

BRAIN COMMUNICATIONS

Detection of short tandem repeat expansions on a targeted neurological gene panel using STRipy improves the diagnostic rate for ataxias

 Carolin K. Scriba,^{1,2*}  Chiara Folland,^{1,*}  Michael Black,² Jessica Baker,² Daniel Abromeit,² Samantha Saw,² Mei-Ting Chiew,² Rebecca Gooding,² Nigel G. Laing,¹  Mark R. Davis² and  Gianina Ravenscroft¹

* These authors contributed equally to this work.

Short tandem repeat expansions are associated with over 50 diseases, many with primary neurological presentations. Despite the prevalence of short tandem repeat expansion disorders, genetically diagnosing these conditions is complicated by a lack of efficient and comprehensive diagnostic screening approaches. We integrated a new short tandem repeat genotyping tool, STRipy, into the analytical workflow for short-read sequencing data generated by the comprehensive neurological disease gene panel used in the Diagnostic Genomics Department, PathWest Laboratory Medicine. We tested STRipy on Versions 6 and 7 of the panel. Version 6 already included probes covering five short tandem repeat expansion loci in the following genes: *CACNA1A*, *PPP2R2B*, *TBP*, *NOP56* and *RFC1*. Additional probes targeting 13 neurological disease short tandem repeat expansion loci were designed and included in Version 7. All expansions detected by STRipy were validated and sized using PCR-based diagnostic techniques. Four hundred and eighteen patients with ataxia were tested on Version 6 of the panel, and 61 (14.6%) had reportable pathogenic variants, including 11 patients with pathogenic repeat expansions detected by STRipy. Sixty-seven ataxia patients were tested on Version 7 of the panel, and 15 (22.4%) had reportable pathogenic variants, including three repeat expansions detected by STRipy. Therefore, STRipy contributed 18.0% and 20.0% of the solved cases from Version 6 and 7 of the ataxia subpanels, respectively. STRipy accurately identified and sized loci with shorter pathogenic repeat thresholds where the expansion was smaller than the read length. In addition to increased diagnostic yield, implementation of STRipy into diagnostic analysis pipelines has streamlined clinical diagnosis of short tandem repeat expansion disorders.

1 Centre for Medical Research, University of Western Australia, Harry Perkins Institute of Medical Research, Perth, WA 6009, Australia

2 Department of Diagnostic Genomics, PathWest Laboratory Medicine, Nedlands, WA 6009, Australia

Correspondence to: Gianina Ravenscroft

Harry Perkins Institute of Medical Research, 6 Verdun St, Nedlands, WA 6009, Australia

E-mail: gina.ravenscroft@perkins.uwa.edu.au

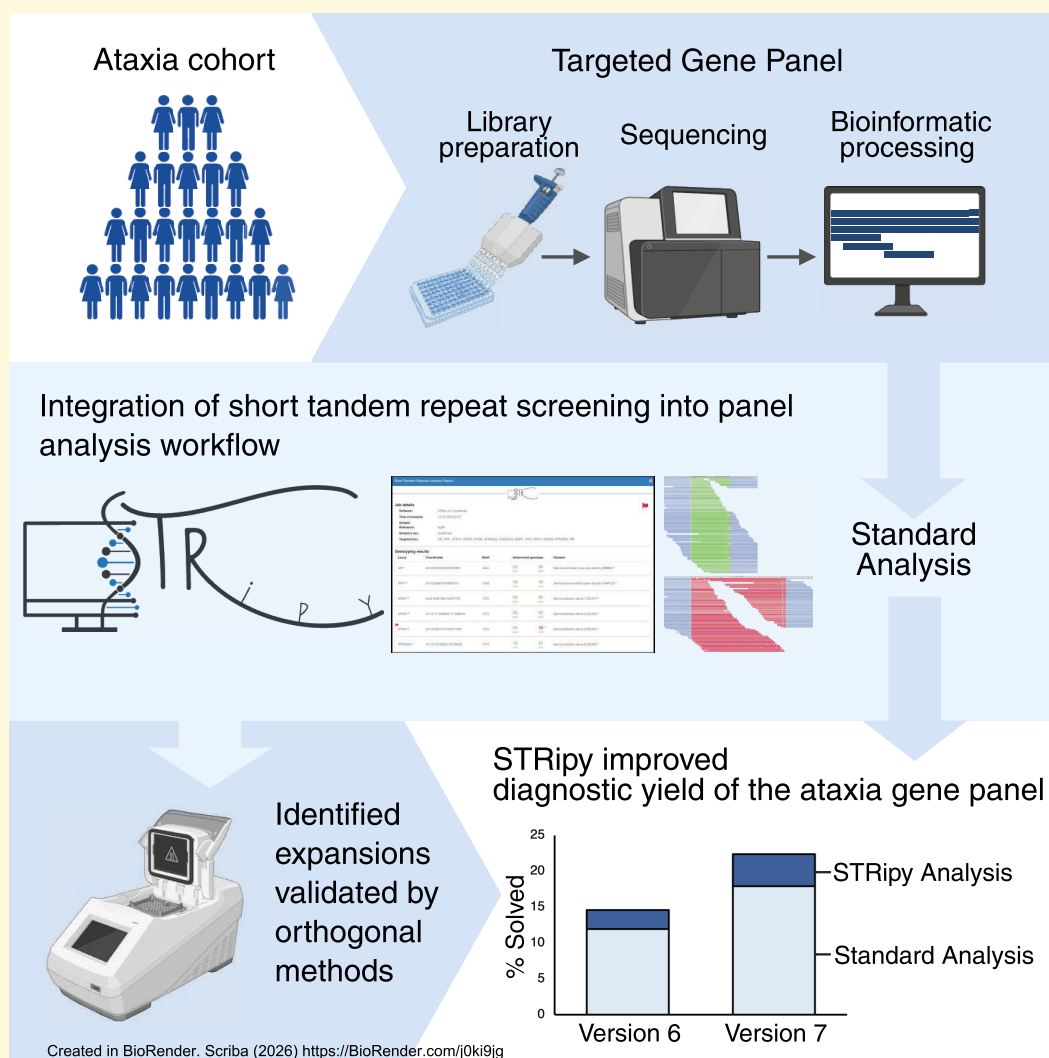
Correspondence may also be sent to: Carolin Scriba

