Emerging Topics in Life Sciences (2023)
https://doi.org/10.1042/ETLS20230018

Review Article

Detection and discovery of repeat expansions in ataxia enabled by next-generation sequencing: present and future

Haloom Rafehi[1,2], Mark F. Bennett[1,2,3] and Melanie Bahlo[1,2]

[1]Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia; [2]Department of Medical Biology, University of Melbourne, Parkville, VIC, Australia; [3]Epilepsy Research Centre, Department of Medicine, University of Melbourne, Austin Health, Heidelberg, VIC, Australia

Correspondence: Haloom Rafehi (rafehi.h@wehi.edu.au)

Hereditary cerebellar ataxias are a heterogenous group of progressive neurological disorders that are disproportionately caused by repeat expansions (REs) of short tandem repeats (STRs). Genetic diagnosis for RE disorders such as ataxias are difficult as the current gold standard for diagnosis is repeat-primed PCR assays or Southern blots, neither of which are scalable nor readily available for all STR loci. In the last five years, significant advances have been made in our ability to detect STRs and REs in short-read sequencing data, especially whole-genome sequencing. Given the increasing reliance of genomics in diagnosis of rare diseases, the use of established RE detection pipelines for RE disorders is now a highly feasible and practical first-step alternative to molecular testing methods. In addition, many new pathogenic REs have been discovered in recent years by utilising WGS data. Collectively, genomes are an important resource/platform for further advancements in both the discovery and diagnosis of REs that cause ataxia and will lead to much needed improvement in diagnostic rates for patients with hereditary ataxia.

Introduction

Hereditary cerebellar ataxias are a heterogenous group of progressive neurological disorders characterised by significant morbidity and mortality. The prevalence of hereditary ataxia ranges from 1.5 to 4.9 cases per 100 000 individuals [1]. While primarily affecting adults, ataxias can also have prenatal, childhood and adolescent onset. Ataxia is primarily a gait movement disorder and usually presents with dysarthria, dysmetria, and impaired oculomotor control. Other frequently co-occurring symptoms include parkinsonism, dementia, dystonia, chorea, vestibulopathy, sleep disorders, peripheral neuropathy, pyramidal symptoms such as weakness and spasticity, ocular abnormalities such as nystagmus or oculomotor apraxia and deafness [2].

The genetics of hereditary ataxias is uniquely complex. While sometimes caused by *de novo* or inherited rare and deleterious mutations, it is most associated with repeat expansion (RE) of short tandem repeats (STRs) [3]. STRs (also known as microsatellites) are repetitive elements of DNA in which motifs 2 to 6 base pairs (bp) in length are repeated in tandem. STR lengths are highly variable between individuals due to their inherent instability, however most variation in STR length is benign. There are currently over 50 REs known to cause disease and of which 18 cause ataxias (summarised in Table 1).

Historically, the discovery and diagnosis of RE disorders has been made difficult by the repetitive nature of the DNA sequence. In the era of clinical exomes and genomes for diagnosis of genetic disorders, the gold standard for diagnosis of ataxias caused by REs remains repeat-primed PCR assays or Southern blots, neither of which is scalable nor readily available for all STR loci. Testing is also

Received: 20 June 2023
Revised: 29 August 2023
Accepted: 12 September 2023

Version of Record published:
21 September 2023

Table 1 Overview of ataxias caused by pathogenic REs[1]

Year of discovery	Disorder name	Inheritance type	chr	gene	location	pathogenic motif	normal repeat range	pathogenic range	mechanism	discovery method	citation
1993	SCA1	AD	6p22	*ATXN1*	exon	CAG	6–38	≥39–88	polyQ, RNA?	linkage; expansion screening	[57]
1994	SCA3	AD	14q32	*ATXN3*	exon	CAG	12–44	≥55–87	polyQ, RNA (foci)	linkage; cloning	[58]
1994	DRPLA	AD	12p13.31	ATN1	exon	CAG	3–35	>–48–93	polyQ, RNA?	linkage; expansion screening	[59,60]
1996	SCA2	AD	12q24	*ATXN2*	exon	CAG	13–31	≥32–500	polyQ, RNA?	linkage; cloning	[47]
1996	SCA7	AD	3p21	*ATXN7*	exon	CAG	4–33	≥37–460	polyQ, RNA?	linkage; cloning	[61]
1996	FRDA	AR	9q21.11	*FXN*	intron	GAA	5–34	≥66–1300	gene silencing	linkage; expansion screening	[62]
1997	SCA6	AD	19p13	*CACNA1A*	exonic	CAG	4–18	≥20–33	polyQ, RNA?	linkage; expansion screening	[63]
1999	SCA17	AD	6q27	*TBP*	exonic	CAG or CAG/CAA	25–40	≥43–66	polyQ, RNA?	linkage; candidate gene analysis	[36]
1999	SCA8	AD	13q21	*ATXN8 (ATXN8OS)*	3’UTR	CAG/CTG	15–50	>74–250	RNA (foci), RAN	linkage; cloning	[64]
1999	SCA12	AD	5q31	*PPP2R2B*	5’UTR	CAG	4–32	≥43–78	RAN (polyG)?	linkage; repeat expansion detection	[65]
2000	SCA10	AD	22q13	*ATXN10*	intron	ATTCT/ATTGT	10–32	>280–4500	RNA (foci)	linkage; expansion screening	[66]
2009	SCA31	AD	16q22	*BEAN1 (TK2)*	intron	(TAAAA), TGGAA/TAGAA	?	≥110–760	RNA (foci, PS), RAN	linkage	[67]
2011	SCA36	AD	20p13	*NOP56*	intron	GGCCTG	5–14	≥650–2500	RNA (foci)	linkage; expansion screening	[68]
2017	SCA37	AD	1p32	*DAB1*	intron	(ATTTT), ATTTC	7–400 (ATTTT)	≥31–75 (ATTTC)	RNA	linkage; expansion screening	[69]
2019	CANVAS	AR	4p14	*RFC1*	intron	AAGGG, ACAGG, AAAGG-AAGGG-AAAGG	-	≥400–2000	unknown	linkage; WGS	[19,32]
2019	GDPAG	AR	2q32.2	*GLS*	5’UTR	CAG	8–16	≥680–1400	gene silencing	candidate gene analysis	[43]
2023	SCA27B	AD	13q33.1	*FGF14*	intron	AAG	10–250	≥300	reduced gene expression	linkage; WGS	[21,34]
2023	-	AD	16q22.1	*THAP11*	exon	CAG	20–38	≥45–100	PolyQ	linkage; LRS	[35]

[1] Table adapted from review paper [70].

limited to the most common REs. Diagnosis of RE disorders is further complicated by variable phenotypes which can overlap with other, more common disorders such as Parkinson's Disease or amyotrophic lateral sclerosis (ALS) [4].

