichment, this level of flexibility is rarely required in routine diagnostics. Due to the high costs of ONT whole genome sequencing protocols, which can provide the high coverage required for repeat analysis, this approach is currently impractical in clinical diagnostics. PCR-based methods for tandem repeat analysis typically report repeat sizes with a variance ranging from ±1 repeat unit for unexpanded alleles to ±3 repeat units for short repeat expansions (<100 repeat units).[48, 49] Most loci analysed by our method showed similar precision (Fig. 3). Repeats with known pathogenic expansions in the validation cohort were successfully sized with high precision (Fig. 4A and Supplementary Table 5). The long expansions in *FMR1* were readily identified and found to show extensive diversity in length (Fig. 4A). This larger than expected heterogeneity probably reflects somatic instability of the expanded repeats that is not adequately detectable by PCR-based methods.[56–59] Our panel revealed a tendency of slightly overestimating the size of the *ATXN8OS*/*ATXN8* and *FMR1* repeat locus compared to PCR-based analysis. This might become a constraint in those rare instances when normal allele sizes are close to the pathogenic range for *ATXN8OS*/*ATXN8* or close to the premutation range (>54) for *FMR1*. In these cases and depending on the determined limits of agreement for all borderline repeat lengths, confirmatory analysis by PCR may become necessary assuming that PCR detects (validated by sanger sequencing and Southern blotting) the correct repeat sizes. To further

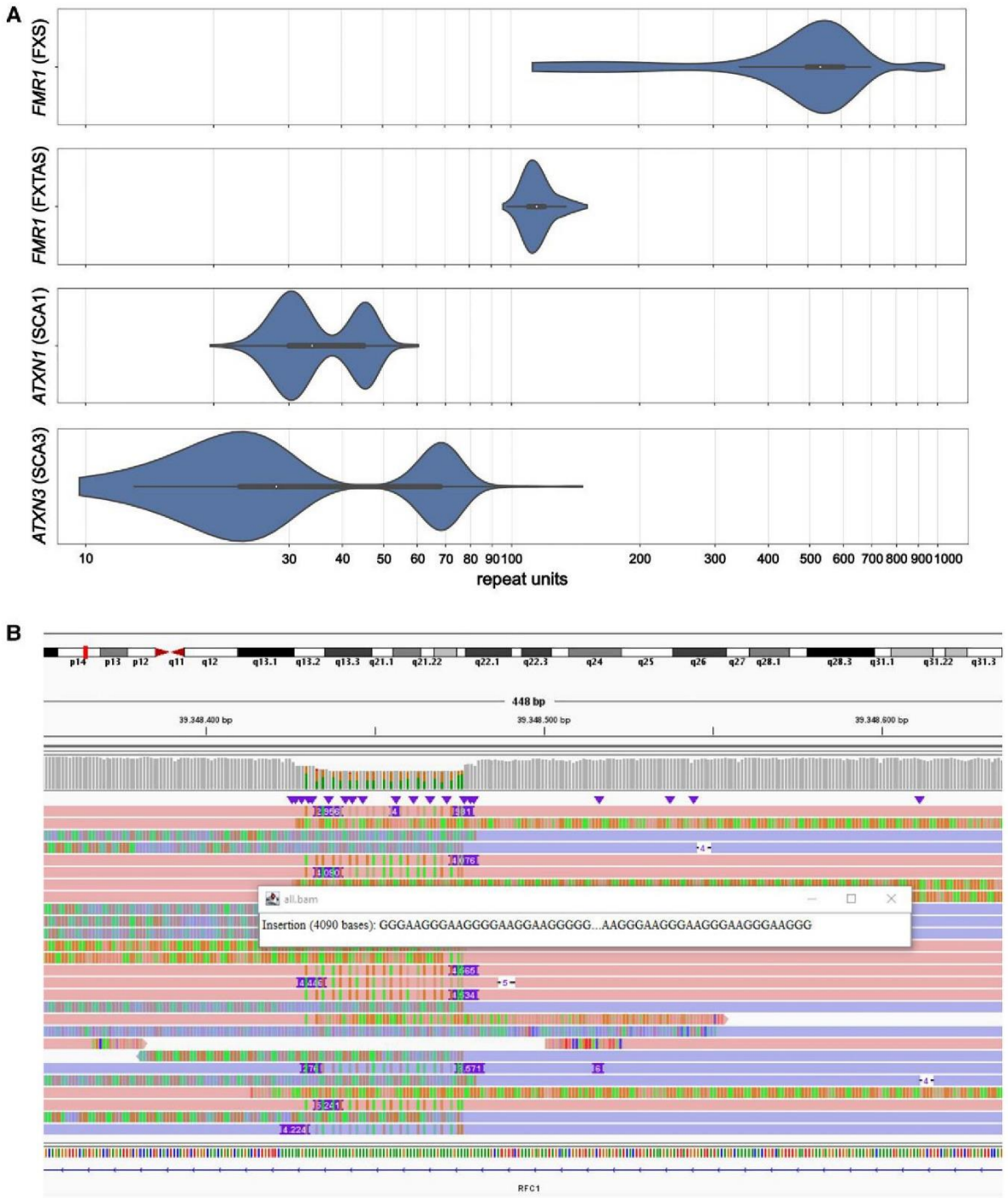

Figure 4 Evaluation of patients with known pathogenic repeat expansions. (**A**) Graphical representation of the repeat size analysis for patients of the validation cohort carrying pathogenic repeat expansions in *FMR1* causing FXS (first plot, male patient) and FXTAS (second plot, male patient) as well as in *ATXN1* and *ATXN3* causing SCA1 and SCA3 (third and fourth plot), respectively. The *x*-axis shows repeat lengths on a logarithmic scale. The *y*-axis represents frequency of the respective allele length. (**B**) Genome browser view showing reads assigned to the *RFC1* locus of a patient with a confirmed CANVAS caused by biallelic pathogenic expansions of the AAGGG motif.

improve Clin-CATS, especially for quantifying similarly sized alleles, we aimed to implement allelic phasing using the combination of the tools Pepper, DeepVariant and WhatsHap.[60–62] However, probably due to the absence of single nucleotide variants, phasing did not allow for the alignment of reads in most cases. To improve Clin-CATS in the future, we will work to implement a combination of advanced data analysis algorithms.[37]

After demonstrating its validity, we applied Clin-CATS for diagnostic testing of 100 patients with adult-onset ataxia. It allowed us to identify pathogenic repeat expansions causative for ataxia in 28

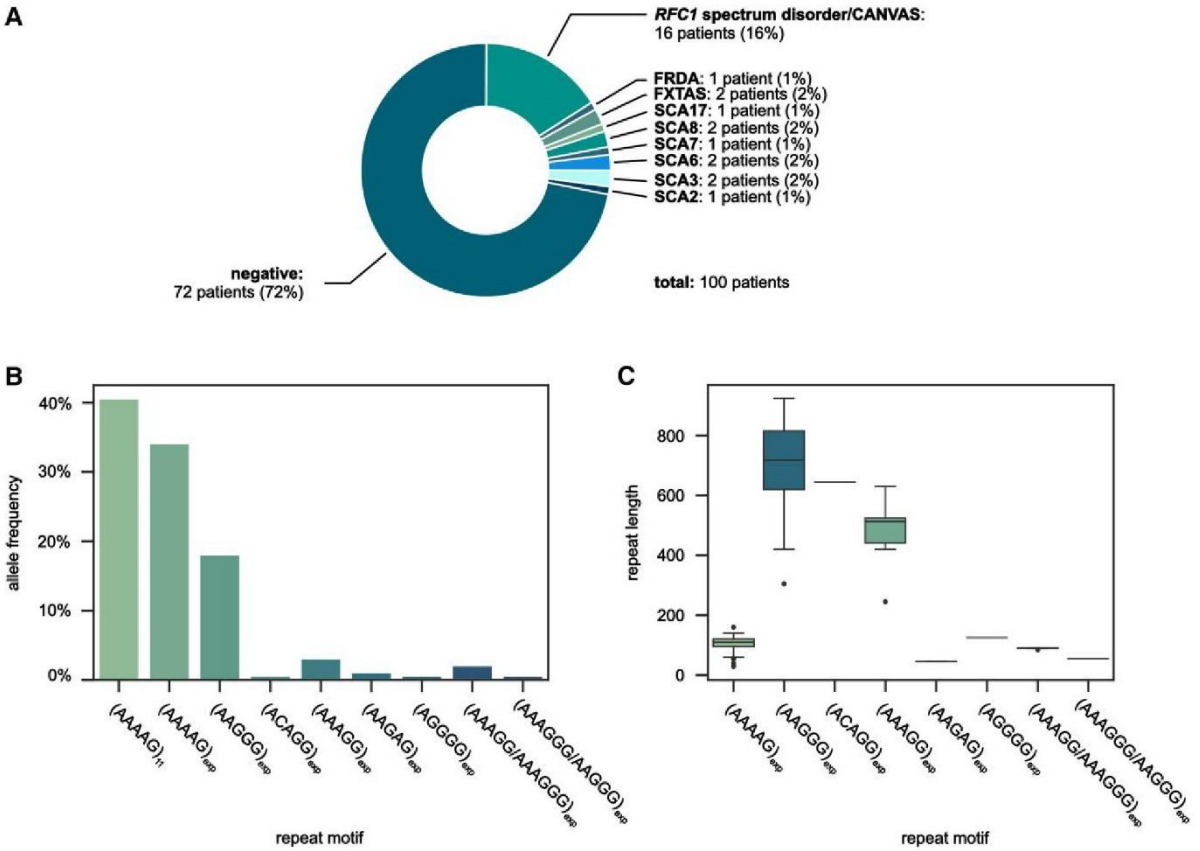

Figure 5 Diagnostic results and composition of the RFC1 locus of 100 ataxia patients. (**A**) Results of the diagnostic parallel repeat analysis of 100 patients with ataxia symptoms (diagnostic cohort). (**B**) Repeat sequence motifs of the *RFC1* locus identified in all patients of the diagnostic cohort and their frequency. (**C**) Repeat length distribution of the different repeat sequence motifs of the *RFC1* locus.

patients (28%) and carrier status for *FRDA* and *RFC1* in seven patients while all remaining loci were negative for pathogenic expansions. Thus, Clin-CATS can diagnose and simultaneously exclude multiple differential diagnoses in a single run, while also determining carrier status for each condition. In deriving epidemiologic information from this cohort, it must be kept in mind that patients were not preselected before analysis, for example, with respect to family history, symptoms other than ataxia in adulthood or previous analyses, so bias cannot be excluded.

Clin-CATS not only enabled the parallel quantification of different repeat loci, but also provided additional information that further improved the patients' diagnoses by differentiating the variable lengths of neighbouring repeat structures, such as in *ATXN8OS/ATXN8* where the variable proximal non-pathogenic $(CTA{\cdot}TAG)_n$ repeat can be excluded. Besides repeat length, we simultaneously determined cytosine methylation of the *FMR1* promotor. Both parameters are known to determine the manifestation of the different *FMR1*-associated phenotypes in males and are usually concordant.[30,31,52] In accordance with previous studies, FXS patients with full mutations showed a hypermethylated *FMR1* promotor in contrast to the patients affected by FXTAS with premutated *FMR1* allele who showed a non-methylated *FMR1* promotor like healthy males (Supplementary Table 8). Based on the combined *FMR1* repeat length and methylation analysis, we identified a rare case in which both parameters were discordant. One patient who presented with FXTAS but no intellectual disability carried a fully expanded but non-methylated *FMR1* allele. Only the simultaneous analysis of both parameters allowed us to assess the genetic findings and diagnose this patient as a so-called high-functioning male.[63] Routine implementation of Clin-CATS should allow us to determine whether this discrepancy of repeat length and methylation is more prevalent in ataxia patients.

Previous studies revealed repeat expansions in *RFC1* as a frequent cause of adult-onset ataxia known as CANVAS in its full presentation[13–15]; inclusion of this locus into the ataxia panel was therefore of high priority. Assessing the *RFC1* repeat composition usually requires a long-range PCR and Sanger sequencing or a repeat-primed PCR. Neither method can span the entire length of the expanded repeat and thus determine its size or potential structural variations, rendering Southern blotting the only option.[24,64] With Clin-CATS, *RFC1* repeat sequence and size were determined in synchrony with nine other tandem repeat loci in a single run.

In our study, the alleles of all patients affected by *RFC1* spectrum disorder showed large expansions usually exceeding 400 repeats. We also identified an allele carrying the AAGGG motif smaller in size (305 repeat units on average) heteroallelic with the non-pathogenic AAAAG motif. Further studies need to prove that the repeat motif rather than repeat size is crucial for pathogenicity. As such these studies need to reveal whether there is a certain threshold for pathogenicity of these AAGGG expansions and verify non-pathogenicity of large AAAGG expansions as they have also been observed in this and previous studies.[13,28] Interestingly, we

identified the rare pathogenic $(ACAGG)_{exp}$ composition with an expansion size of 645 repeat units heteroallelic with the common pathogenic $(AAGGG)_{exp}$ composition in one CANVAS patient of the European population. Before the present study, the size of an ACAGG repeat composition was determined in only a single Asian-Pacific family at ~1000 repeat units.[25] ONT sequencing of the *RFC1* locus allowed us to decipher the novel unknown AGGGG motif with a size of 125 repeat units in the heterozygous state. On the basis of the relatively short size of the expansion, pathogenicity is unlikely but this will require further studies.

As observed in another recent study by Stevanovski *et al.*,[37] we identified short heterogeneous repeat expansions (55–90 repeat units in size) comprised at least two different repeat motives. Four alleles contained AAAGG as the dominant motif interspersed with AAAGGG as a minor motif. Another patient had AAAGGG as the dominant motif intermixed with AAGGG repeat units. Interestingly, large AAAGG expansions (>245 repeat units) showed a homogeneous repeat pattern. These findings demonstrate the polymorphic nature of this region that requires careful analysis. Future studies will be required to evaluate the pathogenicity and origin of heterogeneous repeat motifs. Likewise, the effect of single repeat interruptions, which are not detected by our method, needs to be further investigated, especially in pathogenic $(AAGGG)_{exp}$ repeats, as they may explain phenotypic differences. Overall, we diagnosed pathogenic biallelic *RFC1* expansions in 16 of 100 patients (16%) with clinical symptoms of ataxia being by far the most common diagnosis. This high frequency highlights the relevance of *RFC1* repeat expansions as leading causes of adult-onset cerebellar ataxia. For large pathogenic $(AAGGG)_{exp}$ alleles, enrichment was often lower than for benign expansions in samples with similar average coverage over all enriched loci, indicating that the pathogenic $(AAGGG)_{exp}$ repeat composition may be fragile in preanalytics. This limitation affecting repeat size quantification in a few samples was overcome by collecting blood in tubes containing a DNA stabilizing reagent and by avoiding delays in sample shipment (unpublished observation).

As STRique was not designed to handle repeats as complex as in *RFC1*, the analysis of this locus relied on manual inspection of mapped reads. We are currently working on an automated bioinformatics workflow for the *RFC1* locus using recently developed approaches.[37,65] Overall, our study underscores the importance of genetic testing for the diagnosis of adult-onset ataxia syndromes, given the clinical overlap in phenotypes and the high prevalence of genetic causes even in the absence of a positive family history.[9,66] Parallel repeat analysis enabled a molecular diagnosis of ataxias independent of preconceptions based on clinical presentation or prevalence. Two patients, in whom *RFC1*-associated adult-onset ataxia was clinically considered due to a concomitant pure sensory axonal neuropathy, were diagnosed with SCA3 and SCA6. While sensory neuropathy is associated with SCA3, it is uncommon in patients with SCA6 and may have an independent aetiology in this patient.[67] Equally remarkable is the diagnosis of late-onset FRDA in a 53-year-old patient due to a comparatively short *FXN* expansion of 145–315 repeat units besides a large expansion of 790–1280.[68–70] Clinically, FRDA is often not considered in patients with late-onset ataxia because it commonly manifests before the age of 25. Additionally, characteristic phenotypic features might be missing and atypical features can be present in later manifesting FRDA.[69,71,72] The limited data on the prevalence of FRDA in a respective population add to this issue. Performing Clin-CATS over an extended time period will establish the genetic basis to determine the frequency of pathogenic expansions in *FXN* and *RFC1* from which the prevalence of FRDA and *RFC1* spectrum disease can be inferred in a respective population. These data will improve the clinical evaluation of ataxia patients, deciphering the phenotypical variability and facilitate genetic counselling for relatives of affected patients.

A comparison of the costs for repeat analysis by Clin-CATS and conventional PCR testing is difficult due to individual pricing of reagents, equipment and manpower. According to our estimates, reagent and equipment expenses are currently higher for long-read sequencing-based assays, while labour costs are lower due to reduced laboratory effort, with higher information content of Clin-CATS. ONT's current efforts to combine target enrichment using CRISPR–Cas9 with barcoding of individual samples may allow multiple samples to be run simultaneously on a FlowCell and further reduce costs. Our results demonstrate the power of single-reaction concurrent testing of ataxia-related loci to provide the most efficient and comprehensive analysis in a clinical setting.[73] Clin-CATS determines not only the length of 10 repeat loci in parallel, but also the diagnostically significant methylation of the *FMR1* promoter, repeat interruptions and the composition of the *RFC1* locus. The analysis is readily scalable to additional ataxia loci such as the rare and complex SCA31 and SCA37 where repeat composition determines pathogenicity, and will probably enable the discovery of new disease mechanisms as well as improve our understanding of genotype–phenotype correlations for repeat expansion disorders.

Acknowledgements

A. Benet-Pagès und B. Neitzel are thanked for valuable suggestions and discussion and J. Ehses for proofreading the manuscript and helpful suggestions. H.E. is grateful for a scholarship from the Medical Faculty of the University of Munich (LMU). This study is part of the medical thesis by H.E.

Funding

No funding was received towards this work.

Competing interests

All authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Ashley EA. Towards precision medicine. *Nat Rev Genet*. 2016;17:507-522.
2. Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics. *Front Genet*. 2019;10:426.
3. Dolzhenko E, Bennett MF, Richmond PA, *et al*. Expansion hunter de novo: A computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol*. 2020;21:102.
4. Bahlo M, Bennett MF, Degorski P, Tankard RM, Delatycki MB, Lockhart PJ. Recent advances in the detection of repeat expansions with short-read next-generation sequencing. *F1000Res*. 2018;7:736.

5. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res*. 2019;47:e90.
6. Rajan-Babu IS, Peng JJ, Chiu R, *et al*. Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions. *Genome Med*. 2021;13:126.
7. Paulson H. Repeat expansion diseases. In: Geschwind DH, Paulson HL, Klein C, eds. *Handbook of clinical neurology - Neurogenetics Part II*. Elsevier; 2018:105-123.
8. Chintalaphani SR, Pineda SS, Deveson IW, Kumar KR. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol Commun*. 2021;9:98.
9. Hadjivassiliou M, Martindale J, Shanmugarajah P, *et al*. Causes of progressive cerebellar ataxia: Prospective evaluation of 1500 patients. *J Neurol Neurosurg Psychiatry*. 2017;88:301-309.
10. Sandford E, Burmeister M. Genes and genetic testing in hereditary ataxias. *Genes (Basel)*. 2014;5:586-603.
11. Jayadev S, Bird TD. Hereditary ataxias: Overview. *Genet Med*. 2013;15:673-683.
12. Brusse E, Maat-Kievit J, Van Swieten J. Diagnosis and management of early- and late-onset cerebellar ataxia. *Clin Genet*. 2007;71:12-24.
13. Cortese A, Simone R, Sullivan R, *et al*. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat Genet*. 2019;51:649-658.
14. Akçimen F, Ross JP, Bourassa CV, *et al*. Investigation of the RFC1 repeat expansion in a Canadian and a Brazilian ataxia cohort: Identification of novel conformations. *Front Genet*. 2019;10:1219.
15. Rafehi H, Szmulewicz DJ, Bennett MF, *et al*. Bioinformatics-based identification of expanded repeats: A non-reference intronic pentamer expansion in RFC1 causes CANVAS. *Am J Hum Genet*. 2019;105:151-165.
16. Cortese A, Tozza S, Yau WY, *et al*. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome due to RFC1 repeat expansion. *Brain*. 2020;143:480-490.
17. Bronstein AM, Mossman S, Luxon LM. The neck-eye reflex in patients with reduced vestibular and optokinetic function. *Brain*. 1991;114:1-11.
18. Migliaccio AA, Halmagyi GM, McGarvie LA, Cremer PD. Cerebellar ataxia with bilateral vestibulopathy: Description of a syndrome and its characteristic clinical sign. *Brain*. 2004;127:280-293.
19. Szmulewicz DJ, Waterston JA, MacDougall HG, *et al*. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome (CANVAS): A review of the clinical features and video-oculographic diagnosis. *Ann N Y Acad Sci*. 2011;1233:139-147.
20. Wu TY, Taylor JM, Kilfoyle DH, *et al*. Autonomic dysfunction is a major feature of cerebellar ataxia, neuropathy, vestibular areflexia 'CANVAS' syndrome. *Brain*. 2014;137:2649-2656.
21. Cazzato D, Bella ED, Dacci P, Mariotti C, Lauria G. Cerebellar ataxia, neuropathy, and vestibular areflexia syndrome: A slowly progressive disorder with stereotypical presentation. *J Neurol*. 2016;263:245-249.
22. Traschütz A, Cortese A, Reich S, *et al*. Natural history, phenotypic spectrum, and discriminative features of multisystemic RFC1 disease. *Neurology*. 2021;96:e1369-e1382.
23. Meindl T, Cordts I, Scherzer AL, *et al*. CANVAS: Case report on a novel repeat expansion disorder with late-onset ataxia. *Nervenarzt*. 2020;91:537-540.
24. Cortese A, Reilly MM, Houlden H. RFC1 CANVAS/spectrum disorder. In: Adam MP, Everman DB, Mirzaa GM, *et al*, eds. *GeneReviews®*: University of Washington; 1993-2022. Accessed 2 February 2022. https://www.ncbi.nlm.nih.gov/books/NBK564656/
25. Scriba CK, Beecroft SJ, Clayton JS, *et al*. A novel RFC1 repeat motif (ACAGG) in two Asia-pacific CANVAS families. *Brain*. 2020;143:2904-2910.
26. Beecroft SJ, Cortese A, Sullivan R, *et al*. A Māori specific RFC1 pathogenic repeat configuration in CANVAS, likely due to a founder allele. *Brain*. 2020;143:2673-2680.
27. Tsuchiya M, Nan H, Koh K, *et al*. RFC1 repeat expansion in Japanese patients with late-onset cerebellar ataxia. *J Hum Genet*. 2020;65:1143-1147.
28. Gisatulin M, Dobricic V, Zühlke C, *et al*. Clinical spectrum of the pentanucleotide repeat expansion in the *RFC1* gene in ataxia syndromes. *Neurology*. 2020;95:e2912-e2923.
29. Abramzon Y, Dewan R, Cortese A, *et al*. Investigating RFC1 expansions in sporadic amyotrophic lateral sclerosis. *J Neurol Sci*. 2021;430:118061.
30. Jacquemont S, Hagerman RJ, Leehey M, *et al*. Fragile X premutation tremor/ataxia syndrome: Molecular, clinical, and neuroimaging correlates. *Am J Hum Genet*. 2003;72:869-878.
31. Salcedo-Arellano MJ, Dufour B, McLennan Y, Martinez-Cerdeno V, Hagerman R. Fragile X syndrome and associated disorders: Clinical aspects and pathology. *Neurobiol Dis*. 2020;136:104740.
32. Mitsuhashi S, Frith MC, Mizuguchi T, *et al*. Tandem-genotypes: Robust detection of tandem repeat expansions from long DNA reads. *Genome Biol*. 2019;20:58.
33. Giesselmann P, Brändl B, Raimondeau E, *et al*. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat Biotechnol*. 2019;37:1478-1481.
34. Gilpatrick T, Lee I, Graham JE, *et al*. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol*. 2020;38:433-438.
35. Kovaka S, Fan Y, Ni B, Timp W, Schatz MC. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat Biotechnol*. 2021;39:431-441.
36. Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat Biotechnol*. 2021;39:442-450.
37. Stevanovski I, Chintalaphani SR, Gamaarachchi H, *et al*. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci Adv*. 2022;8:eabm5386.
38. Schöls L, Amoiridis G, Büttner T, Przuntek H, Epplen JT, Riess O. Autosomal dominant cerebellar ataxia: Phenotypic differences in genetically defined subtypes? *Ann Neurol*. 1997;42:924-932.
39. Ruano L, Melo C, Silva MC, Coutinho P. The global epidemiology of hereditary ataxia and spastic paraplegia: A systematic review of prevalence studies. *Neuroepidemiology*. 2014;42:174-183.
40. Labun K, Montague TG, Krause M, Torres Cleuren YN, Tjeldnes H, Valen E. CHOPCHOP V3: Expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res*. 2019;47:W171-W174.
41. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol*. 2019;20:129.
42. Oxford Nanopore Technologies. Guppy protocol: modified base calling. Accessed 2 February 2022. https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revz_14dec2018/modified-base-calling
43. Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094-3100.
44. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. Nanopack: Visualizing and processing long-read sequencing data. *Bioinformatics*. 2018;34:2666-2669.
45. GitHub. https://github.com/erdogant/findpeaks. Accessed 2 February 2022.

46. Oxford Nanopore Technologies. Megalodon. Accessed 2 February 2022. https://nanoporetech.github.io/megalodon
47. Perlman S. Hereditary ataxia overview. In: Adam MP, Everman DB, Mirzaa GM, *et al*, eds. *GeneReviews®*. University of Washington; 1993-2022. Accessed 2 February 2022. https://www.ncbi.nlm.nih.gov/books/NBK1138/
48. Sequeiros J, Martindale J, Seneca S, *et al*. EMQN Best practice guidelines for molecular genetic testing of SCAs. *Eur J Hum Genet*. 2010;18:1173-1176.
49. Sequeiros J, Seneca S, Martindale J. Consensus and controversies in best practices for molecular genetic testing of spinocerebellar ataxias. *Eur J Hum Genet*. 2010;18:1188-1195.
50. Nolin SL, Glicksman A, Tortora N, *et al*. Expansions and contractions of the FMR1 CGG repeat in 5,508 transmissions of normal, intermediate, and premutation alleles. *Am J Med Genet A*. 2019; 179:1148-1156.
51. Orr HT, Chung M, Banfi S, *et al*. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat Genet*. 1993;4:221-226.
52. Cabal-Herrera AM, Tassanakijpanich N, Salcedo-Arellano MJ, Hagerman RJ. Fragile X-Associated Tremor/Ataxia Syndrome (FXTAS): Pathophysiology and clinical implications. *Int J Mol Sci*. 2020;21:4391.
53. Allen EG, He W, Yadav-Shah M, Sherman SL. A study of the distributional characteristics of FMR1 transcript levels in 238 individuals. *Hum Genet*. 2004;114:439-447.
54. Kenneson A, Zhang F, Hagedorn CH, Warren ST. Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate-length and premutation carriers. *Hum Mol Genet*. 2001;10:1449-1454.
55. Opal P, Ashizawa T. Spinocerebellar Ataxia Type 1. In: Adam MP, Everman DB, Mirzaa GM, *et al*, eds. *GeneReviews®*. University of Washington;1993-2023. Accessed 10 January 2023. https://www.ncbi.nlm.nih.gov/books/NBK1184/
56. De Biase I, Rasmussen A, Monticelli A, *et al*. Somatic instability of the expanded GAA triplet-repeat sequence in Friedreich ataxia progresses throughout life. *Genomics*. 2007;90:1-5.
57. Long A, Napierala JS, Polak U, *et al*. Somatic instability of the expanded GAA repeats in Friedreich's ataxia. *PLoS ONE*. 2017;12: e0189990.
58. Sharma R, Bhatti S, Gomez M, *et al*. The GAA triplet-repeat sequence in Friedreich ataxia shows a high level of somatic instability in vivo, with a significant predilection for large contractions. *Hum Mol Genet*. 2002;11:2175-2187.
59. Burman RW, Popovich BW, Jacky PB, Turker MS. Fully expanded FMR1 CGG repeats exhibit a length-and differentiation-dependent instability in cell hybrids that is independent of DNA methylation. *Hum Mol Genet*. 1999;8:2293-2302.
60. Shafin K, Pesout T, Chang PC, *et al*. Haplotype-aware variant calling with PEPPER-margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods*. 2021;18:1322-1332.
61. WhatsHap. Accessed 2 February 2022. https://github.com/whatshap/whatshap
62. DeepVariant. Accessed 02 February 2022. https://github.com/google/deepvariant
63. Hensel CH, Vanzo RJ, Martin MM, *et al*. Abnormally methylated FMR1 in absence of a detectable full mutation in a U.S.A patient cohort referred for Fragile X testing. *Sci Rep*. 2019;9:15315.
64. Dolzhenko E, van Vugt JJFA, Shaw RJ, *et al*. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res*. 2017;27:1895-1903.
65. Benson G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27:573-580.
66. Klockgether T. Sporadic ataxia with adult onset: Classification and diagnostic criteria. *Lancet Neurol*. 2010;9:94-104.
67. Linnemann C, Tezenas du Montcel S, Rakowicz M, *et al*. Peripheral neuropathy in spinocerebellar ataxia type 1, 2, 3, and 6. *Cerebellum*. 2016;15:165-173.
68. Bidichandani SI, Delatycki MB. Friedreich ataxia. In: Adam MP, Everman DB, Mirzaa GM, *et al*, eds. *GeneReviews®*: University of Washington; 1993-2022. Accessed 2 February 2022. https://www.ncbi.nlm.nih.gov/books/NBK1281/
69. Berciano J, Infante J, García A, Polo JM, Volpini V, Combarros O. Very late-onset Friedreich's ataxia with minimal GAA1 expansion mimicking multiple system atrophy of cerebellar type. *Mov Disord*. 2005;20:1643-1645.
70. Fearon C, Lonergan R, Ferguson D, *et al*. Very-late-onset Friedreich's ataxia: Diagnosis in a kindred with late-onset cerebellar ataxia. *Pract Neurol*. 2020;20:55-58.
71. Rota S, Marchina E, Todeschini A, *et al*. Very late-onset Friedreich ataxia with laryngeal dystonia. *Case Rep Neurol*. 2014;6:287-290.
72. Salomão RPA, Gama MTD, Rezende Filho FM, Maggi F, Pedroso JL, Barsottini OGP. Late-Onset Friedreich's Ataxia (LOFA) mimicking Charcot–Marie–Tooth disease type 2: What is similar and what is different? *Cerebellum*. 2017;16:599-601.
73. de Silva RN, Vallortigara J, Greenfield J, Hunt B, Giunti P, Hadjivassiliou M. Diagnosis and management of progressive ataxia in adults. *Pract Neurol*. 2019;19:196-207.