Harry Perkins Institute of Medical Research, 6 Verdun St, Nedlands, WA 6009, Australia

E-mail: carolin.scriba@perkins.uwa.edu.au

Keywords: STRs; ataxia; diagnostic screening; neurogenetics

Graphical Abstract



Introduction

Neurogenetic disorders often exhibit heterogeneity within individual diseases while demonstrating high levels of clinical overlap between separate disease entities.^{1,2} Hundreds of genes have been associated with neurological conditions. For the past 10 years, short-read, massively parallel sequencing (MPS) technologies have enabled neurogenetic diagnostics to evolve from time-intensive single gene tests to high-throughput screens, testing hundreds of genes at once.^{3,4}

Short tandem repeats (STRs) are short sequences of DNA between two and six nucleotides that are repeated consecutively within the genome. These repetitive regions display high rates of instability and are prone to expansion. More than 50 STR expansions have been associated with disease, the majority of which present as primary neurological conditions.⁵ Eighteen of these cause hereditary ataxias, accounting for the most common forms.⁶ Overall, STR expansion diseases are estimated to affect ~1 in 3000 people.⁷ Despite their prevalence, the repetitive nature of these loci has made their detection challenging, and standard

diagnostic practice continues to rely on single gene, PCR-based assays. Therefore, an adaptable high-throughput method for the screening of STR expansion disorders would provide great utility for neurogenetic diagnostics, improving both efficiency and diagnostic yield.

Several tools have been developed to identify STR expansions from short-read sequencing data.⁸⁻¹³ In 2022, the open-source software, STRipy, was released.¹⁴ STRipy provides a graphical user interface that can genotype loci from its catalogue of all known pathogenic STR expansions. The software uses ExpansionHunter as the backend genotyper and integrates the tool REviewer to provide simple visual presentation of the predicted genotype and detection of any phenotype altering interruptions. To enable automated, high-throughput analysis of STR expansion loci, Halman *et al.*¹⁴ also developed a pipeline version of STRipy.

We integrated the STRipy pipeline into the analytical workflow for data generated by the comprehensive MPS neurological disease gene panel used at PathWest's Diagnostic Genomics department. The panel consists of more than 500 clinically relevant genes and

is used to test approximately a thousand patients for neurogenetic conditions every year. Probes targeting additional STR expansion loci were incorporated into the newest version (Version 7) of the panel and 455 patients were screened for pathogenic expansions. We also conducted retrospective screening of ataxia patients who were tested on Version 6 of the panel ($n = 418$), as five STR expansion loci associated with ataxia phenotypes were already covered. Here, we demonstrate that STRipy is a valuable tool for screening known STR expansion disease loci from short-read targeted panel data, resulting in a substantially improved diagnostic yield for ataxias.

Materials and methods

Panel sequencing

The PathWest neurological disease targeted MPS gene panel (Version 6) consisted of 559 genes associated with numerous neurological disease phenotypes. Custom oligo design for target enrichment and panel manufacture were performed by TWIST Biosciences with library preparation, according to manufacturer's instructions (EF workflow kit; Twist Bioscience, product nos. 101901 and 103904). Libraries were sequenced on the NextSeq 550/500 system using the Mid output cartridge (Illumina, Version 2.5, Cat no. SY-415-1002) with an average coverage of 300× across the panel and 50× for repeat regions. Raw sequencing data were processed using the DRAGEN enrichment pipeline, aligning to GRCh38. Analysis of sequencing data for single nucleotide variants (SNV) and small indels was performed using the Alissa Interpret Research software package (Agilent Technologies). Patients were analysed on *in silico* phenotypic subpanels corresponding to their clinical features. Copy number variants (CNVs) were detected as described by Beecroft *et al.*³

Additional probes added to the PathWest neurological targeted gene panel

Four STR loci (SCA6, SCA12, SCA17 and SCA36) in genes associated with spinocerebellar ataxia (SCA) were already covered on

Version 6 of the panel because non-repeat associated pathogenic variants have been reported within these genes. Probes for the *RFC1* locus were also already present as they were introduced to assess if manual inspection of binary alignment map (BAM) files could be used as a screening tool.¹⁵

To facilitate the detection of STR expansions from panel data, additional probes were ordered for Version 7 to cover loci for the following repeat expansion disorders: spinal and bulbar muscular atrophy (SBMA), dentatorubral-pallidoluysian atrophy (DRPLA), SCA1, SCA2, SCA3, SCA7, SCA8, SCA10, *C9orf72* motor neuron disease/frontotemporal dementia (MND/FTD), SCA27B, Huntington disease-like 2 (HDL2), Huntington's disease (HD) and neuronal intranuclear inclusion disease (NIID). Genes and oligo coordinates are displayed in [Supplementary Table 1](#).

Incorporating STRipy pipeline into the diagnostic workflow

For automated, high-throughput analysis of STR expansion loci, we used the STRipy pipeline, a command-line version of STRipy (available at <https://gitlab.com/andreassh/stripipy-pipeline>). Using a docker installation, the STRipy pipeline was integrated into the existing PathWest Diagnostic Genomics in-house bioinformatics analysis workflow ([Fig. 1](#)). The code and computational resources required are provided in the [Supplementary Materials: Methods](#). The STR repeat loci that were analysed included all those associated with neurological conditions listed in the STRipy online database (<https://stripipy.org/database>), including those covered in panel Versions 6 and 7 (see [Supplementary Table 1](#)). Loci that were not covered in Versions 6 and/or 7 were included in the analysis to account for STRipy capturing off-target expansions. A full list of regions analysable by STRipy using data from Version 6 and/or 7 of the panel is available in [Supplementary Table 2](#). The pathogenic thresholds listed in the STRipy online database were used.¹⁴ A minimum of 20× coverage of each repeat region is required for STRipy to detect repeats.