Detection and diagnosis of RE disorders

The advent of whole exome and genome sequencing (WES/WGS) has accelerated the diagnosis and discovery of rare genetic diseases. Accessibility of WES and more recently WGS for clinical diagnosis of rare diseases is constantly increasing [5], however screens remain limited mostly to SNVs and small indels, as conducted by clinical genomics bioinformatics pipelines. In recent years it has become increasingly feasible to detect and accurately size REs using WGS, and to a lesser extent, WES. Early iterations of analysis tools struggled with the repetitive nature of STRs, however this hurdle has been largely overcome since the development of catalogue-based methods such as ExpansionHunter [6], gangSTR [7], STRetch [8] and exSTRa [9], and more recently, visualisation tools such as REViewer [10]. ExpansionHunter, gangSTR and STRetch provide an exact genotype for STRs shorter than the read length (typically 150 bp) or an estimated size for longer STRs, although STRetch has a higher computational burden due to its requirement to make use of an alternative, augmented reference genome. In contrast, exSTRa is an outlier method that determines whether a specific STR is expanded compared with other individuals. These methods are catalogue based, i.e. they will only screen pre-defined STRs and motifs and have been used with great success to diagnose pathogenic REs in disease cohorts [4,11–13]. For example, we recently diagnosed SCA36 in a multigenerational Australian pedigree using ExpansionHunter and exSTRa [14]. SCA36 is a rare form of ataxia caused by an intronic GGCCTG RE in *NOP56,* with no clinical test available in Australia. Diagnosis of SCA36 was made within five days of receiving WGS data. In addition, REViewer is an important tools for visually confirming the composition of REs and can be used to identify interruptions in the motif and to eliminate false positive findings.

One recent study from the UKs 100 000 Genomes Project validated RE screening in WGS compared with PCR for neurological disease cohorts [4]. Compared with PCR, WGS was able to correctly classify expanded alleles with 97.3% sensitivity (95% CI 94.2–99.0) and 99.6% specificity (99.1–99.9) for thirteen pathogenic loci. Screening of WGS from 11 631 patients with suspected genetic neurological disorders with ExpansionHunter yielded 81 pathogenic REs. Follow up analysis with PCR confirmed that 68 were in the pathogenic range, representing an 84% true positive rate. Many of these diagnoses were made in people who did not present with typical symptoms, including children. This included REs for SCA2 in patients with early onset Parkinson's disease and ALS, SCA3 in a complex Parkinson's disease patient, and a SCA1 diagnosis in a hereditary spastic paraplegia patient. This study demonstrates that WGS, which is increasingly generated in both clinical and research settings, is a critical tool for the diagnosis of RE disorders, and is a rapid alternative to the long diagnostic odyssey associated with consecutive testing with PCR and Southern blots. In addition, it highlights the heterogeneity of RE disorder phenotypes as well as the issue of underdiagnosis. RE disorders, especially ataxias, are very rare, and evidence suggests that they are being misdiagnosed as more well-known neurological conditions. For example, a recent study of REs in ALS/FTD identified enrichment for multiple REs in the pathogenic range (SCA1, DM1 and DM2), and others in the intermediate range (including SCA2, SCA17 and Huntington's disease), highlighting the heterogeneity of the RE disorders and potential for misdiagnosis [15]. Furthermore, REs in *RFC1*, which cause CANVAS, have also been identified as a common cause of idiopathic neuropathy [16–18].

While there is demonstrated utility in detecting REs in WGS data, there are some challenges that still need to be addressed. Exonic REs are generally short and can be genotyped with high accuracy, however non-exonic expansions can be very large and the true RE size may be substantially underestimated by tools such as ExpansionHunter [4]. For some loci, such as *NOP56* (SCA36), RE are easily detected in WGS [14]. Some REs such as AAGGG and ACAGG in CANVAS are easily detected but difficult to accurately size [19,20]. However, these motifs are not common and the presence in a homozygous or compound heterozygous state would indicate CANVAS despite an underestimated allele size.

In contrast, *FGF14*-GAA (SCA27B), which has a pathogenic threshold of >300 repeats, is commonly observed with 50–250 repeats in the general population. However, ExpansionHunter is unable to accurately size RE larger than ~100 repeats at this locus and will severely underestimate true positive expansions [21]. SCA37 is uniquely difficult to detect with short-read sequencing techniques due to the combination of a pathogenic TTTCA embedded deep within the expanded reference TTTTA motif and cannot be detected with ExpansionHunter [22] (Figure 1). In addition, false positives are common for some loci, such as the RE that

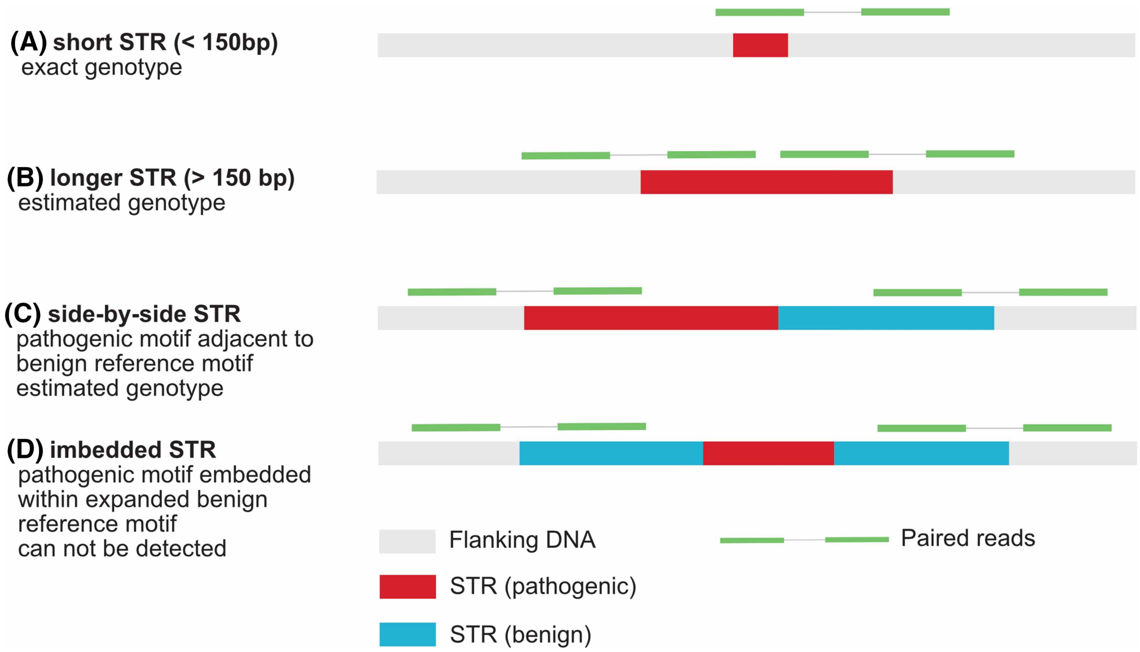

Figure 1. Genotyping of STRs in short-read sequencing data.

