BRAIN
ORIGINAL ARTICLE

Mosaic divergent repeat interruptions in XDP influence repeat stability and disease onset

Joanne Trinh,[1] Theresa Lüth,[1] Susen Schaake,[1] Björn-Hergen Laabs,[2] Kathleen Schlüter,[1] Joshua Laβ,[1] Jelena Pozojevic,[1] Ronnie Tse,[1] Inke König,[2] Roland Dominic Jamora,[3] Raymond L. Rosales,[4] Norbert Brüggemann,[1,5] Gerard Saranza,[6] Cid Czarina E. Diesta,[7] Frank J. Kaiser,[8,9] Christel Depienne,[8] Christopher E. Pearson,[10,11] Ana Westenberger[1] and Christine Klein[1]

While many genetic causes of movement disorders have been identified, modifiers of disease expression are largely unknown. X-linked dystonia-parkinsonism (XDP) is a neurodegenerative disease caused by a SINE-VNTR-Alu(AGAGGG)$_n$ retrotransposon insertion in *TAF1*, with a polymorphic (AGAGGG)$_n$ repeat. Repeat length and variants in *MSH3* and *PMS2* explain ~65% of the variance in age at onset (AAO) in XDP. However, additional genetic modifiers are conceivably at play in XDP, such as repeat interruptions.

Long-read nanopore sequencing of PCR amplicons from XDP patients ($n = 202$) was performed to assess potential repeat interruption and instability. Repeat-primed PCR and Cas9-mediated targeted enrichment confirmed the presence of identified divergent repeat motifs.

In addition to the canonical pure SINE-VNTR-Alu-5′-(AGAGGG)$_n$, we observed a mosaic of divergent repeat motifs that polarized at the beginning of the tract, where the divergent repeat interruptions varied in motif length by having one, two, or three nucleotides fewer than the hexameric motif, distinct from interruptions in other disease-associated repeats, which match the lengths of the canonical motifs. All divergent configurations occurred mosaically and in two investigated brain regions (basal ganglia, cerebellum) and in blood-derived DNA from the same patient. The most common divergent interruption was AGG [5′-SINE-VNTR-Alu(AGAGGG)$_2$*AGG*(AGAGGG)$_n$], similar to the pure tract, followed by AGGG [5′-SINE-VNTR-Alu(AGAGGG)$_2$*AGGG*(AGAGGG)$_n$], at median frequencies of 0.425 (IQR: 0.42–0.43) and 0.128 (IQR: 0.12–0.13), respectively. The mosaic AGG motif was not associated with repeat number (estimate = −3.8342, $P = 0.869$). The mosaic pure tract frequency was associated with repeat number (estimate = 45.32, $P = 0.0441$) but not AAO (estimate = −41.486, $P = 0.378$). Importantly, the mosaic frequency of the AGGG negatively correlated with repeat number after adjusting for age at sampling (estimate = −161.09, $P = 3.44 \times 10^{-5}$). When including the XDP-relevant *MSH3*/*PMS2* modifier single nucleotide polymorphisms into the model, the mosaic AGGG frequency was associated with AAO (estimate = 155.1063, $P = 0.047$); however, the association dissipated after including the repeat number (estimate = −92.46430, $P = 0.079$).

We reveal novel mosaic divergent repeat interruptions affecting both motif length and sequence (DRILS) of the canonical motif polarized within the SINE-VNTR-Alu(AGAGGG)$_n$ repeat. Our study illustrates: (i) the importance of somatic mosaic genotypes; (ii) the biological plausibility of multiple modifiers (both germline and somatic) that can have additive effects on repeat instability; and (iii) that these variations may remain undetected without assessment of single molecules.

1 Institute of Neurogenetics, University of Lübeck and University Hospital Schleswig-Holstein, Lübeck, Germany
2 Institute of Medical Biometry and Statistics, University of Lübeck, Lübeck, Germany
3 Department of Neurosciences, College of Medicine—Philippine General Hospital, University of the Philippines Manila, Manila, Philippines

Received January 11, 2022. Revised March 14, 2022. Accepted April 14, 2022. Advance access publication April 27, 2022

4 Department of Neurology and Psychiatry, University of Santo Tomas and the CNS-Metropolitan Medical Center, Manila, Philippines
5 Department of Neurology, University of Lübeck, Lübeck, Germany
6 Section of Neurology, Department of Internal Medicine, Chong Hua Hospital, Cebu, Philippines
7 Department of Neurosciences, Movement Disorders Clinic, Makati Medical Center, Makati City, Philippines
8 Institute for Human Genetics at the University Hospital Essen, Essen, Germany
9 Center for Rare Diseases (Essenser Zentrum für Seltene Erkrankungen—EZSE) at the University Hospital Essen, Essen, Germany
10 Program of Genetics and Genome Biology, The Hospital for Sick Children, The Peter Gilgan Centre for Research and Learning, Toronto, Canada
11 University of Toronto, Program of Molecular Genetics, Toronto, Canada

Correspondence to: Joanne Trinh, PhD
University of Lübeck, Ratzeburger Allee 160
23538 Lübeck, Germany
E-mail: joanne.trinh@neuro.uni-luebeck.de

Keywords: repeat interruptions; mosaic; genetic modifiers; single-molecule sequencing; X-linked dystonia parkinsonism

Introduction

While multiple genetic causes of movement disorders have been identified in the past decade, disease modifiers are still largely unknown for most conditions.[1] Individual patients carrying the same pathogenic variant may have variable expressivity of the disease, including variable age at onset (AAO), severity, and clinical manifestations. Thus, in addition to the pathogenic variant, there are genetic modifiers influencing expressivity and onset. One emerging example of this broader concept is X-linked dystonia-parkinsonism (XDP), a neurodegenerative movement disorder endemic to the Philippines and first described in 1976.[2,3]

XDP is one of a large, ever-growing class of >60 diseases caused by unstable tandem repeats and in particular a sub-class of inserted repeats.[4] XDP is characterized by adulthood-onset dystonic movements and parkinsonism due to striatal volume loss as a result of an insertion of the retrotransposon SINE-VNTR-Alu (SVA) in intron 32 of the *TAF1* (TATA-binding protein-associated factor 1) gene.[3,5] There is a hexanucleotide repeat domain within the SVA, which consists of the repeat sequence $(AGAGGG)_n$.[6] This hexanucleotide repeat domain varies in numbers ranging from 30 to 55 and is a strong genetic modifier of AAO.