Supplementary Material

PCR-based fragment analysis of repeat expansions

SCA1, 2, 3, 6, 7, 8, 17

PCR was used to amplify the repeat containing regions of interest in the *ATXN1* (SCA1), *ATXN2* (SCA2), *ATXN3* (SCA3), *CACNA1A* (SCA6), *ATXN7* (SCA7), *ATXN8OS/ATXN8* (SCA8) and *TBP* (SCA17) gene. Primer sequences are given in the supporting information (Supplementary Table 4). For the PCR reaction mixture, 50 ng of genomic DNA, dNTPs (0.2 mM), forward and reverse primers (1 µM each) and 0.5 U of Taq polymerase (Qiagen) were combined with Q-Solution (Qiagen) in PCR buffer (Qiagen). Amplification was performed by initial heating to 94°C for 5 minutes followed by 14 cycles of 94°C for 30 seconds, 64°C for 15 seconds, 72°C for 30 seconds, subsequent 25 cycles of 94°C for 30 seconds, 50°C for 30 seconds, 72°C for 30 seconds) and final heating to 72°C for 7 minutes. PCR products were diluted and combined with Hi-Di-Formamide (Life Technologies) and Gene Scan 500 LIZ Size Standard (Applied Biosystems). Amplification products were separated on ABI PRISM™ Genetic Analyzer 3730XL followed by repeat length determination by GeneMarker (SoftGenetics). For patients with only one allele identified in the analysis of the *ATXN2*, *ATXN7* and *ATXN8OS/ATXN8* locus a long-range PCR was additionally performed. For the PCR reaction, 150 ng genomic DNA, dNTPs (0.5 mM), forward, reverse and anchor primers (0.3 µM each) as well as 3.75 U of Taq polymerase (Qiagen) in Expand Long Template PCR buffer (Roche). Analysis of fragments was performed as described before.

FXS and FXTAS

Repeat length of the *FMR1* gene (FXTAS and FXS) were determined using the Asuragen AmplideX PCR/CE FMR1 reagent kit by following provider's instructions.

FRDA

Short-range: Samples were amplified by PCR with primers targeting the GAA repeats in the *FXN* gene. Primer sequences are given in the supporting information. For the PCR reaction mixture 150–200 ng genomic DNA, dNTPs (0,2 mM), 1 µM of primers and 0.5 U of Taq polymerase (Qiagen) were combined with Q-Solution (Qiagen) in PCR buffer (Qiagen). Amplification was performed by initial heating to 95°C for 10 minutes followed by 39 cycles of 95°C for 30 seconds, 62°C for 30 seconds, 72°C for 60 seconds. PCR products were diluted and combined with Hi-Di-Formamide (Life Technologies) and Gene Scan 500 LIZ Size Standard (Applied Biosystems). Amplification products were separated on ABI PRISM™ Genetic Analyzer 3730XL followed by repeat length determination by GeneMarker (SoftGenetics).

Long-range: Samples were amplified by PCR with primers targeting the GAA repeats in the *FXN* gene. Primer sequences are given in the supporting information (Supplementary Table 4). Amplification was performed using the Roche Expand Long Template PCR System following provider's instructions. PCR products were diluted and combined with Hi-Di-Formamide (Life Technologies) and Gene Scan 500 LIZ Size Standard (Applied Biosystems). Repeat length determination was performed by fragment separation on ABI PRISM™ Genetic Analyzer 3730XL followed by evaluation with GeneMarker (SoftGenetics).

FMR1 methylation analysis by MLPA

Methylation level of the *FMR1* locus were determined by MRC-Holland SALSA MS-MLPA Probemix ME029-B3 FMR1/AFF2 following provider's instructions.

Supplementary Table 1. Repeat loci analyzed with corresponding gene, disorder, mode of inheritance, location within the gene, repeat motif and classification of repeat sizes (according to references [1–11]).

Disease	Gene	Inheritance	Genomic region	Repeat sequence	Allele sizes
Disorders associated with expansion of short coding repeats					
SCA1	*ATXN1*	AD	Exon	CAG	Normal: 6-35 Mutable normal: 36-38 Pathogenic (full penetrance): > 38
SCA2	*ATXN2*	AD	Exon	CAG	Normal: ≤ 32 Pathogenic (reduced penetrance): 33-34 Pathogenic (full penetrance): > 34
SCA3	*ATXN3*	AD	Exon	CAG	Normal: < 45 Pathogenic (reduced penetrance): 45-51 Pathogenic (full penetrance): > 51
SCA6	*CACNA1A*	AD	Exon	CAG	Normal: < 19 Pathogenic (reduced penetrance): 19 Pathogenic (full penetrance): 20-33
SCA7	*ATXN7*	AD	Exon	CAG	Normal: 7 – 27 Mutable normal: 28–33 Pathogenic (reduced penetrance): 34–36 Pathogenic (full penetrance): >36
SCA8	*ATXN8OS/ ATXN8*	AD	3'-UTR	(CTA · TAG) (CTG · CAG)	Normal: 15–50 Pathogenic (incomplete penetrance): >50
SCA17	*TBP*	AD	Exon	CAG	Normal: 25-40 Pathogenic (reduced penetrance): 41-48 Pathogenic (full penetrance): 49-66
Disorders associated with expansion of long non-coding repeats					
FRDA	*FXN*	AR	Intron	GAA	Normal: 5-33 Pathogenic (reduced penetrance): 34-65 Pathogenic (full penetrance): 66-1700
CANVAS	RFC1	AR	Intron	See Supplementary Table 2	
FXTAS	FMR1	X-linked	5'-UTR	CGG	Normal: 5-54 Premutation allele (FXTAS-associated): 55-200 Full mutation (FXS-associated): 200-several thousands

Supplementary Table 2. Literature known repeat compositions of the intronic region in *RFC1* and their classification as non-pathogenic, likely non-pathogenic and pathogenic (according to references [8,12–19]).

Non-pathogenic patterns	
Sequence	Repeat units
$(AAAAG)_{widltype}$	11
$(AAAAG)_{expanded}$	12-200
$(AAAGG)_{expanded}$	40-1000
Likely Non-Pathogenic patterns	
Sequences	Repeat units
$(AAGAG)_{expanded}$	Unknown
$(AGAGG)_{expanded}$	Unknown
Pathogenic patterns	
Sequences	Repeat units
$(AAGGG)_{expanded}$	Most frequently 400-2000
$(ACAGG)_{expanded}$	~ 1000
Pattern with unknown significance	
Sequences	Repeat units
$(AACGG)_{expanded}$	unknown
$(AAAGGG)_{expanded}$	unknown

Supplementary Table 3. CRISPR RNAs (crRNAs) to enrich repeat regions within the indicated loci.

Locus	Guide	Sequence 5'→3'
ATXN1	ATXN1_D1	GGTGGAAACTTTTATCGGTT
	ATXN1_U2	ATGTAATCGATCTAAGAACC
ATXN2	ATXN2_D1	GTCGGCTCTGTCTCTACCGA
	ATXN2_U2	CCGGTCACCCGCCGTCAAGC
ATXN3	ATXN3_D1	AGCGCATTCCCAAATAGACG
	ATXN3_U1	GATTACTGCTGAACGCACAT
CACNA1A	CACNA1A_D2	GGTCCAGTTCTGCGTGGAAT
	CACNA1A_U2	TTGGCACTCGGGCATAGACT
ATXN7	ATXN7_D1	ATCTAGGTTAAACTTCCCGC
	ATXN7_U2	CGGTACTTCGTCCTGACACC
ATXN8OS/ATXN8	ATXN8_D1	CATTACAGGTCACGCAAAGA
	ATXN8_U1	ATACTTGGCCATCGTAATTG
TBP	TBP_D1	GAGGTTACTACTGCATGTTG
	TBP_U1	TGAGACGAGTTCCAGCGCAA
FXN	FXN_D1	CACCAGTTTCGAGAATCCTG
	FXN_U2	CTGCTGTAAACCCATACCGG
FMR1	FMR1_D2	ATCACGATCCCAATCTTCTC
	FMR1_U2	TTTAGGCTTGAGCAACGAAC
RFC1	RFC1_D1	TTCGTGGAACTATCTTGGTA
	RFC1_D2	TGATTACAACCATCAAGGAT
	RFC1_U1	TAACTTCCAACAACCTCAAC
	RFC1_U2	GCTCAGTCGTTTTAACCAGG

Supplementary Table 4. Primers for PCR-based fragment analysis.

Primers for Marker PCR		
Gene	**Forward Primer 5'→3'**	**Reverse Primer**
ATXN1	CAACATGGGCAGTCTGAGCCAG	GAACTGGAAATGTGGACGTACTGG
ATXN2	CGTGCGAGCCGGTGTATGGG	GGCGACGCTAGAAGGCCGCT
ATXN3	CCAGTGACTACTTTGATTCG	TGGCCTTTCACATGGATGTGAA
CACNA1A	CACACGTGTCCTATTCCCCTGTGATC	GGGTACCTCCGAGGGCCGCTGG
TBP	GACCCCACAGCCTATTCAGA	TTGACTGCTGAACGGCTGCA
ATXN7	TGTTACATTGTAGGAGCGGAA	CACGACTGTCCCAGCATCACTT
ATXN8OS/ ATXN8	CATCAGATAATTTTTGGAAGGATG	GTCCTTCATGTTAGAAAACCTGG
FXN (short-range)	GAAGAAACTTTGGGATTGGTTGC	CTGCCGCAGCCTCTGGAG
FXN (long-range)	GGAGGGATCCGTCTGGGCAAAGG	CAATCCAGGACAGTCAGGGCTTT

Primers for Triplet Repeat Primed Long-Range PCR (TP-LR PCR)			
Gene	***Anchor Primer***	***Forward Primer 5'→3'***	***Reverse Primer 5'→3'***
SCA2	TACGCATCCCAGTTTGAGACG	TACGCATCCCAGTTTGAGACGCAGCAGCAGCAGCAG	GAGGAGACCGAGGACGA
SCA7	TACGCATCCCAGTTTGAGACG	TACGCATCCCAGTTTGAGACGCAGCAGCAGCAGCAG	CAGGAAGTTTGGAAGCC
SCA8	TACGCATCCCAGTTTGAGACG	TACGCATCCCAGTTTGAGACGCAGCAGCAGCAGCAG	CATTCAGATTGCCTTTTCTGAC

Supplementary Table 5. Coverage of loci analyzed and repeat sizes determined by PCR and Clin-CATS for individuals in the validation cohort. For *RFC1* the motifs identified and repeat lengths determined are given.

Target	*Coverage*	*PCR (Allele 1)*	*PCR (Allele 2)*	*Nanopore (Allele 1)*	*Nanopore (Allele 2)*
1 Patient 280 (gender male)					
ATXN1	284	29	30	30	30
ATXN2	304	22	22	22	22
ATXN3	81	23	23	22	22
CACNA1A	83	11	14	11	14
ATXN7	629	10	10	10	10
ATXN8OS/ATXN8	197	23	27	26	28
FMR1	344	29	–	31	–
FXN	294	8	9	8	8
TBP	322	36	38	38	38
RFC1	476	-	-	AAAAG: 11	AAAAG: 105 (90-115)
Mean Coverage	301				
2 Patient 266 (gender male)					
ATXN1	110	30	30	30	30
ATXN2	108	22	22	23	23
ATXN3	47	23	30	21	30
CACNA1A	18	13	13	11	13
ATXN7	202	10	10	9	9
ATXN8OS/ATXN8	76	24	25	25	25
FMR1	127	29	–	33	–
FXN	135	9	9	9	9
TBP	76	35	36	35	35
RFC1	283	-	-	AAAAG: 11	AAAAG: 11
Mean Coverage	118				
3 Patient 023 (gender male)					
ATXN1	35	29	30	31	31
ATXN2	34	22	23	21	21
ATXN3	18	14	14	13	13
CACNA1A	15	12	13	12	12
ATXN7	53	10	13	10	10
ATXN8OS/ATXN8	27	18	23	18	25
FMR1	67	31	–	30	–
FXN	32	9	9	9	9
TBP	27	37	38	36	36
RFC1	76	-	-	AAAAG: 110 (105-115)	AAAAG: 130 (125-140)
Mean Coverage	38				
4 Patient 928 (gender female)					
ATXN1	214	29	29	30	30
ATXN2	128	22	22	22	22
ATXN3	72	23	26	25	25
CACNA1A	45	11	12	13	13
ATXN7	237	10	10	10	10
ATXN8OS/ATXN8	104	23	28	25	30

FMR1	396	30	30	32	32
FXN	185	19	24	18	23
TBP	150	38	38	38	38
RFC1	324	-	-	AAAAG: 11	AAAAG: 11
Mean Coverage	186				
5 Patient 100 (gender male)					
ATXN1	119	28	32	29	29
ATXN2	97	22	22	22	22
ATXN3	51	21	27	20	27
CACNA1A	49	13	13	13	13
ATXN7	306	10	10	11	11
ATXN8OS/ATXN8	41	19	25	21	25
FMR1	111	23	–	24	–
FXN	135	9	9	9	9
TBP	66	37	38	38	38
RFC1	275	-	-	AAAAG: 11	AAAAG: 60 (50-70)
Mean Coverage	125				
6 Patient 438 (gender female)					
ATXN1	151	29	30	30	30
ATXN2	127	22	22	22	22
ATXN3	39	20	24	21	21
CACNA1A	31	11	12	11	11
ATXN7	250	10	12	11	11
ATXN8OS/ATXN8	171	23	27	23	25
FMR1	167	29	29	34	34
FXN	124	9	19	9	19
TBP	135	37	38	38	38
RFC1	61	-	-	AAGGG: 950 (900-1050)	AAGGG: 795 (700-855)
Mean Coverage	126				
7 Patient 366 (gender female)					
ATXN1	178	29	29	29	29
ATXN2	272	22	23	23	23
ATXN3	80	23	23	23	23
CACNA1A	30	11	13	12	12
ATXN7	321	10	10	9	9
ATXN8OS/ATXN8	253	25	27	28	28
FMR1	362	30	30	32	32
FXN	210	8	9	9	9
TBP	233	36	38	37	37
RFC1	333	-	-	AAAAG: 11	AAAAG: 11
Mean Coverage	227				
8 Patient 004 (gender female)					
ATXN1	148	31	31	32	32
ATXN2	100	22	22	23	23
ATXN3	116	14	23	14	24
CACNA1A	22	7	12	7	11
ATXN7	107	10	10	12	12

ATXN8OS/ATXN8	89	18	27	19	27
FMR1	187	24	30	24	29
FXN	141	10	10	9	9
TBP	116	37	37	37	37
RFC1	138	-	-	AAAAG: 60 (45-75)	AAAAG: 100 (90-110)
Mean Coverage	116				
9 Patient 490 (gender male)					
ATXN1	63	29	31	30	30
ATXN2	50	22	22	22	22
ATXN3	75	14	24	13	24
CACNA1A	35	13	14	14	14
ATXN7	42	10	10	10	10
ATXN8OS/ATXN8	46	18	27	19	29
FMR1	62	30	–	30	–
FXN	57	8	9	9	9
TBP	61	36	36	36	36
RFC1	126	-	-	AAAAG: 11	AAAAG: 11
Mean Coverage	62				
10 Patient 756 (gender male)					
ATXN1	119	30	32	31	31
ATXN2	106	22	22	23	23
ATXN3	25	14	21	14	22
CACNA1A	30	11	12	11	11
ATXN7	59	10	11	10	10
ATXN8OS/ATXN8	167	23	25	25	25
FMR1	60	40	–	42	–
FXN	99	10	11	9	9
TBP	102	38	38	35	35
RFC1		-	-	AAAAG: 120 (110-130)	AAAGG/ AAAGGG: 60 (55-65)
Mean Coverage	85				
11 Patient 697 (gender male)					
ATXN1	547	29	30	30	30
ATXN2	466	22	22	23	23
ATXN3	460	23	64	23	66
CACNA1A	84	12	13	13	13
ATXN7	379	10	10	11	11
ATXN8OS/ATXN8	414	18	27	19	28
FMR1	358	29	–	32	–
FXN	477	22	25	22	25
TBP	509	36	37	36	36
RFC1	480	-	-	AAAGG: 165 (160-170)	AAAGG: 670 (610-710)
Mean Coverage	417				
12 Patient 823 (gender male)					
ATXN1	70	30	30	32	32
ATXN2	129	22	23	22	22
ATXN3	29	20	27	20	27

CACNA1A	12	11	11	10	12
ATXN7	167	10	14	10	10
ATXN8OS/ATXN8	92	19	26	20	28
FMR1	82	104	–	110-127	–
FXN	70	10	21	10	21
TBP	72	36	37	35	35
RFC1	142	-	-	AAAAG: 45 (40-50)	AAAAG: 110 (100-115)
Mean Coverage	87				
13 Patient 573 (gender male)					
ATXN1	71	29	29	29	29
ATXN2	86	22	22	22	22
ATXN3	32	21	23	22	22
CACNA1A	8	11	12	n.d.	n.d.
ATXN7	123	12	13	13	13
ATXN8OS/ATXN8	102	25	25	26	26
FMR1	67	>200	–	358-639	–
FXN	81	10	10	9	9
TBP	87	36	37	37	37
RFC1	143	-	-	AAAAG: 11	AAAAG: 105 (85-110)
Mean Coverage	80				
14 Patient 577 (gender female)					
ATXN1	445	30	43	30	45
ATXN2	400	22	22	22	22
ATXN3	828	20	23	22	22
CACNA1A	247	8	13	8	13
ATXN7	366	10	10	10	10
ATXN8OS/ATXN8	305	28	30	31	31
FMR1	988	31	31	32	32
FXN	460	9	9	9	9
TBP	346	36	37	36	36
RFC1	1429	-	-	AAAAG: 11	AAAAG: 45 (40-50)
Mean Coverage	581				

Supplementary Table 6. Mean coverage of loci of interest, mean differences between PCR repeat length and Clin-CATS as well as 95% limits of agreement overall samples in the validation cohort.

Target	Mean Coverage	Median Coverage	Mean Difference	95% limits of agreement
ATXN1	182	134	0.50	-1.60–2.60
ATXN2	172	118	0.18	-1.24–1.60
ATXN3	140	62	-0.14	-2.45–2.16
CACNA1A	51	31	-0.04	-1.75–1.67
ATXN7	232	220	-0.11	-2.63–2.41
ATXN8OS/ ATXN8	149	103	1.14	-1.04–3.23
FMR1	241	147	1.43	-2.00–4.86
FXN	179	135	-0.29	-1.58–1.01
TBP	164	109	-0.36	-2.63–1.92
RFC1	330	275	n.d.	n.d.
Mean	151			

Supplementary Table 7. Results of the analysis of 100 patients with ataxia symptoms (diagnostic cohort). Pathogenic findings in all regions of interest, confirmatory PCR results for expanded alleles and identified *RFC1* motives on both alleles and their sizes are given.

Pat.	Result Clin-CATS	PCR repeat analysis	*RFC1* motif (allele 1)	Repeat units of RFC1 (allele 1)	*RFC1* motif (allele 2)	Repeat units of RFC1 (allele 2)	Age at analysis
1	negative	-	AAAAG	11	AAAAG	11	68
2	negative	-	AAAAG	11	AAAAG	11	69
3	*SCA17:* 36/52	35/51	AAAAG	11	AAGAG	45 (35-50)	70
4	*RFC1 spectrum disorder*	-	AAGGG	680 (450-775)	AAGGG	790 (750-900)	74
5	*RFC1 spectrum disorder*	-	AAGGG	610 (565-640)	ACAGG	645 (615-695)	65
6	*SCA3:* 22/68	23/67	AAAAG	11	AAAGG	630 (570-660)	69
7	negative	-	AAAAG	90 (85-95)	AAAAG	85 (80-90)	62
8	*SCA6:* 11/21; *carrier for RFC1 spectrum disorder*	11/22	AAAAG	105 (100-110)	AAGGG	430 (405-455)	74
9	negative	-	AAAAG	11	AAAAG	11	78
10	negative; *carrier for RFC1 spectrum disorder*	-	AAAAG	110 (90-120)	AAGGG	305 (260-330)	84
11	negative	-	AAAAG	11	AAAAG	110 (105-110)	67
12	negative	-	AAAAG	40 (35-45)	AAAAG	70 (65-75)	56
13	negative	-	AAAGG/ AAAGGG	90 (80-95)	AAAGG/ AAAGGG	90 (80-95)	66
14	*FXTAS (high function male):* 97-359; *FMR1* promotor methylation: 3.2%	>200	AAAAG	11	AAAAG	11	61
15	negative	-	AAAAG	11	AAAGG	525 (505-545)	30
16	negative	-	AAAAG	105 (85-115)	AAAAG	140 (125-150)	28
17	negative	-	AAAAG	11	AAAAG	11	64
18	negative	-	AAAAG	11	AAAAG	100 (75-110)	71
19	negative	-	AAAAG	11	AAAAG	115 (95-125)	79
20	negative	-	AAAAG	11	AAAAG	105 (85-115)	69
21	negative; *carrier for RFC1 spectrum disorder*	-	AAAAG	11	AAGGG	500 (440-525)	83
22	negative	-	AAAAG	135 (130-145)	AAAAG	125 (120-130)	57
23	negative	-	AAAAG	90 (70-100)	AAAAG	125 (110-140)	61
24	negative	-	AAAAG	11	AAAAG	11	54
25	negative	-	AAAAG	11	AAAAG	11	63
26	negative	-	AAAAG	100 (85-110)	AAAAG	125 (110-130)	74
27	negative	-	AAAAG	30 (25-35)	AAAGG	245 (215-290)	64
28	negative	-	AAAAG	11	AAAAG	115 (80-125)	54
29	negative	-	AAAAG	120 (100-130)	AAAAG	140 (135-145)	71
30	negative	-	AAAAG	11	AAAGG	420 (375-455)	64
31	negative	-	AAAAG	160 (150-185)	AAAAG	125 (85-145)	62
32	negative	-	AAAAG	11	AAAAG	11	59
33	negative	-	AAAAG	115 (95-125)	AAAAG	115 (95-125)	65
34	*RFC1 spectrum disorder*	-	AAGGG	>600	AAGGG	>400	51
35	negative	-	AAAAG	11	AAAAG	11	60

36	negative	-	AAAAG	120 (95-130)	AAAAG	140 (130-150)	81
37	negative	-	AAAAG	30 (25-35)	AAAGG	520 (445-555)	66
38	negative	-	AAAAG	11	AAAAG	95 (70-105)	51
39	negative	-	AAGAG	45 (30-50)	AAAAG	100 (75-110)	51
40	*RFC1 spectrum disorder*	-	AAGGG	850 (810-915)	AAGGG	600 (545-640)	59
41	negative	-	AAAAG	11	AAAAG	11	71
42	*RFC1 spectrum disorder*	-	AAGGG	620 (565-670)	AAGGG	620 (565-670)	80
43	negative	-	AAAAG	11	AAAAG	135 (120-145)	61
44	negative	-	AAAAG	95 (85-95)	AAAAG	125 (125-130)	56
45	*SCA2*: 22/37	22/36	AAAAG	11	AAAGG	505 (465-540)	78
46	negative	-	AAAAG	11	AAAAG	11	52
47	*RFC1 spectrum disorder*	-	AAGGG	725	AAGGG	835	73
48	negative	-	AAAAG	110 (105-115)	AAAAG	135 (130-140)	61
49	*FXTAS:* 90-105; *FMR1* promotor methylation: 0.1%	92	AAAGG/ AAAGG G	90 (75-105)	AAAAG	120 (105-125)	71
50	negative	-	AAAAG	11	AAAAG	120 (110-135)	84
51	negative	-	AAAAG	11	AAAAG	11	53
52	*RFC1 spectrum disorder*	-	AAGGG	>270	AAGGG	>250	66
53	negative	-	AAAAG	95 (75-95)	AAAAG	110 (95-125)	52
54	negative	-	AAAAG	11	AAAAG	60 (50-65)	51
55	negative	-	AAAAG	11	AAAAG	120 (100-145)	77
56	*RFC1 spectrum disorder; carrier for FRDA:* 9/69	10/71	AAGGG	610	AAGGG	755	63
57	negative	-	AAAAG	11	AAAAG	100 (80-115)	78
58	negative	-	AAAAG	11	AAAAG	115 (90-140)	78
59	negative	-	AAAAG	11	AAAAG	11	40
60	negative	-	AAAAG	11	AAAAG	11	46
61	intermediate *ATXN1* allele: 33/36 with 2 CAT interruptions	31/37	AAAAG	11	AAAAG	120 (100-130)	32
62	*RFC1 spectrum disorder*	-	AAGGG	890 (845-925)	AAGGG	685 (575-750)	74
63	negative	-	AAAAG	11	AAAAG	11	63
64	negative	-	AAAAG	11	AAAAG	60 (55-65)	58
65	*SCA3:* 13/66	14/64	AAAAG	11	AAAAG	80 (70-90)	68
66	negative	-	AAAAG	90 (70-100)	AAAAG	105 (90-125)	57
67	negative	-	AAAAG	11	AAAAG	140 (125-150)	61
68	*RFC1 spectrum disorder*	-	AAGGG	745 (640-815)	AAGGG	925 (850-990)	70
69	*RFC1 spectrum disorder*	-	AAGGG	715	AAGGG	865	80
70	intermediate *ATXN1* allele: 33/37 with 2 CAT interruptions	32/36	AAAAG	85 (70-90)	AGGGG	125 (105-135)	64
71	*SCA8:* 26/120-214	23/134	AAAAG	11	AAAGG/ AAAGGG	85 (75-90)	46
72	*SCA8:* 87-240	90/133	AAAAG	11	AAAAG	11	48
73	negative	-	AAAAG	11	AAAAG	55 (50-56)	68
74	*carrier for FRDA:* 9/530-810	10/670	AAAAG	11	AAAAG	115 (95-125)	88
75	negative	-	AAAAG	100 (85-110)	AAAAG	130 (115-145)	59
76	negative	-	AAAAG	11	AAAAG	55 (45-60)	63
77	negative	-	AAAAG	95 (80-100)	AAAAG	140 (125-145)	62
78	negative	-	AAAAG	11	AAAAG	11	64
79	negative	-	AAAAG	11	AAAAG	105 (85-110)	79

80	*RFC1 spectrum disorder*	-	AAGGG	630 (555-700)	AAGGG	790 (750-815)	75
81	negative	-	AAAAG	11	AAAAG	100 (90-110)	58
82	*RFC1 spectrum disorder;* intermediate *ATXN8/ATXN8OS allele:* 29/59	29/60	AAGGG	715 (665-765)	AAGGG	915 (910-915)	63
83	negative	-	AAAAG	11	AAAAG	125 (110-135)	72
84	negative	-	AAAAG	11	AAAAG	11	58
85	*SCA6:* 12/22	12/22	AAAAG	11	AAAAG	11	76
86	negative	-	AAAAG	11	AAAAG	60 (55-65)	59
87	negative	-	AAAAG	11	AAAAG	115 (95-125)	69
88	*RFC1 spectrum disorder*	-	AAGGG	720 (670-765)	AAGGG	815 (790-855)	59
89	negative	-	AAAAG	11	AAAAG	115 (100-125)	72
90	*SCA7:* 10/43	9/47	AAAAG	11	AAAAG	120 (115-130)	37
91	*RFC1 spectrum disorder*	-	AAGGG	750	AAGGG	815	71
92	negative	-	AAAAG	11	AAAAG	115 (110-125)	70
93	negative	-	AAAAG	11	AAAAG	11	63
94	negative; *carrier for RFC1 spectrum disorder*	-	AAAAG	11	AAGGG	420 (430-440)	22
95	*RFC1 spectrum disorder*	-	AAGGG	700 (650-745)	AAGGG	860 (840-880)	82
96	negative	-	AAAAG	11	AAAAG	11	62
97	negative	-	AAAAG	11	AAAAG	95 (85-100)	43
98	*FRDA: 145-315/790-1280*	260/1070	AAAAG	100 (95-110)	AAAGGG/ AAGGG	55 (50-60)	53
99	negative	-	AAAAG	11	AAAAG	11	58
100	negative; *carrier for RFC1 spectrum disorder*	-	AAAAG	11	AAGGG	680 (630-670)	70

Supplementary Table 8. Results of the *FMR1* promotor methylation analysis.

