# *FGF14* repeat length and mosaic interruptions: modifiers of SCA27b?

Joshua Laß MSc[1†], Mirja Thomsen MSc[1†], Max Borsche MD[1,2†], Theresa Lüth PhD[1†], Julia C. Prietzsche BSc[1], Susen Schaake BSc[1], Andona Milovanović MD[3], Hannah Macpherson MSc[4,5], Emil K. Gustavsson PhD[4,5], Paula Saffie Awad MD[6,7], Nataša Dragašević-Mišković MD[3], Björn-Hergen Laabs PhD[8], Inke R. König PhD[8], Ana Westenberger PhD[1], Christopher E. Pearson PhD[9], Norbert Brüggemann MD[1,2], Christine Klein MD[1], Joanne Trinh PhD[1]

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

## Abstract

Deep intronic *FGF14* repeat expansions have been identified as a frequent genetic cause of late-onset cerebellar ataxias, explaining up to 30% of patients. Interruptions between repeats have previously been identified to impact the penetrance in other repeat expansion disorders. Repeat interruptions within *FGF14* have yet to be characterized in detail.

We utilized long-range PCR, Sanger sequencing, repeat-primed PCR, Nanopore, and PacBio sequencing to distinguish the repeat motifs, mosaicism, and number of repeat interruptions present in *FGF14*-related ataxia patients and unaffected individuals.

We identified 28 individuals with an expanded repeat length (≥250 repeats) in the FGF14 gene after a previous screening of 367 patients with late-onset ataxia. Additionally, a cohort of 192 unaffecteds was screened for repeat expansions in the *FGF14* gene where we found 12 expansion carriers. We applied advanced genetic methods to investigate the repeat motif. In total, the 40 individuals had expansions ranging from 232 to 486 repeats (*SD*=60) and 20 had repeat interruptions, including complex motifs such as *GAG*, GAA*GGA*, GAAGAA*A*GAA, GAAA*A*GAAGAA*G*GAAGAA*G*GAA, GAAA*A*GAAGAA*G*GAA, and *GCA*GAAGAAGAAGAA. We calculated the longest pure GAA length from the long-read data for all 40 individuals. When comparing the pure GAA tract between patients and unaffecteds, clusters were apparent based on greater or less than 200 repeats. Five ataxia patients with interruptions still had a remaining pure GAA expansion >200. We observed an association of the pure GAA length with age at onset ($p$=0.012, $R^2$=0.263,). Unaffected individuals had a longer

interruption length compared to the patients ($p$=0.010). Mosaic divergent repeat interruptions were discovered that affect motif length and sequence (mDRILS), which varied in number and mosaicism (frequency: 0.37-0.93). The mDRILS correlated with pure GAA length ($p$=0.022, $R^2$=0.334), with a higher mosaic frequency of interruptions in unaffecteds compared to patients (unaffecteds: 0.90; patients: 0.67; $p$=0.010).

We demonstrate that i) long-read sequencing is required to detect complex repeat interruptions accurately; ii) repeat interruptions in *FGF14* are mosaic, have various lengths, and start positions in the repeat tract and can thereby be annotated as mDRILS, which iii) enabled us to establish a categorization based on remaining pure GAA repeats quantifying the impact of mDRILS on pathogenicity or age at onset, dependent on the interruption length and position, with high accuracy. Finally, we iv) provide evidence that mosaicism stabilizes pure GAA repeats in interrupted *FGF14* repeat expansions.

**Author affiliations:**

1 Institute of Neurogenetics, University of Lübeck and University Hospital Schleswig-Holstein, 23538 Lübeck, Germany

2 Department of Neurology, University of Lübeck and University Hospital Schleswig-Holstein, 23538 Lübeck, Germany

3 Neurology Clinic, University Clinical Center of Serbia, Belgrade, 11000, Serbia

4 Department of Genetics and Genomic Medicine, UCL GOS Institute of Child Health, London, WC1E 6BT, UK

5 Department of Neurodegenerative Diseases, Institute of Neurology, UCL, London, WC1E 6BT, UK

6 Programa de Pós-Graduação em Ciências Médicas, Universidade Federal do Rio Grande do Sul, Porto Alegre, 90035003, Brazil

7 Servicio de Neurología, Clínica Santa María, Santiago, 7520349, Chile

8 Institute of Medical Biometry and Statistics, University of Lübeck, 23538 Lübeck, Germany

9 The Hospital for Sick Children, Genetics & Genome Biology; Department of Molecular Genetics, University of Toronto, Toronto, ON M5G 0A4,Canada

Correspondence to: Dr. Joanne Trinh

Institute of Neurogenetics,

University of Lübeck and University Hospital Schleswig-Holstein,

Campus Lübeck, BMF, Building 67, Room 12,

Ratzeburger Allee 160, 23538 Lübeck, Germany

Corresponding author's phone and fax: +49-451-3101-8202 / +49-451-3101-8204

Corresponding author's e-mail address: joanne.trinh@uni-luebeck.de

**Running title**: *FGF14* repeat length and interruptions

**Keywords:** repeat expansion; interruptions; SCA27b, long-read sequencing

# Introduction

Cerebellar neurodegenerative ataxias are commonly associated with tandem repeat expansions. The recently identified deep intronic repeat expansions of $(GAA)_n$ in the *FGF14* (Fibroblast Growth Factor) gene cause SCA27b.[1,2] The expansion is located in the first intron and leads to a reduction of *FGF14* expression, similar to the GAA expansion of Friedrich's ataxia.[3] In SCA27b, a threshold of ≥250 repeats is considered disease-causing, whereby expansions between 250 and 300 repeats are likely pathogenic with reduced penetrance, and expansions with ≥300 repeats are fully penetrant.[1,2] Different studies suggested repeat interruptions in *FGF14* to be non-pathogenic.[2,4-7] However, evidence for interruption non-pathogenicity relates mainly to family studies thus far, and more in-depth analysis of the specific repeat expansion sequence, motif and interruption length has yet to be performed. One large study used Nanopore sequencing in a subset of individuals and found a similar frequency of GAA*GGA* interruptions in patients and controls, where they concluded that the interruption was non-pathogenic.[6] However, interruptions have not been further characterized in terms of interruption length, position, and mosaicism. Finally, the role of shorter repeat interruptions and the impact of their position within the repeat has yet to be deciphered in detail.

Long-read sequencing has revealed an increased appreciation of the number of loci with expanded repeats, the sequence motifs, and their purity.[8] While many repeat expansion disorders are now characterized, with known cis-elements flanking the unstable repeat, including FGF14 repeat tract purity, modifiers of disease expression are largely unknown.[9,10] The hexanucleotide repeat relevant for X-linked dystonia-parkinsonism (XDP) consists of the hexanucleotide $(AGAGGG)_n$ sequence repeated 30 to 55 times and is a strong genetic modifier of age at onset (AAO).[11,12] A novel mosaic repeat motif pattern that deviates from the known hexanucleotide repeat motif, both in motif length and sequence (mDRILS), modifies repeat stability in XDP.[13] This genetic association in XDP demonstrates the importance of somatic mosaic genotypes and the biological plausibility of multiple germline and somatic modifiers, which may collectively contribute to repeat instability. These variations may remain undetected without assessment of single molecules. Data on the correlation between repeat length and age at onset varies between *FGF14* studies. Even studies with large sample sizes could not consistently demonstrate such an association, which contradicts with the general understanding of repeat expansion disorders.[14-16]

Thus, in-depth genetic methods might shed new light on this relation, which is of importance for clinical care and patient counseling.

In this study, we aimed to delineate the repeat tract sequence, mosaicism, and number of repeat interruptions in *FGF14*-related late-onset cerebellar ataxia patients and unaffecteds (Fig. 1). Our findings show that: i) long-read sequencing is required to detect complex repeat interruptions accurately; ii) repeat interruptions in *FGF14* are mosaic, have various lengths, and start positions in the repeat tract and can thereby be annotated as mDRILS, which iii) enabled us to establish a categorization based on remaining pure GAA repeats quantifying the impact of mDRILS on pathogenicity or age at onset, dependent on the interruption length and position, with high accuracy. Finally, we iv) provide evidence that mosaicism stabilizes pure GAA repeats in interrupted *FGF14* repeat expansions.

