Title

De novo repeat interruptions are associated with reduced somatic instability and mild or absent clinical features in myotonic dystrophy type 1

Running Title

De novo variant repeats in myotonic dystrophy

Authors

Sarah A Cumming[1]*, Mark J Hamilton[1,2]*, Yvonne Robb[3], Helen Gregory[4], Catherine McWilliam[5], Anneli Cooper[1], Berit Adam[1], Josephine McGhie[1], Graham Hamilton[6], Pawel Herzyk[6], Michael R Tschannen[7], Elizabeth Worthey[7,8], Richard Petty[9], Bob Ballantyne[2], The Scottish Myotonic Dystrophy Consortium¶, Jon Warner[10], Maria Elena Farrugia[9], Cheryl Longman[2], Darren G Monckton[1]

Affiliations

[1] Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK, [2] West of Scotland Clinical Genetics Service, Queen Elizabeth University Hospital, Glasgow G51 4TF, UK, [3] Clinical Genetics Service, Western General Hospital, Edinburgh EH4 2XU, UK, [4]Department of Clinical Genetics, Aberdeen Royal Hospital, Aberdeen AB25 2ZA, UK, [5] Human Genetics Unit, Ninewells Hospital, Dundee DD1 9SY, UK, [6] Glasgow Polyomics, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G61 1QH, UK, [7] Human and Molecular Genetics Center, Medical College Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA, [8] Hudson Alpha Institute for Biotechnology, 601 Genome Way, NW, Huntsville AL 35806, USA, [9] Department of Neurology, Institute of Neurological Sciences, Queen Elizabeth University Hospital, Glasgow G51 4TF, UK, [10] Molecular Genetics Service, Molecular Medicine Centre, Western General hospital, Crewe Road, Edinburgh EH4 2XU

¶ The Scottish Myotonic Dystrophy Consortium are:
Cheryl Longman, Douglas Wilcox, Alison Wilcox, Richard Petty, Yvonne Robb, Maria Elena Farrugia, Helen Gregory, Alexis Duncan, Catherine McWilliam, John Dean, Anne-Marie Taylor, Lorna MacLeish, Monika Rahman, Anne McKeown, Kirsten Patterson, Mark Hamilton, Bob Ballantyne, Sarah Cumming and Darren G Monckton
The Scottish Myotonic Dystrophy Consortium is a subgroup of the Scottish Muscle Network; Programme Manager Hugh Kennedy, Programme Support Officer Laura Craig.

*These authors contributed equally to this work

Conflict of Interest

Professor Monckton has been a paid scientific consultant of Biogen Idec, AMO Pharma, Charles River and Vertex Pharmaceuticals. Professor Monckton also has a research contract with AMO Pharma.

Acknowledgement of grant support:

Muscular Dystrophy Association (252524)
Wellcome Trust Institutional Strategic Support Fund (ISSF) (097821/Z/11/Z)
Muscular Dystrophy UK (formerly Muscular Dystrophy Campaign) (MC3/1073)
Chief Scientist Office (CAF/MD/15/01)

Address for correspondence:

E-mail: markhamilton1@nhs.net Tel: +44 141 354 9201

ABSTRACT

Myotonic dystrophy type 1 (DM1) is a multisystem disorder, caused by expansion of a CTG trinucleotide repeat in the 3'-untranslated region of the *DMPK* gene. The repeat expansion is somatically unstable and tends to increase in length with time, contributing to disease progression. In some individuals, the repeat array is interrupted by variant repeats such as CCG and CGG, stabilising the expansion and often leading to milder symptoms. We have characterised three families, each including one person with variant repeats that had arisen *de novo* on paternal transmission of the repeat expansion. Two individuals were identified for screening due to an unusual result in the laboratory diagnostic test, and the third due to exceptionally mild symptoms. The presence of variant repeats in all three expanded alleles was confirmed by restriction digestion of small pool PCR products, and allele structures were determined by PacBio sequencing. Each was different, but all contained CCG repeats close to the 3'-end of the repeat expansion. All other family members had inherited pure CTG repeats. The variant repeat-containing alleles were more stable in the blood than pure alleles of similar length, which may in part account for the mild symptoms observed in all three individuals. This emphasises the importance of somatic instability as a disease mechanism in DM1. Further, since patients with variant repeats may have unusually mild symptoms, identification of these individuals has important implications for genetic counselling and for patient stratification in DM1 clinical trials.

Key Words

Myotonic dystrophy type 1, somatic instability, variant repeats, PacBio sequencing

INTRODUCTION

Myotonic dystrophy type 1 (DM1) is a dominantly inherited, multisystem condition. Features include skeletal muscle weakness and myotonia, cardiac conduction abnormalities, frontal balding, ptosis, cataracts, excessive daytime somnolence and insulin resistance [1]. DM1 results from the expansion of a CTG trinucleotide repeat in the 3'-untranslated region of the *DMPK* gene, with pathogenic alleles ranging from around 50 to over 1,000 repeats [2-4]. Age at onset and severity of symptoms are highly variable, and there is a broad inverse correlation between expansion size and age at onset of symptoms [5-7].

The expanded CTG tract is unstable in the germline, and intergenerational expansions account for the phenomenon of genetic anticipation [8]. Furthermore, the tract is genetically unstable in somatic cells. Somatic mutation is expansion-biased, and correlates inversely with age at onset of symptoms [9]. This confounds genotype-phenotype studies, as Southern blotting of restriction digested genomic DNA fails to take account of the effect of age on repeat length distribution. Small-pool PCR (SP-PCR) can resolve somatic mosaicism, enabling calculation of individual-specific mutation rates [10], and allowing estimation of progenitor allele length, which is the major determinant of age at disease onset [11].

In ~3 to 5% of DM1 patients, the CTG repeat expansion contains interruptions, which may include CCG, CTC or GGC motifs [12-14]. The presence of such variant repeats can affect the mutational dynamics of the expanded DM1 allele, with implications for the clinical phenotype. For example, the usual pattern of anticipation may be lost due to increased stability in the germline. The repeats may also be stabilised in the soma, and

patients with variant repeats may exhibit delayed onset, unusually mild symptoms, or atypical patterns of symptoms [12-15].

Variant repeats may also affect diagnostic testing for DM1. This is usually carried out by triplet primed PCR (TP-PCR) [16,17], in which variant repeats can affect primer binding, resulting an atypical appearance of the amplicon ladder. An additional test, such as TP-PCR from the opposite end of the repeat, or Southern blotting of restriction digested genomic DNA, is therefore recommended to avoid false negatives [17]. In the light of the apparent associations between variant repeats and both unusual TP-PCR results and atypical disease symptoms, we hypothesised that patients with variant repeats might be identifiable within our Scottish DM1 patient cohort on this basis.

MATERIALS AND METHODS

Patient identification and recruitment

Scottish adults with DM1 were recruited as part of the ongoing Genetic Variation in Myotonic Dystrophy Study (DMGV). Ethical approval was obtained for recruitment of patients with DM1 from the four major clinical genetics centres in Scotland (Glasgow, Edinburgh, Aberdeen, Dundee; WOS REC 08/S0703/121). Patients were recruited from annual outpatient review appointments, provided whole blood samples for DNA extraction and completed a standardised symptom questionnaire. Written informed consent was obtained, allowing study team access to medical records. Additional written consent was obtained from DMGV14 for publication of data relating to chorionic villus sampling (CVS) and preimplantation genetic diagnosis (PGD).

