Normal and pathogenic variation of *RFC1* repeat expansions: implications for clinical diagnosis

Natalia Dominik,[1,†] Stefania Magri,[2,†] Riccardo Currò,[1,3] Elena Abati,[1,4] Stefano Facchini,[1,5] Marinella Corbetta,[2] Hannah MacPherson,[1] Daniela Di Bella,[2] Elisa Sarto,[2] Igor Stevanovski,[6,7] Sanjog R. Chintalaphani,[7] Fulya Akcimen,[8] Arianna Manini,[1,4,9] Elisa Vegezzi,[5] Ilaria Quartesan,[3] Kylie-Ann Montgomery,[1] Valentina Pirota,[10,11,12] Emmanuele Crespan,[10] Cecilia Perini,[10] Glenda Paola Grupelli,[10] Pedro J. Tomaselli,[13] Wilson Marques,[13] Genomics England Research Consortium, Joseph Shaw,[1] James Polke,[1] Ettore Salsano,[14] Silvia Fenu,[14] Davide Pareyson,[14] Chiara Pisciotta,[14] George K. Tofaris,[15] Andrea H. Nemeth,[15,16] John Ealing,[17] Aleksandar Radunovic,[18] Seamus Kearney,[19] Kishore R. Kumar,[20,21,22] Steve Vucic,[22,23] Marina Kennerson,[21,24,25] Mary M. Reilly,[1] Henry Houlden,[1] Ira Deveson,[6] Arianna Tucci,[1] Franco Taroni[2] and Andrea Cortese[1,3]

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

Abstract

Cerebellar Ataxia, Neuropathy and Vestibular Areflexia Syndrome (CANVAS) is an autosomal recessive neurodegenerative disease, usually caused by biallelic AAGGG repeat expansions in *RFC1*. In this study, we leveraged whole genome sequencing (WGS) data from nearly 10,000 individuals recruited within the Genomics England sequencing project to investigate the normal and pathogenic variation of the *RFC1* repeat.

We identified three novel repeat motifs, AGGGC (n=6 from 5 families), AAGGC (n=2 from 1 family), AGAGG (n=1), associated with CANVAS in the homozygous or compound heterozygous state with the common pathogenic AAGGG expansion. While AAAAG, AAAGGG and AAGAG expansions appear to be benign, here we show a pathogenic role for large AAAGG repeat configuration expansions (n=5). Long read sequencing was used to fully characterise the entire repeat sequence and revealed a pure AGGGC expansion in six patients, whereas the other patients presented complex motifs with AAGGG or AAAGG interruptions. All pathogenic motifs seem to have arisen from a common haplotype and are predicted to form highly stable G quadruplexes, which have been previously demonstrated to affect gene transcription in other conditions.

The assessment of these novel configurations is warranted in CANVAS patients with negative or inconclusive genetic testing. Particular attention should be paid to carriers of compound AAGGG/AAAGG expansions, since the AAAGG motif when very large (>500 repeats) or in the presence of AAGGG interruptions.

Accurate sizing and full sequencing of the satellite repeat with long read is recommended in clinically selected cases, in order to achieve an accurate molecular diagnosis and counsel patients and their families.

Author affiliations:

1 Department of Neuromuscular Diseases, University College London, WC1N 3BG London, UK

2 Unit of Medical Genetics and Neurogenetics, Fondazione IRCCS Istituto Neurologico Carlo Besta, 20133, Milan, Italy

3 Department of Brain and Behavioral Sciences, University of Pavia, 27100, Pavia, Italy

4 Department of Pathophysiology and Transplantation, University of Milan, 20122, Milan, Italy;

5 IRCCS Mondino Foundation, 27100, Pavia, Italy

6 Genomics Pillar, Garvan Institute of Medical Research, Sydney, 2010, Australia.

7 Centre for Population Genomics, Garvan Institute of Medical Research and Murdoch Children's Research Institute, 2010, Australia

8 Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, 2292, USA

9 Department of Neurology and Laboratory of Neuroscience, IRCCS Istituto Auxologico Italiano, 20145 Milan, Italy

10 Institute of Molecular Genetics IGM-CNR "Luigi Luca Cavalli-Sforza", 27100 Pavia, Italy

11 Department of Chemistry, University of Pavia, 27100 Pavia, Italy

12 G4-INTERACT, USERN, Pavia, Italy

13 Department of Neurology, School of Medicine of Ribeirão Preto, University of São Paulo, 2650 Ribeirão Preto, Brazil

14 Clinic of Central and Peripheral Degenerative Neuropathies Unit, IRCCS Foundation, C. Besta Neurological Institute, 20126 Milan, Italy

15 Nuffield Department of Clinical Neurosciences, University of Oxford, OX3 9DU UK

16 Oxford Centre for Genomic Medicine, Oxford University Hospitals NHS Foundation Trust, OX3 7HE, UK

17 Salford Royal NHS Foundation Trust Greater Manchester Neuroscience Centre, Manchester Centre for Clinical Neurosciences Salford, Greater Manchester, M6 8HD, UK

18 Barts MND Centre, Royal London Hospital, Whitechapel, London, UK, E1 1BB

19 Department of Neurology, Royal Victoria Hospital, Belfast, BT12 6BA, UK

20 Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Darlinghurst, NSW, 2010, Australia

21 Molecular Medicine Laboratory, Concord Hospital, Concord, NSW, 2139, Australia

22 Concord Clinical School, Faculty of Medicine and Health, University of Sydney, Sydney, NSW, Australia

23 Brain and Nerve Research Centre, Concord Hospital, Sydney, NSW, 2139, Australia

24 Northcott Neuroscience Laboratory, ANZAC Research Institute SLHD, Sydney, NSW, 2050, Australia

25 School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Sydney, NSW, 2050, Australia

Correspondence to: Andrea Cortese, MD PhD

Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, Queen Square, WC1N 3BG, London, UK

E-mail: andrea.cortese@ucl.ac.uk

Running title: Normal and pathogenic variation of *RFC1*

Keywords: *RFC1*; CANVAS; ataxia; neuropathy; repeat expansions; long-read sequencing

Abbreviations: CANVAS = Cerebellar Ataxia, Neuropathy and Vestibular Areflexia Syndrome; CCS = Circular Consensus maps; DIG = digoxigenin granules solution; DLS = Direct Label and Stain; EHDN = ExpansionHunterDeNovo; GEL = Genomics England; HMW = high molecular weight; NHS = National Health Service; nt = nucleotides; OGM = Optical genome mapping; PacBio = Pacific Biosciences; QGRS = Quadruplex forming G-Rich Sequences; RFC1 = Replication Factor Complex subunit 1; STR = short-tandem repeat; WGS =whole genome sequencing

Introduction

Cerebellar Ataxia, Neuropathy and Vestibular Areflexia Syndrome (CANVAS) is an autosomal recessive neurodegenerative disease characterized by adult onset and slowly progressive ataxia caused by the concurrent impairment of sensory neurons, the vestibular system and the cerebellum. In most cases, the disease is caused by biallelic AAGGG repeat expansions in the second intron of the Replication Factor Complex subunit 1 (*RFC1*) gene.[1-19] Moreover, additional pathogenic $(AAAGG)_{10\text{-}25}(AAGGG)_n$ and ACAGG configurations were identified in people from Oceania and East Asia suggesting the possibility of genetic heterogeneity, at the repeat locus, underlying this condition.[20-23]

In this study we leveraged whole genome sequencing (WGS) data from the 100,000 Genomes Project to investigate the normal and pathogenic variation of the *RFC1* repeat as well as identifying additional pathogenic motifs causing CANVAS, that were further analysed by targeted long read sequencing.