(**A**) An exact genotype can be determined for STRs that are shorter than the read length (typically 150 bp for modern short-read sequencing). (**B**) For STRs that are longer than the read length, genotypes are estimated; the longer the STR, the more likely that the size will be underestimated. (**C**) pathogenic insertions, such as a side-by-side insertion of a pathogenic motif adjacent to a benign motif (such as in FAME) are easily detected but can be difficult to accurately genotype. (**D**) in contrast, pathogenic motifs embedded within benign motifs, such as SCA37, can be difficult to detect with short-read sequencing data.

causes Fragile X Syndrome in *FMR1* [4]. Use of visualisation tools such as REViewer are essential for identifying false positives. GnomAD now has an STR catalogue for 19 241 genomes (v3.1), which includes 59 known pathogenic REs, but not newly discovered REs such as SCA27B or the putative *THAP11*-CAG. The gnomAD STR catalogue is a useful resource as it contains variations of motifs at specific, known RE loci and has REViewer plots available for all individuals, which can be helpful for researchers to use to compare to their own datasets and potentially mitigate risk of false positives.

Collectively, these data demonstrate that although some loci such as SCA27B and SCA37 require a RE-specific approach, most REs are easily detected with WGS in a homogenous/common approach. This approach now needs to be embedded in standard clinical genomics pipelines, with validation via current PCR or Southern based assays, or, in the future, most likely with long-read sequencing methods, especially for the larger REs. Such an approach will yield significant benefits for patients, their families and health care systems.

Recent discoveries of REs

In addition to short-read sequencing facilitating diagnosis of known RE disorders, the technology has also played a key role in the discovery of novel REs in recent years. Historically, the discovery of pathogenic REs was slow, as discovery relied heavily on linkage analysis and molecular methods such as expansion screening and DNA cloning (Figure 2, Table 1). The first pathogenic REs were discovered in 1991 [23,24]. The early discoveries were biased towards coding regions - however, over time there has been a boom in the discovery of non-coding pathogenic REs.

The first pathogenic RE discovered with WGS was the hexanucleotide GGGGCC RE in *C9ORF72* which causes ALS and frontotemporal dementia (FTD). This was discovered due to a well-powered linkage analysis and a highly significant GWAS hit at chromosome 9p21 for ALS [25] and FTD [26] which highlighted the genomic region requiring further examination. However, this discovery used manual inspection of the reads and extensive prior knowledge, and was not scalable.

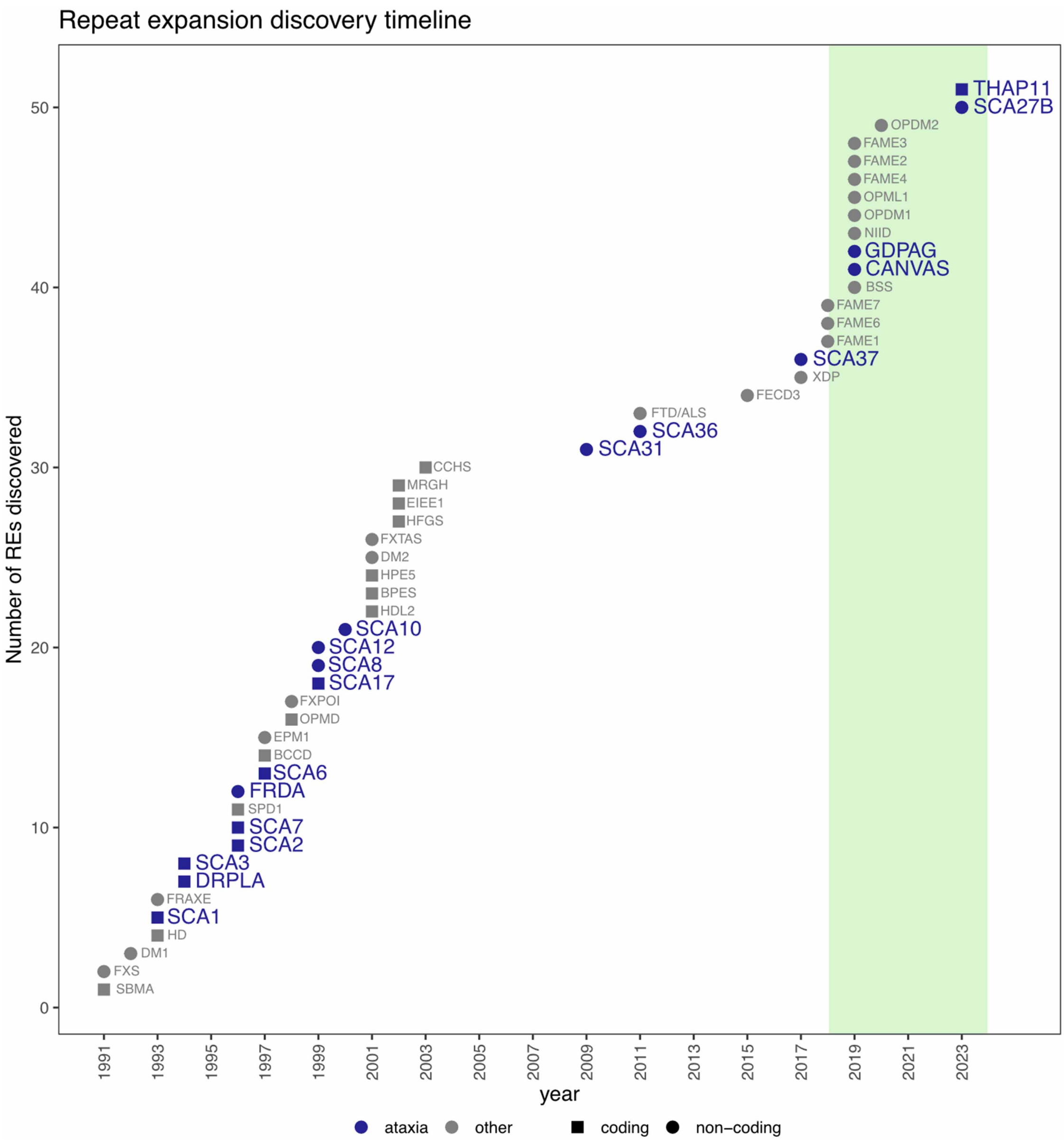

Figure 2. Timeline of the discovery of RE disorders.