To date, there are four putative modifiers of XDP expressivity associated with AAO and/or disease severity.[6–8] These modifiers are the length of the hexanucleotide repeat polymorphism and modifiers of AAO related to variants in the DNA repair genes *MSH3* and *PMS2*.[7,8] Both types of modifiers are characteristic for repeat expansion disorders in which expanded repeats of various lengths may be transcribed.

Genetic modifiers, such as DNA repair genes, are present in the germline and are inherited genetic modifiers. Mosaic modifiers, which exist in every patient, necessarily originate as post-zygotic mutations, have not been studied extensively. We previously described the presence of somatic repeat length mosaicism in XDP with a higher number of repeats detected in the cerebellum and basal ganglia compared to blood of the same patient.[9] The mosaic lengths between tissues arose as post-zygotic mutations. In this study, we identified novel mosaic repeat motif patterns that deviate from the known hexanucleotide repeat motif both in motif length and sequence, and investigate whether they act as new genetic modifiers of repeat instability and AAO of this neurodegenerative disorder.

Materials and methods

Patient demographics

The study was approved by the Ethics Committees of the University of Lübeck, Germany and at the Metropolitan Medical Center, Manila, Philippines. For the analysis of genomic variants within the SVA and the detection of variations of the hexanucleotide repeat domain, $n = 202$ patients with XDP were investigated and included different brain regions (basal ganglia, cerebellum) and blood-derived DNA from one patient with XDP. As XDP follows an X-linked recessive inheritance pattern, all patients were male. The mean AAO was 41.93 (SD = ±8.56) years, and the mean age at examination (AAE) was 47.5 (SD = ±9.84) years (Supplementary Table 1).

DNA extraction

All DNA was extracted from the Blood and Cell Culture DNA Midi kit (Qiagen). Summary of experimental procedures are described in Supplementary Table 1.

Nanopore sequencing of PCR amplicon

Long-range PCR was performed for the SVA (3.2 kb) as previously described.[9] The master mix and the amplification conditions are presented in Supplementary Tables 2 and 3, respectively (XDP-16153 F: 5′-GTTCCATTGTGTGGTTGTACCAGCGTTTGTTC-3′, XDP-19345R: 5′-CACATGAAAAGATGCCC AACATCATTAGCCATTAG-3′).[10] The libraries were prepared with the Ligation Sequencing Kit (LSK109) and the samples were barcoded with the Native 96 Barcoding Kit (EXP-NBD196), using 400 ng of each patient-derived PCR product. Subsequently, the library was sequenced on a GridION (R9.4.1 flow cell), and in total three flow cells, with one library per flow cell, were used for the sequencing of the multiplexed PCR amplicons.

Nanopore sequencing of Cas9-mediated targeted enrichment

In addition, the *TAF1* SVA was enriched by an amplification-free Cas9-mediated approach, as previously described.[9] For Cas9 enrichment, crRNAs were designed with ChopChop (https://

chopchop.cbu.uib.no). Two crRNAs were used upstream of the *TAF1* SVA insertion (crRNA 1 and 2) and two crRNAs were used downstream (crRNA 3 and 4) for a 5.5 kb product specifically around the SVA (Supplementary Table 4). Blood-, basal ganglia- and cerebellum-derived DNA from one patient was used for the Cas9-enrichment. The libraries were generated with the Ligation Sequencing Kit (LSK109) and no barcoding was performed. The sequencing was performed on the MinION (R9.4.1 flow cell). For the blood-derived DNA, four flow cells were loaded with five libraries (4 × 5 µg and 1 × 1 µg of input DNA). For the basal ganglia-derived sample, four flow cells were loaded with five libraries (3 × 3 µg, 1 × 2 µg and 1 × 1 µg of input DNA). Lastly, for the cerebellum-derived sample we used three flow cells and five libraries (4 × 5 µg and 1 × 10 µg of input DNA).

Flow cell loading

All libraries were prepared with the Ligation Sequencing Kit (SQK-LSK109), loaded on a R9.4.1 flow cell and sequenced on MinION/GridION.

Detection of repeats

Base-calling was performed with the most updated Guppy version 5.0.11. For the detection of the repeat length, the super accurate model (dna_r9.4.1_450bps_sup.cfg) was used. All reads were mapped to the reference sequence with the software Minimap2 (v2.17) (Supplementary Fig. 1). Samtools (v1.9) was used for coverage determination and filtering (>1500×). We filtered for Phred score Q > 12. Motif mismatch detection was achieved with 'Noise-canceling repeat finder' (NCRF) (v1.01.02).[11] The detailed commands are listed in the Supplementary material, and more information as well as the corresponding reference files used for the alignment are provided at: https://github.com/nanopol/xdp_sva/.