# Methods

## Participant recruitment

The sample consisted of *n*=367 patients with ataxia of unknown genetic cause and *n*=192 unaffecteds. Patients were recruited at the outpatient clinics at the tertiary referral centers for ataxia and vertigo at the Department of Neurology, University of Lübeck, Lübeck, Germany, at the Department for Neurodegenerative Diseases and Movement Disorders at the Neurology Clinic at the University Clinical Center of Serbia, Belgrade, Serbia and at a movement disorders center in Santiago, Chile. The main inclusion criteria was the presence of progressive cerebellar ataxia without a known cause. Patients with secondary forms of ataxia (lesion, toxic, inflammatory, and paraneoplastic) and known repeat-expansion SCAs (SCA1, 2, 3, 6, and 17) were excluded. Peripheral blood samples were collected from participants following informed consent and approval by the local ethics committees. Most of the patients with SCA27b were previously reported (Table 2). [1,17-19] Genomic DNA was extracted using the QIAamp DNA Blood Mini Kit (Qiagen) according to the manufacturer's instructions.

## Long-range PCR

Long-range PCR was performed to amplify the GAA repeat region of *FGF14* using flanking primers FGF14-F (5’-CAGTTCCTGCCCACATAGAGC-3') and FGF14-R (5’-AGCAATCGTCAGTCAGTGTAAGC-3'). The predicted product is 315 bp long and includes 50 GAA repeats based on the hg38 reference. A 25 µL PCR reaction was set up using the Platinum SuperFi II Master Mix (Thermo Fisher Scientific), 5% DMSO, 0.5 µM forward and reverse primers, and 100 ng of genomic DNA. A touchdown PCR protocol was designed (Supplementary Table 1). PCR products were visualized using agarose gel electrophoresis. For fragment analysis, an M13F-tail (CACGACGTTGTAAAACGAC) was attached to the forward primer, and a third FAM-labeled primer (FAM-M13F) was added to analyze products by capillary electrophoresis on a Genetic Analyzer 3500XL (Applied Biosystems). Fragment sizes were determined using GeneMapper software (Applied Biosystems) with a GeneScan™ 1200 LIZ™ Size Standard. For samples with repeat lengths greater than 250, corresponding to products with a size greater than ~900 bp, repeat-primed PCR (RP-PCR), Sanger sequencing, Nanopore, and PacBio sequencing were conducted.

## Repeat-primed PCR (RP-PCR)

RP-PCR was used to analyze the repeat composition and test for possible interruptions of the $(GAA)_n$ repeat. The design included a locus-specific primer (FGF14-R), a repeat-containing primer with an M13F-tail designed to amplify the GAA/TCC or GAAGGA/CTTCCT motif (FGF14-RP-GAA: 5'-M13F-CTTCTTCTTCTTCTTCTTCTT-3'; FGF14-RP-GAAGGA: 5'-M13F-TCCTTCTCCTTCTCCTTC-3'), and a FAM-labeled M13-primer (FAM-M13F). Cycling conditions are displayed in Supplementary Table 2. The products were analyzed on a Genetic Analyzer 3500XL, producing a ladder-like pattern indicative of repeat expansions of the respective motif.

## Nanopore sequencing

Two Nanopore sequencing workflows were used to analyze the repeat expansion in *FGF14*. The first approach involved PCR amplification of the repeat tract, utilizing the long-range PCR products (see above) as an input.

To sequence all individuals on a single R10.4.1 flow cell, the long-range PCR products were multiplexed using the Native Barcoding Kit NBD114-96 (ONT). A 200 fmol input of the amplified PCR product was used. Library preparation was performed with the SQK-LSK114 kit (ONT). The final library was subsequently loaded onto an R10.4.1 flow cell (FLO-MIN114) and sequenced on a GridION platform as previously described.[20]

A Cas9 enrichment was performed to validate the results of the first PCR-amplification-based approach as previously described.[13] In total, four CRISPR RNAs (crRNA) were designed using the ChopChop tool (https://chopchop.cbu.uib.no, accessed on 10 October 2023) (Supplementary Table 3). Two crRNAs located upstream and two crRNAs downstream of the *FGF14* repeat expansion were used for efficient DNA cleavage. The enrichment resulted in a 3.2 kb product, which served as input for the library preparation using the SQK-LSK114 kit. The final library was sequenced on a GridION platform with an R10.4.1 flow cell.

## Sanger sequencing

Sanger sequencing was performed on the long-range PCR products to confirm their specificity and the repeat motif sequence. Sequencing reactions were prepared using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) with primers FGF14-F and FGF14-R. The sequencing conditions included 25 cycles of 96 °C for ten seconds, 55 °C for five seconds, and 60 °C for three minutes. Sequencing products were purified using sodium acetate precipitation and analyzed on a Genetic Analyzer 3500XL (Applied Biosystems).

## Pacbio sequencing

For each sample, a 1.3X cleanup was performed using PacBio SMRTbell beads. Quality control checks were conducted using the Qubit 1x dsDNA HS kit (Thermo Fisher) for concentration and the Femto Pulse with the Genomic DNA 165kb Kit (Agilent) for fragment analysis. Samples were then processed using an adapted version of the standard PacBio multiplexed amplicon library protocol (102-359-000) and the SMRTbell Prep Kit 3.0. A final concentration of 78.38 fmol of each sample was added, and reaction volumes were divided by six for End Repair and A-tailing and by 5.4 for Adapter Ligation. During ligation, samples were barcoded with unique SMRTbell adapters, and an additional step of incubation at 65 °C for ten minutes was added before the four °C hold. Samples were pooled before a 1.2X Pronex bead clean and elution in 40 µL. Nuclease treatment was performed in a total volume of 50 µL at 37 °C for 15 min and then held at four °C. A second 1.2X Pronex bead cleanup was carried out with final elution in 15 µL. The final library underwent quality assessment using the Qubit and Femto Pulse. Sequencing was carried out as a standard amplicon library on the Sequel IIe using binding kit 3.2. The library was loaded at 70 pM with a movie time of 30 hours.

## Bioinformatic analysis

For Nanopore sequencing, base-calling was performed using Dorado's (version 7.2.13) super accuracy model (dna_r10.4.1_e8.2_400bps_sup@v4.3.0). Read quality was analyzed with the Nanostat software (version 1.5.0). For PacBio sequencing, the BAM files were demultiplexed with the software Lima (version 2.9.0) and converted to FASTQ files using Samtools (version 1.15). For both Nanopore and PacBio sequencing, Minimap2 (version 2.22) was used to align the reads to the reference sequence with parameters for long Nanopore sequencing reads.[21] SAM-to-

BAM conversion and BAM file handling were conducted with the Samtools software (version 1.15).[22] The next step was the sorting and indexing of the reads with Samtools.

The trinucleotide repeats were analyzed with the "Noise-Cancelling Repeat Finder" (NCRF, version 1.01.02).[23] Only reads with a maximum noise of 80% and a minimum of 15 detected repeat units were included in the analysis. A minimum threshold of 200 repeats was set to filter for the long allele. A maximum threshold of 100 repeats was applied to assign reads to the short allele. The repeat length was determined with the median repeat length of all reads. The detection of interruptions and their frequency were done with the summary output in R, as previously described.[20] The interruption frequency was then calculated by dividing the number of reads with interruption by the total number of reads for that individual. The scripts and reference file are provided at: https://github.com/joshua21997/FGF14-repeat-expansion.

### Statistical analysis

The graphical representation and statistical analyses were performed in R (version 4.3.0) and Biorender. Visualization was done with the ggplot2 package (version 3.4.4). Mann-Whitney U-tests were performed for pairwise comparisons between patients and unaffecteds, with the significance level set to $\alpha = 0.05$ for the Mann-Whitney U-tests. The adjusted significance level for multiple testing based on Bonferroni is $\alpha = 0.013$. To compare the different methods, Bland-Altman plots were used. Additionally, correlation analysis using a linear regression model implemented with the lm-function was performed.