Outputs consisted of html files summarizing results for each sample in an interactive report ([Fig. 2](#)) and a text file that indicated

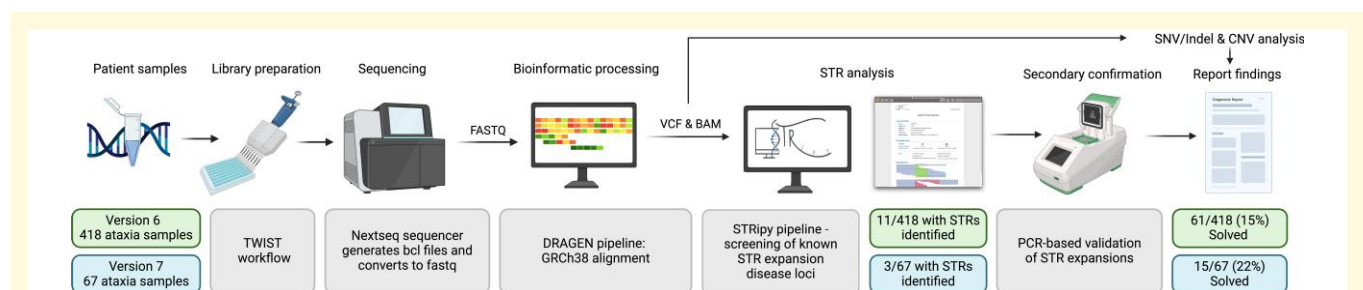
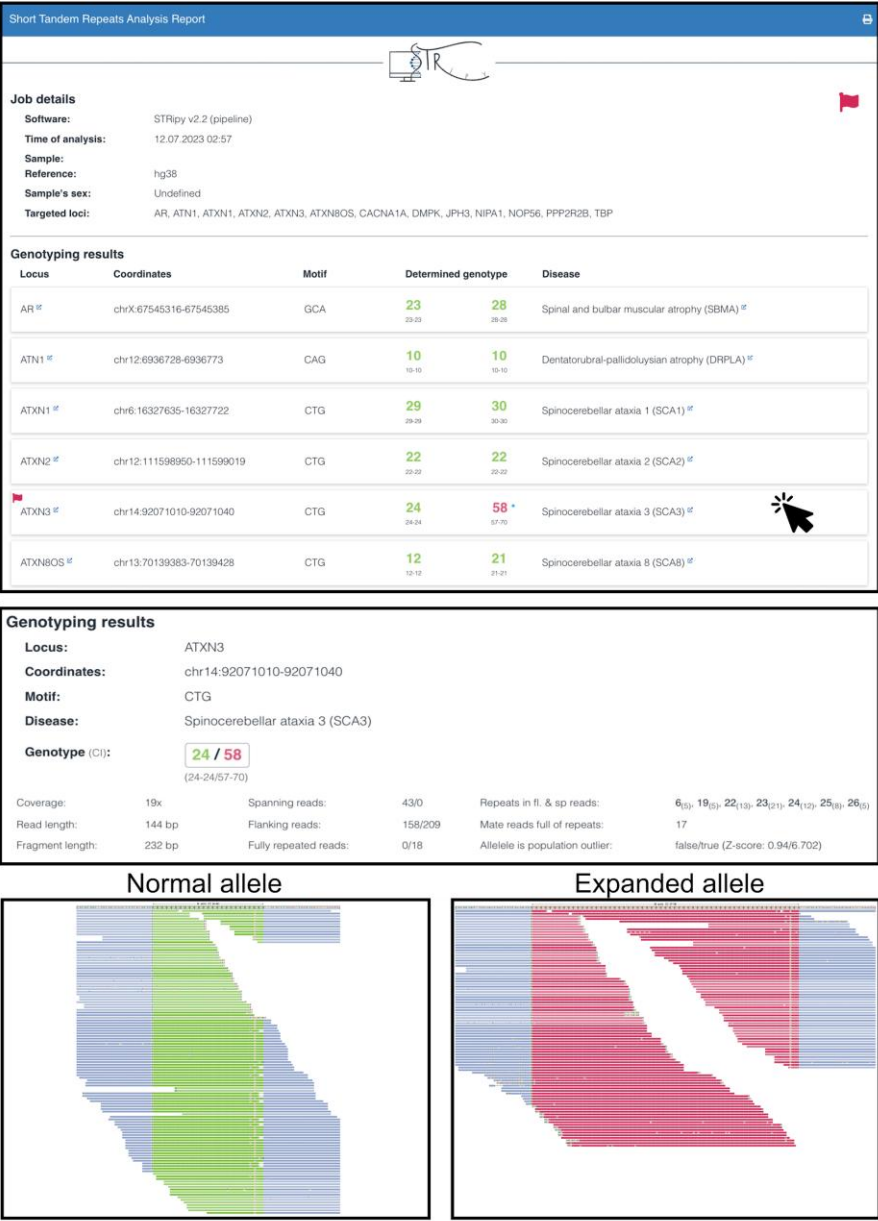


Figure 1 Integration of STRipy into the PathWest diagnostic analysis workflow. For Version 6 of the panel, STRipy was run retrospectively on existing sequencing data. This figure illustrates the updated diagnostic workflow used for Version 7 onwards. STRipy was integrated such that it runs in parallel with SNV/Indel and CNV analysis. The STRipy pipeline takes BAM files as input and generates user-friendly html reports. Abbreviations: VCF = variant call file; BAM = binary alignment map; STR = short tandem repeat; PCR = polymerase chain reaction; SNV = single nucleotide variant; CNV = copy number variant. Created in BioRender. Folland, C. (2025) <https://BioRender.com/jk55i8g>.

A STRipy html Report



B Orthogonal Validation

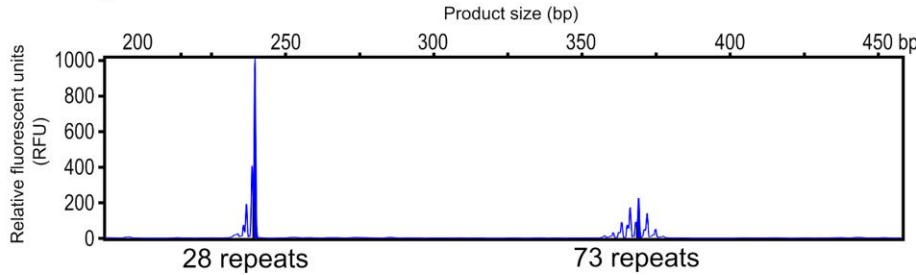


Figure 2 Example of expansion detected by STRipy. (A) Example html report generated by STRipy from panel data. This report shows a pathogenic expansion at the SCA3 locus. The html report summarizes genotyping results from all loci screened by STRipy. Any loci with expansions that pass the pathogenic threshold are labelled with a red flag icon that is easy to identify. The report is interactive; selecting a tab of interest produces a drop-down view including the genotype, coverage, read length and the number of spanning, flanking and fully repeated reads. There are also REViewer visualizations of the normal and expanded allele, which assist with interpretation of the result. (B) Orthogonal validation of the SCA3 repeat expansion using flanking PCR confirms presence of an expanded allele at 73 repeats.

any intermediate or pathogenic results for all samples in a sequencing run. The html files show the gene, motif, coordinates, estimated number of repeats, threshold indicators, disease name and repeat alignment map for each STR locus analysed.

