


# Homozygous *RFC1* AAGGG Repeat Expansions Are Common in Idiopathic Peripheral Neuropathy

Zitian Tang, BS <sup>1†</sup>, Sinem S. Ovunc, MD,<sup>2†</sup> Ryo Iwase, MD, PhD <sup>2</sup>, Elle Mehinovic, MSc,<sup>1</sup> Simone Thomas, MSc <sup>3</sup>, Jenna Ulibarri, BS,<sup>1</sup> Zefan Li, BS <sup>1</sup>, Dustin Baldrige, MD, PhD <sup>4</sup>, Carlos Cruchaga, PhD <sup>5</sup>, Menghan Liu, MSc <sup>5</sup>, Matt Johnson, MA, MSc <sup>5</sup>, Jeffrey Milbrandt, MD, PhD <sup>1,6,7</sup>, Brian Callaghan, MD, MSc <sup>2</sup>, PNRR Study Group,<sup>3</sup> Ahmet Höke, MD, PhD <sup>3</sup>, Peter K. Todd, MD, PhD <sup>2,8</sup> and Sheng Chih Jin, PhD <sup>1,4</sup>

**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 homozygous (biallelic) *RFC1* expansions in more diverse patient populations remains unclear—partly due to the absence of accurate, user-friendly computational pipelines 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 polymerase chain reaction (PCR)-based confirmation in a large, well-characterized US cohort consisting of 788 patients with iPN (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 97.2% concordance with repeat-primed PCR-based validation.

**Results:** Biallelic *RFC1* expansions were present in only 1 out of 778 controls but present in 18 out of 788 (2.3%) patients with iPN (Fisher's exact  $p = 7 \times 10^{-5}$ ), including 6.9% (10/144) of pure sensory, 1.1% (4/369) of SFN, and 1.5% (4/266) of sensorimotor neuropathy. These data indicate that motor nerve involvement should not exclude patients from *RFC1* repeat screening. Monoallelic expansions were observed at a nominally higher frequency in iPN than in controls (9.1% vs 7.2%), but this difference did not reach statistical significance (Fisher's exact  $p = 0.17$ ). We also found no evidence of second mutations or expansions on the other allele among monoallelic carriers.

**Interpretation:** Our approach provides a robust, cost-effective method for detecting *RFC1* expansions from WGS data. Our findings indicate that homozygous (biallelic) AAGGG repeat expansions in *RFC1* contribute to development of iPN. Heterozygous expansions may also confer disease risk, but future studies are needed to assess this observation and explore any phenotypic differences with biallelic cases.

ANN NEUROL 2026;00:1–14

View this article online at [wileyonlinelibrary.com](https://www.wileyonlinelibrary.com). DOI: 10.1002/ana.78226

Received Apr 11, 2025, and in revised form Feb 26, 2026. Accepted for publication Mar 19, 2026.

Address correspondence to Dr Jin, Genetics and Pediatrics, Washington University School of Medicine, 4515 McKinley Ave., St. Louis, MO, 63110. E-mail: [jin810@wustl.edu](mailto:jin810@wustl.edu); Dr Todd, Department of Neurology, University of Michigan, 4148 BSRB, 109 Zina Pitcher Place, Ann Arbor, MI 48109-2200. E-mail: [petertod@med.umich.edu](mailto:petertod@med.umich.edu); Dr Höke, Johns Hopkins School of Medicine, 855 N. Wolfe St., Rangos 248, Baltimore, MD 21205. E-mail: [ahoke1@jh.edu](mailto:ahoke1@jh.edu)

<sup>†</sup>These authors contributed equally to the work presented here and should therefore be regarded as co-first authors.

From the <sup>1</sup>Department of Genetics, Washington University School of Medicine, St. Louis, MO; <sup>2</sup>Department of Neurology, University of Michigan, Ann Arbor, MI; <sup>3</sup>Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD; <sup>4</sup>Department of Pediatrics, Washington University School of Medicine, St. Louis, MO; <sup>5</sup>NeuroGenomics and Informatics Center, Washington University School of Medicine in St. Louis, St. Louis, MO; <sup>6</sup>Needleman Center for Neuro-metabolism and Axonal Therapeutics, St. Louis, MO; <sup>7</sup>McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO; and <sup>8</sup>Veterans Affairs Medical Center, Ann Arbor, MI

Additional supporting information can be found in the online version of this article.

Peripheral neuropathy (PN) is among the most common neurological disorders, with an estimated prevalence of 12 to 20% in the United States and up to 30% in older adults.<sup>1</sup> 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.<sup>2</sup> Approximately one-third of all PN cases are classified as idiopathic PN (iPN), wherein no definitive etiology can be determined despite thorough and appropriate clinical evaluation.<sup>3</sup> As upward of 18% of late-onset iPN may be explained by genetic causes,<sup>4</sup> neurology clinics increasingly implement gene-panel based testing for potential pathogenic single nucleotide variants (SNVs) and copy number variations (CNVs).<sup>5</sup> Despite these efforts, most patients with iPN 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 approximately 8% of the human genome.<sup>6–8</sup> 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.<sup>6,9,10</sup> Genomic instability caused by pathogenic STR expansion contributes to over 70 human disorders, many of which are neurological.<sup>11</sup> Recent large-scale initiatives such as the 1000 Genomes Project<sup>12</sup> and the Simons Genome Diversity Project<sup>13</sup> 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.<sup>12,13</sup> Nonetheless, current STR detection software—such as ExpansionHunter Denovo (EHdn),<sup>14</sup> ExpansionHunter (EH),<sup>15</sup> GangSTR,<sup>16</sup> and exSTRa<sup>17</sup>—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 key contributor to PN pathogenesis. In 2019, a biallelic non-reference AAGGG repeat expansion in the second intron of the Replication Factor C Subunit 1 (*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.<sup>18–20</sup> *RFC1* encodes the principal subunit of the replication factor C complex,

essential for DNA polymerases and repair through ATP-dependent loading of proliferating cell nuclear antigen onto DNA strands.<sup>21,22</sup> Despite its ubiquitous expression and critical role in genomic integrity maintenance, particularly in DNA damage recognition and repair pathways, the pathological mechanism linking *RFC1* expansions to disease remains unclear.<sup>21,22</sup> Pathogenic alleles typically involve replacement of the wild-type (AAAAG)<sub>11</sub> repeat motif with over 250 repeats of AAGGG.<sup>19,23</sup> Since the original discovery, additional pathogenic repeat motifs, including ACAGG, AAGGC, AGGGC, and AGAGG, have been identified either in the homozygous or compound heterozygous state with the AAGGG expansion.<sup>24,25</sup> It is now recognized that AAAGG repeats, initially considered non-pathogenic, may cause disease at substantially expanded repeat lengths.<sup>25</sup> The identification of rare compound heterozygous cases with *RFC1* truncating or splicing mutations instead of wild-type (AAAAG)<sub>11</sub> repeat alongside a single-allele expansion aligns with the recessive inheritance pattern of CANVAS, and suggest a potential loss-of-function mechanism.<sup>26–28</sup> However, to date, 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.<sup>18,19,28,29</sup>