Repeat-primed PCR

Repeat-primed PCR (RP PCR) with a FAM-tagged primer was performed for validation. The master mix and the amplification conditions are presented in Supplementary Tables 5 and 6. Supplementary Fig. 2 shows a schematic of the primer binding locations in the *TAF1* intron 32 and the hexanucleotide repeat domain of the *TAF1* SVA. For the fragment analysis a total of 1 µl of the RP PCR products with 10.7 µl HiDi Formamide (Applied Biosystems) and 0.3 µl GeneScan™ 600 LIZ™ Dye Size Standard (Applied Biosystems) were used for capillary electrophoresis on an ABI 3500×L Genetic Analyzer (Applied Biosystems). The output was analysed using GeneMapper software (version 4.1, Applied Biosystems).

Statistical analyses

Frequencies of deletions were estimated using NCRF, and Spearman's correlation coefficient was employed to estimate the correlation and the corresponding *P*-values reported. Box plots were used to show the distribution. We aimed to adjust the mosaic frequencies for the age at sampling and designed a linear regression model predicting the mosaic AGGG frequency by age at sampling. We observed that the age has an impact on the mosaic AGGG frequency (Supplementary Table 7). Thus, we obtained the residuals of the regression model and used this as an adjusted predictor for repeat number or AAO. Regression models were used to assess the correlation between AAO and the frequency of the most common DRIL, AGG [5′-SINE-VNTR-Alu$(AGAGGG)_2$AGG$(AGAGGG)_n$], adjusted for age, three single nucleotide polymorphisms (SNPs) in *MSH3* and *PMS2*, or for three SNPs in *MSH3* and *PMS2* and the SVA repeat number.

Data availability

The data presented in this study are available on SRA (SAMN24775867-SAMN24775962, SAMN24115523-SAMN24115530). The bioinformatic commands to quantify the *TAF1* SVA $(AGAGGG)_n$ repeat length are described at: https://github.com/nanopol/xdp_sva/.

Results

Deep sequencing of the PCR amplicon of the 3.2 kb *TAF1* SVA region in blood-derived DNA with nanopore yielded a mean coverage of 10 915X per sample (SD = ±8207) (Supplementary Fig. 1 and Table 1). We detected a prominent occurrence of deletions in every patient ($n = 202$), with a mean frequency of 0.97 (SD = ±0.113) at the beginning of the repeat tract, consistently present on the plus- and minus-strands in all patients (Fig. 1A and B). At the single nucleotide resolution, three deletions (deletions 1, 2, 3) were found at the 5′ end at positions 11, 14, and 17 of the repeat motif (Fig. 1C). These detected deletions lead to divergent repeat motifs that occur at the second and/or third motif of the expanded $(AGAGGG)_n$ tandem repeat (Fig. 2A).

The divergent motifs were detected as multiple combinations and at various frequencies in every patient, indicating somatic mosaicism (Fig. 2A). The most frequently detected in all of the analysed patients in this study was the divergent repeat AGG, with the pattern $(AGAGGG)_2$AGG$(AGAGGG)_n$, having a median frequency 0.425 (IQR: 0.42–0.43). This occurred with near equal frequency to the pure uninterrupted $(AGAGGGG)_n$ tract. The second most detected divergent motif was AGGG, where the resulting repeat motif pattern was $(AGAGGG)_2$AGGG$(AGAGGG)_n$, with a median frequency of 0.128 (IQR: 0.12–0.13) (Fig. 2B). Other divergent repeat motifs and patterns were detected at lower frequencies (Fig. 2B). It is noteworthy that each of the divergent repeat motifs shifted the hexameric repeat frame of the repeat tract. Most change the trinucleotide repeat frame. However, the $(AGAGGG)_2$AGG$(AGAGGG)_n$ divergent repeat motif retained the trinucleotide frame. Curiously, the AGG motif was not associated with AAO. The significance, if any, that this was the most common form, nearly equal to the pure tract, and its retention of the trinucleotide frame, is unknown.

Further genetic validation was performed using repeat-primed PCR and Cas9-targeted enrichment of the divergent repeat motifs. Using repeat-primed PCR targeting the divergent repeat motif patterns, we observed a signal that indicated the presence of divergent motifs at the beginning 5′ region of the AGAGGG repeat tract (Supplementary Fig. 2A–D).

Lastly, Cas9-targeted enrichment was performed to avoid errors from PCR amplification as another validation. We found comparable frequencies of somatic divergent repeat motifs in the blood, basal ganglia, and cerebellum-derived DNA from the same patient (Fig. 3A–C). To discern both repeat size and mosaic status on the same DNA fragments via long-read sequencing, we investigated the mosaicism by directly measuring how much of this somatic instability is derived from the canonical hexamer repeat tract or AGGG motif repeat tract. We observed that the there is more variability in repeat number [range: 21–53 (blood); 19–68 (basal ganglia); 19–60 (cerebellum)] when looking at the canonical hexamer repeat

Table 1 Residuals and coefficients of linear model predicting repeat number with mosaic AGGG frequency adjusted for age at sampling

Residuals				
Min	**1Q**	**Median**	**3Q**	**Max**
−10.2224	−2.6164	−0.3089	2.2163	13.0807
Coefficients				
	Estimate	**SE**	**t-value**	**Pr(>|t|)**
Intercept	41.4221	0.2804	147.70	$<2 \times 10^{-16}$
Beta_{res}	−161.0907	37.9957	−4.24	3.44×10^{-5}

Model: RN = Intercept + beta_{res} × res. 1Q = first quartile; 3Q = third quartile; Max = maximum; Min = minimum; res = residuals; RN = repeat number; SE = standard error.