Expansion controls

Available positive control samples with expansions that had been detected and sized through standard NATA-accredited diagnostic workflows were tested on Version 7 of the panel. The data were analysed by the STRipy pipeline to determine whether the software was able to accurately detect known pathogenic expansions.

Running STRipy on panel data

Version 6: Sequencing data generated from all ataxia samples previously tested on Version 6 of the panel ($n = 418$, 197 XX, 222 XY) were retrospectively tested using the STRipy pipeline.

Version 7: Following the incorporation of additional probes to Version 7 of the panel, the STRipy pipeline was implemented for routine screening of sequencing data for all patients tested on the panel ($n = 455$, 220 XX, 235 XY), including 67 (36 XX, 31 XY) ataxia patients.

Validation of STR expansions detected by STRipy

DNA samples from patients where STRipy indicated the presence of an expanded allele from the panel data were tested using the standard diagnostic workflow for repeat disorders to confirm the presence and size of the expanded allele. Individual techniques are detailed in the [Supplementary Materials: Methods](#) and [Supplementary Table 3](#).

Assessment of expansion in *C9orf72*

Due to the high prevalence of *C9orf72* expansions in MND/FTD patients, the diagnostic workflow first screens for the expansion,

and patients testing negative for the expansion then cascade to gene panel testing. Eighty-two samples from clinically diagnosed MND/FTD patients negative for the *C9orf72* expansion were tested on Version 7 of the panel and analysed using flanking PCR to size alleles in the normal range. Samples indicating a single normal allele were tested by RP-PCR to determine whether the patient was homozygous for the normal sized allele or harboured an expansion ([Supplementary Materials: Methods](#)).

Statistical analysis

Statistical comparison of STRipy output and validated allele sizes for the *C9orf72* repeat region was performed with Prism (GraphPad Software, Version 10.2.0(335)). The correlation between validated allele sizes at the *C9orf72* locus and STRipy predictions made with four or more spanning reads was assessed using Pearson's product-moment correlation.

Results

Retrospective screening of the ataxia samples tested on Version 6 of the panel

Of the 418 patient samples tested on Version 6 of the panel, 12% ($n = 50$) had genetically resolved outcomes following standard analysis for indels, SNVs and CNVs. The cohort data was then run through the STRipy pipeline. STRipy indicated the presence of pathogenic expansions in 11 samples across multiple loci: four expansions associated with SCA6, three expansions associated with SCA36, and to our surprise two expansions associated with SCA1, and two expansions associated with SCA3 ([Table 1](#)). We did not expect to observe coverage of *ATXN1* and *ATXN3*, associated with SCA1 and SCA3, as probes for these loci were not included in Version 6 of the panel. Inspection of the REviewer outputs and BAM files at the SCA1 and SCA3 loci for these samples confirmed the presence of correctly mapped reads, with

Table 1 Pathogenic expansions detected on the Version 6 ataxia subpanel

Sample ID	Sex	Locus details				STRipy output				Confirmed size	
		Locus	Normal Range	Pathogenic threshold	Repeat	Allele 1 (repeats)	Spanning reads	Allele 2 (repeats)	Spanning reads	Allele 1 (repeats)	Allele 2 (repeats)
V6POS1	F	SCA1	6–32	≥39	CAG	30 (30–30)	35	50 (49–54)	0	30 (±1)	42 (±1)
V6POS2	M	SCA1	6–32	≥39	CAG	29 (29–29)	31	44 (44–44)	4	29 (±1)	44 (±1)
V6POS3	F	SCA3	12–44	≥56	CAG	24 (24–24)	43	58 (57–70)	0	28 (±1) ^a	73 (±2) ^a
V6POS4	F	SCA3	12–44	≥56	CAG	15 (15–15)	90	51 (51–60)	0	19 (±1) ^a	76 (±2) ^a
V6POS5	M	SCA36	3–14	≥25	GGCCTG	6 (6–6)	34	21 (19–23)	0	7 (±1)	>90
V6POS6	M	SCA36	3–14	≥25	GGCCTG	3 (3–3)	47	27 (27–31)	0	4 (±1)	>90
V6POS7	F	SCA36	3–14	≥25	GGCCTG	10 (10–10)	40	24 (24–27)	0	10 (±1)	>90
V6POS8	M	SCA6	4–18	≥21	CAG	13 (13–13)	44	22 (22–22)	43	13 (±1)	22 (±1)
V6POS9	F	SCA6	4–18	≥21	CAG	12 (12–12)	44	22 (22–22)	40	12 (±1)	22 (±1)
V6POS10	F	SCA6	4–18	≥21	CAG	13 (13–13)	43	22 (22–22)	30	13 (±1)	22 (±1)
V6POS11	F	SCA6	4–18	≥21	CAG	13 (13–13)	38	22 (22–22)	38	13 (±1)	22 (±1)

^aATXN3 (associated with SCA3) has a polymorphic region directly preceding the repeat locus where SNPs common in the population can result in additional repeats. Standard diagnostic practice includes this region for calculating the number of repeats from the PCR assay as the sequence is not detected by this technique. This results in four additional repeats calculated by the PCR assay compared to STRipy.

sequence matching that of the region flanking the repeat. Expansions in both SCA1 (*ATXN1*) and SCA3 (*ATXN3*) consist of CAG repeat motifs.^{5,6} This is also the repeat motif present in SCA6 (*CACNA1A*), SCA12 (*PPP2R2B*) and SCA17 (*TBP*),^{5,6} all of which are targeted in Version 6 of the panel. We hypothesize that non-specific binding of these probes captures the SCA1 and SCA3 loci, including them into the enriched library where they undergo clonal amplification and sequencing. This would provide enough coverage of these regions for STRipy to call the presence of an expansion at these loci.