We identified three novel pathogenic repeat configurations, AAGGC, AGGGC, and AGAGG, either in the homozygous or compound heterozygous state with AAGGG repeat, showing a similar or larger size compared to the common AAGGG expansion. In addition, pathogenic uninterrupted or interrupted AAAGG expansions were identified, with significantly larger sizes compared to the more frequent non-pathogenic AAAGG repeat.

Materials and methods

Whole Genome Sequencing data analysis

The 100,000 Genomes Project, run by Genomics England (GEL), was established to sequence whole genomes of patients of the National Health Service (NHS) of the United Kingdom, affected by rare diseases and cancer.[24] In this study, we leveraged GEL WGS data and screened for the presence of pentanucleotide expansions in *RFC1* in 893 samples from patients diagnosed with ataxia and 8107 controls, all aged 30 years or older. Repeat expansions were detected using ExpansionHunterDeNovo (EHDN) v0.9.0. We considered all motifs composed of 5 or 6 nucleotides occurring at the *RFC1* locus. Repeat motifs which were present, in the homozygous or compound heterozygous state with the AAGGG expansion in ataxia cases, but absent or significantly less frequent in controls, were considered to be possibly pathogenic and were further assessed as described below.

Structural variants were detected using Manta25 as described in https://re-docs.genomicsengland.co.uk/genomic_data/.

Predicted genetic ancestries for samples in GEL are based on PCA analysis, using as reference populations the five macro-ethnicities of 1000 Genomes project (European, African, South Asian, East Asian, American). Samples where none of the components reaches 95% are classified as Mixed.

RP-PCR

Samples identified to carry novel pathogenic repeat motifs by EHDN were tested by RP-PCR. In addition, we screened a cohort of 540 samples with genetically confirmed RFC1 CANVAS, as defined by the presence of a positive RP-PCR for the AAGGG expansion and the absence of an amplifiable PCR product from the flanking PCR, looking for expansions of different repeat motifs on the second allele. RP-PCR for AAAAG, AAAGG and AAGGG expansions were performed as previously described.[1] Moreover, the following primers were used: AGGGC-Rv: 5'-CAGGAAACAGCTATGACCAACAGAGCAAGACTCTGTTTCAAAAAGGGCAGGGCAGGGCAGGGCA-3'; AAGGC -Rv; 5'-AAGGC: CAGGAAACAGCTATGACCAACAGAGCAAGACTCTGTTTCAAAAA GGCAAGGCAAGGCAA-3'; or AGAGG-Rv: 5'-CAGGAAACAGCTATGACCAACAGAGCAAGACTCTGTTTCAAAAAGGAGAGGAGAGGAGAGGAGA-3', depending on the configuration tested. The PCR conditions for AGGGC and AAGGC were modified to 30s denaturation per cycle as opposed to 10s for all the other configurations.

Southern Blotting

Briefly, 5µg of high molecular weight (HMW) DNA was enzymatically digested with EcoRI for 3 h and size fractionated on a 1.2% agarose gel for 15 hours. The gel was washed in depurination, denaturing and neutralising solutions for 45 minutes each, after which the blot was assembled to transfer DNA from the gel onto a positively charged membrane using an upward transfer method for 15 h. The DNA was UV crosslinked to the membrane and hybridised with a mixture of salmon sperm and *RFC1* probe in digoxigenin granules solution (DIG) (Roche) overnight. The membrane was then washed, blocked and anti-DIG antibody was added after which detection buffer and a chemiluminescent CDP-STAR substrate (Roche) was used to visualize hybridization fragments.

Targeted RFC1 long read sequencing

We performed long read sequencing to establish the precise repeat sequence in patients carrying novel likely pathogenic expansion of *RFC1*. Given the technical hurdle in sequencing large repeat expansions, samples were sequenced on different platforms, including Oxford Nanopore and Pacific Biosciences (PacBio) and target enrichment was performed with either clustered regularly interspaced short palindromic repeats CRISPR/CRISPR-associated protein-9 nuclease (Cas9) system or ReadUntil programmable selective sequencing.

Samples were extracted from blood using the Qiagen MagAttract HMW DNA Kit and quality was checked using readouts from a Thermo Scientific NanoDrop. For CRISPR/Cas9 targeted sequencing, fragment lengths were assessed using the Agilent Femto Pulse Genomic DNA 165kb kit and only samples with the majority of fragments over 25kb were used., Libraries were prepared from 5µg of input DNA for each sample for both the Pacific Biosciences No-Amp Targeted Sequencing Utilizing the CRISPR-Cas9 System protocol (Version 09) and the Oxford Nanopore Ligation sequencing gDNA Cas9 enrichment (SQK-LSK109) protocol (Version: ENR_9084_v109_revT_04Dec2018). Libraries were sequenced on the Oxford Nanopore PromethION or MinION platforms or the Pacific Biosciences (PacBio) Sequel IIe, respectively. For the Oxford Nanopore Ligation sequencing gDNA Cas9 enrichment, we used four CRISPR-Cas9 guides from Nakamura et al. [22], which were RFC1-F1: 5′-GACAGTAACTGTACCACAATGGG-3′, RFC1-R1: 5′-

CTATATTCGTGGAACTATCTTGG-3′, RFC1-F2: 5′-ACACTCTTTGAAGGAATAACAGG-3′ and RFC1-R2: 5′-TGAGGTATGAAT CATCCTGAGGG-3′, except for cases IV-1, XI-1, XII-1, for which only the two guides RFC1-F2 and RFC1-R2 were used. The guides RFC1-F3: 5′-GAAACTAAATAGAACCAGCC-3′ RFC1-R3: 5′-GACTATGGCTTACCTGAGTG-3′, which were designed in house, were used for Pacific Biosciences No-Amp Targeted Sequencing and up to 10 samples were multiplexed using PacBio barcoded adapters. Libraries loaded onto the PromethION and MinION were run for 72 hours with standard loading protocols. Sequel IIe libraries were run for a movie time of 30 hours with an immobilisation time of 4 hours. All libraries were loaded neat.

Programmable targeted sequencing was performed as described previously.[26] HMW DNA was sheared to ~20kb fragment size using Covaris G-tubes. Sequencing libraries were prepared from ~3-5ug of HMW DNA, using native library prep kit SQK-LSK110, according to the manufacturer's instructions. Each library was loaded onto a FLO-MIN106D (R9.4.1) flow cell and run on an ONT MinION device with live target selection/rejection executed by the *ReadFish* software package.[27] Detailed descriptions of software and hardware configurations used for ReadFish experiments are provided in a recent publication that demonstrates the suitability of this approach for profiling tandem repeats.[26] The target used in this study was the *RFC1* gene locus ±50 kb. Samples were run for a maximum duration of 72 h, with nuclease flushes and library reloading performed at approximately 24 and 48 h timepoints for targeted sequencing runs, to maximise sequencing yield.