Patient	Description	Methylation level [%]
Validation cohort		
928	healthy female	42.0
438	healthy female	45.4
366	healthy female	43.9
004	healthy female	42.3
577	healthy female	42.6
	Average	43.2 (95%-CI: 42.0–44.5)
100	healthy male	0.0
490	healthy male	0.1
280	healthy male	0.0
266	healthy male	0.2
023	healthy male	0.0
756	healthy male	0.1
697	SCA3 male	0.1
	Average	0.1 (95%-CI: 0.0–0.1)
573	FXS male	69.4
823	FXTAS male	0.6
Diagnostic cohort		
49	FXTAS male	0.1
14	High functioning male	3.2

Supplementary Table 9. Age at disease onset and symptoms of patients with *RFC1* spectrum disorder. Clinical presentation of patients with *RFC1* spectrum disorder including age at disease manifestation, presence of chronic chough, sensory neuro(no)pathy, cerebellar ataxia, bilateral vestibulopathy and autonomic dysfunctions.

Patient	Age at disease manifestation	Cough	PNP	Sensory ataxia	Bilateral Vestibulopathy	Autonomic Dysfunction
4	70	+	+	+	+	-
5	61	+	+	+	-	+
34	37	+	+	+	+	-
40	data not available	data not available	+	+	+	data not available
42	67	+	+	+	+	-
47	40	+	+	+	+	-
52	63	+	+	+	-	+
56	60	-	+	+	-	-
62	60	-	+	+	+	-
68	68	+	+	+	+	-
69	59	+	+	+	+	+
80	68	-	+	+	+	+
82	62	-	+	+	+	-
88	45	-	+	+	+	-
91	65	+	+	+	-	-
95	67	-	+	+	+	-

References

1. Opal P, Ashizawa, T. Spinocerebellar Ataxia Type 1. In: Adam MP, Ardinger HH, Pagon RA, *et al*. *GeneReviews*® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022.https://www.ncbi.nlm.nih.gov/books/NBK1184/. Accessed 13 August 2022.

2. Pulst, S. Spinocerebellar Ataxia Type 2. In: Adam MP, Ardinger HH, Pagon RA, *et al*. *GeneReviews*® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022. https://www.ncbi.nlm.nih.gov/books/NBK1275/. Accessed 13 August 2022.

3. Paulson, H, Shakkottai V. Spinocerebellar Ataxia Type 3. In: Adam MP, Ardinger HH, Pagon RA, *et al*. *GeneReviews*® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022. https://www.ncbi.nlm.nih.gov/books/NBK1196/. Accessed 13 August 2022.

4. Casey HL, Gomez MC. Spinocerebellar Ataxia Type 6. In: Adam MP, Ardinger HH, Pagon RA, *et al*. *GeneReviews*® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022. https://www.ncbi.nlm.nih.gov/books/NBK1140/. Accessed 13 August 2022.

5. Toyoshima Y, Onodera O, Yamada, M, Tsuji S, Takahashi, H. Spinocerebellar Ataxia Type 17. In: Adam MP, Ardinger HH, Pagon RA, *et al*. *GeneReviews*® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022. https://www.ncbi.nlm.nih.gov/books/NBK1438/. Accessed 13 August 2022.

6. Hunter JE, Berry-Kravis E, Hipp H, Todd PK. FMR1 Disorders. In: Adam MP, Ardinger HH, Pagon RA, *et al*. *GeneReviews*® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022. https://www.ncbi.nlm.nih.gov/books/NBK1116/. Accessed 13 August 2022.

7. Bidichandani, Sanjay, Delatycki, Martin B. Friedreich Ataxia. In: Adam MP, Ardinger HH, Pagon RA, *et al*. *GeneReviews*® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022. https://www.ncbi.nlm.nih.gov/books/NBK1281/. Accessed 13 August 2022.

8. Cortese A, Reilly, Mary M, Houlden, Henry. RFC1 CANVAS / Spectrum Disorder. In: Adam MP, Ardinger HH, Pagon RA, *et al*. *GeneReviews*® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022. https://www.ncbi.nlm.nih.gov/books/NBK564656/. Accessed 13 August 2022.

9. Perlman S. Hereditary Ataxia Overview. In: Adam MP, Ardinger HH, Pagon RA, *et al*. *GeneReviews*® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022. https://www.ncbi.nlm.nih.gov/books/NBK1138/. Accessed 13 August 2022.

10. La Spada AR. Spinocerebellar Ataxia Type 7. In: Adam MP, Ardinger HH, Pagon RA, *et al*. *GeneReviews*® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022. https://www.ncbi.nlm.nih.gov/books/NBK1256/. Accessed 13 August 2022.

11. Cleary JD, Subramony S, Ranum LP. Spinocerebellar Ataxia Type 8. In: Adam MP, Ardinger HH, Pagon RA, *et al*. *GeneReviews*® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022. https://www.ncbi.nlm.nih.gov/books/NBK1268. Accessed 13 August 2022.

12. Cortese A, Simone R, Sullivan R, *et al*. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat Genet*. 2019;51(4):649-658.

13. Cortese A, Tozza S, Yau WY, et al. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome due to RFC1 repeat expansion. *Brain*. 2020;143(2):480-490.

14. Akçimen F, Ross JP, Bourassa CV, *et al.* Investigation of the RFC1 Repeat Expansion in a Canadian and a Brazilian Ataxia Cohort: Identification of Novel Conformations. *Front Genet*. 2019;10:1219.

15. Scriba CK, Beecroft SJ, Clayton JS, *et al.* A novel RFC1 repeat motif (ACAGG) in two Asia-Pacific CANVAS families. *Brain*. 2020;143(10):2904-2910.

16. Beecroft SJ, Cortese A, Sullivan R, *et al.* A Māori specific RFC1 pathogenic repeat configuration in CANVAS, likely due to a founder allele. *Brain*. 2020;143(9):2673-2680.

17. Tsuchiya M, Nan H, Koh K, *et al.* RFC1 repeat expansion in Japanese patients with late-onset cerebellar ataxia. *J Hum Genet*. 2020;65(12):1143-1147.

18. Abramzon Y, Dewan R, Cortese A, *et al.* Investigating RFC1 expansions in sporadic amyotrophic lateral sclerosis. *J Neurol Sci*. 2021;430:118061.

19. Stevanovski I, Chintalaphani SR, Gamaarachchi H, *et al.* Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci Adv*. 2022;8(9):eabm5386.

6 Paper II: Methylation of the 4q35 D4Z4 repeat defines disease status in facioscapulohumeral muscular dystrophy

Hannes Erdmann[±], Florentine Scharf[±], Stefanie Gehling, Anna Benet-Pagès, Sibylle Jakubiczka, Kerstin Becker, Maria Seipelt, Felix Kleefeld, Karl Christian Knop, Eva-Christina Prott, Miriam Hiebeler, Federica Montagnese, Dieter Gläser, Matthias Vorgerd, Tim Hagenacker, Maggie C. Walter, Peter Reilich, Teresa Neuhann, Martin Zenker, Elke Holinski-Feder, Benedikt Schoser, Angela Abicht.
Brain. 2023;146(4):1388–1402.

[±] These authors contributed equally.

DOI:

10.1093/brain/awac336

Impact factor of Brain (2021):

15.255

Rank by Journal Citation Indicator Category Clinical Neurology (2021):

6/267 (JCI Percentile 97.94%)

https://doi.org/10.1093/brain/awac336 BRAIN 2023: 146; 1388–1402 | 1388

BRAIN
ORIGINAL ARTICLE

Methylation of the 4q35 D4Z4 repeat defines disease status in facioscapulohumeral muscular dystrophy

Hannes Erdmann,[1,2,†] Florentine Scharf,[1,†] Stefanie Gehling,[1] Anna Benet-Pagès,[1,3] Sibylle Jakubiczka,[4] Kerstin Becker,[1] Maria Seipelt,[5] Felix Kleefeld,[6] Karl Christian Knop,[7] Eva-Christina Prott,[8] Miriam Hiebeler,[2] Federica Montagnese,[2] Dieter Gläser,[9] Matthias Vorgerd,[10] Tim Hagenacker,[11] Maggie C. Walter,[2] Peter Reilich,[2] Teresa Neuhann,[1] Martin Zenker,[4] Elke Holinski-Feder,[1,12] Benedikt Schoser[2] and Angela Abicht[1,2]

[†]These authors contributed equally to this work.

Genetic diagnosis of facioscapulohumeral muscular dystrophy (FSHD) remains a challenge in clinical practice as it cannot be detected by standard sequencing methods despite being the third most common muscular dystrophy. The conventional diagnostic strategy addresses the known genetic parameters of FSHD: the required presence of a permissive haplotype, a size reduction of the D4Z4 repeat of chromosome 4q35 (defining FSHD1) or a pathogenic variant in an epigenetic suppressor gene (consistent with FSHD2). Incomplete penetrance and epistatic effects of the underlying genetic parameters as well as epigenetic parameters (D4Z4 methylation) pose challenges to diagnostic accuracy and hinder prediction of clinical severity.

In order to circumvent the known limitations of conventional diagnostics and to complement genetic parameters with epigenetic ones, we developed and validated a multistage diagnostic workflow that consists of a haplotype analysis and a high-throughput methylation profile analysis (FSHD-MPA). FSHD-MPA determines the average global methylation level of the D4Z4 repeat array as well as the regional methylation of the most distal repeat unit by combining bisulphite conversion with next-generation sequencing and a bioinformatics pipeline and uses these as diagnostic parameters. We applied the diagnostic workflow to a cohort of 148 patients and compared the epigenetic parameters based on FSHD-MPA to genetic parameters of conventional genetic testing. In addition, we studied the correlation of repeat length and methylation level within the most distal repeat unit with age-corrected clinical severity and age at disease onset in FSHD patients. The results of our study show that FSHD-MPA is a powerful tool to accurately determine the epigenetic parameters of FSHD, allowing discrimination between FSHD patients and healthy individuals, while simultaneously distinguishing FSHD1 and FSHD2. The strong correlation between methylation level and clinical severity indicates that the methylation level determined by FSHD-MPA accounts for differences in disease severity among individuals with similar genetic parameters. Thus, our findings further confirm that epigenetic parameters rather than genetic parameters represent FSHD disease status and may serve as a valuable biomarker for disease status.

1 Medical Genetics Center (MGZ), 80335 Munich, Germany
2 Friedrich-Baur-Institute, Department of Neurology, Klinikum der Universität, Ludwig-Maximilians-Universität, 80336 Munich, Germany
3 Institute of Neurogenomics, Helmholtz Center Munich, 85764 Neuherberg, Germany

Received May 13, 2022. Revised August 06, 2022. Accepted August 31, 2022. Advance access publication September 14, 2022

4 Institute of Human Genetics, Universitätsklinikum Magdeburg, Otto-von-Guericke Universität, 39120 Magdeburg, Germany
5 Department of Neurology, Universitätsklinikum Marburg, Philipps-University Marburg, 35043 Marburg, Germany
6 Department of Neurology and Experimental Neurology, Charité Berlin, 10117 Berlin, Germany
7 Neurologische Praxis Neuer Wall, 20354 Hamburg, Germany
8 Praxis für Humangenetik, 42103 Wuppertal, Germany
9 Genetikum, 89231 Neu-Ulm, Germany
10 Department of Neurology, Berufgenossenschaftliches Universitätsklinikum Bergmannsheil, Ruhr-Universität Bochum, 44789 Bochum, Germany
11 Department of Neurology and Center for Translational Neuro- and Behavioral Sciences (C-TNBS), University Hospital Essen, 45147 Essen, Germany
12 Department of Medicine IV, Klinikum der Universität, Ludwig-Maximilians-Universität, 80336 Munich, Germany

Correspondence to: Angela Abicht
Medical Genetics Center, Bayerstr. 3–5
80335 Munich, Germany
E-mail: angela.abicht@mgz-muenchen.de

Correspondence may also be addressed to: Benedikt Schoser
Friedrich-Baur-Institute, Department of Neurology
LMU Klinikum, Ziemenssenstr. 1
80336 München, Germany
E-mail: benedikt.schoser@med.uni-muenchen.de

Keywords: FSHD; methylation; bisulphite sequencing; epigenetics

Introduction

Facioscapulohumeral muscular dystrophy (FSHD; OMIM #158900) is a hereditary progressive myopathy characterized by initial asymmetric weakness and atrophy of facial, shoulder girdle and upper arm muscles with a descending involvement of the distal lower extremities and possibly the pelvic girdle.[1–3] Despite this distinct clinical presentation, the phenotype may vary in terms of the pattern of muscle affection, incomplete symptoms or atypical features complicating the differentiation from other myopathies or neurological diseases.[4–6] FSHD presents with an autosomal-dominant mode of inheritance affecting both males and females and can manifest at all ages, mostly within the second or third decade of life.[7] *De novo* cases are found in about 30% of patients with adult-onset FSHD and in about 70% of patients with early-onset FSHD.[8,9] A high degree of variability regarding the age at disease onset, impairment and disease progression is observed between individuals, even within the same family carrying identical genetic features and in monozygotic twins.[6,10,11] Despite its high prevalence[1,12,13] and numerous therapeutic approaches,[14–17] diagnostic confirmation of individuals affected by FSHD remains challenging.[2,18–22]

At the molecular level, FSHD is mediated by a loss of repression of the silenced *DUX4* gene in somatic cells as a result of structural and epigenetic rearrangements of the subtelomere D4Z4 macrosatellite repeat region on chromosome 4q35.[23–25] Stable expression of the *DUX4* gene causes damage, dystrophic changes and atrophy in skeletal muscle via different pathways.[23,26,27] Molecular prerequisite for a stable *DUX4* transcript is a specific permissive haplotype (4qA and haplotype variant 4qAL) that provides a polyadenylation signal (PAS) for the *DUX4* mRNA within the most distal repeat unit (RU) in the FSHD locus.[28] Currently, two subtypes of FSHD are distinguished based on their molecular background. In the most common form, FSHD1, accounting for about 95% of cases, *DUX4* derepression is linked to a contraction of the D4Z4 macrosatellite repeat array to less than 12 RU.[29–31] In rare FSHD2, *DUX4* expression is associated with global hypomethylation of the D4Z4 repeat array that is usually caused by genetic defects in genes encoding for proteins involved in epigenetic suppression. To date, known FSHD2-causing epigenetic suppressor genes include the structural maintenance of chromosomes flexible hinge domain containing 1 gene (*SMCHD1*), the methyltransferase 3B gene (*DNMT3B*) and the ligand-dependent nuclear receptor interacting factor 1 gene (*LRIF1*).[32–34] Also in FSHD2, manifestation of the disease is linked to stable *DUX4* expression and therefore requires the presence of at least one permissive allele on chromosome 4. Genetic diagnosis has conventionally been based on (i) confirmation of the presence of a permissive haplotype; followed by (ii) determination of the D4Z4 repeat length by Southern blotting[35] and, in patients without D4Z4 repeat contraction, sequencing of *SMCHD1* and related epigenetic suppressor genes.[32,36] However, this strategy comes with limitations: (i) Southern blotting for repeat size analysis requires large amounts of high molecular weight DNA, which can only be obtained by elaborate pre-analytics and freshly drawn blood for DNA extraction. (ii) Repeat contractions (especially moderate ones) on permissive haplotypes have no full penetrance. They are not only found in FSHD1 patients but also in 1–2% of healthy individuals.[20,37] Additionally, current diagnostic protocols cannot distinguish whether a repeat contraction is in *cis* or *trans* to the permissive haplotype. (iii) Assessing the clinical relevance of variants in *SMCHD1* and other epigenetic suppressor genes is difficult because their functional relevance is co-determined by structural and epigenetic parameters of both 4q35 alleles.[38] (iv) In some patients with FSHD phenotype, neither repeat contraction nor pathogenic variants in the known epigenetic suppressor genes can be identified, suggesting that additional factors are associated with disease that are not captured by the conventional analytic strategy.[39]

Two recent approaches, molecular combing and single-molecule optical mapping, improved FSHD testing by deciphering the architecture of the FSHD locus as they simultaneously

determine haplotype and repeat length, also of large D4Z4 arrays, and as they detect complex rearrangements.[40–42] However, because the tests are also based on repeat length, some of the previously described limitations remain.

To overcome these limitations, the use of methylation as a diagnostic parameter has been proposed.[43–45] Hypomethylation of the CpG-rich (73%) D4Z4 repeat was described early in FSHD patients, and different protocols have been used since.[37,46,47] A current protocol based on bisulphite sequencing with subsequent vector cloning of individual fragments and sequencing of reaction products made use of the hypomethylation and its different distribution observed for FSHD1 versus FSHD2 to distinguish between healthy individuals and patients affected by either FSHD1 or FSHD2. This assay allows determination of the local methylation status of the most distal repeat unit of alleles carrying the permissive haplotype. In addition, the global methylation status of the whole D4Z4 repeat array on chromosome 4q35 is determined.[45,48] Based on the methylation profile, individuals with isolated distal hypomethylation will have an epigenetic diagnosis of FSHD1, whereas individuals with global and distal hypomethylation will have an epigenetic diagnosis of FSHD2.

In addition to being discussed as diagnostic marker, methylation has also been considered as a marker of disease severity. The most accurate prognostic parameter for FSHD1 disease status known to date is the repeat size of the D4Z4 repeat array, as it shows a mild inverse correlation with disease severity and a mild positive correlation with age at disease onset.[49–52] However, its relevance is limited because a large phenotypical variance is observed for individuals carrying similarly sized contracted alleles. Moreover, FSHD2 patients are not represented. Using the above-mentioned methylation assay, a qualitative association of disease severity and methylation level within the distal repeat unit has been shown.[53] Therefore, methylation level might be suitable as biomarker for disease severity needed for upcoming therapeutic approaches.[14,16,54]

In this study, we developed and implemented a multistage diagnostic approach for the diagnosis of FSHD based on epigenetics. The diagnostic workflow consists of (i) a haplotype analysis by two independent assays, one of them novel and capturing the region of the poly-A signal, to confirm or exclude permissive alleles; and (ii) a high-throughput methylation profile analysis (FSHD-MPA) that uses regions and primers reported by Jones *et al.*[45] but combines bisulphite conversion reactions with next-generation sequencing (NGS), and a bioinformatics pipeline. We applied this diagnostic workflow in a cohort of 148 patients and compare the epigenetic results to genetic parameters of conventional genetic testing (repeat-size analysis and sequencing of epigenetic suppressor genes) and to the patient's phenotype. By correlating distal methylation level of the D4Z4 repeat array and age-corrected clinical severity, we verify methylation profiles not only as a diagnostic parameter but also as a biomarker for FSHD disease status.

Materials and methods

Patients and study approval

In total, 224 individuals assigned to three cohorts were analysed within this study. The 'establishment cohort' of 56 individuals (Supplementary Tables 4–6) with known FSHD disease status was used to establish the laboratory and bioinformatic procedure of FSHD-MPA and to determine thresholds for pathogenic methylation levels within the different methylation assays. This cohort included 24 unaffected controls (Supplementary Table 6) and 32 FSHD patients based on a classic clinical phenotype and known genetic parameters (presence of a permissive haplotype, a D4Z4 repeat size reduction <12 RU defining FSHD1 in 29 patients; Supplementary Table 4), or a pathogenic variant in the epigenetic suppressor gene *SMCHD1* (defining FSHD2 in three patients; Supplementary Table 5).

A 'diagnostic cohort' consisted of 148 individuals (Supplementary Tables 7–10) that were referred for either symptomatic ($n = 145$) or predictive testing ($n = 3$). Symptomatic individuals were reported with a phenotype compatible with FSHD, asymptomatic individuals were referred for predictive testing because of a positive family history of FSHD. The diagnostic outcome was analysed by comparing the results of FSHD-MPA with Southern blotting and sequencing of epigenetic suppressor genes whenever possible.

A 'genotype–phenotype cohort' of 70 FSHD-MPA-positive patients (patients of the diagnostic cohort and additional patients shown in Supplementary Table 11) was assembled to study the correlation between age at disease onset and clinical severity with repeat size and methylation level. Standardized phenotype data were collected from patient records, including the age at disease onset, clinical signs and symptoms and family history. The age-corrected clinical severity score (CSS) was calculated as previously established for all patients with detailed clinical description by[21,55]

$$age - corrected\ CSS = \frac{2 \times CSS}{age\ at\ examination} \times 1000 \quad (1)$$

In total, complete phenotype datasets were available for 66 patients to calculate the CSS.[21,55] For an additional four patients, only the age at disease onset was available. The age at disease onset was recorded within a 20-year interval because of an individually variable experienced onset of disease and difficulty assessing the parameter retrospectively. Patients with different haplotypes (4qA or 4qAL) were analysed separately as the assays target different regions and have specific and different thresholds for pathogenic results.

Informed consent was obtained from all participants. All genetic analyses and investigations were performed in accordance with the guidelines of the Declaration of Helsinki and approved by local institutions (Bayerische Landesärztekammer, vote no. 2019-210).

Multistage diagnostic workflow

A multistage diagnostic workflow that was established and applied to the diagnostic cohort. Based on the phenotype description, we first performed a haplotype analysis by two independent assays to confirm or exclude the presence of at least one permissive allele. Patients who did not have a permissive allele were diagnosed as FSHD-negative. Second, a high-throughput methylation profile analysis (FSHD-MPA) was carried out to determine distal and global D4Z4 methylation levels and to diagnose FSHD1 and FSHD2 based on epigenetic parameters. Analyses of FSHD underlying genetic parameters (D4Z4 repeat contraction, pathogenic variants in epigenetic suppressor genes) were carried out to further confirm the FSHD diagnosis based on epigenetic parameters or to identify alternative diagnoses in patients in whom FSHD is considered unlikely (Fig. 1A).

Determination of permissive haplotypes 4q161 and 4qA/4qAL

Two independent assays were used to identify the presence of permissive haplotypes: (i) haplotype assay A: allele-specific Sanger sequencing of a single nucleotide polymorphism (SNP) containing a region proximal to the D4Z4 repeat array (p13E11) to identify the

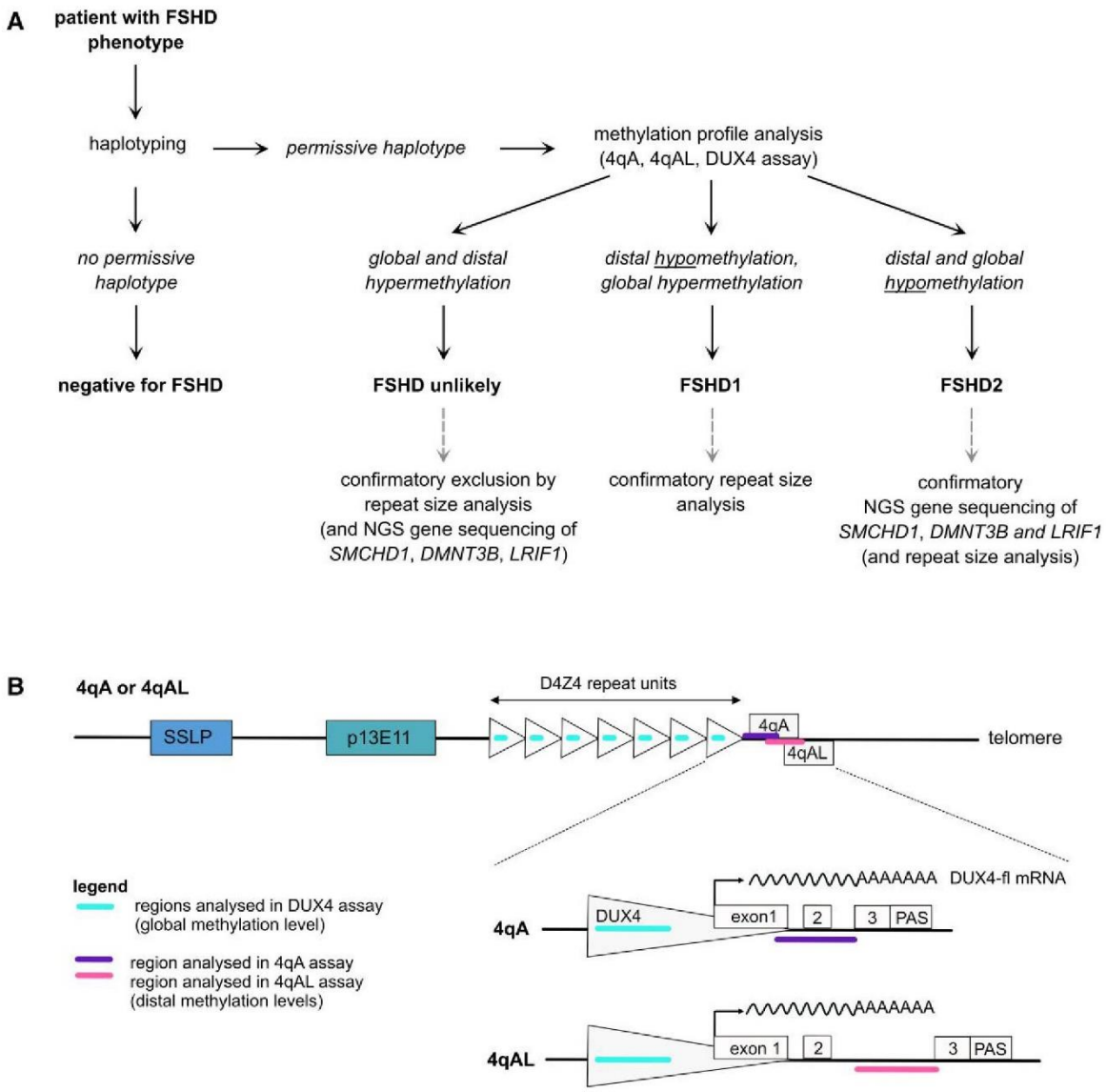

Figure 1 Diagnostic workflow of FSHD testing based on methylation profile analysis. (**A**) Multistage diagnostic workflow that consists of (i) haplotype analysis to confirm or exclude permissive alleles; and (ii) high-throughput methylation profile analysis (FSHD-MPA) using three different methylation assays (DUX4, 4qA, 4qAL) to detect global and distal methylation level of the FSHD locus. Patients with a distal hypomethylation (4qA or 4qAL assay, covering CpGs within the most distal repeat unit of the haplotypes 4qA and 4qAL) are assigned as compatible with FSHD1, those with a distal (4qA and/or 4qAL assay) and global hypomethylation (DUX4 assay, covering CpGs within each D4Z4 repeat unit of chromosome 4) are assigned as compatible with FSHD2. Patients with a hypermethylated D4Z4 region are considered to be not affected by FSHD. (**B**) Schematic representation of the architecture of a D4Z4 repeat array on chromosome 4 with regions assayed by methylation profile analysis. Distal methylation status is determined within the last repeat unit in the 4qA (*top*) and 4qAL (*bottom*) assay. Global methylation status of the locus is assayed as average over all D4Z4 repeat units within this array (lines in triangles).

presence of the most frequent permissive 4qA161 subhaplotype as previously described[56]; and (ii) haplotype Assay B: KASP genotyping assay (LGC Biosearch Technologies) to detect a SNP in the intronic region of the most distal D4Z4 repeat present in all permissive 4qA and 4qAL haplotypes (chromosomal position chr4: 190175588, reference genome GRch38/hg38.p11). The assay was performed on a Roche LightCycler 480 instrument following the manufacturer's instruction using two designed probes:

(5′-CCCCCGCGCCACCGTCGCCCGCCCGCCCGGGCCCCTGCAGCC TCCCAGCTGCCAGC[G/A]CGGAGCTCCTGGCGGTCAAAAGCATACC TCT GTCTGTCTTTGCCCGCTTCCTGG-3′). Patients without permissive haplotype were diagnosed negative for FSHD.

FSHD-MPA

FSHD-MPA consists of three different assays: DUX4, 4qA and 4qAL methylation assay. The DUX4 assay determines the methylation status of a 59 CpGs containing region present in each D4Z4 repeat unit. It represents the global methylation status of the 4q35 region. The 4qA and 4qAL assays determine the regional methylation status of the most distal repeat unit on the permissive haplotypes 4qA and 4qAL, which differ by a 2.2 kb large intronic extension present in the latter. They cover regions of 56 CpGs (4qA assay) and 31 CpGs (4qAL assay), respectively. Amplification of non-permissive alleles or similar regions as in chromosome 10 is avoided by nested PCRs

using region-specific primers as reported by Jones *et al.*[45] Following the protocol of Jones *et al.*,[45] 1 µg of gDNA was converted using the Epitec Bisulfite Kit (Qiagen) following the manufacturer's instructions. Amplification of 150 ng of converted gDNA was performed by nested PCR with three sets of primers (4qA, 4qAL and DUX4 assay) using HotStarTaq Plus Polymerase (Qiagen) as described.[45] Primer sequences used in the assays are given in Supplementary Table 1. After quality control of the amplicons by fragment analysis, library preparation for NGS sequencing was performed on 10 ng of DNA using NEBNext Ultra Library Prep Kit according to the manufacturer's instructions. Pooled samples were sequenced by NGS using an Illumina MiSeq system.