All expansions detected by STRipy, including for SCA1 and SCA3, were confirmed and accurately sized by orthogonal diagnostic methods. The concordance between allele sizes determined by STRipy and the confirmatory diagnostic assays is shown in Table 1.

The implementation of the STRipy pipeline for the few loci covered by Version 6 of the panel alone increased the diagnostic yield of the ataxia *in silico* panel from 12.0% to 14.6% overall ($n = 61$), with expansion detected by STRipy contributing 18.0% of the solved cases. Of note, STRipy detected a *NOP56* expansion in one patient exhibiting both ataxia and neuropathy with a strong autosomal dominant family history. The patient had previously been shown to harbour a *PMP22* duplication, following original panel data analysis, and therefore was unlikely to undergo further testing.

Implementation of STRipy pipeline for Version 7 of the neurological gene panel

In Version 7 of the panel, we incorporated additional oligos to target repeat loci that were previously not covered on Version 6 (Supplementary Table 1). We first tested known positive controls that were available for the expansions associated with DRPLA, HD, SBMA, SCA2, SCA7, SCA8, SCA27B and the *C9orf72* repeat associated with MND/FTD. Excluding the *C9orf72* expansion, STRipy indicated the presence of an expansion for all positive control samples tested, with varying degrees of accuracy in relation to the repeat size (Table 2).

Since implementation of Version 7 of the panel, 455 samples have been screened, of which 67 samples were analysed on the *in silico* ataxia subpanel. The overall diagnostic yield for the ataxia panel was 22.4% ($n = 15$), including three samples with expansions detected by STRipy (SCA1, SCA2 and SCA3), confirmed and sized by accredited diagnostic testing (Table 3). Incorporating STRipy into the analysis workflow increased diagnostic yield for the ataxia subpanel from 17.9% to 22.4%, contributing 20% of solved cases.

In addition to this, STRipy was able to detect the presence of, and accurately genotype, the CAG repeat expansion associated with HD in a sample that was tested on the dystonia/Parkinson's disease subpanel (Table 3).

Accuracy of STRipy calls and utility for screening

STRipy displayed greater accuracy for loci with shorter pathogenic repeat thresholds, where the size of the expansion was smaller than the read length. All pathogenic expansions called by STRipy with

Table 2 Control samples tested on Version 7 of the neurological gene panel

Sample ID	Sex	Locus	Locus details			STRipy output				Confirmed size	
			Normal Range	Pathogenic threshold	Repeat	Allele 1 (repeats)	Spanning reads	Allele 2 (repeats)	Spanning reads	Allele 1 (repeats)	Allele 2 (repeats)
MNDCNTRL	F	MND/FTD	1-19	>24	GGGGCC	2 (2-2)	32	19 (16-23)	0	2 (±1)	>50
DRPLACNTRL	F	DRPLA	7-23	≥49	CAG	10 (10-10)	125	49 (49-56)	0	10 (±1)	68 (±1)
HDCTRL1	F	HD	9-26	≥36	CAG	20 (20-20)	78	40 (40-40)	8	20 (±1)	40 (±1)
HDCTRL2	M	HD	9-26	≥36	CAG	22 (22-22)	76	53 (53-65)	0	22 (±1)	98 (±1)
SBMACNTRL	M	SBMA	17-35	≥40	CAG	42 (42-42)	12	NA	NA	41 (±1)	NA
SCA2CNTRL	M	SCA2	13-31	≥35	CAG	22 (22-22)	87	37 (37-37)	19	22 (±1)	36 (±2)
SCA7CNTRL	M	SCA7	4-35	≥36	CAG	13 (13-13)	25	49 (49-52)	0	13 (±1)	46 (±2)
SCA8CNTRL	F	SCA8	2-37	≥80	CTG	59 (43-66)	0	70 (58-83)	0	101 (±3)	154 (±3)

Table 3 Pathogenic expansions detected on Version 7 of the neurological panel

Sample ID	Sex	Locus details				STRipy output				Confirmed size	
		Locus	Normal Range	Pathogenic threshold	Repeat	Allele 1 (repeats)	Spanning reads	Allele 2 (repeats)	Spanning reads	Allele 1 (repeats)	Allele 2 (repeats)
V7POS1	M	SCA1	6–32	≥39	CAG	30 (30–30)	53	48 (48–51)	0	31 (±1)	42 (±1)
V7POS2	F	SCA2	13–31	≥35	CAG	22 (22–22)	46	38 (38–38)	13	22 (±1)	38 (±2)
V7POS3	M	SCA3	12–44	≥56	CAG	18 (18–18)	148	51 (51–60)	0	22 (±1) a	77 (±1) a
V7POS4	M	HD	9–26	≥36	CAG	28 (28–28)	46	40 (40–40)	11	28 (±1)	40 (±1)

^aATXN3 (associated with SCA3) has a polymorphic region directly preceding the repeat locus where SNPs common in the population can result in additional repeats. Standard diagnostic practice includes this region for calculating the number of repeats from the PCR assay as the sequence is not detected by this technique. This results in four additional repeats calculated by the PCR assay compared to STRipy.

spanning reads were completely concordant with PCR-based sizing (Tables 1–3).

For some larger pathogenic expansions, STRipy could be used in a screening capacity to trigger further testing and accurate sizing. For example, STRipy could reliably size smaller alleles and indicate the presence of larger expansions associated with SCA1, SCA3 and SCA36, which could then be validated and sized using orthogonal methods.

For loci with a broader range of non-pathogenic allele sizes, such as *FGF14* and *RFC1*, the utility of STRipy is limited to the detection of smaller, normal alleles to exclude these loci as differential diagnoses, where indicated by the phenotype. The *RFC1* locus presents an added challenge for detecting pathogenic alleles as there are multiple repeat motifs associated with the locus. When alternate repeats are present, the coverage of the locus drops, likely due to reduced target enrichment as probes compliment the reference sequence. We have previously shown that manually assessing the BAM at this locus is useful for screening for the presence of biallelic expansions.¹⁵ STRipy improves the genotyping of normal alleles but is also limited by sequence coverage and read length for the assessment of expanded alleles. Though not comprehensive, this screening can reduce the number of samples that require testing by orthogonal methods.