PCR amplification and Southern blotting of expanded DM1 alleles

Small pool PCR amplification of the CTG repeats and Southern blotting was carried out essentially as described [18], using the flanking primers DM-C and DM-DR [19].

Where necessary, PCRs were supplemented with 10% DMSO (Sigma-Aldrich UK) and the annealing temperature was reduced to 63.5°C. Expanded alleles were screened for AciI-sensitive variant repeats by digestion with AciI (New England Biolabs UK Ltd; restriction site 5'-CCGC-3'). When DMSO had been added to the PCRs, the amplicons were first purified using the QIAquick PCR purification kit (Qiagen UK). The probe used for Southern blotting was a PCR product with 56 CTGs amplified using DM-C and DM-DR. Repeat lengths were estimated using CLIQS 1D gel analysis software (TotalLab UK Ltd.) by comparison against the molecular weight marker. The lower boundary of the expanded alleles was used to estimate the inherited repeat length (the estimated progenitor allele length; ePAL) [19], the major determinant of age at onset of symptoms [11]. The densest point of the distribution of alleles was also used to estimate the modal allele length.

Whole genome amplification of DNA extracted from single cells

Single cells biopsied from a 3-day embryo were collected into PBS, lysed with 200 mM NaOH and 50 mM dithiothreitol at 65°C for 10 minutes, then neutralised using 200 mM tricine. Multiple displacement amplification was then carried out using the REPLI-g® kit (Qiagen). The appropriate amount of whole genome amplified (WGA) template for PCR was determined empirically by serial dilution.

Library preparation for PacBio RS II sequencing

Expanded DM1 alleles were sequenced using the PacBio RS II platform (Pacific Biosciences Inc.) [20]. Material for sequencing was generated by PCR using 250 ng genomic DNA template per patient. For each sample, a different, barcoded forward primer was used. These consisted of the forward flanking primer DM-C, with a 5'-end extension encoding an IonXpress™ barcode (Thermo Fisher Scientific UK).

Amplification conditions were as for non-barcoded primers. Amplicons were concentrated using 1.8X volume Agencourt® AMPure® XP beads (Beckman Coulter UK). The expanded alleles were excised from 1% agarose gels, based on prior estimates of the range of allele lengths obtained by SP-PCR. Amplicons were purified using the QIAquick gel extraction kit (Qiagen UK), quantified using the Qubit® dsDNA HS assay kit (Thermo Fisher Scientific UK) and combined to form an equimolar pool, based on estimated modal allele lengths. The amplicon pool was concentrated further using 1.8X volume Agencourt® AMPure® XP beads, and eluted in 10 mM Tris, pH 8.0. Generation of SMRTbell™ templates and subsequent sequencing were performed at the Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI, USA, or the Earlham Institute, Norwich, UK. Circular consensus sequence (CCS) reads [21] were generated at Milwaukee or Earlham using the CCS algorithm in the SMRT™ Portal provided by PacBio.

Bioinformatic analysis

PacBio sequence reads were analysed using open source tools on the Galaxy instance of Glasgow Polyomics, University of Glasgow [22,23]. CCS reads were demultiplexed by barcode using the Je-demultiplex tool [24], then mapped against DM1-specific reference sequences using BWA-MEM [25,26] and visualised using Tablet [27]. Since we had included a 5'-end barcode only, reverse and complement reads were also demultiplexed to increase the yield of sequence reads for each patient.

Data from all subjects in the three families described have been deposited in the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/). Accession numbers SCV000747869 to SCV000747879.

RESULTS

Two hundred and fifty one adults with DM1 were recruited from annual review appointments. In three families (Fig 1), one individual was identified to be screened for variant repeats, because of an unusual TP-PCR trace or an unusual pattern of symptoms. In all cases, the individual identified for variant screening had been diagnosed with DM1 after requesting a genetic test in the context of a known family history of the condition. None was the index case in their family. All other members of the three families had classical DM1 symptoms, and nothing unusual was noted regarding their molecular diagnostic tests. Clinical summaries are provided in Table 1, with further detail in Supplementary data.

Patient DMGV14 (Family 1, Fig 1) underwent predictive testing for DM1 at the age of 18. TP-PCR from the 3'-flank of the CTG repeat [16] failed to detect an expansion, though Southern blot of restriction digested genomic DNA later confirmed the presence of an expanded allele. At age 27, bidirectional TP-PCR was undertaken in another diagnostic laboratory. This showed a typical ladder of peaks within the affected range on 5'-TP-PCR, but on 3'-TP-PCR a shorter ladder corresponding to ~50 CTG repeats and at a reduced intensity was seen (data not shown). At age 33, DMGV14 had no detectable muscle signs of DM1, and was in full-time employment in an office environment.

DMGV182 (Family 2, Fig 1) requested genetic testing for DM1 at age 43. He denied DM1-specific symptoms, although volunteered a history of jaw discomfort and "slowness" on biting down. Bidirectional TP-PCR from the 3'-end detected ~60 repeats, whereas from the 5'-end, greater than 150 repeats were seen (data not shown). In view of the patient's mild symptoms and atypical TP-PCR result, electromyography (EMG)

192 studies and an ophthalmic examination were requested. EMG showed no myotonia in
193 peripheral muscles, though there was increased insertional activity suggestive of
194 increased muscle membrane irritability. Mild myotonia was detected in masseter.
195 Ophthalmic examination revealed bilateral early posterior subcapsular cataracts.
196 DMGV15 (Family 3, Fig 1) underwent predictive testing for DM1 at age 22. She had
197 not noted any muscle symptoms and had no typical DM1 features. Southern blot
198 analysis of restriction digested genomic DNA confirmed the presence of a CTG repeat
199 expansion. Bidirectional TP-PCR on blood DNA from DMGV15 gave a characteristic
200 ladder consistent with an expanded repeat in the 5'-direction, and a ladder with reduced
201 intensity in the 3'-direction (data not shown). An experienced nurse specialist (YR)
202 noted the clinical discordance between DMGV15 and her classically affected brother,
203 DMGV54, and suggested she be screened for variant repeats. At age 46, DMGV15 had
204 no clear signs or symptoms of DM1, and was in full time employment.
205 Blood DNA samples from all available members of the three families were PCR
206 amplified using flanking DM1 primers. Amplicons were digested with AciI, to screen
207 for CCG or CGG variant repeats. Both alleles from most individuals were amplified
208 successfully (Fig 1), however in the case of DMGV14, the expanded allele only
209 amplified in the presence of 10% DMSO, suggesting it has a particularly high G + C
210 content, possibly indicative of variant repeats. In all three individuals with putative
211 variant repeats (DMGV14, 182 and 15), the expanded allele amplicons digested with
212 AciI. Those of all other family members remained undigested (Fig 1). These data
213 suggest that in each of these three families, variant repeats have arisen *de novo* during
214 paternal transmission of the repeat expansion.