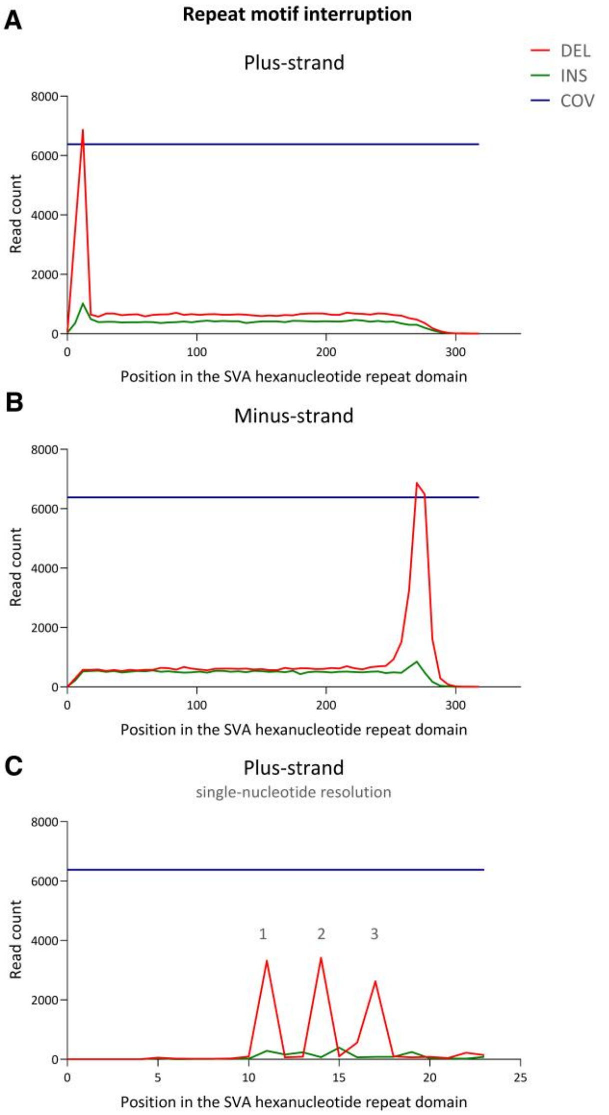

Figure 1 Detection of divergent repeat interruptions within the *TAF1* SVA hexanucleotide repeat domain. (**A**) Plus-strand reads show deletions or insertions within the SVA hexanucleotide repeat domain. (**B**) Minus-strand reads show deletions or insertion within the SVA hexanucleotide repeat domain. (**C**) Single-nucleotide resolution of the deletions detected at the 5′ end of the SVA hexanucleotide repeat domain. COV = coverage (number of reads covering the *TAF1* SVA hexanucleotide repeat domain); DEL = deletion; INS = insertion.

tract (wt-wt-wt) compared to the AGGG motif (1-2-wt) [range: 41–52 (blood); 47–84 (basal ganglia); 24–53 (cerebellum)]. Furthermore, there was a higher quartile coefficient of dispersion for the canonical repeat tract (range: 0.042–0.054) compared to the AGGG motif repeat tract (range: 0.019–0.031) (Fig. 3D and E).

We focused our further analyses on the $(\text{AGAGGG})_2$ *AGGG*(AGAGGG)n and the $(\text{AGAGGG})_2$*AGG*(AGAGGG)n repeat combinations as they were the most frequently detected. We investigated these divergent repeat motifs and their influence on repeat tract length (i.e. repeat number). The frequency of the AGGG negatively correlated with repeat tract length ($r = -0.48$, $P = 9.5 \times 10^{-13}$): the higher the frequency of AGGG, the shorter the repeat tract (Fig. 4A). This same effect was not observed for the AGG (Fig. 4B). The AGGG positively correlated with AAO ($r = 0.34$, $P = 9.5 \times 10^{-7}$), whereas the AGG did not show a correlation with AAO (Fig. 4C and D). Since somatic mosaicism may change with age, which may confound analyses, we adjusted for age at sampling. After adjusting for age at sampling, using the residuals of the regression model as an adjusted predictor for repeat number, we found that the mosaic AGGG frequency in blood DNAs was associated with repeat number (estimate = −161.09, $P = 3.44 \times 10^{-5}$) (Table 1). After adjusting for age at sampling, using the residuals of the regression model as an adjusted predictor for AAO, we found that the mosaic AGGG frequency was not associated with AAO (estimate = 138.9471, $P = 0.09$) (Table 2). When including genetics of the XDP-relevant *MSH3/PMS2* SNPs into the model, the mosaic AGGG frequency was associated with AAO (estimate = 155.1063, $P = 0.047$), however, the association dissipated after including the repeat number (estimate = −92.46430, $P = 0.079$) (Table 2).

We assessed the AGG repeat motif using residuals from the age at sampling in a linear model and did not observe an association with repeat number (estimate = −3.8342, $P = 0.869$) or AAO (estimate = −1.8236, $P = 0.97$). The canonical hexamer repeat motif frequency was associated with repeat number (estimate = 45.32, $P = 0.0441$) but not AAO (estimate = −41.486, $P = 0.378$) (Supplementary Tables 8 and 9).

Discussion

A thorough analysis of the expanded *TAF1* SVA repeat tract in 202 XDP patients revealed novel mosaic divergent repeat interruptions affecting both motif length and sequence (DRILS) of the canonical motif polarized within the expanded repeat tract.

