Biallelic Expansion of an intronic Repeat in the *RFC1* Gene is a common cause of Late-Onset Ataxia

Andrea Cortese (1)*, Roberto Simone (2), Rosin Sullivan (1)[$], Jana Vandrovcova (1)[$], Huma Tariq (1), Yau Way Yan (1), Jack Humphrey (1), Zane Jaunmuktane (2), Prasanth Sivakumar (1), James Polke (3), Muhammad Ilyas (4), Eloise Tribollet (1), Pedro J Tomaselli (5), Grazia Devigili (6), Ilaria Callegari (7), Maurizio Versino (7,8), Vincenzo Salpietro (1), Stephanie Efthymiou (1), Diego Kaski (1), Nick W Wood (1), Nadja S Andrade (9), Elena Buglo (10), Adriana Rebelo (10), Alexander M Rossor (1), Adolfo Bronstein (2), Pietro Fratta (1), Wilson J Marques (5), Stephan Züchner (10), Mary M Reilly (1)[#], Henry Houlden (1)*[,#]

(1) Department of Neuromuscular Disease and (2) Department of Clinical and Movement Neurosciences (3) Neurogenetics Laboratory, UCL Institute of Neurology and The National Hospital for Neurology, Queen Square, London, WC1N 3BG, UK.
(4) Department of Biotechnology, Islamabad University, Islamabad and Punjab University, Lahore, Pakistan.
(5) Department of Neurology, School of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil.
(6) UO Neurologia I, Fondazione IRCCS Istituto Neurologico "Carlo Besta", Milano
(7) IRCCS Mondino Foundation, Pavia, Italy (8) Department of Brain and Behavioral Sciences, University of Pavia, Pavia, Italy
(9) Department of Psychiatry and Behavioural Sciences, Center for Therapeutic Innovation, (10) Dr. John T. Macdonald Foundation Department of Human Genetics and John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, USA.
[$] equal contribution
[#] these authors jointly directed this project
* corresponding authors

Correspondence to
Andrea Cortese, MD PhD
UCL Institute of Neurology, Department of Neuromuscular Disease,
Queen Square, WC1N 3BG
London, United Kingdom
Email: andrea.cortese@ucl.ac.uk

Henry Houlden, MD PhD
UCL Institute of Neurology, Department of Neuromuscular Disease,
Queen Square, WC1N 3BG
London, United Kingdom
Email: h.houlden@ucl.ac.uk

Abstract

Late-onset ataxia is common, often idiopathic and can result from cerebellar, proprioceptive or vestibular impairment, when in combination also termed cerebellar ataxia, neuropathy, vestibular areflexia syndrome (CANVAS). We used non-parametric linkage analysis and genome sequencing to identify a biallelic intronic AAGGG repeat expansion in the replication factor C subunit-1 (*RFC1*) as the cause of familial CANVAS and a frequent cause of late-onset ataxia, particularly if sensory neuronopathy and bilateral vestibular areflexia coexisted. The expansion, which occurs in the polyA tail of an AluSx3 element and differs in terms of both size and nucleotide sequence from the reference $(AAAAG)_{11}$ allele, does not affect *RFC1* expression in patient peripheral and brain tissue suggesting no overt loss-of-function. These data, along with the European expansion carrier frequency of 0.7%, implies that biallelic AAGGG expansion in *RFC1* is a frequent cause of late-onset ataxia.

INTRODUCTION

Late-onset ataxia, postural imbalance and falls are a frequent reason for neurological consultation. Physiologically, motor coordination is achieved under visual control thanks to the cerebellar integration of proprioceptive information conveyed by large-fibre sensory neurons and vestibular inputs. Failure of any, or a combination of these systems can result in ataxia (1–6). Both acquired and genetic causes are known but a large proportion remain idiopathic.

Previous studies suggest that, there is a spectrum of clinical signs from pure idiopathic late-onset cerebellar degeneration (ILOCA) through to the combined degeneration of the cerebellum and its vestibular and sensory afferents, which has been named cerebellar atrophy, neuropathy and vestibular areflexia syndrome (CANVAS) (**Figure 1A)** (7). CANVAS is an adult-onset slowly progressive neurological disorder characterized by imbalance, sensory neuropathy (neuronopathy), bilateral vestibulopathy(8), chronic cough, and occasionally autonomic dysfunction. (9). Typically, sensory action potentials and somatosensory potentials are absent throughout, brain MRI shows cerebellar atrophy and vestibular testing is consistent with impaired vestibular function bilaterally (10,9,11–17). Late-onset ataxia and CANVAS are usually sporadic, but occasionally occur in siblings, raising the possibility of recessive transmission. However, initial attempts to identify the underlying genetic defect by whole-exome sequencing were unsuccessful.

Using non-parametric linkage analysis and whole-genome sequencing we identified a recessive intronic AAGGG repeat expansion in replication factor C subunit 1 (*RFC1*) as a the cause of familial CANVAS. The expansion occurs in the polyA tail of an AluSx3 element and differs in terms of both size and nucleotide sequence from the reference $(AAAAG)_{11}$ allele. Screening of additional sporadic cases with late-onset ataxia confirmed the presence of the mutated AAGGG repeat expansion in 22% of them, and in higher percentages in if sensory neuronopathy and/or bilateral vestibular areflexia coexisted, suggesting that it represents a frequent and certainly underrecognized cause of late-onset ataxia.

RESULTS

Genetic study

We genotyped 29 individuals, 23 affected and 6 unaffected, from 11 families (**Figure 1B).** The majority of the families consisted of affected sibships except two first-degree cousins from non-consanguineous families (Fam 5b-2 and Fam 6b-1). None of the families had convincing evidence of vertical disease transmission.

Assuming a recessive mode of inheritance, non-parametric linkage analysis identified a single peak at position 4q14 with cumulative maximum HLOD of 5.8 (**Figure 2A**). Haplotype analysis defined a 1.7 MB region between markers rs6814637 and

rs10008483 (chr4:38977921-40712231) where within single families affected siblings shared the same maternal and paternal alleles as opposed to unaffected brothers and sisters, who had at most one of them (**Figure 2B**). The region contains 21 known HGNC genes **(Supplementary Table 1).** Homozygosity mapping in consanguineous families showed that the previously identified 1.7 MB region is encompassed in a larger run of homozygosity of 12 MB shared by the affected siblings (**Supplementary Figure 1**). Of interest, inside the 1.7Mb region four SNPs (rs2066790, rs11096992, rs17584703 and rs6844176, bold highlighted), mapping inside a region encompassing all exons of replication factor C subunit 1 (*RFC1*) and the last exon of WD repeat domain 19 (*WDR19*) were shared by all affected individuals from different families except for individual Fam 5b-2, raising the possibility of a founder haplotype (**Figure 2C and 2D**).

Whole-exome sequencing was previously performed in seven individuals (Fam 1-1, Fam 1-2, Fam 1-3, Fam 3-1, Fam 3-2, Fam 4-2, Fam 4-3) from three unrelated families (Fam1, Fam3, Fam4), but did not identify recurrent non-synonymous variants within the coding regions of the genes encompassed in the 1.7 Mb region (data not shown). We next performed whole-genome sequencing (WGS) in an additional 6 affected individuals (Fam 2-1, Fam 8-2, Fam 8-3, Fam 6a-2, Fam 5a-2, Fam 7-1), 1 unaffected subject (Fam 8-1) from four unrelated families and one sporadic case (s9). Analysis for non-synonymous variants and copy number variants did not reveal changes recurring in the affected families. By visually inspecting the aligned paired reads inside the 1.7 Mb region we noted in all CANVAS patients a reduced read depth in a region encompassing a simple tandem $(AAAAG)_{11}$ repeat at position chr4:39350045-39350103 (**Figure 3A**). Inside the microsatellite region, the reference $(AAAAG)_{11}$ repeat was replaced in patients by a variable number of AAGGG repeated units, which were detected on the reads mapped to either side of the short tandem repeat. However, none of the reads could span across the microsatellite region from one side to the other, suggesting the presence of a biallelic expansion of the AAGGG repeat unit (**Figure 3B**). WGS from an unaffected sibling (Fam 8-1) showed an equal distribution of interrupted reads containing the mutated AAGGG repeated unit change as well as reads containing the AAAAG repeat.

We then performed repeat-primed PCR (RPPCR) with primers targeting the mutant AAGGG pentanucleotide unit and confirmed the presence of an AAGGG repeat expansion in all affected members from 11 families, as well as in unaffected carriers (**Figure 3C**). Flanking PCR using standard conditions failed to amplify the region in all patients suggesting the presence of a large expansion on both alleles, as opposed to their unaffected siblings for whom at least one allele could be amplified by PCR (data not shown).