215 Variant repeat interruptions may stabilise the repeat array, reducing the rate of repeat
216 expansion over time [13]. In order to determine whether this is the case for DMGV14,
217 182 and 15, SP-PCRs were compared against those from five DM1 patients of similar
218 ePAL and age to the three variant repeat patients, but whose expanded alleles contain no
219 AciI-sensitive variant repeats (data not shown) (Fig 2A). The ePAL and mode were
220 determined for each patient (Table 2). The difference between the two measures
221 (ΔCTG) may be used as a simple measure of somatic instability. The repeat length
222 change for DMGV14, 182 and 15 is less than for any of the patients that lack variant
223 repeats (Table 2, Fig 2A).
224 DMGV14 has had *in vitro* fertilization and PGD. As part of the PGD protocol, a single
225 cell was removed for DM1 testing. We obtained WGA material from these assays, and
226 also genomic DNA from a previous CVS. In order to determine whether DMGV14's
227 expanded allele was stabilised in the germline, SP-PCR and AciI digestion were carried
228 out (Fig 2B). From eight separate fertilisations, one expanded allele was approximately
229 the same overall length as DMGV14's, and the remaining seven were substantially
230 longer, including one with over 1,300 repeats. All embryos also had a longer stretch of
231 pure CTG repeats at the 5'-end than DMGV14.
232 The expanded alleles from all available members of the three families were next
233 sequenced using the PacBio RSII platform. Reads were aligned against a DM1
234 reference sequence, comprising 600 CTG repeats and 72 bp of 3'-flanking sequence.
235 The aligned reads from DMGV14, 15 and 182 contained CCG mismatches close to the
236 3'-end of the CTG repeat expansion (Fig 3). Most [12-15], but not all [28,29], of the
237 DM1 variant repeats characterised to date have been near the 3'-end of the repeat array.
238 The 5'-ends of the variant repeat-containing reads, and the entire length of the reads

from all other family members, generally consisted of pure CTG repeats (Fig 3, Fig S2). However, each individual read might contain one or more sequence variants, including but not limited to CCG. These had no consistent pattern of distribution, and most likely resulted from a mixture of sporadic somatic variants and PCR and/or sequencing errors. A high percentage of reads from all patients also lacked a G residue in the immediate 3'-flank (Fig 3, Fig S1, Fig S2), which most likely results from a common sequencing error, since it was not seen in Sanger sequenced, PCR amplified DM1 alleles (data not shown). It also appeared to be site-specific, as the mean percentage of reads missing a G was higher for data generated in Wisconsin (61%) than at Earlham (14.5%).

Sequence reads from DMGV14 were aligned against the reference sequence described above. A large number of CCGCTG hexamers was present towards the 3'-end of the repeats (Fig 3). These were variable in number between reads, as was the number of CTG repeats at each end. Aligned reads (603 in total) were examined in detail to determine the consensus pattern of variant repeats as NM_004409.4(DMPK):c.*224_283CTG[180_240]CCGCTG[53_67]CTG[53_67]. (Fig 3). This is broadly consistent with the AciI digestion, which generated AciI-resistant fragments equivalent to ~225 and ~70 CTG repeats (Fig 1). DNA from the WGA samples E1 and E2 was also sequenced, and these reads also contained CCGCTG hexamers close to the 3'-end (Fig S1). Assuming the expanded alleles from the CVS sample, and the WGA samples E3 to E7, also had CCGCTG hexamers near the 3'-end, allele structure was estimated for each (Fig S1) based on the AciI digestion experiment (Fig 2). Reads from the expanded alleles of DMGV14's other family members DMGV165, DMGV83 and DMGV57, contained no germline CCG repeats (Fig 3, Fig S2).

When DMGV182's reads were mapped against the reference sequence described above, CCG variant repeats were visible in a highly variable distribution close to the 3'-flank (Fig 3). Aligned reads (163 in total) were examined to determine the average allele structure. Many of these reads (~17%) contained no CCG repeats at all in the cluster near the 3'-end. Around 26% contained a single CCG, ~26% had two, ~26% had from 3 to 9 CCGs separated by one or more CTG repeats, and ~4% contained from 6 to 26 consecutive CCG repeats (Fig 3). The average structure of the reads is broadly consistent with the AciI digestion, where AciI-resistant fragments corresponding to ~245 CTG and ~60 CTG repeats were generated (Fig 1). Sequence reads from the other family members, DMGV234, 184 and 242 showed no evidence of germline CCG variant repeats (Fig 3, Fig S2), consistent with the AciI digestion (Fig 1).

Around 17% of the aligned PacBio sequence reads from DMGV182 appear to contain no variant repeats in the cluster near the 3'-flank, a much higher percentage than for DMGV14 (<1%). To test whether PCR and/or sequencing errors are responsible for the high percentage of sequence reads that lack CCGs, a single molecule AciI digestion experiment was performed. Multiple reactions using from 7.5 to 50 pg template per reaction were carried out, digested with AciI, blotted and hybridised as before, generating 215 distinct bands over several experiments. No undigested bands were seen. In ~30% of bands, complete digestion occurred, and ~70% of bands were only partially digested by AciI (Fig 4). This suggests that ~70% of individual bands blotted contain a mixture of molecules with and without AciI sites (Fig 4). From this result we infer that at least a single restriction site was present in DMGV182's original germline allele. We therefore estimated the germline allele structure to be NM_004409.4(DMPK):c.*224_283CTG[200_300]CCG[1]CTG[41_59].

When expanded alleles from DMGV15 were aligned against the reference sequence described above, a block of CCG$(CTG)_2$ nonamer variant repeats was visible towards the 3'-end of the reads (Fig 3). For 338 aligned sequence reads, an average structure was determined as NM_004409.4(DMPK):c.*224_283CTG[260_320]CCGCTGCTG[10_14]CTG[15_23]. This is broadly consistent with the AciI digest, which generated an AciI-resistant fragment equivalent to ~245 CTG repeats. A second predicted 135 bp digestion-resistant fragment may be hidden by the non-disease causing allele.

All 251 individuals recruited to DMGV were screened for variant repeats by digestion with AciI. Eighteen individuals in total, including the three described here, had AciI sensitive variant repeats, giving an overall prevalence of 7.2%. This included seven apparently independent occurrences from a total of 169 families (4.1%). No other example of *de novo* gain of variant repeats has been identified to date in this cohort.

DISCUSSION

In this study, we have identified three DM1 patients with CCG variant repeats generated by apparent *de novo* mutations. The variant repeats appear to stabilise the expanded alleles in the blood, and all three patients have symptoms that are milder than expected. We also describe the first use of PacBio SMRT sequencing to study CTG repeat expansions in DM1. PacBio sequencing was previously used to sequence repeat expansions in the fragile X gene [30], and spinocerebellar ataxia types 10 [31,32] and 31 [33]. We have now used this technology to characterise DM1 mutant allele structures in greater detail than was previously possible using cloned DNA fragments [12,13].

Our findings add further evidence for a major contribution of somatic instability to disease progression in DM1. We have previously shown that the principal genetic determinant of age at onset of symptoms in DM1 is the progenitor allele length, and that age at onset is further modified by individual-specific differences in the level of somatic instability [11]. Furthermore, somatic instability is greater in tissues most severely affected, for example skeletal muscle and cerebral white matter [34,35], suggesting tissue-specific differences in expansion rates may account in part for the pattern of symptoms. In the present three cases, reduced somatic expansion was accompanied by milder symptoms, consistent with somatic instability as a key driver of DM1 pathophysiology.

The major factors influencing somatic instability of expanded trinucleotide repeats are not currently fully understood, although there is evidence for a modifying effect of sequence variants in genes involved in DNA mismatch repair [36,37], as well as epigenetic changes at the repeat locus itself [38]. In other trinucleotide repeat disorders, variant repeat motifs have been described acting as 'anchors', reducing the likelihood of misalignment events during DNA processing [39,40]. Consistent with previous studies [13], our data suggest that in DM1 variant repeats have a comparatively major stabilising effect, also increasing the stability of the neighbouring pure CTG sequence.