Reads were quality and adapter trimmed using cutadapt v3.4 and TrimGalore v0.6.1. Reads were mapped using bwameth v0.2.2 against the sequences of the nested PCR products (4qA/4qAL/DUX4) (Supplementary Table 2).

After mapping, known CpG positions are extracted from sequencing data and counted: a C corresponds to a methylated CpG; a T to an unmethylated CpGs, in which the C underwent conversion. From these counts, mean methylation levels were calculated over all reads and CpGs. Overlapping regions from paired end reads were only considered once. Only samples with more than 5000 reads within each assay, an average coverage of all CpGs within one assay of at least 1000× and less than 5 CpGs with a coverage below 500× were considered for analysis.

Cut-offs for FSHD1/2 positive or negative classifications were defined as the 99.9% CI of the methylation levels of 4qA and 4qAL and the 99% CI of the methylation levels of DUX4 in the establishment cohort. The area between the thresholds for positive and negative predictions has been defined as inconclusive (grey zone) to prevent overfitting. Validity of the approach has been confirmed using a 3-fold cross-validation. Determined cut-offs (Supplementary Table 3) serve for the assessment of patients as positive or negative for FSHD: patients with isolated distal hypomethylation were diagnosed as FSHD1 (4qA assay or 4qAL assay); patients with distal (4qA and/or 4qAL assay) and global hypomethylation (DUX4 assay) were diagnosed as FSHD2; patients with distal and global hypermethylation—corresponding to the epigenetic suppression of *DUX4* expression in healthy individuals—were diagnosed negative for FSHD. In each diagnostic run, controls with confirmed negative and positive result for FSHD1 and FSHD2 are included as quality control.

Next-generation sequencing

Analysis of *SMCHD1* and *DNMT3B* as well as of *LRIF1* included in a custom panel (Agilent SureSelectXT or Twist Human Comprehensive Exome + Mitochondrial Genome) comprising 2896 and 19 182 genes, respectively, was performed by NGS using an Illumina NextSeq 500 system or Illumina NovaSeq 6000 system. Only regions covered with at least 20× were considered for assessment. Only variants (single-nucleotide polymorphisms/small insertions and deletions (INDELs)) in the coding and flanking intronic regions (±50 bp) were evaluated. Variants were classified according to the ACMG (American College of Medical Genetics and Genomics) guidelines.[57,58]

Extraction of genomic DNA and Southern blotting for D4Z4 repeat length analysis is described in the Supplementary Material.

Statistical analysis

Statistical analyses were performed using the software R v.4.0.2. To study whether methylation levels of patients affected by FSHD1 and FSHD2 are significantly lowered to healthy individuals within the establishment cohort, *P*-values were calculated using a one-tailed *t*-test. For the 4qA and 4qAL assay, the group of healthy individuals was compared to the group of patients affected by FSHD1 and FSHD2. For the DUX4 assay, the group of FSHD2 patients was compared to healthy individuals and FSHD1 patients, respectively. To study the correlation of repeat length and age at disease onset or clinical severity within the genotype–phenotype cohort, all patients with pathogenic repeat contractions (<12 RU) were considered independent of their epigenetic classification as affected by FSDH1, FSHD2 or both when all required clinical data were available. For the correlation analysis of methylation level and age at disease onset or clinical severity, hypomethylated distal methylation level determined in the 4qA or 4qAL assay of all patients with FSHD phenotype independent of their classification as affected by FSHD1 or FSHD2 were considered. Analysis was performed separately for the 4qA and 4qAL haplotype. In patients carrying a hypomethylated 4qA and 4qAL allele, methylation status of both alleles was considered. Correlation analyses were performed by Pearson's correlation test, 95% CIs of the correlation coefficients were determined and *P*-values were calculated to test the significance of the correlation.

Data availability

Anonymized data from this study are available from the corresponding author on reasonable request.

Results

Establishment of FSHD-MPA

In the establishment cohort of 56 individuals with known disease status based on genetic parameters, we determined methylation levels using three different methylation assays (DUX4, 4qA, 4qAL) (Figs 1B and 2 and Supplementary Tables 4–6). While healthy individuals showed high methylation levels within all three assays (4qA, 4qAL and DUX4), 24 of 29 FSHD1 patients showed a regional reduction of the methylation level of the distal repeat unit (4qA or 4qAL assay) without reduction of the global methylation level of the whole D4Z4 repeat array on chromosome 4q35 (DUX4 assay). Three of 29 FSHD1 patients showed additional reduction of the global methylation level, although no pathogenic variant in *SMCHD1* was detected. FSHD2 patients showed a global hypomethylation (DUX4 assay) including the distal repeat unit (4qA and/or 4qAL assay). Healthy individuals and FSHD patients significantly differed in their methylation levels (Fig. 2) within the 4qA and 4qAL assay ($P < 0.001$) and FSHD2 patients showed significant differences from FSHD1 patients ($P = 0.03$) and from healthy individuals ($P = 0.01$) within the DUX4 assay. This allowed defining assay-specific thresholds for normal, inconclusive and pathogenic results (Supplementary Table 3) based on the 99.9% (4qA and 4qAL assay) and 99% (DUX4 assay) CIs of the methylation levels of the different groups within the three assays.

FSHD-MPA in a diagnostic cohort

Our multistage diagnostic workflow for the diagnosis of FSHD based on epigenetic parameters (Fig. 1) gave the following results for a diagnostic cohort of 148 patients (Fig. 3A): in 36 patients (24%), an isolated distal hypomethylation, and in 14 patients (10%), a global hypomethylation of the D4Z4 repeat array was detected, leading to the epigenetic diagnosis of FSHD1 and FSHD2, respectively.

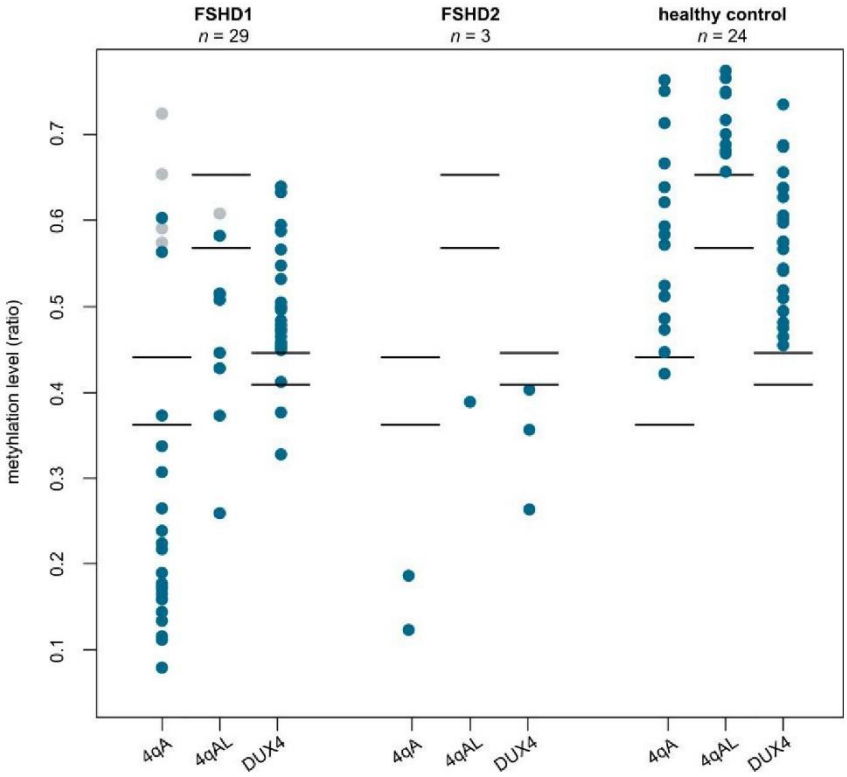

Figure 2 Methylation profiles of the D4Z4 locus of healthy individuals and FSHD1 and FSHD2 patients. Methylation levels in the establishment cohort of 56 individuals with known disease status based on genetic parameters. Methylation levels were determined in the three assays for FSHD1 patients (*left*), FSHD2 patients (*middle*) and unaffected controls (*right*). Thresholds for pathogenic and normal results are indicated by bold horizontal black lines. Grey dots represent hypermethylated 4qA alleles being heterozygous with hypomethylated 4qAL alleles or vice versa.

Eighty patients (54%) were tested negative based on the absence of a permissive allele or hypermethylation in FSHD-MPA. For 18 patients (12%), global and/or distal methylation levels were within the grey zone, leading to inconclusive results that require further analyses.

FSHD-MPA results indicating FSHD1

For 23 of 36 patients diagnosed with FSHD1 based on isolated distal hypomethylation in FSHD-MPA, material was available for D4Z4 repeat size analysis by Sothern blotting. In all cases, a contracted allele with less than 12 RU was detected, so the diagnosis based on epigenetic parameters was consistent with that of genetic parameters (Fig. 3B and Supplementary Table 7). Five FSHD1 patients (A3, A5, A19, A32, A36) showed not only a distal hypomethylation (defining FSHD1 based on epigenetic parameters) but also a mild reduction of the global methylation level within the inconclusive range.

FSHD-MPA results indicating FSHD2

In 14 patients (B1–B14), FSHD-MPA revealed a global hypomethylation of the D4Z4 repeat region leading to the epigenetic diagnosis of FSHD2 (Fig. 3C and Supplementary Table 8). Sequencing all patients for pathogenic alterations in *SMCHD1* and *DNMT3B* led to the identification of potentially causative variants in *SMCHD1* in seven patients (B1–B7; Table 1), consistent with the genetic presentation of FSHD2. Three variants are classified as likely pathogenic according to ACMG diagnostic criteria. The remaining four are classified as variants of uncertain significance (class 3). However, their complete absence from population databases and their bioinformatic prediction strongly suggest pathogenicity, even though the current evidence is insufficient for a formal classification as probably pathogenic (class 4). One of these patients (B5) showed an additional contraction of the D4Z4 repeat to 9 RU. Thus, this patient had combined genetic features of FSHD1 and FSHD2. In two patients (B4, B6), the D4Z4 repeat size could not be determined because no additional DNA was available.

Interestingly, of the seven patients epigenetically diagnosed with FSHD2 without any variants in *SMCHD1* or *DNMT3B* but with a global hypomethylation in FSHD-MPA, five (B8–B12) carried a moderate repeat contraction (6 to 9 RU). Based on genetic parameters, they would have been classified as FSHD1 patients. Two of the patients (B13, B14) with a reduced global methylation and diagnosis of FSHD2 based on epigenetic parameters would have been diagnosed negative for FSHD (uncontracted D4Z4 repeat sizes and absence of pathogenic variants in *SMCHD1*, *DNMT3B* and additionally *LRIF1*).

FSHD-MPA with inconclusive results

Twelve of 18 patients with inconclusive results based on FSHD-MPA showed a mild isolated reduction of the distal methylation within the intermediate range (patients I1–I12; Supplementary Table 9 and Fig. 3D). In six of these patients (I1–I6), a contracted D4Z4 allele was identified by Southern blotting, consistent with the diagnosis of FSHD1 based on genetic parameters. Five patients (I7–I11) had uncontracted D4Z4 repeats, making the diagnosis of FSHD unlikely.

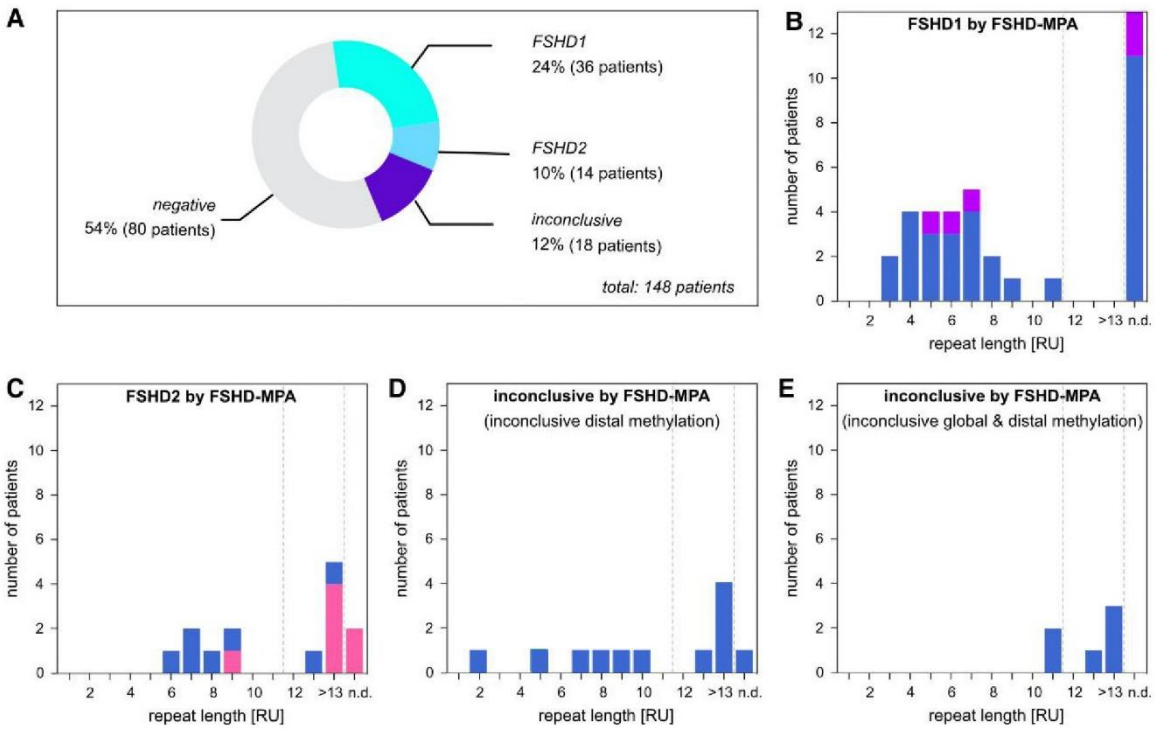

Figure 3 Results of the analysis of patients with suspected FSHD. (**A**) Diagnostic results for 148 patients analysed by FSHD-MPA. (**B**) Patients with FSHD1 diagnosis according to FSHD-MPA with their D4Z4 repeat sizes determined by Southern blotting. Patients with regional distal hypomethylation are indicated in blue, patients with additional mildly reduced global methylation within the inconclusive range are indicated in purple. (**C**) Patients with FSHD2 according to FSHD-MPA and their D4Z4 repeat size determined by Southern blotting. Patients carrying a potentially causal variant in *SMCHD1* are indicated in pink and patients without a causal variant in *SMCHD1* and *DNMT3B* are indicated in blue. (**D**) Patients with inconclusive results in FSHD-MPA showing isolated distal reduction of methylation within the grey zone and their D4Z4 repeat size determined by Southern blotting. (**E**) Repeat length of patients with inconclusive global and distal methylation in FSHD-MPA without potentially pathogenic variant in *SMCHD1* and *DNMT3B* with their D4Z4 repeat size determined by Southern blotting. First vertical dashed line indicates threshold of contracted repeat arrays compatible with FSHD1. RU = repeat units; n.d. = not determined.

In one patient (I12), the genetic diagnosis remained unsolved because no material was available for repeat lengths analysis. FSHD-MPA revealed 6 of 18 patients (I13–I18; Fig. 3E) with mildly reduced global and additionally distal methylation levels, all within the inconclusive range. Two of these patients (I15, I18) would have been diagnosed as FSHD1 based on genetic parameters (mild D4Z4 repeat contraction with 11 RU; negative results of *SMCHD1* and *DNMT3B* sequencing). In one patient (I17) with an uncontracted D4Z4 repeat array and no variant in *SMCHD1* and *DNMT3B*, two variants of uncertain significance (class 3 according to ACMG) in the *RYR1* gene (Table 2) were identified and a *RYR1*-associated myopathy was discussed as underlying cause of the clinical symptoms. However, there is not enough clinical and genetic evidence to confirm this differential diagnosis. The underlying cause of the patient's symptoms and the mildly reduced global and distal methylation remains unsolved. In the three remaining patients (I13, I14, I16) the absence of a repeat contraction and potentially pathogenic variant in *SMCHD1* or *DNMT3B* questions an FSHD diagnosis, although no other diagnosis could be established.

Patients with negative results based on absence of a permissive haplotype or negative FSHD-MPA

Eighty patients out of the diagnostic cohort were tested negative based on the absence of a permissive allele or a negative result in FSHD-MPA (Supplementary Table 10). In 14 patients, the result was negative based on the absence of a permissive haplotype. To confirm specific amplification of the distal D4Z4 region by FSHD-MPA, the first step of this analysis (bisulphite conversion and nested PCR) was performed. Only the DUX4 fragment was detected in the absence of 4qA or 4qAL fragments. In the remaining patients FSHD was considered unlikely based on negative results in FSHD-MPA. In 10 cases with a negative FSHD-MPA result but strong clinical suspicion of FSHD, D4Z4 repeat size analysis was performed to confirm the negative result of FSHD-MPA. Repeat analysis showed uncontracted alleles in all cases, and an additional sequence analysis of *SMCHD1* and *DNMT3B* carried out in four of them was also negative. In six patients (N7, N22, N27, N35, N63 and N66), alternative diagnoses appear to be very likely (Table 2). In another patient (N49) an alternative diagnosis is possible that needs to be confirmed. Three patients (N16, N37, N55) of the diagnostic cohort were predictively tested. Although two of these patients (N37 and N55) inherited genetic parameters of FSHD1 and FSHD2 from affected parents, respectively (Supplementary Table 10), the result of FSHD-MPA was negative consistent with the current asymptomatic status.

FSHD-MPA analysis of a family with contracted D4Z4 arrays of incomplete penetrance

In one family of our study (Fig. 4), individuals within three generations (grandfather (I:1, Z17), daughter (II:1, Z18), granddaughter

Table 1 Likely causative *SMCHD1* variants identified within this study

Patient	Variant (NM_015295.3, NG_031972.1)	Position/type of variant	Predicted consequence	ACMG classification	Clinical database ClinVar	Population database
B1	c.5843A > C p.His1948Pro	Exon 46/ missense	1 bp substitution in exon 46, change of amino acid from histidine to proline at a weakly conserved position that show moderate physicochemical differences[a]	Uncertain significance (PM2, PS3)	No entry	No entry
B2	c.5556_5561delinsT p.Lys1852Asnfs*17	Exon 45/ frameshift	6 bp deletion for T nucleotide insertion in exon 45, frameshift and PTC 17 codons downstream, NMD predicted[a]	Likely pathogenic (PVS1, PM2, PS3)	No entry	No entry
B3	c.4966 + 5G > T	Intron 39/ splice donor variant	1 bp substitution within the splice donor site, bioinformatics prediction of splice donor weakening[a]	Uncertain significance (PM2, PS3)	No entry	No entry
B4	c.2753T > A p.Leu918*	Exon 22/ nonsense	1 bp substitution in exon 22, generation of PTC, NMD predicted[a]	Likely pathogenic (PVS1, PM2, PS3)	No entry	No entry
B5	c.1846A > G p.Lys616Glu	Exon 14/ missense	1 bp substitution in exon 14, change of amino acid at a highly conserved position from lysine to glutamate differing mildly in their physicochemical properties (pathogenic according to bioinformatics prediction)[a]	Uncertain significance: (PM2, PS3, PP3)	1 entry in ClinVar: uncertain significance	No entry
B6	c.2409_2410del p.Tyr804Cysfs*8	Exon 19/ frameshift	2 bp deletion in exon 19, frameshift and PTC 8 codons downstream, NMD predicted[a]	Likely pathogenic (PVS1, PM2, PS3)	No entry	No entry
B7	c.1787G > C p.Trp596Ser	Exon 13/ missense	1 bp substitution in exon 13, change of amino acid at a highly conserved position from tryptophan to serine differing largely in their physicochemical properties (pathogenic according to bioinformatics prediction)[a]	Uncertain significance (PM2, PS3, PP3)	No entry	No entry
N56	c.1754G > A p.Arg585His	Exon 13/ missense	1 bp substitution in exon 14, change of amino acid at a highly conserved position from arginine to histidine differing in their physicochemical properties (pathogenic according to bioinformatics prediction)[a]	Uncertain significance: (PM2, PP3, PP4)	No entry	No entry
Z14	c.5145_5146del p.Thr1716fs	Exon 41/ frameshift	2 bp deletion in exon 41, frameshift and PTC one codons downstream, NMD predicted[a]	Likely pathogenic (PVS1, PM2)	No entry	No entry

ACMG = American College of Medical Genetics and Genomics; PTC = premature termination codon; NMD = nonsense-mediated decay; bp = base pair.
[a]Predicted consequences, not confirmed by experimental studies.

(III:1, Z19) and grandson (III:2, Z20)) carried a contracted D4Z4 repeat array of 2 RU in addition to a permissive haplotype. While the daughter (II:1, CSS = 4.5) and the grandchildren (III:1, CSS = 3.5 and III:2, CSS = 4) showed severe clinical impairment from FSHD beginning in childhood, the grandfather (I:1) was clinically unaffected. We performed FSHD-MPA for all carriers of a contracted D4Z4 array to evaluate whether healthy individuals can be distinguished from clinically affected patients based on the methylation profiles. While the unaffected grandfather had a negative result for FSHD-MPA showing hypermethylated D4Z4 repeat arrays, his daughter and grandchildren showed a highly hypomethylated distal repeat unit consistent with their severe clinical phenotype.

FSHD-MPA results in correlation to the clinical phenotype

After verifying methylation as a qualitative diagnostic parameter, we analysed the correlation of methylation status to the clinical phenotype in a cohort of 70 FSHD-MPA-positive patients independent of their classification as affected by FSHD1 or FSHD2 and compared it to the correlation of the D4Z4 repeat length in 46 patients with D4Z4 repeat contraction. First, we analysed the correlation of the age at disease onset with D4Z4 repeat length and distal methylation level (4qA/4qAL assay of FSHD-MPA) (Fig. 5), respectively. In general, the more contracted the D4Z4 repeat and the lower the methylation level of the most distal repeat unit is, the earlier

Table 2 Variants identified by NGS sequencing that were discussed as underlying alternative diagnoses in patients of this study

Patient	Gene (OMIM)	Variant(s)	Zygosity/mode of inheritance/exon/type of variant	ACMG classification	Gene-associated diseases and their case-specific assessment as alternative diagnosis to FSHD
I17	*RYR1* (180901)	NM_000540.2: c.5335C>T: (p.Pro1779Ser)	Heterozygous/AR or AD/exon 34/ missense variant	Unclear significance (PM2)	*RYR1*-associated myopathy discussed but not confirmed based on currently available clinical and genetic data
		c.7210G>A: (p.Glu2404Lys)	Heterozygous/AR or AD/exon 42/ missense variant	Unclear significance (PM2)	
N7	*VCP* (601023)	NM_007126.3: c.464G>A (p.Arg155His)	Heterozygous/AD/ exon 5/missense variant	Pathogenic (PS3, PS4, PM1, PM2, PM5, PP1, PP2, PP3)	Inclusion body myopathy with Paget disease of bone and/or frontotemporal dementia (IBMPFD) likely
N22	*FLNC* (102565)	NM_001458.4: c.8130G>A (p.Trp2710*)	Heterozygous/AD/ exon 48/nonsense variant	Pathogenic (PVS1, PS3, PS4, PM2, PP5)	Myofibrillar myopathy type 5 likely
N27	*DNM2* (602378)	NM_001005360.2: c.1856C>G (p.Ser619Trp)	Heterozygous/AD/ exon 17/missense variant	Pathogenic (PS3, PM2, PM5, PM6, PP5)	Centronuclear myopathy likely
N35	*PYROXD1* 617220)	NM_024854.3: c.464A>G (p.ASn155Ser)	Homozygous/AR/exon 5/missense variant	Pathogenic (PS3, PM2, PM3, PP1, PP3)	Myofibrillar myopathy type 8 likely
N49	*CAPN3* (114240)	NM_000070.2: c.550del (p.Thr184Argfs*36)	Heterozygous/AR or AD/exon 4/ frameshift variant	Pathogenic (PVS1, PS3, PS4, PM3, PM2)	Limb-girdle muscular dystrophy (LGMD R1 or LGMD D4) possible
		c.2219G>T (p.Asp707Tyr)	Heterozygous/AR or AD/exon 20	Unclear significance (PM2, PM5)	
N63 and N66 (twins)	*DMD* (300377)	NM004006.2: c.(649 + 1_650-1)_(2168 + 1_2169-1)dup	Heterozygous/XLR/ exon 8–17	Likely pathogenic (PS4, PP1, PM2, PP3)	Becker muscular dystrophy likely

ACMG = American College of Medical Genetics and Genomics; AD = autosomal dominant; AR = autosomal recessive; XLR = X-linked recessive.

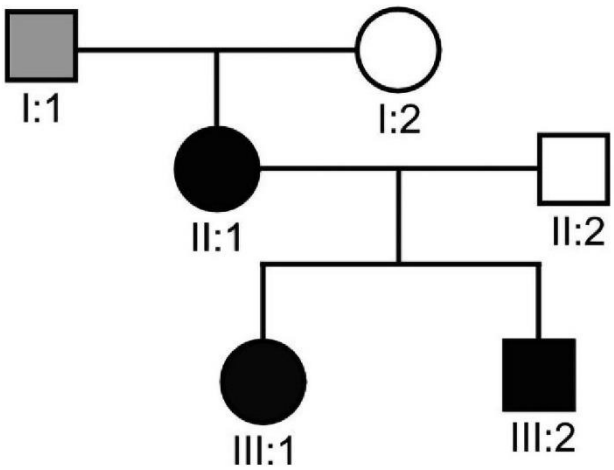

Figure 4 Family with carriers of a D4Z4 repeat contraction with incomplete penetrance.

the disease manifests. A very weak and non-significant correlation was found for repeat length and age at disease onset (Pearson's correlation coefficient $r = 0.21$, 95% CI: −0.09–0.47, $P = 0.17$; Fig. 5C), while a weak correlation was found for distal methylation level [4qA haplotype: Pearson's $r = 0.32$, 95% CI: 0.05–0.55 (Fig. 5A-); 4qAL haplotype: Pearson's $r = 0.38$, 95% CI: −0.07–0.70 (Fig. 5B)]. While the correlation for the 4qA haplotype was significant at the 95% significance level ($P = 0.02$), this criterion was narrowly missed for the 4qAL haplotype ($P = 0.10$). Furthermore, the lower the distal methylation or smaller the D4Z4 repeat size, the more severe the disease (Fig. 6). Repeat length and age-corrected disease severity were linked moderately and significantly (Pearson's $r = -0.48$, 95% CI: −0.21 to −0.68, $P < 0.01$; Fig. 6C). A moderate and strong correlation (both significant at the 95% level) was found for the methylation within the distal repeat unit revealed by the 4qA (Pearson's $r = -0.53$, 95% CI: −0.28 to −0.71, $P < 0.01$; Fig. 6A) and 4qAL assay of FSHD-MPA (Pearson's $r = -0.70$, 95% CI: −0.38 to −0.87, $P < 0.01$; Fig. 6B), respectively. We performed a regression analysis to obtain a functional description of the linkage of age-corrected CSS and methylation level with the distal repeat unit, assuming a linear relationship between both parameters. Using these equations, we independently determined the threshold values for pathogenic hypomethylation by setting the disease severity to exactly 0. In agreement with the threshold values determined within the establishment of the method, both limits are within the intermediate range slightly above the validated pathogenic threshold (4qA: 0.363 versus 0.362, 4qAL 0.617 versus 0.568).