We next screened a cohort of 150 patients diagnosed with sporadic late onset ataxia and we identified additional 33 (22%) sporadic cases carrying the recessive AAGGG repeat expansion, as defined by a positive RPPCR for AAGGG repeat unit and the absence of PCR amplifiable products by standard flanking PCR. The percentage of positive cases raised to

63% (32/51) if considering cases with late onset cerebellar ataxia and sensory neuronopathy and to 92% (11/12) in cases will full CANVAS syndrome. Taking advantage of two informative single-nucleotide polymorphisms rs11096992 and rs2066790, by PCR and direct sequencing we observed that all additional sporadic cases but individual s23 shared the same haplotype as familial CANVAS cases.

By long-range PCR we were able to amplify and confirm by Sanger sequencing in all patients the presence of the AAGGG expansion (**Figure 3D**). However long-range PCR did not allow sizing of the repeat expansion as PCR is error-prone and contraction of repeated regions during PCR cycling have been previously demonstrated (18). Therefore, Southern blots were conducted in 34 cases and confirmed the presence of biallelic large expansions in all of them. Biallelic expansions could be visualized as two distinct bands in subjects carrying expansions of different sizes, or one thick band if the expanded alleles had a similar size (**Supplementary Figure 2**). Four unaffected siblings from four families were also included and they all carried one expanded and one normal allele. Although the expansion size could vary across different families, ranging from around 400 to 2000 repeats, in the majority of cases approximately 1000 repeats were observed. Repeat size was relatively stable in siblings within single families. There was no association between age at onset and the number of AAGGG repeat units on either the smaller or larger allele (n= 34; r=-0.006, p=0.97 and r=-0.04, p=0.81, respectively)

Polymorphic conformations and allelic distribution of the short tandem repeat locus in the normal population

Recessive AAGGG expansion, as defined by the combination of positive RPPCR targeting the AAGGG repeat and the absence of a PCR amplifiable product on flanking PCR, were not observed in 304 healthy controls screened. RPPCR analysis targeting the AAGGG repeat showed that 0.7% (4 out of 608 chromosomes tested) carried an AAGGG expansion in heterozygous state. Southern blot analysis was performed in all of them and confirmed the presence of an expanded allele ranging from 200 to 880 repeats (mean 720 ± 360) The chr4:39350045-39350103 locus, where the expansion resides, was shown to be highly polymorphic in the normal population and, besides the rare AAGGG expansion allele $(AAGGG)_{exp}$, three other conformations were observed: $(AAAAG)_{11}$, $(AAAAG)_{exp}$, $(AAAGG)_{exp}$ (**Figure 4A**). The $(AAAGG)_{exp}$ was often interrupted by XX By a combinatory approach of flanking PCR, RPPCR targeting one of the three possible nucleotide sequences, as well as Southern blot and Sanger sequencing in selected cases, we observed an allelic distribution of 75.5% (n = 459) for the $(AAAAG)_{11}$ allele, 13.0% (n = 79) for $(AAAAG)_{exp}$ allele, 7.9% (n = 48) for $(AAAGG)_{exp}$ allele, and, as per above, 0.7% (n=4) for the $(AAGGG)_{exp}$ allele (**Figure 4B**). Average size of $(AAAAG)_{exp}$ ranged from 40 to 400 repeats (mean 160 ± 72) and $(AAAGG)_{exp}$ ranged from 40 to 880 (mean 236 ± 181) (**Figure 4C**).

Eight healthy subjects had biallelic repeat expansions of a distinct repeated unit: $(AAAAG)_{exp}/(AAGGG)_{exp}$ in one case, $(AAAGG)_{exp}/(AAGGG)_{exp}$ in one case and

$(AAAAG)_{exp}$ /$(AAAGG)_{exp}$ in six cases. 22 cases likely had two expansions of the repeated AAAAG unit and nine of the repeated AAGGG unit, as defined by a positive RPPCR for the target repeat and two distinct bands on the southern blot, although we cannot exclude that one of the two alleles may be characterized by a distinct nucleotide sequence, which was not considered in the present study. Indeed, 9 additional subjects had no PCR amplifiable product on flanking PCR and were negative for RPPCR targeting the AAAAG, AAAGG, or AAGGG repeated units suggesting the potential existence of other possible allelic conformations in 3% (n=18) of tested chromosomes. Southern blot could not be performed because of insufficient amount of DNA in these cases.

The haplotype associated in most patients with the AAGGG repeat expansion has an allelic carrier frequency in 1000genome control population of 18%. Based on rs11096992 and rs2066790 markers genotyping, the disease-associated haplotype rs2066790 (AA), rs11096992 (AA) was absent in recessive state from healthy individuals who carried two (AAAAG)11 alleles, two $(AAAAG)_{exp}$ alleles or a compound $(AAAAG)_{11}$/$(AAAAG)_{exp}$ genotype, but was observed in three out of nine carriers of two $(AAAGG)_{exp}$ alleles and one healthy subject with $(AAGGG)_{exp}$/$(AAAGG)_{exp}$ alleles, suggesting its possible association with both $(AAGGG)_{exp}$ and $(AAAGG)_{exp}$ configurations of the repeated unit, but not $(AAAAG)_{11}$ or $(AAAAG)_{exp.}$

Clinical features of patients carrying the recessive AAGGG repeat expansion

The clinical features of 56 cases carrying the recessive intronic AAGGG repeat expansion, including 23 familial and 33 sporadic cases, are summarized in **Table 1** and detailed in **Supplementary Table 2.** All cases were of Caucasian ancestry. Apart from a higher frequency of vestibular areflexia in familial CANVAS, clinical features were otherwise similar in familial and sporadic cases; hence data are presented together. Mean age of onset was 54 ± 9 (35-73) years and mean disease duration at examination 11 ± 7 (1-30) years. The most common complain at disease onset was unsteadiness, which was reported by 84% of patients, and frequently described as being worse in the dark. 37% of patients complained of chronic cough, which in some cases could precede by decades the onset of the walking difficulties. The neurologic examination invariably showed signs in keeping with a large fibre sensory neuropathy, 80% of patients had signs of cerebellar involvement, and overall 54% had evidence of bilateral vestibular areflexia. 23% of patients had concurrent autonomic nervous system involvement, particularity affecting micturition and defecation. Nerve conduction studies confirmed the presence of a non-length-dependent sensory neuropathy in all cases tested, as opposed to an entirely normal motor conduction study in most patients. Cerebellar atrophy was identified in 35 (83%) of 42 cases who underwent an MRI or CT scan.

Neuropathological examination

Pathological examination was conducted in a patient with CANVAS who carried the biallelic AAGGG repeat expansion and compared with a patient with genetically

confirmed Friedreich's ataxia, one patient with spinocerebellar ataxia 17 (SCA17) and one case with *C9orf72*-related frontotemporal dementia (FTD), as well as control brains (**Figure 5**). The patient with CANVAS showed severe, widespread depletion of Purkinje cells with associated prominent Bergmann gliosis, whilst cell density in the granule cell layer was well preserved. Loss of Purkinje cells was also observed in Friedreich's ataxia, SCA17 and, to a much lesser extent, in *C9orf72*-related FTD, but not in control brain. Similar to Friedreich's ataxia and control brain, and as opposed to SCA17 and a *C9orf72*-related FTD which were tested as positive controls, immunostaining for p62 showed no pathological cytoplasmic or intranuclear inclusions in the cerebellar cortex of the patient with CANVAS. Examination of the brain, in addition to prominent cerebellar atrophy, revealed age-related changes in the form of neurofibrillary tangle tau pathology and amyloid-β pathology (**Supplementary Figure 3**).

Eight nerve biopsies and 10 muscle biopsies were also available for the assessment from patients carrying the homozygous AAGGG repeat expansion. In all nerve biopsies, there was prominent widespread depletion of myelinated fibres and the muscle biopsies confirmed chronic denervation with re-innervation (**Supplementary Figure 4**).

Fluorescence *in situ* hybridization using sense $(AAGGG)_5$ and anti-sense $(TTCCC)_5$ repeat specific oligonucleotides was performed on vermis post-mortem tissue from one CANVAS patients, disease and healthy controls. As opposed to SH-SY5Y cells transfected with pcDNA3.1/CT-GFP TOPO vector containing either $(TTCCC)_{94}$ or $(AAGGG)_{54}$ where intranuclear and cytoplasmic inclusion were clearly detectable, we did not observe the presence of endogenous RNA foci in any of the samples examined (**Supplementary Figure 5**).

RNA-sequencing

We performed whole transcriptome analysis in order to assess the presence of changes in *RFC1* expression, as well as *in-cis* and *in-trans* effects at more distant genomic regions. RNA-seq data showed that *RFC1* mRNA was unchanged in CANVAS (n=4) and control (n=4) fibroblasts (P = 0.42) and in CANVAS (n=2) and control (n=3) lymphoblasts (P = 0.45). We also performed RNA-seq from frontal cortex and cerebellar vermis from autopsied brains from one CANVAS patient, Friedreich's ataxia cases (n=3) and controls without evidence of neurological disease (n=3). In the single CANVAS patient, *RFC1* appears to be unchanged in both cortex and cerebellum compared to the other samples (**Figure 6A)**. However, frataxin gene (*FXN*) was clearly down regulated in Friedreich's ataxia frontal cortex and cerebellum compared to controls (cerebellum P = 0.007; log2 fold change = -1.2; frontal cortex P = 0.0003; log2 fold change = -1.3) (**Figure 6A)**. The single CANVAS sample resembled the controls for *FXN* expression.