## Data availability

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

# Results

## Long-read sequencing can robustly detect FGF14 repeat length

Screening of *n*=367 patients with late-onset ataxia and *n*=192 unaffecteds by fragment analysis identified 40 with an expanded repeat length (≥250 repeats) in the *FGF14* gene, of which 28 were patients (266-476 repeats) and twelve unaffecteds (256-326 repeats) (Table 1 and 2). Most of the patients with late-onset ataxia were included in previous publications (Table 2). [1,17-19] Nanopore sequencing was performed on these 40 individuals to investigate repeat length and composition. The median read length of the long allele determined from Nanopore sequencing was 1012 bp (IQR:899-1112 bp), and the median q-score was 14.6 (IQR:14.2-15.6). The detected repeat number with Nanopore sequencing ranged from 232 to 414, and the median repeat length was 309 (IQR:279-370) (Table 2). The fragment analysis detected comparable repeat lengths for the long allele (median repeat length: 322, IQR:286-376) (Fig. 2A). The Bland-Altman analysis between Nanopore sequencing and fragment analysis showed a small bias, and three of 40 individuals were out of the limits of agreement (Fig. 2B).

PacBio sequencing was performed on 16/40 individuals with the *FGF14* repeat expansion as a second validation of the repeat length. The median q-score of the PacBio sequencing was 38.7 (IQR:37.7-41.5), and the median read length was 1162 bp (IQR:1108-2310 bp). The median repeat number of the long allele with PacBio sequencing was 318 (IQR:273-342).

Comparing the repeat tract length, the PacBio sequencing results were concordant with the Nanopore sequencing results (one of 16 individuals was outside the limits of agreement) and with the fragment analysis results (one of 16 individuals was outside the limits of agreement) (Fig. 2C-F).

## The pure GAA tract predicts disease manifestation and age at onset

Next, we assessed the pure GAA tract length, without interruptions, in patients and unaffected. The range of the pure GAA tract was 11 to 486 repeat units across these individuals. For individuals with interruptions, the longest GAA tract was used for further analyses. The comparison between patients and unaffecteds resulted in the identification of four distinct clusters (Fig. 3A and 4A). Recent literature has identified SCA27b patients with a lower repeat number than 250.[6,24] Therefore, we apply a threshold of 200 pure GAA repeat units. The median

AAO of SCA27b patients is 60 years (range 21-87 years).[25] The first group consisted of affected patients with a pure GAA length of >200 repeat units and diagnosed with SCA27b (affected, Fig. 3A and 4A indicated in red). In the second group were unaffecteds with a pure GAA tract >200, indicating pre-manifesting carrier status (pre-manifesting, Fig. 3A and 4A indicated in yellow). Of note, these five individulas had an age at examination of 35, 50, 51, 53, and 80, respectively. The third group consisted of unaffecteds with a pure GAA length <200 and no diagnosis (unaffected, Fig. 3A and 4A indicated in blue). The last group consisted of patients with a pure GAA tract <200 repeat units (other ataxia, Fig. 3A and 4A indicated in green). Next, we tested the relationship between the pure GAA lengths and AAO. There is only a correlation between GAA repeat length and AAO ($p$=0.012, $R^2$=0.262) (Fig. 4B) in patients with a pure GAA tract >250 repeat units.

## Novel repeat interruptions were found through long-read sequencing

Seven different repeat interruptions were detected (Table 2). Among these, the most frequently observed interruption motif was (GAA*GGA*)$_n$, characterized by a single adenosine to guanine conversion (A-to-*G*). This interruption motif was found in twelve individuals (six affected and six unaffected). The (GAA*GGA*)$_n$ interruption motif was repeated two to 158 times across people.

The repeat interruption was found at the 5'end of the repeat tract, (GAA)$_{1\text{-}20}$(GAA*GGA*)$_{2\text{-}158}$ (GAA)$_n$.

Another variant repeat motif was (GAA)$_3$GAAA*A*GAAGAA*G*GAAGAA*G*GAA(GAA)$_n$. This motif was identified in one unaffected and included one deletion and two insertions. A similar motif, (GAA)$_3$*G*AAA*A*GAAGAA*G*GAA(GAA)$_n$, was identified in another unaffected. The shortest interruption motif observed was (GAA)$_7$(*GAG*)$_3$(GAA)$_n$, detected in one patient.

Two motifs were exclusively detected by Nanopore sequencing. One motif was (GAA)$_5$(GAAGAA*A*GAA)$_3$(GAA)$_n$, which resulted from an insertion of an adenosine. The other motif was (GAA)$_4$(GAA*GAG*)$_2$(GAA)$_n$. Both motifs were detected in patients.

The most complex motif was (GAA)$_{22}$(*GCA*GAAGAAGAA)$_3$(*GCA*GAA)(*GCA*GAAGAAGAAGAA)$_{24}$(*GCA*GAA)$_{12}$ (*GCAGCA*GAA)$_3$(*GCA*GAA)$_{13}$(*GCAGCA*GAA*GCA*GAA)$_4$(*GCA*GAAGAAGAAGAA)$_6$ (GAA)$_{22}$*GAG*(GAA)$_n$. This motif was observed in two affected members and one unaffected

member of one family. However, the affected members of this family have a phenotype that differs from reported *FGF14* SCA27b patients.[4]

## Mosaic divergent repeat interruptions are associated with the pure GAA stability

Given the detection of mDRILS in XDP, we investigated the mosaicism of *FGF14* repeat interruptions in the long-read data. We calculated the mosaic frequency of the repeat motifs for each individual. The mosaicism obtained from Nanopore sequencing was similar to that from PacBio sequencing ($p$=0.017, $R^2$=0.363) (Fig. 5A,B). We utilized PacBio sequencing data for mosaic frequency calculations due to its higher q-score. Notably, the repeat interruption mosaicism was negatively associated with pure GAA length ($p$=0.022, $R^2$=0.334) (Fig. 4C). The higher the mosaic frequency, the shorter the pure GAA length.

To compare the interruption frequency between unaffecteds and patients, we only included patients with the SCA27B phenotype; thus, a previously published Chilean family with a different phenotype was excluded.[19] Unaffected individuals had higher interruption frequencies (0.90, IQR:0.87-0.91) compared to patients (0.67, IQR:0.54-0.70) (p=0.010, R=-2.576) (Fig. 6A). The most common interruption motif (GAA*GGA*)$_n$ had a mosaic frequency ranging from 0.37 to 0.93. Patients with (GAA*GGA*)$_n$ exhibited a lower median frequency of 0.61 (IQR:0.5-0.72) compared to unaffecteds, who had a median frequency of 0.90 (IQR:0.89-0.92) ($p$=0.018, $Z$=2.359) (Fig. 6B). The mosaic frequencies of the GAAA*A*GAAGAA*G*GAAGAA*G*GAA motif and the GAAA*A*GAAGAA*G*GAA motif were similar, at 0.82 and 0.84, respectively. The shortest interruption motif, *GAG*, had a slightly lower frequency of 0.70 compared to the other motifs. In contrast to the other motifs, the most complex motif had the overall lowest frequency with 0.46.

Further investigation of repeat interruptions was performed to rule out DNA contamination from the non-expanded short allele. The median repeat number on the short allele was 18 (IQR:17-19). Repeat interruptions on the short allele were seen in three individuals, two patients, and one unaffected (Supplementary Table 4). However, the interruption motifs in the short allele did not overlap with the expanded allele patterns.

## Discussion

Following our discovery of mDRILS in X-linked dystonia-parkinsonism,[13] we now observe that mDRILS can also act as modifiers of penetrance and age of onset (AAO) in a much more common condition, SCA27b. We consider this an important finding of translational relevance, as it suggests a novel, shared mechanism across different repeat expansion disorders, that has the potential to predict disease manifestation in individual repeat expansion carriers in a personalized fashion and represents a somatic phenomenon that may be amenable to environmental and lifestyle changes.

