Article

Heterozygous and Homozygous *RFC1 AAGGG* Repeat Expansions are Common in Idiopathic Peripheral Neuropathy

Zitian Tang,[1,†] Sinem S. Ovunc,[2,†] Elle Mehinovic,[1] Simone Thomas,[3] Jenna Ulibarri,[1] Zefan Li,[1] Dustin Baldridge,[4] Carlos Cruchaga,[5] Matt Johnson,[5] Jeffrey Milbrandt,[1,6,7] Brian Callaghan,[2] PNRR Study Group,[3] Ahmet Höke,[3,*] Peter K. Todd, [2,8,*] Sheng Chih Jin,[1,4,*]

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

[*]**Co-corresponding authors**

Author affiliations

1 Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

2 Department of Neurology, University of Michigan, Ann Arbor, MI, USA

3 Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

4 Department of Pediatrics, Washington University School of Medicine, St. Louis, MO, USA

5 NeuroGenomics and Informatics Center, Washington University School of Medicine in St. Louis, St. Louis, MO, USA

6 Needleman Center for Neuro-metabolism and Axonal Therapeutics, St. Louis, MO, USA

7 McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA

8 Veterans Affairs Medical Center, Ann Arbor, MI, USA.

Correspondence to:

Sheng Chih Jin, Ph.D.

Assistant Professor of Genetics and Pediatrics

Washington University School of Medicine

Article

4515 McKinley Ave.

St. Louis, MO, 63110, USA

E-mail: jin810@wustl.edu

Peter K. Todd, M.D., Ph.D.

Sackett Endowed Professor of Neurology, Associate Chair for Research

University of Michigan

Staff Physician, Ann Arbor VA Medical Center

4148 BSRB

109 Zina Pitcher Place

Ann Arbor, MI 48109-2200, USA

E-mail: petertod@med.umich.edu

Ahmet Höke, M.D., Ph.D. FRCPC

Professor, Neurology and Neuroscience

W. W. Smith Charitable Trust Professor of Neuroimmunology

Director, Neuromuscular Division

Director, Merkin Peripheral Neuropathy and Nerve Regeneration Center

Editor-in-Chief, Annals of Clinical and Translational Neurology

Johns Hopkins School of Medicine

855 N. Wolfe St., Rangos 248

Baltimore, MD, 21205, USA

E-mail: ahoke1@jh.edu

Running head: *RFC1* tandem repeat in idiopathic peripheral neuropathy

Article

Abstract

Objective

Biallelic intronic AAGGG repeat expansions in *RFC1* cause Cerebellar Ataxia with Neuropathy and Vestibular Areflexia Syndrome and may also contribute to isolated sensory neuropathy. The clinical significance of both heterozygous and biallelic *RFC1* expansions in more diverse patient populations remains unclear—partly due to the absence of accurate, user-friendly computational tools specifically tailored for tandem repeat analysis.

Methods

To discern the relationship between *RFC1* expansions and idiopathic peripheral neuropathy (iPN), we performed whole-genome sequencing (WGS) followed by PCR-based confirmation in a large, well-characterized U.S. cohort consisting of 788 iPN patients (369 pure small fiber neuropathy (SFN), 266 sensorimotor, 144 pure sensory, and 9 pure motor). We developed an integrative pipeline combining ExpansionHunter Denovo and Expansion Hunter coupled with unsupervised clustering to reliably detect and genotype *RFC1* expansions from short-read WGS data, achieving 98.2% concordance with repeat-primed PCR based validation.

Results

Biallelic *RFC1* expansions were absent in 879 controls but present in 2.8% of iPN patients (Fisher's exact $p = 5.9\text{x}10^{-8}$), including 6.2% of pure sensory, 2.2% of SFN, and 1.5% of sensorimotor neuropathy, indicating that motor nerve involvement should not exclude patients from *RFC1* repeat screening. We also observed a markedly increased frequency of monoallelic expansions in iPN compared to controls (13.2% versus 2.5%; Fisher's exact $p = 3.4\text{x}10^{-17}$), without evidence of secondary mutations or expansions on the other allele.

Article

Interpretation

Our approach provides a robust, cost-effective method for detecting *RFC1* expansions from WGS data. Our findings indicate that both heterozygous and homozygous AAGGG repeat expansions in *RFC1* can contribute to development of iPN.

Article

Introduction

Peripheral neuropathy (PN) is among the most common neurological disorders, with an estimated prevalence of 12–20% in the United States and up to 30% in older adults.[1] PN causes a range of debilitating symptoms including pain, sensory loss, imbalance, and weakness. These manifestations, although rarely lethal, contribute to significant humanistic and socioeconomic burden, reflected in patients' reduced quality of life and severe impairments in daily functioning.[2] Approximately one-third of all PN cases are classified as idiopathic (iPN), wherein no definitive etiology can be determined despite thorough and appropriate clinical evaluation.[3] As upwards of 18% of late-onset iPN may be explained by genetic causes,[4] neurology clinics increasingly implement gene-panel based testing for potential pathogenic single nucleotide variants (SNVs) and copy number variations (CNVs).[5] Despite these efforts, the majority of iPN patients remain undiagnosed even after gene panel testing.

Short tandem repeats (STRs), defined as repetitive units of 1 to 6 base pairs (bp) adjacent to each other, constitute around 8% of the human genome.[6–8] Due to replication slippage and mismatch repair errors that occur during DNA replication and RNA transcription, STRs have among the highest mutation rates of all genetic variant classes.[6,9,10] Genomic instability caused by pathogenic STR expansion contributes to over 70 human disorders, many of which are neurological.[11] Recent large-scale initiatives such as the 1000 Genomes Project[12] and the Simons Genome Diversity Project[13] have cataloged extensive polymorphic STRs in healthy individuals, revealing that most loci exhibit multiple common alleles and that each genome carries thousands of STR variants relative to the reference. The increasing accuracy of short-read, high-coverage whole-genome sequencing (WGS), coupled with specialized bioinformatic tools, has further refined our understanding of the prevalence and impact of

STR expansions on human health.[12,13] Nonetheless, current STR detection software—such as ExpansionHunter Denovo (EHdn)[14], ExpansionHunter (EH)[15], GangSTR[16], and exSTRa[17]—faces several limitations, including reliance on predefined repeat catalogs, incomplete assessment of novel motifs, limited ability to capture zygosity, and complex computational requirements that pose hurdles for routine clinical use.

STR expansions have emerged as a potential contributor to PN pathogenesis. In 2019, a biallelic non-reference AAGGG repeat expansion in the second intron of the *RFC1* gene was identified as the most common cause of cerebellar ataxia, neuropathy, and vestibular areflexia syndrome (CANVAS), a multisystemic disorder characterized by progressive imbalance, sensory neuronopathy and cerebellar atrophy.[18–20] This non-reference expansion replaces the wild-type $(AAAAG)_{11}$ repeat at this locus in the reference genome, with pathogenic expansions typically exceeding 250 AAGGG repeats.[19,21] Subsequent to this discovery, several other pathogenic repeat motifs, including ACAGG, AAGGC, AGGGC, and AGAGG expansions, have been identified either in the homozygous or compound heterozygous state with the AAGGG expansion.[22,23] It is now recognized that AAAGG repeats, initially considered non-pathogenic, may cause disease at substantially expanded repeat lengths.[23] The identification of rare compound heterozygous cases with *RFC1* truncating or splicing mutations instead of wild-type $(AAAAG)_{11}$ repeat alongside a single-allele expansion aligns with the recessive inheritance pattern of CANVAS, and suggests a potential loss-of-function mechanism.[24–26] However, no alterations in *RFC1* expression, splicing, or DNA repair function have been correlated with the repeat expansions, thus the precise mechanism by which the *RFC1* repeat expansions causes disease remains unknown.[18,19,26,27]

More recently, this same biallelic expansion was also observed in an Italian patient cohort with chronic idiopathic axonal polyneuropathy.[18–20] These studies suggested biallelic AAGGG repeat expansions as a causative phenomenon isolated to sensory neuropathy, but absent in those with sensory-motor neuropathy. Nevertheless, whether the results of this standalone study is representative of the impact of AAGGG repeats on susceptibility to PN pathogenesis within the broader population, and whether repeats in the heterozygosity state confer additional risk for neuropathy, remains largely unexplored.