There were no differentially expressed genes between patient and control fibroblasts, whereas 132 differentially expressed genes were identified between patient and

control lymphoblasts. Gene Ontology analysis showed enrichment for immune terms, whose relevance to the disease will warrant further work. Notably, only eight differentially expressed genes were located on chromosome 4 and were all separated by at least 25Mb from the locus of the repeat expansion. Analysis of differentially expressed genes in frontal cortex and vermis was not possible due to the limited numbers of CANVAS samples (n=1).

Splicing analysis was performed in lymphoblasts. We identified 145 exons in 108 genes that had evidence of differential exon usage in CANVAS patients compared to healthy controls. Motif analysis for the alternatively spliced exons showed enrichment of motifs targeted by SRSF proteins, and in particular of SRSF3. *RFC1* did not show aberrant splicing of its coding exons in mature mRNA. Also, no reads containing the AAGGG or TTCCC repeated unit mapping to intron 2 of *RFC1* pre-mRNA transcript were detected and no anti-sense or non-coding transcript was observed at *RFC1* locus in any of the tissues examined. Gene ontology analysis of alternatively spliced genes found enrichment for focal adhesion and non-specific cellular response terms. Lists of differentially expressed genes and differentially expressed exons in lymphoblasts, their normalised count values in brain samples and motif analysis for the alternatively spliced exons are provided in **Supplementary Table 4**.

RFC1 expression in patients' tissue

Quantitative reverse transcriptase PCR was performed using two sets of primers **(Figure 6B)** and, concordantly with RNA-seq data, did not show any significant decrease of *RFC1* mRNA (RefSeq NM_002913) level in patients' fibroblasts (n=5), lymphoblasts (n=2), muscle (n=6), frontal cortex and cerebellar vermis (n=1) compared to healthy controls or Friedreich's ataxia cases **(Figure 6C)**. Exon 2 and 3 were correctly spliced in the mature *RFC1* mRNA as shown by RNA-seq, qRT-PCR and sequencing. However, assessment of pre-mRNA expression by qRT-PCR showed a consistent increase of intron 2 retention (IR) in patients' lymphoblasts (n=2), muscle (n=6) (p=0.0077), cerebellar and frontal cortex (n=1) compared to healthy controls (**Supplementary Figure 6**). The low level of *RFC1* expression in fibroblasts prevented the assessment of pre-mRNA processing.

Western blot showed that *RFC1* protein (Uniprot P35251-1) was not decreased in patients' fibroblasts (n=5), lymphoblasts (n=4) or brain (n=1) compared to healthy controls or Friedreich's ataxia cases **(Figure 6D and Supplementary Figure 7)**. Assessment of *RFC1* protein expression in muscle could not be performed due to limited tissue availability.

Since *RFC1* play a key role in DNA damage recognition and recruitment of DNA repair enzymes, we assessed whether patients' derived fibroblasts have an impaired response to DNA damage. Patients' fibroblasts did not show an increased susceptibility to DNA-damage and their treatment with double-stranded DNA break inducing agents, UV and Methyl methanesulfonate, triggered a grossly normal response to DNA-damage (**Supplementary Figure 8**).

DISCUSSION

We identified a recessive repeat expansion in intron 2 of *RFC1* as the cause of CANVAS and late-onset ataxia. Twenty-three cases from 11 families and 33 sporadic cases carried the biallelic AAGGG repeat expansion. Notably, out of 150 cases from a single centre diagnosed with late-onset ataxia 22% resulted positive for the biallelic AAGGG repeat expansion and the percentage was higher if only patients with sensory neuronopathy and cerebellar involvement (62%), CANVAS disease (92%) and familial CANVAS disease (100%) were considered, highlighting that a higher diagnostic can be achieved in cases with well-defined clinical features and positive family history. Not since the discovery two decades ago of the most common genes casing ataxia (19–22) and Charcot-Marie-Tooth (CMT) genes, (23–26) has a novel gene explained percentages above 10% of genetically undetermined cases (27,28).

We determined that the allelic carrier frequency of the AAGGG repeat expansion in healthy controls was 0.7%, which is similar to the allelic carrier frequency of the GAA expansion in *FXN* gene ranging from 0.9 to 1.6%, and which in the biallelic state causes the most common recessive ataxia, Friedreich's ataxia. Together, this data suggests that the recessive AAGGG expansion in *RFC1* may represent a frequent cause of late onset ataxia in the general population, with an estimated prevalence at birth of the recessive trait of ~1/20,000.

The expansion resides at the 3'-end of a deep intronic AluSx3 element and it increases the polyA-tail size from 11 to over 400 repeated units, but also alters its sequence. Of interest, expansions in terminal and mid A stretches of Alu elements have been previously identified to cause Friedreich's ataxia (19), SCA37 (29), more recently benign adult familial myoclonic epilepsy (BAFME) (30) and now CANVAS and late-onset ataxia. Together, these observations suggest that variations and expansion of this highly polymorphic regions of Alu elements represent a common mechanism underlying different inherited neurological disorders. Notably, both SCA37 and BAFME are characterized by expansion of a mutated repeated unit, ATTTC and TTTCA, respectively (29,30). In this study, as well as in BAFME and SCA37, the presence in normal population of large expansions of the reference repeated unit suggests that the nucleotide change rather than the size of the expansion may be the driving pathogenic mechanism

Alu elements are repetitive elements about 300 base pairs long highly conserved within primate genomes. The 3'-end of an Alu element has a longer A-rich region that plays a critical role in its amplification mechanism (31). Active elements degrade rapidly on an evolutionary time scale by A-tail shortening or heterogeneous base interruptions accumulating in the A-tail, such as G insertions. We hypothesize that the mutation of the AAGGG repeated unit occurred as part of the inactivation process by G interruption of the polyA tail of the retrotransposon AluSx3. As known, repetitive DNA motives, particularly

G-rich regions, can form secondary or tertiary nucleotide structures such as hairpins, parallel and antiparallel G-quadruplexes and, if transcribed, DNA-RNA hybrids also known as R loops. These structures have been shown to increase the exposure of single-stranded DNA to damaging environmental agents and can initiate repeat expansion and perpetrate genomic instability across meiotic and mitotic divisions or after DNA damage (32).

Since the same ancestral haplotype is shared by the majority of familial and positive cases as well as some healthy carriers of two $(AAAGG)_{exp}$ alleles, we speculate that nucleotide change AAAAG to AAAGG or AAGGG may represent an ancestral founder event, which was followed by the pathologic expansion of the repeated unit, whose size seems to correlate positively with its GC content. However, the identification of two patients (fam 5b-2 and s23) with a recessive AAGGG repeat expansion who share only one allele of the common haplotype implies that repeat expansions of the mutated AAGGG unit can occur also on a different genetic background. Interestingly, fam 5b-2 was also found to carry the largest repeat expansion (10kb or 2,000 repeats) among the cohort of patients tested.

In the majority of the patients the expansion encompassed 1,000 repeats, but as low as 400 AAGGG repeats were shown to be sufficient to cause disease. The size of expanded alleles was relatively stable in siblings within single families, but no parent of the affected patients was available to assess whether this also applies across generations. We did not observe a correlation between age of onset of the neuropathy and size of the repeat expansion, although the disease course was very slowly progressive and initial symptoms might have been neglected in some patients but reported by others.

So far, approximately 40 neurological or neuromuscular genetic disorders have been associated with nucleotide repeat expansions. Two of them are known to be inherited in a recessive mode, namely Friedreich's ataxia and myoclonic epilepsy type 1, and are both associated loss-of-function of the repeat hosting gene (33–35).

A remarkable aspect of the recessive expansion described here is that our data does not suggest a direct mechanism of loss of function for the *RFC1* gene. We did not observe a reduced level of *RFC1* expression at either transcript or protein level in CANVAS patients, although as a known loss of function control we were able to detect a significant reduction of *FXN* transcript in post-mortem brain from patients with Friedreich's ataxia. Also, RNA-seq data did not show a clear effect on the expression of neighbouring or distant genes. We cannot exclude that the repeat expansion may cause more subtle tissue-specific alterations of *RFC1* transcript and protein or alter the structural organization of the chromatin.