For the analysis of *FGF14* intronic repeat expansions, we propose a new concept to investigate and interpret their potential pathogenicity and challenge the previously held, more simplistic view of repeat interruptions, abolishing the pathogenic effect of the expanded repeat in general. Specifically, we propose four different scenarios, all related to the length of the pure GAA repeat tract: first, people with ataxia aged >60 years and an uninterrupted *FGF14* intronic repeat length of >200 repeats have SCA27b, despite the presence of an interrupted repeat, typically at the 5' end of the expansion. Second, people with the same molecular constellation and age <60 years are considered to be in their premanifesting phase of SCA27b. Third, people with repeat interruptions resulting in a pure GAA repeat tract of <200 will not go on to develop SCA27b. Fourth, in patients previously presumed with SCA27b and carrying a large repeat interruption with <200 uninterrupted GAA repeats, the diagnosis needs to be revisited and likely revised to another type of ataxia (Fig. 2A). An example is the Chilean family with two members previously diagnosed with SCA27b but with a phenotype that differs from reported *FGF14* SCA27b patients,[4] where we have identified a pure GAA tract with 65 repeat units. Otherwise, our data implies that not all interruptions can be considered non-pathogenic, as, for example, five ataxia patients with interruptions still had a remaining pure GAA expansion >200. As many studies screened in unspecified ataxia cohorts for repeat expansions in *FGF14* to debunk the clinical phenotype of SCA27b without considering interruptions, our data demonstrates that 14% (4/28) of our ataxia patients with *FGF14* repeat expansions had rather non-pathogenic expansions if the presence of mDRILS is taken into account. Consideration of the pure GAA repeat tract and mDRILS is required, especially if the phenotype is not entirely consistent with the typical clinical picture of SCA27b (i.e. pure late-onset cerebellar ataxia +/- episodic features).

Notably, this situation in SCA27b extends beyond our previous findings in X-linked dystonia-parkinsonism, where the pathogenic insertion in the *TAF1* gene is the clear-cut cause in all affected individuals, and mDRILS in the hexanucleotide repeats within this insertion act as a modifier of AAO only. Intriguingly, in addition to the penetrance-determining length of the pure GAA tract in SCA27b, we can assume a similar AAO-modifying effect by mDRILS in *FGF14* as well, with a higher mDRILS frequency resulting in shorter uninterrupted GAA tracts, which, in turn, are associated with a later AAO (Fig. 3B). As a relationship between repeat length and age at onset is well-established in the field of repeat expansion disorders and, therefore, is expected to occur likewise in *FGF14*-related disease, our study shows the importance of taking into account mDRILS to better detect phenotypic correlations and impact on repeat stability.[14]

The mDRILS in both the XDP-relevant and *FGF14* repeats may affect repeat instability by modifying the propensity to form unusual mutagenic DNA structures, as observed with the interruptions of *FMR1* (FXS) and *ATXN1* (SCA1).[26] mDRILS may also lead to pathogenic variations of splice forms, translation (exonization), repeat-associated non-AUG (RAN) translation, ribosomal frameshifting, ribosomal pausing, transcriptional slippage in the repeat, or repeat instability.[13,27-32]

Overall, all methods (RP-PCR, Sanger sequencing, Nanopore sequencing, and PacBio sequencing) were concordant and detected the same interruptions. We confidently identified five different repeat interruption motifs within *FGF14.* Nanopore sequencing analysis performed thus far in *FGF14*-related disease could demonstrate that interruptions are present in both unaffected and ataxia patients but have not yet explored 1) the interruptions' position and 2) their impact on repeat expansion pathogenicity at the individual level.[6]

Limitations of our study include the relatively small sample size compared to other studies. Moreover, we cannot independently support our estimate of pathogenicity based on advanced genetic methods, as we do not investigate FGF14 expression, and further biomarkers robustly distinguishing SCA27b from other ataxias are not available to date. Thus, the development of such, preferably easy-accessible biomarkers would be highly beneficial to the field.

In conclusion, this study highlights the importance of an in-depth, multi-method assessment of repeat tract purity in repeat expansion disorders. The correlation of mosaic interruptions with the repeat stability and indirectly patient status suggests a protective effect of mDRILS. Long-read sequencing can uncover the length of the pure GAA motif along with repeat interruptions.

However, it is warranted that the assessment of repeat interruptions, the pure GAA tract, and mDRILS be extended to other repeat expansion disorders and that their potential protective mechanism(s) be elucidated thoroughly. While our findings refine both the diagnosis of (*FGF14*-related) ataxia and the prognosis in non-manifesting or not-yet-manifesting repeat expansion carriers, it also highlights the necessity of including repeat interruption analysis not only in the research setting but also in diagnostic testing for SCA27b to avoid false-positive or false-negative testing results and interpretation thereof.

## Acknowledgment

We express our deepest gratitude to the families and patients who have participated in this study.

## Funding

This study was supported by the German Research Foundation (FOR2488 to J.T., C.K., Heisenberg grant, J.T., BR4328.2-1, GRK1957, N.B.) and the Else Kröner Fresenius Foundation (EKFS, J.T.).

## Competing interests

C.K. serves as a medical advisor to Centogene, Takeda, and Retromer Therapeutics and received speaking honoraria from Desitin and Bial. A.W. serves as an advisor for medical writing to CENTOGENE GmbH. N.B. received honoraria from Abbott, Abbvie, Biogen, Biomarin, Bridgebio, Centogene, Esteve, Ipsen, Merz, Teva, and Zambon. M.B. receives honoraria by Bial. The remaining authors report no disclosures.

## Supplementary material

Supplementary material is available online.

# References

1. Rafehi H, Read J, Szmulewicz DJ, et al. An intronic GAA repeat expansion in FGF14 causes the autosomal-dominant adult-onset ataxia SCA50/ATX-FGF14. *Am J Hum Genet*. Jan 5 2023;110(1):105-119. doi:10.1016/j.ajhg.2022.11.015
2. Pellerin D, Danzi MC, Wilke C, et al. Deep Intronic FGF14 GAA Repeat Expansion in Late-Onset Cerebellar Ataxia. *N Engl J Med*. Jan 12 2023;388(2):128-141. doi:10.1056/NEJMoa2207406
3. Ohshima K, Montermini L, Wells RD, Pandolfo M. Inhibitory effects of expanded GAA.TTC triplet repeats from intron I of the Friedreich ataxia gene on transcription and replication in vivo. *J Biol Chem*. Jun 5 1998;273(23):14588-95. doi:10.1074/jbc.273.23.14588
4. Pellerin D, Iruzubieta P, Tekgul S, et al. Non-GAA Repeat Expansions in FGF14 Are Likely Not Pathogenic-Reply to: "Shaking Up Ataxia: FGF14 and RFC1 Repeat Expansions in Affected and Unaffected Members of a Chilean Family". *Mov Disord*. Aug 2023;38(8):1575-1577. doi:10.1002/mds.29552
5. Ouyang R, Wan L, Pellerin D, et al. The genetic landscape and phenotypic spectrum of GAA-FGF14 ataxia in China: a large cohort study. *EBioMedicine*. Apr 2024;102:105077. doi:10.1016/j.ebiom.2024.105077
6. Mohren L, Erdlenbruch F, Leitao E, et al. Identification and characterisation of pathogenic and non-pathogenic FGF14 repeat expansions. *Nat Commun*. Sep 3 2024;15(1):7665. doi:10.1038/s41467-024-52148-1
7. Ando M, Higuchi Y, Yuan J, et al. Clinical variability associated with intronic FGF14 GAA repeat expansion in Japan. *Ann Clin Transl Neurol*. Jan 2024;11(1):96-104. doi:10.1002/acn3.51936
8. Gall-Duncan T, Sato N, Yuen RKC, Pearson CE. Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences. *Genome Res*. Jan 2022;32(1):1-27. doi:10.1101/gr.269530.120
9. Cleary JD, Pearson CE. The contribution of cis-elements to disease-associated repeat instability: clinical and experimental evidence. *Cytogenet Genome Res*. 2003;100(1-4):25-55. doi:10.1159/000072837
10. Pellerin D, Del Gobbo GF, Couse M, et al. A common flanking variant is associated with enhanced stability of the FGF14-SCA27B repeat locus. *Nat Genet*. Jul 2024;56(7):1366-1370. doi:10.1038/s41588-024-01808-5
11. Aneichyk T, Hendriks WT, Yadav R, et al. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell*. Feb 22 2018;172(5):897-909 e21. doi:10.1016/j.cell.2018.02.011
12. Rakovic A, Domingo A, Grutz K, et al. Genome editing in induced pluripotent stem cells rescues TAF1 levels in X-linked dystonia-parkinsonism. *Mov Disord*. Jul 2018;33(7):1108-1118. doi:10.1002/mds.27441
13. Trinh J, Luth T, Schaake S, et al. Mosaic divergent repeat interruptions in XDP influence repeat stability and disease onset. *Brain*. Mar 1 2023;146(3):1075-1082. doi:10.1093/brain/awac160
14. Paulson H. Repeat expansion diseases. *Handb Clin Neurol*. 2018;147:105-123. doi:10.1016/B978-0-444-63233-3.00009-9