Discussion

Ideally, FSHD diagnosis would rely on a characteristic clinical phenotype and the detection of *DUX4* mRNA or *DUX4* protein as this is the disease-causing consequence of epigenetic derepression of the genetic locus.[59,60] However, *DUX4* expression is not detectable in peripheral blood and only low and heterogeneous in affected muscles, with a small number of myocytes generating a large amount of *DUX4* protein.[61,62] Consequently, FSHD diagnosis continues to be based on genetic and epigenetic markers being associated with disease manifestation in combination with careful assessment of the patient's clinical presentation. To determine

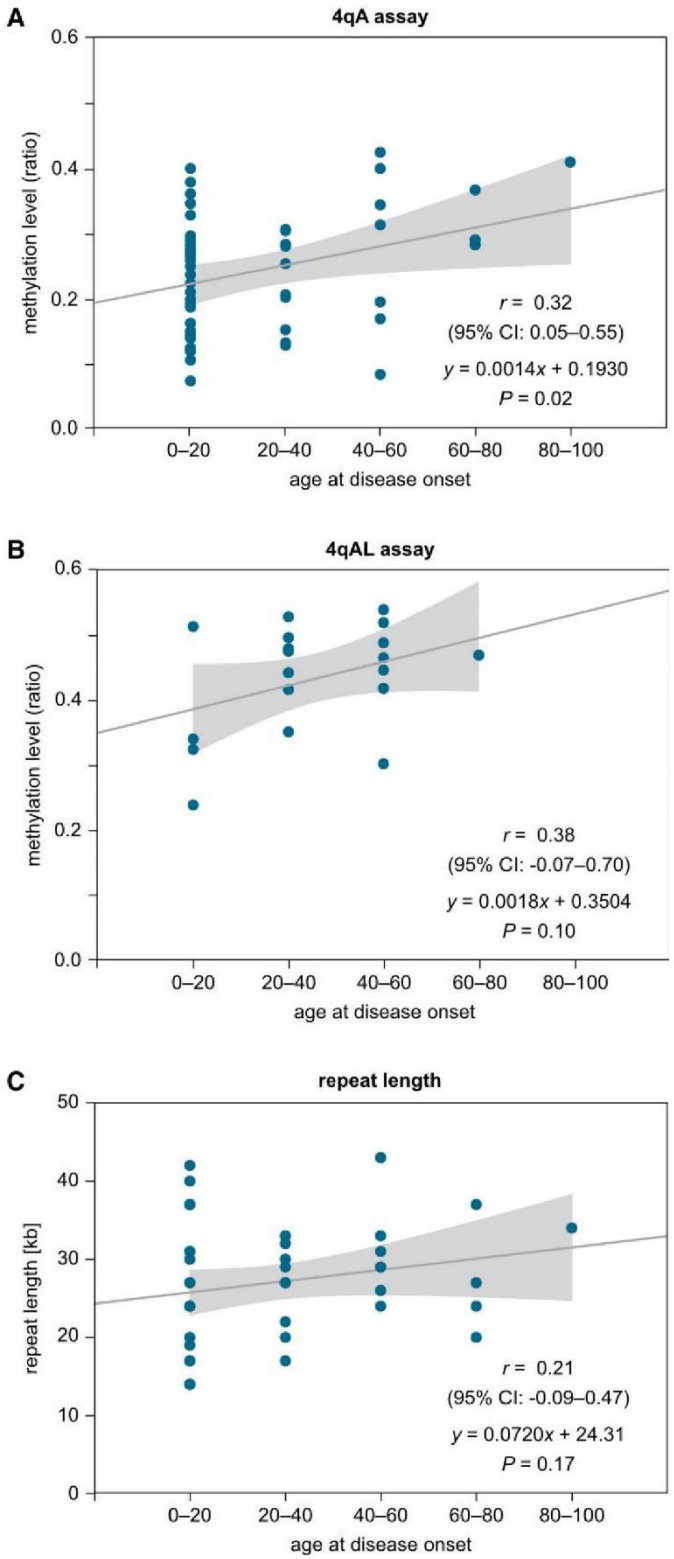

Figure 5 Correlation of age at disease onset with genetic and epigenetic parameters. Correlation and linear regression of age at disease onset with methylation level (**A**) of the 4qA assay, (**B**) of the 4qAL assay, as well as (**C**) with repeat length with respective Pearson's correlation coefficients, their 95% CIs and *P*-values. Highlighted areas around the regression lines indicate 95% CI of the regression.

the epigenetics of the D4Z4 array of chromosome 4q35 for diagnosis of FSHD, we implemented a methylation profile analysis using primers and regions reported by Jones *et al.*[45] (FSHD-MPA). In contrast to the original method, FSHD-MPA sequencing was performed by direct NGS of the bisulphite sequencing (BSS) products instead of cloning them into a vector for Sanger sequencing. In addition to

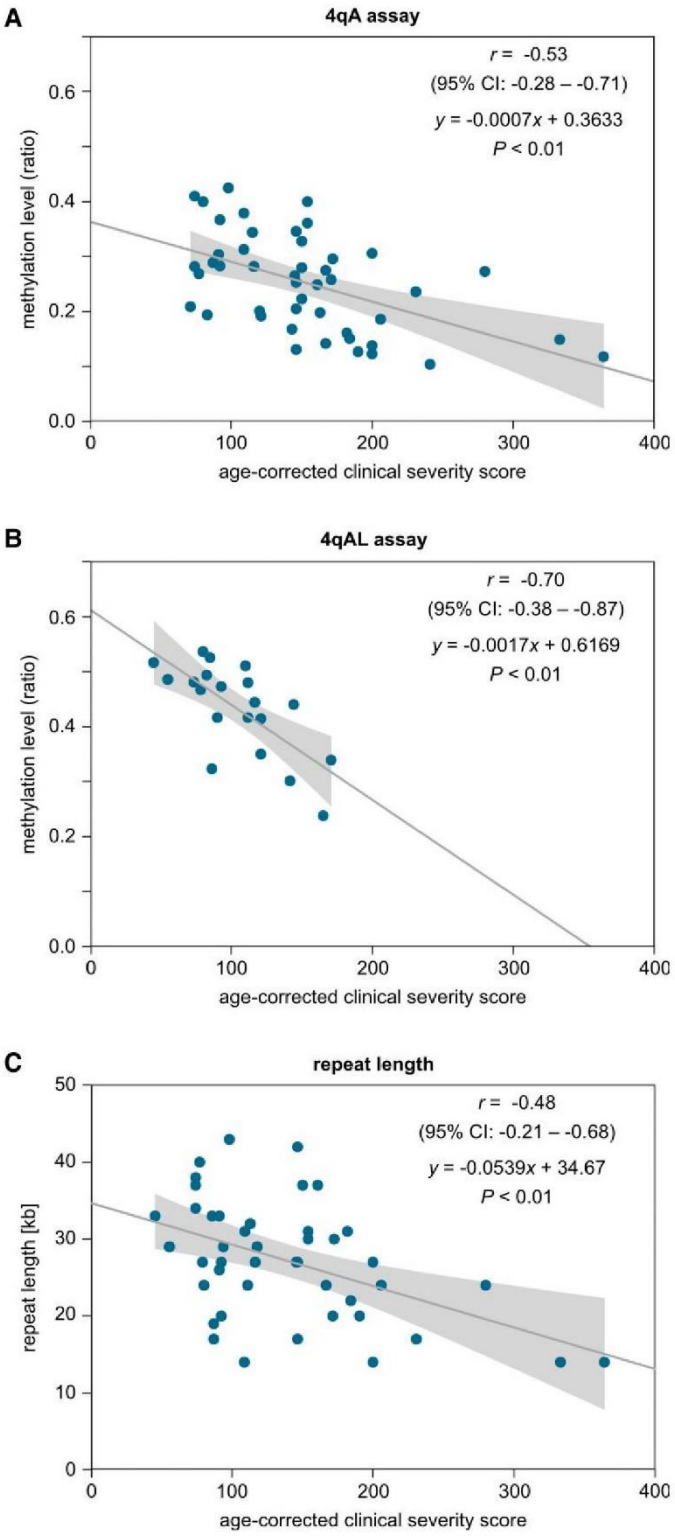

Figure 6 Correlation of clinical severity with genetic and epigenetic parameters. Correlation and linear regression of age-corrected clinical severity with methylation level (**A**) of the 4qA haplotype, (**B**) of the 4qAL haplotype, as well as (**C**) with the repeat length with respective Pearson's correlation coefficients, their 95% CIs and *P*-values. Highlighted areas around the regression lines indicate 95% CI of the regression.

the reduction of laboratory effort, which enables a high throughput analysis, the sequencing depth is increased by at least 100× to ensure a statistically sufficient representation of the regions. Thus, accuracy of the methylation levels determined by our novel method is further increased. Compared to Southern blotting, the current standard in FSHD testing, only small amounts of DNA (~1.5 µg)

are required, which can be extracted from frozen blood and do not need to be of high molecular weight.

A cohort of 56 individuals (establishment cohort) with 32 FSHD patients and 24 healthy controls allowed us to define thresholds for both distal methylation level (based on two methylation assays covering CpGs within the most distal repeat unit of the haplotypes 4qA and 4qAL, respectively[45]) and the global methylation level (based on a DUX4 assay covering CpGs within each D4Z4 repeat unit of chromosome 4; Supplementary Table 3). While isolated distal hypomethylation indicates FSHD1, additional global hypomethylation indicates FSHD2.

To further analyse FSHD-MPA in a diagnostic setting, we established a multistage diagnostic workflow (Fig. 1) that consisted of a haplotype analysis (step 1) followed by FSHD-MPA (step 2) and subsequently evaluated a larger cohort of 148 patients with clinically suspected FSHD (diagnostic cohort; Fig. 3).

FSHD-MPA reliably identifies FSHD patients

Based on the defined thresholds, methylation profiles reliably allowed for the diagnosis of FSHD. All 35 patients with available positive result for FSHD based on genetic parameters (confirmed D4Z4 repeat contraction in FSHD1, causative epigenetic suppressor gene variant in FSHD 2) were detected by FSHD-MPA based on their methylation profile. In seven patients who were tested negative by FSHD MPA, this result was confirmed as further diagnostic testing revealed that the diagnosis of a different neuromuscular disorder was likely (Table 2). Thus, FSHD-MPA was confirmed to be a reliable diagnostic tool to confirm or exclude the diagnosis of FSHD.

FSHD-MPA identifies FSHD patients that otherwise might be missed

FSHD-MPA can detect positive cases of FSHD that might have been missed by established methods. It is known that in a proportion of FSHD2 patients, it is not possible to identify a pathogenic variant in the causative genes known to date.[34] Two patients positive for FSHD2 in FSHD-MPA would have been tested negative based on their uncontracted D4Z4 array and the absence of pathogenic variants in *SMCHD1*, *DMT3B* and *LRIF1*. Although FSHD2 cannot be confirmed, the clinical data of both patients strongly support this diagnosis. Four of seven variants in *SMCHD1* were classified as variants of uncertain significance (Table 1). Only together with the positive FSHD-MPA, a pathogenicity of these variants and the diagnosis of FSHD2 is further supported. In general, FSHD2 is a rare disease compared to FSHD1. Consequently, the number of FSHD2 patients in this study was limited. Further work should be focused on a larger group of FSHD2 patients, to reassure the sufficient detection of this condition.

A few percent of FSHD patients carry complex rearrangements such as 4q-10q translocations, p13-E11 deletion and other non-canonical variants that might escape Southern blotting and can only be resolved by molecular combing or single-molecule optical mapping.[41,63,64] From a conceptual perspective, FSHD-MPA is able to diagnose FSHD in these patients, because a complete open reading frame of *DUX4* is a prerequisite for FSHD and the 4q/4qAL assay is targeted to this region. However, there might be very complex structural variants, and further studies need to be carried out for experimental confirmation that FSHD-MPA recognizes FSHD in these patients.

FSHD-MPA identifies FSHD1 and FSHD2 simultaneously and indicates an epigenetic overlap

In contrast to diagnostics based on genetic parameters that detect either FSHD1 based on D4Z4 repeat length or FSHD2 based on sequencing of epigenetic suppressor genes, FSHD-MPA makes it possible to detect both FSHD subtypes simultaneously. This is a diagnostic advantage; however, the distinction between patients with a repeat contraction and pathogenic variants in epigenetic suppressor genes is not consistently predicted by FSHD-MPA. As such, 5 of 14 patients with global hypomethylation indicating FSHD2 within the diagnostic cohort resembled the genetic picture of FSHD1 because contracted D4Z4 repeat arrays in the absence of likely pathogenic variants in *SMCHD1* and *DNMT3B* were identified. A technical artefact, e.g. artificial lowering of global methylation in the presence of a very short D4Z4 repeat and strong distal hypomethylation, seems unlikely in these cases because the contracted D4Z4 repeats are at the upper size range of FSHD1 (Supplementary Table 8, patients B8–B12) and distal methylation is only moderately reduced. Rather, our results suggest the presence of additional, yet unknown parameters influencing the methylation status of the FSHD locus. This is especially illustrated by one patient (B8), who carries one contracted and one uncontracted permissive allele with different haplotypes (4qA and 4qAL). However, instead of showing a methylation profile with monoallelic distal hypomethylation, the patient revealed hypomethylation within all three assays (4qA, 4qAL and DUX4). The biallelic global hypomethylation was not explained by pathogenic variants in known epigenetic suppressor genes. It is known that FSHD patients can have genetic features of both FSHD1 (repeat contraction) and FSHD2 (variant in epigenetic suppressor genes), as did one patient (B5) in our study.[65,66] Likewise, some FSHD2 patients with global hypomethylation are known to have neither a pathogenic variant in epigenetic suppressor genes nor a repeat contraction. Analogously, it is likely that there are some patients with repeat contraction and without variants in known epigenetic suppressor genes who have global hypomethylation corresponding to epigenetic FSHD2 for yet unknown reasons. These are likely to be frequently overlooked, as diagnosis usually ends once a repeat contraction is detected. Overall, our findings are in line with the hypothesis that FSHD forms a molecular disease spectrum where the genetic diagnosis of FSHD1 and FSHD2 represents two extremes of a epigenetic continuum.[66] Despite this epigenetic overlap, delineation of the two entities remains an important basis for genetic counselling, consensus scales of severity and classification of patients in clinical care.

Some FSHD-MPA results remain inconclusive

To a certain extent, FSHD-MPA revealed inconclusive results predominantly showing mild distal hypomethylation compatible with FSHD1 but above the diagnostic cut-off. Southern blotting confirmed FSHD1 in the majority of patients (5 of 7 patients) carrying a 4qA haplotype, while the diagnosis was excluded in the majority of patients (3 of 4 patients) carrying a 4qAL haplotype. Although FSHD-MPA can distinguish permissive and non-permissive haplotypes, it cannot distinguish homozygous permissive alleles. Given the high prevalence of the 4qA haplotype of up to 40% in the European population,[63,67] it is likely that hypomethylation of one allele is diluted by a hypermethylated non-pathogenic allele leading to an inconclusive result. Most challenging is the interpretation of six patients showing global and distal hypomethylation in the inconclusive range. In these cases, diagnostic evaluation of

epigenetic data requires careful consideration of clinical and genetic findings. The diagnostic precision of FSHD-MPA could be further increased by a larger establishment cohort and independent methylation profile analysis of two permissive homozygous alleles.

Distal methylation level as a biomarker for disease severity

Our study not only demonstrated that the methylation profiles of the D4Z4 repeat array are a precise diagnostic parameter, but also a biomarker for FSHD severity and a prognostic parameter for age at disease onset. In general, distal methylation level (4qA or 4qAL assay) showed stronger correlation with clinical parameters than D4Z4 repeat length (Figs 5 and 6). The correlations of both repeat length and distal methylation with age at disease onset are weaker than the correlations with age-corrected clinical severity. This is likely a consequence of the approximate survey of age at disease onset, which may also cause a non-significance for the correlations (*P*-values higher than 5%) of this parameter with the distal 4qAL methylation level and repeat length for the given sample size, respectively. The weaker correlation of clinical parameters with distal methylation in the 4qA assay compared with the 4qAL assay likely results from the higher prevalence of the 4qA haplotype in the general population and thus the higher likelihood of homozygous individuals in which only one of the parental alleles is hypomethylated.[67] Consequently, in the 4qA assay, the subset of these individuals and their higher average distal methylation level attenuates the correlation. To conclude, distal methylation level is a more reliable parameter compared to D4Z4 repeat size. It is also more universal as it accounts for FSHD1 as well as for FSHD2 patients and additionally distinguishes asymptomatic carriers of contracted alleles from affected ones.

A striking phenotypic variability was observed in a three-generation family with FSHD1 (Fig. 4, patients Z17–Z20). Four family members were carrying a permissive and contracted allele with two repeat units. All individuals would have been diagnosed with FSHD1 based on repeat analysis. However, the grandfather was clinically unaffected, which could be explained by a somatic mosaicism of the repeat contraction with uncontracted D4Z4 alleles not being resolved by Southern blotting or a penetrance defect.[68,69] Independent of a possible somatic mosaicism FSHD-MPA was able to differentiate healthy individuals from clinically affected ones in this family. This family and the patients tested predictively are indicative that methylation—measured by the FSHD-MPA—is not only a marker of disease severity, but also potentially an important prognostic marker, and for the first time might allow accurate predictive testing for FSHD. To verify this, longitudinal studies are needed to rule out the possibility that FSHD manifests later in life and that the methylation profiles determined by FSHD-MPA are stable over the lifetime. This is suggested by the fact that age-corrected clinical severity, rather than unadjusted clinical severity, shows a high correlation with distal methylation.

The fact that thresholds for pathogenic methylation levels in the 4qA and 4qAL assays could be independently derived from correlation analysis underlines that the epigenetic parameter is the main representation of disease status and outperforms repeat length. As such, it may reflect *DUX4* expression, although it cannot be distinguished whether it is directly associated with it or just another consequence of epigenetic derepression of the FSHD genetic locus due to the only partly understood FSHD pathomechanism. Repeat contractions, variants in *SMCHD1*, *DMT3B* and *LRIF1* and yet unknown factors might rather be disease modifiers acting on the epigenetic structure of the locus than being disease-causing by themselves. As such, patients genetically diagnosed with FSHD1 and FSHD2 show an epigenetic overlap as observed in our and other studies.[65,66] Additionally, this explains why individuals carrying repeat-contractions on permissive alleles or having pathogenic variants within *SMCHD1* in combination with large D4Z4 arrays are healthy and show a hypermethylated FSHD locus.[38,41,66] Because epigenetic patterns are not inherited in a Mendelian manner, this is also consistent with variable disease manifestations in relatives carrying the same genetic features.[70]

The results of our study refute the recent questioning of the importance of methylation in the diagnosis and pathogenesis of FSHD.[71,72] Contrary results within different epigenetic tests are likely the consequence of unspecific amplification of regions other than those associated with the disease and do not reflect irrelevancy of methylation as a diagnostic parameter.[48] A standardization and international harmonization of diagnostic parameters with respect to the region analysed and method used needs to be established in order to avoid further controversy and confusion for genetic counsellors, clinicians and patients.

In summary, we implemented an NGS-based bisulphite sequencing reaction method (FSHD-MPA) to determine the average and the distal methylation level of the D4Z4 repeat array of chromosome 4q35. We demonstrate that the method reliably identifies individuals affected by FSHD and overcomes current limitations of genetic testing. Additionally, we have verified methylation levels in the D4Z4 distal repeat as the most accurate biomarker for disease severity and have shown that epigenetic rather than genetic parameters determine disease status. Novel long-read sequencing or optical genome mapping technologies may further refine diagnosis and improve prognostic yield by separately analysing the methylation of two alleles with the same haplotype and by assessing genetic and epigenetic parameters at the same time.[73,74]

Acknowledgements

We thank Christina Rapp for her experimental work to establish the method, Ariane Hallermayr for her assistance in analysing data and Jakob Koegst as well as Karl Akbari for advices regarding statistical analysis. H.E. is grateful to the Faculty of Medicine of the Ludwig-Maximilians-Universität München for a fellowship. This study is part of the medical thesis of H.E.

Funding

The study was funded by the Medical Genetics Center without external funding.

Competing interests

There are no conflicts of interests for any of the authors.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Deenen JCW, Arnts H, van der Maarel SM, *et al.* Population-based incidence and prevalence of facioscapulohumeral dystrophy. *Neurology*. 2014;83:1056-1059.

2. Padberg GW, Lunt PW, Koch M, Fardeau M. Diagnostic criteria for facioscapulohumeral muscular dystrophy. *Neuromuscul Disord*. 1991;1:231-234.
3. Mul K, Lassche S, Voermans NC, Padberg GW, Horlings CG, van Engelen BG. What's in a name? The clinical features of facioscapulohumeral muscular dystrophy. *Pract Neurol*. 2016;16:201-207.
4. Ricci G, Ruggiero L, Vercelli L, *et al.* A novel clinical tool to classify facioscapulohumeral muscular dystrophy phenotypes. *J Neurol*. 2016;263:1204-1214.
5. Kottlors M, Kress W, Meng G, Glocker FX. Facioscapulohumeral muscular dystrophy presenting with isolated axial myopathy and bent spine syndrome. *Muscle Nerve*. 2010;42:273-275.
6. Ricci G, Mele F, Govi M, *et al.* Large genotype–phenotype study in carriers of D4Z4 borderline alleles provides guidance for facioscapulohumeral muscular dystrophy diagnosis. *Sci Rep*. 2020;10:21648.
7. Ricci G, Scionti I, Sera F, *et al.* Large scale genotype–phenotype analyses indicate that novel prognostic tools are required for families with facioscapulohumeral muscular dystrophy. *Brain*. 2013;136:3408-3417.
8. Goselink RJM, Voermans NC, Okkersen K, *et al.* Early onset facioscapulohumeral dystrophy—A systematic review using individual patient data. *Neuromuscul Disord*. 2017;27:1077-1083.
9. Zatz M, Marie SK, Passos-Bueno MR, *et al.* High proportion of new mutations and possible anticipation in Brazilian facioscapulohumeral muscular dystrophy families. *Am J Hum Genet*. 1995;56:99-105.
10. Tawil R, Storvick D, Feasby TE, Weiffenbach B, Griggs RC. Extreme variability of expression in monozygotic twins with FSH muscular dystrophy. *Neurology*. 1993;43:345.
11. Ruggiero L, Mele F, Manganelli F, *et al.* Phenotypic variability among patients with D4Z4 reduced allele facioscapulohumeral muscular dystrophy. *JAMA Netw*. 2020;3:e204040.
12. Mostacciuolo M, Pastorello E, Vazza G, *et al.* Facioscapulohumeral muscular dystrophy: Epidemiological and molecular study in a north-east Italian population sample. *Clin Genet*. 2009;75:550-555.
13. Lunt PW, Harper PS. Genetic counselling in facioscapulohumeral muscular dystrophy. *J Med Genet*. 1991;28:655-664.
14. Ansseau E, Vanderplanck C, Wauters A, Harper SQ, Coppée F, Belayew A. Antisense oligonucleotides used to target the DUX4 mRNA as therapeutic approaches in FacioScapuloHumeral muscular dystrophy (FSHD). *Genes (Basel)*. 2017;8:93.
15. Lim KRQ, Bittel A, Maruyama R, *et al.* DUX4 Transcript knockdown with antisense 2′-*O*-methoxyethyl gapmers for the treatment of facioscapulohumeral muscular dystrophy. *Mol Ther*. 2021;29:848-858.
16. Marsollier AC, Ciszewski L, Mariot V, *et al.* Antisense targeting of 3′ end elements involved in *DUX4* mRNA processing is an efficient therapeutic strategy for facioscapulohumeral dystrophy: A new gene-silencing approach. *Hum Mol Genet*. 2016;25:1468-1478.
17. Rashnonejad A, Amini-Chermahini G, Taylor NK, Wein N, Harper SQ. Designed U7 snRNAs inhibit DUX4 expression and improve FSHD-associated outcomes in DUX4 overexpressing cells and FSHD patient myotubes. *Mol Ther Nucleic Acids*. 2020;23:476-486.
18. Sacconi S, Salviati L, Bourget I, *et al.* Diagnostic challenges in facioscapulohumeral muscular dystrophy. *Neurology*. 2006;67:1464-1466.
19. Ricci G, Zatz M, Tupler R. Facioscapulohumeral muscular dystrophy: More complex than it appears. *Curr Mol Med*. 2014;14:1052-1068.
20. Scionti I, Greco F, Ricci G, *et al.* Large-scale population analysis challenges the current criteria for the molecular diagnosis of fascioscapulohumeral muscular dystrophy. *Am J Hum Genet*. 2012;90:628-635.
21. Ricci E, Galluzzi G, Deidda G, *et al.* Progress in the molecular diagnosis of facioscapulohumeral muscular dystrophy and correlation between the number of KpnI repeats at the 4q35 locus and clinical phenotype. *Ann Neurol*. 1999;45:751-757.
22. Wohlgemuth M, Lemmers RJ, Jonker M, *et al.* A family-based study into penetrance in facioscapulohumeral muscular dystrophy type 1. *Neurology*. 2018;91:e444-e454.
23. Lemmers Richard JLF, van der Vliet Patrick J, Rinse K, *et al.* A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science*. 2010;329:1650-1653.
24. Snider L, Asawachaicharn A, Tyler AE, *et al.* RNA Transcripts, miRNA-sized fragments and proteins produced from D4Z4 units: New candidates for the pathophysiology of facioscapulohumeral dystrophy. *Hum Mol Genet*. 2009;18:2414-2430.
25. Greco A, Goossens R, van Engelen B, van der Maarel SM. Consequences of epigenetic derepression in facioscapulohumeral muscular dystrophy. *Clin Genet*. 2020;97:799-814.
26. Bosnakovski D, Chan SSK, Recht OO, *et al.* Muscle pathology from stochastic low level DUX4 expression in an FSHD mouse model. *Nat Commun*. 2017;8:550.
27. Geng LN, Yao Z, Snider L, *et al.* DUX4 Activates germline genes, retroelements, and immune mediators: Implications for facioscapulohumeral dystrophy. *Dev Cell*. 2012;22:38-51.
28. Lemmers RJLF, Wohlgemuth M, van der Gaag KJ, *et al.* Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. *Am J Hum Genet*. 2007;81:884-894.
29. Wijmenga C, Hewitt JE, Sandkuijl LA, *et al.* Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nat Genet*. 1992;2:26-30.
30. Deutekom JCTV, Wljmenga C, Tlenhoven EAEV, *et al.* FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. *Hum Mol Genet*. 1993;2:2037-2042.
31. Preston MK, Tawil R, Wang LH. *Facioscapulohumeral Muscular Dystrophy. GeneReviews® [Internet]*. Accessed 2 February 2022. University of Washington, Seattle; 1993–2022. https://www.ncbi.nlm.nih.gov/books/NBK1443/
32. Lemmers RJLF, Tawil R, Petek LM, *et al.* Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat Genet*. 2012;44:1370-1374.
33. van den Boogaard ML, Lemmers RJLF, 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;98:1020-1029.
34. Hamanaka K, Šikrová D, Mitsuhashi S, *et al.* Homozygous nonsense variant in *LRIF1* associated with facioscapulohumeral muscular dystrophy. *Neurology*. 2020;94:e2441-e2447.
35. Lemmers RJLF, De Kievit P, van geel M, *et al.* Complete allele information in the diagnosis of facioscapulohumeral muscular dystrophy by triple DNA analysis. *Ann Neurol*. 2001;50:816-819.
36. Jia FF, Drew AP, Nicholson GA, Corbett A, Kumar KR. Facioscapulohumeral muscular dystrophy type 2: An update on the clinical, genetic, and molecular findings. *Neuromuscul Disord*. 2021;31:1101-1112.
37. de Greef JC, Lemmers RJLF, van Engelen BGM, *et al.* Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. *Hum Mutat*. 2009;30:1449-1459.