Other mechanisms have also been explored to account for milder symptoms associated with variant repeats in DM1. The primary cellular pathology in DM1 results from the toxicity of mRNAs that contain expanded CUG repeats. These repeats adopt a hairpin secondary structure [41], and sequester several key regulatory RNA-binding proteins, including muscleblind-like protein 1 (MBNL1), in the form of ribonuclear foci. Perturbations in the relative levels of different splicing factors lead to dysregulation of

alternative splicing of a range of key proteins (reviewed in [42]). Variant repeats within the CUG expansion may alter mRNA secondary structure, which may in turn affect affinity for effector proteins in the DM1 cascade [13]. In addition, a unique, highly polarised pattern of hypermethylation has been described in patients with variant repeats near the 3'-end of the array [43], which could affect local gene expression as well as influencing repeat instability.

In all three cases we describe here, as well as a recently described *de novo* CTC variant repeat [15], the DM1 expansion was paternally inherited. While this may be due to chance, the larger number of cell divisions in male gametogenesis does markedly increase the chance of replication-associated errors [44]. In a previously reported family with inherited CCGCTG variant repeats, expansion of the variant hexamer within the repeat array was observed during paternal transmission [13]. It may therefore be the case that, a single *de novo* substitution having occurred sporadically, subsequent DNA processing errors in postpubertal spermatogenesis facilitated further expansion of the variant sequence to produce the larger blocks seen in families 1 and 3.

In eight separate germline transmissions of DMGV14's CCGCTG variant repeats, the pure CTG repeats at the 5'-end always expanded, and in most cases the overall allele length also increased, including one allele that had over 1,300 repeats. Although the necessary step of WGA could have introduced artefactual changes in the repeat, this seems unlikely, since all PCRs generated a single discrete band for the expanded allele. Furthermore, both the uncut and digested fragment lengths were concordant between trophectoderm and blastomere cells where both were available for a single embryo. The results contrast previously described germline transmissions of variant repeat-containing alleles, where size increases after maternal transmission were only ~50

repeats [13], or where multiple intergenerational contractions occurred in a family [12,15]. While the phenotype that would be associated with the larger germline expansions of DMGV14's allele cannot be predicted, this finding urges caution against counselling patients that variant repeats are unlikely to be associated with congenital onset DM1 on transmission. Characterisation of a greater number of variant repeat families is therefore a priority, to facilitate more accurate genetic counselling of affected individuals regarding implications for prospective pregnancy.

DMGV182's expanded allele was unusual in that ~17% of sequence reads contained no CCGs in the variant-containing zone near the 3'-end (Fig 3). However, in a single molecule SP-PCR and AciI digestion experiment, all bands were at least partially digested by AciI (Fig 4), suggesting there are no alleles that lack variant repeats. One possible explanation is that variant repeats were present in the genomic DNA template, but were sometimes lost during PCR. Partial digestion of a band might result from slipped-strand products with complementary loopouts disrupting the AciI cut site in some molecules (Fig S3). Slipped-strand DNA structures form in disease-associated triplet repeats [40,45], and have recently also been shown to occur *in vitro* during PCR amplification of DM1 alleles [46]. PCR slippage errors might also generate a subset of amplicons that have lost their variant repeats, and hence do not digest (Fig S3). The sequence reads that lacked CCG variant repeats may have been generated by PCR slippage errors, or by errors in the generation of ccs reads from the raw sequence data.

The three cases described, of *de novo* variant repeats accompanied by mild symptoms occurring within known DM1 families, highlight the importance of awareness of variant repeats among clinical genetic services. The cases reported were identifiable from abnormal diagnostic TP-PCR traces, although clinicians should also be mindful of the

possibility of false negative results on TP-PCR, particularly if undertaken in a single direction. Furthermore, there are implications for genetic counselling, since progression of disease and transmission of the expanded allele to offspring may be significantly different in those with variant repeats compared to pure CTG repeats, although accurate predictions cannot be made based on current data. Observations to date also suggest that screening for variant repeats would be an important component of patient stratification for clinical trials, since such individuals may be statistical outliers in terms of disease severity and thus could confound interpretation of trial data, especially where cohorts are small.

ACKNOWLEDGEMENTS AND AFFILIATIONS

The authors thank all participants in the DMGV study for their co-operation, in particular the members of the three families described here. Electromyography in patient DMGV182 was undertaken by Dr. Gregory Moran, Department of Clinical Neurophysiology, Western General Hospital, Edinburgh.

REFERENCES

1. Turner C, Hilton-Jones D. Myotonic dystrophy: diagnosis, management and new therapies. *Curr Opin Neurol* 2014;27:599-606.

2. Brook JD, McCurrach ME, Harley HG*, et al.* Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell* 1992;68:799-808.

3. Mahadevan MS, Amemiya C, Jansen G*, et al.* Structure and genomic sequence of the myotonic dystrophy (DM kinase) gene. *Hum Mol Genet* 1993;2:299-304.

4. Fu YH, Pizzuti A, Fenwick RG, Jr.*, et al.* An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science* 1992;255:1256-8.

406 5. Jaspert A, Fahsold R, Grehl H, Claus D. Myotonic dystrophy: correlation of clinical
407 symptoms with the size of the CTG trinucleotide repeat. *J Neurol* 1995;242:99-
408 104.
409 6. Harley HG, Rundle SA, MacMillan JC*, et al.* Size of the unstable CTG repeat
410 sequence in relation to phenotype and parental transmission in myotonic dystrophy.
411 *Am J Hum Genet* 1993;52:1164-74.
412 7. Hunter A, Tsilfidis C, Mettler G*, et al.* The correlation of age of onset with CTG
413 trinucleotide repeat amplification in myotonic dystrophy. *J Med Genet*
414 1992;29:774-9.
415 8. Harley HG, Rundle SA, Reardon W*, et al.* Unstable DNA sequence in myotonic
416 dystrophy. *Lancet* 1992;339:1125-8.
417 9. Wong LJ, Ashizawa T, Monckton DG, Caskey CT, Richards CS. Somatic
418 heterogeneity of the CTG repeat in myotonic dystrophy is age and size dependent.
419 *Am J Hum Genet* 1995;56:114-22.
420 10. Higham CF, Morales F, Cobbold CA, Haydon DT, Monckton DG. High levels of
421 somatic DNA diversity at the myotonic dystrophy type 1 locus are driven by ultra-
422 frequent expansion and contraction mutations. *Hum Mol Genet* 2012;21:2450-63.
423 11. Morales F, Couto JM, Higham CF*, et al.* Somatic instability of the expanded CTG
424 triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and
425 modifier of disease severity. *Hum Mol Genet* 2012;21:3558-67.
426 12. Musova Z, Mazanec R, Krepelova A*, et al.* Highly unstable sequence interruptions
427 of the CTG repeat in the myotonic dystrophy gene. *Am J Med Genet A*
428 2009;149A:1365-74.

13. Braida C, Stefanatos RK, Adam B*, et al.* Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dystrophy type 1 patients. *Hum Mol Genet* 2010;19:1399-412.