Pathogenic expansions in *C9orf72* are the underlying cause of a significant proportion of MND/FTD, including up to 50% of familial cases and 10% of sporadic cases.¹⁶ The pathogenic threshold for *C9orf72* expansions is generally reported as >60 hexanucleotide repeats, though this threshold is not well defined, and expansions often reach thousands of repeats.¹⁷ REviewer visualizations for this locus indicated poor mapping quality within some samples analysed by STRipy, prompting further assessment.

Patients clinically diagnosed with MND/FTD are currently tested for the expansion by a PCR-based assay prior to panel testing. Eighty-two patients tested negative following flanking PCR and were consequently tested on Version 7 of the panel to identify alternate causes of disease. This enabled comparison of *C9orf72* non-pathogenic genotypes between STRipy and the PCR-based assay.

There was substantial discordance between predicted STRipy genotypes and PCR-based assessment of *C9orf72* in several samples, making interpretation of STRipy calls for this locus challenging. The *C9orf72* locus shows greater variability in the number of spanning reads for smaller alleles than expected (Fig. 3). This was reflected in the discordant results seen for some samples with smaller alleles (Fig. 3) and may indicate inconsistencies with

sequencing or read mapping in this GC-rich region. Alleles with four or more spanning reads at the *C9orf72* locus ($n = 50$) showed a high level of correlation ($r = 0.9925$, $P < 0.0001$) between STRipy genotype calls and PCR-based sizing (Fig. 3).

Discussion

STR expansions are a common cause of multiple inherited neurological disorders.⁷ Presently, molecular diagnostics for STR expansions involve individual, repeat specific assays. Testing for these disorders therefore relies heavily on clinical assessment, which is often confounded by the phenotypic heterogeneity and clinical overlap between disorders.⁵ This often results in iterative testing, prolonging patient uncertainty. Due to lack of feasibility, diagnostic testing for many of the repeat loci is not routinely available at diagnostic laboratories in Australia. This requires send-away of samples and screening by an international service provider, often with considerable costs and lengthy time delays.

Here we used STRipy to screen short-read panel data for pathogenic expansions in a diagnostic setting. We demonstrate that integration of STRipy into panel analysis provides a high-throughput method for screening multiple STR expansion loci concurrently and significantly improves diagnostic yield for ataxia patients. Expansions were identified in 2.6% and 4.5% of all patients tested on Versions 6 and 7 of the ataxia subpanel, improving the diagnostic yield of these panels by 18% and 20%, respectively. This is comparable to other studies that have used ExpansionHunter to analyse exome sequencing data, identifying STR expansions in 0.24–4.4% of cases.^{18–20} We were also able to identify a pathogenic expansion associated with HD in a patient tested on the dystonia/Parkinson's subpanel. The HD phenotype has considerable heterogeneity,^{21,22} which can lead to a delay in diagnosis for patients seeking informed family planning and clinical management. Our results show that detection of STR expansions using panel data is comparable to that of exome data and demonstrate that integrating STR calling into diagnostic analysis pipelines can increase diagnostic yield and improve testing efficiency.

Both read length and locus coverage impact detection of STR expansions from short-read sequencing.^{18,23} The primary advantages of targeted gene panels are their adaptability and the depth of coverage afforded. The high read depth provides greater confidence in the calls made. STRipy performed best for loci with lower pathogenic thresholds, showing complete concordance with orthogonal