38. Lemmers RJLF, van der Vliet PJ, Vreijling JP, *et al.* Cis D4Z4 repeat duplications associated with facioscapulohumeral muscular dystrophy type 2. *Hum Mol Genet*. 2018;27:3488-3497.
39. Lemmers RJLF, O'Shea S, Padberg GW, Lunt PW, van der Maarel SM. Best practice guidelines on genetic diagnostics of facioscapulohumeral muscular dystrophy: Workshop 9th June 2010, LUMC, Leiden, the Netherlands. *Neuromuscul Disord*. 2012;22:463-470.
40. Nguyen K, Walrafen P, Bernard R, *et al.* Molecular combing reveals allelic combinations in facioscapulohumeral dystrophy. *Ann Neurol*. 2011;70:627-633.
41. Nguyen K, Puppo F, Roche S, *et al.* Molecular combing reveals complex 4q35 rearrangements in facioscapulohumeral dystrophy. *Hum Mutat*. 2017;38:1432-1441.
42. 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;23:1506-1514.
43. Gaillard MC, Roche S, Dion C, *et al.* Differential DNA methylation of the D4Z4 repeat in patients with FSHD and asymptomatic carriers. *Neurology*. 2014;83:733-742.
44. Hartweck LM, Anderson LJ, Lemmers RJ, *et al.* A focal domain of extreme demethylation within D4Z4 in FSHD2. *Neurology*. 2013; 80:392-399.
45. Jones TI, Yan C, Sapp PC, *et al.* Identifying diagnostic DNA methylation profiles for facioscapulohumeral muscular dystrophy in blood and saliva using bisulfite sequencing. *Clin Epigenetics*. 2014;6:23.
46. Hewitt JE, Lyle R, Clark LN, *et al.* Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystrophy. *Hum Mol Genet*. 1994;3:1287-1295.
47. van Overveld PGM, Lemmers RJFL, Sandkuijl LA, *et al.* Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. *Nat Genet*. 2003;35:315-317.
48. Gould T, Jones TI, Jones PL. Precise epigenetic analysis using targeted bisulfite genomic sequencing distinguishes FSHD1, FSHD2, and healthy subjects. *Diagnostics*. 2021;11:1469.
49. Lunt PW, Jardine PE, Koch MC, *et al.* Correlation between fragment size at D4F104S1 and age at onset or at wheelchair use, with a possible generational effect, accounts for much phenotypic variation in 4q35-facioscapulohumeral muscular dystrophy (FSHD). *Hum Mol Genet*. 1995;4:951-958.
50. Tawil R, Forrester J, Griggs RC, *et al.* Evidence for anticipation and association of deletion size with severity in facioscapulohumerd muscular dystrophy. *Ann Neurol*. 1996;39:744-748.
51. Mul K, Voermans NC, Lemmers RJLF, *et al.* Phenotype–genotype relations in facioscapulohumeral muscular dystrophy type 1. *Clin Genet*. 2018;94:521-527.
52. Katz NK, Hogan J, Delbango R, Cernik C, Tawil R, Statland JM. Predictors of functional outcomes in patients with facioscapulohumeral muscular dystrophy. *Brain*. 2021;144:3451-3460.
53. Jones TI, King OD, Himeda CL, *et al.* Individual epigenetic status of the pathogenic D4Z4 macrosatellite correlates with disease in facioscapulohumeral muscular dystrophy. *Clin Epigenetics*. 2015;7:37.
54. Lim KRQ, Yokota T. Genetic approaches for the treatment of facioscapulohumeral muscular dystrophy. *Front Pharmacol*. 2021; 12:642858.
55. Van Overveld PGM, Enthoven L, Ricci E, *et al.* Variable hypomethylation of D4Z4 in facioscapulohumeral muscular dystrophy. *Ann Neurol*. 2005;58:569-576.
56. Tsumagari K, Chen D, Hackman JR, Bossler AD, Ehrlich M. FSH dystrophy and a subtelomeric 4q haplotype: A new assay and associations with disease. *J Med Genet*. 2010;47:745-751.
57. Richards S, Aziz N, Bale S, *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17:405-423.
58. Nykamp K, Anderson M, Powers M, *et al.* Sherloc: A comprehensive refinement of the ACMG–AMP variant classification criteria. *Genet Med*. 2017;19:1105-1117.
59. Dixit M, Ansseau E, Tassin A, *et al.* *DUX4*, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of *PITX1*. *Proc Natl Acad Sci*. 2007;104: 18157-18162.
60. Jones TI, Chen JCJ, Rahimov F, *et al.* Facioscapulohumeral muscular dystrophy family studies of DUX4 expression: evidence for disease modifiers and a quantitative model of pathogenesis. *Hum Mol Genet*. 2012;21:4419-4430.
61. Snider L, Geng LN, Lemmers RJLF, *et al.* Facioscapulohumeral dystrophy: Incomplete suppression of a retrotransposed gene. *PLOS Genet*. 2010;6:e1001181.
62. Signorelli M, Mason AG, Mul K, *et al.* Evaluation of blood gene expression levels in facioscapulohumeral muscular dystrophy patients. *Sci Rep*. 2020;10:17547.
63. Lemmers RJLF, van der Vliet PJ, van der Gaag KJ, *et al.* Worldwide population analysis of the 4q and 10q subtelomeres identifies only four discrete interchromosomal sequence transfers in human evolution. *Am J Hum Genet*. 2010;86:364-377.
64. van Deutekom JCT, Bakker E, Lemmers RJLF, *et al.* Evidence for subtelomeric exchange of 3.3 kb tandemly repeated units between chromosomes 4q35 and 10q26: implications for genetic counselling and etiology of FSHD1. *Hum Mol Genet*. 1996;5: 1997-2003.
65. Larsen M, Rost S, El Hajj N, *et al.* Diagnostic approach for FSHD revisited: *SMCHD1* mutations cause FSHD2 and act as modifiers of disease severity in FSHD1. *Eur J Hum Genet*. 2015; 23:808-816.
66. Sacconi S, Briand-Suleau A, Gros M, *et al.* FSHD1 and FSHD2 form a disease continuum. *Neurology*. 2019;92:e2273-e2285.
67. Lemmers RJ, van der Vliet PJ, Balog J, *et al.* Deep characterization of a common D4Z4 variant identifies biallelic DUX4 expression as a modifier for disease penetrance in FSHD2. *Eur J Hum Genet*. 2018;26:94-106.
68. van der Maarel SM, Deidda G, Lemmers RJLF, *et al.* *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*. 2000;66:26-35.
69. Lemmers RJLF, van der Wielen MJR, Bakker E, Padberg GW, Frants RR, van der Maarel SM. Somatic mosaicism in FSHD often goes undetected. *Ann Neurol*. 2004;55:845-850.
70. Calandra P, Cascino I, Lemmers RJLF, *et al.* Allele-specific DNA hypomethylation characterises FSHD1 and FSHD2. *J Med Genet*. 2016;53:348-355.
71. Salsi V, Magdinier F, Tupler R. Does DNA methylation matter in FSHD? *Genes (Basel)*. 2020;11:258.
72. Nikolic A, Jones TI, Govi M, *et al.* Interpretation of the epigenetic signature of facioscapulohumeral muscular dystrophy in light of genotype–phenotype studies. *Int J Mol Sci*. 2020;21:2635.
73. Sharim H, Grunwald A, Gabrieli T, *et al.* Long-read single-molecule maps of the functional methylome. *Genome Res*. 2019;29:646-656.
74. Mitsuhashi S, Nakagawa S, Takahashi Ueda M, Imanishi T, Frith MC, Mitsuhashi H. Nanopore-based single molecule sequencing of the D4Z4 array responsible for facioscapulohumeral muscular dystrophy. *Sci Rep*. 2017;7:14789.

Supplementary Material

Extraction of genomic DNA (gDNA)

Genomic DNA (gDNA) was obtained from total peripheral EDTA blood samples by extraction of white blood cells with a Biomek FX system (Beckman Coulter) using the NucleoMag® Blood 3 ml Kit (Machery-Nagel, #REF 744502.1) or by manual extraction by the Qiagen FlexiGene DNA Kit as by manufacturer's instructions. All DNA samples showed high purity as determined by optical density measurements (A260/A280 > 1.9 and A260/A230 > 2.0).

Southern Blotting for D4Z4 Repeat Length Analysis

DNA was digested with *Eco*RI, *Eco*RI + BlnI and *Apo*I (= Isoschizomer to *Xap*I) (*Eco*RI, *Apo*I: New England Biolabs, Frankfurt, Germany; *Bln*I: Takara Company Europe GmbH, Frankfurt, Germany) and separated by electrophorese for 40 h at 1,2 V/cm on 0,5 % agarose gels in 1 x TAE buffer. DNA was transferred to Hybond XL membranes (GE Healthcare, Freiburg, Germany) and hybridized with radioactively labelled probe p13E-11. Bands were than visualized by autoradiography. Fragment lengths were determined by comparison to a standard marker included in each run (λ DNA-Mono Cut Mix; New England Biolabs, Frankfurt, Germany). Only fragments smaller than 48 kb after digestion with *Eco*RI were considered for further interpretation. Fragments, being reduced by 3 kb after additional cleavage with *Bln*I and not being detectable after cleavage with + *Xap*I or, *Apo*I, respectively, were regarded as 4q35-type fragments. Fragments, being reduced by 5 kb after cleavage with *Xap*I or *Apo*I, respectively and not being detectable after cleavage with *Eco*RI + BlnI were regarded as 10q26-type fragments.[39]

Supplementary Table 1. Sequences of primers used for amplifying BSS converted DNA in the 4qA, 4qAL and DUX4 assay. R represents adenine or guanine. For each variable position there is one primer containing adenine and one containing guanine.

4qA assay	**Sequence 5'→3'**
1. PCR	
BSS1438F	GTTTTGTTGGAGGAGTTTTAGGA
BSS3742R	AACATTCAACCAAAATTTCACRAAA
2. PCR	
BSS1438F	GTTTTGTTGGAGGAGTTTTAGGA
BSS3626R	AACAAAAATATACTTTTAACCRCCAAAAA
4qAL	
1. PCR	
BSS4qALF	TTATTTATGAAGGGGTGGAGTTTGTT
BSS3742R	AACATTCAACCAAAATTTCACRAAA
2. PCR	
BSS4qALF	TTATTTATGAAGGGGTGGAGTTTGTT
BSS3626R	AACAAAAATATACTTTTAACCRCCAAAAA
DUX4	
1. PCR	
BSS167F	TTTTGGGTTGGGTGGAGATTTT
BSS1036R	AACACCRTACCRAACTTACACCCTT
2. PCR	
BSS475F	TTAGGAGGGAGGGAGGGAGGTAG
BSS1036R	AACACCRTACCRAACTTACACCCTT

Supplementary Table 2. Sequences of the nested PCR products amplified in the 4qA, 4qAL and DUX4 FSHD-MPA assay.

Fragment	Sequence 5'→3'
4qA	GTTTTGTTGGAGGAGTTTTAGGACGCGGGGTTGGGACGGGGTCGGGTGGTTCGGGGTAGGG CGGTGGTTTTTTTTCGCGGGGAATATTTGGTTGGTTACGGAGGGGCGTGTTTTCGTTTCGTTT TTTTATCGGGTTGATCGGTTTGGGATTTTTGTTTTTTAGGTTTAGGTTCGGTGAGAGATTTTA TATCGCGGAGAATTGTTATTTTTTTTTGGGTATTTCGGGGATTTTAGAGTCGGTTTAGGTATT AGTAGGTGGGTCGTTTATTGCGTACGCGCGGGTTTGCGGGTAGTCGTTTGGGTTGTGGGAGT AGTTCGGGTAGAGTTTTTTTGTTTTTTTATTAGTTTATTTCGTCGTTTGATCGTTTTTTTTTTA TTTTTATTTTTTATTTTCGGAAAACGCGTCGTTTTTTGGGTTGGGTGGAGATTTTCGTTTCGC GAAATATCGGGTTTCGCGTAGCGTTCGGGTTTGATATCGTTTCGGCGGTTCGTTTTTTTTGCG TTTTCGCGTTATCGTCGTTCGTTCGTTCGGGTTTTTGTAGTTTTTTAGTTGTTAGCGCGGAGT TTTTGGCGGTTAAAAGTATATTTTTGTT
4qAL	TTATTTATGAAGGGGTGGAGTTTGTTTGTTTGTGGGTTTTTATAAGGGCGGTTGGTTGGTTG GTTGGTTGTTCGGGTAGGTTTTTTGGTTGTATTTGTCGTAGTGTATAGTTCGGTTGAGGTGTA CGGGAGTTCGTCGGTTTTTTTTTGTTCGCGTTCGTTCGTGAAATTTCGGTCGGGGTTTATCGC GATGGTTTTTTCGATATTTTCGGATAGTATTTTTTTCGCGGAAGTTCGGGGACGAGGACGGC GACGGAGATTCGTTTGGATTTCGAGTTAAAGCGAGGTTTTGCGAGTTTGTAGTTTTTTAGTT GTTAGCGCGGAGTTTTTGGCGGTTAAAAGTATATTTTTGTT
DUX4	TTAGGAGGGAGGGAGGGAGGTAGGGAGGTAGGGAGGAACGGAGGGAAAGATAGAGCGAC GTAGGGATTGGGGGCGGGCGGGAGGGAGTCGGGGACGGGGGGAGGAAGGTAGGGAGGAA AAGCGGTTTTCGGTTTTCGGGAGTAGCGGGATTTTCGTTTTTCGGGAAAACGGTTAGCGTTC GGCGCGGGTTGAGGGTTGGGTTTATAGTCGTCGCGTCGGTCGGCGGGGTATTATTTATTCGT TTCGGTTTCGTGGTTTAGGGAGTGGGCGGTTTTTTTCGGGATAAAAGATCGGGATTCGGGTT GTCGTCGGGTTTTTATTCGCGCGGTTTATAGATCGTATATTTTTAGGTTGAGTTTTGTAACGC GGCGCGAGGTCGATAGTTTCGGTTACGGAGGAGTTATACGTAGGACGACGGAGGCGTGATT TTGGTTTTCGCGTGGTTTTGTTTTTCGTAAGGCGGTTTGTTGTTTACGTTTTTTCGGTTTTCGA AAGGTTGGTTATGTCGATTGTTTGTTTTCGGAGTTTTGCGGGTATTCGGAAATATGTAGGGA AGGGTGTAAGTTCGGTACGGTGTT

Supplementary Table 3. Cutoff values for hypomethylation (positive , negative, and inconclusive) for the 4qA, 4qAL and DUX4 FSHD-MPA assay.

Assay	positive	inconclusive	negative
4qA	< 0.362	0.362–0.441	> 0.441
4qAL	< 0.568	0.568–0.653	> 0.653
DUX4	< 0.409	0.409–0.446	> 0.446

Supplementary Table 4. Establishment cohort: Patients defined as FSHD1 based on genetic parameters (permissive haplotype, D4Z4 repeat size reduction < 12 RU) with their epigenetic parameters based on FSHD-MPA. If determined: Result of additional sequencing of *SMCHD1* and *DNMT3B*. Highlighted in green: hypermethylation, highlighted in yellow: inconclusive methylation, highlighted in red: pathogenic hypomethylation.

Patient ID	Permissive Haplo-type(s)	D4Z4 Repeat Size of contracted allele [kb (RU)]	Methylation level by FSHD-MPA			Result of additional *SMCHD1*, *DNMT3B* sequencing
			4qA assay	4qAL assay	DUX4 assay	
VPA1	4qA	27 (6)	0.1594	/	0.4740	/
VPA2	4qA	24 (5)	0.1116	/	0.4962	/
VPA3	4qA	13 (2)	0.0792	/	0.4494	/
VPA4	4qA	38 (10)	0.3069	/	0.4710	/
VPA5	4qA, 4qAL	33 (8)	0.5910	0.5151	0.5664	/
VPA6	4qA, 4qAL	9 (1)	0.5745	0.2590	0.5001	/
VPA7	4qA	24 (5)	0.1587	/	0.6331	negative
VPA8	4qA	23 (5)	0.1895	/	0.4528	/
VPA9	4qA	20 (4)	0.1742	/	0.3275	negative
VPA10	4qA	14 (2)	0.2172	/	0.5879	/
VPA11	4qA, 4qAL	28 (6)	0.3370	0.4281	0.4122	negative
VPA12	4qA, 4qAL	33 (8)	0.6539	0.5078	0.5477	/
VPA13	4qA, 4qAL	20 (4)	0.1338	0.6083	0.3763	negative
VPA14	4qA	27 (6)	0.1441	/	0.4982	/
VPA15	4qAL	15 (2)	/	0.4462	0.5950	/
VPA16	4qAL	30 (7)	/	0.5151	0.4518	/
VPA17	4qA	27 (6)	0.1651	/	0.4573	/
VPA18	4qA	20 (4)	0.3728	/	0.5049	/
VPA19	4qAL	30 (7)	/	0.3726	0.4649	/
VPA20	4qAL	25 (5)	/	0.5145	0.4729	/
VPA21	4qA	33 (8)	0.6033	/	0.5320	/
VPA22	4qA, 4qAL	40 (11)	0.7242	0.5824	0.6398	/
VPA23	4qA	20 (4)	0.1716	/	0.4979	negative
VPA24	4qA	32 (8)	0.1771	/	0.4514	/
VPA25	4qA	20 (4)	0.1157	/	0.4784	/
VPA26	4qA	30 (7)	0.2239	/	0.4531	/
VPA27	4qA	28 (6)	0.2646	/	0.4552	/

VPA28	4qA	38 (10)	0.5635	/	0.4837	/
VPA29	4qA	41 (11)	0.2386	/	0.4585	/

Supplementary Table 5. Establishment cohort: Patients defined as FSHD2 based on genetic parameters (permissive haplotype, pathogenic variant in SMCHD1) with their D4Z4 repeat size, variant in *SMCHD1* and epigenetic parameters based on FSHD-MPA. Highlighted in green: hypermethylation, highlighted in yellow: inconclusive methylation, highlighted in red: pathogenic hypomethylation.

Patient ID	Permissive Haplotype(s)	Methylation level by FSHD-MPA			D4Z4 Repeat Size [kb (RU)]	Pathogenic *SMCHD1* variants
		4qA assay	4qAL assay	DUX4 assay		
VPB1	4qA	0.1231	/	0.3561	38 (10)	c.4454C>T (p.Pro1485Leu)
VPB2	4qA	0.1862	/	0.4029	>48	c.866delA (p.Asn289Thrfs*23)
VPB3	4qAL	/	0.3886	0.2634	36 (9)	c.3274_3276+1del: p.Lys1092del

Supplementary Table 6. Establishment cohort: Unaffected individuals, their permissive haplotypes and their epigenetic parameters based on FSHD-MPA. Highlighted in green: hypermethylation. Highlighted in yellow: inconclusive methylation.

Patient ID	Haplotype(s)	Methylation level		
		4qA assay	4qAL assay	DUX4 assay
VN1	4qA	0.4731	/	0.4943
VN2	4qA	0.7133	/	0.7350
VN3	4qA	0.6391	/	0.6560
VN4	4qAL	/	0.7479	0.6274
VN5	4qA	0.7633	/	0.6855
VN6	4qAL	/	0.7498	0.5976
VN7	4qAL	/	0.7657	0.5414
VN8	4qAL	/	0.7170	0.5099
VN9	4qAL	/	0.6566	0.6373
VN10	4qA	0.5120	/	0.4752
VN11	4qAL	/	0.6883	0.6019
VN12	4qA	0.5246	/	0.6877
VN13	4qA	0.4217	/	0.5190
VN14	4qA	0.6215	/	0.5443
VN15	4qA	0.4861	/	0.6062
VN16	4qA	0.4469	/	0.4549
VN17	4qAL	/	0.7744	0.6384
VN18	4qA, 4qAL	0.5936	0.6811	0.5668

VN19	4qA	0.5838	/	0.5739
VN20	4qA	0.6664	/	0.5097
VN21	4qA	0.5720	/	0.4949
VN22	4qA	0.7507	/	0.5758
VN23	4qAL	/	0.7005	0.4649
VN24	4qAL	/	0.6779	0.4814

Supplementary Table 7. Diagnostic cohort: Patients diagnosed as FSHD1 based on epigenetic parameters of FSHD-MPA, their permissive haplotypes, their methylation level in the three different assays, their repeat length of the D4Z4 array determined by Southern Blotting, eventual results of *SMCHD1*/*DNMT3B* sequencing and clinical data (family history, age at disease onset, (age corrected) clinical severity score (CSS)). Symbols: + positive/present; - not present/negative; / not determined/not known. Highlighted in green: hypermethylation, highlighted in yellow: inconclusive methylation, highlighted in red: pathogenic hypomethylation.

Patient	Haplotype (4qA/4qA161)	Haplotype by FSHD-MPA	Methylation level by FSHD-MPA			*SMCHD1*/ *DNMT3B* sequencing	Repeat length 4q35 [kb (RU)]	Epigenetic diagnosis	Genetic diagnosis	Familiy history	Age at disease onset	CSS	Age at Assessment	Age-corrected CSS
			DUX4 assay	4qA assay	4qAL assay									
A1	+/+	4qA/4qAL	0.483	0.667	0.489	/	/	FSHD1	/	/	/	/	65	/
A2	+/+	4qA	0.454	0.273	-	/	24 (5)	FSHD1	FSHD1	+	0-20	3.5	25	280
A3	+/+	4qA	0.419	0.313	-	negative	31 (7)	FSHD1	FSHD1	+	40-60	3	55	109
A4	+/+	4qA	0.474	0.194	-	/	/	FSHD1	/	+	40-60	2.5	60	83
A5	+/+	4qA	0.445	0.205	-	/	27 (6)	FSHD1+	FSHD1	-	20-40	3	41	146
A6	+/+	4qA	0.452	0.361	-	negative	30 (7)	FSHD1	FSHD1	-	0-20	3	39	154
A7	+/+	4qA, 4qAL	0.536	0.536	0.239	negative	/	FSHD1	/	-	0-20	5	60	167
A8	+/+	4qA	0.568	0.269	-	/	40 (11)	FSHD1	FSHD1	-	0-20	3	78	77
A9	+/+	4qA, 4qAL	0.59	0.602	0.352	/	/	FSHD1	/	-	20-40	2.5	41	122
A10	+/+	4qA	0.491	0.344	-	negative	/	FSHD1	/	/	40-60	3.5	61	115
A11	+/+	4qA	0.597	0.131	-	negative	17 (3)	FSHD1	FSHD1	-	20-40	3	41	146
A12	+/+	4qA	0.566	0.127	-	negative	20 (4)	FSHD1	FSHD1	+	20-40	4	42	190
A13	+/+	4qA	0.624	0.151	-	/	22 (4)	FSHD1	FSHD1	+	20-40	3.5	38	184
A14	+/+	4qA	0.671	0.306	-	/	/	FSHD1	/	-	20-40	3	30	200
A15	+/+	4qA	0.47	0.283	-	/	27 (6)	FSHD1	FSHD1	-	20-40	3	65	92
A16	+/+	4qAL	0.626	-	0.433	/	/	FSHD1	/	/	/	/	82	/
A17	+/+	4qAL	0.501	-	0.54	/	/	FSHD1	/	-	40-60	2.5	62	81
A18	+/+	4qA	0.597	0.29	-	/	24 (5)	FSHD1	FSHD1	-	60-80	3.5	62	/
A19	+/+	4qA	0.445	0.28	-	/	/	FSHD1	/	-	20-40	3	40	150

A20	+/+	4qA	0.594	0.142	-	/	24 (5)	FSHD1	FSHD1	-	0-20	4	48	167
A21	+/+	4qAL	0.626		0.483	/	32 (8)	FSHD1	FSHD1	+	20-40	3.5	62	113
A22	+/+	4qA	0.557	0.235	-	/	/	FSHD1	/	-	/	/	39	/
A23	+/+	4qA/4qAL	0.662	0.249	0.722	/	37 (9)	FSHD1	FSHD1	-	0-20	4.5	56	161
A24	+/+	4qA	0.459	0.327+	-	/	31 (7)	FSHD1	FSHD1	+	/	/	22	/
A25	+/+	4qA/4qAL	0.58	0.236	0.71	/	17 (3)	FSHD1	FSHD1	+	0-20	3	26	231
A26	+/+	4qA	0.476	0.258	-	/	20 (4)	FSHD1	FSHD1	+	0-20	3	35	171
A27	+/+	4qA/4qAL	0.661	0.779	0.447	/	29 (7)	FSHD1	FSHD1	+	40-60	3	51	118
A28	+/+	4qA, 4qAL	0.511	0.653	0.47	/	27 (6)	FSHD1	FSHD1	+	60-80	3	76	79
A29	+/+	4qAL	0.518	-	0.489	/	29 (7), 45(13)	FSHD1	FSHD1	+	40-60	1.5	54	56
A30	+/+	4qA	0.472	0.253	-	/	/	FSHD1	/	-	20-40	3	41	146
A31	+/+	4qAL	0.55	-	0.325	/	19 (4)	FSHD1	/	+	0-20	1	23	87
A32	+/+	4qAL	0.419	-	0.514	/	24 (5)	FSHD1	FSHD1	+	0-20	1.5	27	111
A33	+/+	4qAL	0.557	-	0.419	/	26 (6)	FSHD1	FSHD1	-	40-60	2.5	55	91
A34	+/+	4qAL	0.559	-	0.529	/	33 (8)	FSHD1	/	-	20-40	1.5	35	86
A35	+/+	4qAL	0.613	-	0.419	/	/	FSHD1	/	-	40-60	3.5	62	113
A36	+/+	4qA	0.435	0.201	-	/	/	FSHD1	/	+	20-40	1.5	25	120

Supplementary Table 8. Diagnostic cohort: Patients (B) diagnosed for FSHD2 by FSHD-MPA with methylation level determined, repeat length of the D4Z4 array determined by Southern Blotting, results of *SMCHD1*/*DNMT3B* sequencing and clinical data (family history, age at disease onset, (age corrected) clinical severity score (CSS)). Symbols: + positive/present; - not present/negative; / not determined/not known. **LRIF1* was additionally sequenced and evaluated. Highlighted in green: hypermethylation, highlighted in yellow: inconclusive methylation, highlighted in red: pathogenic hypomethylation.

Pat.	Haplotype (4qA/4qA161)	Haplotype by FSHD-MPA	Methylation level			*SMCHD1*/ *DNMT3B* sequencing	Repeat length 4q35 [kb (RU)]	Epigenetic diagnosis	Genetic diagnosis	Familiy history	Age at disease onset	CSS	Age at Assessment	Age-corrected CSS
			DUX4 assay	4qA assay	4qAL assay									
B1	+/+	4qA	0.198	0.196	-	*SMCHD1*: c.5843A>C p.His1948Pro	>48	FSHD2	FSHD2	-	0-20	4	49	163
B2	+/+	4qAL	0.303	-	0.417	*SMCHD1*: c.5556_5561delinsT p.Lys1852Asnfs*17	>48	FSHD2	FSHD2	+	20-40	2.5	41	122
B3	+/+	4qA, AqAL	0.330	0.082	0.466	*SMCHD1*: c.4966+5G>T	>48	FSHD2	FSHD2	/	40-60	/	59	/
B4	+/+	4qA	0.223	0.069	-	*SMCHD1*: c.2753T>A p.Leu918*	/	FSHD2	FSHD2	/	/	/	72	/
B5	+/+	4qA	0.272	0.120	-	*SMCHD1*: c.1846A>G p.Lys616Glu	37 (9)	FSHD2	FSHD1/2	+	0-10	/	17	/
B6	+/+	4qA	0.213	0.072	-	*SMCHD1*: c.2409_2410del p.Tyr804Cysfs*8)	/	FSHD2	FSHD2	+	0-20	/	54	/
B7	+/+	4qA, 4qAL	0.363	0.192	0.596	*SMCHD1*: c.1787G>C p.Trp596Ser	>48	FSHD2	FSHD2	-	0-20	3.5	58	121
B8	+/+	4qAL, 4qA	0.272	0.296	0.341	negative	30 (7)	FSHD2	FSHD1	+	20-40	2.5	29	172
B9	+/+	4qAL	0.405	-	0.476	negative	29 (7)	FSHD2	FSHD1	-	20-40	1.5	32	94
B10	+/+	4qA	0.350	0.304	-	negative	33 (8)	FSHD2	FSHD1	+	20-40	1.5	33	91
B11	+/+	4qAL	0.384	-	0.443	negative	27 (6)	FSHD2	FSHD1	+	20-40	4	55	145
B12	+/+	4qA	0.332	0.223	-	negative	37 (9)	FSHD2	FSHD1	+	0-20	1.5	20	150
B13	+/+	4qAL	0.361	-	0.502	negative*	>48	FSHD2	negative	+	20-40	/.	35	/
B14	+/+	4qA	0.376	0.265	-	negative*	48 (13)	FSHD2	negative	+	0-20	4	55	145

Supplementary Table 9. Diagnostic cohort: Patients (I) with inconclusive results regarding FSHD based on FSHD-MPA with methylation level determined, repeat length of the D4Z4 array determined by Southern Blotting, results of *SMCHD1*/*DNMT3B* sequencing and clinical data (family history, age at disease onset, (age corrected) clinical severity score (CSS)). Abbreviation: presymp. = pressymptomatic. Symbols: + positive/present; - not present/negative; / not determined/not known; * Patient with positive family history of FSHD that shows implied facial weakness and 6-fold elevated CK levels. Highlighted in green: hypermethylation, highlighted in yellow: inconclusive methylation, highlighted in red: pathogenic hypomethylation.

Pat.	Haplotype (4qA/4qA161)	Haplotype by FSHD-MPA	Methylation level			*SMCHD1/ DNMT3B* sequencing	Repeat length 4q35 [kb (RU)]	Epigenetic diagnosis	Genetic diagnosis	Familiy history	Age at disease onset	CSS	Age at Assessment	Age-corrected CSS
			DUX4 assay	4qA assay	4qAL assay									
I1	+/+	4qA	0.600	0.400	-	/	24 (5), >48	inconclusive for FSHD1	FSHD1	-	40-60	2	50	80
I2	+/+	4qA	0.603	0.379	-	negative	14 (2), >48	inconclusive for FSHD1	FSHD1	-	0-20	2.5	46	109
I3	+/+	4qA	0.650	0.400	-	negative	31 (7), >48	inconclusive for FSHD1	FSHD1	-	0-20	2	26	154
I4	+/+	4qA	0.641	0.410	-	/	34 (8), >48	inconclusive for FSHD1	FSHD1	-	80-100	3	81	74
I5	+/+	4qA	0.517	0.398	-	/	38 (10) , >48	inconclusive for FSHD1	FSHD1	+	/	presymp.*	26	presymp.*
I6	+/+	4qA/4qAL	0.710	0.794	0.651	negative	35 (9) , >48	inconclusive for FSHD1	FSHD1	-	/	/	48	/
I7	+/+	4qAL	0.539	-	0.587	negative	>48, >48	inconclusive for FSHD1	negative	+	FSHD unlikely		68	FSHD unlikely
I8	+/+	4qA/4qAL	0.505	0.477	0.605	negative	>50, >50	inconclusive for FSHD1	negative	-	FSHD unlikely	/	48	FSHD unlikely
I9	+/+	4qAL	0.547	-	0.634	negative	>48, >48	inconclusive for FSHD1	negative	-	FSHD unlikely	/	68	FSHD unlikely
I10	+/+	4qA	0.525	0.392	-	/	>48, >48	inconclusive for FSHD1	negative	-	FSHD unlikely	/	79	FSHD unlikely
I11	+/+	4qA	0.505	0.416	-	/	48 (13) , >48	inconclusive for FSHD1	negative	-	FSHD unlikely	/	69	FSHD unlikely
I12	+/+	4qA	0.465	0.421	-	/	/	inconclusive for FSHD1	/	/	/	/	69	/
I13	+/+	4qA, 4qAL	0.421	0.335	0.804	negative	>48, >48	inconclusive for FSHD	negative	+	FSHD unlikely	/	29	FSHD unlikely
I14	+/+	4qA	0.361	0.367	-	negative	>48, >48	inconclusive for FSHD	/	+	/	/	38	/
I15	+/+	4qA	0.426	0.425	-	/	42 (11), >48	inconclusive for FSHD	FSHD1	+	40-60	2.5	51	98
I16	+/+	4qAL	0.410	-	0.652	negative	>48, >48	inconclusive for FSHD	negative	/	FSHD unlikely	/	80	FSHD unlikely
I17	+/+	4qA	0.368	0.377	-	negative	46 (13) , >48	inconclusive for FSHD	negative	/	FSHD unlikely	/	37	FSHD unlikely
I18	+/+	4qA	0.405	0.346	-	negative	42 (11), 46 (13)	inconclusive for FSHD	FSHD1		0-20	3	41	146

Supplementary Table 10. Diagnostic cohort: Patients with negative FSHD diagnosis based on absence of a permissive haplotype or negative FSHD-MPA result with their haplotype, repeat length of the D4Z4 array determined by Southern Blotting results of *SMCHD1*/*DNMT3B* sequencing and if identified differential diagnosis likely accounting for the symptoms. Symbols: + positive/present; - not present/negative; / not determined/not known, * asymptomatic, predictive testing. [#]Presence of DUX4 fragment after BSS and nested PCR was confirmed without quantifying methylation level and in absence of 4qA and 4qAL fragments. Highlighted in green: hypermethylation.