15. Wilke C, Pellerin D, Mengel D, et al. GAA-FGF14 ataxia (SCA27B): phenotypic profile, natural history progression and 4-aminopyridine treatment response. *Brain*. Oct 3 2023;146(10):4144-4157. doi:10.1093/brain/awad157
16. Iruzubieta P, Pellerin D, Bergareche A, et al. Frequency and phenotypic spectrum of spinocerebellar ataxia 27B and other genetic ataxias in a Spanish cohort of late-onset cerebellar ataxia. *Eur J Neurol*. Dec 2023;30(12):3828-3833. doi:10.1111/ene.16039
17. Borsche M, Thomsen M, Szmulewicz DJ, et al. Bilateral vestibulopathy in RFC1-positive CANVAS is distinctly different compared to FGF14-linked spinocerebellar ataxia 27B. *J Neurol*. Feb 2024;271(2):1023-1027. doi:10.1007/s00415-023-12050-0
18. Milovanovic A, Dragasevic-Miskovic N, Thomsen M, et al. RFC1 and FGF14 Repeat Expansions in Serbian Patients with Cerebellar Ataxia. *Mov Disord Clin Pract*. Jun 2024;11(6):626-633. doi:10.1002/mdc3.14020
19. Saffie Awad P, Lohmann K, Hirmas Y, et al. Shaking Up Ataxia: FGF14 and RFC1 Repeat Expansions in Affected and Unaffected Members of a Chilean Family. *Mov Disord*. Jun 2023;38(6):1107-1109. doi:10.1002/mds.29390
20. Luth T, Labeta J, Schaake S, et al. Elucidating Hexanucleotide Repeat Number and Methylation within the X-Linked Dystonia-Parkinsonism (XDP)-Related SVA Retrotransposon in TAF1 with Nanopore Sequencing. *Genes (Basel)*. Jan 11 2022;13(1)doi:10.3390/genes13010126
21. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. Sep 15 2018;34(18):3094-3100. doi:10.1093/bioinformatics/bty191
22. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. Aug 15 2009;25(16):2078-9. doi:10.1093/bioinformatics/btp352
23. Harris RS, Cechova M, Makova KD. Noise-cancelling repeat finder: uncovering tandem repeats in error-prone long-read sequencing data. *Bioinformatics*. Nov 1 2019;35(22):4809-4811. doi:10.1093/bioinformatics/btz484
24. Hengel H, Pellerin D, Wilke C, et al. As Frequent as Polyglutamine Spinocerebellar Ataxias: SCA27B in a Large German Autosomal Dominant Ataxia Cohort. *Mov Disord*. Aug 2023;38(8):1557-1558. doi:10.1002/mds.29559
25. Pellerin D, Danzi M, Renaud M, et al. GAA-FGF14-Related Ataxia. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, eds. *GeneReviews((R))*. 1993.
26. Yousuf A, Ahmed N, Qurashi A. Non-canonical DNA/RNA structures associated with the pathogenesis of Fragile X-associated tremor/ataxia syndrome and Fragile X syndrome. *Front Genet*. 2022;13:866021. doi:10.3389/fgene.2022.866021
27. Stochmanski SJ, Therrien M, Laganiere J, et al. Expanded ATXN3 frameshifting events are toxic in Drosophila and mammalian neuron models. *Hum Mol Genet*. May 15 2012;21(10):2211-8. doi:10.1093/hmg/dds036
28. Gaspar C, Jannatipour M, Dion P, et al. CAG tract of MJD-1 may be prone to frameshifts causing polyalanine accumulation. *Hum Mol Genet*. Aug 12 2000;9(13):1957-66. doi:10.1093/hmg/9.13.1957
29. Toulouse A, Au-Yeung F, Gaspar C, Roussel J, Dion P, Rouleau GA. Ribosomal frameshifting on MJD-1 transcripts with long CAG tracts. *Hum Mol Genet*. Sep 15 2005;14(18):2649-60. doi:10.1093/hmg/ddi299
30. Aviner R, Lee TT, Masto VB, Li KH, Andino R, Frydman J. Polyglutamine-mediated ribotoxicity disrupts proteostasis and stress responses in Huntington's disease. *Nat Cell Biol*. Jun 2024;26(6):892-902. doi:10.1038/s41556-024-01414-x

31. Stein KC, Morales-Polanco F, van der Lienden J, Rainbolt TK, Frydman J. Ageing exacerbates ribosome pausing to disrupt cotranslational proteostasis. *Nature*. Jan 2022;601(7894):637-642. doi:10.1038/s41586-021-04295-4

32. Parsons MA, Sinden RR, Izban MG. Transcriptional properties of RNA polymerase II within triplet repeat-containing DNA from the human myotonic dystrophy and fragile X loci. *J Biol Chem*. Oct 9 1998;273(41):26998-7008. doi:10.1074/jbc.273.41.26998

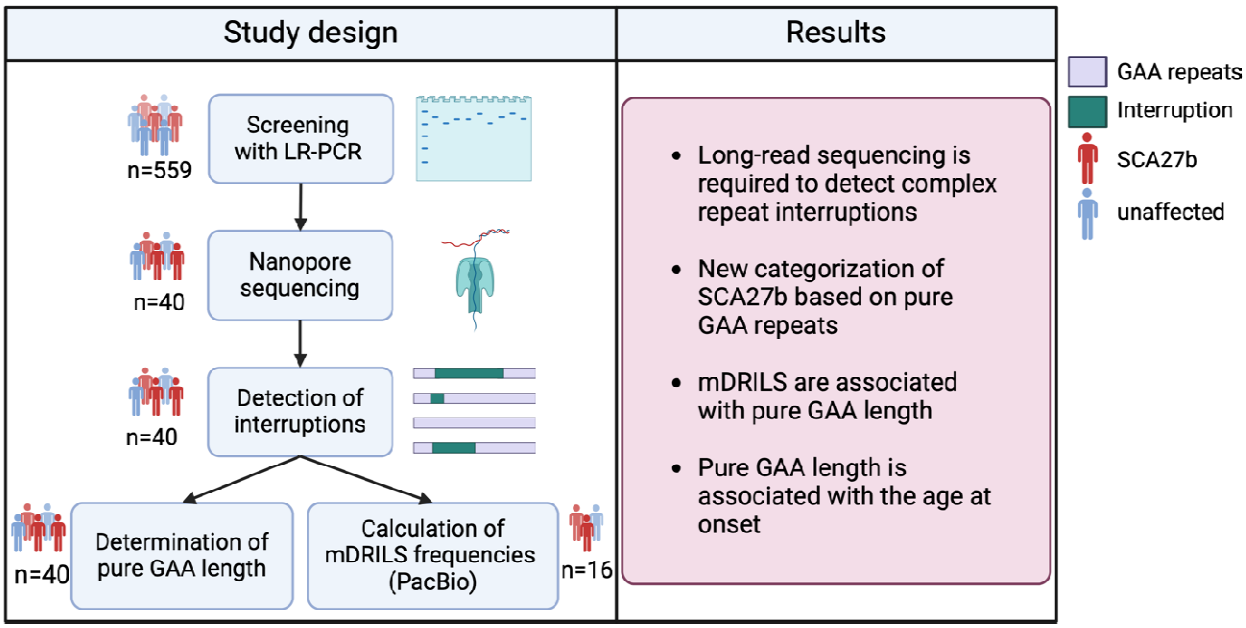


**Figure 1. Overview of the study**. We performed initial screening of the FGF14 repeat expansions with long-range PCR, then a subset with long-read sequencing to detect interruptions and the pure GAA tract. Legend: LR-PCR = long-range PCR, PacBio = Pacbio sequencing, mDRILS = mosaic divergent repeat interruption affecting length and sequence. The figure was created with bioRender.