More recently, this same biallelic expansion was also observed in an Italian patient cohort with chronic idiopathic axonal polyneuropathy.<sup>18–20</sup> 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 are 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 US iPN cohort (the Peripheral Neuropathy Research Registry [PNRR]), coupling short-read WGS with confirmatory repeat-primed polymerase chain reaction (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 used unsupervised clustering to robustly classify *RFC1* genotypes to assess the impact of biallelic AAGGG expansions and the potential risk associated with monoallelic expansion carriers. By benchmarking our pipeline's performance against repeat-primed PCR results — the current clinical gold standard for *RFC1* repeat expansion testing in the United States — 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 at Washington University School of Medicine generated short-read WGS data at 30X Illumina short-read 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),<sup>30</sup> and the mean coverage from samtools depth. The post-alignment QC metrics are shown in Supplementary Table S1.

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] = 13.8), a sex ratio slightly skewed toward male patients (487/788, 61.8%) over female patients (301/788, 38.2%), and a majority of White ethnic composition (731/788, 92.8%). 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 (Protocol no. HUM00244538) approved the protocols.

### *iPN Categorization*

For comparison purposes, we classified the iPN cohort into 4 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 motor iPN subcategory (N = 9) includes patients with motor involvement (see Table 1).

**TABLE 1. Demographic and Clinical Characteristics of Study Participants**

	iPN Cases	Controls
Total sample size	788	778
Age	62.2 ± 13.8	63.1 ± 14.7
Sex		
M	487 (61.8%)	368 (47.3%)
F	301 (38.2%)	410 (52.7%)
Ethnicity		
White	731 (92.8%)	705 (90.6%)
Black/African American	33 (4.2%)	69 (8.9%)
Asian	10 (1.3%)	2 (0.3%)
More than one race	7 (0.9%)	2 (0.3%)
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%)	/

A total of 788 iPN cases and 854 controls were included. Demographic data show comparable age and sex distribution between iPN cases and controls. The iPN cases were predominantly classified into 4 sub-cohorts: small fiber neuropathy only (46.8%), sensorimotor (33.8%), sensory only (18.3%), and motor only (1.1%).

iPN = idiopathic peripheral neuropathy.

### *Control Cohort and Exclusion Criteria*

To obtain an unbiased estimate of *RFC1* expansion carrier frequency in the North American population and to determine whether *RFC1* repeat expansions are enriched in our patients with iPN, we incorporated control cohorts matched for ethnic background and age distribution. We started from two control cohorts: (i) cognitively normal elderly individuals enrolled as controls in the Knight Alzheimer's Disease Research Center (ADRC) at Washington University (N = 484), and (ii) unaffected parents aged ≥40 years from the Undiagnosed Mendelian Disorders (UMD) study at Washington University (N = 395). To ensure a fair comparison in the subsequent case-control analysis, we excluded controls with PN, chemo-induced PN (CIPN), type I/II diabetes, or vitamin B12 deficiency (Supplementary Fig S1).

For Knight ADRC controls, individuals were classified based on cognitive testing and the Clinical Dementia

Rating (CDR). Cognitively healthy individuals with CDR = 0 were classified as healthy controls. Among these healthy controls, we manually reviewed health history and clinical diagnosis records according to the ADRC 2025 April clinical core module freeze to exclude any instances of the phenotypes listed above.

For UMD controls, the Institute of Informatics, Data Science & Biostatistics (I2DB) at Washington University linked participant medical records numbers (MRNs) to Epic medical records and extracted diagnosis data from multiple sources (inpatient/outpatient diagnoses, problem list history, medical history, and billing diagnosis history) to capture each participant's complete diagnostic history. We then exclude participants with the following ICD-10 codes: G60.x, G61.x, G62.x (PN and CIPN); E08.42, E09.42, E10.42, E11.42, E13.42 (diabetes); and D51.x (vitamin B12 deficiency).

After applying these exclusions across both cohorts, 778 of 879 controls were retained for the case-control analysis. All PCR-free short-read WGS for controls were generated at Washington University on Illumina NovaSeq platforms. Overall, the control cohort closely resembles the iPN cohort in mean age ( $63.1 \pm 14.7$  years) and racial composition (705/778, 90.6% White), as well as in sequencing QC metrics (see Tables 1 and Supplementary Table S1).

### Novel STR Expansion Detection Pipeline Based on WGS

Initial STR expansion detection was performed using EHdn<sup>14</sup> and EH<sup>15</sup> on BAM files generated from patient sample sequencing (Fig 1A). These data were aligned to the GRCh38 reference genome using Parabricks *pbrun germline* pipeline (version 4.0.0–1),<sup>31</sup> which utilized BWA-MEM<sup>32</sup> and STAR<sup>33</sup> as the sequencing aligners. We then used EHdn (version 0.9.1) to analyze the aligned BAM files and performed genome-wide STR scanning without reliance on predefined catalogs. We utilized ANNOVAR<sup>34</sup> to annotate genes where STRs are located, and the RepeatMasker database<sup>35</sup> to annotate known gene-motif pairs. Next, we prioritized motifs detected in the *RFC1* gene and created an *RFC1* STR catalog using gene information, repeat motifs, and genomic coordinates from EHdn output. We provided this catalog alongside the BAM files to EH (version 5.0.0) 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% of carriers detected by both tools. For each motif, we then defined carriers as the union of samples identified by EHdn

and those with at least one allele labeled with “INREPEAT” by EH.