14. Santoro M, Masciullo M, Pietrobono R*, et al.* Molecular, clinical, and muscle studies in myotonic dystrophy type 1 (DM1) associated with novel variant CCG expansions. *J Neurol* 2013;260:1245-57.

15. Pesovic J, Peric S, Brkusanin M*, et al.* Molecular genetic and clinical characterization of myotonic dystrophy type 1 patients carrying variant repeats within DMPK expansions. *Neurogenetics* 2017;18:207-18.

16. Warner JP, Barron LH, Goudie D*, et al.* A general method for the detection of large CAG repeat expansions by fluorescent PCR. *J Med Genet* 1996;33:1022-6.

17. Kamsteeg EJ, Kress W, Catalli C*, et al.* Best practice guidelines and recommendations on the molecular diagnosis of myotonic dystrophy types 1 and 2. *Eur J Hum Genet* 2012;20:1203-8.

18. Gomes-Pereira M, Bidichandani SI, Monckton DG. Analysis of unstable triplet repeats using small-pool polymerase chain reaction. *Methods Mol Biol* 2004;277:61-76.

19. Monckton DG, Wong LJ, Ashizawa T, Caskey CT. Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. *Hum Mol Genet* 1995;4:1-8.

20. Eid J, Fehr A, Gray J*, et al.* Real-time DNA sequencing from single polymerase molecules. *Science* 2009;323:133-8.

21. Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res* 2010;38:e159.

22. Giardine B, Riemer C, Hardison RC*, et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 2005;15:1451-5.

23. Afgan E, Baker D, van den Beek M*, et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* 2016;44:W3-W10.

24. Girardot C, Scholtalbers J, Sauer S, Su SY, Furlong EE. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinformatics* 2016;17:419.

25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754-60.

26. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.arXiv:1303.3997 [q-bio.GN].

27. Milne I, Stephen G, Bayer M*, et al.* Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform* 2013;14:193-202.

28. Dryland PA, Doherty E, Love JM, Love DR. Simple Repeat-Primed PCR Analysis of the Myotonic Dystrophy Type 1 Gene in a Clinical Diagnostics Environment. *J Neurodegener Dis* 2013;2013:857564.

29. Botta A, Rossi G, Marcaurelio M*, et al.* Identification and characterization of 5' CCG interruptions in complex DMPK expanded alleles. *Eur J Hum Genet* 2017;25:257-61.

30. Loomis EW, Eid JS, Peluso P*, et al.* Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res* 2013;23:121-8.

31. Landrian I, McFarland KN, Liu J*, et al.* Inheritance patterns of ATCCT repeat interruptions in spinocerebellar ataxia type 10 (SCA10) expansions. *PLoS One* 2017;12:e0175958.

32. Schule B, McFarland KN, Lee K*, et al.* Parkinson's disease associated with pure ATXN10 repeat expansion. *NPJ Parkinsons Dis* 2017;3:27.

33. Doi K, Monjo T, Hoang PH*, et al.* Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing. *Bioinformatics* 2014;30:815-22.

34. Anvret M, Ahlberg G, Grandell U*, et al.* Larger expansions of the CTG repeat in muscle compared to lymphocytes from patients with myotonic dystrophy. *Hum Mol Genet* 1993;2:1397-400.

35. Jinnai K, Mitani M, Futamura N*, et al.* Somatic instability of CTG repeats in the cerebellum of myotonic dystrophy type 1. *Muscle Nerve* 2013;48:105-8.

36. Zhao XN, Usdin K. The Repeat Expansion Diseases: The dark side of DNA repair. *DNA Repair (Amst)* 2015;32:96-105.

37. Morales F, Vasquez M, Santamaria C*, et al.* A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. *DNA Repair (Amst)* 2016;40:57-66.

38. Pearson CE, Nichol Edamura K, Cleary JD. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet* 2005;6:729-42.

498 39. Choudhry S, Mukerji M, Srivastava AK, Jain S, Brahmachari SK. CAG repeat
499 instability at SCA2 locus: anchoring CAA interruptions and linked single
500 nucleotide polymorphisms. *Hum Mol Genet* 2001;10:2437-46.
501 40. Pearson CE, Eichler EE, Lorenzetti D*, et al.* Interruptions in the triplet repeats of
502 SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA
503 (S-DNA) formation. *Biochemistry* 1998;37:2701-8.
504 41. Koch KS, Leffert HL. Giant hairpins formed by CUG repeats in myotonic dystrophy
505 messenger RNAs might sterically block RNA export through nuclear pores. *J*
506 *Theor Biol* 1998;192:505-14.
507 42. Pettersson OJ, Aagaard L, Jensen TG, Damgaard CK. Molecular mechanisms in
508 DM1 - a focus on foci. *Nucleic Acids Res* 2015;43:2433-41.
509 43. Santoro M, Fontana L, Masciullo M*, et al.* Expansion size and presence of
510 CCG/CTC/CGG sequence interruptions in the expanded CTG array are
511 independently associated to hypermethylation at the DMPK locus in myotonic
512 dystrophy type 1 (DM1). *Biochim Biophys Acta* 2015;1852:2645-52.
513 44. Rahbari R, Wuster A, Lindsay SJ*, et al.* Timing, rates and spectra of human
514 germline mutation. *Nat Genet* 2016;48:126-33.
515 45. Pearson CE, Sinden RR. Alternative structures in duplex DNA formed within the
516 trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry*
517 1996;35:5041-53.
518 46. Gomes-Pereira M, Monckton DG. Ethidium Bromide Modifies The Agarose
519 Electrophoretic Mobility of CAG•CTG Alternative DNA Structures Generated by
520 PCR. *Front Cell Neurosci* 2017;11:153.
521

Titles and Legends to Figures

Figure 1. One member of each family has AciI-sensitive variant repeats. Each family tree shows only affected individuals; the proband is marked with an arrow. The individuals suspected to have variant repeat interruptions are shown in grey. ID = patient code, Age = age at sampling. The panels show small pool PCR products from 500 pg template DNA, undigested (-) or digested with the restriction enzyme AciI that recognises CCG or CGG variant repeats (+) and Southern blotted. The expanded alleles from DMGV14, 182 and 15 each contain AciI-sensitive variant repeats and have been digested; all other expanded alleles remain uncut. The non-disease associated allele (N), and molecular weight marker (bp) are indicated. The equivalent number of triplet repeats in undigested fragments (rpts) for each molecular weight marker was determined by subtracting the length of the sequence flanking the repeat (106 bp), and dividing by three.

Figure 2. Expanded alleles containing variant repeats are stabilised in blood DNA, but not in the germline. A. The panels on the left show small pool PCR products from 300 pg template DNA from the three patients with variant repeats. The panels on the right show small pool PCR products from five patients without known variant repeats and with broadly similar ages and repeat lengths. The white dashed lines show the estimated progenitor allele length and the mode. The expanded alleles from the three patients with variant repeats are stabilised compared to those without. ID = patient code, Age = age at sampling. The non-disease causing allele (N), molecular weight marker (bp) and the equivalent number of triplet repeats (rpts) are indicated. **B.** The panels show small pool PCR products from 500 pg genomic DNA (DMGV14, CVS), or an

empirically determined equivalent of whole genome amplified DNA, undigested (-) or digested with the restriction enzyme AciI that recognises CCG or CGG variant repeats (+) and Southern blotted. CVS = chorionic villus sample from an affected pregnancy, E1 to E7 = whole genome amplified samples from seven embryos generated by IVF. DNA was amplified from blastomere (blast) or trophectoderm (troph). The non-disease causing allele (N), size in base pairs (bp) and the number of triplet repeats in undigested fragments (rpts) are indicated.