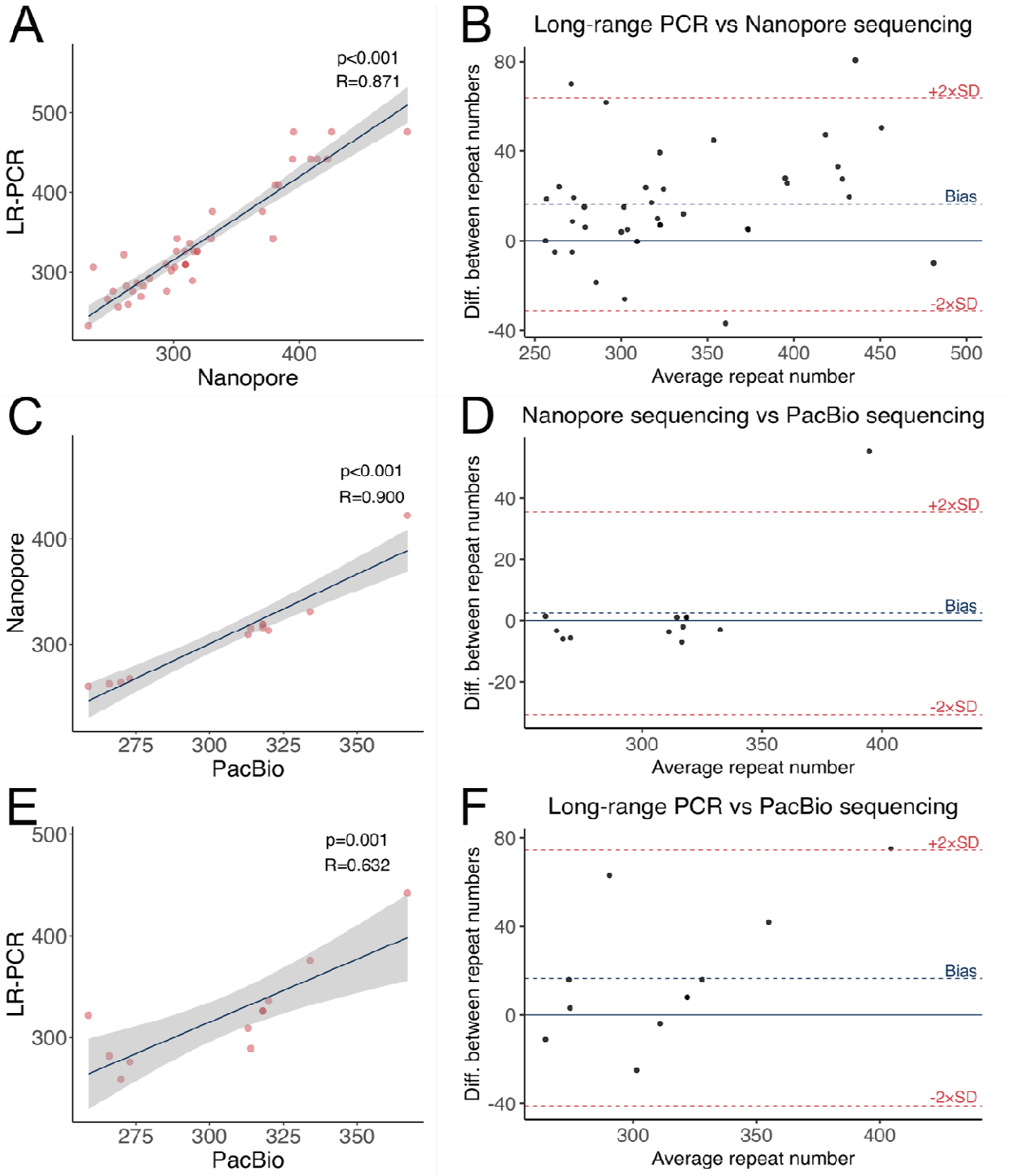


**Figure 2. Comparison of the repeat length between different methods. (A)** Correlation between LR-PCR and Nanopore. **(B)** Bland Altman Plots between LR-PCR and Nanopore **(C)** Correlation between Nanopore and PacBio. **(D)** Bland Altman Plots between Nanopore and PacBio. **(E)** Correlation between LR-PCR and PacBio. **(F)** Bland Altman Plots between LR-PCR and PacBio. A linear regression model was used for statistical analysis. Legend: LR-PCR = long-range PCR, Nanopore = Oxford Nanopore Technology sequencing, PacBio = PacBio sequencing.

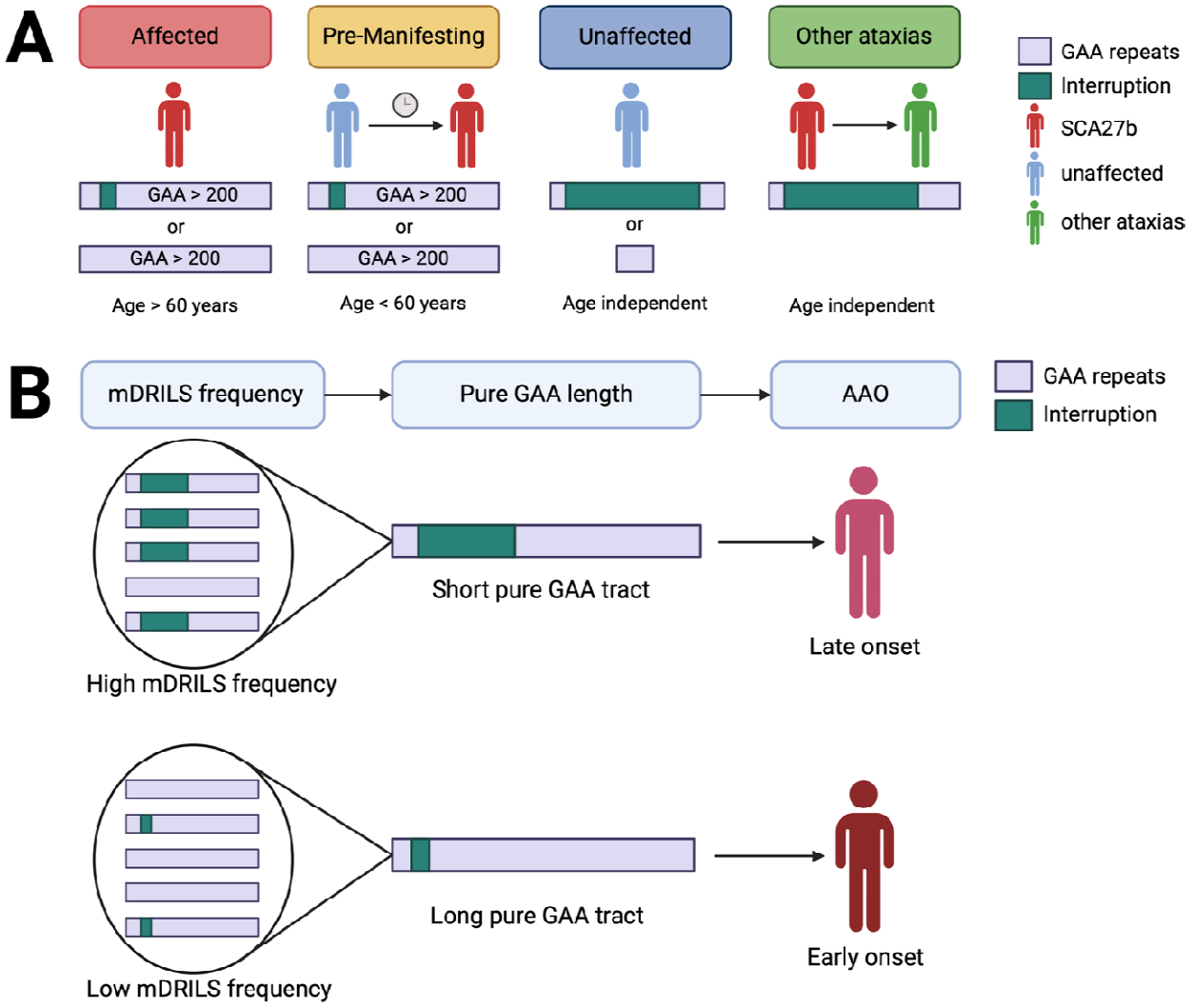


**Figure 3. Schematic illustration of: (A) New categorization of the individuals based on the pure GAA length.** Affected individuals with SCA27b are represented in red, who have either >200 pure GAA repeats, or an interruption with >200 pure GAA repeats. Pre-manifesting carriers with an earlier age (<60 years of age) are in yellow. Unaffecteds have long repeat interruptions and short pure GAA (<200 repeats) tracts. Lastly, misdiagnoses of SCA27b ataxia can occur with a long repeat interruption. **(B) Relationship between mDRILS frequency, pure GAA length, and age at onset.** mDRILS are modifiers of the repeat stability, and are more frequent in late-onset SCA27b compared to early-onset SCA27b. Legend: mDRILS = mosaic divergent repeat interruptions affection length and sequence, AAO = age at onset. These figures were created with bioRender.