RFC1 encodes the large subunit of replication factor C, a five subunit DNA polymerase accessory protein. It loads *PCNA* onto DNA and activates DNA polymerases

delta and epsilon to promote the coordinated synthesis of both strands during replication or after DNA damage (36–38). It is interesting to note that mutations in many of the genes involved in DNA repair have been already associated with degenerative neurological disorders, including ataxia-telangiectasia, xeroderma pigmentosum, Cockayne syndrome and ataxia oculomotor apraxia 1 and 2 (39). Interestingly, ataxia and neuropathy are common clinical features to all of them suggesting a particular susceptibility of cerebellum and peripheral nerves to DNA damage. However, our preliminary study did not show an impaired response to DNA damage in patients' derived fibroblasts.

In fact, late-onset Mendelian disorders represent a unique interpretative challenge, as risk variants may exert subtle effects rather than a clear loss of function of the mutated gene that are compatible with normal developmental until adult or old age (40). To this regard, although unusual in the context of a recessive mode of inheritance, other mechanisms, including the production of toxic RNA containing the expanded repeat, and the translation of a repeat-encoded polypeptide, should be considered (41). We did not observe in patients brain the presence of RNA foci of either the sense or anti-sense repeated unit. However, we were able to detect a consistent increase across different tissues of the retention of intron 2 in *RFC1* pre-mRNA. Retention of the repeat-hosting intron was recently identified as a common event associated with other disease-causing GC-rich intronic expansions, such as in myotonic dystrophy type 2 and *C9orf72*-ALS/FTD but not AT-rich repeat expansions such as in Friedreich's ataxia (42). Intron retention and abnormal pre-mRNA processing bear potential effects on nuclear retention and nucleocytoplasmic transport of the pre-mRNA, which, if efficiently exported to the cytoplasm, would be accessible to the translationary machinery.

Notwithstanding the enormous progress in Mendelian gene identification during the last decade, up to 40% of patients with ataxia and inherited neuropathy remain genetically undiagnosed and the percentage can rise up to 80-90% in particular subtypes, such as late-onset ataxia (2,5,43) and hereditary sensory neuropathies (27,28). Our paper, together with other studies from recent years (30,44–46), provides evidence that the combined used of whole-genome sequencing and classical genetic investigations such as linkage analysis, can provide a powerful tool to unravel a significant part of the missing heritability hidden in non-coding regions of the human genome

ACKNOWLEDGMENTS

AC is funded by the inherited neuropathy consortium, which is a part of the NIH Rare Diseases Clinical Research Network (RDCRN) (U54NS065712) and Wellcome Trust (204841/Z/16/Z). AMR is funded by a Wellcome Trust Postdoctoral Fellowship for Clinicians (110043/Z/15/Z). HH is also supported by Rosetrees Trust, Ataxia UK, The MSA Trust, Brain Research UK, MDUK, The Muscular Dystrophy Association (MDA), Higher Education Commission (HEC) of Pakistan and The Wellcome Trust (Synaptopathies Strategic Award). The INC (U54NS065712) is a part of the NCATS Rare Diseases Clinical Research Network (RDCRN). RDCRN is an initiative of the Office of Rare Diseases Research (ORDR), NCATS, funded through a collaboration between NCATS and the NINDS. Stephan Zuchner is thankful to the National Institute of Health (4R01NS075764) for its support. This research was also supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre (BRC). Neuromuscular and brain tissue samples were obtained from University College London Hospitals NHS Foundation Trust as part of the UK Brain Archive Information Network (BRAIN UK) which is funded by the Medical Research Council and Brain Tumour Research and the NIH funded NeuroBioBank.. We also thank Francesca Launchbury from UCL IQPath laboratory for the technical assistance in histology slide preparation.

AUTHORS CONTRIBUTION

AC designed the study, collected clinical data, performed the genetic analysis which led to the discovery of the AAGGG repeat expansions, analysed the data, drafted the manuscript together with contributions from JV, RS, RS, JH. RS, ASN, ET, EB, AR, YW, MI performed the investigation on *RFC1* expression. JV performed the computational genetic analysis; SR and HT collected and analysed the genetic data in healthy controls; PJT, WJM, AB, GD, IC, MV, DK, VS, SE, AMR contributed with collection of clinical data and patients' samples. HJ, SP, PF performed the RNA-seq analysis; ZJ performed the pathological investigation; RS, AMR, PF, JP contributed to the design of the study. SZ contributed to the design of the study and analysed the data. HH, MMR designed the study, collected patients' clinical data and biological samples and analysed the data. All authors revised the manuscript.

COMPETING INTEREST STATEMENT

The authors declare no competing financial interests.

REFERENCES

1. Harding AE. “Idiopathic” late onset cerebellar ataxia. A clinical and genetic study of 36 cases. J Neurol Sci. 1981 Aug;51(2):259–71.

2. Muzaimi MB, Thomas J, Palmer-Smith S, Rosser L, Harper PS, Wiles CM, et al. Population based study of late onset cerebellar ataxia in south east Wales. J Neurol Neurosurg Psychiatry. 2004 Aug;75(8):1129–34.

3. Sghirlanzoni A, Pareyson D, Lauria G. Sensory neuron diseases. Lancet Neurol. 2005 Jun;4(6):349–61.

4. Strupp M, Feil K, Dieterich M, Brandt T. Bilateral vestibulopathy. Handb Clin Neurol. 2016;137:235–40.

5. Abele M, Bürk K, Schöls L, Schwartz S, Besenthal I, Dichgans J, et al. The aetiology of sporadic adult-onset ataxia. Brain J Neurol. 2002 May;125(Pt 5):961–8.

6. Kirchner H, Kremmyda O, Hüfner K, Stephan T, Zingler V, Brandt T, et al. Clinical, electrophysiological, and MRI findings in patients with cerebellar ataxia and a bilaterally pathological head-impulse test. Ann N Y Acad Sci. 2011 Sep;1233:127–38.

7. Migliaccio AA, Halmagyi GM, McGarvie LA, Cremer PD. Cerebellar ataxia with bilateral vestibulopathy: description of a syndrome and its characteristic clinical sign. Brain J Neurol. 2004 Feb;127(Pt 2):280–93.

8. Szmulewicz DJ, Roberts L, McLean CA, MacDougall HG, Halmagyi GM, Storey E. Proposed diagnostic criteria for cerebellar ataxia with neuropathy and vestibular areflexia syndrome (CANVAS). Neurol Clin Pract. 2016 Feb;6(1):61–8.

9. Wu TY, Taylor JM, Kilfoyle DH, Smith AD, McGuinness BJ, Simpson MP, et al. Autonomic dysfunction is a major feature of cerebellar ataxia, neuropathy, vestibular areflexia “CANVAS” syndrome. Brain J Neurol. 2014 Oct;137(Pt 10):2649–56.

10. Szmulewicz DJ, Merchant SN, Halmagyi GM. Cerebellar ataxia with neuropathy and bilateral vestibular areflexia syndrome: a histopathologic case report. Otol Neurotol Off Publ Am Otol Soc Am Neurotol Soc Eur Acad Otol Neurotol. 2011 Oct;32(8):e63-65.

11. Szmulewicz DJ, McLean CA, Rodriguez ML, Chancellor AM, Mossman S, Lamont D, et al. Dorsal root ganglionopathy is responsible for the sensory impairment in CANVAS. Neurology. 2014 Apr 22;82(16):1410–5.

12. Cazzato D, Bella ED, Dacci P, Mariotti C, Lauria G. Cerebellar ataxia, neuropathy, and vestibular areflexia syndrome: a slowly progressive disorder with stereotypical presentation. J Neurol. 2016 Feb;263(2):245–9.

13. Rust H, Peters N, Allum JHJ, Wagner B, Honegger F, Baumann T. VEMPs in a patient with cerebellar ataxia, neuropathy and vestibular areflexia (CANVAS). J Neurol Sci. 2017 Jul 15;378:9–11.

14. Pelosi L, Leadbetter R, Mulroy E, Chancellor AM, Mossman S, Roxburgh R. Peripheral nerve ultrasound in cerebellar ataxia neuropathy vestibular areflexia syndrome (CANVAS). Muscle Nerve. 2017 Jul;56(1):160–2.

15. Pelosi L, Mulroy E, Leadbetter R, Kilfoyle D, Chancellor AM, Mossman S, et al. Peripheral nerves are pathologically small in cerebellar ataxia neuropathy vestibular areflexia syndrome: a controlled ultrasound study. Eur J Neurol. 2018 Apr;25(4):659–65.

16. Taki M, Nakamura T, Matsuura H, Hasegawa T, Sakaguchi H, Morita K, et al. Cerebellar ataxia with neuropathy and vestibular areflexia syndrome (CANVAS). Auris Nasus Larynx. 2018 Aug;45(4):866–70.

17. Infante J, García A, Serrano-Cárdenas KM, González-Aguado R, Gazulla J, de Lucas EM, et al. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome (CANVAS) with chronic cough and preserved muscle stretch reflexes: evidence for selective sparing of afferent Ia fibres. J Neurol. 2018 Jun;265(6):1454–62.