Figure 3. PacBio sequencing confirms that CCG variant repeats have arisen *de novo* in each family. For each family, the top panel shows the 3'-end of PacBio sequence reads for both father and child, zoomed-out (left) and zoomed-in (right). Mismatches compared to the reference sequence (usually the C in a CCG repeat) appear black. The approximate number of reads in the zoomed-out panels is shown to the left of the top panel. The junction between the repeats and the 3'-flank, where a G nucleotide is frequently missing from the sequence reads, is marked ΔG. The distance in repeats (rpts) from the 3'-flank is marked below the zoomed-out panel showing reads from the individual with variant repeats. For each family, the schematic diagram below the sequence read panels shows the average allele structure determined by scoring reads from the individual with variant repeats.

Figure 4. Single molecule PCR products from DMGV182 are always at least partially digested by AciI. The panels show single molecule PCR products from DMGV182 undigested (-) or digested with the restriction enzyme AciI that recognises CCG or CGG variant repeats (+) and Southern blotted. The pairs of panels on the left (1 to 3) show examples of molecules that appear completely digested by AciI. The pairs of panels on the right (4 to 6) show examples of molecules only partially digested by AciI.

570 The white arrows in the right-hand panel of each pair show the digestion product(s) that
571 correspond to each PCR product. This contrasts with the PacBio sequencing data for the
572 same sample, where 17% of CCS reads did not contain CCG variant repeats near the 3'-
573 end.

Family	DMGV ID	Age at last review	Self-reported age at symptom onset	MIRS	Neuromuscular assessment	Cardiac abnormality	Cataract	Other diagnoses	In full time education or employment?	Age at DNA sampling (years)	Progenitor allele length (repeats)	Modal allele size (repeats)
1	14	33	Denies symptoms	1	No clinically apparent weakness or myotonia	-	-	Hypothyroid	Y	25.5	381	418
1	57	36	5	4	Marked facial weakness with ptosis. Distal weakness with relative sparing of deltoids. Grip myotonia. Uses bilateral ankle foot orthoses.	-	-	Paraumbilical fistula Horseshoe kidney Malone procedure for faecal incontinence	N	20.5	597	922
1	165	62	28	4	Dysarthria. Walks with a stick indoors, wheelchair for outdoors	First degree heart block	+	Diverticulosis Hypokalaemia Ischaemic heart disease Barrett's oesophagus	N	59	383	811
1	83	71	38	4	Grip MRC grade 2/5, proximal power 4/5	Implantable cardiac defibrillator *in situ*	+	Seen by speech and language therapist for swallowing issues	N	46	105	131
2	182	37	Denies symptoms	1	No clinically apparent weakness or myotonia. Mild masseter myotonia and peripheral muscle membrane irritability on EMG	-	+	Dermal fibrosis	Y	33.5	293	368
2	184	31	20	2	Walks independently. Grip myotonia.	-	-	Oligospermia Bowel symptoms with bacterial overgrowth Low grade neutropenia Low immunoglobulin G Recurrent pilomatrixoma	Y	28	288	652
2	206	69	60	2	Walks independently, mild myotonia only. Jaw weakness	Electrical cardioversion for atrial flutter	-	Investigated for abnormal liver function tests Moderate pharyngeal dysphagia Borderline hypercalcaemia	N (Retired)	70	90	131
2	242	65	ND	2	Walks independently, no myotonia	-	-	Osteopenia Recurrent primary hyperparathyroidism	N (Retired)	65	80	99

3	15	46	Denies symptoms	1	No clinically apparent weakness or myotonia	Mitral valve replacement for congenital heart anomaly	-	None	**Y**	39	303	379
3	54	43	35	4	Bilateral ankle foot orthoses for foot drop. Distal weakness with poor grip strength, forearm weakness and wasting with relative sparing of deltoid.	-	-	None	N	40	146	230
3	234	53	ND	ND	Severe generalised muscle weakness, marked grip and percussion myotonia, bilateral ptosis	ND	ND	Sudden death at age 54 secondary to respiratory failure	ND	ND	496	663

Table 1: Summary of clinical features in families 1, 2 and 3. Individuals found to carry variant repeat alleles are highlighted in grey. MIRS = Muscle Impairment Rating Scale; MRC = Medical Research Council; ND = no data, EMG = electromyography.

Table 2: Somatic instability of repeat expansions with and without CCG variant repeat interruptions

Patient ID	Age at sampling (years)	Variant repeats	ePAL (repeats)	Mode (repeats)	ΔCTG
DMGV14	25.5	Y	381	418	37
DMGV182	33.5	Y	294	359	65
DMGV15	39	Y	327	385	58
DMGV82	28	N	337	533	196
DMGV158	33	N	277	643	366
DMGV159	21.5	N	346	490	144
DMGV184	28	N	308	629	321
DMGV262	34	N	304	516	212