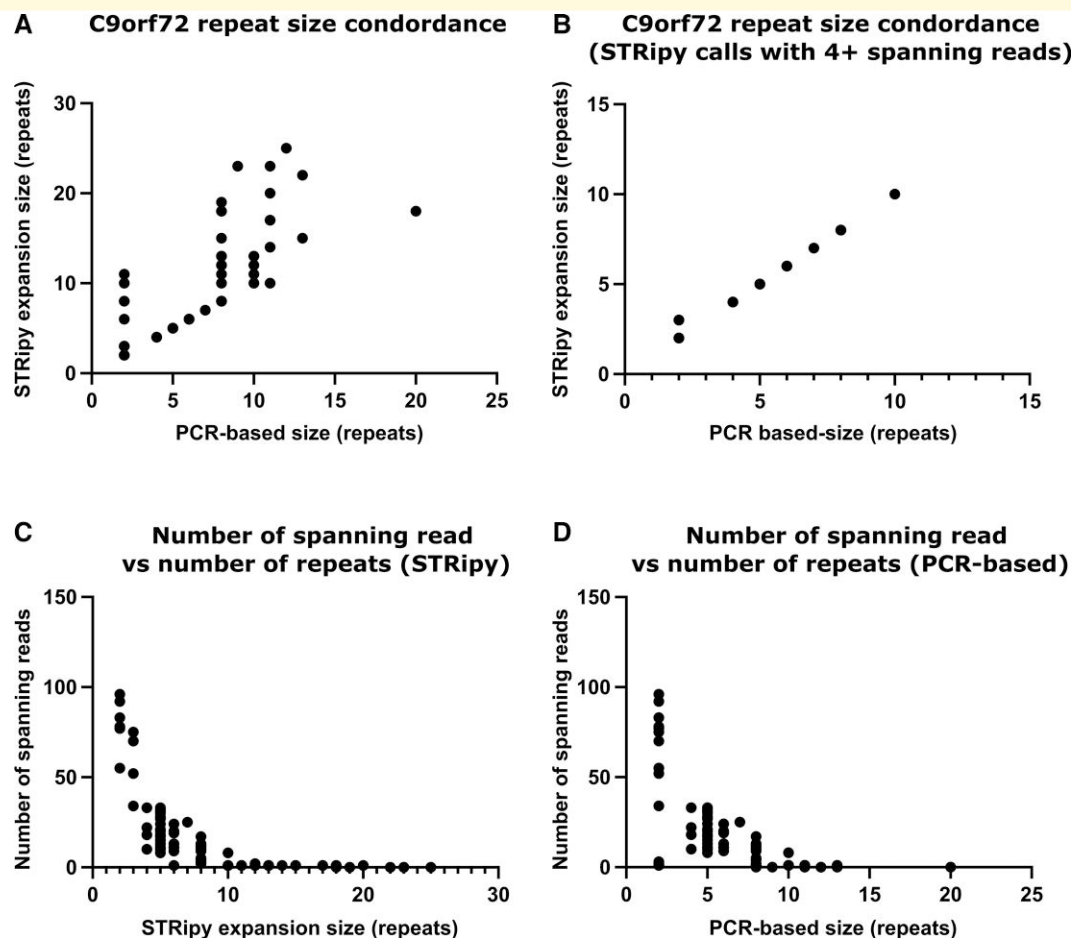


Figure 3 C9orf72 repeat size concordance. All graphs relate to the C9orf72 repeat locus associated with MND/FTD, each data point represents allele 2 from a patient tested for C9orf72 ($n = 82$). (A) Displays the number of repeats predicted by STRipy compared to the validated allele size, using flanking PCR. (B) Shows the correlation between the number of repeats predicted by STRipy compared to the validated allele size using flanking PCR but only includes alleles where four or more spanning reads were present for the STRipy prediction ($n = 50$). Pearson's correlation coefficient was 0.9925, $P < 0.0001$, indicating a strong correlation between STRipy genotype calls and PCR-based sizing. (C) Displays the number of spanning reads available and the number of repeats predicted by STRipy for each allele. (D) Shows the number of spanning reads utilized for the STRipy prediction compared to the validated allele size using flanking PCR.

techniques where the size of the expanded allele was smaller than the read length.

STRipy was also able to successfully identify expansions in *ATXN1*, *ATXN3* and *NOP56* that exceeded the read length. Though unable to size larger expansions, the ability to accurately genotype smaller alleles provides a useful screening tool for expansions that have a narrow range for the normal allele present within the population. For example, pathogenic expansions in the *NOP56* gene associated with SCA36 generally exceed 1000 hexanucleotide repeats (~6 kb); however, the reference range within population controls is between 3 and 14 repeats.^{24,25} In all *NOP56* cases detected by STRipy above this range, no spanning reads were present for the larger allele, indicating that the expansions were likely larger than predicted, and this was confirmed by orthogonal methods. As such, alleles indicated by STRipy to harbour more repeats than the upper bound of the reference range, particularly where no spanning reads are mapped, warrant further testing. For large expansions where the normal range is broader and may exceed read length,

such as *FGF14*, *C9orf72* and *RFC1*, the utility of STRipy is limited to excluding the presence of an expansion where two normal alleles are detected with adequate spanning reads.

The capabilities of any analysis software are limited by the type and quality of data input. The ability to size larger expansions from the targeted panel data was reduced when compared to genome-wide sequencing data.¹⁴ From the alignment file, STRipy extracts all reads that overlap the STR locus and a 2 kb region flanking the repeat.¹⁴ Next, reads directly flanking the STR region, as well as fully repeated reads that have a pair in the extracted region, are included in the analysis file.¹⁴ The probes used here to capture regions of interest for the panel target regions of ~200 bp. Targeting larger regions encompassing the repeat may improve capacity to resolve larger expansions from panel data by generating more reads with in-repeat read mates. This may also produce more accurate genotypes for the *C9orf72* locus.

The phenotypic heterogeneity observed across numerous STR expansion disorders complicates differential diagnosis and

individual test selection. Integrating STR calling into short-read data analysis provides a comprehensive first-line screen. The testing captures a broader range of patient phenotypes, including patients with fewer differential features indicative of specific STR expansions who may not have otherwise been tested. One expansion that is likely to have been missed was discovered through the retrospective screening of ataxia patients tested on Version 6 of the panel. This patient had a pathogenic expansion in *NOP56*, in addition to a reported *PMP22* duplication associated with Charcot–Marie–Tooth (CMT) neuropathy. Clinical notes for the patient indicated both neuropathic and ataxic phenotypes. However, without the additional screening, all clinical features would have been attributed to the *PMP22* duplication, hindering informative familial testing. Dual diagnoses are often difficult to determine due to a blended phenotype. Despite this they have been reported to contribute up to 7% of diagnoses in patients tested using exome sequencing.^{26–28} As genomic testing becomes more comprehensive, identification of dual diagnoses may increase, improving the ability to delineate blended phenotypes.

Integration of STR screening tools into short-read MPS workflows also improves accessibility to STR expansion testing. Accredited diagnostic testing for *SCA36* expansions in *NOP56* is not currently available in Australia and many other countries.²⁹ Prior to implementing STRipy into the sequencing analysis pipeline, our laboratory did not offer testing for *SCA36*. STRipy detected pathogenic *NOP56* expansions in three ataxia patients tested on Version 6 of the panel, contributing ~5% of all genetically resolved ataxia cases. MPS gene panels are the first-line test for neurological disorders in many diagnostic laboratories. Since their widespread adoption, significant advances have been made in the analysis of short-read MPS data. Incorporation of additional tools is important to fully utilize data already generated. These tools provide substantial value without significant cost. Implementing STR detection software into diagnostic analysis pipelines facilitates screening for *SCA36* and other rare STR expansions where routine diagnostic testing may not be readily available. Command-line implementation of STRipy enabled concurrent genotyping of all repeat loci for every sample tested. Individual reports were produced, which display all genotypes and alignment visualizations in a user-friendly and interactive display. This enabled high-throughput screening of STR expansions with minimal additional training or workload for analysts.

Conclusion

Here, we demonstrate that screening pathogenic STR expansions using STRipy on a neurological gene panel accurately detects expansions and improves the diagnostic success rate for ataxias in a clinical diagnostic laboratory. STRipy offers a simple computational approach that comprehensively screens many loci in a single workflow, improving testing efficiency and reducing the cycles of iterative testing. This is particularly beneficial for diagnosing atypical phenotypes, identifying dual diagnoses, and for rare disorders where testing may not be routinely available. Preliminary work has also demonstrated the diagnostic utility of STRipy for diagnosing muscle-related repeat expansions from panel data, including

identifying patients with pathogenic expansions in *DMPK* and *CNBP*, associated with myotonic dystrophy types 1 and 2 (DM1 and DM2).

Supplementary material

Supplementary material is available at *Brain Communications* online.

Acknowledgements

We thank Andreas Halman for useful discussions and advice regarding installation and running of the STRipy pipeline. We thank the patients and their clinicians for their involvement with this work. The graphical abstract was created in BioRender: Scriba (2026) <https://BioRender.com/j0ki9jg>.