Here we describe analysis of STR expansions in a well-characterized U.S. iPN cohort (the Peripheral Neuropathy Research Registry, or PNRR), coupling short-read WGS with confirmatory repeat-primed PCR (RP-PCR). We developed and applied a novel automated pipeline that integrates EHdn and EH to improve the sensitivity and specificity of STR detection. Additionally, we employed unsupervised clustering to robustly classify *RFC1* genotypes and performed case-control burden tests to assess the impact of biallelic and monoallelic repeat expansion. By benchmarking our pipeline's performance against repeat-primed PCR results — the current clinical gold standard for *RFC1* repeat expansion testing in the US. — we aim to not only to validate *RFC1* STR motif detection in the iPN population but also pave the way for broader clinical application of STR-based diagnostics in neuropathy and other neurological disorders.

Methods

iPN Patient Cohort

DNA samples (whole blood, buffy coat, or saliva) were obtained from 816 patients with iPN enrolled in the PNRR at the Johns Hopkins Hospital. The McDonnell Genome Institute (MGI) at Washington University School of Medicine generated short-read WGS data at 30X

coverage for all samples. We conducted extensive quality control (QC), excluding 18 samples that failed either sequencing or alignment and removing 10 samples not sent for PCR analysis, resulting in a final cohort of 788 iPN cases (**Table 1**). We calculated read length, number of sequencing reads, insert size, percentage reads or bases mapped, duplicate rate, and mean error rate based on results from samtools stats (samtools version 1.10), and the mean coverage from samtools depth. The post-alignment QC metrics are shown in **Supplementary Table 1**.

The demographic data for the iPN cohort are detailed in **Table 1**, with an average age at visit of 62.2 years old (standard deviation [SD] = 13.6), a sex ratio slightly skewed toward male (61.8%) over female (38.2%), and a majority White ethnic composition (92.4%). The institutional review boards of Washington University School of Medicine (Protocol no. 202401194), Johns Hopkins University School of Medicine (Protocol no. NA_00047435), and University of Michigan approved the protocols (Protocol no. HUM00244538).

iPN Categorization

For comparison purposes, we classified the iPN cohort into four subcategories. The sensory iPN subcategory (N = 144) includes patients with large and small fiber sensory involvement based on electrophysiology and neurological examination. The SFN subcategory (N = 369) comprises patients with involvement of only small sensory fibers as determined by normal nerve conduction studies but with an abnormal skin biopsy and/or neurological examination. The sensorimotor iPN subcategory (N = 266) consists of patients with a sensorimotor involvement as determined by electrophysiology and neurological examination. Finally, the

sensory and SFN subcategory (N = 513) includes patients with either sensory iPN or SFN (**Table 1**).

Control Cohort

To effectively filter out STRs present in the general population, we incorporated a control cohort matched for ethnic background and age distribution. The control cohort (N = 879) comprises 484 cognitively normal elderly individuals enrolled as controls in the Knight Alzheimer's Disease Research Center (ADRC) at Washington University and 395 unaffected parents aged 40 years or older from the Undiagnosed Mendelian Disease (UMD) study at Washington University. Overall, the control cohort closely resembles the iPN cohort in mean age (63.9 ± 14.8 years) and racial composition (85.9% White) (**Table 1**).

Novel STR Expansion Detection Pipeline based on WGS

Initial STR expansion detection was performed using ExpansionHunter Denovo (EHdn) and ExpansionHunter (EH) on raw FASTQ files generated from patient sample sequencing. These data were aligned to the GRCh38 reference genome using Parabricks *pbrun germline* pipeline (version 4.0.0-1)[28], which utilized BWA-MEM[29] and STAR[30] as the sequencing aligners. We then used EHdn[14] (v0.9.1) to analyze the aligned BAM files and performed genome-wide STR scanning without reliance on predefined catalogs. We utilized ANNOVAR[31] to annotate genes where STRs located, and the RepeatMasker database[32] to annotate known gene-motif pairs. Next, we prioritized motifs detected in the *RFC1* gene and created a *RFC1* STR catalog, which we provided alongside the BAM files to EH[15] (v5.0.0) for further refinement of our STR detection. EH acts not only as a genotyping tool but also an *in silico* validation of EHdn calls. In the final analysis, we retained only STR motifs detected

by both tools, requiring at least 10% consensus in the detected carriers. For each motif, we defined carriers as the union of samples identified by EHdn and those labeled with "INREPEAT" or "SPANNING" calls by EH.

STR Motif Validation and Carriers Identification

To validate the STR motifs identified in the previous step, we generated SAM files for each carrier in the potential STR expansion regions (**Figure 1A**). Here, we defined the repeat window (i.e., region-of-interest, or ROI) as 90 bp upstream and downstream from the *RFC1* repeat region (defined by RepeatMasker annotation) in the reference genome. BLAT (version v.36) was run on every read within the SAM files, generating one PSL file per carrier per STR motif. Next, we initialized an SQLite database (Python 3.6.3, SQLite3 2.6.0) containing all data from the SAM files to enable precise and efficient information retrieval and querying in subsequent steps. An in-house Python script calculated the top three BLAT matches and their corresponding scores, which were integrated as a new database column.

We identified eligible carriers for a specific STR motif by extracting all unique samples listed in the database. For each sample, we retained only reads overlapping the ROI with a minimum mapping quality of one (**Figure 1A**). We then cross-referenced the top BLAT alignment of each read to confirm it fell within the target region.

To quantify STR length, we determined the total length of each read from its CIGAR string and identified the longest consecutive stretch of the motif (allowing one mismatch per motif unit through a Hamming distance approach). All rotational variants of the motif (e.g., ACAG, CAGA, AGAC, GACA) were treated as equivalent. By this process, we calculated the STR

length as the total number of consecutive motif units identified. The algorithm flags any user-defined repeating patterns within the reads that might confound STR analysis. These patterns are normalized to their canonical forms to prevent double counting.

We next quantified STR expansions by comparing each read's STR length to the predetermined threshold. Here, the threshold is defaulted to be 10% of the read length and can be changed by user inputs. Only reads carrying repeats longer than this threshold are considered candidate reads. The proportion of candidate reads relative to the total read count, multiplied by 100, provides the percentage of STR-containing reads for each sample (equation 1). We also monitored the maximum STR length within each sample by tracking the largest continuous STR pattern across all reads. Finally, we calculated the mean STR length by dividing the sum of all STR lengths by the total read count (equation 2). Samples with either a zero percentage of STR-containing reads or a zero mean STR length were classified as non-carriers in the short-read WGS analysis.

$$\%\ Reads\ with\ STR = \frac{Num\ Reads\ with\ (STR\ Length \geq Threshold\ Length)}{Total\ Read\ Num} \times 100\% \quad (1)$$

$$Average\ STR\ Length = \frac{Total\ STR\ Lengths}{Total\ Read\ Num} \quad (2)$$

Post-Analysis Case and Control Carriers Phenotype Confirmation

For cases carrying *RFC1* AAGGG expansion, we checked the iPN subtype, disease history, and other neurologically relevant test results to confirm their clinical phenotypes (see **Results**).