Repeat interruptions exist in other repeat expansion disorders.[12–15] For example, interruptions with the same motif length,

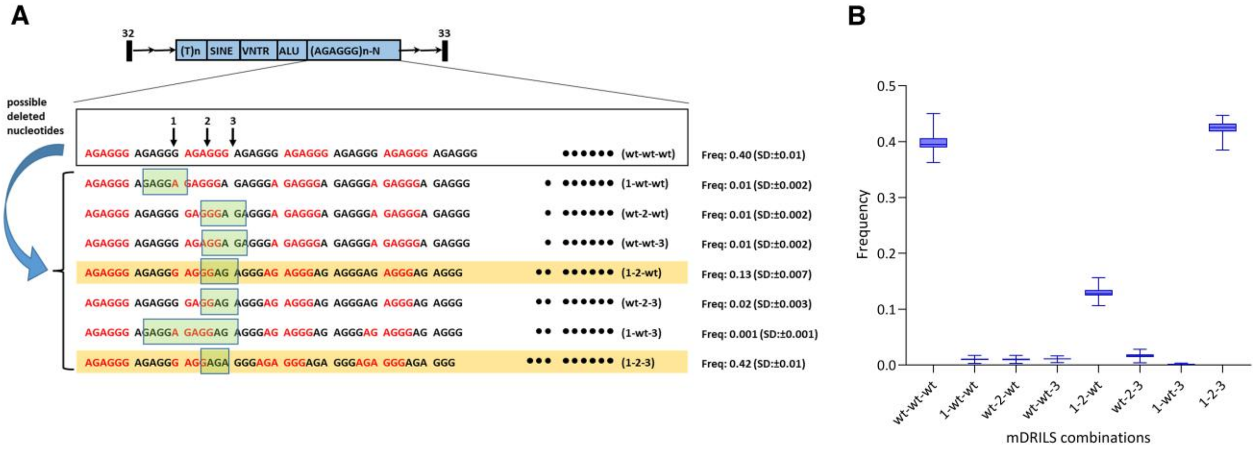

Figure 2 Overview of mDRILS combinations within the *TAF1* SVA hexanucleotide repeat motif and the corresponding detected frequencies. (**A**) Deleted nucleotides change the size and sequence of a single repeat unit of the hexameric tract. Boxed outline: change in repeat unit. Shaded highlight: common haplotypes. Black dots: indicating shifts in the repeat tract frame. The mosaic frequency of each DRIL is displayed alongside the interrupted repeat units. (**B**) The box plot shows frequencies of the combinations of deletions. The line and box represent median and interquartile range, respectively, and the whiskers represent the range.

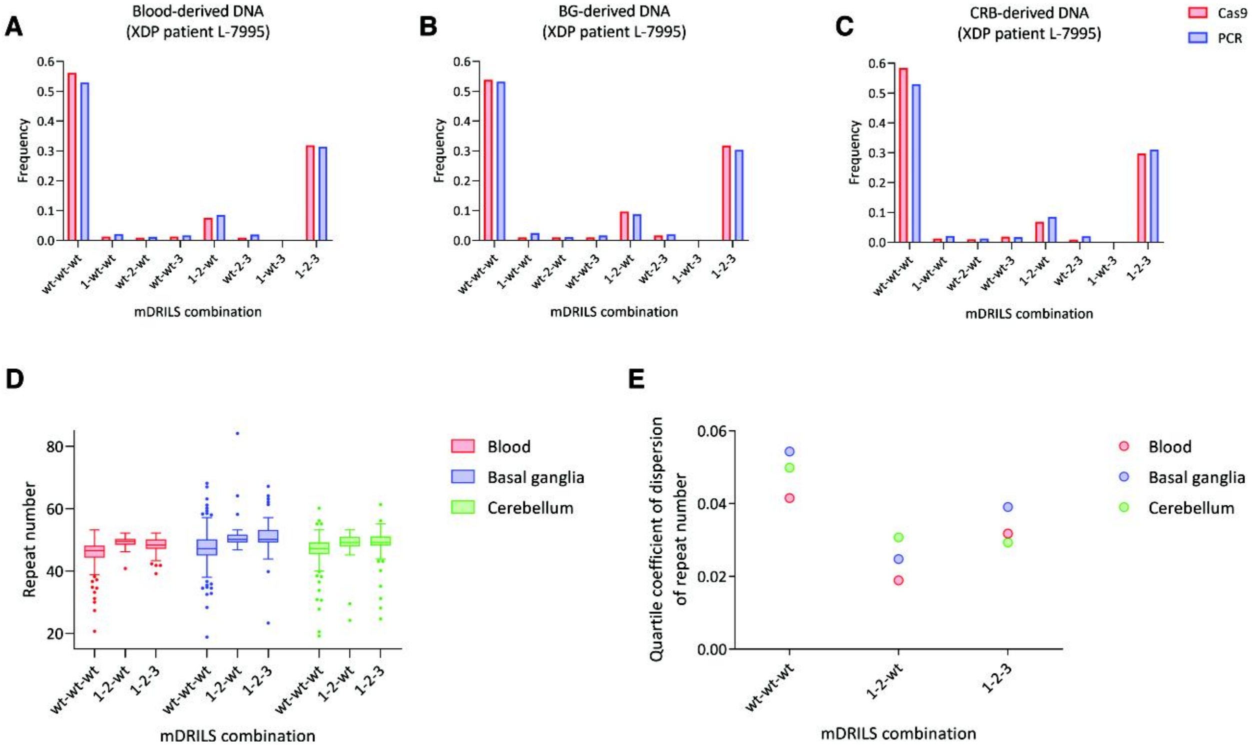

Figure 3 Comparison of mDRILS frequencies detected in blood- and brain-derived DNA, enriched for the SVA insertion by PCR or amplification-free Cas9-enrichment. The bars represent a single frequency value for each mDRILS detected from (**A**) blood-derived DNA, (**B**) basal ganglia (BG)-derived DNA and (**C**) cerebellum (CRB)-derived DNA. (**D**) Repeat number distribution of each mDRILS, and (**E**) quartile coefficient of dispersion of each mDRILS. The bar charts show the detected frequencies of the different combinations of mDRILS. Box plot centre line represents median and box limits are upper and lower quartiles.