Bioinformatic analysis

Alignment to the hg38 reference of Nanopore reads, PacBio CCS and PacBio subreads was done using minimap2[28] with additional options "-r 10000 -g 20000 -E 4,0". For PacBio sequences, the recommended step of generating Circular Consensus maps (CCS) from subreads was not always possible because of low depth of the sequencing data. The only CCS we could obtain was for the AAGGG allele of Case V-1.After alignment, we used PacBio scripts [https://github.com/PacificBiosciences/apps-scripts] to extract the repeat region (extractRegions.py) and obtain waterfall plots (waterfall.py) for the following motifs: AAGGG, AGAGG, AGGGC, AAGGC and AAAGG.

For programmable targeted sequencing, raw ONT sequencing data was converted to BLOW5 format using *slow5tools* (v0.3.0)[29] then base-called using *Guppy* (v6). Resulting FASTQ files

were aligned to the *hg38* reference genome using *minimap2* (v2.14-r883). The short-tandem repeat (STR) site within RFC1 locus was genotyped using a process validated in our recent manuscript.[27] This method involves local haplotype-aware assembly of ONT reads spanning a given STR site and annotation of STR size, motif and other summary statistics using Tandem Repeats Finder (4.09), followed by manual inspection and motif counting.

Haplotype analysis

We used SHAPEITv4[30] with default parameters to phase a 2Mb region (chr4:38020000-40550000) encompassing the *RFC1* gene. To maximise available haplotype information, the entire Rare Diseases panel in Genomics England (78195 samples from patients affected by rare diseases) were jointly phased. Input data format was an aggregate VCF file with a total of 551795 variants.

The estimation of the haplotype age is based on the online application *Genetic Mutation Age Estimator* (https://shiny.wehi.edu.au/rafehi.h/mutation-dating/) [31]. The method requires as input a list of ancestral segments for sampled individuals. We used the five individuals with pathogenic expansions (see Figure 3): AAGGG hom, ACAGG hom, case VII-1, case I-1, case III-3.

Optical genome mapping (OGM)

Patients for whom whole blood was available were subjected to BioNano optical genome mapping to gather additional information on the precise size of the expanded repeat. Ultra HMW genomic DNA was isolated as described by the Bionano Prep SP Frozen Human Blood DNA Isolation Protocol v2. Homogeneous ultra HMW DNA was labelled using the Bionano Prep Direct Label and Stain (DLS) Protocol provided with the kit and the homogeneous labelled DNA was loaded onto a Saphyr chip. Optical mapping was performed at theoretical coverage of 400x. Molecule files (.bnx) were aligned to hg38 with Bionano Solve script "align_bnx_to_cmap.py" from Bionano Solve v3.6 (https://bionano.com/software-downloads/) using standard parameters. For each sample, molecules overlapping both markers flanking the repeat expansion were extracted (Marker IDs: 7723 and 7724). Intermarker distances were analysed by decomposing into two Gaussian components and using the gaussian

mean as allele size and repeat expansion size was calculated as the difference between the gaussian mean and the intermarker distance of a non-expanded allele (6858 bp).

G-quadruplexes analysis

The propensity of the different repeat configurations in *RFC1* to form G-quadruplexes (G4)[32] was predicted using the Quadruplex forming G-Rich Sequences (QGRS) Mapper[33] and G4-Hunter software[34], through which the likelihood to form a stable G4 is rated in terms of G-score values. Putative G4s were identified considering the following restriction parameters: for QGRS, maximum sequence length of 30 nucleotides (nt), minimum G-tetrads number in a G4 of 2, loops lengths in the range of 0−36 nt, G-Score values > 15; For G4-Hunter threshold of 1.5 with a window size of 20 nt.

Data availability

Anonymized data are available from the corresponding author.

Results

Identification of novel pathogenic repeat motifs in RFC1 in the 100,000 Genome project

Out of 893 cases diagnosed with adult-onset ataxia (over the age of 30 years) recruited as part of the 100,000 Genome project, 124 cases had at least one AAGGG repeat expansion and 48 had biallelic AAGGG repeat expansions, thus confirming a diagnosis of CANVAS and disease spectrum.

To identify additional likely pathogenic repeat motifs in *RFC1* we specifically looked for rare repeat configurations present in patients diagnosed with adult-onset ataxia (over the age of 30 years) or in compound heterozygous state with the known pathogenic AAGGG repeat expansion, but absent or significantly less frequent in controls under the same conditions (**Table 1**).

We identified three cases carrying repeat expansions AAGGC (Case I-1), AGGGC (case II-1), or AGAGG (case VII-1) repeat motifs which were absent in non-neurological controls. AAGGC was present in the homozygous state, while AGGGC and AGAGG were in the

compound heterozygous state with the AAGGG expansion. One additional case of self-reported Asian ancestry carried the previously reported rare pathogenic ACAGG repeat expansion in the homozygous state.

AAAAG, AAAGGG, AAGAG expansions were found with similar frequency in patients and controls, supporting their non-pathogenic significance, while there was a higher percentage of compound heterozygous AAGGG/AAAGG carriers in ataxia cases (p=0.05).

All predicted genetic ancestries for individuals carrying rare homozygous or compound heterozygous expansions in RFC1 are reported in **Supplementary Table 2**. Patients carrying AAGGC (Case I-1) and AGGGC (case II-1) expansion were of predicted South Asian and mixed ethnicity, respectively; ACAGG expansion carrier was confirmed to be East Asian based on the predicted genetic ancestry, while other repeat configurations were mostly identified in individuals of European or mixed ethnicity.

We did not identify any loss-of-function variant or structural variant in the RFC1 gene in individuals carrying heterozygous AAGGG repeat expansions.

The presence of AGGGC, AAGGC or AGAGG repeat expansions was confirmed by RP-PCR in all three cases, and the AAGGC repeat segregated with the disease in family I as it was also present in the affected sister I-2. (**Figure 1A**)

Additionally, one case with isolated cerebellar ataxia carried the AAGGG expansion along with an ACGGG repeat, which was also absent in controls. However, Sanger sequencing showed that the ACGGG expansion was only 50 repeats, which is considerably below the lower limit of pathogenicity for the pathogenic AAGGG motifs of 250 repeats and was therefore considered likely non-pathogenic in this case. Notably, the patient exhibited isolated cerebellar ataxia but no neuropathy, which is unusual in RFC1 disease.

We next screened by RP-PCR an internal cohort of 540 DNA samples from cases with sensory neuropathy, ataxia or CANVAS and identified five additional cases carrying an AGGGC expansion (cases III-1, IV-1, V-1, V-2 and VI-1) and three cases carrying AAAGG expansions on the second allele (cases X-1, XI-1, XII-1) (**Table 2**). We did not identify additional AGAGG or AAGGC repeat expansion carriers. All cases were of self-reported Caucasian ethnicity.

Based on Southern blotting, OGM or long-read sequencing **(Figure 1B-C),** when available, we observed that the repeat size of the rare AGGGC, AAGGC and AGAGG was >600 repeats in all cases (mean±SD, 892±247) (**Figure 2A**).

Also, enough DNA for Southern blotting was available from 5 patients with CANVAS spectrum as defined by the presence of sensory neuropathy and at least one of the additional features of the full syndrome (cerebellar dysfunction, vestibular areflexia, cough) (cases VI-X) and 8 controls carrying compound heterozygous AAGGG/AAAGG expansions (**Figure 2B**).