Ataxias are shown in blue, non-ataxia disorders are shown in grey. Circle presents non-coding loci and squares represent coding loci. The green shading indicates the start of the rapid discovery of pathogenic REs facilitated with short-read sequencing.

In 2018, the discovery of the RE that causes FAME1 had major repercussions for the discovery of REs. The FAMEs are a group of epilepsies with very distinct phenotypes and tight linkage analysis which remained unsolved for decades, until the discovery of the TTTTA/TTTCA RE in *SAMD12* in FAME1 [27]. This RE was discovered by a number of methods which included WGS, and the identification of this pathogenic motif had a flow on effect and within only two years, FAME 2,3 and 4 were all solved using a mixture of WGS, long-read sequencing and traditional RE detection methods, often with prior information from mapping efforts [28–30]. All were caused by the TTTTA/TTTCA motif in different genes. Of note, the first TTTTA/TTTCA disorder reported was SCA37, which was discovered a year before FAME1. It is the only TTTTA/TTTCA motif to date known to cause ataxia rather than epilepsy. It is not clear why this specific motif causes FAME in some instances, and SCA37 in others, however it has been postulated that cell-specific gene expression may play a role. Ataxia genes generally have elevated expression in the cerebellum (Figure 3) when compared with randomly chosen set of genes of the same size from the human genome. Given the importance of Purkinje cell degeneration in the cerebellum in ataxia pathology [31], we postulate that elevated expression in these cells might contribute to the disease phenotype.

In 2019, another major breakthrough was published: the discovery of the recessively inherited AAGGG RE in *RFC1* [19,32]. This discovery was made simultaneously by two teams. The first paper published was by a UK team who relied on strong linkage analysis to manually screen the read data from WGS using a visual inspection approach. This was only practical due to a highly significant and narrow linkage region that could be manually screened [32]. The second paper, published by our team, used a novel tool called ExpansionHunter Denovo (EHDN) [33] and was the first time a RE was discovered using an unbiased method with a genome-wide approach [19]. Prior to the publication of EHDN, all RE detection tools were catalogue based, i.e. they could only screen for expansions of a pre-defined list of STRs (location and motif). EHDN is a catalogue-free method that leverages the lower read quality inherent to regions of repetitive DNA to rapidly identify 'anchored' and repeat-rich reads amongst aligned, unaligned and misaligned reads. These reads are anchored to a genomic location by their high quality read pair which aligns uniquely to the flanking DNA. Using this method, we discovered the AAGGG RE in CANVAS which is a non-reference motif that was not present in STR catalogues at the time and thus not detectable at that time with catalogue-based methods with the current catalogue.

Since 2018, 15 new pathogenic REs have been discovered, nine of which were published in 2019 and most of which relied on WGS for discovery. Three of these REs cause ataxia, including the recent discovery SCA27B, which is caused by a GAA RE in *FGF14* that was discovered with EHDN by two groups simultaneously [21,34]. Like CANVAS, SCA27B is a surprisingly common form of adult onset ataxia, which facilitated its discovery.

Recently, a study used long-read sequencing to identify a coding polyglutamine (polyQ) RE in the gene *THAP11*, in a five-generation Chinese family with autosomal dominant ataxia [35]. They report a repeat length of 45–55 repeats in adult-onset patients and one RE of 100 repeats in a childhood-onset patient, indicating anticipation can occur at this locus. This is the first discovery of a coding ataxia RE since SCA17 in 1999 [36].

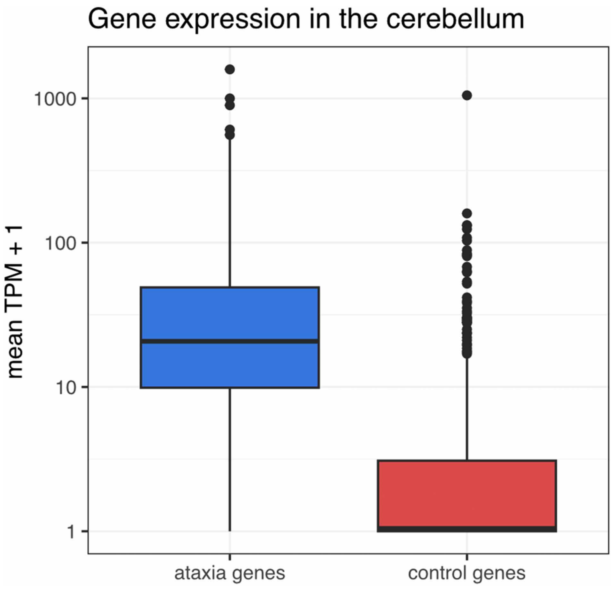

Figure 3. Ataxia genes are preferentially expressed in the cerebellum.
Cerebellar gene expression, presented as the mean transcripts per million plus one, was obtained from GTEx (v8 RNASeQCv1.1.9). Mean cerebellar expression levels of ataxia genes are shown compared with a control gene set as a box plot. An ataxia gene list consisting of 372 was curated based on genes from OMIM, PanelApp UK and PanelApp Australia (curated 2023-03-22). The control set comprises 372 randomly selected non-ataxia genes.

Discovering new REs

While short-read sequencing has been critical to the discovery of many REs in recent years, there are challenges that need to be addressed. One of the key difficulties for RE discovery is the lack of large STR genotype databases sourced from ancestry-diverse populations. GnomAD, which contains 76 156 genomes of diverse ancestries (v3.1.2 data set, GRCh38), has been pivotal for genomics of rare disease describing SNP and indel frequencies, but such extensive catalogues do not exist for STRs.

Some smaller scale databases exist, mainly based on analysis of the 1000 Genomes Project [37]. For example, the Illumina genome-wide polymorphic STR catalogue was generated using STR-finder, a tool that infers STRs from short-read sequencing at population levels [6]. This was generated using 2504 unrelated individuals from 1000 Genomes Project and contains 175 000 curated and polymorphic STRs, and is an important STR resource [37]. While the 1000 Genomes Project is a stratified sample that sought to maximised genetic diversity, it is too small to sufficiently capture the rare but crucial STR variation across populations that is needed for the discovery of rare pathogenic variants. More recently, a population reference panel of STR variation was generated using 3550 individuals from the 1000 Genomes Project and H3Africa cohorts, identifying over 1.7 million STR loci [38]. These catalogues are useful resources for STR discovery, especially for exonic STRs which are generally smaller, easier to genotype, and more evolutionarily conserved.