but varying repeat sequence from the canonical repeat motif, have been observed in fragile X syndrome (FXS), Huntington's disease, spinal cerebellar ataxia (SCA)1, SCA2, SCA3, SCA8, SCA10, SCA17, SCA31, myotonic dystrophy, and Friedreich's ataxia.[4,16] The presence of these interruptions can modify AAO by years. Interruptions of the $(GAA)_n$ tract in Friedreich's ataxia[17] can delay AAO by 9 years. A pure HTT $(CAG)_n$ tract can hasten disease by 13–29 years, whereas a multiply-interrupted tract can delay disease by 3–6 years.[18] AGG interruptions polarized to the 5′-end of the $(CGG)_n$ of *FMR*1, is associated with FXS, fragile X-associated tremor/ataxia syndrome, and fragile X-associated primary ovarian insufficiency. The CAT interruptions of the (CAG)n of *ATXN1* in SCA1, the CAA interruptions in SCA2, GAA of Friedreich's ataxia, and the various variant repeats at the expanded $(CTG)_n$ in *DMPK*, associated

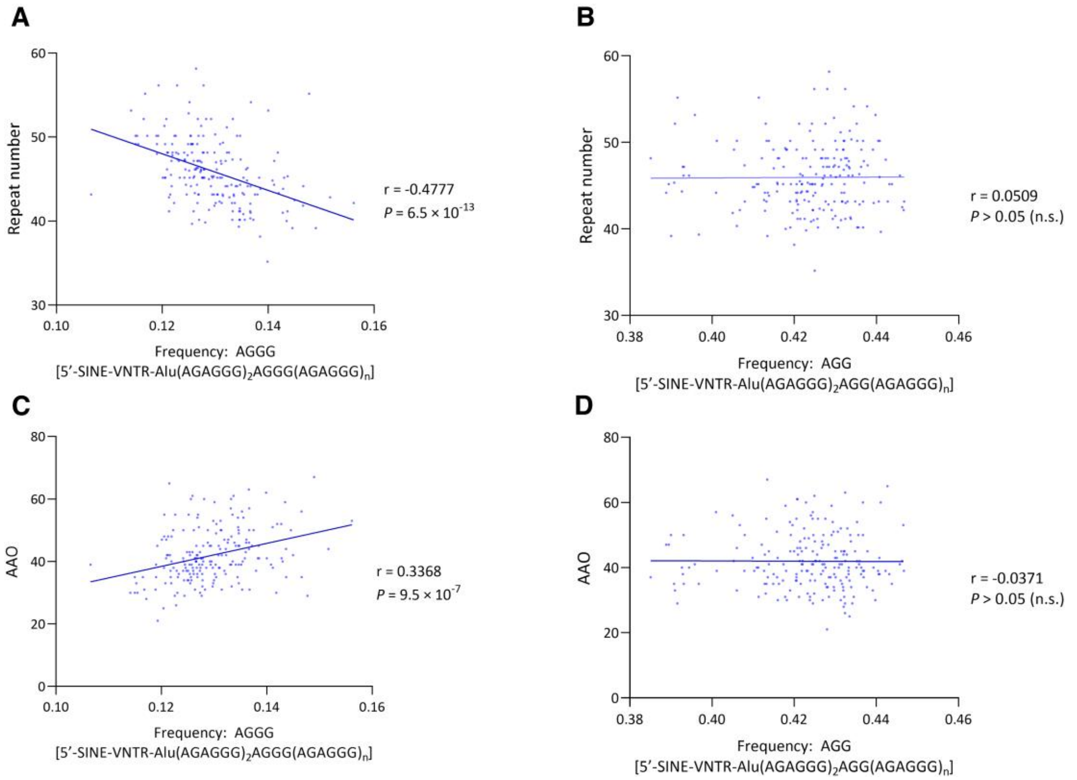

Figure 4 Relationship between mDRILs, AAO, and repeat number in patients with XDP. (**A**) The correlation between hexanucleotide repeat number and the mosaic frequency for divergent motif AGGG [5′-SINE-VNTR-Alu(AGAGGG)$_2$AGGG(AGAGGG)$_n$]. (**B**) The correlation between hexanucleotide repeat number and mosaic frequency for divergent motif AGG [5′-SINE-VNTR-Alu(AGAGGG)$_2$AGG(AGAGGG)$_n$]. (**C**) The correlation between AAO and mosaic frequency for divergent motif AGGG [5′-SINE-VNTR-Alu(AGAGGG)$_2$AGGG(AGAGGG)$_n$]. (**D**) The correlation between age at onset (AAO) mosaic frequency for divergent motif AGG [5′-SINE-VNTR-Alu(AGAGGG)$_2$AGG(AGAGGG)$_n$]. r = Spearman's rank correlation coefficient; *P* = Spearman's exploratory *P*-value.

with myotonic dystrophy. In addition to affecting the AAO, interruptions can also alter clinical presentation, as in Friedreich's ataxia, myotonic dystrophy type 1 (DM1), and SCA10.[12,15,19] However, these disruptions in purity were all by interrupting repeat units of the same motif length. In all cases, with the exception of DM1 and SCA8, the change of the repeat motif was of a single nucleotide replacement in the motif (i.e. CAG → CAA in the case of HTT). For DM1 the variant repeats were of a variety of sequence, all the same number of nucleotides as the canonical motif, and were present in 8.4% of DM1 individuals with expansions.[19] Moreover, in each of these cases the interruptions are reported as germline repeat configurations, and the possibility that they may vary somatic as putative mosaic repeat disruptions have not been looked at. XDP divergent repeat motifs were found to be polarized at the beginning (5′ end) of the *TAF1* SVA (AGAGGG)$_n$ repeat tract in a somatic mosaic fashion, indicating a new mechanism. We postulate that these deletions stabilize the repeats, especially as a higher frequency of the AGGG divergent repeat is associated with shorter repeats. It can be inferred that the loss of divergent repeat motif indirectly delays the AAO. However, the association dissipates when including other modifiers and repeat numbers into the model. One possible explanation is that the mosaic deletions affect repeat stability and thus repeat number.