In CANVAS patients, the AAAGG expansions were always ≥600 repeats (mean±SD, 979±257), and were significantly larger compared to AAAGG expansions (238±142 repeat units) found in controls (p<0.0001), suggesting that, although the AAAGG repeat is usually small and non-pathogenic, as shown in **Figure 2A**, larger AAAGG repeat expansion occur and may have a pathogenic role.

Long read sequencing confirms the sequence of the expanded repeats

To gain further insight into the exact sequence of the novel pathogenic motifs, we performed targeted long-read sequencing (**Figure 1D and Supplementary Table 1**).

We confirmed the presence of uninterrupted $AGGGC_{1240}$ in case II-1 and $AGGGC_{3200}$ in case III-1. Moreover, long-read sequencing was able to accurately define the exact repeat composition of AGAGG and AAGGC expansions, which revealed the presence of mixed repeat motifs $(AAGGC)_{900}(AAGGG)_{940}$ and $(AGAGG)_{470}(AAAGG)_{470}$, in case I-1 and VII-1, respectively. Long-read sequencing was also performed in five cases carrying large AAAGG expansion and showed the presence of uninterrupted AAAGG motifs in three (Cases X-1, XI-1 and XII-1), with sizes of 980, 800 and 600 repeat units, respectively, while two probands (case VIII-1 and IX-1) carried complex $(AAAGG)_{610}$ $(AAGGG)_{390}$ and $(AAAGG)_{700}(AAGGG)_{200}$ repeats.

All pathogenic repeat configurations share an ancestral haplotype

Subsequently, we looked at the inferred haplotypes associated with the novel pathogenic repeat motifs. A region of 66kb (**Figure 3**, between markers B and C, chr4:39302305-39366034, hg38) is shared among all pathogenic alleles. It is worth noting that a larger region of 207 kb (between markers A and C), and which contains *WDR19* and *RFC1* genes, is shared among all pathogenic alleles but one (Case III-1), where the haplotype becomes the same as the wild type allele. This suggests a probable more recent recombination event at marker B for Case III-1. The larger shared region identified in carriers of the novel pathogenic configurations, as well

as in AAGGG and AAAGG carriers, supports the existence of an ancestral haplotype which gave rise to these expanded alleles. Notably, non-pathogenic $AAAAG_{(9\text{-}11)}$ and expanded AAAAG repeat originated from a different haplotype.

We estimated that the ancestral haplotype which gave rise to different pathogenic repeat configurations in RFC1 likely dates back to 56,100 years ago(CI at 95% 27,680-115,580 years)

Clinical features of patients carrying novel pathogenic repeat configurations in *RFC1*

We found 14 patients from 12 families carrying novel pathogenic *RFC1* repeat configurations. Demographic and clinical characteristics of patients are summarized in **Table 2**. All patients were Europeans, apart from patient I-1 and I-2 who were from India and patient X-1 who was from Australia. Mean age of onset was 51.5 ± 13.7 (24-73) years, and mean disease duration at examination was 17.2 years ± 8.7 (3-34) years. Six patients had isolated sensory neuropathy, which was associated with cough in four of them, one patient had sensory neuropathy and vestibular dysfunction, while seven cases had full CANVAS. Additional features were observed in some cases including early onset and rapid progression (case I-1), cognitive impairment (III-1, VI-1), muscle cramps (I-1, II-1, III-1 and IV-1), and REM sleep behaviour disorder with positive dopamine transporter scan (DatScan) (IX-1). Autonomic dysfunction was observed in six cases and it two of them (II-1, III-1), who both carried AGGGC expansion, was severe and led to syncopal episodes. Detailed descriptions of the clinical features are provided in **supplementary note.**

Pathogenic configurations in *RFC1* are predicted to form G quadruplexes

As repetitive G-rich sequences are known to form G quadruplexes (G4)[32,35,36], a secondary DNA structure which act as transcriptional regulator by impeding transcription factor binding to duplex-DNA or stalling the progression of RNA polymerase, we set out to evaluate the propensity of the different repeat configurations in *RFC1* to form G4.

All pathogenic repeat configurations showed high G4 scores, in the range observed for the well-known G4-forming regions of the *cMYC*[37] and *HRAS1*[38] genes, as predicted by QGRS-Mapper and G4Hunter in contrast to the non-pathogenic AAAAG (**Table 3**).

Discussion

We leveraged WGS data from nearly 10,000 individuals recruited to the Genomics England sequencing project to investigate the normal and pathogenic variation of the RFC1 repeat. We identified three novel repeat configurations associated with CANVAS and disease spectrum, including AGGGC, AAGGC and AGAGG. Notably, we also showed a pathogenic role for large uninterrupted or interrupted AAAGG expansions, while AAAAG, AAGAG and AAAGGG expansions are likely always benign (**Figure 4**).

Most pathogenic repeat expansion were found in individuals of Caucasian ancestry, however ACAGG seems to be common in East Asians, while AAGGC was identified in a family of South Asian ancestry. Interestingly most pathogenic repeats seem to have arisen from a shared region 207 Kb, supporting their origin from a common ancestor who lived ~50,000 years ago. *Rafehi et al.*[2] had previously identified a larger ancestral haplotype in Australian patients affected by CANVAS of 360 Kb and estimated that the most recent common ancestor lived approximately 25,880 (CI: 14,080–48,020) years ago.[2] In our study, the inclusion of additional pathogenic repeat configurations and multiple ethnicities allowed the identification of a smaller core haplotype and has extended further back in time the origin of the common ancestor carrying a pathogenic repeat in *RFC1*. It is reasonable to believe that occurrence of subsequent A-G transition and A-G or G-C transversion in the polyA tail of the AluxSx3 element on the ancestral haplotype has favoured the further expansion of GC rich motifs over the millennia. Since the most significant recent wave out of Africa is estimated to have taken place about 70,000–50,000 years ago, we can speculate that the repeat containing haplotype spread with the migration of early modern humans from Africa through the Near East and to the rest of the world.

Patients showed clinical features undistinguishable from those of patients carrying biallelic AAGGG expansion. In some cases, however, the disease appeared to be more severe due to symptomatic dysautonomia, early cerebellar involvement or disabling gait disturbance.

The identification of these motifs has direct clinical implications. Given their frequency, RP-PCR for AAAGG and AGGGC should be considered in all cases. Particular attention should be paid to carriers of compound AAGGG/AAAGG expansions and accurate sizing and full sequencing of the satellite through long read is recommended to establish its possible pathogenicity. In addition, depending on availability, Southern blotting, genome optical mapping, or long read sequencing, are warranted in patients with a suggestive clinical

phenotype but inconclusive screening, such as in cases with absence of PCR amplifiable product on flanking PCR but negative RP-PCR for AAGGG expansion).

The findings of this study highlight the genetic complexity of *RFC1*-related disease and lend support to the hypothesis that the size and GC content of the pathogenic repeat is more important than the exact repeat motif. Consistently, all pathogenic repeat configurations are rich in G content and are predicted to form highly stable G quadruplexes, which have been previously demonstrated to affect gene transcription in other pathogenic conditions.[35,36]

Both Nanopore or Pacbio sequencing platforms and either the targeted CRISPR/Cas9 or adaptive selection approach were used to increase accuracy of the sequencing of *RFC1* repeat locus. Despite several attempts and similarly to other large satellites, long read sequencing of the *RFC1* repeat remains challenging and, depending on the specific configurations, size, and DNA quality, only few reads were available for analysis in some cases. Notably, an uneven coverage at the *RFC1* locus across samples was also observed in a recent study looking at *RFC1* repeat composition through Nanopore sequencing. The authors attributed the variability to variable degrees of DNA fragmentation depending on the delay between blood sampling and DNA extraction. Hopefully, constant advancements in long read sequencing platforms and decrease in cost (currently ~1,000 USD per sample) will soon translate into increased accessibility and higher accuracy of this technology.