18. Hommelsheim CM, Frantzeskakis L, Huang M, Ülker B. PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications. Sci Rep. 2014 May 23;4:5052.

19. Campuzano V, Montermini L, Moltò MD, Pianese L, Cossée M, Cavalcanti F, et al. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science. 1996 Mar 8;271(5254):1423–7.

20. Orr HT, Chung MY, Banfi S, Kwiatkowski TJ, Servadio A, Beaudet AL, et al. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nat Genet. 1993 Jul;4(3):221–6.

21. Pulst SM, Nechiporuk A, Nechiporuk T, Gispert S, Chen XN, Lopes-Cendes I, et al. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nat Genet. 1996 Nov;14(3):269–76.

22. Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S, et al. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet. 1994 Nov;8(3):221–8.

23. Lupski JR, de Oca-Luna RM, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ, et al. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell. 1991 Jul 26;66(2):219–32.

24. Hayasaka K, Himoro M, Sato W, Takada G, Uyemura K, Shimizu N, et al. Charcot-Marie-Tooth neuropathy type 1B is associated with mutations of the myelin P0 gene. Nat Genet. 1993 Sep;5(1):31–4.

25. Bergoffen J, Scherer SS, Wang S, Scott MO, Bone LJ, Paul DL, et al. Connexin mutations in X-linked Charcot-Marie-Tooth disease. Science. 1993 Dec 24;262(5142):2039–42.

26. Züchner S, Mersiyanova IV, Muglia M, Bissar-Tadmouri N, Rochelle J, Dadali EL, et al. Mutations in the mitochondrial GTPase mitofusin 2 cause Charcot-Marie-Tooth neuropathy type 2A. Nat Genet. 2004 May;36(5):449–51.

27. Fridman V, Bundy B, Reilly MM, Pareyson D, Bacon C, Burns J, et al. CMT subtypes and disease burden in patients enrolled in the Inherited Neuropathies Consortium natural history study: a cross-sectional analysis. J Neurol Neurosurg Psychiatry. 2015 Aug;86(8):873–8.

28. Murphy SM, Laura M, Fawcett K, Pandraud A, Liu Y-T, Davidson GL, et al. Charcot-Marie-Tooth disease: frequency of genetic subtypes and guidelines for genetic testing. J Neurol Neurosurg Psychiatry. 2012 Jul;83(7):706–10.

29. Seixas AI, Loureiro JR, Costa C, Ordóñez-Ugalde A, Marcelino H, Oliveira CL, et al. A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of DAB1, Mapping to SCA37, Causes Spinocerebellar Ataxia. Am J Hum Genet. 2017 Jul 6;101(1):87–103.

30. Ishiura H, Doi K, Mitsui J, Yoshimura J, Matsukawa MK, Fujiyama A, et al. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Nat Genet. 2018 Apr;50(4):581–90.

31. Deininger P. Alu elements: know the SINEs. Genome Biol. 2011 Dec 28;12(12):236.

32. Haeusler AR, Donnelly CJ, Rothstein JD. The expanding biology of the C9orf72 nucleotide repeat expansion in neurodegenerative disease. Nat Rev Neurosci. 2016;17(6):383–95.

33. Dürr A, Cossee M, Agid Y, Campuzano V, Mignard C, Penet C, et al. Clinical and genetic abnormalities in patients with Friedreich's ataxia. N Engl J Med. 1996 Oct 17;335(16):1169–75.

34. Lazaropoulos M, Dong Y, Clark E, Greeley NR, Seyer LA, Brigatti KW, et al. Frataxin levels in peripheral tissue in Friedreich ataxia. Ann Clin Transl Neurol. 2015 Aug;2(8):831–42.

35. Paulson H. Repeat expansion diseases. Handb Clin Neurol. 2018;147:105–23.

36. Majka J, Burgers PMJ. The PCNA-RFC families of DNA clamps and clamp loaders. Prog Nucleic Acid Res Mol Biol. 2004;78:227–60.

37. Tomida J, Masuda Y, Hiroaki H, Ishikawa T, Song I, Tsurimoto T, et al. DNA damage-induced ubiquitylation of RFC2 subunit of replication factor C complex. J Biol Chem. 2008 Apr 4;283(14):9071–9.

38. Overmeer RM, Gourdin AM, Giglia-Mari A, Kool H, Houtsmuller AB, Siegal G, et al. Replication factor C recruits DNA polymerase delta to sites of nucleotide excision repair but is not required for PCNA recruitment. Mol Cell Biol. 2010 Oct;30(20):4828–39.

39. McKinnon PJ. Maintaining genome stability in the nervous system. Nat Neurosci. 2013 Nov;16(11):1523–9.

40. Higuchi Y, Hashiguchi A, Yuan J, Yoshimura A, Mitsui J, Ishiura H, et al. Mutations in MME cause an autosomal-recessive Charcot-Marie-Tooth disease type 2. Ann Neurol. 2016 Apr;79(4):659–72.

41. La Spada AR, Taylor JP. Repeat expansion disease: progress and puzzles in disease pathogenesis. Nat Rev Genet. 2010 Apr;11(4):247–58.

42. Sznajder ŁJ, Thomas JD, Carrell EM, Reid T, McFarland KN, Cleary JD, et al. Intron retention induced by microsatellite expansions as a disease biomarker. Proc Natl Acad Sci U S A. 2018 17;115(16):4234–9.

43. Gebus O, Montaut S, Monga B, Wirth T, Cheraud C, Alves Do Rego C, et al. Deciphering the causes of sporadic late-onset cerebellar ataxias: a prospective study with implications for diagnostic work. J Neurol. 2017 Jun;264(6):1118–26.

44. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron. 2011 Oct 20;72(2):245–56.

45. Renton AE, Majounie E, Waite A, Simón-Sánchez J, Rollinson S, Gibbs JR, et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron. 2011 Oct 20;72(2):257–68.

46. Aneichyk T, Hendriks WT, Yadav R, Shin D, Gao D, Vaine CA, et al. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. Cell. 2018 Feb 22;172(5):897-909.e21.

47. Manole A, Jaunmuktane Z, Hargreaves I, Ludtmann MHR, Salpietro V, Bello OD, et al. Clinical, pathological and functional characterization of riboflavin-responsive neuropathy. Brain J Neurol. 2017 01;140(11):2820–37.

48. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

49. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002 Jan;30(1):97–101.

50. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma Oxf Engl. 2009 Jul 15;25(14):1754–60.

51. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

52. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep;38(16):e164.

53. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.

54. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 18;536(7616):285–91.

55. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014 Jun 26;15(6):R84.

56. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24–6.

57. Weis J, Brandner S, Lammens M, Sommer C, Vallat J-M. Processing of nerve biopsies: a practical guide for neuropathologists. Clin Neuropathol. 2012 Feb;31(1):7–23.

58. Dubowitz V, Sewry C, Oldfors A. Muscle Biopsy—A Practical Approach , 4th edn. Elsevier Limited, Philadelphia. 4th ed. Philadelphia: Elsevier Limited; 2013.

59. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma Oxf Engl. 2013 Jan 1;29(1):15–21.

60. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinforma Oxf Engl. 2015 Jan 15;31(2):166–9.

61. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

62. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Res. 2012 Oct;22(10):2008–17.

63. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinforma Oxf Engl. 2016 01;32(19):3047–8.

64. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016 08;44(W1):W83-89.

65. Paz I, Kosti I, Ares M, Cline M, Mandel-Gutfreund Y. RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W361-367.

66. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic Acids Res. 2003 Jan 1;31(1):439–41.

67. Podhorecka M, Skladanowski A, Bozko P. H2AX Phosphorylation: Its Role in DNA Damage Response and Cancer Therapy. J Nucleic Acids. 2010 Aug 3;2010.

68. Sharma A, Singh K, Almasan A. Histone H2AX phosphorylation: a marker for DNA damage. Methods Mol Biol Clifton NJ. 2012;920:613–26.

FIGURES LEGENDS

Figure 1. a, Clinical spectrum of idiopathic late-onset ataxia from isolated cerebellar, vestibular and sensory variants to full-blown CANVAS. ILOCA: idiopathic late-onset cerebellar ataxia. CANVAS: cerebellar ataxia, neuropathy, vestibular areflexia syndrome. **b, Pedigrees of CANVAS families.** Squares indicate males and circles females. Diagonal lines are used for deceased individuals. CANVAS patients are indicated with filled symbols. Black dots indicate genotyped individuals. Red dots indicate patients enrolled for whole-genome sequencing study.