For control analysis, the pipeline initially detected 32 control samples as *RFC1* STR expansion carriers. We obtained the clinical phenotyping information for these carriers and

excluded 10 with conditions associated with peripheral neuropathy, such as diabetes, autoimmune disorders, and neurodegenerative conditions, resulting in 22 final control carriers.

Genotype Determination for STR Carriers

To distinguish between potential biallelic and monoallelic STR expansions among carriers, we conducted three clustering approaches using "% Reads with STR" and "Average STR Length" as input features. Before clustering, we standardized both features using z-score normalization to ensure comparable scales.

We determined the optimal number of clusters (K) using both the elbow and Silhouette methods (R version 4.4.1, NbClust 3.0.1), which consistently indicated an optimal K value of 2. We then applied three clustering algorithms with K = 2 to identify potential biallelic and monoallelic carriers: 1) k-means with multiple 25 random starts to ensure stable cluster assignments; 2) hierarchical clustering using Ward's method (ward.D2) with Euclidean distance that minimizes within-cluster variance; and 3) Gaussian Mixture Models using the Mclust algorithm (mclust 6.1.1). Any sample identified as biallelic by at least one clustering method was considered biallelic, and all remaining samples were classified as monoallelic carriers (**Figure 1A**). In cases of inconsistent results between random seeds or methods, we performed manual review using the Integrative Genomics Viewer (IGV) to confirm biallelic or monoallelic status.

Repeat-primed PCR

Article

Repeat-primed (RP) PCR targeting the repeat locus was performed as previously described to identify carriers of the AAGGG expansion.[18–20] Samples exhibiting a smear or distinct band on the gel underwent further confirmation via capillary electrophoresis conducted by Genewiz Inc. The resulting data files were analyzed using the Peak Scanner Fragment Analysis Software. Samples displaying either a clear smear or thick band in AAGGG RP-PCR were considered candidate carriers. Those showing bands in the AAAAG RP-PCR assay were classified as likely monoallelic carriers, while those without bands were classified as likely biallelic carriers. Fragment analysis was subsequently employed to verify expansion lengths. Long AAGGG expansions were characterized by the presence of distinct sawtooth patterns, whereas short AAGGG expansions displayed peaks without evidence of a sawtooth pattern (**Figure 1B**).

Cases with a positive RP-PCR result for the AAGGG expansion were subsequently assessed for the wild-type allele status through AAAAG RP-PCR. Flanking PCR was conducted for cases requiring additional evaluation. Detailed information on primers and cycling conditions for all PCR and RP-PCR experiments is provided in **Supplementary Table 3**.

Benchmarking of WGS results against RP-PCR

We systematically compared WGS analysis results to those obtained via PCR-based methods. Our WGS pipeline included intermediate EH outputs as well as final pipeline calls, all of which were evaluated alongside PCR data (**Figure 2**). We illustrated the number of biallelic (**Figure 2B**) and monoallelic (**Figure 2C**) carriers detected by RP-PCR (PCR), our computational pipeline (WGS), all EH calls (EH_All), and EH “INREPEAT” calls only (EH_IRR). For samples identified as carriers by only PCR or WGS pipeline, we re-examined

the PCR and fragment analysis results and conducted additional inspection of the short-read WGS data using IGV. During IGV inspection, samples containing at least one correctly-aligned read — confirmed by directly BLATing the read within IGV — and exhibiting more than five consecutive repeat motifs were classified as carriers. Carriers missed by detection pipeline but rescued by IGV inspection were labeled accordingly in **Supplementary Table 2**.

In addition, we compared the genotyping results from WGS analysis and the PCR method. Samples consistently classified as either biallelic or monoallelic carriers were assigned a final genotype directly, whereas samples with discordant results underwent additional IGV inspection and PCR verification. Due to the intrinsic limitations associated with short-read WGS and RP-PCR, some samples continued to show discrepant carrier status or genotype between the two methods. We designated these challenging individuals as 'Edge cases' (**Figure 1C**; **Supplementary Table 2**).

Results

WGS Study and PCR Validation of *RFC1* STR Carriers

The workflow overview for this study is illustrated in **Figure 1**. Consecutively enrolled patients were selected from the PNRR cohort that were seen at the Johns Hopkins Hospital.[33] A clinical diagnosis of iPN was established by examining physicians using a combination of symptoms and signs and exclusionary laboratory testing as defined by the American Academy of Neurology[34]. Only individuals meeting the criteria for iPN, defined as a slowly progressive, symmetric, distal axonal polyneuropathy of unknown cause, were selected for WGS (see **Methods**). Patients with any other confirmed causes of distal symmetric peripheral neuropathies, such as amyloidosis, chronic renal failure, alcohol abuse, vitamin deficiencies, or inherited neuropathies (based on genetic diagnosis or neuropathy in a first-degree family

member) were excluded, as were primary demyelinating neuropathies. The cohort consisted of 788 patients with iPN enrolled in the PNRR, each undergoing 30X Illumina short-read WGS. This cohort was 38.4% female and 92.4% European-ancestry individuals, with an average age at onset of 62.2 years **(Table 1**). A more detailed description of clinical features is available in **Table 2**.

In our WGS analysis, we classified carriers of $(AAGGG)_{exp}$ STR in the *RFC1* gene as either biallelic and monoallelic by applying unsupervised clustering methods on the percentage of reads containing the motif and the average detected STR length (**Method**). To address the absence of comprehensive benchmarking STR detection from short-read sequencing against RP-PCR on a large scale, we performed PCR analysis on the same cohort analyzed by WGS (**Figure 1B**). Combining both WGS and PCR analyses, we identified 128 shared carriers of $(AAGGG)_{exp}$ *RFC1* STRs. Of the shared carriers, 22 iPN patients were classified as biallelic by both methods. Overall, WGS and PCR methods produced matching genotype classifications in 98.2% (774/788) of cases, with a Cohen's kappa coefficient of 0.95, indicating excellent concordance (**Figure 2A**).

Our hybrid workflow, coupled with unsupervised clustering, demonstrated high precision in detecting STRs when compared to standard WGS analysis using EHdn or EH alone. As shown in **Figure 2B-C**, EH results without filtering (EH All) identified 64 biallelic carriers and 715 monoallelic carriers, far exceeding the 22 biallelic and 104 monoallelic detected by both PCR and our workflow, suggesting a high rate of false positives. However, restricting EH results to in-repeat reads only (EH IRR) appeared overly stringent, as it detected only 4 biallelic carriers. In contrast, our workflow, which incorporates results from EH

"INREPEAT" and "SPANNING" calls, eliminated potential false carriers while implementing our in-house genotyping filters and further IGV examination, identified 22 biallelic carriers, closely aligning with PCR results while maintaining a balance between sensitivity and specificity.

After PCR validation, we confirmed the genotype for 126 (16.0%) $(AAGGG)_{exp}$ *RFC1* STR carriers in the entire iPN cohort, of which 22 (2.8%) were biallelic carriers and 104 (13.2%) were monoallelic carriers (**Figure 3**). We observed a significantly higher biallelic carrier frequency in iPN cases than controls (Fisher's exact $p = 5.9x10^{-8}$, odds ratio [OR] = 51.6 (Haldane correction); CI = [3.1 - 852.6]; **Figure 3A**). Stratified analysis revealed the frequency of biallelic carriers was significantly higher in the pure sensory iPN subcategory (6.2%) compared to sensorimotor IPN patients (1.5%; Fisher's exact $p = 1.5x10^{-2}$, OR = 4.3; CI = [1.2 - 19.7]) and pure SFN patients (2.2%; Fisher's exact $p = 2.8x10^{-2}$, OR = 3.0; CI = [1.0 - 9.1]). The elevated biallelic carrier frequency in the sensory subtype aligned with previous reports, but the frequency of biallelic repeats in this population was lower than previously described.[20,35] Interestingly, no significant differences in biallelic carrier frequency were observed between pure SFN and sensorimotor iPN subcategories or between the combined sensory/SFN group and sensorimotor iPN patients.