Despite these difficulties, there are strategies that can be implemented to improve the discovery of pathogenic REs in ataxia. First, while there are thousands of different motifs, only a small number have been associated with disease. For example, all known coding REs to date are CAG (polyQ), GCN (polyA) or CCG, and a first pass analysis might focus specifically on these motifs within coding regions. There are over 6000 known CGG sites, 93% of which are highly variable between individuals, and could potentially be candidates for pathogenic REs [39]. Likewise, motifs such as AAG, TTTCA, GGCCTG and others can be prioritised for intronic STRs.

Second, STRs in genes already known to cause ataxia should be prioritised — *CACNA1A*, *FGF14*, *RFC1* and *GLS* are all examples of genes that can cause ataxia by point mutations/indels, structural variants and also STRs [40–43]. Furthermore, expression in the cerebellum is known to be important for cerebellar ataxia — genes with no expression in the cerebellum can be excluded from analysis as a useful initial filter (Figure 3).

Non-catalogue methods such as EHDN are widely used for the discovery of novel REs. Other non-catalogue tools have recently been published. This includes STRling [44], which is the first tool to report the coordinates of novel STR expansions to base pair accuracy, and superSTR, which is a rapid non-reference-based method that identifies expanded motifs [45] and thus can be applied to RNAseq data and other organisms, without reference genomes. In addition, STRling can detect RE that are smaller than the read length, which is especially important for exonic RE which are often pathogenic under 150 bp. In contrast, EHDN cannot detect STRs shorter than the read length and is therefore biased against detecting exonic REs.

There is still utility in catalogue-based methods such as ExpansionHunter as a primary discovery tool. In contrast with intronic STRs which can be hundreds to thousands of repeats in length, exonic STRs are usually very short, and even small increases in length can be pathogenic. For example, in SCA2, alleles between 32–36 are incompletely penetrated, and alleles greater than or equal to 37 are fully penetrant [46,47]. However, 37 repeats span 111 base pairs, and this would not be detected with tools such as EHDN. Given that exonic STRs are generally highly conserved and well characterised, using a reference-based method such as ExpansionHunter, which can accurately size shorter REs, with a catalogue such as the Illumina polymorphic STR catalogue is a good approach to capturing exonic REs.

REs that resemble SCA37, in which the pathogenic motif is embedded deep within the reference motif, are difficult to discover with short-read sequencing, as reads from the STR cannot be uniquely mapped to the locus. However, the expansion of the reference motif can be detected with tools such as ExpansionHunter and EHDN, in which case such loci can be short listed for further investigative work, such as long-read sequencing and tools such as superSTR may identify the unmapped read enrichment for pathogenic repeats.

Interruptions

Further complexities arise in the form of interrupted motifs; however the impact of these interruptions is poorly understood. Motif interruptions refer to occasional interruptions in the motif sequence, and are distinct from motifs changes, i.e. the existing motif changes from one motif to a repetitive stretch of another motif. Generally, pathogenic REs are pure and the motif is not interrupted. Uninterrupted motifs are more unstable, and more susceptible to expansion during gametogenesis [48,49]. In addition, there are reports that interruptions change the disease phenotype. An example is SCA2 — in which pure CAG expansions greater than 33 repeats in *ATXN2* cause ataxia, but expansions interrupted by sporadic CAA motifs may cause Parkinson's Disease (PD) instead [50,51]. However, the research is conflicting and multiple studies do not find an association with PD and interrupted *ATXN2* motifs [52], possibly due to differences in familial or sporadic PD and ethnic diversity. In addition, severe neurodegeneration of the dopaminergic substantia nigra is often observed in SCA2 and in rare cases can cause parkinsonism, raising the possibility that people with SCA2 are being misdiagnosed with PD. The mechanism is unclear as both CAA and CAG code for glutamine and there are reports of interrupted CAG motifs in *ATXN2* with symptoms consistent with SCA2 [53]. RNA toxicity may be impacted, although this remains poorly understood [54,55].

Current RE detection methods do not detect interruptions, however motif purity can be checked using REViewer, as long as the interruption is not too deeply embedded within the STR where it may not be able to be identified. The accumulation of large genomics disease cohorts makes large-scale screening of motif purity increasingly accessible and may help address concerns regarding the impact of interruptions on disease progression.

Conclusion

Advances in methodologies for screening of REs in WGS data has resulted in significant progress in the discovery and diagnosis of these complex genetic variations in ataxia. However, these tools remain under-utilised in clinical diagnosis pipelines. Genomics is now widely used in the clinical setting, and patients will continue to miss out on rapid genetic diagnosis until such pipelines are implemented as a first-line screen for REs in ataxia. In addition, WGS is increasingly being generated for disease cohorts in a research setting, including ataxia, which will facilitate further discoveries of pathogenic REs. Although significant strides have been made using short-read sequencing, the nature of long REs means that WGS will always have its limitations. Long-read sequencing has the potential to address these limitations as the read length is sufficient to completely capture the RE within a single read. However, this is still an emerging technology and has limitations, including affordability, scalability and technical issues (reviewed in [56]), and will likely not displace short-read genomics in the short term. As costs continue to decrease and the technology continues to improve, short-read sequencing, with support from long-read sequencing, will continue to be critical for the discovery and diagnosis of pathogenic REs in coming years.

Summary

- Multiple studies have shown that short-read sequencing data can be used to report whether an individual has a pathogenic RE. This information can be used to suggest further clinical investigation and to improve diagnostic rates of RE disorders.
- In recent years, we have witnessed the success of utilising short-read sequencing data for the discovery of REs, which includes the discovery of pathogenic REs that cause CANVAS and SCA27B, two very common causes of late-onset ataxia.
- RE detection pipelines are ready for application in clinical pipelines where genomics is increasingly being used for genetic diagnosis.
- The rapid accumulation of both short and long-read genomes data in the research setting and continual improvement of RE detection methods suggests that further RE discovery are likely in the coming years.

Competing Interests

The authors declare that there are no competing interests associated with the manuscript.

Funding

H.R. was supported by an NHMRC Emerging Leadership 1 grant (1194364) and M.B. was supported by an NHMRC Leadership 1 grant (1195236). Additional funding was provided by the Independent Research Institute Infrastructure Support Scheme and the Victorian State Government Operational Infrastructure Program.

Open Access

Open access for this article was enabled by the participation of University of Melbourne in an all-inclusive *Read & Publish* agreement with Portland Press and the Biochemical Society under a transformative agreement with CAUL.

Author Contributions

H.R. planned, wrote and edited the manuscript. M.F.B reviewed and edited the manuscript. M.B planned, reviewed and edited the manuscript.