In conclusion, the study expanded the genetic heterogeneity underlying *RFC1* CANVAS and disease spectrum, and identified three additional pathogenic AAGGC, AGGGC and AGAGG repeat motifs. We also demonstrated a pathogenic role for large uninterrupted or interrupted AAAGG expansions, thereby highlighting the importance of sizing and, if possible, full sequencing of the *RFC1* satellite expansion in clinically selected cases, in order to correctly diagnose and counsel patients and their families.

Appendix 1

Genomics England Research Consortium

J. C. Ambrose, P. Arumugam, E. L. Baple, M. Bleda, F. Boardman-Pretty, J. M. Boissiere, C. R. Boustred, H. Brittain, M. J. Caulfield, G. C. Chan, C. E. H. Craig, L. C. Daugherty, A. de Burca, A. Devereau, G. Elgar, R. E. Foulger, T. Fowler, P. Furió-Tarí, E. Gustavsson, J. M. Hackett, D. Halai, A. Hamblin, S. Henderson, J. E. Holman, T. J. P. Hubbard, K. Ibáñez, R.

Jackson, L. J. Jones, D. Kasperaviciute, M. Kayikci, L. Lahnstein, K. Lawson, S. E. A. Leigh, I. U. S. Leong, F. J. Lopez, F. Maleady-Crowe, J. Mason, E. M. McDonagh, L. Moutsianas, M. Mueller, N. Murugaesu, A. C. Need, C. A. Odhams, C. Patch, D. Perez-Gil, D. Polychronopoulos, J. Pullinger, T. Rahim, A. Rendon, P. Riesgo-Ferreiro, T. Rogers, M. Ryten, B. Rugginini, K. Savage, K. Sawant, R. H. Scott, A. Siddiq, A. Sieghart, D. Smedley, K. R. Smith, A. Sosinsky, W. Spooner, H. E. Stevens, A. Stuckey, R. Sultana, E. R. A. Thomas, S. R. Thompson, C. Tregidgo, A. Tucci, E. Walsh, S. A. Watters, M. J. Welland, E. Williams, K. Witkowska, S. M. Wood, M. Zarowiecki.

Further details are available in the Supplementary material.

Funding

This work was supported by the Medical Research Council (MR/T001712/1), Fondazione Cariplo (grant n. 2019-1836), the Inherited Neuropathy Consortium, Fondazione Regionale per la Ricerca Biomedica (Regione Lombardia, project ID 1751723), National Ministry of Health (Ricerca Corrente 2021-2022) and Italian Ministry for Universities and Research (MUR, 20229MMHXP) to awarded to A.C. This work has also been supported by a Medical Research Future Fund Genomics Health Futures Mission grant (APP2007681) awarded to M.L.K. and S.V. and grant CP 20/2018 from the Fondazione Regionale per la Ricerca Biomedica to F.T. F.A. was supported by NIH Intramural Research Program, the US National Institute on Aging. E.A. was partially supported by the Telethon Foundation and by the Rotary Club Milano Ovest.

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain* online

References

1. Cortese A, Simone R, Sullivan R, et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. Nat Genet. 2019;51(4):649–58.

2. Rafehi H, Szmulewicz DJ, Bennett MF, et al. Bioinformatics-Based Identification of Expanded Repeats: A Non-reference Intronic Pentamer Expansion in RFC1 Causes CANVAS. Am J Hum Genet. 2019 Jun 12;

3. Cortese A, Reilly MM, Houlden H. RFC1 CANVAS / Spectrum Disorder. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Stephens K, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2021 Jan 5]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK564656/

4. Cortese A, Curro' R, Vegezzi E, Yau WY, Houlden H, Reilly MM. Cerebellar ataxia, neuropathy and vestibular areflexia syndrome (CANVAS): genetic and clinical aspects. Pract Neurol. 2022 Feb;22(1):14–8.

5. Cortese A, Tozza S, Yau WY, et al. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome due to RFC1 repeat expansion. Brain. 2020 Feb 1;143(2):480–90.

6. Currò R, Salvalaggio A, Tozza S, et al. RFC1 expansions are a common cause of idiopathic sensory neuropathy. Brain. 2021 Jun 22;144(5):1542–50.

7. Kumar KR, Cortese A, Tomlinson SE, et al. RFC1 expansions can mimic hereditary sensory neuropathy with cough and Sjögren syndrome. Brain. 2020 Oct 1;143(10):e82.

8. Ronco R, Perini C, Currò R, et al. Truncating Variants in RFC1 in Cerebellar Ataxia, Neuropathy, and Vestibular Areflexia Syndrome. Neurology. 2023 Jan 31;100(5):e543–54.

9. Benkirane M, Da Cunha D, Marelli C, et al. RFC1 nonsense and frameshift variants cause CANVAS: clues for an unsolved pathophysiology. Brain. 2022 Nov 21;145(11):3770–5.

10. Huin V, Coarelli G, Guemy C, et al. Motor neuron pathology in CANVAS due to RFC1 expansions. Brain. 2021 Dec 20;awab449.

11. Traschütz A, Cortese A, Reich S, et al. Natural History, Phenotypic Spectrum, and Discriminative Features of Multisystemic RFC1 Disease. Neurology. 2021 Mar 2;96(9):e1369–82.

12. Aboud Syriani D, Wong D, Andani S, et al. Prevalence of RFC1-mediated spinocerebellar ataxia in a North American ataxia cohort. Neurol Genet. 2020 Jun;6(3):e440.

13. Beijer D, Dohrn MF, De Winter J, et al. RFC1 repeat expansions: A recurrent cause of sensory and autonomic neuropathy with cough and ataxia. Eur J Neurol. 2022 Jul;29(7):2156–61.

14. Gisatulin M, Dobricic V, Zühlke C, et al. Clinical spectrum of the pentanucleotide repeat expansion in the RFC1 gene in ataxia syndromes. Neurology. 2020 Sep 1;

15. Tagliapietra M, Cardellini D, Ferrarini M, et al. RFC1 AAGGG repeat expansion masquerading as Chronic Idiopathic Axonal Polyneuropathy. J Neurol. 2021 Nov;268(11):4280–90.

16. Montaut S, Diedhiou N, Fahrer P, et al. Biallelic RFC1-expansion in a French multicentric sporadic ataxia cohort. J Neurol. 2021 Sep;268(9):3337–43.

17. Van Daele SH, Vermeer S, Van Eesbeeck A, et al. Diagnostic yield of testing for RFC1 repeat expansions in patients with unexplained adult-onset cerebellar ataxia. J Neurol Neurosurg Psychiatry. 2020 Jul 30;

18. Ghorbani F, de Boer-Bergsma J, Verschuuren-Bemelmans CC, et al. Prevalence of intronic repeat expansions in RFC1 in Dutch patients with CANVAS and adult-onset ataxia. J Neurol. 2022 Nov;269(11):6086–93.

19. Erdmann H, Schöberl F, Giurgiu M, et al. Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing. Brain. 2022 Oct 13;awac377.