DRILs may arise by deletions or insertions of single nucleotides from the canonical AGAGGG motif. If the mosaic divergent repeats (mDRILs) arose from deletion events, the repeat tract length will be shorter than the canonical AGAGGG repeat tract. If mDRILs arose from insertion events, the repeat tract length will be longer than the canonical hexameric repeat tract. Our analyses on *n* = 202 suggest that this is indeed an insertion event (Supplementary Fig. 3A and B). It is noteworthy, that in the XDP hexameric repeat tract the divergent motifs occur at the extreme 5′-end of the repeat at the first and second repeat unit of the tract. This polarization is similar to polarized variant AGG and CAA repeat motifs of the *FMR1* CGG and the *HTT* CAG repeat tracts, respectively.[18,20] The polarity of the XDP mutations, coupled with their being present somatically, suggests that they may arise through a possible biological mechanism. Recently, it has been shown that FAN1 exo-nuclease pauses during excision of CGG and CAG slip-out DNAs.[21] These pauses are particularly intense at the polar ends of the repeat tract, proximal to the variant interrupting CAA repeat motifs of CAG slip-outs. One might imagine a polarized error leading to sequence alterations of the repeat. Other interruption-specific pathways involve mismatch repair proteins.[22] Future studies will reveal how DRILS may arise. Our findings in XDP patients revealed that there is mosaicism for each of the different divergent repeat configurations within a given patients sample of the repeat tract, supporting either mechanism as a dynamic process.

The DRILS in the XDP-relevant repeat may affect repeat instability by modifying the propensity to form unusual mutagenic DNA structures, as has been observed for the interruptions of *FMR1* (FXS) and *ATXN1* (SCA1).[16] DRILs can also lead to pathogenic spliceoforms, translation (exonization), RAN-translation, or repeat instability.[4] As many other repeat disorders (e.g. *RFC1, SCA31, SCA8*) are currently being investigated with third generation sequencing technologies, the presence of emerging repeat unit variations like

Table 2 Residuals and coefficients of linear model predicting age at onset with mosaic AGGG frequency

Adjusted for age at sampling[a] Residuals				
Min	**1Q**	**Median**	**3Q**	**Max**
−20.185	−5.661	−1.585	5.303	24.657
Coefficients	**Estimate**	**SE**	**t-value**	**Pr(>|t|)**
Intercept	41.9447	0.6058	69.234	$<2\times10^{-16}$
beta$_{res}$	138.9471	82.0822	1.693	0.0921
Adjusted for age at sampling and genetic modifiers[b] **Residuals**				
−16.7619	−5.7232	−0.9387	5.6250	21.9343
Min	**1Q**	**Median**	**3Q**	**Max**
Coefficients	**Estimate**	**SE**	**t-value**	**Pr(>|t|)**
Intercept	44.2336	1.2167	36.354	$<2\times10^{-16}$
beta$_{res}$	155.1063	77.4761	2.002	0.046679
beta$_{rs245013}$	−1.9713	1.0305	−1.913	0.057234
beta$_{rs33003}$	−1.6813	0.9739	−1.726	0.085866
beta$_{rs62456190}$	3.8217	1.0563	3.618	0.000379
Adjusted for age at sampling, genetic modifiers, and repeat number[c] **Residuals**				
Min	**1Q**	**Median**	**3Q**	**Max**
−16.7062	−3.5218	0.1044	3.5887	15.9468
Coefficients	**Estimate**	**SE**	**t-value**	**Pr(>|t|)**
Intercept	108.70240	3.99377	27.218	$<2\times10^{-16}$
beta$_{RN}$	−1.55022	0.09415	−16.465	$<2\times10^{-16}$
beta$_{res}$	−92.46430	52.29976	−1.768	0.07865
beta$_{rs245013}$	−2.05898	0.66629	−3.090	0.00230
beta$_{rs33003}$	−2.06651	0.63010	−3.280	0.00123
beta$_{rs62456190}$	4.56800	0.68445	6.674	2.57×10^{-10}

1Q = first quartile; 3Q = third quartile; Max = maximum; Min = minimum; res = residuals; RN = repeat number; SE = standard error.

[a] Model; AAO = Intercept + beta$_{res}$ × res.

[b] Model; AAO = Intercept + beta$_{res}$ × res + beta$_{rs245013}$ × rs245013 + beta$_{rs33003}$ × rs33003 + beta$_{rs62456190}$ × rs62456190.

[c] Model; AAO = Intercept + beta$_{RN}$ × RN + beta$_{res}$ × res + beta$_{rs245013}$ × rs245013 + beta$_{rs33003}$ × rs33003 + beta$_{rs62456190}$ × rs62456190.

DRILs will become apparent.[4] It is possible that repeat motif variations, like DRILS, also arise in expansions of other repeats and have been missed due to the sequencing method applied.

General limitations of nanopore sequencing are possible artifacts by amplification during the long-range PCR, sequencing, or software.[11] Thus, for further confirmation, we validated the presence of divergent repeat motifs in two ways: RP-PCR and Cas9-targeted enrichment. However, somatic mosaicism can be difficult to detect and in this case it is evident only with Cas9 enrichment and PCR amplicon long-read sequencing. Still, replication in other cohorts and using different technologies is warranted.