Figure 2. Identification of CANVAS locus. a, Non-parametric multipoint linkage analysis identifies a unique locus associated with the disease in chromosomal region 4p14 with maximal HLOD score of 5.8. **b,** Schematic representation of shared haplotypes within single families. Light blue bars indicate a genomic region shared by affected siblings in a family and for which unaffected siblings are discordant. Two red dashed lines define a 1.7 Mb region common to the different families. Single-nucleotide polymorphisms defining the haplotypes are represented on the top line. **c,** Fine-mapping inside the 1.7 Mb region identifies a recessive haplotype shared by all distinct families (green highlighted), except for individual fam 5b-2, who likely shares only one allele (light green highlighted). **d**, Schematic representation of the candidate 1.7 Mb region encompassing all 24 exons and flanking regions of Replication Factor Subunit C (*RFC1*) and the last exon and flanking intron of WD Repeat-Containing Protein 19 (*WDR19*).

Fig. 3 | A recessive expansion of a mutated AAGGG repeated unit in intron 2 of *RFC1* causes CANVAS and late-onset ataxia in familial and sporadic cases. a. A reduced read depth of whole genome sequencing is observed in CANVAS patients (n=6) in a region corresponding to a short tandem AAAAG repeat in intron 2 of *RFC1*. STR: short tandem repeat. **b**, Visualization on IGV of reads aligned to the short repeat and flanking region show in patients (n=6) the presence of a mutated AAGGG repeat unit (representative image). Reads from both sides are interrupted and are unable to cover the entire length of the microsatellite region. Note that, as per IGV default setting, AAGGG repeated units not mapping to the $(AAAAG)_{11}$ reference sequence are soft-clipped and do not contribute to the coverage of the STR in Figure 3A, which is virtually absent. However, ≥20 reads containing the AAGGG repeated unit could be observed in each patient if soft-clipped reads are shown. **c,** Repeat-primed PCR (RPPCR) targeting the mutated AAGGG repeated unit. FAM-labelled PCR products are separated on an ABI3730 DNA Analyzer. Electropherograms are visualized on GENEMAPPER at 2,000 relative fluorescence units. Representative plots from a patient carrying the AAGGG repeat expansion and one non-carrier are shown. RPPCR experiments were repeated independently at least twice with similar results. **d,** Sanger sequencing of long-range PCR reactions confirms in patients the AAAAG to AAGGG nucleotide change of the repeated unit.

Figure 4. Polymorphic configurations of the repeat expansion locus and allelic distribution in healthy controls

A. Schematic representation of the repeat expansion locus in intron 2 of Replication factor C subunit 1 and its main allelic variants. **B.** Estimated allelic frequencies in 608 chromosomes from 304 healthy controls. **C.** Average size and standard deviation of $(AAAGG)_{exp}$ and $(AAAGG)_{exp}$ expansions in healthy controls and $(AAAGG)_{exp}$ in CANVAS patients

Figure 5. Pathology of cerebellar degeneration in a patient with CANVAS carrying the recessive AAGGG repeat expansion

(**A-E**) Haematoxylin and Eosin (H&E) stained sections and (**A1-E1**) Sections immunostained for p62.

In a control brain (**A**), age-matched for the patient with CANVAS syndrome, there is well preserved density of Purkinje cells (yellow arrow) and also granule cell layer is densely populated with small neurocytes (green asterisk). In CANVAS syndrome (**B**) there is severe, widespread depletion of Purkinje cells with associated prominent Bergmann gliosis (blue arrow), whilst cell density in the granule cell layer is well preserved. In a patient with genetically confirmed Friedreich's ataxia (**C**), there is patchy depletion of Purkinje cells associated with Bergmann gliosis and unremarkable appearance of the granule cell layer. In a patient with genetically confirmed spinocerebellar ataxia 17 (SCA17) (**D**), there is widespread Purkinje cell loss with only occasional Purkinje cells remaining; also in this patient granule cell layer is densely populated with small neurocytes. In a patient with frontotemporal dementia due to *C9orf72* expansion (**E**), the Purkinje cell loss is patchy and granule cell layer is unremarkable. Immunostaining for p62 shows no pathological cytoplasmic or intranuclear inclusions in the cerebellar cortex in the control patient (**A1**), the patient with CANVAS syndrome (**B1**) and also in the patient with Friedreich's ataxia (**C1**). In SCA17 patient, there are scattered discrete intranuclear p62 immunoreactive inclusions in the small neurones within granule cell layer (**D1**; high-power view of a representative intranuclear inclusion is demonstrated in the inset within **D1**). In the patient with *C9orf72* expansion, there are frequent characteristic perinuclear p62 positive inclusions in the granule cell layer (**E1** and high-power view of a representative inclusion is shown in the inset within **E1**). Scale bar: 100μm in A-E, 30μm in A1-E1 and 5μm in insets in **D1** and **E1**.

Figure 6. *RFC1* expression is not affected by the AAGGG repeat expansion. A. Plots showing expression levels of *RFC1* and *FXN* as Fragments Per Kilobase Million (FPKM) in controls (Ctrl), patients with Friedreich's ataxia (FRDA) and one CANVAS patient. **B.** Mapping on *RFC1* transcript 1 of the primers used for assessment by qRT-PCR of *RFC1 mRNA* (cF1-cR1 and cF2-cR2) and pre-mRNA (cF1/iR1) expression. Blue arrows indicate primers mapping to exonic and intronic regions of canonical *RFC1* transcript. Primers spanning across exonic junctions are connected by dotted lines. A red triangle indicates the

site of the AAGGG repeat expansion. **C.** Expression levels of the canonical coding *RFC1* mRNA as measured by qRT-PCR using two separate set of primers cF1-cR1 and cF2-cR2. **D**. *RFC1* protein levels as measured by Western blotting using the polyclonal antibody (ab193559) and normalized to beta-actin in fibroblasts (FBs), lymphoblasts (LBLs), and post-mortem cerebellum (CBM) and frontal cortex (FCX) from patients with CANVAS compared to healthy controls (Ctrl) and Friedreich's Ataxia (FRDA) cases. Bar graphs show mean ± SD

Supplementary Table 4. Lists of differently expressed genes and exons from RNAseq experiments. Differential gene expression was assessed with DESeq2 (1.8.2) (**LBLs_DESeq_hits**) and differential splicing was assessed with DEXSeq (**LBLs_DEXSeq_hits**) running on R (3.3.2) (R project for statistical computing) between CANVAS (n=2) and control (n=3) lymphoblasts. The thresholds for significance for differential expression and splicing were set at a Benjamini-Hochberg false discovery rate of 10%**.** Motif analysis was conducted on 49 alternatively spliced exons in lymphoblasts identified by unambiguous sequences with known strand using RBPmap (**LBLs_RBP_motif_hit_counts**). Normalised count values in brain samples for previously identified differently expressed genes and exons in lymphoblasts are provided in **Normcount_Brain_LBLs_DESeq_hits** and **Normcount_Brain_LBLs_DEXSeqhit** tables, respectively. CANVAS cerebellar ataxia, neuropathy, vestibular areflexia syndrome, CBM cerebellum, Ctrl control, FCX frontal cortex, FDR false discovery rate, FRDA Friedreich's ataxia, LBLs lymphoblasts

TABLES

Table 1. Clinical features of patients with familial or sporadic late-onset ataxia carrying the recessive AAGGG repeat expansion in *RFC1*.

	Familial cases (N=23)	Sporadic cases (N=33)	All cases (N=56)	P-value
Male	12 (52%)	11 (52%)	27 (48%)	NS
Age of onset	53 ± 8	54 ± 10	54 ± 9	NS
Disease duration at examination	13 ± 9	10 ± 6	11 ± 7	NS
Sensory neuropathy	23 (100%)	33(100%)	56 (100%)	NS
Cerebellar syndrome	18 (78%)	27 (82%)	45 (80%)	NS
Bilateral vestibular impairment	17 (74%)	13 (39%)	30 (53%)	0.01
Dysautonomia	4 (17%)	9 (27%)	13 (23%)	NS
Cough	7 (30%)	14 (42%)	21 (37%)	NS
SAPs UL **Reduced** **Absent**	 6/21 (29%) 15/21 (71%)	 4/31 (13%) 27/31 (87%)	 10/46 (22%) 36/46 (78%)	NS
SAPs LL **Reduced** **absent**	 2/21 (10%) 19/21 (90%)	 1/31 (3%) 30/31 (97%)	 3/52 (6%) 49/52 (94%)	NS
Normal motor conduction	19/21 (90%)	26/31 (84%)	45/52 (87%)	NS
Cerebellar atrophy at CT/MRI scan	14/17 (82%)	21/25 (84%)	35/42 (83%)	NS
Full-blown CANVAS syndrome	15 (65%)	11 (33%)	26 (46%)	0.02

cMAP compound motor action potential, CT computed tomography, LL lower limbs, MRI magnetic resonance imaging, NS not significant, SAP sensory action potential, UL upper limbs.

METHODS

Patients

For the initial linkage study, we enrolled 29 individuals, 23 affected and 6 unaffected, from 11 families with a clinical diagnosis of CANVAS across four Centres: National Hospital for Neurology and Neurosurgery (London, UK), C. Mondino National Neurological Institute (Pavia, Italy), C. Besta Neurological Institute and Department of Neurology, School of Medicine (Ribeirão Preto, Brazil).