20. Beecroft SJ, Cortese A, Sullivan R, et al. A Māori specific RFC1 pathogenic repeat configuration in CANVAS, likely due to a founder allele. Brain. 2020 Sep 1;143(9):2673–80.

21. Scriba CK, Beecroft SJ, Clayton JS, et al. A novel RFC1 repeat motif (ACAGG) in two Asia-Pacific CANVAS families. Brain. 2020 Oct 1;143(10):2904–10.

22. Nakamura H, Doi H, Mitsuhashi S, et al. Long-read sequencing identifies the pathogenic nucleotide repeat expansion in RFC1 in a Japanese case of CANVAS. J Hum Genet. 2020 May;65(5):475–80.

23. Miyatake S, Yoshida K, Koshimizu E, et al. Repeat conformation heterogeneity in cerebellar ataxia, neuropathy, vestibular areflexia syndrome. Brain. 2022 Apr 29;145(3):1139–50.

24. 100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, Martin A, et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med. 2021 Nov 11;385(20):1868–80.

25. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016 Apr 15;32(8):1220–2.

26. Stevanovski I, Chintalaphani SR, Gamaarachchi H, et al. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. Sci Adv. 2022 Mar 4;8(9):eabm5386.

27. Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. Nat Biotechnol. 2021 Apr;39(4):442–50.

28. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100.

29. Gamaarachchi H, Samarakoon H, Jenner SP, et al. Fast nanopore sequencing data analysis with SLOW5. Nat Biotechnol. 2022 Jul;40(7):1026–9.

30. Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nat Commun. 2019 Nov 28;10(1):5436.

31. Gandolfo LC, Bahlo M, Speed TP. Dating Rare Mutations from Small Samples with Dense Marker Data. Genetics. 2014 Aug 1;197(4):1315–27.

32. Frasson I, Pirota V, Richter SN, Doria F. Multimeric G-quadruplexes: A review on their biological roles and targeting. Int J Biol Macromol. 2022 Apr 15;204:89–102.

33. Kikin O, D'Antonio L, Bagga PS. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W676-682.

34. Bedrat A, Lacroix L, Mergny JL. Re-evaluation of G-quadruplex propensity with G4Hunter. Nucleic Acids Res. 2016 Feb 29;44(4):1746–59.

35. Varshney D, Spiegel J, Zyner K, Tannahill D, Balasubramanian S. The regulation and functions of DNA and RNA G-quadruplexes. Nat Rev Mol Cell Biol. 2020 Aug;21(8):459–74.

36. Wang E, Thombre R, Shah Y, Latanich R, Wang J. G-Quadruplexes as pathogenic drivers in neurodegenerative disorders. Nucleic Acids Res. 2021 May 21;49(9):4816–30.

37. Dickerhoff J, Dai J, Yang D. Structural recognition of the MYC promoter G-quadruplex by a quinoline derivative: insights into molecular targeting of parallel G-quadruplexes. Nucleic Acids Research. 2021 Jun 4;49(10):5905–15.

38. Cogoi S, Shchekotikhin AE, Xodo LE. HRAS is silenced by two neighboring G-quadruplexes and activated by MAZ, a zinc-finger transcription factor with DNA unfolding property. Nucleic Acids Res. 2014 Sep 1;42(13):8379–88.

Figure legends

Figure 1 Long read sequencing defines the precise sequence of the novel pathogenic RFC1 motifs. Pedigrees with arrow and P indicating proband **(panel A**), RP-PCR plots and, where available, southern blot images and optical genome mapping plots (**panel B**), long-read sequencing results (**panel C**) of representative patients with AAGGC, AGGGC, AGAGG and AAAGG expansions (cases I-1, III-1, VII-1, XII-1). In case III-1 only partial reads, which did not span the entire RFC1 repeat locus, could be obtained from the AAGGG allele.

Figure 2 RFC1 repeat expansion size. A) Comparison of repeat sizes of alleles carrying AAGGG, AAAGG, AAGGC, AGGGC and AGAGG expansion from this and previous studies.[1,5,6] The dotted lines refer to the smallest pathogenic of 250 AAGGG repeat identified so far **B**) Comparison of the AAAGG repeat sizes in the compound heterozygous state with the AAGGG expansion, in patients with CANVAS and disease spectrum versus controls.

Figure 3 A shared ancestral haplotype in patients with pathogenic RFC1 motifs. Graphical representation of the haplotypes associated with AAGGG, ACAGG and novel pathogenic repeat motifs identified in this study. For each SNP, the reference allele is represented in blue, while the alternative allele is represented in yellow. The repeat expansion locus is marked with a red line (R). There is a shared region (B-C,-rs2066782-rs6851075, chr4:39302305-39366034, hg38) of 66kb for all novel configurations. A larger region of 207kb (A-C, rs148316325-rs6851075, chr4:39158847-39366034, hg38), which is flanked by two recombination hotspots (arrows), is also shared among all but one allele for Case III-1 suggesting a recombination event at B (rs2066782) in this family. The shared haplotype lies in a region of low-recombination rate (HapMap data) and is delimited by small peaks at A and C. A smaller increase in recombination rate is also visible at B.

Figure 4 Normal and pathogenic significance of repeat expansion motifs at RFC1 locus.

Table 1 Normal and pathogenic variation of the RFC1 repeat locus on the 100,000 Genome Project

	Hereditary ataxia (N=893)	Non neurological controls (N=8107)	P values
Rare homozygous (<1%) repeat expansions present in ataxia cases and absent in controls			
ACAGG (hom)	1 (0.01%)	0 (0 %)	-
AAGGC (hom)	1 (0.01%)	0 (0 %)	-
Repeat expansion found in compound heterozygous state with AAGGG expansions (allele 1/allele2)			
AAGGG/AAAAG	21 (2.3%)	248 (3%)	Ns
AAGGG/AAAGGG	5 (0.6%)	32 (0.4%)	Ns
AAGGG/AAGAG	3 (0.3%)	16 (0.2%)	Ns
AAGGG/**AAAGG**	10 (1.1%)	47 (0.6%)	0.05
AAGGG/ACGGG*	1 (0.01%)	0 (0%)	-
AAGGG/**AGAGG**	1 (0.01%)	0 (0 %)	-
AAGGG/**AGGGC**	1 (0.01%)	0 (0 %)	-

Ns=not significant. *Small $(ACGGG)_{50}$ expansion in typical non-pathogenic range (10-220). Novel pathogenic repeat motifs identified in this study are bold highlighted.