Our study illustrates: (i) the importance of underexplored dynamic somatic mosaic genotypes (repeat tract length, motif length, and motif sequence) present in every individual ($n = 202$) in this case; (ii) the biological plausibility of multiple modifiers (both germline and somatic) that can have effects on repeat instability and expressivity; and (iii) that these variations may remain undetected with older technologies that do not assess single molecules. Importantly, this study sheds light on another putative modifier of XDP expressivity associated with AAO and potentially disease severity. Mosaic repeat deletions present as a novel disease mechanism that is also clinically relevant for other repeat expansion disorders and future genetic counselling with implications beyond XDP.

Funding

The study was supported by the Deutsche Forschungsgemeinschaft (DFG FOR2488; to J.T., I.R.K., N.B., A.W., and C.K.), by funding by the Collaborative Center for X-linked Dystonia-Parkinsonism at Massachusetts General Hospital (to N.B., A.W., and C.K.), the Else-Kröner Fresenius Foundation (to J.T. and N.B.), by intramural funds from the University of Lübeck (to J.T. and C.K.), a career development award from Peter Engelhorn Foundation (to J.T.) and by a career development award from the Hermann and Lilly Schilling

Foundation (to C.K.). J.T. and T.L. had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Posey JE, O'Donnell-Luria AH, Chong JX, *et al*. Insights into genetics, human biology and disease gleaned from family based genomic studies. *Genet Med*. 2019;21:798–812.
2. Lee LV, Pascasio FM, Fuentes FD, Viterbo GH. Torsion dystonia in Panay, Philippines. *Adv Neurol*. 1976;14:137–151.
3. Pauly MG, Ruiz Lopez M, Westenberger A, *et al*. Expanding data collection for the MDSGene database: X-linked dystonia-parkinsonism as use case example. *Mov Disord*. 2020;35:1933–1938.
4. 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*. 2022;32:1–27.
5. Domingo A, Westenberger A, Lee LV, *et al*. New insights into the genetics of X-linked dystonia-parkinsonism (XDP, DYT3). *Eur J Hum Genet*. 2015;23:1334–1340.
6. Bragg DC, Mangkalaphiban K, Vaine CA, *et al*. Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. *Proc Natl Acad Sci U S A*. 2017;114:E11020–E11028.
7. Laabs BH, Klein C, Pozojevic J, *et al*. Identifying genetic modifiers of age-associated penetrance in X-linked dystonia-parkinsonism. *Nat Commun*. 2021;12:3216.
8. Westenberger A, Reyes CJ, Saranza G, *et al*. A hexanucleotide repeat modifies expressivity of X-linked dystonia parkinsonism. *Ann Neurol*. 2019;85:812–822.
9. Reyes CJ, Laabs BH, Schaake S, *et al*. Brain regional differences in hexanucleotide repeat length in X-linked dystonia-parkinsonism using nanopore sequencing. *Neurol Genet*. 2021;7:e608.
10. Kawarai T, Pasco PM, Teleg RA, *et al*. Application of long-range polymerase chain reaction in the diagnosis of X-linked dystonia-parkinsonism. *Neurogenetics*. 2013;14:167–169.
11. Harris RS, Cechova M, Makova KD. Noise-cancelling repeat finder: uncovering tandem repeats in error-prone long-read sequencing data. *Bioinformatics*. 2019;35:4809–4811.
12. Nethisinghe S, Kesavan M, Ging H, *et al*. Interruptions of the FXN GAA repeat tract delay the age at onset of Friedreich's ataxia in a location dependent manner. *Int J Mol Sci*. 2021;22:7507.
13. McFarland KN, Liu J, Landrian I, *et al*. Paradoxical effects of repeat interruptions on spinocerebellar ataxia type 10 expansions and repeat instability. *Eur J Hum Genet*. 2013;21:1272–1276.
14. Stolle CA, Frackelton EC, McCallum J, *et al*. Novel, complex interruptions of the GAA repeat in small, expanded alleles of two affected siblings with late-onset Friedreich ataxia. *Mov Disord*. 2008;23:1303–1306.
15. Matsuura T, Fang P, Pearson CE, *et al*. Interruptions in the expanded ATTCT repeat of spinocerebellar ataxia type 10: repeat purity as a disease modifier? *Am J Hum Genet*. 2006;78:125–129.
16. Pearson CE, Eichler EE, Lorenzetti D, *et al*. Interruptions in the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation. *Biochemistry*. 1998;37:2701–2708.
17. Al-Mahdawi S, Ging H, Bayot A, *et al*. Large interruptions of GAA repeat expansion mutations in Friedreich ataxia are very rare. *Front Cell Neurosci*. 2018;12:443.
18. Wright GEB, Black HF, Collins JA, *et al*. Interrupting sequence variants and age of onset in Huntington's disease: clinical implications and emerging therapies. *Lancet Neurol*. 2020;19:930–939.
19. Wenninger S, Cumming SA, Gutschmidt K, *et al*. Associations between variant repeat interruptions and clinical outcomes in myotonic dystrophy type 1. *Neurol Genet*. 2021;7:e572.
20. 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.
21. Deshmukh AL, Caron MC, Mohiuddin M, *et al*. FAN1 exo- not endo-nuclease pausing on disease-associated slipped-DNA repeats: A mechanism of repeat instability. *Cell Rep*. 2021;37:110078.
22. Rolfsmeier ML, Dixon MJ, Lahue RS. Mismatch repair blocks expansions of interrupted trinucleotide repeats in yeast. *Mol Cell*. 2000;6:1501–1507.