Figure 1

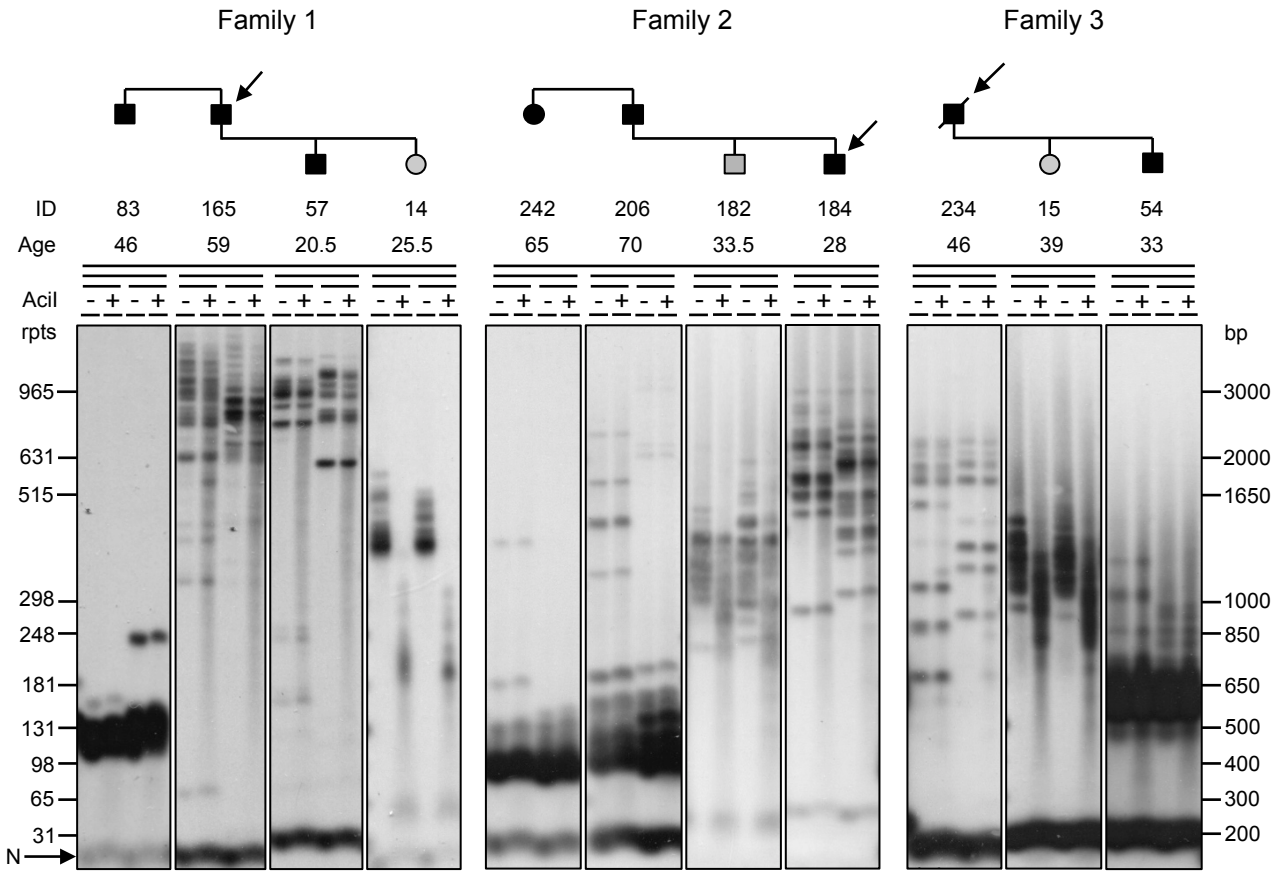

Family 1
Family 2
Family 3
ID
83
165
57
14
242
206
182
184
234
15
54
Age
46
59
20.5
25.5
65
70
33.5
28
46
39
33
AciI
- + - + - + - + - + - + - + - + - + - + - + - + - +
rpts
bp
965
631
515
298
248
181
131
98
65
31
N
3000
2000
1650
1000
850
650
500
400
300
200

Figure 2

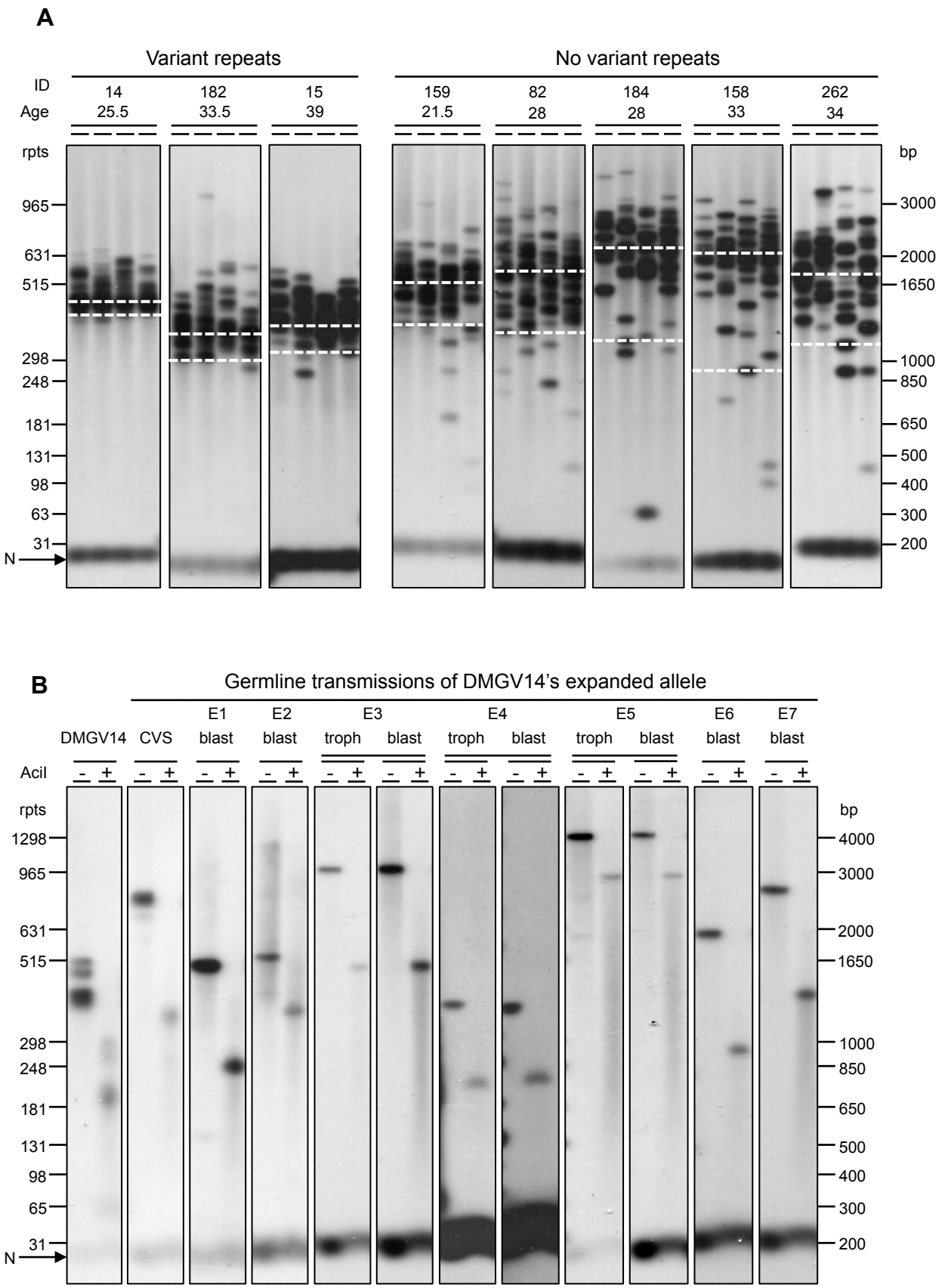

A
Variant repeats
No variant repeats
ID
Age
14
25.5
182
33.5
15
39
159
21.5
82
28
184
28
158
33
262
34
rpts
bp
965
631
515
298
248
181
131
98
63
31
N
3000
2000
1650
1000
850
650
500
400
300
200
B
Germline transmissions of DMGV14's expanded allele
DMGV14
CVS
E1
blast
E2
blast
E3
troph
blast
E4
troph
blast
E5
troph
blast
E6
blast
E7
blast
AcIl
- +
rpts
bp
1298
965
631
515
298
248
181
131
98
65
31
N
4000
3000
2000
1650
1000
850
650
500
400
300
200