1 **Table 2 Clinical and demographic features of patients carrying novel pathogenic repeat configurations in *RFC1***

	RFC1 genotype	Sex	Ethnicity	Phenotype	Age of onset	Disease duration (years)	Chronic cough	Cerebellar syndrome	Sensory neuropathy	Bilateral vestibular areflexia	Dysautonomia	Use of walking aids (age)	Additional features
AAGGC expansion													
Case I-1	Allele 1: $(AAGGG)_{510}$ $(AAGGC)_{880}$ Allele 2: $(AAGGG)_{940}$ $(AAGGC)_{900}$	F	Caucasian (Indian)	CANVAS	24	17	Yes	Yes	Yes	Yes	No	Stick (36)	Cramps, pyramidal signs
Case I-2	Allele 1: $(AAGGG)_n(AAGGC)_n$ Allele 2: $(AAGGG)_n(AAGGC)_n$	F	Caucasian (Indian)	Sensory neuropathy + cough	34	8	Yes	N/A	Yes	N/A	N/A	No	/
AGGGC expansion													
Case II-1	Allele 1: $(AGGGC)_{1240}$ Allele 2: $(AAGGG)_{930}$	M	Mixed (Lebanese)	Sensory neuropathy + vestibular dysfunction	53	11	Yes	No	Yes	Yes	Yes	No	Cramps
Case III-1	Allele 1: $(AGGGC)_{3200}$ Allele 2: $(AAGGG)_{1000}$	M	Caucasian (British)	CANVAS	71	12	Yes	Yes	Yes	N/A	Yes	Wheelchair (81)	Cramps, cognitive/behavioural abnormalities after age 80
Case IV-1	Allele 1: $(AGGGC)_{1875}$ / Allele 2: $(AAGGG)_{500}$	M	Caucasian (Italian)	CANVAS	41	34	No	Yes	Yes	Yes	Yes	Wheelchair (72)	Cramps
Case V-1	Allele 1: $(AGGGC)_n$/ Allele 2: $(AAGGG)_n$	F	Caucasian (Italian)	Sensory neuropathy + cough	60	13	Yes	No	Yes	No	No	No	/
Case V-2	Allele 1: $(AGGGC)_n$/ Allele 2: $(AAGGG)_n$	F	Caucasian (Italian)	Sensory neuropathy	40	20	No	No	Yes	No	No	No	/
Case VI-1	Allele 1: $(AGGGC)_n$/ Allele 2: $(AAGGG)_n$	F	Caucasian (Italian)	Sensory ganglionopathy + cough	62	23	Yes	No	Yes	N/A	Yes	No	Voice and hand tremor, urinary incontinence
AGAGG expansion													
Case VII-1	Allele 1: $(AAAGG)_{470}$ $(AGAGG)_{470}$ / Allele 2: $(AAGGG)_{1140}$	F	Caucasian (British)	CANVAS	50	24	Yes	Yes	Yes	Yes	No	Walker (69), wheelchair(74)	/
AAAGG expansion													
Case VIII-1	Allele 1: $(AAAGG)_{610}$ $(AAGGG)_{390}$ / Allele 2: $(AAGGG)_{110}$	M	Caucasian (British)	CANVAS	55	20	Yes	Yes	Yes	N/A	Yes	Walker and wheelchair (74)	Cognitive impairment since age 72

	0												
Case IX-1	Allele 1: $(AAGGG)_{700}$ $(AAAGG)_{200}$ / Allele 2: $(AAGGG)_{1170}$	M	Caucasian (British)	CANVAS	45	31	Yes	Yes	Yes	Yes	Yes	Walker (75)	RBD, positive DatScan
Case X-1	Allele 1: $(AAAGG)_{980}$/ Allele 2: $(AAGGG)_{1010}$	M	Caucasian (Australian)	CANVAS	58	15	Yes	Yes	Yes	Yes	N/A	N/A	/
Case XI-1	Allele 1: $(AAAGG)_{800}$/ Allele 2: $(AAGGG)_{500}$	F	Caucasian (Italian)	Sensory ganglionopathy + cough	73	3	Yes	No	Yes	No	No	Stick (77)	/
Case XII-1	Allele 1: $(AAAGG)_{600}$/ Allele 2: $(AAGGG)_{390}$	M	Caucasian (Italian)	Sensory ganglionopathy + cough	56	10	Yes	No	Yes	No	No	No	/

1
2

Table 3 Pathogenic RFC1 motifs are predicted to form G quadruplexes

Gene - Analyzed Sequences	QGRS-Mapper Score	G4Hunter Score
RFC1 – $(AGGGC)_{10}$	42	1.83
RFC1 – $(AAGGG)_{10}$	42	2.00
RFC1 - $(AAGGC)_{10}$	21	1.82
RFC1 - $(AAAGG)_{10}$	21	0.94
RFC1 – $(AGAGG)_{10}$	21	1.12
RFC1 - $(AAAAG)_{10}$	No putative G4 identified	
c-MYC – TGGGGAGGTGGGGAGGGTGGGGAAGG	41	2.59
HRAS-1 – TCGGGTTGCGGCGCAGGCACGGGCG	41	1.19

G-score values comparison between repeat configurations found in *RFC1* and well-known G4s-forming sequences.

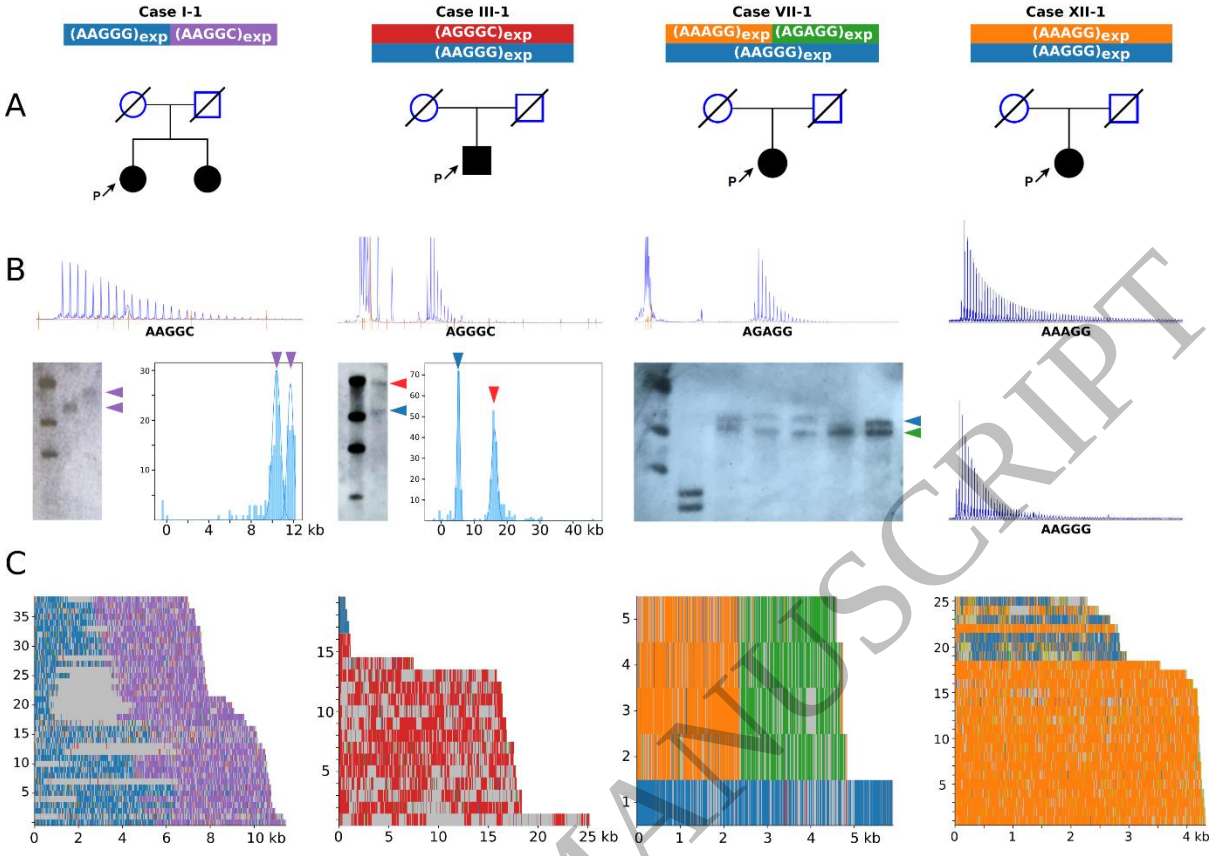


Figure 1

543x381 mm (x DPI)