An additional 150 patients with sporadic CANVAS or late onset ataxia (onset after 35 years of age) were identified from the neurogenetic database of the National Hospital for Neurology and Neurosurgery (London, UK). For the experimental procedures, patients' samples are generally refereed at as CANVAS and no distinction between samples from patients with full-blown CANVAS or other more limited variants of late-onset ataxia is made. A skin biopsy was performed in five (fam 1-3, fam 2-2, fam 5a-2, fam 5b-2, fam 6b-1) genetically confirmed subjects and six age and gender matched controls. Fibroblast cultures were maintained according to standard procedures (47). Epstein-Barr virus–transformed lymphoblast cultures from two patients (fam 8-2, fam 11-2) were generated and maintained. Epstein-Barr virus–transformed lymphoblast cultures from three age and gender matched healthy controls were provided by the European Collection of Authenticated Cell Cultures (ECACC) (Salisbury, UK)

Paraffin-embedded and snap-frozen cerebellar (vermis) and frontal cortex from post-mortem brain from one sporadic CANVAS patient carrying the biallelic AAGGG repeat expansion (s16), three patients with genetically confirmed Friedreich's ataxia, one patient with genetically confirmed spinocerebellar ataxia 17, one patient with genetically confirmed *C9orf72*-related FTD and three neurologically healthy controls were obtained from the Queen Square Brain Bank for Neurological Disorders (London, UK).

Eight nerve biopsies and 10 muscle biopsies were obtained from patients carrying the homozygous AAGGG repeat expansion and healthy controls for pathological examination. Muscle biopsy tissue from six patients (fam 6b-1, s1, s2, s18, s19, s22) and five controls was also used for qRT-PCR.

The study was approved by the UCL Institute of Neurology Institutional Review Board and all subjects gave written informed consent to participate. The study has complied with all relevant ethical regulations.

SNP genotyping and linkage analysis

Genotype calls were generated by the UCL genomics genotyping facility using InfiniumCoreExome arrays (Illumina, Carlsbad, CA, USA). Raw data were processed and QC'ed using GenomeStudio (Illumina, Carlsbad, CA, USA). All individual passed the 99% call rate threshold and were included in the subsequent analysis using PLINK 1.9 software (48). Uninformative markers or markers with missing genotypes > 10% were removed and the resulting dataset was further pruned to remove markers in high linkage equilibrium. Finally, the dataset was thinned to include 1cM spaced markers covering all autosomes. In

total 3476 markers were included. For fine-mapping analyses all available informative markers were included.

Parametric linkage analysis was performed using MERLIN (49) assuming a highly penetrant recessive model of inheritance and disease allele frequency less than 1:10,000. MERLIN software was also used to obtain the most likely haplotypes in the candidate region. All genotyped individuals were included for haplotype analysis.

Single nucleotide polymorphisms rs11096992 and rs2066790 were genotyped in sporadic CANVAS patients and unaffected individuals by PCR followed by Sanger sequencing. Primers sequences, concentrations and PCR thermocycling conditions are provided in **Supplementary Table 3**

Whole Genome Sequencing

Whole Genome Sequencing was performed by deCODE genetics, Inc. (deCODE genetics, Reykjavik, Iceland). Paired-end sequencing reads (100bp) were generated using HiSeq4000 (Illumina, San Diego, CA, USA) and aligned to GRCH37 using Burrows-Wheeler Aligner (50). The mean coverage per sample was 35x. Variants were called according to the GATK UnifiedGenotyper (51) workflow and annotated using ANNOVAR (52). Variants were prioritised based on segregation, minor allele frequency (<0.0001 in the 1000 Genomes Project (53), NHLBI GO Exome Sequencing project (Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [September 2017], or gnomAD (54), evolutionary conservation and in-silico prediction of pathogenicity for coding variants. Copy number analysis was performed using LUMPY (55) with default parameters. The candidate region on chromosome 4 was also visually inspected for any copy number or structural variants using IGV (56).

Repeat-primed PCR

Repeat-primed PCR was performed in order to provide qualitative assessment of the presence of an expanded AAGGG repeat as well expansions of the reference AAAAG allele or the AAAGG variant. The repeat-primed PCR was designed such that the reverse primers binds at different points within the repeat expansion to produce multiple amplicons of incremental size. 25 to 27 nucleotides flanking the repeat were added in order to increase binding affinity of the reverse primer to the polymorphic (A/AA/-) 3'end of the microsatellite and flanking region and give preferential amplification of the larger PCR product, thus allowing sizing of the expansion in some cases. Primers sequences, concentrations and PCR thermocycling conditions are provided in **supplementary table 3**

Reverse primers were used in equimolar concentrations. Fragment length analysis was performed on an ABI 3730xl genetic analyser (Applied Biosystems, Foster City, CA, USA), and data were analysed using GeneMapper software. Expansions with a characteristic ''saw-tooth'' pattern were identified and put forward for Southern blotting where sufficient DNA allowed.

Southern Blotting

Five μg of gDNA was digested for three hours with EcoRI (10U) prior to electrophoresis. DNA was transferred to positively charged nylon membrane (Roche Applied Science) by capillary blotting and was crosslinked by exposure by ultraviolet light. Digoxygenin (DIG)-labelled probes were prepared by PCR amplification of a genomic fragment cloned into a pGEM®-T Easy Vector using PCR DIG Probe Synthesis Kit (Roche Applied Science). Primer pairs used for cloning of gDNA fragment and PCR amplification of digoxigenin-labelled probe and PCR conditions are shown in **Supplementary Table 3**. Filter hybridization was undertaken as recommended in the DIG Application Manual (Roche Applied Science) except for the supplementation of DIG Easy Hyb buffer with 100 mg/ml denatured fragmented salmon sperm DNA. After prehybridization at 46°C for three hours, hybridization was allowed to proceed at 46°C overnight. A total of 600 μl of PCR products containing the labelled oligonucleotide probe was used in 30 ml of hybridization solution. Membranes were washed initially in 23 standard sodium citrate (SSC) and 0.1% sodium dodecyl sulfate (SDS), while the oven was being ramped from 48°C to 65°C and then washed three times in fresh solution at 65°C for 15 min. Detection of the hybridized probe DNA was carried out as recommended in the DIG Application Manual with CSPD ready-to-use (Roche Applied Science) as a chemiluminescent substrate. Signals were visualized on Fluorescent Detection Film (Roche Applied Science) after 1 hr. All samples were electrophoresed against DIG-labelled DNA molecular-weight markers II and III (Roche Applied Science). Pentanucleotide repeat number was estimated after subtraction of the wild-type allele fragment size (5,037 bp). Sizes of the detected bands were recorded for each individual and number of expanded repeated unit was estimated using the formula repeated pentanucleotides unit = (size of the expanded band in bp – 5000 bp)/5.

Neuropathological examination

The formalin fixed cerebellar tissue was embedded in paraffin wax, from which 5μm thick sections were cut for routine haematoxylin and eosin staining and immunohistochemistry. The sections were immunostained for p62 (Abcam, ab56416, 1:500), TDP43 (Novus Biologicals, 2E2-D3, 1:500), α-synuclein (Abcam, 4D6, 1:1000), phospho-Tau (AT-8, Innogenetics, 1:100) and anti βA4 (DAKO 6F3D, 1:50). Immunostaining, together with appropriate controls, was performed on a Roche Ventana Discovery automated staining platform following the manufacturer's guidelines, using biotinylated secondary antibodies and streptavidin-conjugated horseradish peroxidase and diaminobenzidine as the chromogen. Assessment of neuronal density in the cerebellar cortex was performed semi-quantitatively. Nerve and muscle biopsy specimens were performed and analysed according to standard procedures (57,58). In brief, all nerve biopsies were examined after processing for paraffin histology (immunostaining for neurofilaments was performed with SMI31 antibody (Sternberger, 1:5000) and in resin blocks (semithin resin sections were stained with methylene blue azure – basic fuchsin). The muscle biopsies were examined with routine histochemical stains after freezing in isopentane cooled in liquid nitrogen.

qRT-PCR

Total RNA was extracted from fibroblasts, lymphoblasts and brain regions using 1 ml of Qiazol (Qiagen) and 200 µl chloroform. Aqueous phase was loaded and purified on columns using the RNeasy Lipid Tissue Mini kit (Qiagen) and treated with RNAse-free DNAse I (Qiagen). cDNA was synthesized using 500 ng of total RNA for all samples, with a Superscript III first strand cDNA synthesis kit (Invitrogen) and an equimolar mixture of oligo dT and random hexamer primers. Real-time qRT–PCR was carried out using Power SYBR Green Master Mix (Applied Biosystems) and measured using a QuantStudio 7 Flex Real-Time PCR platform (Applied Biosystems). Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) was used as housekeeping gene to normalize across different samples. Amplified transcripts were quantified using the comparative Ct method and presented as normalized fold expression change ($2^{-\Delta\Delta Ct}$). Oligonucleotide sequences and thermocycling conditions are provided in **Supplementary Table 3.**