### STR Motif Validation and Carriers Identification

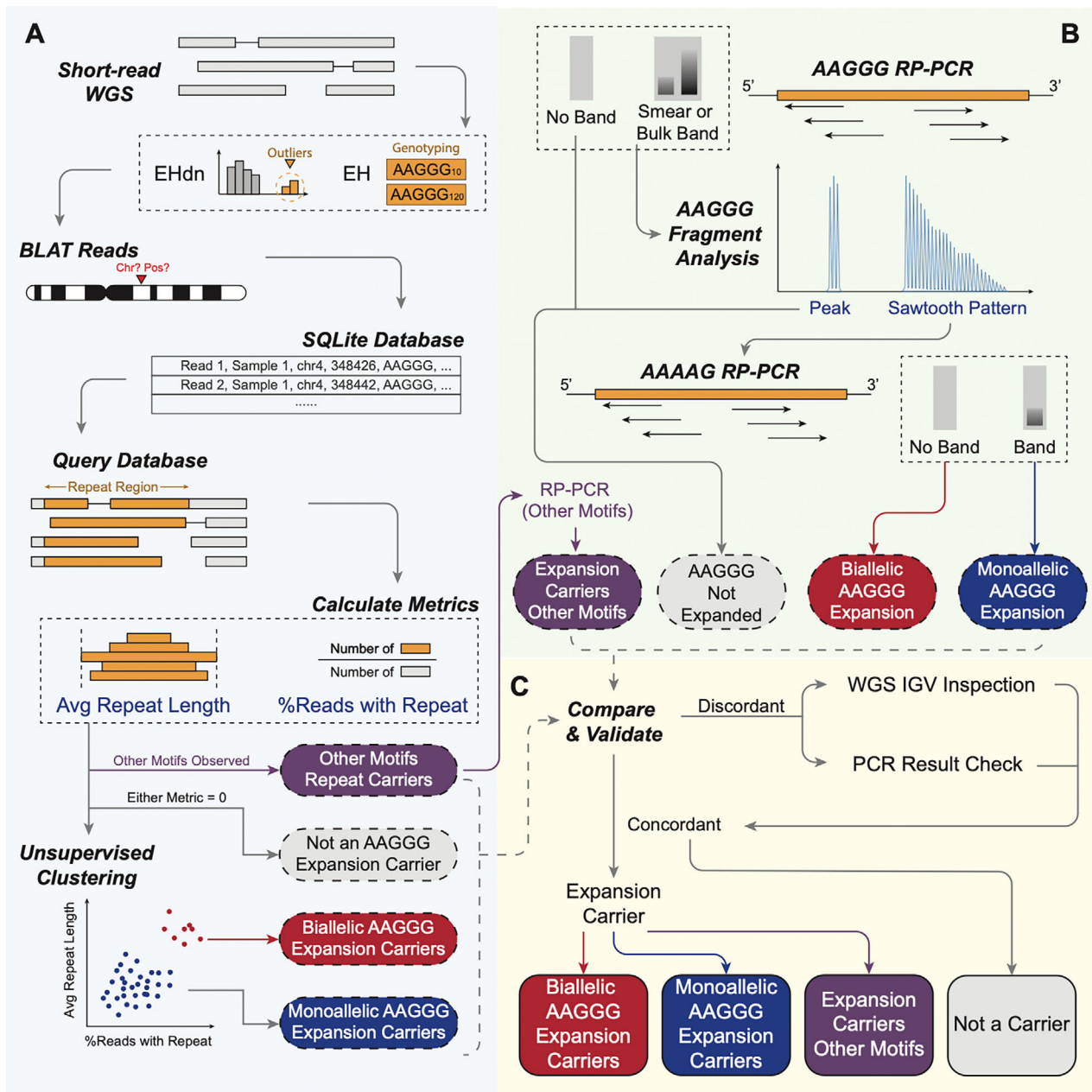
To validate the STR motifs identified in the previous step, we generated SAM files for each carrier in the potential STR expansion regions (see Fig 1A). Here, we defined the repeat window (ie, region-of-interest [ROI]) as 90 bp upstream and downstream from the *RFC1* repeat region in the reference genome. We initialized an SQLite database (Python 3.6.3 and SQLite3 2.6.0) containing all data from the SAM files to enable precise and efficient information retrieval and querying in subsequent steps. BLAT (version v.36) was run on every read within the SAM files, generating one PSL file per carrier per STR motif. We then used an in-house Python script to calculate the top 3 BLAT matches and their corresponding scores, which were integrated as a new column into the database.

We identified eligible carriers for a specific STR motif by extracting all unique samples listed in the database. To quantify STR length, we determined the total length of each read from its CIGAR string and identified the longest consecutive stretch of the queried motif, allowing one mismatch per motif unit through a Hamming distance approach. All rotational variants of the motif (eg, ACG, CGA, and GAC) were treated as the same motif. The algorithm also reports other motifs recognized that might interrupt the queried motif or reside on the other allele. RP-PCR-validated alternative motif expansion carriers are listed in Supplementary Table S4.

We next compared each read's STR length to a predetermined threshold (defaulted to be 10% of the read length). The proportion of candidate reads, carrying repeats longer than this threshold, relative to the total read count and multiplied by 100, provides the percentage of STR-containing reads for each sample (Equation 1). On the other hand, we calculated the mean STR length by dividing the sum of STR lengths in all reads 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.

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

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



**FIGURE 1:** Methodological workflow for *RFC1* repeat expansion detection. (A) 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. PCR = polymerase chain reaction; WGS = whole-genome sequencing.

### Genotype Determination for STR Carriers

To distinguish between potential biallelic and monoallelic STR expansions among carriers, we conducted 3 clustering approaches based on “% Reads with STR” and “Average STR Length,” as defined by Equations 1 and 2 above. 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 software version 4.4.1, NbClust 3.0.1; Supplementary

Fig S2A). We then applied three clustering algorithms 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). An example clustering result is shown in Supplementary Figure S2B. Any sample identified as biallelic by at least one clustering method was considered biallelic, and all

remaining samples were classified as monoallelic carriers. 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 (see Fig 1A). Samples with no supporting reads of the AAGGG motif were excluded from the carrier list.

For comparison and benchmarking purposes, we also determined the genotypes for individuals detected as carriers by ExpansionHunter. As previously described, we only keep EH carriers with at least one allele's length greater than 11 copies and were estimated from in-repeat reads (INREPEAT). Among these carriers, we further classified samples with alleles' lengths estimated from in-repeat or flanking reads (INREPEAT/INREPEAT and INREPEAT/FLANKING) as biallelic carriers, and those from spanning reads (INREPEAT/SPANNING) as monoallelic carriers.

### Repeat-Primed PCR

RP-PCR targeting the repeat locus was performed, as previously described, to identify carriers of the AAGGG expansion.<sup>18–20</sup> 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. Fragment analysis was subsequently used to impute expansion lengths. Long AAGGG expansions were inferred by the presence of distinct sawtooth patterns, whereas short AAGGG repeats displayed a limited number of peaks without evidence of an extended sawtooth pattern (Fig 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. Those showing bands in the AAAAG RP-PCR assay were classified as likely monoallelic carriers, whereas those without bands were classified as likely biallelic carriers. Flanking PCR was also conducted on these cases to assess for the presence or absence of biallelic expanded repeats. To clarify cases with additional repeat motifs observed in WGS data, we performed motif-specific RP-PCR for AAAGG, ACAAG, AACGG, AAGAG, and AAAGGG, and fragment analysis to clarify their expansion status. Detailed information on primer sequences and thermocycling conditions for all PCR experiments is provided in Supplementary Table S3.