Funding

GR is supported by an National Health and Medical Research Council (NHMRC) Emerging Leadership level 2 (EL2) Fellowship (APP2007769). C.K.S. and C.F. are supported by the Australian Government research training program fees offset and stipend scholarship. C.K.S. is supported by the top-up scholarship from the Australasian Genomic Technologies Association. C.F. is supported by Jean Rogerson HDR Scholarship and the Jock and Marjorie Hetherington HDR Top-Up Scholarship. Medical Research Future Fund (MRFF) Genomics Health Futures Mission Grant (2007681).

Competing interests

The authors report no competing interests.

Data availability

The data that support the findings of this study are available upon request from the corresponding author.

References

- McDonald CM. Clinical approach to the diagnostic evaluation of hereditary and acquired neuromuscular diseases. *Phys Med Rehabil Clin N.* 2012; 23(3):495–563.
- Pang SYY, Teo KC, Hsu JS, *et al.* The role of gene variants in the pathogenesis of neurodegenerative disorders as revealed by next generation sequencing studies: A review. *Transl Neurodegener.* 2017;6:27.
- Beccroft SJ, Yau KS, Allcock RJN, *et al.* Targeted gene panel use in 2249 neuromuscular patients: The Australasian referral center experience. *Ann Clin Transl Neur.* 2020;7(3):353–362.
- Schofield D, Alam K, Douglas L, *et al.* Cost-effectiveness of massively parallel sequencing for diagnosis of paediatric muscle diseases. *Npj Genom Med.* 2017;2:4.
- Chintalaphani SR, Pineda SS, Deveson IW, Kumar KR. An update on the neurological short tandem repeat expansion disorders and the emergence of

- long-read sequencing diagnostics. *Acta Neuropathol Commun.* 2021;9(1):98.
6. Rafehi H, Bennett MF, Bahlo M. Detection and discovery of repeat expansions in ataxia enabled by next-generation sequencing: Present and future. *Emerg Top Life Sci.* 2023;7(3):349-359.
 7. Ibanez K, Polke J, Hagelstrom RT, *et al.* Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: A retrospective diagnostic accuracy and prospective clinical validation study. *Lancet Neurol.* 2022;21(3):234-245.
 8. 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.
 9. 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.
 10. Dolzhenko E, van Vugt J, Shaw RJ, *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* 2017;27(11):1895-1903.
 11. Dolzhenko E, Weisburd B, Ibanez K, *et al.* REViewer: Haplotype-resolved visualization of read alignments in and around tandem repeats. *Genome Med.* 2022;14(1):84.
 12. Dashnow H, Lek M, Phipson B, *et al.* STRetch: Detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* 2018;19(1):121.
 13. Dashnow H, Pedersen BS, Hiatt L, *et al.* STRling: A k-mer counting approach that detects short tandem repeat expansions at known and novel loci. *Genome Biol.* 2022;23(1):257.
 14. Halman A, Dolzhenko E, Oshlack A. STRipy: A graphical application for enhanced genotyping of pathogenic short tandem repeats in sequencing data. *Hum Mutat.* 2022;43(7):859-868.
 15. Scriba CK, Stevanovski I, Chintalaphani SR, *et al.* RFC1 in an Australasian neurological disease cohort: Extending the genetic heterogeneity and implications for diagnostics. *Brain Commun.* 2023;5(4):fcad208.
 16. Umoh ME, Fournier C, Li Y, *et al.* Comparative analysis of C9orf72 and sporadic disease in an ALS clinic population. *Neurology.* 2016;87(10):1024-1030.
 17. Gossye H, Engelborghs S, Van Broeckhoven C, van der Zee J. C9orf72 frontotemporal dementia and/or amyotrophic lateral sclerosis. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, eds. *GeneReviews(R)* [Internet]. University of Washington, Seattle; 1993.
 18. Yoon JG, Lee S, Cho J, *et al.* Diagnostic uplift through the implementation of short tandem repeat analysis using exome sequencing. *Eur J Hum Genet.* 2024;32:584-587.
 19. van der Sanden BPGH, Corominas J, de Groot M, *et al.* Systematic analysis of short tandem repeats in 38,095 exomes provides an additional diagnostic yield. *Genet Med.* 2021;23(8):1569-1573.
 20. Mereaux JL, Davoine CS, Coutelier M, *et al.* Fast and reliable detection of repeat expansions in spinocerebellar ataxia using exomes. *J Med Genet.* 2023;60(7):717-721.
 21. Ross CA, Tabrizi SJ. Huntington's disease: From molecular pathogenesis to clinical treatment. *Lancet Neurol.* 2011;10(1):83-98.
 22. Mehrabi NF, Waldvogel HJ, Tippet LJ, Hogg VM, Synek BJ, Faull RL. Symptom heterogeneity in Huntington's disease correlates with neuronal degeneration in the cerebral cortex. *Neurobiol Dis.* 2016;96:67-74.
 23. Tang H, Kirkness EF, Lippert C, *et al.* Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes. *Am J Hum Genet.* 2017;101(5):700-715.
 24. Kobayashi H, Abe K, Matsuura T, *et al.* Expansion of intronic GGCCTG hexanucleotide repeat in causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. *Am J Hum Genet.* 2011;89(1):121-130.
 25. García-Murias M, Quintáns B, Arias M, *et al.* Costa da Morte' ataxia is spinocerebellar ataxia 36: Clinical and genetic characterization. *Brain.* 2012;135:1423-1435.
 26. Posey JE, Rosenfeld JA, James RA, *et al.* Molecular diagnostic experience of whole-exome sequencing in adult patients. *Genet Med.* 2016;18(7):678-685.
 27. Balci TB, Hartley T, Xi Y, *et al.* Debunking occam's razor: Diagnosing multiple genetic diseases in families by whole-exome sequencing. *Clin Genet.* 2017;92(3):281-289.
 28. Wojcik MH, Reuter CM, Marwaha S, *et al.* Beyond the exome: What's next in diagnostic testing for Mendelian conditions. *Am J Hum Genet.* 2023;110(8):1229-1248.
 29. Rafehi H, Szmulewicz DJ, Pope K, *et al.* Rapid diagnosis of spinocerebellar ataxia 36 in a three-generation family using short-read whole-genome sequencing data. *Mov Disord.* 2020;35(9):1675-1679.