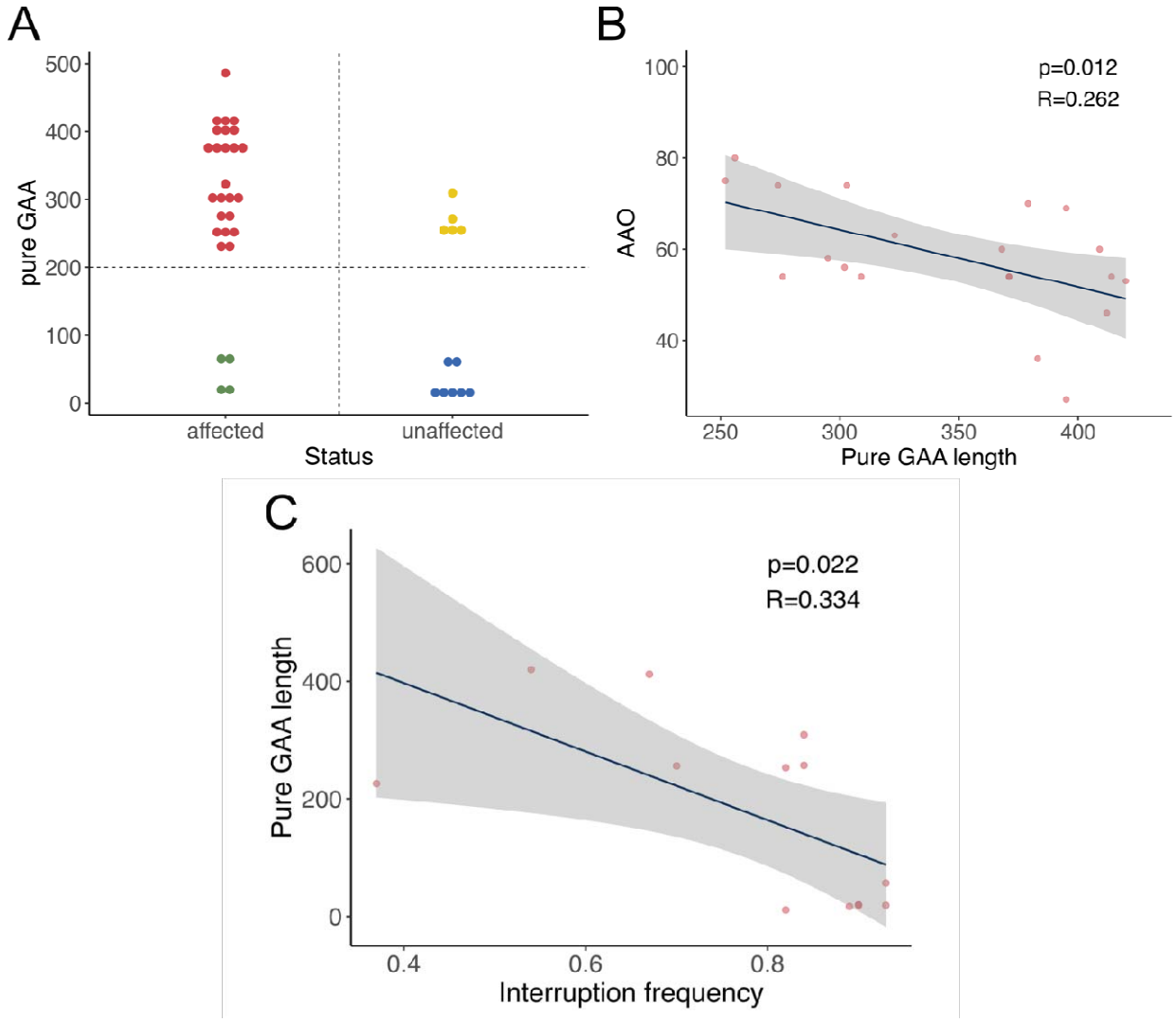


**Figure 4. Analysis of the pure GAA length. (A)** Differences between affected and unaffected repeat expansion carriers in pure GAA length. **(B)** Correlation between age at onset (AAO) and pure GAA length. **(C)** Correlation between interruption frequency and pure GAA length. A linear regression model was used for statistical analysis. The adjusted significance level is $\alpha = 0.013$.

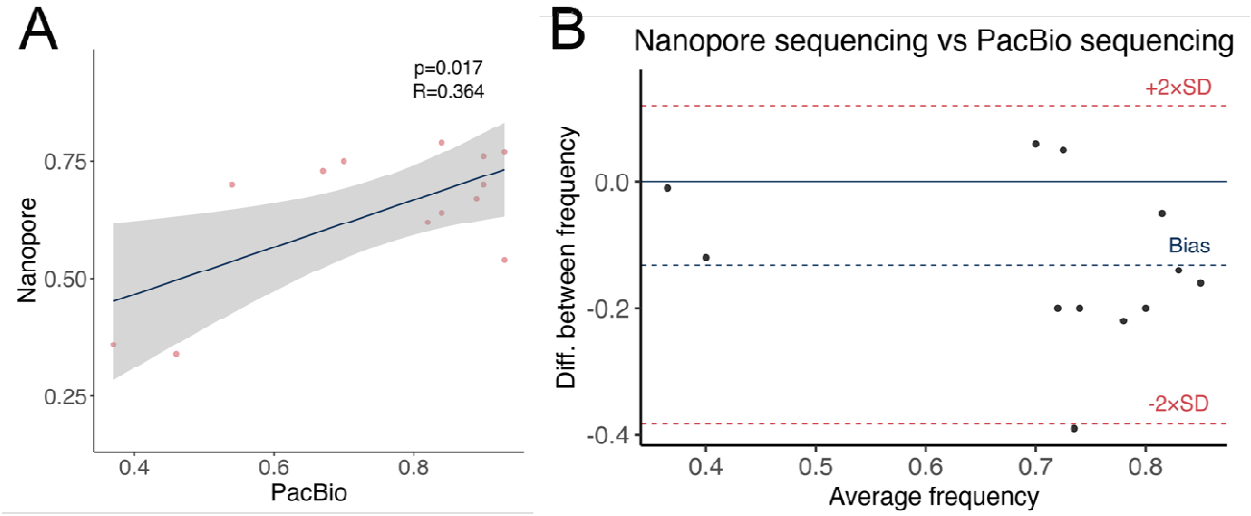


**Figure 5. Comparison of the interruption frequency of all motifs between Nanopore and PacBio sequencing. (A)** Correlation between Nanopore and PacBio sequencing **(B)** Bland-Altman plot between Nanopore and PacBio sequencing. A linear regression model was used for statistical analysis.