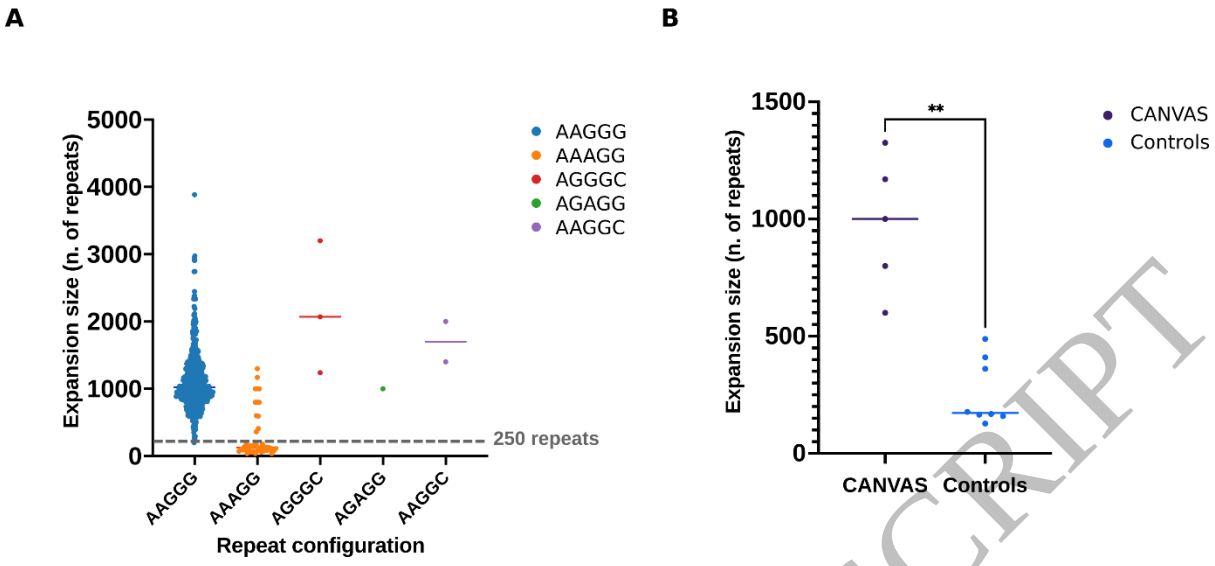


Figure 2
543x248 mm (x DPI)

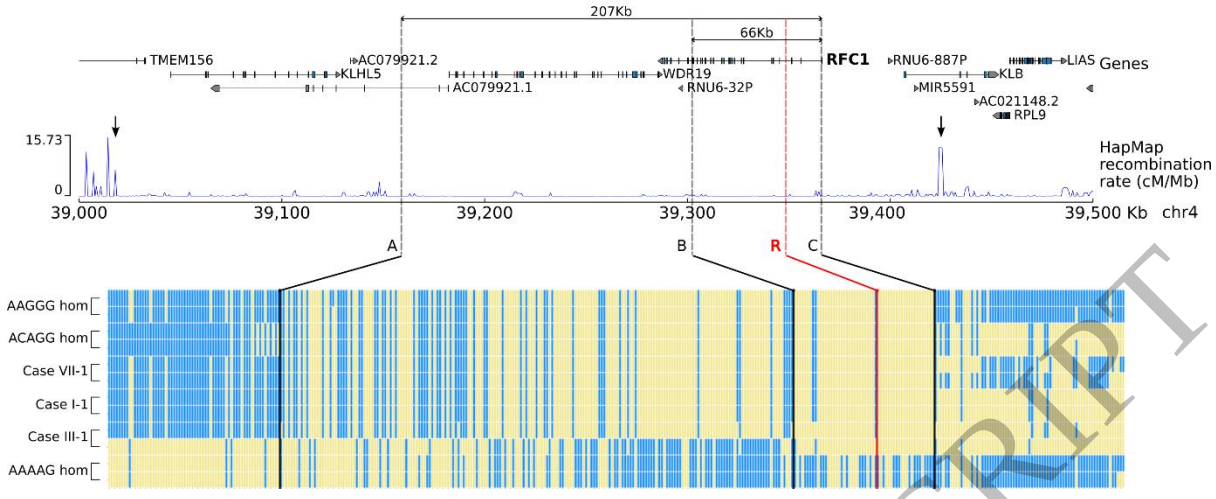
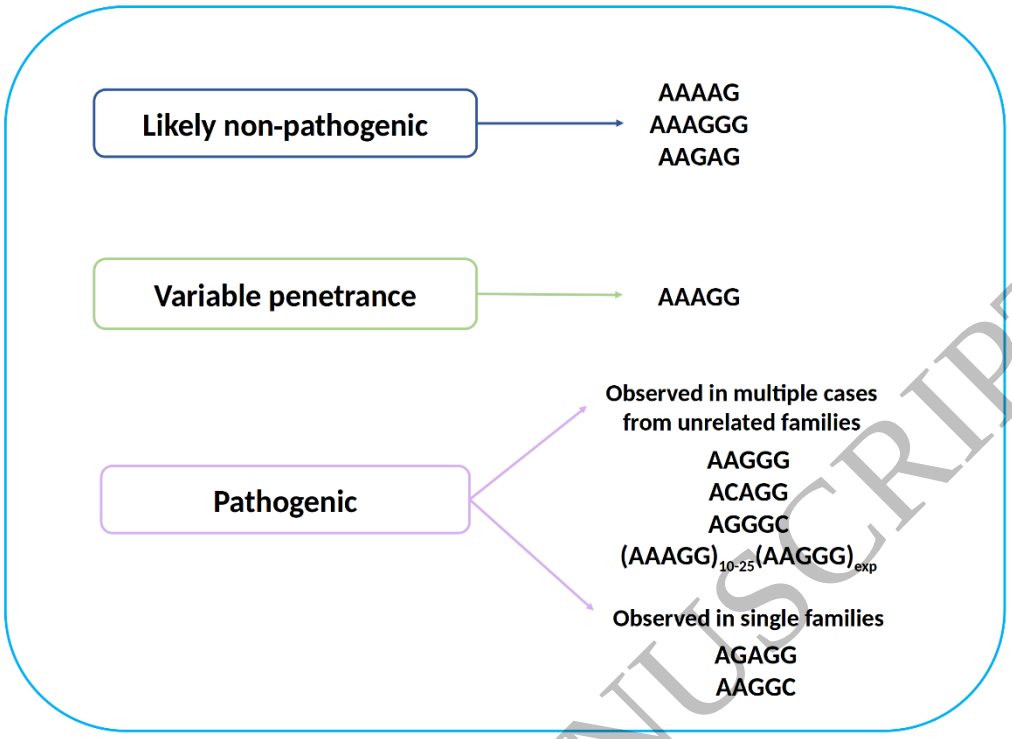


Figure 3
543x223 mm (x DPI)



1

2 *Figure 4*
3 *559x347 mm (x DPI)*