### Benchmarking of WGS Results Against RP-PCR

We systematically compared WGS analysis results to those obtained via RP-PCR methods (Fig 1C). The WGS pipeline included intermediate EH outputs as well as final calls

generated by our computational pipeline, all of which were compared with PCR data. For samples identified as carriers by only PCR or WGS pipeline, we re-examined the PCR gel electrophoresis 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 5 consecutive repeat motifs were classified as carriers. Carriers missed by the detection pipeline but rescued by IGV inspection were labeled accordingly in Supplementary Table S2.

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 2 methods. In this scenario, we followed RP-PCR classification and ignored the discrepant WGS classification.

### Compound Heterozygous Deleterious Variants near RFC1 Gene

Raw FASTQ files of WGS data from the PNRR cohort were processed using Parabricks *pbrun germline* pipeline<sup>31</sup> (version 4.0.0–1), which utilized GATK *HaplotypeCaller*<sup>36</sup> as the variant calling tool. The resulting VCF files were split by chromosome, and the following analysis was based on chromosome 4 only (chromosome where the *RFC1* gene is located).

Variants in VCF files were annotated using SnpEff<sup>37</sup> (version 4.3) with multiple reference databases. After initial annotation against the Hg38 reference genome, the Ensembl<sup>38</sup> GRCh38.86 database was used for gene annotation. Variant types were classified using SnpSift,<sup>39</sup> 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.

## Results

### WGS Study 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.<sup>40</sup> 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.<sup>41</sup> Only individuals meeting the criteria for iPN, defined as a slowly progressive, symmetric, distal axonal polyneuropathy of unknown cause, were selected for WGS. 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, each undergoing 30X Illumina short-read WGS. This cohort was 38.2% (301/788) female patients and 92.8% (731/788) European-ancestry individuals, with an average age at onset of 62.2 years (see Table 1).

Using our WGS workflow, we performed case-control analysis and classified carriers of (AAGGG)<sub>exp</sub> STR in the *RFC1* gene as either biallelic or monoallelic by applying unsupervised clustering methods on the percentage of reads containing the motif and the average detected STR length (see Fig 1 and Methods). Our pipeline identified a significantly higher frequency of biallelic (AAGGG)<sub>exp</sub> *RFC1* STR carriers in patients with iPN (18/788, 2.3%) than controls (1/778, 0.1%; Fisher's exact  $p = 0.00007$ , odds ratio [OR] = 18.0, confidence interval [CI] = 2.82–747.43; Fig 2A). In contrast, we observed no significant difference in monoallelic *RFC1* STR frequency in patients with iPN (72/788, 9.1%) compared with controls (56/778, 7.2%; Fisher's exact  $p = 0.17$ , OR = 1.3, CI = 0.89–1.90; Fig 2B). To address the absence of large-scale comprehensive benchmarking STR detection from short-read sequencing against RP-PCR, we performed PCR analysis on the same cohort analyzed by WGS.

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 Figures 2C and 2D, EH carriers with all types of supporting reads (last row in each plot) identified 10 biallelic and 148 monoallelic unique carriers that did not overlap with the shared biallelic and monoallelic carriers detected by PCR and our WGS workflow, suggesting a high rate of false-positive results. Restricting EH results to at least one allele having INREPEAT supporting reads (third row in each plot) helped eliminate some of these false carriers. Our workflow, building on top of EH INREPEAT calls and implementing our in-house genotyping and filtering criteria and further IGV examination, closely aligned with PCR

results while maintaining a balance between sensitivity and specificity (see below).

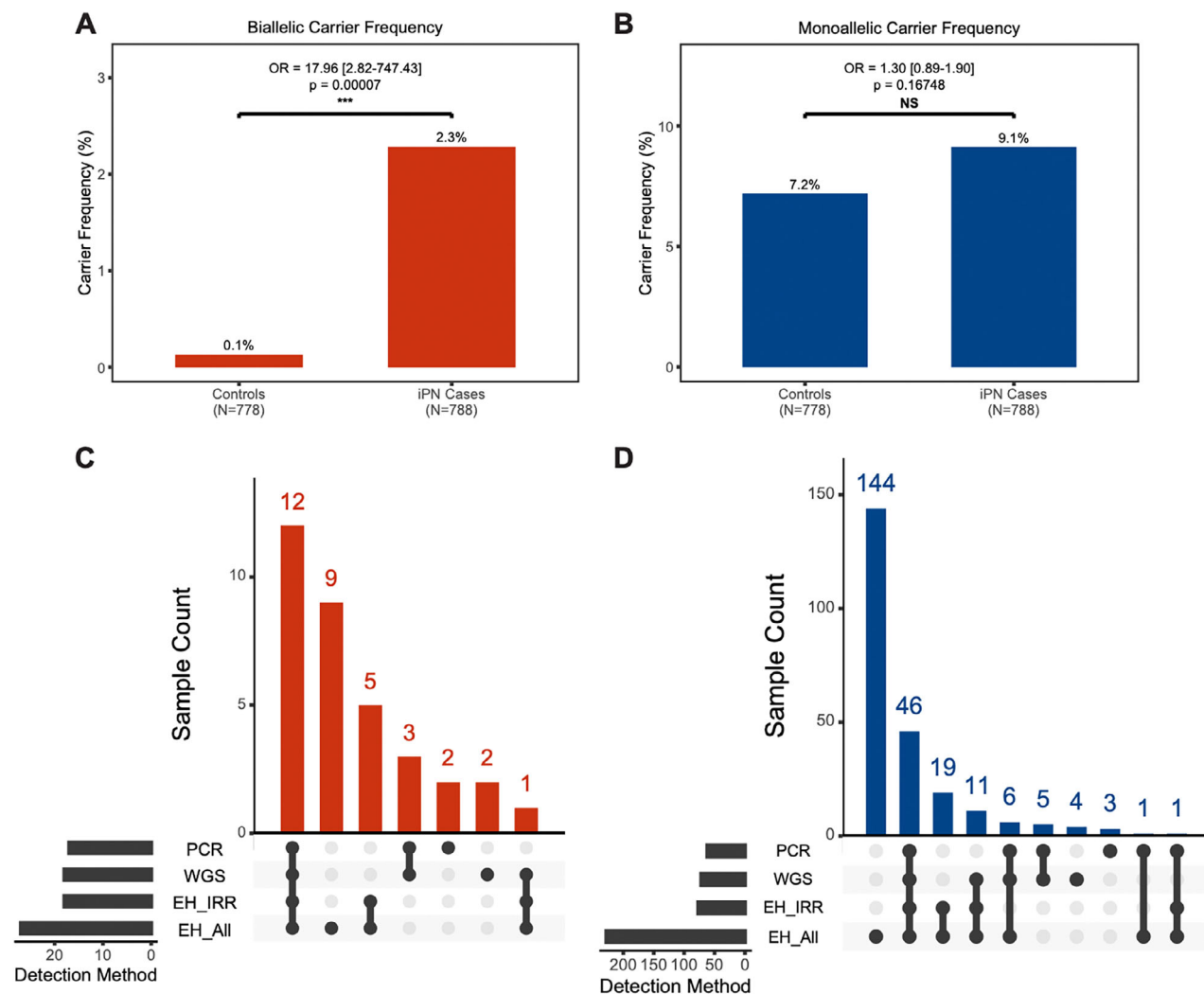
### PCR Validation of *RFC1* STR Carriers