Pat.	Haplotype (4qA/4qA161)	Haplotype FSHD-MPA	Methylation level			Epigenetic diagnosis	*SMCHD1/ DNMT3B* sequencing	Repeat length 4q35 [kb (RU)]	Alternative Diagnosis
			DUX4 assay	4qA assay	4qAL assay				
N1	+/+	4qA	0.531	0.602	-	negative	negative	/	/
N2	+/+	4qAL	0.775	0.843	0.763	negative	/	/	/
N3	+/+	4qA	0.667	0.747	-	negative	negative	/	/
N4	+/+	4qA	0.710	0.672	-	negative	/	/	/
N5	+/+	4qA	0.665	0.677	-	negative	negative	>48, >48	/
N6	+/-	4qA	0.629	0.566	-	negative	negative	>48, >48	/
N7	+/+	4qA	0.599	0.522	-	negative	negative	/	*VCP* associated inclusion body myopathy likely
N8	+/+	4qA	0.530	0.561	-	negative	/	/	/
N9	+/-	4qA	0.639	0.677	-	negative	/	/	/
N10	+/+	4qAL	0.524	-	0.663	negative	/	/	/
N11	+/+	4qA	0.715	0.541	-	negative	negative	/	/
N12	+/+	4qA	0.600	0.661	-	negative	/	/	/
N13	+/+	4qA	0.532	0.590	-	negative	negative	/	/
N14	+/+	4qA	0.745	0.568	-	negative	negative	/	/
N15	+/+	4qA	0.676	0.706	-	negative	/	/	/
N16*	+/+	4qAL	0.538	-	0.691	negative	/	>48, >48	/
N17	+/+	4qAL	0.651	-	0.714	negative	/	>48, >48	/
N18	+/+	4qA	0.674	0.594	-	negative	/	/	/
N19	+/+	4qA	0.632	0.593	-	negative	/	/	/

N20	+/+	4qA	0.616	0.612	-	negative	/	/	/
N21	-/-	non-permissive	0.749	-	-	negative	negative	/	/
N22	+/+	4qA, 4qAL	0.673	0.686	0.806	negative	/	/	*FLNC*-associated myopathy likely
N23	+/+	4qA	0.589	0.660	-	negative	/	/	/
N24	+/+	4qA	0.677	0.672	-	negative	/	>48, >48	/
N25	-/-	non-permissive	0.725	-	-	negative	/	/	/
N26	+/+	4qA	0.609	0.733	-	negative	/	/	/
N27	+/+	4qA	0.678	0.645	-	negative	negative	>48, >48	*DNM2* centronuclear myopathy likely
N28	+/+	4qA	0.556	0.538	-	negative	/	>48, >48	/
N29	+/+	4qA	0.688	0.613	-	negative	/	/	/
N30	+/+	4qA	0.659	0.693	-	negative	/	/	/
N31	+/+	4qA	0.582	0.607		negative	/	/	/
N32	+/+	4qAL	0.616	-	0.676	negative	/	/	/
N33	+/+	4qA	0.725	0.665	-	negative	/	/	/
N34	+/+	4qA	0.609	0.470	-	negative	/	/	/
N35	+/+	4qA	0.701	0.759	-	negative	/	/	*PYROXD1*-associated myofibrillar myopathy likely
N36	+/+	4qA	0.581	0.581	-	negative	/	/	/
N37*	+/+	4qA	0.449	0.610	-	negative	/	38 (10) , >48	
N38	+/+	4qA	0.703	0.789	-	negative	/	/	/
N39	+/+	4qA	0.541	0.601	-	negative	/	/	/
N40	+/+	4qA	0.762	0.589	-	negative	/	/	/
N41	+/+	4qA	0.803	0.695	-	negative	/	/	/
N42	+/+	4qA, 4qAL	0.753	0.676	0.771	negative	/	/	/

N43	+/+	4qA	0.475	0.569	-	negative	/	>48, >48	/
N44	+/+	4qA	0.572	0.614	-	negative	/	/	/
N45	+/+	4qA	0.738	0.817	-	negative	/	/	/
N46	+/+	4qA	0.715	0.680	-	negative	negative	/	/
N47	+/+	4qA	0.452	0.582	-	negative	negative	/	/
N48	+/+	4qA	0.620	0.490	-	negative	/	/	/
N49	+/+	4qA, 4qAL	0.696	0.637	0.734	negative	/		*CAPN3* limb-girdle myopathy possible
N50	+/+	4qA	0.588	0.653	-	negative	/	/	/
N51	+/+	4qA	0.668	0.655	-	negative	negative	/	/
N52	+/+	4qA	0.736	0.560	-	negative	negative	/	/
N53	+/+	4qA, 4qAL	0.601	0.556	0.672	negative	/	/	/
N54	+/+	4qAL	0.546	-	0.723	negative	/	/	/
N55*	+/+	4qA	0.526	0.534	-	negative	*SMCHD1:* c.1754G>A p. Arg585His (class 3)	/	/
N56	+/+	4qA	0.754	0.599	-	negative	/	/	/
N57	+/+	4qA	0.735	0.542	-	negative	negative	/	/
N58	+/+	4qA	0.723	0.697	-	negative	/	/	/
N59	+/+	4qA	0.673	0.599	-	negative	/	/	/
N60	+/+	4qA	0.578	0.574	-	negative	/	/	/
N61	-/-	non-permissive	0.542	-	-	negative	/	/	/
N62	+/+	4qAL	0.7	-	0.814	negative	/	/	/
N63	-/-	non-permissive	/≠	-	-	negative	/		Becker muscular dystrophy likely
N64	-/-	non-permissive	/≠	-	-	negative	/	/	/
N65	+/+	4qA	0.66	0.574	-	negative	/	/	/

N66	-/-	non-permissive	/≠	-	-	negative	/	/	Becker muscular dystrophy likely
N67	-/-	non-permissive	/≠	-	-	negative	/	/	/
N68	-/-	non-permissive	/≠	-	-	negative	/	/	/
N69	-/+	non-permissive	/≠	-	-	negative	/	/	/
N70	+/+	4qA	0.756	0.657	-	negative	/	/	/
N71	+/+	4qA	0.65	0.644	-	negative	/	/	/
N72	-/-	non-permissive	/≠	-	-	negative	/	/	/
N73	+/+	4qA	0.447	0.484	-	negative	/	/	/
N74	-/+	non-permissive	0.674	-	-	negative	/	/	/
N75	-/-	non-permissive	/≠	-	-	negative	/	/	/
N76	+/+	4qA	0.733	0.81	-	negative	/	/	/
N77	+/+	4qAL	0.63	-	0.653	negative	negative	>48, >48	/
N78	+/+	4qA	0.687	0.693	-	negative	/	>48, >48	/
N79	-/-	non-permissive	/≠	-	-	negative	/	/	/
N80	-/-	non-permissive	/≠	-	-	negative	/	/	/

Supplementary Table 11. Patients (Z) with confirmed FSHD diagnosis or positive family history for FSHD included in the genotype-phenotype cohort with methylation level determined, repeat length of the D4Z4 array determined by Southern Blotting, results of *SMCHD1/DNMT3B* sequencing and clinical data (family history, age at disease onset, (age corrected) clinical severity score (CSS)). Abbreviation: asymp. = asymptomatic. Symbols: + positive/present; - not present/negative; / not determined/not known. **LRIF1* was additionally sequenced and evaluated. Highlighted in green: hypermethylation, highlighted in yellow: inconclusive methylation, highlighted in red: pathogenic hypomethylation.

Pat.	Haplotype (4qA/4qA161)	Haplotype by FSHD-MPA	Methylation level			*SMCHD1/ DNMT3B* sequencing	Repeat length 4q35 [kb (RU)]	Epigenetic diagnosis	Genetic diagnosis	Familiy history	Age at disease onset	CSS	Age at Assessment	Age-corrected CSS
			DUX4 assay	4qA assay	4qAL assay									
Z1	+/+	4qA	0.423	0.282	-	/	27 (6) , >48	FSHD1+	FSHD1	+	0-20	2.5	43	116
Z2	+/+	4qA	0.584	0.161	-	/	31 (7) , >48	FSHD1	FSHD1	+	0-20	2	22	182
Z3	+/+	4qA	0.502	0.104	-	/	/	FSHD1	/	-	0-20	3.5	29	241
Z4	+/+	4qA	0.492	0.209	-	/	/	FSHD1	/	+	0-20	2	56	71
Z5	+/+	4qA	0.58	0.168	-	/	/	FSHD1	/	-	40-60	4	56	143
Z6	+/+	4qA	0.55	0.282	-	/	37 (9) , >48	FSHD1	/	+	60-80	3	81	74
Z7	+/+	4qA	0.556	0.289	-	/	17 (3) , >48	FSHD1	FSHD1	-	0-20	2	46	87
Z8	+/+	4qA, 4qAL	0.465	0.275	0.677	/	/	FSHD1	/	-	0-20	2.5	30	167
Z9	+/+	4qAL	0.523	-	0.52	/	33 (8) , >48	FSHD1	FSHD1	+	40-60	1	44	45
Z10	+/+	4qA, 4qAL	0.572	0.706	0.497	/	/	FSHD1	/	+	20-40	1	24	83
Z11	+/+	4qA	0.401	0.138	-	/	27 (6) , >48	FSHD2	FSHD1 (FSHD2 unknown)	+	0-20	2	20	200
Z12	+/+	4qA	0.579	0.186	-	/	24 (5) , >48	FSHD1	FSHD1	/	0-20	3.5	34	206
Z13	+/+	4qAL	0.532	-	0.484	/	38 (10) , >48	FSHD1	/	-	/	2	54	74
Z14	+/+	4qAL	0.383	-	0.303	*SMCHD1*:c.5145_5146del p.(Thr1716fs) (class)	/	FSHD2	FSHD2	-	40-60	4	56	143
Z15	+/+	4qA	0.561	0.367	-	/	20 (4) , >48	FSHD1	FSHD1	-	60-80	3	65	92
Z16	+/+	4qA	0.434	0.328	-	/	/	FSHD1	/	+	0-20	3	40	150
Z17	+/+	4qA	0.557	0.524		negative*	14 (2) , >48	negative	FSHD1	+	/	asympt.	71	asympt.
Z18	+/+	4qA	0.472	0.123		negative*	14 (2) , >48	FSHD1	FSHD1	+	0-20	4.5	45	200
Z19	+/+	4qA	0.555	0.149		negative*	14 (2) , >48	FSHD1	FSHD1	+	0-20	3.5	21	333
Z20	+/+	4qA	0.502	0.118		negative*	14 (2) , >48	FSHD1	FSHD1	+	0-20	4	22	364

7 References

1 Erdmann H, Schöberl F, Giurgiu M, et al. Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing. *Brain*. 2023;146(5):1831-1843. doi:10.1093/brain/awac377

2 Erdmann H, Scharf F, Gehling S, et al. Methylation of the 4q35 D4Z4 repeat defines disease status in facioscapulohumeral muscular dystrophy. *Brain*. 2023;146(4):1388-1402. doi:10.1093/brain/awac336

3 Erdmann H, Scharf F, Hallermayr A, et al. Reply: An epigenetic basis for genetic anticipation in facioscapulohumeral muscular dystrophy type 1. *Brain*. Published online 22 June 2023. doi:10.1093/brain/awad216

4 Reilich P, Schlotter B, Montagnese F, et al. Location matters – Genotype-phenotype correlation in LRSAM1 mutations associated with rare Charcot-Marie-Tooth neuropathy CMT2P. *Neuromuscul Disord*. 2021;31(2):123-133. doi:10.1016/j.nmd.2020.11.011

5 Schoeberl F, Abicht A, Kuepper C, et al. Sensory neuropathy due to RFC1 in a patient with ALS: more than a coincidence? *J Neurol*. 2022;269(5):2774-2777. doi:10.1007/s00415-021-10835-9

6 Erdmann H, Abicht A. Häufige intronische Repeat-Expansionen in FGF14 – eine weitere lang gesuchte Ursache bei spätmanifester Ataxie. *DGNeurologie*. 2023;6(2):159-160. doi:10.1007/s42451-023-00535-1

7 Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921. doi:10.1038/35057062

8 Biscotti MA, Olmo E, Heslop-Harrison JS (Pat). Repetitive DNA in eukaryotic genomes. *Chromosome Res*. 2015;23(3):415-420. doi:10.1007/s10577-015-9499-z

9 Depienne C, Mandel JL. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *Am J Hum Genet*. 2021;108(5):764-785. doi:10.1016/j.ajhg.2021.03.011

10 Dumbovic G, Forcales SV, Perucho M. Emerging roles of macrosatellite repeats in genome organization and disease development. *Epigenetics.* 2017;12(7):515-526. doi:10.1080/15592294.2017.1318235

11 Gatchel JR, Zoghbi HY. Diseases of Unstable Repeat Expansion: Mechanisms and Common Principles. *Nat Rev Genet*. 2005;6(10):743-755. doi:10.1038/nrg1691

12 Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Annu Rev Genet*. 2010;44(1):445-477. doi:10.1146/annurev-genet-072610-155046

13 Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet*. 2018;19(5):286-298. doi:10.1038/nrg.2017.115

14 Chintalaphani SR, Pineda SS, Deveson IW, Kumar KR. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol*. 2021;9(1):98. doi:10.1186/s40478-021-01201-x

15 Paulson H. Chapter 9 - Repeat expansion diseases. In: Geschwind DH, Paulson HL, Klein C, eds. *Handb Clin Neurol*. Vol 147. Elsevier; 2018:105-123. doi:10.1016/B978-0-444-63233-3.00009-9

16 Salzberg SL, Yorke JA. Beware of mis-assembled genomes. *Bioinformatics*. 2005;21(24):4320-4321. doi:10.1093/bioinformatics/bti769

17 Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. 2012;13(1):36-46. doi:10.1038/nrg3117

18 Mantere T, Kersten S, Hoischen A. Long-Read Sequencing Emerging in Medical Genetics. *Front Genet*. 2019;10:426. doi:10.3389/fgene.2019.00426

19 Giesselmann P, Brändl B, Raimondeau E, et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat Biotechnol*. 2019;37(12):1478-1481. doi:10.1038/s41587-019-0293-x

20 Stevanovski Igor, Chintalaphani Sanjog R., Gamaarachchi Hasindu, et al. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci Adv*. 2022;8(9):eabm5386. doi:10.1126/sciadv.abm5386

21 Gilpatrick T, Lee I, Graham JE, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol*. 2020;38(4):433-438. doi:10.1038/s41587-020-0407-5

22 Sato N, Amino T, Kobayashi K, et al. Spinocerebellar Ataxia Type 31 Is Associated with “Inserted” Penta-Nucleotide Repeats Containing (TGGAA)n. *Am J Hum Genet*. 2009;85(5):544-557. doi:10.1016/j.ajhg.2009.09.019

23 Seixas AI, Loureiro JR, Costa C, et al. A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of DAB1, Mapping to SCA37, Causes Spinocerebellar Ataxia. *Am J Hum Genet*. 2017;101(1):87-103. doi:10.1016/j.ajhg.2017.06.007

24 Cortese A, Simone R, Sullivan R, et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat Genet*. 2019;51(4):649-658. doi:10.1038/s41588-019-0372-4

25 Sulovari Arvis, Li Ruiyang, Audano Peter A., et al. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc Natl Acad Sci USA*. 2019;116(46):23243-23253. doi:10.1073/pnas.1912175116

26 Lemmers Richard J. L. F., van der Vliet Patrick J., Klooster Rinse, et al. A Unifying Genetic Model for Facioscapulohumeral Muscular Dystrophy. *Science*. 2010;329(5999):1650-1653. doi:10.1126/science.1189044

27 Geng LN, Yao Z, Snider L, et al. DUX4 Activates Germline Genes, Retroelements, and Immune Mediators: Implications for Facioscapulohumeral Dystrophy. *Dev Cell*. 2012;22(1):38-51. doi:10.1016/j.devcel.2011.11.013

28 Jones TI, Chen JCJ, Rahimov F, et al. Facioscapulohumeral muscular dystrophy family studies of DUX4 expression: evidence for disease modifiers and a quantitative model of pathogenesis. *Hum Mol Genet*. 2012;21(20):4419-4430. doi:10.1093/hmg/dds284

29 Bosnakovski D, Chan SSK, Recht OO, et al. Muscle pathology from stochastic low level DUX4 expression in an FSHD mouse model. *Nat Commun*. 2017;8(1):550. doi:10.1038/s41467-017-00730-1

30 Rashnonejad A, Amini-Chermahini G, Taylor NK, Wein N, Harper SQ. Designed U7 snRNAs inhibit DUX4 expression and improve FSHD-associated outcomes in DUX4 overexpressing cells and FSHD patient myotubes. *Mol Ther Nucl Acids*. 2020;23:476-486. doi:10.1016/j.omtn.2020.12.004

31 Marsollier AC, Ciszewski L, Mariot V, et al. Antisense targeting of 3′ end elements involved in DUX4 mRNA processing is an efficient therapeutic strategy for facioscapulohumeral dystrophy: a new gene-silencing approach. *Hum Mol Genet*. 2016;25(8):1468-1478. doi:10.1093/hmg/ddw015

32 Lim KRQ, Bittel A, Maruyama R, et al. DUX4 Transcript Knockdown with Antisense 2′-O-Methoxyethyl Gapmers for the Treatment of Facioscapulohumeral Muscular Dystrophy. *Mol Ther*. 2021;29(2):848-858. doi:10.1016/j.ymthe.2020.10.010

33 Deenen JCW, Arnts H, van der Maarel SM, et al. Population-based incidence and prevalence of facioscapulohumeral dystrophy. *Neurology*. 2014;83(12):1056. doi:10.1212/WNL.0000000000000797

34 Mostacciuolo M, Pastorello E, Vazza G, et al. Facioscapulohumeral muscular dystrophy: epidemiological and molecular study in a north-east Italian population sample. *Clin Genet*. 2009;75(6):550-555. doi:10.1111/j.1399-0004.2009.01158.x

35 Gaillard MC, Roche S, Dion C, et al. Differential DNA methylation of the D4Z4 repeat in patients with FSHD and asymptomatic carriers. *Neurology*. 2014;83(8):733. doi:10.1212/WNL.0000000000000708

36 Jones TI, Yan C, Sapp PC, et al. Identifying diagnostic DNA methylation profiles for facioscapulohumeral muscular dystrophy in blood and saliva using bisulfite sequencing. *Clin Epigenetics*. 2014;6(1):23. doi:10.1186/1868-7083-6-23

37 Hartweck LM, Anderson LJ, Lemmers RJ, et al. A focal domain of extreme demethylation within D4Z4 in FSHD2. *Neurology*. 2013;80(4):392. doi:10.1212/WNL.0b013e31827f075c

38 Gould T, Jones TI, Jones PL. Precise Epigenetic Analysis Using Targeted Bisulfite Genomic Sequencing Distinguishes FSHD1, FSHD2, and Healthy Subjects. *Diagnostics*. 2021;11(8). doi:10.3390/diagnostics11081469

39 Salsi V, Magdinier F, Tupler R. Does DNA Methylation Matter in FSHD? *Genes*. 2020;11(3):258. doi:10.3390/genes11030258

40 Jayadev S, Bird TD. Hereditary ataxias: overview. *Genet Med*. 2013;15(9):673-683. doi:10.1038/gim.2013.28

41 Bird TD. Hereditary Ataxia Overview. GeneReviews®. Published April 28, 2021. Accessed April 28, 2021. https://www.ncbi.nlm.nih.gov/books/NBK1138/

42 Brusse E, Maat-Kievit J, Van Swieten J. Diagnosis and management of early- and late-onset cerebellar ataxia. *Clin Genet*. 2007;71(1):12-24. doi:10.1111/j.1399-0004.2006.00722.x

43 Pellerin D, Danzi MC, Wilke C, et al. Deep Intronic FGF14 GAA Repeat Expansion in Late-Onset Cerebellar Ataxia. *N Engl J Med*. 2023;388:128-141. doi:10.1056/NEJMoa2207406

44 Kenneson A, Zhang F, Hagedorn CH, Warren ST. Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate-length and premutation carriers. *Hum Mol Genet*. 2001;10(14):1449-1454. doi:10.1093/hmg/10.14.1449

45 Kraan CM, Godler DE, Amor DJ. Epigenetics of fragile X syndrome and fragile X-related disorders. *Dev Med Child Neurol*. 2019;61(2):121-127. doi:10.1111/dmcn.13985

46 Groh M, Lufino MMP, Wade-Martins R, Gromak N. R-loops Associated with Triplet Repeat Expansions Promote Gene Silencing in Friedreich Ataxia and Fragile X Syndrome. *PLOS Genet*. 2014;10(5):e1004318. doi:10.1371/journal.pgen.1004318

47 Delatycki MB, Bidichandani SI. Friedreich ataxia- pathogenesis and implications for therapies. *Neurobiol Dis*. 2019;132:104606. doi:10.1016/j.nbd.2019.104606

48 Lieberman AP, Shakkottai VG, Albin RL. Polyglutamine Repeats in Neurodegenerative Diseases. *Annu Rev Pathol Mech Dis*. 2019;14(1):1-27. doi:10.1146/annurev-pathmechdis-012418-012857

49 Miller JW, Urbinati CR, Teng-umnuay P, et al. Recruitment of human muscleblind proteins to (CUG)n expansions associated with myotonic dystrophy. *EMBO J*. 2000;19(17):4439-4448. doi:10.1093/emboj/19.17.4439

50 Cleary JD, Pattamatta A, Ranum LPW. Repeat-associated non-ATG (RAN) translation. *J Biol Chem*. 2018;293(42):16127-16141. doi:10.1074/jbc.R118.003237

51 Bañez-Coronel M, Ayhan F, Tarabochia AD, et al. RAN Translation in Huntington Disease. *Neuron*. 2015;88(4):667-677. doi:10.1016/j.neuron.2015.10.038

52 Zu T, Pattamatta A, Ranum LPW. Repeat-Associated Non-ATG Translation in Neurological Diseases. *Cold Spring Harb Perspect Biol*. 2018;10(12). doi:10.1101/cshperspect.a033019

53 Swinnen B, Robberecht W, Van Den Bosch L. RNA toxicity in non-coding repeat expansion disorders. *EMBO J*. 2020;39(1):e101112. doi:10.15252/embj.2018101112

54 Cortese A, Tozza S, Yau WY, et al. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome due to RFC1 repeat expansion. *Brain*. 2020;143(2):480-490. doi:10.1093/brain/awz418

55 Akçimen F, Ross JP, Bourassa CV, et al. Investigation of the RFC1 Repeat Expansion in a Canadian and a Brazilian Ataxia Cohort: Identification of Novel Conformations. *Front Genet*. 2019;10:1219. doi:10.3389/fgene.2019.01219

56 Rafehi H, Szmulewicz DJ, Bennett MF, et al. Bioinformatics-Based Identification of Expanded Repeats: A Non-reference Intronic Pentamer Expansion in RFC1 Causes CANVAS. *Am J Hum Genet*. 2019;105(1):151-165. doi:10.1016/j.ajhg.2019.05.016

57 Huin V, Coarelli G, Guemy C, et al. Motor neuron pathology in CANVAS due to RFC1 expansions. *Brain*. 2022;145(6):2121-2132. doi:10.1093/brain/awab449

58 Schmitt G da S, Lima FD de, Matos PCAAP, et al. Dysautonomia in RFC1-related disorder: Clinical and neurophysiological evaluation. *Clin Neurophysiol*. 2022;142:68-74. doi:10.1016/j.clinph.2022.07.501

59 Davies K, Szmulewicz DJ, Corben LA, Delatycki M, Lockhart PJ. RFC1-Related Disease. *Neurol Genet*. 2022;8(5):e200016. doi:10.1212/NXG.0000000000200016

60 Traschütz A, Wilke C, Haack TB, Bender B, RFC1 Study Group, Synofzik M. Sensory axonal neuropathy in RFC1-disease: tip of the iceberg of broad subclinical multisystemic neurodegeneration. *Brain*. 2022;145(3):e6-e9. doi:10.1093/brain/awac003

61 Ronco R, Perini C, Currò R, et al. Truncating Variants in RFC1 in Cerebellar Ataxia, Neuropathy, and Vestibular Areflexia Syndrome. *Neurology*. 2023;100(5):e543–e554. doi:10.1212/WNL.0000000000201486

62 Benkirane M, Da Cunha D, Marelli C, et al. RFC1 nonsense and frameshift variants cause CANVAS: clues for an unsolved pathophysiology. *Brain*. 2022;145(11):3770–3775. doi:10.1093/brain/awac280

63 Montaut S, Diedhiou N, Fahrer P, et al. Biallelic RFC1-expansion in a French multicentric sporadic ataxia cohort. *J Neurol*. 2021;268(9);3337–3343. doi:10.1007/s00415-021-10499-5

64 Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet*. 2020;21(10):597-614. doi:10.1038/s41576-020-0236-x

65 Eid J, Fehr A, Gray J, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*. 2009;323(5910):133-138. doi:10.1126/science.1162986

66 Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. *Nat Biotechnol*. 2016;34. doi:10.1038/nbt.3423

67 Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol*. 2016;17(1):239. doi:10.1186/s13059-016-1103-0

68 Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol*. 2021;39(11):1348-1365. doi:10.1038/s41587-021-01108-x

69 Mul K, Lassche S, Voermans NC, Padberg GW, Horlings CG, van Engelen BG. What's in a name? The clinical features of facioscapulohumeral muscular dystrophy. *Pract Neurol*. 2016;16(3):201. doi:10.1136/practneurol-2015-001353

70 Kottlors M, Kress W, Meng G, Glocker FX. Facioscapulohumeral muscular dystrophy presenting with isolated axial myopathy and bent spine syndrome. *Muscle Nerve*. 2010;42(2):273-275. doi:10.1002/mus.21722

71 Ricci G, Ruggiero L, Vercelli L, et al. A novel clinical tool to classify facioscapulohumeral muscular dystrophy phenotypes. *J Neurol*. 2016;263(6):1204-1214. doi:10.1007/s00415-016-8123-2

72 Ricci G, Scionti I, Sera F, et al. Large scale genotype-phenotype analyses indicate that novel prognostic tools are required for families with facioscapulohumeral muscular dystrophy. *Brain*. 2013;136(Pt 11):3408-3417. doi:10.1093/brain/awt226

73 Dixit Manjusha, Ansseau Eugénie, Tassin Alexandra, et al. DUX4, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of PITX1. *Proc Natl Acad Sci USA*. 2007;104(46):18157-18162. doi:10.1073/pnas.0708659104

74 Wijmenga C, Hewitt JE, Sandkuijl LA, et al. Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nat Genet*. 1992;2(1):26-30. doi:10.1038/ng0992-26

75 Gabriëls J, Beckers MC, Ding H, et al. Nucleotide sequence of the partially deleted D4Z4 locus in a patient with FSHD identifies a putative gene within each 3.3 kb element. *Gene*. 1999;236(1):25-32. doi:10.1016/S0378-1119(99)00267-X

76 Hamanaka K, Šikrová D, Mitsuhashi S, et al. Homozygous nonsense variant in *LRIF1* associated with facioscapulohumeral muscular dystrophy. *Neurology*. 2020;94(23):e2441. doi:10.1212/WNL.0000000000009617

77 van den Boogaard ML, Lemmers RJLF, 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):1020-1029. doi:10.1016/j.ajhg.2016.03.013

78 Lemmers RJLF, Goeman JJ, van der Vliet PJ, et al. Inter-individual differences in CpG methylation at D4Z4 correlate with clinical variability in FSHD1 and FSHD2. *Hum Mol Genet*. 2015;24(3):659-669. doi:10.1093/hmg/ddu486

79 Lemmers RJLF, O'Shea S, Padberg GW, Lunt PW, van der Maarel SM. Best practice guidelines on genetic diagnostics of Facioscapulohumeral muscular dystrophy: Workshop 9th June 2010, LUMC, Leiden, The Netherlands. *Neuromuscul Disord*. 2012;22(5):463-470. doi:10.1016/j.nmd.2011.09.004

80 Lemmers RJLF, De Kievit P, van geel M, et al. Complete allele information in the diagnosis of facioscapulohumeral muscular dystrophy by triple DNA analysis. *Ann Neurol*. 2001;50(6):816-819. doi:10.1002/ana.10057

81 Lemmers RJLF. Analyzing Copy Number Variation Using Pulsed-Field Gel Electrophoresis: Providing a Genetic Diagnosis for FSHD1. In: White SJ, Cantsilieris S, eds. *Genotyping: Methods and Protocols*. Springer New York; 2017:107-125. doi:10.1007/978-1-4939-6442-0_7

82 Nguyen K, Walrafen P, Bernard R, et al. Molecular combing reveals allelic combinations in facioscapulohumeral dystrophy. *Ann Neurol*. 2011;70(4):627-633. doi:10.1002/ana.22513

83 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;23(11):1506-1514. doi:10.1016/j.jmoldx.2021.07.021

84 Dai Y, Li P, Wang Z, et al. Single-molecule optical mapping enables quantitative measurement of D4Z4 repeats in facioscapulohumeral muscular dystrophy (FSHD). *J Med Genet*. 2020;57(2):109. doi:10.1136/jmedgenet-2019-106078

85 Nguyen K, Puppo F, Roche S, et al. Molecular combing reveals complex 4q35 rearrangements in Facioscapulohumeral dystrophy. *Hum Mutat*. 2017;38(10):1432-1441. doi:10.1002/humu.23304

86 Tsumagari K, Chen D, Hackman JR, Bossler AD, Ehrlich M. FSH dystrophy and a subtelomeric 4q haplotype: a new assay and associations with disease. *J Med Genet*. 2010;47(11):745. doi:10.1136/jmg.2009.076703

87 Scionti I, Fabbri G, Fiorillo C, et al. Facioscapulohumeral muscular dystrophy: new insights from compound heterozygotes and implication for prenatal genetic counselling. *J Med Genet*. 2012;49(3):171. doi:10.1136/jmedgenet-2011-100454

88 de Greef JC, Lemmers RJLF, van Engelen BGM, et al. Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. *Hum Mutat*. 2009;30(10):1449-1459. doi:10.1002/humu.21091

89 Lemmers RJLF, van der Vliet PJ, Vreijling JP, et al. Cis D4Z4 repeat duplications associated with facioscapulohumeral muscular dystrophy type 2. *Hum Mol Genet*. 2018;27(20):3488-3497. doi:10.1093/hmg/ddy236

90 Nikolic A, Jones TI, Govi M, et al. Interpretation of the Epigenetic Signature of Facioscapulohumeral Muscular Dystrophy in Light of Genotype-Phenotype Studies. *Int J Mol Sci*. 2020;21(7):2635. doi:10.3390/ijms21072635

91 Jones TI, King OD, Himeda CL, et al. Individual epigenetic status of the pathogenic D4Z4 macrosatellite correlates with disease in facioscapulohumeral muscular dystrophy. *Clin Epigenetics*. 2015;7(1):37. doi:10.1186/s13148-015-0072-6

92 Frommer M, McDonald LE, Millar DS, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA*. 1992;89(5):1827-1831. doi:10.1073/pnas.89.5.1827

93 Li Y, Tollefsbol TO. DNA Methylation Detection: Bisulfite Genomic Sequencing Analysis. In: Tollefsbol TO, ed. *Epigenetics Protocols*. Humana Press; 2011:11-21. doi:10.1007/978-1-61779-316-5_2

94 Russell J Butterfield, Diane M Dunn, Brett Duval, Sarah Moldt, Robert B Weiss. Deciphering D4Z4 CpG methylation gradients in fascioscapulohumeral muscular dystrophy using nanopore sequencing. *bioRxiv*. [Preprint] doi:10.1101/2023.02.17.528868

95 Hiramuki Y, Kure Y, Saito Y, 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. doi:10.1186/s12967-022-03743-7

96 Zheng F, Qiu L, Chen L, et al. An epigenetic basis for genetic anticipation in facioscapulohumeral muscular dystrophy type 1. *Brain*. Published online June 23. doi:10.1093/brain/awad215

Appendix A: Paper III: Reply: An epigenetic basis for genetic anticipation in facioscapulohumeral muscular dystrophy type 1

Hannes Erdmann[±], Florentine Scharf[±], Ariane Hallermayr, Hayk Barsegehyan, Maggie C. Walter, Elke Holinski-Feder, Benedikt Schoser, Angela Abicht. *Brain*. Published online 22 June 2023.