Figure 3

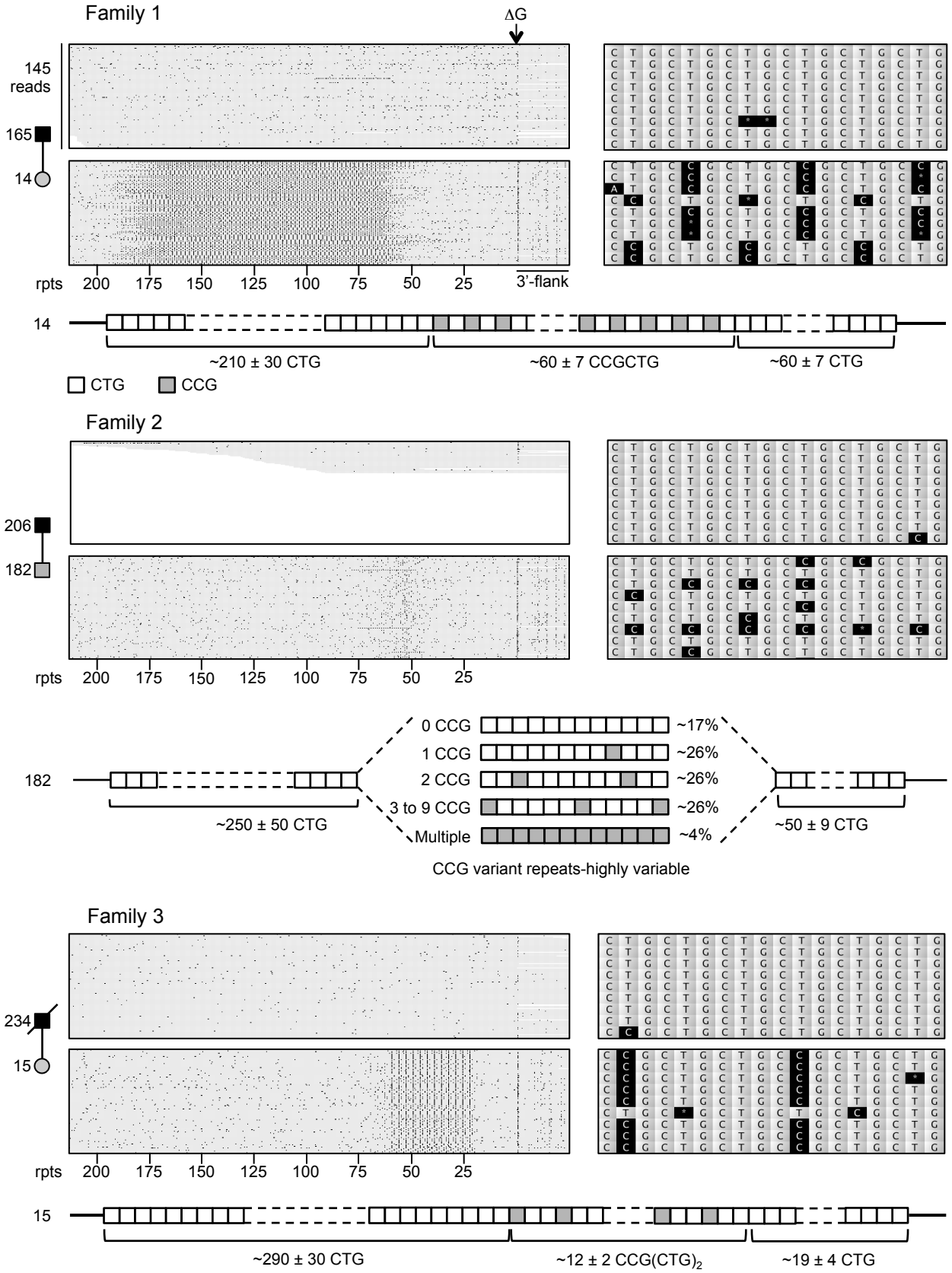
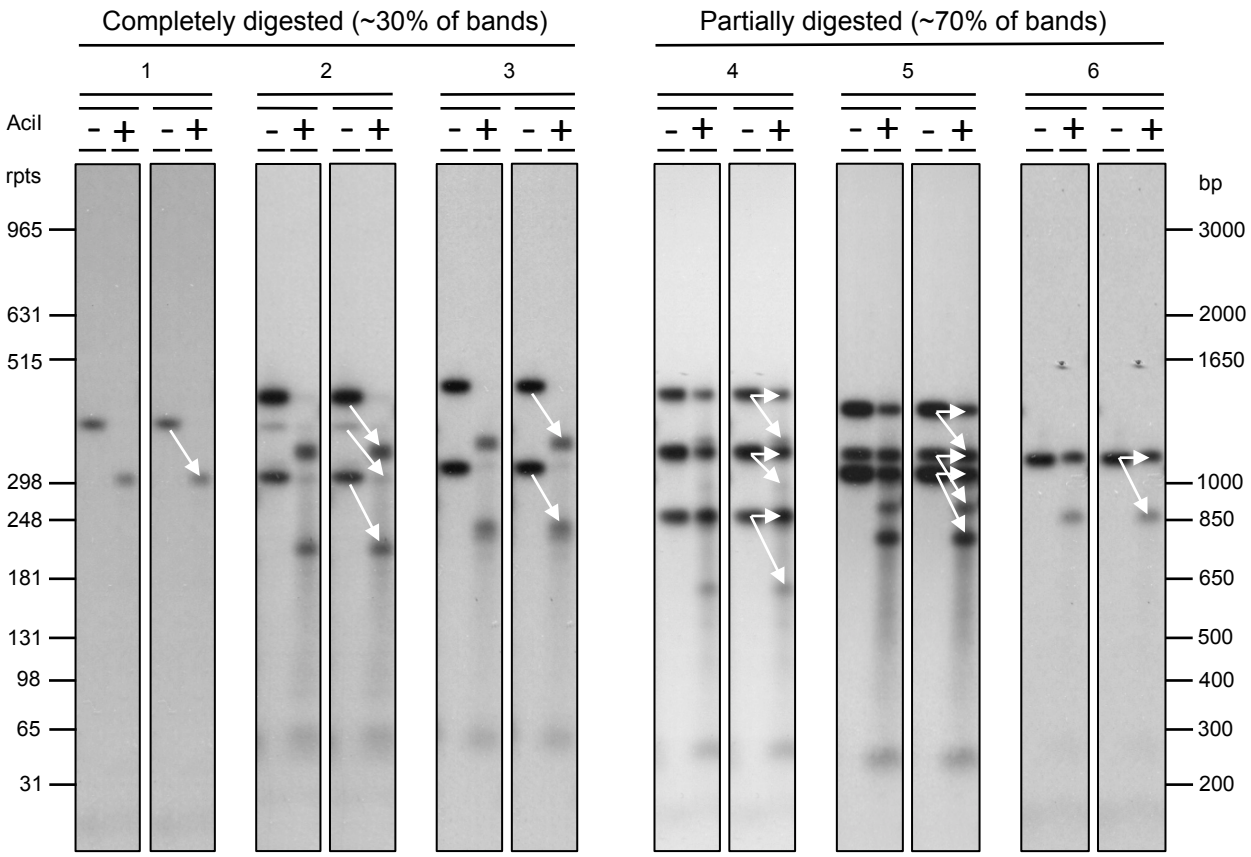

Family 1
ΔG
145 reads
165
14
rpts 200 175 150 125 100 75 50 25
3'-flank
14
~210 ± 30 CTG
~60 ± 7 CCGCTG
~60 ± 7 CTG
CTG
CCG
Family 2
206
182
rpts 200 175 150 125 100 75 50 25
182
0 CCG ~17%
1 CCG ~26%
2 CCG ~26%
3 to 9 CCG ~26%
Multiple ~4%
~250 ± 50 CTG
~50 ± 9 CTG
CCG variant repeats-highly variable
Family 3
234
15
rpts 200 175 150 125 100 75 50 25
15
~290 ± 30 CTG
~12 ± 2 CCG(CTG)2
~19 ± 4 CTG

Figure 4


Completely digested (~30% of bands)
Partially digested (~70% of bands)
1
2
3
4
5
6
AciI
- + - + - + - + - + - + - + - + - + - + - +
rpts
965
631
515
298
248
181
131
98
65
31
bp
3000
2000
1650
1000
850
650
500
400
300
200