Next, we sought to validate WGS-based genotyping results by PCR. Overall, WGS and PCR methods produced matching genotype classifications in 97% (766/788) of iPN cases, with a Cohen's kappa coefficient of 0.87, indicating excellent concordance (Fig 3A). Based on PCR validation results, we confirmed 10.2% (N = 80) of the iPN cohort as *RFC1* (AAGGG)<sub>exp</sub> repeat carriers (Fig 3B). Among the carriers, 77.5% (N = 62) carried heterozygous repeat expansions (example shown in Supplementary Fig S3) and 22.5% (N = 18) were biallelic repeat expansion carriers (example shown in Supplementary Fig S4). In 5 samples that were classified as carriers by PCR only, we carefully reviewed the reads in IGV and found no reads carrying the altered AAGGG motif (example shown in Supplementary Fig S5). We believe that this is due to the inherent alignment challenges in repetitive regions, where pure AAGGG-containing reads likely mis-map to other genomic locations with similar repeat sequences. A detailed summary of the carrier status and genotyping results for all carriers is provided in Supplementary Table S2.

Stratified analysis based on PCR-validated carriers revealed that biallelic expansion carriers frequency was significantly higher in the pure sensory iPN subcategory (10/144, 6.9%) compared with patients with pure SFN (4/369, 1.1%; Fisher's exact  $p = 0.0008$ , OR = 6.8, CI = 1.9–30.1) and patients with sensorimotor iPN (4/266, 1.5%; Fisher's exact  $p = 0.008$ , OR = 4.9, CI = 1.4–21.7; Fig 3C). No significant differences in biallelic carrier frequency were observed between pure SFN and sensorimotor iPN subcategories or between patients with sensorimotor iPN and patients with pure motor iPN. The elevated biallelic carrier frequency in the sensory subtype aligned with previous reports, but the frequency of biallelic repeats in this population (6.9%) was lower than previously described (34%).<sup>20,42</sup> In contrast, a stratified analysis on monoallelic (AAGGG)<sub>exp</sub> *RFC1* STR carriers revealed no significant differences among sensory, SFN, sensorimotor, and sensory/SFN categories (Fig 3D).

### PCR Validation of Atypical Repeat Carriers

We saw several different types of atypical repeat carriers in our analysis. For instance, we observed 4 samples with stretches of AAGGG mixing with other repeat motifs on both alleles. To differentiate them from true AAGGG expansion carriers, we performed additional motif-specific RP-PCR and fragment analysis for motifs AAAGG, ACAAG,



**FIGURE 2:** WGS case-control analysis results and comparison across different detection methods. (A) Biallelic *RFC1* AAGGG carrier frequency significantly higher in iPN cases than in controls ( $p < 0.001$ , OR = 2.82-747.43). (B) Monoallelic *RFC1* AAGGG carrier frequency not significantly different between iPN cases and controls ( $p > 0.05$ , OR = 0.89-1.90). (C) Upset plot showing the number of biallelic carriers identified in WGS workflow (WGS), PCR workflow (PCR), EH results supported by IRR/IRR or IRR/FL reads (EH\_IRR), and EH results supported by IRR/IRR, IRR/FL, and FL/FL reads (EH\_All). (D) Upset plot showing the number of monoallelic carriers identified in WGS, PCR, EH results supported by IRR/SP reads (EH\_IRR), and EH results supported by IRR/SP and FL/SP reads (EH\_All). EH = ExpansionHunter; FL = flanking; iPN = idiopathic peripheral neuropathy; IRR = in-repeat; OR = odds ratio; PCR = polymerase chain reaction; SP = spanning; WGS = whole-genome sequencing.