[±] These authors contributed equally.

DOI:

10.1093/brain/awad216

Impact factor of Brain (2021):

15.255

Rank by Journal Citation Indicator Category Clinical Neurology (2021):

6/267 (JCI Percentile 97.94%)

Reply: An epigenetic basis for genetic anticipation in facioscapulohumeral muscular dystrophy type 1

Hannes Erdmann,[1,2,†] Florentine Scharf,[1,†] Ariane Hallermayr,[1,3] Hayk Barsegehyan,[1,4,5] Maggie C. Walter,[2] Elke Holinski-Feder,[1,3] Benedikt Schoser[2] and Angela Abicht[1,2]

[†]**These authors contributed equally to this work.**

Author affiliations:

1 Medical Genetics Center (MGZ), 80335 Munich, Germany

2 Friedrich-Baur-Institute, Department of Neurology, Klinikum der Universität, Ludwig-Maximilians-Universität, 80336 Munich, Germany

3 Department of Medicine IV, Klinikum der Universität, Ludwig-Maximilians-Universität, 80336 Munich, Germany

4 Center for Genetic Medicine Research, Children's National Research Institute, Children's National Hospital, Washington, DC 20010, USA

5 Department of Genomics and Precision Medicine, School of Medicine and Health Sciences, The George Washington University, Washington, DC 20037, USA

Correspondence to: Angela Abicht

Medical Genetics Center

Bayerstr. 3–5

80335 Munich, Germany

E-mail: Angela.Abicht@mgz-muenchen.de

Correspondence may also be addressed to: Benedikt Schoser

Friedrich-Baur-Institute, Department of Neurology, LMU Klinikum

Ziemenssenstr. 1

80336 München, Germany

E-mail: Benedikt.Schoser@med.uni-muenchen.de

Given the urgent need to establish biomarkers and improve molecular diagnosis of facioscapulohumeral muscular dystrophy (FSHD), we were delighted to read the letter of Zheng and colleagues demonstrating that differences in the distal methylation of the FSHD locus explain intergenerational phenotypic variations.[1] In 77 families, Zheng and co-workers show that FSHD symptoms tend to manifest earlier and more severe in descending generations yielding evidence for clinical anticipation in this cohort. Anticipation is usually linked to repeat expansion disorders, where instable microsatellites tend to expand from one generation to the following, causing earlier onset and more severe phenotypes by the underlying pathomechanism. The situation in FSHD is more complex, considering its unique pathomechanism. As D4Z4 repeat sizes are usually stable between generations, anticipation has been assumed to be implausible because of the missing genetic correlate.[2] An important phenomenon explaining anticipation in some families is somatic mosaicism, described in previous works.[3] Mosaicism is often missed by linear gel electrophoresis due to the method's limitation in properly separating large fragments, as it would be required to detect the presence of more than two 4q alleles. In the study by Zheng and coworkers, somatic mosaicism is ruled out as an explanation for anticipation. Repeat sizes were determined by pulsed-field gel electrophoresis, and detected mosaicism was excluded from the study. Instead, Zheng and coworkers showed that epigenetic changes, i.e. notably lower methylation levels of a CpG island within the distal repeat unit (Figure 1B, pink line), are the missing link rationalizing anticipation. The molecular mechanism of this epigenetic difference remains to be determined and adds to the not yet understood peculiarities of FSHD.

The letter of Zheng and co-worker prompted us to further investigate the molecular structure and epigenetic landscape of the FSHD locus in one multigenerational family presented in our original publication and to discuss the findings in terms of the clinical status of family members (Fig. 1A, Table 1).[4] Informed consent to participate in this study was obtained from all patients, and the

study was approved by local institutions (Bayerische Landesärztekammer, 2019-210) by the guidelines of the Declaration of Helsinki.

A reexamination of the family members showed that the grandfather (I:1) is now paucisymptomatic with a mild facial weakness only. His daughter (II:1) is severely affected and meanwhile wheelchair-bound, while the grandchildren (III:1 and III:2) show unchanged severe affection of the upper and lower limb. The grandchildren (III) show higher age-corrected clinical severity scores than the mother, earlier involvement of the lower limb, and a higher clinical severity score (CSS) for age 22, for which clinical records are available for all family members. Accordingly, there is evidence for clinical anticipation not only between the grandfather and his daughter, but also between his daughter and his grandchildren. However, the differences of the age-corrected CSS might overestimate the intergenerational difference in clinical severity, as the score tends to underestimate the extent of the disease in strongly affected patients with early disease onset and long disease duration, as is the case for the daughter (II:1).[5] This effect also seems to be present in the study of Zheng *et al.*, resulting in a mild overestimation of the extent of anticipation in severely affected families.

To further investigate potential molecular differences between the family members that explain the phenotypical differences, we used a novel long-read sequencing method based on CRISPR/Cas9 target enrichment and Oxford nanopore technology (ONT) long-read sequencing to analyze the FSHD locus.[6] This method allows to analyze all parameters associated with FSHD (haplotype, repeat length, and methylation profile) simultaneously. It overcomes the limitation of current bisulfite sequencing (BSS) methods in analyzing methylation, which are the restriction to a few regions within the FSHD locus and their analysis in an average manner over all permissive alleles. For all family members, we determined the epigenetic profile of the whole locus (Fig. 1D–G) as well as the median distal methylation for each 4qA allele separately and over both alleles (all family members were homozygous for the 4qA haplotype). The analyzed region is similar to that of the 4qA assay of our BSS-based methylation profile analysis (FSHD-MPA, blue line, Fig. 1 B).

Our analysis confirms the presence of one contracted (2 repeat units) and one uncontracted (> 13 repeat units) 4qA allele in all family members. The exact size of the uncontracted allele remains elusive due to its long size (> 48 kb) not being fully captured by the analysis. Interestingly, the

analysis yields evidence for a somatic mosaicism in the grandfather (I:1) as the underlying mechanism for the extremely mild affection (Table 1, Fig. 1D) that escaped the detection by Southern blotting due to the usage of linear gel electrophoresis in the previous analysis. About 42% of the cells are carrying the hypomethylated, two repeat unit long D4Z4 repeat array. Given the size bias of nanopores for shorter fragments, the percentage of mosaicism might be slightly lower. The median distal methylation over both alleles is significantly higher ($p < 0.01$) than that of the other family members, which is in agreement with our previous FSHD-MPA analysis. This result indicates that extremely mild FSHD conditions may escape FSHD-MPA, as distal methylation is determined on average across all permissive alleles. This is particularly the case when patients are homozygous for the permissive allele and carry the contracted allele in only a few cells. Correlating with the severe clinical phenotype and in line with the results from FSHD-MPA, the descendants showed strongly hypomethylated contracted D4Z4 repeat arrays with an allele frequency of about 50% (Table 1, Fig. 1E–G dark blue dots). In contrast to FSHD-MPA, ONT long-read sequencing allowed for the determination of the distal methylation specifically for the contracted allele. Interestingly, as in the study by Zheng *et al.* for "CpG6", the differences between the median methylation of the contracted allele of the family members (generation II and III) correlate with their differences of the clinical severity and might be the molecular link to the more severe phenotype of the children. However, as differences between the median methylation levels are small and insignificant in Wilcoxon test, larger studies need to prove that such mild phenotypical differences are resolved by that region. Additionally, the whole distal FSHD locus need to be screened for the region that reflects the clinical status the best.

Our long-read sequencing data align with recent findings of other groups who developed similar methods for deciphering the epigenetic landscape of the FSHD locus.[7,8] Particularly, Fig. 1E and 1G show an ascending methylation gradient with low methylation at the proximal and high methylation at the distal end of the D4Z4 array (turquoise dots). The data available indicate that the methylation reaches a maximum at around 10 RU and keeps constant after this point (Fig. 1C, light blue line).[7,8] This finding is in line with the upper limit for pathogenicity of 10 RU in FSHD1 patients. In patients with pathogenic variants in epigenetic suppressor genes such as *SMCHD1*, *DMNT3B* and *LRIF1* (Fig. 1C, pink line), the slope of the methylation gradient is lower, which might be the reason that patients with repeat arrays longer than 10 RU develop FSDH2 if the array is not as long that the gradient reaches saturation. Other parameters are likely

to influence the slope of the gradient leading to the known phenotypical differences in FSHD patients with similar genetic features. One of these parameters could be a lowered parental methylation profile affecting activation and suppression of *DUX4* in the early embryonic development. Consequently, methylation levels in the descendants might be further reduced by mechanisms to be determined, causing anticipation, as reported by Zheng *et al*. These recent and initial findings by long-read sequencing show that the epigenetic landscape is the missing link between the FSHD phenotype and the underlying genetic parameter. This provides another line of evidence for methylation's superiority as a FSHD biomarker.

The verification of anticipation reported in the Zheng study and the findings in the family we studied have implications for clinical practice. First, it highlights the importance of studying all relatives, even asymptomatic ones, as they might develop mild symptoms at a higher age. Somatic mosaicism and the late manifestation of mild symptoms should be considered in parents of patients with *de novo* FSHD.

To conclude, there is evidence for clinical anticipation in FSHD1 due to lower distal methylation levels in the descendant generation as one possible mechanism. Additionally, as in the family we studied, somatic mosaicism is another mechanism that can lead to more severe phenotypes in descendant generations. Most importantly, distal methylation is a precise biomarker for the clinical status of FSHD patients, as also highlighted by Zheng *et al.* Therefore, the diagnosis of FSHD should generally be based on epigenetics rather than genetic parameters. Future studies need to focus on the possibility of predictive testing based on epigenetics, the development of long-read sequencing-based diagnostic methods - including the identification of the set of CpGs that is the most informative for FSHD – as well as on the continued analysis of the relationship between phenotype, genetic and epigenetic parameter.

Data availability

Anonymized data from this study are available from the corresponding author on reasonable request.

Acknowledgements

We want to thank all family members who studied in this article for their participation and support of our study. In addition, we thank Yi-Wen Chen (Center for Genetic Medicine Research, Children's National Hospital, Washington DC) for providing the sequences of the gRNAs to enrich the FSHD locus by CRISPR/Cas9.

Funding

No funding was received for this work.

Competing interests

All authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Zheng F, Qiu L, Chen L, et al. An epigenetic basis for genetic anticipation in facioscapulohumeral muscular dystrophy type 1. *Brain*. 2023, *accepted*.
2. Flanigan KM, Coffeen CM, Sexton L, Stauffer D, Brunner S, Leppert MF. Genetic characterization of a large, historically significant Utah kindred with facioscapulohumeral dystrophy. *Neuromuscul Disord*. 2001;11(6):525-529.
3. Lemmers RJLF, van der Wielen MJR, Bakker E, Padberg GW, Frants RR, van der Maarel SM. Somatic mosaicism in FSHD often goes undetected. *Ann Neurol*. 2004;55(6):845-850.
4. Erdmann H, Scharf F, Gehling S, et al. Methylation of the 4q35 D4Z4 repeat defines disease status in facioscapulohumeral muscular dystrophy. *Brain*. 2023;146(4):1388-1402.

5. Sacconi S, Briand-Suleau A, Gros M, et al. FSHD1 and FSHD2 form a disease continuum. *Neurology*. 2019;92(19):e2273.

6. Gilpatrick T, Lee I, Graham JE, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol*. 2020;38(4):433-438. doi:10.1038/s41587-020-0407-5

7. Hiramuki Y, Kure Y, Saito Y, 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.

8. Russell J Butterfield, Diane M Dunn, Brett Duval, Sarah Moldt, Robert B Weiss. Deciphering D4Z4 CpG methylation gradients in fascioscapulohumeral muscular dystrophy using nanopore sequencing. *bioRxiv*. [Preprint] doi:10.1101/2023.02.17.528868

Figure legends

Figure 1 Study of the epigenetic profile of the FSHD locus in a family with varying phenotypes. (**A**) Family tree of the family analyzed. (**B**) Region used for the determination of methylation levels within the most distal repeat unit (dark blue line) as well as the position of "CpG6" analyzed in the Zheng study (pink line). (**C**) Schematic and simplified model of the methylation gradient observed for the 4q35 D4Z4 repeat array in preliminary studies. (**D–G**) Epigenetic landscape of the FSHD locus determined within a sliding window of 250 base pairs (left) and box-plots of the methylation status of the CpGs within the distal repeat array for the contracted (dark blue) and uncontracted allele (turquoise) as well as for the combination of both alleles (light blue) for the grandfather (I:1, **D**), his daughter (II:1, **E**), granddaughter (III:1, **F**) and grandson (III:2, **G**). The x-axis indicates the chromosomal position as the distance to the polyadenylation signal (PAS) in exon 3 of *DUX4*.

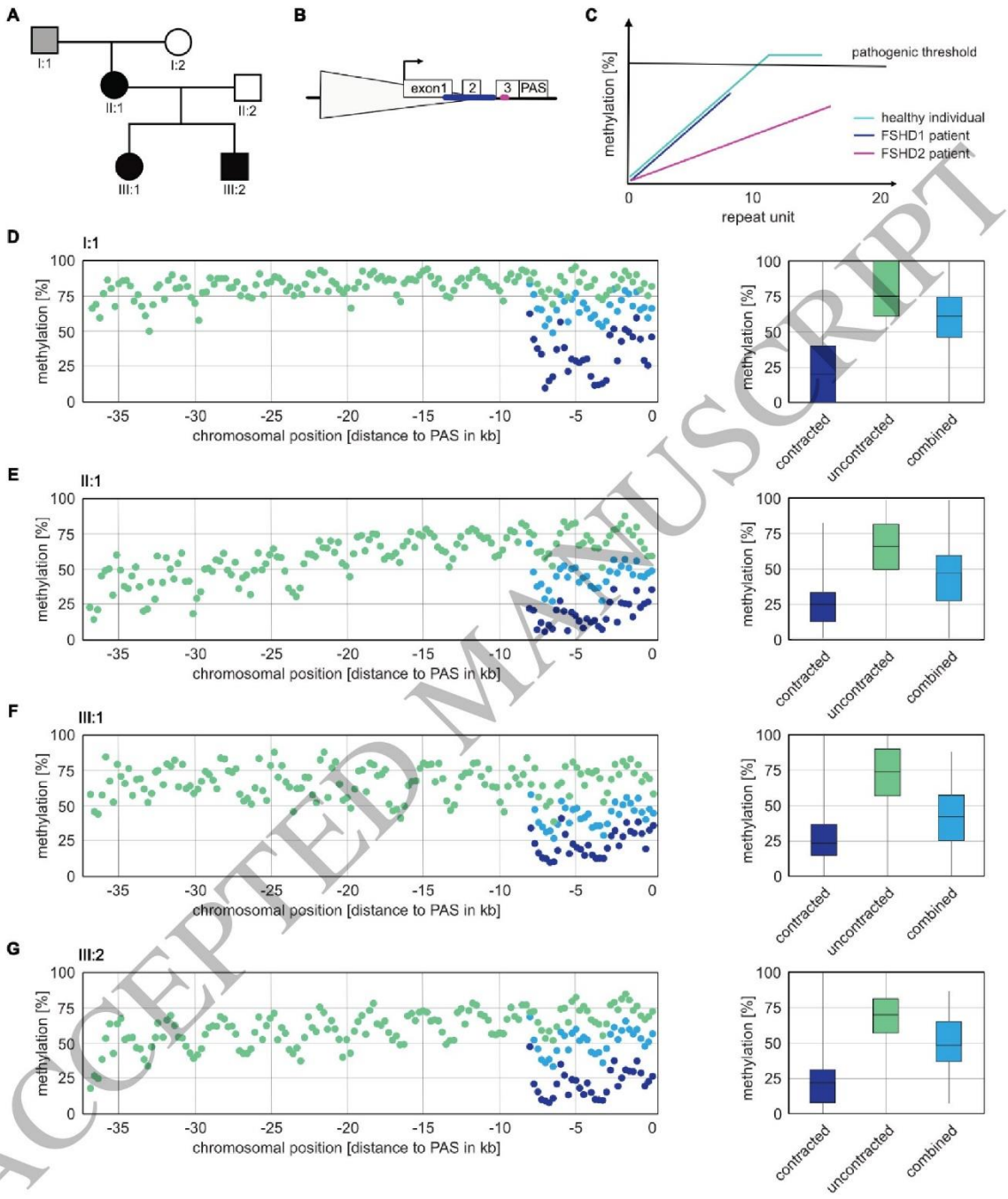

Figure 1
159x192 mm (x DPI)

Table 1 Clinical, genetic and epigenetic parameter determined for the multigenerational family

Characteristic	Patient I:1	Patient II:1	Patient III:1	Patient III:2
Clinical data				
Current age	73	46	22	24
Gender	male	female	female	male
CSS	0.5	5	3.5	4
Age-corrected CSS	14	217	318	333
Age of lower limb involvement	–	24	17	14
CSS in the age of 22	0	1.5	3.5	4
Genetic data				
Repeat size[a]	2 RU	2 RU	2 RU	2 RU
Haplotype	4qA/4qA	4qA/4qA	4qA/4qA	4qA/4qA
average distal methylation within *DUX4* in % (FSHD-MPA)[b]	52.4 (negative)	12.3 (positive)	14.9 (positive)	11.8 (positive)
median distal methylation in % (ONT long-read sequencing)[c]	61.3 (20.0/75.4)	47.4 (25.0/66.7)	42.0 (23.1/73.7)	48.3 (21.4/70.0)
Ratio read count contracted/uncontracted allele	1 : 3.8	1.3 : 1	1.3 : 1	1.1 : 1

[a]Determined by Southern blotting with linear gel electrophoresis and by ONT long-read sequencing. Size of the second allele extents 13 RU.
[b]Distal methylation ranging from the end of exon 1 to the end of intron 2 (chromosomal position chr4_KQ983257v1_fix: 201.476-202.069) assayed within the 4qA assay by FSHD-MPA.[4] Methylation is determined over both 4qA alleles. Parenthesis gives the diagnostic result for FSHD.
[c]Determined for the analogous region as for (b) by ONT long-read sequencing. First value represents the methylation over both alleles and the values in parenthesis the methylation of the contracted (first value) as well as of the uncontracted allele (second value).

Supplementary Material

Extraction of genomic DNA (gDNA)

Genomic DNA (gDNA) was obtained from total peripheral EDTA blood samples by extraction of white blood cells with a Biomek FX system (Beckman Coulter) using the NucleoMag® Blood 3 ml Kit (Machery-Nagel, #REF 744502.1) as per manufacturer's instructions. All DNA samples showed high purity as determined by optical density measurements (A260/A280 > 1.9 and A260/A230 > 2.0).

Library preparation and Oxford Nanopore Cas9-targeted sequencing

Library preparation was performed following the protocols of Gilpatrick *et al.* [1], Giesselmann *et al.*[2] and the Oxford Nanopore Cas9-targeted sequencing protocol as described in the following. In order to enrich the 4q35 locus, two specific crRNAs were used with one of them cutting upstream and the other downstream. Both, crRNAs and tracrRNAs, were purchased from IDT. An equimolar pool (100 µM) of crRNAs was prepared to enrich the regions of interest. For each target one pair (for *RFC1* two pairs) of crRNAs upstream and downstream to the ROI were used. Equimolar amounts (100 µM) of the crRNA pool and the tracrRNA (IDT) were mixed and diluted to 10 µM with nuclease-free duplex buffer (IDT). Incubation at 95°C for 5 minutes and subsequent cool down at room temperature for 10 minutes yielded the desired gRNA complex. Ribonucleoprotein (RNP) complex was formed by incubating a mixture of gRNA complex (10 µM), Hifi Cas9 Nuclease V3 (IDT, 64 µmol) and Cut Smart buffer (New England Biolabs) for 30 minutes at room temperature.

In parallel, 6 µg of input gDNA were dephosphorylated by incubation with Quick calf intestinal alkaline phosphatase (CIP, New England Biolabs) in 1x Cut Smart buffer (New England Biolabs) at 37°C for 10 minutes followed by enzyme deactivation at 80°C for 2 minutes. To the dephosphorylated DNA, the preformed RNP complex, dATP (New England Biolabs) and Taq polymerase (New England Biolabs) were added. The reaction mixture was incubated at 37°C for 15 minutes to enable the Cas9 reaction, before dA-tailing of the DNA ends was performed by incubation at 72°C for 5 minutes.

Adapter ligation was performed by combining ligation buffer (Oxford Nanopore SQK-LSK109 kit), nuclease-free water, NEBNext Quick T4 DNA ligase (New England Biolabs) and adapter mix (Oxford Nanopore SQK-LSK109 kit). The resulting mixture was incubated at room temperature for 10 minutes. DNA was purified by AMPure XP beads (Beckman Coulter, 0.3x) according to manufacturer's protocol in TE buffer (IDT, pH 8.0) and eluted in elution buffer (Oxford Nanopore SQK-LSK109 kit). The library was mixed with sequencing buffer (Oxford

Nanopore SQK-LSK109 kit) and loading beads (Oxford Nanopore SQK-LSK109 kit) before applying it on a primed flow cell (Oxford Nanopore FLO-MIN106D R9) for sequencing on the GridION X5 sequencer.

Analysis of Oxford Nanopore sequencing data

Base calling and methylation calling from ONT raw data was performed using Guppy basecaller (v6.3.9). Reads were aligned against the human reference genome (GRCh38/hg38) using Minimap2 (v2.17).[3,4] Reads mapping uniquely within the region chr4_KQ983257v1_fix:198,188-200,221 were extracted and haplotyped as reads originating from the contracted allele (based on the mapping start within repeat unit 2 as well as the presence of a specific sequence corresponding to the p13E11 element) or the uncontracted allele (all reads exceeding the length of the contracted reads). All other reads were discarded as they could not be assigned unambiguously to one allele or the other. Percent methylation per CpG has been extracted from BAM files using modbam2bed (v0.9.5).[5]

Supplementary Table 1. Mean read counts for the contracted and uncontracted alleles for all patients analyzed.

Patient	Read counts contracted allele	Read counts uncontracted allele	Combined read count
I:1	13	49	62
II:1	26	20	46
III:1	46	36	82
III:2	49	46	95

References

1. Gilpatrick T, Lee I, Graham JE, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol*. 2020;38(4):433-438.

2. Giesselmann P, Brändl B, Raimondeau E, et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat Biotechnol*. 2019;37(12):1478-1481. doi:10.1038/s41587-019-0293-x

3. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094-3100.

4. Oxford Nanopore Technologies. Guppy protocol: modified base calling. Accessed 05 June 2023. https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revz_14dec2018/modified-base-calling.

5. Github: modbam2bed. Accessed 05 June 2023. https://github.com/epi2me-labs/modbam2bed.

Acknowledgements

First and foremost, I would like to thank ANGELA ABICHT for the opportunity to prepare my thesis under her guidance. Thank you for providing these interesting topics, the outstanding and continuous support and encouragement far beyond this thesis, and the numerous hours of spontaneous discussion of research and papers. Furthermore, I would like to thank BENEDIKT SCHOSER for the dedicated supervision of my thesis, the extensive and outstanding support, and the always helpful advice in all kind of questions. SABINE KRAUSE I would like to thank for her helpful input and for completing my thesis committee.

I am grateful to DIETER WOLF for his great supervision and encouragement within the repeats project and to ELKE HOLINSKI-FEDER for her continuous support and interest in the research projects as well as her important inputs.

RAFAELA, MADALINA and VERONIKA I would like to thank for the great collaboration when establishing the repeat analysis. I am grateful to STEFFI for introducing me to FSHD and its diagnostics. I want to thank FLO for the outstanding and pleasant collaboration within the FSHD project. Thanks to ARIANE for answering questions, her support and for the amazing hikes at the weekends.

SABRINA A., I want to thank you for introducing me as a chemist to a genetics lab and answering all kinds of questions. For the helpful input and for always sharing interesting results from the routine diagnostics, I want to thank STEPHI and CHRISTINE. I am grateful to ANNA for her support and surveillance in establishing the methods. I want to thank MANU, MARTIN, KRISTINA, EVGENIA, AND THE WHOLE MGZ TEAM for the assistance and for performing several research experiments on top of the routine work. For the excellent collaboration in tasks and projects outside of research, I want to thank BIRGIT, BIENE, SABRINA L. and BEATE.

For teaching me as a scientist and for being outstanding mentors, I am grateful to ERIC JACOBSEN, SAMI LAKHDAR, DAVID LUPTON, HERBERT MAYR and ANDREAS RÖTHELI.

I want to thank HANNES and SILKE for embarking on the adventure of studying medicine as a second program and handling together the challenges and peculiarities of this program.

For financial support and a fellowship, I want to thank the Medical Faculty of LMU München.

I owe my deepest gratitude to all my friends, especially to BEN, HANNES, JAKOB, JANINA, KARO, KILIAN, MATHIAS, MINA, NIELS, SABRINA and SILKE, for the shared moments, the fun times, the mutual support, endless consultations on decisions and waiting for me when I am late again. I want to thank MY PARENTS, MY BROTHER ROBERT, MY GRANDMOTHER WALBURGA and MY GRANDPARENTS for their constant support and love.

Curriculum Vitae

removed for privacy

removed for privacy

removed for privacy