For monoallelic carriers, we observed significantly higher frequency in iPN cases (13.2%) compared to controls (2.5%; Fishers exact $p = 3.4x10^{-17}$, OR = 5.9; CI = [3.7 - 10.0]; **Figure 3B**). Among sub-cohorts, the SFN group showed the highest frequency of 14.6%, though no significant differences were observed between sensory, SFN, sensorimotor, and sensory/SFN categories. Our control cohort exhibited a lower biallelic carrier frequency (0%) than that

reported in a previous short-read WGS study[36] (0.15%), while monoallelic carriers (2.5%) in our control cohort exhibited a higher frequency than the 0.7% reported by the 1000 Genomes project[37] , but a lower rate than observed in TopMed (5.9%) or the 100K GP (7.2%)[36] – both of which may contain some cases with iPN.

Analysis of *RFC1* AAGGG expansions carrier status for all 788 iPN patients (**Figure 4)** reveals distinct patterns when samples are stratified by zygosity (**Figure 1C, 4B**). Despite the biallelic and monoallelic carriers we have discussed, edge cases constituted 10.0% (N = 14) of all carriers, warranting further classification (**Figure 4B-C**). Among these edge cases, two samples presented persistently discordant genotyping results after multiple rounds of validation through both WGS and PCR methods; thus, they were classified as carriers with discrepant genotyping results. A detailed summary of all classification results is provided in **Supplementary Table 2.**

Compound Heterozygous Deleterious Variants near *RFC1* Gene

Raw FASTQ files of WGS data from the PNRR cohort were processed using Parabricks *pbrun germline* pipeline[28] (version 4.0.0-1), which utilized GATK *HaplotypeCaller*[38] as the variant calling tool to generate VCF files. VCF files were split by chromosome and the following analysis were based on chromosome 4 only (chromosome where *RFC1* gene located).

Variants in VCF files were annotated using snpEff[39] (version 4.3) with multiple reference databases. After initial annotation against the Hg38 reference genome, the Ensembl[40] GRCh38.86 database was used for gene annotation. Variant types were classified using

SnpSift[41], and the functional impact of variants on chromosome 4 was predicted with SIFT4G Annotator. This tool classifies amino acid substitutions as deleterious or tolerated based on sequence homology and physical properties of amino acids, with variants receiving a SIFT score < 0.05 considered potentially deleterious. As a result, no potential deleterious genetic variations were identified in the *RFC1* gene locus.

Clinical Spectrum of Patients Carrying *RFC1* STRs

We next analyzed the clinical characteristics of the confirmed 22 biallelic and 104 monoallelic *RFC1* (AAGGG)exp in comparison to 648 non-carriers (**Table 2**). A two-way chi-squared test (degrees of freedom DOG = 2) revealed a significant difference in the prevalence of sensory iPN ($p < 0.001$), with biallelic carriers showing markedly higher rates (41%) than monoallelic carriers (15%) or non-carriers (18%). No significant differences were observed across the three groups with respect to gender, age, presence of severe pain, sensorimotor iPN, or pure SFN.

Given that AAGGG repeat expansions classically cause CANVAS, which features ataxia and vestibular features, we looked for additional clinical features which might capture subclinical phenotypes that might suggest an *RFC1* expansion when evaluating a patient with idiopathic neuropathy. A previous study reported no association between age at onset and the number of AAGGG repeat units in biallelic carriers.[19] In our study, the average age of onset in biallelic carriers and monoallelic carriers was mid 50s, lower than the average age of the entire cohort (62.2 years old). Several of the SFN patients had a relatively acute onset neuropathy in their 20s and 30s. The most common clinical features in our cohort included distal sensory loss and difficulty in walking due to balance issues.

Article

Discussion

This study represents the largest genetic investigation to date of a single, consistently and thoroughly phenotyped iPN cohort. It is also the first comprehensive comparative analysis of two complementary molecular approaches: WGS and RP-PCR. We developed an integrated pipeline that utilizes EHdn and EH algorithms for *RFC1* STR detection from short-read WGS, achieving 98.2% agreement with PCR-based validation results. Additionally, this pipeline incorporates unsupervised clustering methods to accurately classify patients' *RFC1* genotypes as biallelic carriers or monoallelic carriers.

Using this pipeline, we observed that biallelic $(AAGGG)_{exp}$ in *RFC1* accounted for 2.8% of iPN cases and was absent in controls. The frequency of biallelic carriers was significantly higher among patients with pure sensory iPN (6.2%). However, compared to prior findings by Curro et al.[18], we observed a much lower overall rate of biallelic expansions in pure sensory iPN (6.2% versus 34%). The lower overall rate of biallelic expansions in iPN we observe is largely consistent with unpublished clinical observations of several large subspecialty clinics focused on peripheral neuropathy. The reason for this discrepancy is unclear. Previous studies focused primarily on patients from England or Italy, and expansions do appear more prevalent in European populations.[36] However, the PNRR cohort is heavily biased towards Caucasians and thus likely reflects a genetic background similar to prior European studies. Differences in selection and screening criteria and repeat assessment methodology as well as the smaller size of the Italian cohort, which included only 125 cases, may also explain the discrepancy.[18] Similarly, identification of biallelic *RFC1* repeat expansions in SFN and sensorimotor iPN patients in our large cohort differs from observations in smaller and ataxia focused cohorts, where a prior study[18] reported no biallelic carriers among 100 patients with sensorimotor iPN. In contrast, our larger cohort revealed

biallelic *RFC1* expansions in 1.5% of such patients. Thus, we suggest that the presence of motor nerve involvement should not preclude patients from undergoing *RFC1* repeat screening.

Interestingly, monoallelic *RFC1* $(AAGGG)_{exp}$ showed a significantly higher prevalence in our iPN cohort (13.2%) compared to controls (2.5%; $p = 3.4x10^{-17}$), suggesting that a single expanded *RFC1* allele may confer an increased risk for developing iPN. Consistent with this notion, 18.8% of alleles in the iPN cohort carried *RFC1* expansions, which substantially exceeded the 7.3% reported by Kristina et al. in a large multi-ethnic disease cohort from TOPMed and the 100K Genomes Project.[36] Prior studies suggest that some CANVAS cases have compound heterozygous mutations- with a AAGGG repeat on one allele and a protein-altering mutation on the other allele. However, we observed no such potential damaging missense, loss-of-function variants, or small insertions/deletions near the *RFC1* region in our monoallelic *RFC1* carriers or in our controls. We did observe short (typically 2-5 copies) alternative repeat sequence elements within the other allele in 17 out of 104 (16%) of our monoallelic carriers that might provide an alternative explanation for the increased frequency in the iPN population. However, all such cases had AAAAG repeat alleles as measured by repeat-primed PCR and thus met our criteria for monoallelic carrier status (**Supplementary Figure 1**).

This finding of enriched monoallelic cases in iPN has implications for our understanding of how non-reference AAGGG repeat expansions might elicit disease. If the repeat elicits disease through a loss of function mechanism, this would suggest that even haploinsufficiency for *RFC1* may enhance the risk of developing neuropathy. Alternatively,

heterozygous AAGGG repeat expansions may induce a dose-dependent gain of function toxicity that would induce more fulminant symptoms and greater penetrance in homozygosity.[26] This latter model would be consistent with recent studies in CANVAS patient derived glutamatergic iNeurons, where selective removal of the repeat suppressed disease relevant phenotypes, but reprovision of RFC1 protein did not.[26] Future work will be needed to more clearly ascertain the impact of heterozygous mutations on peripheral sensory neuronal dysfunction and disease pathogenesis.