Acknowledgements

We acknowledge the work and contribution of Liam G Fearnley.

Abbreviations

ALS, amyotrophic lateral sclerosis; EHDN, ExpansionHunter Denovo; FTD, frontotemporal dementia; PD, Parkinson's disease; REs, repeat expansions; STRs, short tandem repeats.

References

1 Ruano, L., Melo, C., Silva, M.C. and Coutinho, P. (2014) The global epidemiology of hereditary ataxia and spastic paraplegia: a systematic review of prevalence studies. *Neuroepidemiology* **42**, 174–183 https://doi.org/10.1159/000358801

2 Rosenthal, L.S. (2022) Neurodegenerative cerebellar ataxia. *Continuum* **28**, 1409–1434 https://doi.org/10.1212/CON.0000000000001180

3 Krygier, M. and Mazurkiewicz-Bełdzińska, M. (2021) Milestones in genetics of cerebellar ataxias. *Neurogenetics* **22**, 225–234 https://doi.org/10.1007/s10048-021-00656-3

4 Ibañez, K., Polke, J., Hagelstrom, R.T., Dolzhenko, E., Pasko, D., Thomas, E.R.A. et al. (2022) 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.* **21**, 234–245 https://doi.org/10.1016/S1474-4422(21)00462-2

5 Austin-Tse, C.A., Jobanputra, V., Perry, D.L., Bick, D., Taft, R.J., Venner, E. et al. (2022) Best practices for the interpretation and reporting of clinical whole genome sequencing. *NPJ Genom. Med.* **7**, 27 https://doi.org/10.1038/s41525-022-00295-z

6 Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S. et al. (2019) Expansionhunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 https://doi.org/10.1093/bioinformatics/btz431

7 Mousavi, N., Shleizer-Burko, S., Yanicky, R. and Gymrek, M. (2019) Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 https://doi.org/10.1093/nar/gkz501

8 Dashnow, H., Lek, M., Phipson, B., Halman, A., Sadedin, S., Lonsdale, A. et al. (2018) STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* **19**, 121 https://doi.org/10.1186/s13059-018-1505-2

9 Tankard, R.M., Bennett, M.F., Degorski, P., Delatycki, M.B., Lockhart, P.J. and Bahlo, M. (2018) Detecting expansions of tandem repeats in cohorts sequenced with short-read sequencing data. *Am. J. Hum. Genet.* **103**, 858–873 https://doi.org/10.1016/j.ajhg.2018.10.015

10 Dolzhenko, E., Weisburd, B., Ibañez, K., Rajan-Babu, I.-S., Anyansi, C., Bennett, M.F. et al. (2022) REViewer: haplotype-resolved visualization of read alignments in and around tandem repeats. *Genome Med.* **14**, 84 https://doi.org/10.1186/s13073-022-01085-z

11 Rafehi, H., Green, C., Bozaoglu, K., Gillies, G., Delatycki, M.B., Lockhart, P.J. et al. (2023) Unexpected diagnosis of myotonic dystrophy type 2 repeat expansion by genome sequencing. *Eur. J. Hum. Genet.* **31**, 122–124 https://doi.org/10.1038/s41431-022-01166-y

12 Rajan-Babu, I.-S., Peng, J.J., Chiu, R., Study, I., Study, C., Li, C. et al. (2021) Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions. *Genome Med.* **13**, 126 https://doi.org/10.1186/s13073-021-00932-9

13 Eratne, D., Schneider, A., Lynch, E., Martyn, M., Velakoulis, D., Fahey, M. et al. (2021) The clinical utility of exome sequencing and extended bioinformatic analyses in adolescents and adults with a broad range of neurological phenotypes: an Australian perspective. *J. Neurol. Sci.* **420**, 117260 https://doi.org/10.1016/j.jns.2020.117260

14 Rafehi, H., Szmulewicz, D.J., Pope, K., Wallis, M., Christodoulou, J., White, S.M. et al. (2020) Rapid diagnosis of spinocerebellar ataxia 36 in a three-generation family using short-read whole-genome sequencing data. *Mov. Disord.* **35**, 1675–1679 https://doi.org/10.1002/mds.28105

15 Henden, L., Fearnley, L.G., Grima, N., McCann, E.P., Dobson-Stone, C., Fitzpatrick, L. et al. (2023) Short tandem repeat expansions in sporadic amyotrophic lateral sclerosis and frontotemporal dementia. *Sci. Adv.* **9**, eade2044 https://doi.org/10.1126/sciadv.ade2044

16 Currò, R., Salvalaggio, A., Tozza, S., Gemelli, C., Dominik, N., Galassi Deforie, V. et al. (2021) RFC1 expansions are a common cause of idiopathic sensory neuropathy. *Brain* **144**, 1542–1550 https://doi.org/10.1093/brain/awab072

17 Yuan, J.-H., Higuchi, Y., Ando, M., Matsuura, E., Hashiguchi, A., Yoshimura, A. et al. (2022) Multi-type RFC1 repeat expansions as the most common cause of hereditary sensory and autonomic neuropathy. *Front. Neurol.* **13**, 986504 https://doi.org/10.3389/fneur.2022.986504

18 Kumar, K.R., Cortese, A., Tomlinson, S.E., Efthymiou, S., Ellis, M., Zhu, D. et al. (2020) RFC1 expansions can mimic hereditary sensory neuropathy with cough and Sjögren syndrome. *Brain* **143**, e82 https://doi.org/10.1093/brain/awaa244