Western blotting

Cells and tissues were lysed in radioimmunoprecipitation assay (RIPA) buffer supplemented with complete EDTA-free protease inhibitor cocktail (Roche). Brain lysates were homogenised on ice using a tissue ruptor with disposable probes (Qiagen). Protein lysate concentrations were measured by the BCA protein assay (Bio-Rad). After adding 5µl of sample buffer (Bio-Rad) and 2µl of NuPAGE reducing agent (Invitrogen) and boiling at 95 °C for 5 min, 15-30 µg proteins for each sample were separated on 4-12% SDS–polyacrylamide gel (Bio-Rad) in MES buffer and transferred onto nitrocellulose membranes (GE-Healthcare) using a Turbo Transfer Pack (Bio-Rad). After blocking in 5% milk, immunoblotting was performed incubating over night at 4°C with the following primary antibodies: anti-RFC1 (GTX129291, GeneTex 1:1000), anti- β-actin (A2228, Sigma, 1:2000). Secondary antibodies were as follows: IRDye-800CW or IRDye-680CW conjugated goat anti-rabbit, donkey anti-mouse, IgG (Li-COR Bioscience). Signals of *RFC1* bands were normalized to those of the corresponding β-actin bands as internal controls. Signals were digitally acquired by using an Odyssey Fc infrared scanner (Li-COR Bioscience) and quantified using Image Studio software (Li-COR Bioscience).

RNA-sequencing

Reads were aligned to the hg38 human genome build using STAR (2.4.2a) (59). BAM files were sorted, and duplicate reads flagged using NovoSort (1.03.09) (Novocraft). The aligned reads overlapping human exons (Ensembl 82) were counted using HTSeq (0.1) (60) . For each gene and each sample, the fragments per kilobase of exon per million mapped reads (FPKM) was calculated. Any gene with a mean FPKM across all samples in a dataset < 1 was discarded from further analysis. Differential gene expression was assessed with DESeq2 (1.8.2) (61) and differential splicing was assessed with DEXSeq (62), running on R (3.3.2) (R project for statistical computing). The thresholds for significance for differential expression and splicing were set at a Benjamini-Hochberg false discovery

rate of 10%. Quality control reports were collated using MultiQC (63). Gene ontology enrichment testing was done using g:Profiler (64) with both GO and KEGG ontologies, with minimum term size of 5 genes and all p-values Bonferroni corrected for multiple testing. Motif analysis was conducted on 49 alternatively spliced exons in lymphoblasts identified by unambiguous sequences with known strand using RBPmap (65). Prediction of non-coding RNAs sequences in intron 2 of *RFC1* was tested by Rfam (66)

Statistical analyses

Clinical variables were compared between familial and sporadic cases with two-tailed Student's t test (continuous variables) and Chi^2 (categorical variables). Correlation between repeat expansion size and age of onset of neuropathy was calculated using Pearson's correlation coefficient. FPKM of *FXN* and *RFC1* was compared using the two-tailed Student's t test. The relative expression of *RFC1* transcript 1 versus *GAPDH* as measured by qRT-PCR was compared with two-tailed Student's t test. Statistical analysis of the results of the western blot analysis was performed with two-tailed Student's t test after confirmation of equality of variances. P values of < 0.05 were considered to be significant.

Cloning of *RFC1* repeat expansion locus

The *RFC1* locus containing the AAGGG repeat expansion was amplified by long-range PCR from genomic DNA from a CANVAS patient carrying the homozygous AAGGG repeat expansion and a healthy control carrying two $(AAAAG)_{11}$ alleles. PCR products were cloned into the pcDNA3.1/TOPO vector (Invitrogen) according to manufacturer's instructions. Primers and thermocycling conditions are provided in **Supplementary Table 3**. The size of the insert was determined by digestion with BstXI. Integrity of repeats and their orientation was confirmed by DNA sequencing (Eurofins Genomics, Louisville, KY, USA), which revealed uninterrupted 94x (CCCTT) and 54x (AAGGG) repeats in mutant clones, as well as 11x (CTTTT) and 11x (AAAAG) repeat sequences in wild-type clone. Once confirmed, the four clones used for experimental procedures were amplified using a maxi-prep plasmid purification system.

RNA *in situ* hybridization

Paraffin-embedded formalin-fixed post-mortem Vermis sections from a CANVAS case, 2 healthy and 2 cerebellar degeneration age-matched controls were deparaffinized in xylene twice for 10min, then rehydrated in 100%, 90% and 70% ethanol, then in phosphate-buffered saline (PBS). About 10^5 SH-SY5Y cells were seeded on coverslips in 24-well plates and transfected using lipofectamine 3000 (Invitrogen) with plasmids expressing wild-type sense $(TTTTC)_{11}$, wild-type anti-sense $(AAAAG)_{11}$, mutant sense $(TTCCC)_{94}$ or mutant anti-sense $(AAGGG)_{54}$ repeat sequences and were analyzed after 24 hours. Cells were fixed in 4% methanol-free paraformaldehyde (Pierce) for 10 min at room temperature, dehydrated in a graded series of alcohols, air dried and rehydrated

in PBS, permeabilised for 10 min in 0.1% Triton X100 in PBS, briefly washed in 2×SSC and incubated for 30 min in pre-hybridisation solution (40 % formamide, 2×SSC, 1 mg/ml tRNA, 1 mg/ml salmon sperm DNA, 0.2 % BSA, 10 % dextran sulphate, 2 mM ribonucleoside vanadyl complex) at 57 °C. Hybridisation solution (40 % formamide, 2×SSC, 1mg/ml tRNA, 1 mg/ml salmon sperm DNA, 0.2 % BSA, 10 % dextran sulphate, 2 mM ribonucleoside vanadyl complex, 0.2 ng/μl $(AAGGG)_5$ or $(CCCTT)_5$ LNA probe, 5′ TYE563-labeled (Exiqon), was heated at 95 °C for 10 min prior to incubation with sections for 1 h at 57 °C. Cells were washed for 30 min at 57 °C with high-stringency buffer (2x SSC, 0.2% Triton X100, 40% formamide) and then for 20 min each, in 0.2x SSC buffer. Nuclei were stained by DAPI. Coverslips were then dehydrated in 70% then 100% EtOH and mounted onto slides in Vectashield mounting medium. Images were acquired using an LSM710 confocal microscope (Zeiss) using a plan-apochromat 63x oil immersion objective.

Response to DNA damage

Fibroblasts were grown in 10 cm dishes in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum. Asynchronous cell cultures were grown to approximately 80% confluency and treated with UV, methyl methanesulfonate or untreated. For UV irradiation, cells were washed with PBS, and exposed to 30 or 120 J/m^2 UV light (254 nm) using a Stratalinker UV crosslinker®. For genotoxin treatment, methyl methanesulfonate (Sigma-Aldrich, St. Louis, MO, USA) was added to the culture media to give a final concentration of 1mM and cells were exposed for 8 hours. After UV irradiation or genotoxin treatment cells were allowed to recover for 24 hours prior to analysis.

Cells were homogenized in RIPA Buffer containing 50 mM Tris pH 7.4, 150 mM NaCl, 1% Triton X-100, 0.5% Na deoxycholate, 0.1% SDS, 1 mM EDTA, and protease inhibitor. Samples were sonicated and centrifuged before protein levels were quantified using a BCA assay (Thermo Fisher Scientific Pierce, Rockford, IL, USA). For Western blot analysis, protein (5 μg) was size separated by SDS-PAGE, transferred to nitrocellulose membranes, and subjected to standard immunoblotting procedures using the following antibodies: γH2AX (Abcam, USA; 1:1000), β-Actin (Sigma-Aldrich, St. Louis, MO, USA; 1:1000). γH2AX has been extensively used as a marker for DNA double strand breaks (DSBs). It is one of the initial markers of DSB being common to all DNA repair pathways. Secondary HRP-conjugated antibodies were purchased from PorteinTech and used at a 1:2000 concentration. Antibody staining was detected by ECL (Thermo Fisher Scientific Pierce, Rockford, IL, USA) and visualized by X-ray film.

Cell viability was assessed using CellTiter-Glo® Luminescent Cell Viability Assay following manufacturers protocol. For cell-viability assessment, 20,000cells/well were seeded in 96-well plates prior to treatment and treated as previously described.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability

The genotyping microarray data and sequence data obtained by whole-genome sequencing and RNA sequencing are available on request from the corresponding authors (A.C.; H.H.). They are not publicly available because some of the study participants did not give full consent for releasing data publicly. Since whole-genome sequence data are protected by the Personal Information Protection Law, availability of these data is under the regulation by the institutional review board. The data obtained RNA sequencing have been deposited on SRA under accession number SUB5043763.