There are several cases in our cohort of monoallelic *RFC1* repeat expansion in young individuals with relatively acute onset SFN. Although the presentation of these patients resembled monophasic acute-onset post-infectious SFN, they lacked a history of prior infection, and their neuropathy persisted for many years. This finding suggests that acute onset SFN patients who do not recover in a timely manner should be evaluated for *RFC1* repeat expansions. It also raises the possibility that monoallelic expansions may serve as a risk factor for poor recovery from peripheral nerve injury or neurotoxic agents.

Comparison between short-read WGS with PCR revealed excellent concordance for most cases. Aside from pure AAGGG repeat carriers, we found some carriers with complex repeat structures, such as mixtures of (AAGGG)exp, (AAAGG)exp, and (ACAGG)exp (**Supplementary table 2**). These cases posed classification challenges for PCR, which typically targets only a few predefined motifs. Our WGS pipeline identified alternative repeat motifs that might be missed by PCR targeted to a fixed set of one or two motifs, indicating its utility for comprehensive screening. Our automated pipeline offers significant practical advantages for clinical implementation. The in-house pipeline, written in Python, Bash, and R, can be easily implemented in clinical settings by personnel with basic bioinformatic

training. While long-read sequencing offers a more definitive characterization of STR expansions, its cost and technical requirements currently limit routine use. Our approach provides a practical, cost-effective alternative for detecting *RFC1* expansions, with WGS complementing PCR by flagging cases for targeted follow-up validation. We anticipate this tool will be broadly adopted for both research and clinical applications, facilitating more accurate and efficient genetic analyses in the field of human and clinical genetics. By making our pipeline publicly available through the cloud-based platform (GitHub), we aim to facilitate broader adoption of this methodology in both research and clinical settings, providing additional diagnostic yield and deepening the understanding of the genetic basis of PN. Ultimately, we hope to facilitate more accurate and efficient genetic analyses in the field of human and medical genetics.

This study has some limitations. Both WGS and PCR methods face challenges in precisely quantifying repeat number, particularly in the presence of multiple repeat motifs. Classifying carriers can be particularly challenging when reads are interrupted by various motif combinations, since capture success for both PCR and WGS depends on motif composition and regional complexity. While long-read sequencing would address these shortcomings, it remains prohibitively expensive for large-scale use and is therefore best reserved for cases where WGS and PCR cannot resolve allelic information or when precise measurement of repeat length is required.

In conclusion, our findings underscore the diagnostic value of *RFC1* expansion screening in iPN patients, where biallelic expansions account for roughly 2.8% of cases, and monoallelic expansions occur at a significantly higher rate in iPN patients versus controls. Overall, our data support testing for *RFC1* repeat expansions in patients with iPN. Moreover, our finding

that monoallelic expansions may serve as a risk allele for the development of peripheral neuropathy has important implications for how we counsel our patients both before and after testing. We are hopeful that a deeper understanding of genotype-phenotype relationships will impact our ability to resolve the pathogenic mechanisms underlying CANVAS and peripheral neuropathy as a whole – providing a way forward for developing therapies for these currently untreatable conditions.

Acknowledgements

We would like to thank the patients, their families, and the Foundation of Peripheral Neuropathy for their support of the PNRR and their willingness to participate in research. We appreciate obtaining access to genetic and phenotypic data from the participants. We also thank the PNRR Study Group members at Johns Hopkins who enroll patients in the registry. These include neuromuscular physicians Sarah Berth, Vinay Chaudhry, David Cornblath, Leana Doherty, Lindsey Hayes, Hristelina Ilieva, Thomas Lloyd, Mohammad Khoshnoodi, Brett McCray, Brett Morrison, Bipasha Mukherjee-Clavin, Lyle Ostrow, Michael Polydefkis, Ricardo Roda, and Charlotte Sumner.

We thank the Genome Technology Access Center at the McDonnell Genome Institute at Washington University in St. Louis for DNA sequencing.

We are grateful to all of the families participating in the Undiagnosed Mendelian Disorders (UMD) Research Project at Washington University in St. Louis, as well as the principal investigator and team members of the UMD project (M. Shinawi, D. Baldridge, C. Gurnett, T. Turner, B. Leon Ricardo). We appreciate obtaining access to genetic and phenotypic data from the UMD project.

Article

PNRR is supported by the Foundation for Peripheral Neuropathy. SCJ is supported by the Hydrocephalus Association Innovator Award, Cerebral Palsy Alliance Research Foundation Project Grant (PRG03121), WashU Children's Discovery Institute Faculty Scholar Award (CDI-FR-2021-926), and NIH U19NS130607, R01NS111029, and R01NS131610. AH is supported by NIH (R21NS135481, P30 MH075673-011), Wellcome Trust, Dr. Miriam and Sheldon G. Adelson Medical Research Foundation, and Dr. Richard Merkin Family Foundation. JM is supported by the NIH U19NS130607 and the Bob and Signa Hermann Fund for Neuropathy Research. PKT and SSO's work on this project was supported by the NIH (R21NS129096) and the Taubman Institute at the University of Michigan.

Author Contributions

Z.T., S.C.J., P.K.T., A.H., J.M., S.S.O. contributed to the conception and design of the study; Z.T., S.S.O., S.C.J., P.K.T., A.H., J.M., B.C., E.M., S.T., J.U., Z.L., D.B., C.C., M.J. contributed to the acquisition and analysis of data; Z.T. S.S.O., B.C., A.H., P.K.T., S.C.J. contributed to the interpretation of studies included in the manuscript; Z.T., S.C.J., S.S.O. contributed to drafting the manuscript or preparing the figures and tables; P.K.T., A.H., B.C., Z.T., S.S.O., J.M., E.M., S.T., J.U., Z.L, D.B., C.C., M.J.,S.C.J. contributed to manuscript editing and approval of the final draft.

Potential Conflicts of Interests

PKT served as a consultant to Denali Therapeutics and shares a patent on an ASO with Ionis Pharmaceuticals. Neither of these relationships are directly relevant to this manuscript.

AH served as consultant to Pfizer, Sangamo and Sanofi. None of these relationships are directly relevant to this manuscript.

Article

JM is co-founder of Disarm Therapeutics, a division of Eli Lilly, and member of SAB of Asha Therapeutics. Neither relationship is directly relevant to this manuscript.

Data Availability

All raw genomic data supporting the findings of this study have been deposited in the Database of Genotypes and Phenotypes (dbGaP) under the accession number phs003788.v1.p1.The automated STR pipelines and associated codes for patient classification and statistical analysis are available on GitHub at https://github.com/jinlab-washu.

Article

References

1. Lehmann HC, Wunderlich G, Fink GR, Sommer C. Diagnosis of peripheral neuropathy. Neurol. Res. Pract. 2020;2(1):20.

2. Callaghan B, Kerber K, Langa KM, et al. Longitudinal patient-oriented outcomes in neuropathy. Neurology 2015;85(1):71–79.

3. Singer MA, Vernino SA, Wolfe GI. Idiopathic neuropathy: new paradigms, new promise. J Peripher Nerv Syst 2012;17 Suppl 2:43–49.

4. Senderek J, Lassuthova P, Kabzińska D, et al. The genetic landscape of axonal neuropathies in the middle-aged and elderly: Focus on *MME*. Neurology 2020;95(24):e3163–e3179.