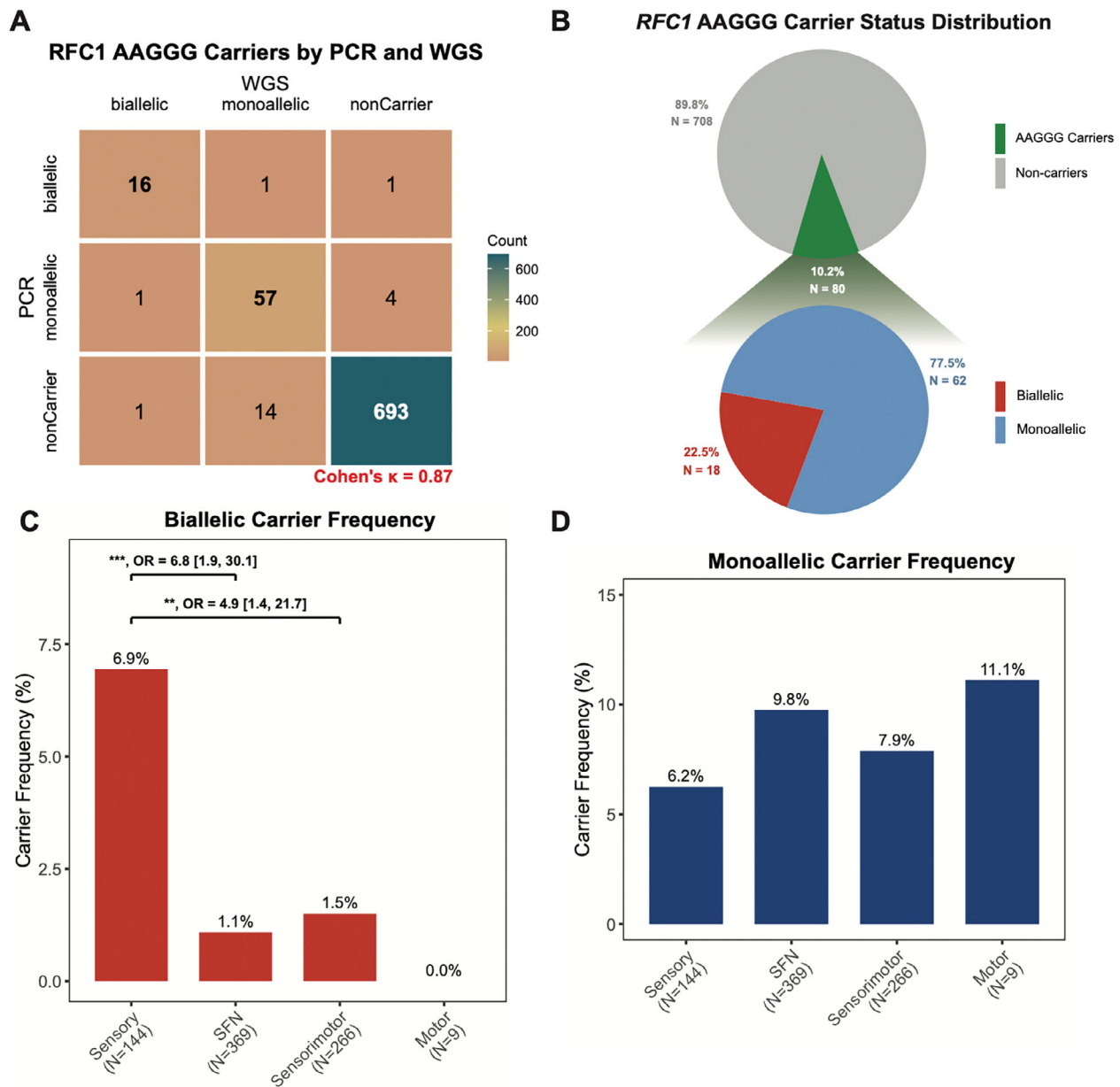
AAAGGG, AAGAG, and AACGG. In these samples, all motifs (including AAGGG) showed peaks rather than the typical decremental sawtooth pattern in fragment analysis, exhibiting no definitive evidence of true AAGGG expansions (Fig 4). Therefore, we excluded these samples from the AAGGG expansion carrier list in our analysis.

In addition, we also observed 5 patients carrying *RFC1* expansions with other motifs on one or more alleles (see Supplementary Table S4). Although these alternative *RFC1* motifs were rare in our cohort, a recent long-read sequencing study reported their presence in self-reported Black or African American participants from the All of Us program.<sup>43</sup> Given the high prevalence of peripheral neuropathy in the general population, it is possible that these

novel *RFC1* motifs contribute to disease risk, a question that warrants further investigation.

### Clinical Spectrum of Patients Carrying *RFC1* STRs

We next analyzed the clinical characteristics of the confirmed 18 biallelic and 62 monoallelic *RFC1* (AAGGG) exp in comparison to 708 non-carriers (see Fig 3B and Table 2). A two-way chi-squared test (degrees of freedom DOF = 2) revealed a significant difference in the prevalence of sensory iPN ( $p < 0.001$ ), with biallelic carriers showing markedly higher rates (10/18, 56%) than monoallelic carriers (8/62, 13%) or non-carriers (126/708, 18%). No significant differences were observed across the



**FIGURE 3:** RP-PCR validation and *RFC1* AAGGG carrier frequency distribution across iPN subtypes. (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. Cohen's kappa coefficient is labeled on the lower right corner. (B) Distribution of final *RFC1* AAGGG carrier status in the study cohort. Top left: Overall carrier status showing 10.2% (N = 80) carriers and 89.8% (N = 708) non-carriers among iPN cases; bottom right: distribution by expansion type among carriers: biallelic carriers, monoallelic carriers, and edge cases. (C) Biallelic AAGGG carrier prevalence showing a significantly higher frequency in pure sensory neuropathy (6.9%) compared with other subtypes ( $p < 0.001$ ). (D) Monoallelic AAGGG carrier frequency showing no statistically significant differences across iPN subtypes. iPN = idiopathic peripheral neuropathy; PCR = polymerase chain reaction; RP-PCR = repeat-primed polymerase chain reaction; WGS = whole-genome sequencing.

3 groups with respect to sex, age, presence of severe pain, sensory motor iPN, pure SFN, or motor iPN.

Given that AAGGG repeat expansions classically cause CANVAS, which is characterized by ataxia and vestibular features, we looked for additional clinical features which might capture subclinical phenotypes suggestive of 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.<sup>19</sup> 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 repeat-positive biallelic SFN patients had a relatively acute onset of neuropathy in their 20s and



**TABLE 2. Clinical Characteristics of Confirmed Biallelic and Monoallelic *RFC1* AAGGG Carriers Versus Non-Carriers in iPN Cases**

	Biallelic (N = 18)	Monoallelic (N = 62)	Non-carriers (N = 708)	<i>p</i>
Male	9 (50%)	36 (58%)	424 (60%)	NS
Age at visit	65 ± 10	62 ± 14	62 ± 14	NS
Painful neuropathy	14 (78%)	44 (71%)	490 (70%)	NS
Sensory neuropathy	10 (56%)	8 (13%)	126 (18%)	< 0.001
Sensorimotor	4 (22%)	20 (32%)	242 (3%)	NS
Small fiber neuropathy	4 (22%)	33 (53%)	332 (47%)	NS
Motor neuropathy	0 (0%)	1 (2%)	8 (1%)	NS

Chi-squared tests were conducted for sex, age at visit, painful status, and percentage carriers in each subtype across all groups. Pure sensory iPN was significantly more prevalent in biallelic carriers (56%) compared with monoallelic carriers (13%) and non-carriers (18%) ( $p < 0.001$ ), while age at visit, sex distribution, and painful neuropathy prevalence were similar across groups. iPN = idiopathic peripheral neuropathy; NS = not significant.

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.<sup>44</sup> However, the PNRR cohort is heavily biased toward Caucasians and thus likely reflects a genetic background similar to prior European studies. Differences in selection and screening criteria, repeat assessment methodology, and the smaller size of the Italian cohort, which included only 125 cases, may also explain the discrepancy.<sup>18</sup> Similarly, identification of biallelic *RFC1* repeat expansions in patients with SFN and patients with sensorimotor iPN in our large cohort differs from observations in smaller and ataxia focused cohorts, where a prior study<sup>18</sup> reported no biallelic carriers among 100 patients with sensorimotor iPN. In contrast, our larger cohort revealed biallelic *RFC1* expansions in 1.5% (N = 4) of such patients. Thus, we suggest that the presence of motor nerve involvement should not preclude patients from undergoing *RFC1* repeat screening.