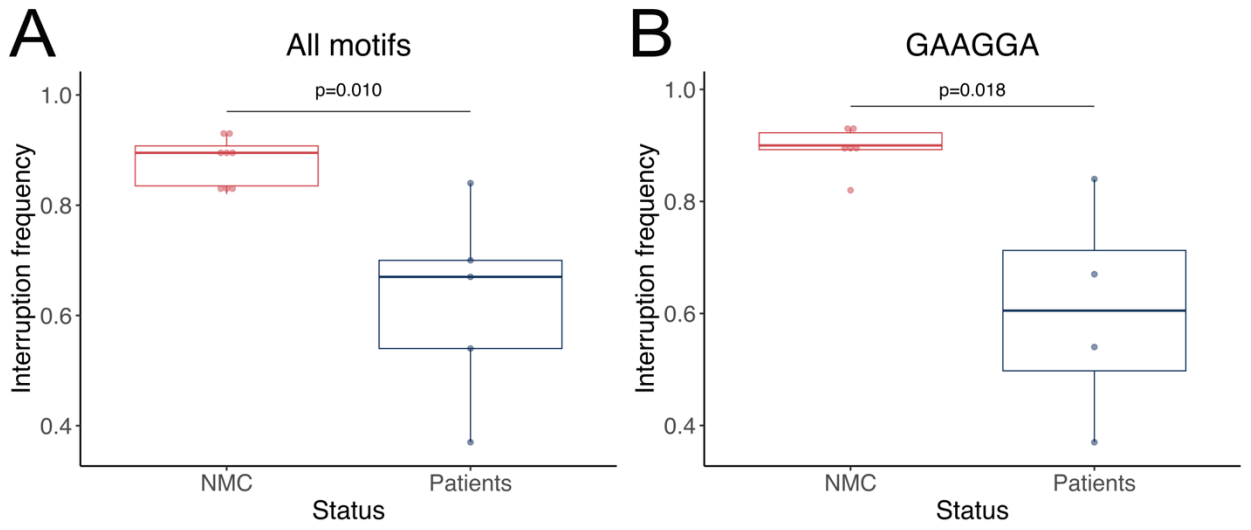


**Figure 6. Analysis of the interruption frequency.** Differences between non-manifesting repeat expansion carriers (NMC) and patients in interruption frequency of all motifs **(A)** and the $(GAA\underline{GGA})_n$ motif **(B)**. Mann-Whitney-U-tests were performed for statistical analysis. The adjusted significance level is $\alpha = 0.013$.

**Table 1: Overview of the n=40 individuals with a repeat number > 250.**

| | Affected | Unaffected |
|---|---|---|
| AAO (Median (IQR)) | 56 years(50;68) | - |
| AAS (Median (IQR)) | 70 years (57;75) | 53 years (49;60) |
| RN (Nanopore) (Median (IQR)) | 322 (297;386) | 301 (264;315) |
| RN (PacBio) (Median (IQR)) | 320 (313;334) | 314 (268;318) |
| RN (LR-PCR) (Median (IQR)) | 342 (305;417) | 309 (285;326) |
| Frequency (Median (IQR)) | 0.61 (0.50;0.72) | 0.90 (0.89;0.92) |
| Interruption length (Median (IQR)) | 24 (3;44) | 147 (77;148) |

Legend: AAO = age at onset, AAS = age at sampling, RN = repeat number, IQR = interquartile range.

**Table 2: Overview of the individuals with an expanded *FGF14* repeat expansion.**

| ID | Status | Previously Published (PMID) | RN (ONT) | RN (LR-PCR) | RN (PacBio) | Interruption frequency (PacBio) | Pure GAA | Interruption in the long allele (ONT) | Confirmed |
|---|---|---|---|---|---|---|---|---|---|
| L-3013 | NMC | | 316 | 326 | 318 | 0.90 | 20 | GAA*GGA* | PacBio, Sanger |
| L-3329 | NMC | | 271 | 286 | 276 | | 271 | - | Sanger |
| L-3344 | NMC | | 319 | 326 | 318 | 0.90 | 19 | GAA*GGA* | PacBio, Sanger |
| L-3479 | NMC | | 260 | 322 | 259 | 0.93 | 57 | GAA*GGA* | PacBio, Sanger |
| L-3501 | NMC | | 319 | 326 | 318 | 0.82 | 19 | GAA*GGA* | PacBio, Sanger |
| L-3656 | NMC | | 294 | 309 | - | | 11 | GAA*GGA* | - |
| L-3657 | NMC | | 309 | 326 | - | | 309 | - | Sanger |
| L-8761 | NMC | | 264 | 259 | 270 | 0.84 | 257 | *GAAAA*GAAGAA*GGAA* | PacBio, Sanger |
| L-8782 | NMC | | 256 | 256 | - | | 256 | - | - |
| L-8934 | NMC | | 315 | 289 | 314 | 0.89 | 17 | GAA*GGA* | PacBio, Sanger, Cas9 |
| L-10408 | NMC | 36493768 | 263 | 282 | 266 | 0.82 | 253 | *GAAAA*GAAGAA*GGAA*GAA*GGAA* | PacBio, Sanger |
| L-14575 | Patient | 36493768, 37861706 | 426 | 476 | 366 | 0.54 | 420 | GAA*GGA* | PacBio, Sanger |
| L-14630 | Patient | 36493768 | 313 | 336 | 320 | 0.84 | 309 | GAA*GGA* | PacBio, Sanger, Cas9 |
| L-14846 | Patient | | 371 | 376 | - | | 371 | - | Sanger |
| L-14853 | Patient | 38487929 | 302 | 326 | - | | 20 | - | Sanger |
| L-14894 | Patient | | 298 | 302 | - | | 302 | GAA*GGA* | Sanger |
| L-14904 | Patient | 38487929 | 247 | 266 | - | | 247 | - | Sanger |
| L-14911 | Patient | 38487929 | 395 | 476 | - | | 395 | - | Sanger |
| L-14995 | Patient | 37246629 | 309 | 309 | 313 | 0.46 | 66 | *GCA*GAAGAAGAAGAA | PacBio, Sanger |
| L-14996 | NMC | 37246629 | 309 | 309 | - | | 66 | *GCA*GAAGAAGAAGAA | - |
| L-14997 | Patient | 37246629 | 309 | 309 | - | | 66 | *GCA*GAAGAAGAAGAA | Sanger |
| L-15166 | Patient | 36493768 | 331 | 376 | 334 | 0.37 | 226 | GAA*GGA* | PacBio |
| L-15713 | Patient | 36493768, 37861706 | 303 | 342 | - | | 303 | - | Sanger |
| L-15739 | Patient | 36493768, 37861706 | 252 | 276 | - | | 252 | - | Sanger |
| L-15754 | Patient | 36493768 | 274 | 269 | - | | 274 | - | Sanger |
| L-15764 | Patient | 36493768, 37861706 | 330 | 342 | 342 | | 323 | GAA*GAG* | Not confirmed |
| L-15891 | Patient | 36493768 | 301 | 306 | - | | 20 | GAA*GGA* | Sanger |
| L-17665 | Patient | 36493768 | 422 | 442 | 367 | 0.67 | 412 | GAA*GGA* | PacBio, Sanger |
| L-17672 | Patient | 36493768, 37861706 | 414 | 442 | - | | 414 | - | Sanger |
| L-18362 | Patient | 36493768 | 486 | 476 | - | | 486 | - | Sanger |
| L-18384 | Patient | 36493768, 37861706 | 409 | 442 | - | | 409 | - | Sanger |
| L-20363 | Patient | 36493768, 37861706 | 381 | 409 | 390 | | 368 | GAA*GAAA*GAA | Not confirmed |
| L-20409 | Patient | | 379 | 342 | - | | 379 | - | Sanger |
| L-20618 | Patient | 38487929 | 371 | 376 | - | | 371 | - | Sanger |
| L-20619 | Patient | | 395 | 442 | - | | 395 | - | Sanger |
| L-20635 | Patient | 38487929 | 236 | 306 | - | | 236 | - | - |
| L-20648 | Patient | 38487929 | 383 | 409 | - | | 383 | - | - |
| L-20656 | Patient | 38487929 | 276 | 282 | - | | 276 | - | Sanger |
| L-22657 | Patient | | 295 | 276 | - | | 295 | - | Sanger |
| L-22867 | Patient | | 267 | 276 | 273 | 0.7 | 256 | *GAG* | PacBio, Sanger |

Legend: RN = repeat number, ONT = Oxford Nanopore Technology sequencing, LR-PCR = long-range PCR, PacBio = PacBio sequencing, NMC = Non-manifesting carrier, Sanger = Sanger sequencing, Cas9 = Cas9 enrichment with Nanopore sequencing. Repeat motif variations are indicated by the underlined and italicized nucleotides.