5. PEPAN - Overview: Comprehensive Peripheral Neuropathy Gene Panel, Varies [Internet]. [date unknown];[cited 2023 Dec 4] Available from: https://www.mayocliniclabs.com/test-catalog/Overview/617688

6. Gymrek M. A genomic view of short tandem repeats. Curr Opin Genet Dev 2017;44:9–16.

7. Gall-Duncan T, Sato N, Yuen RKC, Pearson CE. Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences. Genome Res 2022;32(1):1–27.

8. Deynze KV, Mumm C, Maltby CJ, et al. Enhanced Detection and Genotyping of Disease-Associated Tandem Repeats Using HMMSTR and Targeted Long-Read Sequencing [Internet]. 2024;2024.05.01.24306681.[cited 2024 Sep 9] Available from: https://www.medrxiv.org/content/10.1101/2024.05.01.24306681v1

9. Fan H, Chu J-Y. A Brief Review of Short Tandem Repeat Mutation. Genomics Proteomics Bioinformatics 2007;5(1):7–14.

10. Zegura SL. Human Evolutionary Genetics: Origins, Peoples and Disease. Am J Hum Genet 2005;76(6):1087–1088.

11. Malik I, Kelley CP, Wang ET, Todd PK. Molecular mechanisms underlying nucleotide repeat expansion disorders. Nat Rev Mol Cell Biol 2021;22(9):589–607.

12. Willems T, Gymrek M, Highnam G, et al. The landscape of human STR variation. Genome Res 2014;24(11):1894–1904.

13. Mallick S, Li H, Lipson M, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 2016;538(7624):201–206.

14. Dolzhenko E, Bennett MF, Richmond PA, et al. ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. Genome Biol 2020;21(1):102.

Article

15. Dolzhenko E, Deshpande V, Schlesinger F, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics 2019;35(22):4754–4756.

16. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res 2019;47(15):e90.

17. Tankard RM, Bennett MF, Degorski P, et al. Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. The American Journal of Human Genetics 2018;103(6):858–873.

18. Cortese A, Tozza S, Yau WY, et al. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome due to RFC1 repeat expansion. Brain 2020;143(2):480–490.

19. 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):649–658.

20. Currò R, Salvalaggio A, Tozza S, et al. *RFC1* expansions are a common cause of idiopathic sensory neuropathy. Brain 2021;144(5):1542–1550.

21. Currò R, Dominik N, Facchini S, et al. Role of the repeat expansion size in predicting age of onset and severity in RFC1 disease. Brain 2024;147(5):1887–1898.

22. Scriba CK, Beecroft SJ, Clayton JS, et al. A novel RFC1 repeat motif (ACAGG) in two Asia-Pacific CANVAS families. Brain 2020;143(10):2904–2910.

23. Dominik N, Magri S, Currò R, et al. Normal and pathogenic variation of RFC1 repeat expansions: implications for clinical diagnosis. Brain 2023;146(12):5060–5069.

24. Ronco R, Perini C, Currò R, et al. Truncating Variants in *RFC1* in Cerebellar Ataxia, Neuropathy, and Vestibular Areflexia Syndrome [Internet]. Neurology 2023;100(5)[cited 2025 Apr 1] Available from: https://www.neurology.org/doi/10.1212/WNL.0000000000201486

25. Weber S, Coarelli G, Heinzmann A, et al. Two RFC1 splicing variants in CANVAS. Brain 2023;146(3):e14–e16.

26. Maltby CJ, Krans A, Grudzien SJ, et al. AAGGG repeat expansions trigger RFC1-independent synaptic dysregulation in human CANVAS neurons. Science Advances 2024;10(36):eadn2321.

27. Martelli A, Napierala M, Puccio H. Understanding the genetic and molecular pathogenesis of Friedreich's ataxia through animal and cellular models. Dis Model Mech 2012;5(2):165–176.

28. NVIDIA. Clara Parabricks [Internet]. 2025;Available from: https://www.nvidia.com/en-us/clara/genomics/

29. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. 2013;[cited 2025 Feb 27] Available from: http://arxiv.org/abs/1303.3997

Article

30. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29(1):15–21.

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

32. Smit A, Hubley R, Green P. RepeatMasker Open-4.0 [Internet]. 2013;Available from: http://www.repeatmasker.org

33. Thomas S, Ajroud-Driss S, Dimachkie MM, et al. Peripheral Neuropathy Research Registry: A prospective cohort. Journal of the Peripheral Nervous System 2019;24(1):39–47.

34. England JD, Gronseth GS, Franklin G, et al. Practice Parameter: Evaluation of distal symmetric polyneuropathy: Role of laboratory and genetic testing (an evidence-based review). Neurology 2009;72(2):185–192.

35. Tagliapietra M, Cardellini D, Ferrarini M, et al. RFC1 AAGGG repeat expansion masquerading as Chronic Idiopathic Axonal Polyneuropathy. J Neurol 2021;268(11):4280–4290.

36. Ibañez K, Jadhav B, Zanovello M, et al. Increased frequency of repeat expansion mutations across different populations. Nat Med 2024;30(11):3357–3368.

37. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. Nature 2015;526(7571):68–74.

38. Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples [Internet]. 2018;201178.[cited 2025 Apr 4] Available from: https://www.biorxiv.org/content/10.1101/201178v3

39. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118}; iso-2; iso-3. Fly 2012;6(2):80–92.

40. Harrison PW, Amode MR, Austine-Orimoloye O, et al. Ensembl 2024. Nucleic Acids Research 2024;52(D1):D891–D899.

41. Cingolani P, Patel VM, Coon M, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift [Internet]. Front. Gene. 2012;3[cited 2025 Apr 4] Available from: http://journal.frontiersin.org/article/10.3389/fgene.2012.00035/abstract

Article

Figure Legends

Figure 1. Methodological workflow for RFC1 repeat expansion detection. (**A**) Whole genome sequencing (WGS) bioinformatic analysis pipeline for identifying potential *RFC1* repeat expansion carriers. (**B**) PCR-based workflow for detecting, genotyping, and classifying *RFC1* AAGGG expansion status. (**C**) Reconciliation process for samples with discordant carrier status between WGS and PCR methods.

Figure 2. Comparison of *RFC1* AAGGG carriers across different detection methods. (**A**) Concordance heatmap to show the number of biallelic carriers, monoallelic carriers, and non-carriers detected by PCR and WGS methods. The number inside each grid represents the number of carriers with corresponding classifications. Larger numbers were labeled in darker blue, whereas smaller numbers were shown in lighter yellow. After the comparison, we regarded all carriers that received different classification by the two methods as discordant cases, which then were evaluated separately by further experimental or manual checks. Upset plots ((**B**) Biallelic carriers; (**C**) Monoallelic carriers) showing the number of carriers identified in EH in-repeat results, WGS workflow, PCR workflow, and unfiltered EH result, ordered from the largest to the smallest group. small circles at the bottom label the method(s) used to detect each set of carriers. The total number of carriers (i.e., set size) for each method appears in the bottom-left panel.

Figure 3. *RFC1* AAGGG carrier frequency distribution across iPN subtypes. (**A**) Biallelic AAGGG carrier prevalence showing significantly higher frequency in pure sensory neuropathy (6.2%) compared to other subtypes ($p<0.05$), with no carriers identified in the control group. (**B**) Monoallelic AAGGG carrier frequency was significantly higher in IPN

cases (13.2%) versus controls (2.5%) (p=3.4e-17), with no statistically significant differences across IPN subtypes.

*1 motor IPN carrier didn't show in this figure due to motor IPN sub-cohort size too small.

Figure 4. Distribution of final *RFC1* AAGGG carrier status distribution in the study cohort. (**A**) Overall carrier status showing 17.8% (N=140) carriers and 82.2% (N=648) non-carriers among IPN cases. (**B**) Distribution by expansion type among carriers (N=140): biallelic carriers, monoallelic carriers, and edge cases. (**C**) Classification of edge cases (N=14): carriers with discrepant genotyping results from WGS and PCR, carriers missed by WGS, and missed by PCR.