Monoallelic *RFC1* (AAGGG)<sub>exp</sub> showed a nominally higher prevalence in our iPN cohort (72/788, 9.1%) compared with controls (56/778, 7.2%;  $p = 0.17$ ). This effect was significant in a comparison of the entire PNRR to an unfiltered control population (N = 879; see Supplementary Fig S1). However, after controlling for potentially confounding covariates such as diabetes and B12 deficiency, this difference no longer reached statistical significance. Consistent with this notion, the monoallelic prevalence observed in iPN cases also exceeded the 7.3% reported by Ibañez K, et al. in a large multi-ethnic cohort

derived from TOPMed and the 1000 Genomes Project.<sup>44</sup> Because this difference did not reach statistical significance, whether a single expanded *RFC1* allele contributes to iPN risk remains to be determined. Prior studies suggest that some CANVAS cases have compound heterozygous mutations—with an AAGGG repeat on one allele and a protein-altering mutation on the other allele. Nevertheless, 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 (Methods). We did observe short (typically 2–5 copies) alternative repeat sequence elements within the other allele in 19 out of 62 (30.6%) of our monoallelic carriers that might provide an alternative explanation for the slightly increased frequency in the iPN population. All such cases had AAAAG repeat alleles as measured by repeat-primed PCR and thus met our criteria for monoallelic carrier status.

Our control cohort included rare *RFC1* AAGGG biallelic carriers (1/778, 0.1%), which was similar to the 0.15% reported in a previous short-read WGS study.<sup>44</sup> Likewise, monoallelic carriers in our controls occurred at a frequency of 7.2% (56/778), comparable to the 7.3% reported in TOPMed<sup>44</sup> and higher than the 0.7% reported in the 1000 Genomes project.<sup>45</sup> Interestingly, we observed distinct control carrier frequencies when we analyzed *RFC1* STRs using different EH catalogs, including the EHdn-defined region (chr4:39,348,426–39,349,038), the RepeatMasker-defined region (chr4:39,348,275–39,348,633), and coordinates from Ibañez et al.<sup>44</sup> (chr4:39,348,424–39,348,479; Supplementary Fig S6). This finding highlighted the technical sensitivity of STR

detection algorithms, where using a larger ROI, as defined in raw EHDn output, led to a significantly lower carrier frequency compared with other catalogs used. A limitation of this case-control study is that 50 of our UMD control subjects lacked diagnostic records at our institution, as they were recruited solely as sequenced parents of the enrolled probands rather than as independently evaluated clinical participants. As such, our numbers may underestimate the impact of monoallelic *RFC1* (AAGGG)<sub>exp</sub> on the risk of developing iPN. This factor will need to be carefully considered in future large database analyses.

Although the prevalence of monoallelic *RFC1* (AAGGG)<sub>exp</sub> did not reach statistical significance, the observed trend warrants consideration of how non-reference AAGGG repeat expansions might contribute to disease. If disease is mediated primarily through a loss-of-function mechanism, haploinsufficiency could only modestly increase susceptibility. Alternatively, heterozygous AAGGG expansions could exert a dose-dependent toxic gain-of-function effect, with greater penetrance and severity in biallelic carriers.<sup>28</sup> This latter model is consistent with recent work in CANVAS patient-derived glutamatergic iNeurons, where selective removal of the repeat suppressed disease relevant phenotypes, but reexpression of *RFC1* protein did not.<sup>28</sup> Larger cohorts will be required to determine whether monoallelic expansions are truly enriched in iPN and, if so, to define the clinical impact of heterozygous expansions 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 infection, and their neuropathy persisted for many years. This finding suggests that patients with acute onset SFN 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 and PCR revealed excellent concordance in most cases. Aside from AAGGG repeat carriers, we found some carriers with other repeat structures, such as (AAAGGG)<sub>exp</sub>, (AAAGG)<sub>exp</sub>, or a mixture of them with (AAGGG)<sub>exp</sub> (see Supplementary Table S4). These cases were initially identified by WGS and subsequently validated by RP-PCR. Our WGS pipeline successfully identified these alternative repeat motifs that would likely have been overlooked by conventional PCR assays designed for a limited set of target sequences, demonstrating the superior sensitivity of WGS for comprehensive variant 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. Our approach is practical and cost-effective 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, because capture success for both PCR and WGS depends on motif composition and regional complexity. Whereas long-read sequencing would address these shortcomings through a more definitive characterization of STR expansions, it requires significantly larger DNA input and remains prohibitively expensive for large-scale clinical 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.

Additionally, the limited number of biallelic *RFC1* carriers identified in this study precludes definitive conclusions regarding the spectrum of affected fiber types and optimal patient selection for screening. Whereas our findings suggest that *RFC1* repeat expansion can occur across the full spectrum of patients with iPN, including pure small fiber and motor presentations, we cannot firmly determine whether specific clinical features predict *RFC1* expansion carrier status. Furthermore, as noted, the absence of medical records for 50 UMD parent controls limits our ability to fully characterize the monoallelic expansion frequency difference between cases and controls. Larger-scale studies with increased numbers of confirmed biallelic expansion carriers will be necessary to establish whether *RFC1* screening should be uniformly applied to all patients with iPN or targeted to specific phenotypic subgroups.

In conclusion, our findings underscore the diagnostic value of *RFC1* expansion screening in patients with iPN, where biallelic expansions account for roughly 2.3%

of cases, and monoallelic expansions occur at a significantly higher frequency in patients with iPN versus controls. Overall, our data support testing for *RFC1* repeat expansions in patients with iPN. Moreover, if future studies with greater statistical power confirm monoallelic expansions as a risk factor for PN, this would have important implications for genetic counseling and interpretation of *RFC1* testing results. Considering the current difficulties in obtaining *RFC1* testing in the United States, our pipeline will be very useful as a rapid screening test for *RFC1* repeat expansion carriers. We are hopeful that applying this pipeline will lead to a deeper understanding of genotype–phenotype relationships and will improve our ability to resolve the pathogenic mechanisms underlying CANVAS and PN as a whole – providing a way forward for developing therapies for these currently untreatable conditions.

## Acknowledgment

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 also 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 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. Baldrige, C. Gurnett, T. Turner, and B. Leon Ricardo). We appreciate obtaining access to genetic and phenotypic data from the UMD project.