19 Rafehi, H., Szmulewicz, D.J., Bennett, M.F., Sobreira, N.L.M., Pope, K., Smith, K.R. et al. (2019) Bioinformatics-based identification of expanded repeats: a non-reference intronic pentamer expansion in RFC1 causes CANVAS. *Am. J. Hum. Genet.* **105**, 151–165 https://doi.org/10.1016/j.ajhg.2019.05.016
20 Scriba, C.K., Beecroft, S.J., Clayton, J.S., Cortese, A., Sullivan, R., Yau, W.Y. et al. (2020) A novel RFC1 repeat motif (ACAGG) in two Asia-Pacific CANVAS families. *Brain* **143**, 2904–2910 https://doi.org/10.1093/brain/awaa263
21 Rafehi, H., Read, J., Szmulewicz, D.J., Davies, K.C., Snell, P., Fearnley, L.G. et al. (2023) An intronic GAA repeat expansion in FGF14 causes the autosomal-dominant adult-onset ataxia SCA50/ATX-FGF14. *Am. J. Hum. Genet.* **110**, 105–119 https://doi.org/10.1016/j.ajhg.2022.11.015
22 Rosenbohm, A., Pott, H., Thomsen, M., Rafehi, H., Kaya, S., Szymczak, S. et al. (2022) Familial cerebellar ataxia and amyotrophic lateral sclerosis/frontotemporal dementia with DAB1 and C9ORF72 repeat expansions: an 18-year study. *Mov. Disord.* **37**, 2427–2439 https://doi.org/10.1002/mds.29221
23 Oberlé, I., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A. et al. (1991) Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science* **252**, 1097–1102 https://doi.org/10.1126/science.252.5009.1097
24 Verkerk, A.J., Pieretti, M., Sutcliffe, J.S., Fu, Y.H., Kuhl, D.P., Pizzuti, A. et al. (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 https://doi.org/10.1016/0092-8674(91)90397-h
25 Shatunov, A., Mok, K., Newhouse, S., Weale, M.E., Smith, B., Vance, C. et al. (2010) Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study. *Lancet Neurol.* **9**, 986–994 https://doi.org/10.1016/S1474-4422(10)70197-6
26 Van Deerlin, V.M., Sleiman, P.M.A., Martinez-Lage, M., Chen-Plotkin, A., Wang, L.-S., Graff-Radford, N.R. et al. (2010) Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. *Nat. Genet.* **42**, 234–239 https://doi.org/10.1038/ng.536
27 Ishiura, H., Doi, K., Mitsui, J., Yoshimura, J., Matsukawa, M.K., Fujiyama, A. et al. (2018) Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat. Genet.* **50**, 581–590 https://doi.org/10.1038/s41588-018-0067-2
28 Corbett, M.A., Kroes, T., Veneziano, L., Bennett, M.F., Florian, R., Schneider, A.L. et al. (2019) Intronic ATTTC repeat expansions in STARD7 in familial adult myoclonic epilepsy linked to chromosome 2. *Nat. Commun.* **10**, 4920 https://doi.org/10.1038/s41467-019-12671-y
29 Florian, R.T., Kraft, F., Leitão, E., Kaya, S., Klebe, S., Magnin, E. et al. (2019) Unstable TTTTA/TTTCA expansions in MARCH6 are associated with familial adult myoclonic epilepsy type 3. *Nat. Commun.* **10**, 4919 https://doi.org/10.1038/s41467-019-12763-9
30 Yeetong, P., Pongpanich, M., Srichomthong, C., Assawapitaksakul, A., Shotelersuk, V., Tantirukdham, N. et al. (2019) TTTCA repeat insertions in an intron of YEATS2 in benign adult familial myoclonic epilepsy type 4. *Brain* **142**, 3360–3366 https://doi.org/10.1093/brain/awz267
31 Huang, M. and Verbeek, D.S. (2019) Why do so many genetic insults lead to Purkinje cell degeneration and spinocerebellar ataxia? *Neurosci. Lett.* **688**, 49–57 https://doi.org/10.1016/j.neulet.2018.02.004
32 Cortese, A., Simone, R., Sullivan, R., Vandrovcova, J., Tariq, H., Yan, Y.W. et al. (2019) Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat. Genet.* **51**, 649–658 https://doi.org/10.1038/s41588-019-0372-4
33 Dolzhenko, E., Bennett, M.F., Richmond, P.A., Trost, B., Chen, S., van Vugt, J.J.F.A. et al. (2020) Expansionhunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol.* **21**, 102 https://doi.org/10.1186/s13059-020-02017-z
34 Pellerin, D., Danzi, M.C., Wilke, C., Renaud, M., Fazal, S., Dicaire, M.-J. et al. (2023) Deep intronic FGF14 GAA repeat expansion in late-onset cerebellar ataxia. *N. Engl. J. Med.* **388**, 128–141 https://doi.org/10.1056/NEJMoa2207406
35 Tan, D., Wei, C., Chen, Z., Huang, Y., Deng, J., Li, J. et al. (2023) CAG repeat expansion in THAP11 is associated with a novel spinocerebellar ataxia. *Mov. Disord.* **38**, 1281–1293 https://doi.org/10.1002/mds.29412
36 Koide, R., Kobayashi, S., Shimohata, T., Ikeuchi, T., Maruyama, M., Saito, M. et al. (1999) A neurological disease caused by an expanded CAG trinucleotide repeat in the TATA-binding protein gene: a new polyglutamine disease? *Hum. Mol. Genet.* **8**, 2047–2053 https://doi.org/10.1093/hmg/8.11.2047
37 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M. et al. (2015) A global reference for human genetic variation. *Nature* **526**, 68–74 https://doi.org/10.1038/nature15393
38 Jam, H.Z., Li, Y., DeVito, R., Mousavi, N., Ma, N., Lujumba, I. et al. (2023) A deep population reference panel of tandem repeat variation. *bioRxiv* https://doi.org/10.1101/2023.03.09.531600
39 Annear, D.J., Vandeweyer, G., Elinck, E., Sanchis-Juan, A., French, C.E., Raymond, L. et al. (2021) Abundancy of polymorphic CGG repeats in the human genome suggest a broad involvement in neurological disease. *Sci. Rep.* **11**, 2515 https://doi.org/10.1038/s41598-021-82050-5
40 Ronco, R., Perini, C., Currò, R., Dominik, N., Facchini, S., Gennari, A. et al. (2023) Truncating variants in RFC1 in cerebellar ataxia, neuropathy, and vestibular areflexia syndrome. *Neurology* **100**, e543–ee54 https://doi.org/10.1212/WNL.0000000000201486
41 Brusse, E., de Koning, I., Maat-Kievit, A., Oostra, B.A., Heutink, P. and van Swieten, J.C. (2006) Spinocerebellar ataxia associated with a mutation in the fibroblast growth factor 14 gene (SCA27): a new phenotype. *Mov. Disord.* **21**, 396–401 https://doi.org/10.1002/mds.20708
42 Indelicato, E. and Boesch, S. (2021) From genotype to phenotype: expanding the clinical spectrum of CACNA1A variants in the era of next generation sequencing. *Front. Neurol.* **12**, 639994 https://doi.org/10.3389/fneur.2021.639994
43 van Kuilenburg, A.B.P., Tarailo-Graovac, M., Richmond, P.A., Drögemöller, B.I., Pouladi, M.A., Leen, R. et al. (2019) Glutaminase deficiency caused by short tandem repeat expansion in GLS. *N. Engl. J. Med.* **380**, 1433–1441 https://doi.org/10.1056/NEJMoa1806627
44 Dashnow, H., Pedersen, B.S., Hiatt, L., Brown, J., Beecroft, S.J., Ravenscroft, G. et al. (2022) STRling: a k-mer counting approach that detects short tandem repeat expansions at known and novel loci. *Genome Biol.* **23**, 257 https://doi.org/10.1186/s13059-022-02826-4
45 Fearnley, L.G., Bennett, M.F. and Bahlo, M. (2022) Detection of repeat expansions in large next generation DNA and RNA sequencing data without alignment. *Sci. Rep.* **12**, 13124 https://doi.org/10.1038/s41598-022-17267-z
46 Riess, O., Laccone, F.A., Gispert, S., Schöls, L., Zühlke, C., Vieira-Saecker, A.M. et al. (1997) SCA2 trinucleotide expansion in German SCA patients. *Neurogenetics* **1**, 59–64 https://doi.org/10.1007/s100480050009
47 Pulst, S.M., Nechiporuk, A., Nechiporuk, T., Gispert, S., Chen, X.N., Lopes-Cendes, I. et al. (1996) Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nat. Genet.* **14**, 269–276 https://doi.org/10.1038/ng1196-269
48 Choudhry, S., Mukerji, M., Srivastava, A.K., Jain, S. and Brahmachari, S.K. (2001) CAG repeat instability at SCA2 locus: anchoring CAA interruptions and linked single nucleotide polymorphisms. *Hum. Mol. Genet.* **10**, 2437–2446 https://doi.org/10.1093/hmg/10.21.2437