Article

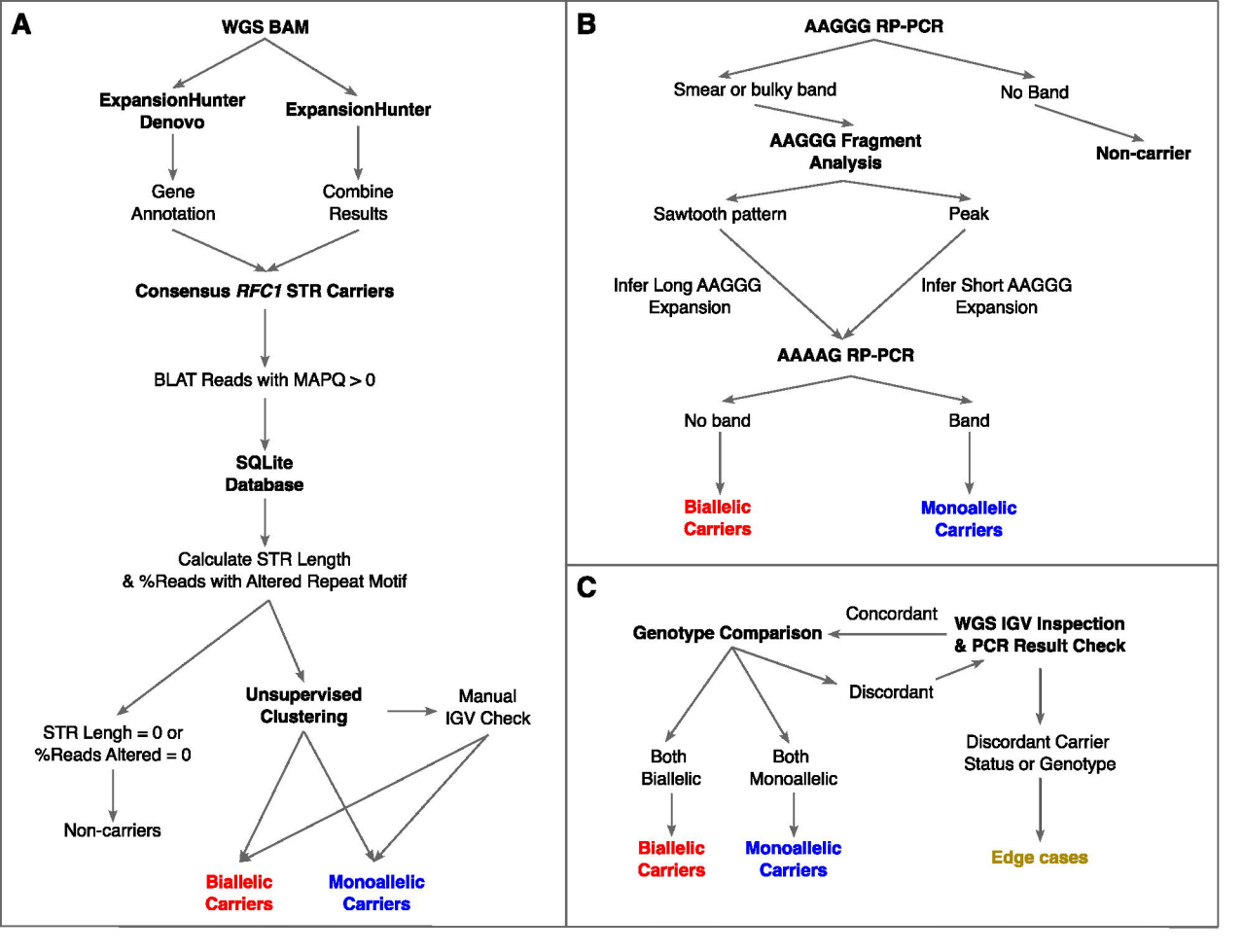
Tables

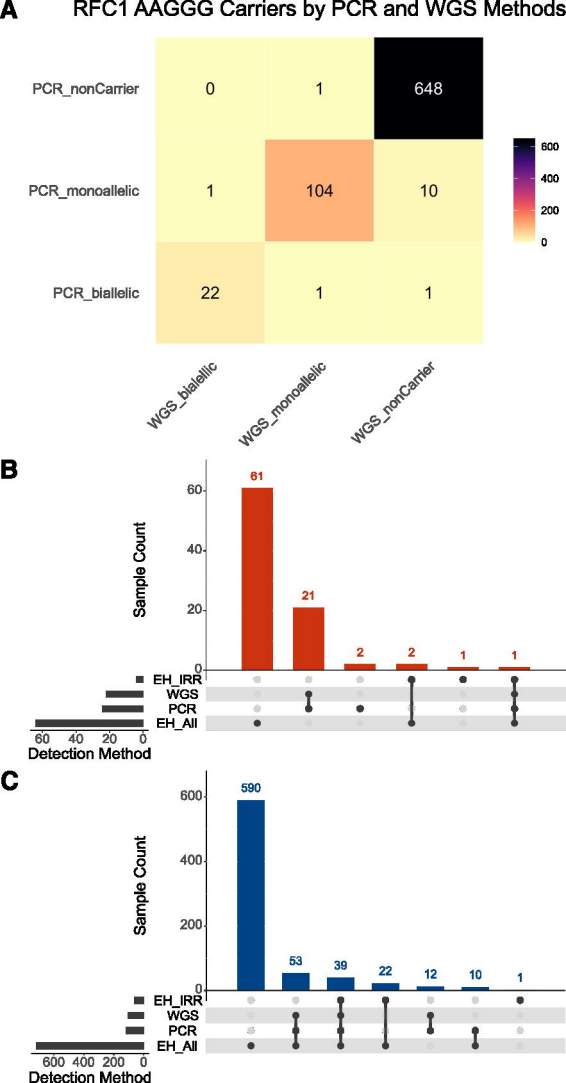
Table 1. Demographic and clinical characteristics of study participants. A total of 788 IPN cases and 879 controls were included. Demographic data show comparable age and gender distribution between IPN cases and controls. IPN cases were predominantly classified into four sub-cohorts: small fiber neuropathy only (46.8%), sensorimotor (33.8%), sensory only (18.3%), and motor only (1.1%).

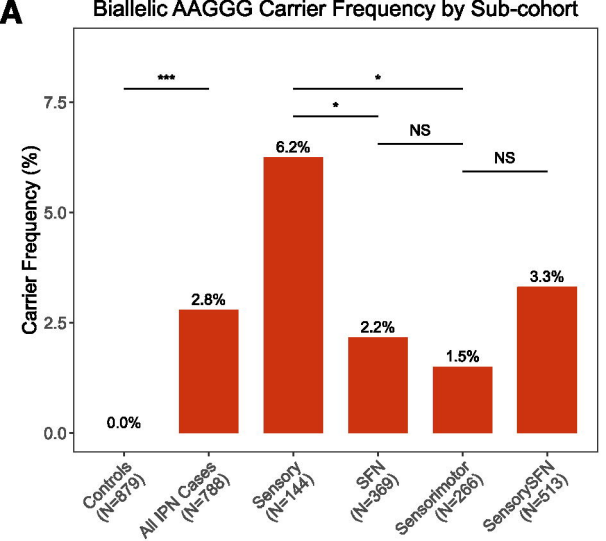
	iPN Cases	**Controls**
Total Sample Size	788	879
Age	62.2 ± 13.8	63.9 ± 14.8
Gender		
Male	487 (61.8%)	406 (46.2%)
Female	301 (38.2%)	473 (53.8%)
Ethnicity		
White	741 (92.9%)	755 (85.9%)
Black/African American	33 (4.1%)	104 (11.8%)
Asian	10 (1.3%)	7 (0.8%)
American Indian/Alaska Native	1 (0.1%)	9 (1.0%)
More than one race	7 (0.9%)	4 (0.5%)
Not Known	6 (0.7%)	0 (0%)
iPN Sub-cohorts		
Small Fiber Neuropathy Only	369 (46.8%)	/
Sensorimotor	266 (33.8%)	/
Sensory Only	144 (18.3%)	/
Motor Only	9 (1.1%)	/