PNRR is supported by the Foundation for Peripheral Neuropathy. This work was supported by the Hydrocephalus Association Innovator Award (S.C.J.), Cerebral Palsy Alliance Research Foundation Project Grant (PRG03121; S.C.J.), WashU Children's Discovery Institute Faculty Scholar Award (CDI-FR-2021-926; S.C.J.), and NIH U19NS130607, R01NS111029, and R01NS131610 (S.C.J.). NIH (R21NS135481, P30MH075673-011; A.H.), Wellcome Trust (A.H.), Dr. Miriam and Sheldon G. Adelson Medical Research

Foundation (A.H.), and Dr. Richard Merkin Family Foundation (A.H.). NIH U19NS130607 (J.M.) and the Bob and Signa Hermann Fund for Neuropathy Research (J.M.). This project was supported by the NIH (R21NS129096) and the Taubman Institute at the University of Michigan (P.K.T. and S.S.O.).

## 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., R.I., S.C.J., P.K.T., A.H., J.M., M.L., 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.C.J., S.S.O. contributed to drafting the text or preparing the figures and tables. All authors contributed to editing the manuscript.

## Potential Conflicts of Interests

P.K.T. 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. A.H. served as consultant to Pfizer, Sangamo and Sanofi. None of these relationships are directly relevant to this manuscript. J.M. 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/ztang99/STR-detection-genotyping>.

## References

1. Lehmann HC, Wunderlich G, Fink GR, Sommer C. Diagnosis of peripheral neuropathy. *Neurol Res Pract* 2020;2:20.
2. Callaghan B, Kerber K, Langa KM, et al. Longitudinal patient-oriented outcomes in neuropathy. *Neurology* 2015;85:71–79.
3. Singer MA, Vernino SA, Wolfe GI. Idiopathic neuropathy: new paradigms, new promise. *J Peripher Nerv Syst* 2012;17:43–49.
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:e3163–e3179.
5. PEPAN - Overview: Comprehensive Peripheral Neuropathy Gene Panel, Varies [Internet], [date unknown]; [cited 2023 Dec 4] Available at: <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;32:1–27.
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://doi.org/10.1101/2024.05.01.24306681v1>.
9. Fan H, Chu J-Y. A brief review of short tandem repeat mutation. *Genomics Proteomics Bioinformatics* 2007;5:7–14.
10. Zegura SL. Human evolutionary genetics: origins, peoples and disease. *Am J Hum Genet* 2005;76:1087–1088.
11. Malik I, Kelley CP, Wang ET, Todd PK. Molecular mechanisms underlying nucleotide repeat expansion disorders. *Nat Rev Mol Cell Biol* 2021;22:589–607.
12. Willems T, Gymrek M, Highnam G, et al. The landscape of human STR variation. *Genome Res* 2014;24:1894–1904.
13. Mallick S, Li H, Lipson M, et al. The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* 2016;538:201–206.
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;21:102.
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;35:4754–4756.
16. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucl Acids Res* 2019;47:e90.
17. Tankard RM, Bennett MF, Degorski P, et al. Detecting expansions of tandem repeats in cohorts sequenced with short-read sequencing data. *Am J Hum Genet* 2018;103:858–873.
18. Cortese A, Tozza S, Yau WY, et al. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome due to RFC1 repeat expansion. *Brain* 2020;143:480–490.
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;51:649–658.
20. Currò R, Salvalaggio A, Tozza S, et al. RFC1 expansions are a common cause of idiopathic sensory neuropathy. *Brain* 2021;144:1542–1550.
21. Davies K, Szmulewicz DJ, Corben LA, et al. RFC1 -related disease: molecular and clinical insights. *Neurol Genet* 2022;8:e200016.
22. Uhlen M, Karlsson MJ, Zhong W, et al. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* 2019;366:eaax9198.
23. 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;147:1887–1898.
24. Scriba CK, Beecroft SJ, Clayton JS, et al. A novel RFC1 repeat motif (ACAGG) in two Asia-Pacific CANVAS families. *Brain* 2020;143:2904–2910.
25. Dominik N, Magri S, Currò R, et al. Normal and pathogenic variation of RFC1 repeat expansions: implications for clinical diagnosis. *Brain* 2023;146:5060–5069.
26. Ronco R, Perini C, Currò R, et al. Truncating variants in RFC1 in cerebellar ataxia, neuropathy, and vestibular areflexia syndrome [Internet]. *Neurology* 2023;100(0)[cited 2025 Apr 1]:e543–e554. <https://doi.org/10.1212/WNL.0000000000201486>.
27. Weber S, Coarelli G, Heinzmann A, et al. Two RFC1 splicing variants in CANVAS. *Brain* 2023;146:e14–e16.
28. Maltby CJ, Krans A, Grudzien SJ, et al. AAGGG repeat expansions trigger RFC1-independent synaptic dysregulation in human CANVAS neurons. *Sci Adv* 2024;10:eadn2321.
29. 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;5:165–176.
30. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
31. NVIDIA. Clara Parabricks [Internet], 2025. Available at: <https://www.nvidia.com/en-us/clara/genomics/>.
32. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet], 2013. [cited 2025 Feb 27] Available at: <http://arxiv.org/abs/1303.3997>.
33. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29:15–21.
34. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucl Acids Res* 2010;38:e164.
35. Smit A, Hubley R, Green P. RepeatMasker Open-4.0 [Internet], 2013. Available at: <http://www.repeatmasker.org>.
36. Poplin R, Ruano-Rubio V, DePristo MA, et al. *Scaling accurate genetic variant discovery to tens of thousands of samples* [Internet]. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, 2018. [cited 2025 Apr 4] <https://doi.org/10.1101/201178v3>
37. 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<sup>1118</sup>*; iso-2; iso-3. *Fly* 2012;6:80–92.
38. Harrison PW, Amode MR, Austine-Orimoloye O, et al. Ensembl 2024. *Nucl Acids Res* 2024;52:D891–D899.
39. 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 Genet* 2012;3:35 [cited 2025 Apr 4]. <https://doi.org/10.3389/fgene.2012.00035>.
40. Thomas S, Ajroud-Driss S, Dimachkie MM, et al. Peripheral neuropathy research registry: a prospective cohort. *J Peripher Nerv Syst* 2019;24:39–47.
41. 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;72:185–192.
42. Tagliapietra M, Cardellini D, Ferrarini M, et al. RFC1 AAGGG repeat expansion masquerading as chronic idiopathic axonal polyneuropathy. *J Neurol* 2021;268:4280–4290.
43. Danzi MC, Xu IRL, Fazal S, et al. Detailed tandem repeat allele profiling in 1,027 long-read genomes reveals genome-wide patterns of pathogenicity [Internet], 2025;2025.01.06.631535.[cited 2025 Sep 22]. <https://doi.org/10.1101/2025.01.06.631535v2>.
44. Ibañez K, Jadhav B, Zanovello M, et al. Increased frequency of repeat expansion mutations across different populations. *Nat Med* 2024;30:3357–3368.
45. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature* 2015;526:68–74.