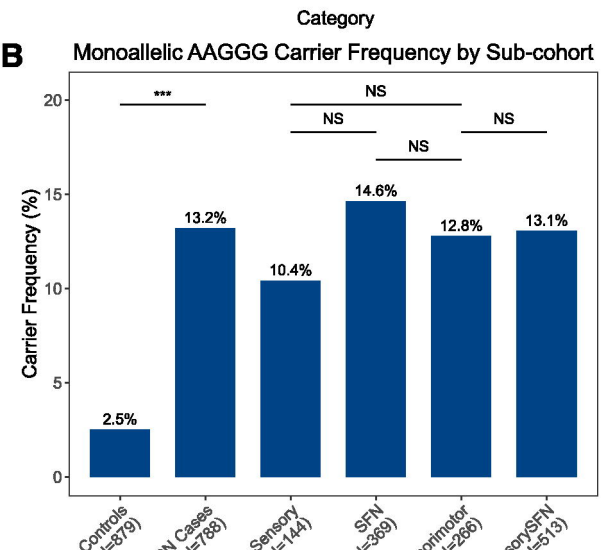
Table 2. Clinical characteristics of confirmed biallelic and monoallelic *RFC1* AAGGG carriers versus non-carriers in IPN cases. Chi-squared tests were conducted for gender, age at visit, painful status, and percentage carriers in each subtype across all groups. Pure sensory iPN was significantly more prevalent in biallelic carriers (41%) compared to monoallelic carriers (15%) and non-carriers (18%) ($p<0.001$), while age at visit, gender distribution, and painful neuropathy prevalence were similar across groups (NS).

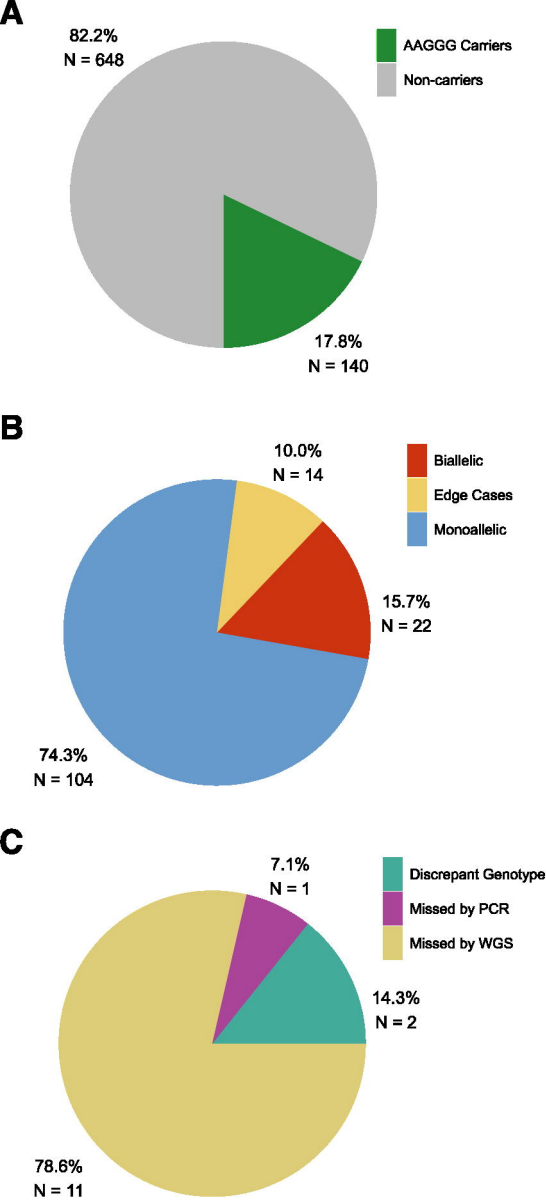
	Biallelic (N = 22)	**Monoallelic (N = 104)**	**Non-Carriers (N = 648)**	**P-value**
Male	12 (55%)	62 (60%)	405 (63%)	NS
Age at Visit	63 ± 13	62 ± 14	62 ± 14	NS
Painful neuropathy	16 (73%)	74 (71%)	469 (72%)	NS
Sensory neuropathy	9 (41%)	15 (15%)	117 (18%)	<0.001
Sensorimotor	4 (18%)	34 (33%)	222 (34%)	NS
Small fiber neuropathy	8 (36%)	54 (52%)	302 (47%)	0.048


A
WGS BAM
ExpansionHunter Denovo
ExpansionHunter
Gene Annotation
Combine Results
Consensus RFC1 STR Carriers
BLAT Reads with MAPQ > 0
SQLite Database
Calculate STR Length & %Reads with Altered Repeat Motif
STR Lengh = 0 or %Reads Altered = 0
Non-carriers
Unsupervised Clustering
Manual IGV Check
Biallelic Carriers
Monoallelic Carriers
B
AAGGG RP-PCR
Smear or bulky band
No Band
Non-carrier
AAGGG Fragment Analysis
Sawtooth pattern
Peak
Infer Long AAGGG Expansion
Infer Short AAGGG Expansion
AAAAG RP-PCR
No band
Band
Biallelic Carriers
Monoallelic Carriers
C
Genotype Comparison
Concordant
WGS IGV Inspection & PCR Result Check
Discordant
Both Biallelic
Both Monoallelic
Discordant Carrier Status or Genotype
Biallelic Carriers
Monoallelic Carriers
Edge cases


A
RFC1 AAGGG Carriers by PCR and WGS Methods
PCR_nonCarrier
PCR_monoallelic
PCR_biallelic
0
1
648
1
104
10
22
1
1
600
400
200
0
WGS_bialelic
WGS_monoallelic
WGS_nonCarrier
B
Sample Count
61
21
2
2
1
1
EH_IRR
WGS
PCR
EH_All
Detection Method
C
Sample Count
590
53
39
22
12
10
1
EH_IRR
WGS
PCR
EH_All
Detection Method


A
Biallelic AAGGG Carrier Frequency by Sub-cohort
Carrier Frequency (%)
7.5
5.0
2.5
0.0
***
*
*
NS
NS
0.0%
2.8%
6.2%
2.2%
1.5%
3.3%
Controls (N=879)
All IPN Cases (N=788)
Sensory (N=144)
SFN (N=369)
Sensorimotor (N=266)
SensorySFN (N=513)
Category


B
Monoallelic AAGGG Carrier Frequency by Sub-cohort
Carrier Frequency (%)
20
15
10
5
0
***
NS
NS
NS
NS
2.5%
13.2%
10.4%
14.6%
12.8%
13.1%
Controls (N=879)
All IPN Cases (N=788)
Sensory (N=144)
SFN (N=369)
Sensorimotor (N=266)
SensorySFN (N=513)
Category


A
82.2%
N = 648
AAGGG Carriers
Non-carriers
17.8%
N = 140
B
10.0%
N = 14
Biallelic
Edge Cases
Monoallelic
15.7%
N = 22
74.3%
N = 104
C
7.1%
N = 1
Discrepant Genotype
Missed by PCR
Missed by WGS
14.3%
N = 2
78.6%
N = 11