49 Gao, R., Matsuura, T., Coolbaugh, M., Zühlke, C., Nakamura, K., Rasmussen, A. et al. (2008) Instability of expanded CAG/CAA repeats in spinocerebellar ataxia type 17. *Eur. J. Hum. Genet.* **16**, 215–222 https://doi.org/10.1038/sj.ejhg.5201954

50 Charles, P., Camuzat, A., Benammar, N., Sellal, F., Destée, A., Bonnet, A.M. et al. (2007) Are interrupted SCA2 CAG repeat expansions responsible for parkinsonism? *Neurology* **69**, 1970–1975 https://doi.org/10.1212/01.wnl.0000269323.21969.db

51 Gwinn-Hardy, K., Chen, J.Y., Liu, H.C., Liu, T.Y., Boss, M., Seltzer, W. et al. (2000) Spinocerebellar ataxia type 2 with parkinsonism in ethnic Chinese. *Neurology* **55**, 800–805 https://doi.org/10.1212/wnl.55.6.800

52 Ross, O.A., Rutherford, N.J., Baker, M., Soto-Ortolaza, A.I., Carrasquillo, M.M., DeJesus-Hernandez, M. et al. (2011) Ataxin-2 repeat-length variation and neurodegeneration. *Hum. Mol. Genet.* **20**, 3207–3212 https://doi.org/10.1093/hmg/ddr227

53 Costanzi-Porrini, S., Tessarolo, D., Abbruzzese, C., Liguori, M., Ashizawa, T. and Giacanelli, M. (2000) An interrupted 34-CAG repeat SCA-2 allele in patients with sporadic spinocerebellar ataxia. *Neurology* **54**, 491–493 https://doi.org/10.1212/wnl.54.2.491

54 Sobczak, K. and Krzyzosiak, W.J. (2005) CAG repeats containing CAA interruptions form branched hairpin structures in spinocerebellar ataxia type 2 transcripts. *J. Biol. Chem.* **280**, 3898–3910 https://doi.org/10.1074/jbc.M409984200

55 Li, P.P., Moulick, R., Feng, H., Sun, X., Arbez, N., Jin, J. et al. (2021) RNA toxicity and perturbation of rRNA processing in spinocerebellar ataxia type 2. *Mov. Disord.* **36**, 2519–2529 https://doi.org/10.1002/mds.28729

56 Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. and Gouil, Q. (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 https://doi.org/10.1186/s13059-020-1935-5

57 Orr, H.T., Chung, M.Y., Banfi, S., Kwiatkowski, Jr, T.J., Servadio, A., Beaudet, A.L. et al. (1993) Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat. Genet.* **4**, 221–226 https://doi.org/10.1038/ng0793-221

58 Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S. et al. (1994) CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat. Genet.* **8**, 221–228 https://doi.org/10.1038/ng1194-221

59 Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K. et al. (1994) Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nat. Genet.* **6**, 9–13 https://doi.org/10.1038/ng0194-9

60 Nagafuchi, S., Yanagisawa, H., Sato, K., Shirayama, T., Ohsaki, E., Bundo, M. et al. (1994) Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. *Nat. Genet.* **6**, 14–18 https://doi.org/10.1038/ng0194-14

61 Lindblad, K., Savontaus, M.L., Stevanin, G., Holmberg, M., Digre, K., Zander, C. et al. (1996) An expanded CAG repeat sequence in spinocerebellar ataxia type 7. *Genome Res.* **6**, 965–971 https://doi.org/10.1101/gr.6.10.965

62 Campuzano, V., Montermini, L., Moltò, M.D., Pianese, L., Cossée, M., Cavalcanti, F. et al. (1996) Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271**, 1423–1427 https://doi.org/10.1126/science.271.5254.1423

63 Zhuchenko, O., Bailey, J., Bonnen, P., Ashizawa, T., Stockton, D.W., Amos, C. et al. (1997) Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. *Nat. Genet.* **15**, 62–69 https://doi.org/10.1038/ng0197-62

64 Koob, M.D., Moseley, M.L., Schut, L.J., Benzow, K.A., Bird, T.D., Day, J.W. et al. (1999) An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nat. Genet.* **21**, 379–384 https://doi.org/10.1038/7710

65 Holmes, S.E., O'Hearn, E.E., McInnis, M.G., Gorelick-Feldman, D.A., Kleiderlein, J.J., Callahan, C. et al. (1999) Expansion of a novel CAG trinucleotide repeat in the 5′ region of PPP2R2B is associated with SCA12. *Nat. Genet.* **23**, 391–392 https://doi.org/10.1038/70493

66 Matsuura, T., Yamagata, T., Burgess, D.L., Rasmussen, A., Grewal, R.P., Watase, K. et al. (2000) Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nat. Genet.* **26**, 191–194 https://doi.org/10.1038/79911

67 Sato, N., Amino, T., Kobayashi, K., Asakawa, S., Ishiguro, T., Tsunemi, T. et al. (2009) Spinocerebellar ataxia type 31 is associated with "inserted" penta-nucleotide repeats containing (TGGAA)n. *Am. J. Hum. Genet.* **85**, 544–557 https://doi.org/10.1016/j.ajhg.2009.09.019

68 Kobayashi, H., Abe, K., Matsuura, T., Ikeda, Y., Hitomi, T., Akechi, Y. et al. (2011) Expansion of intronic GGCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. *Am. J. Hum. Genet.* **89**, 121–130 https://doi.org/10.1016/j.ajhg.2011.05.015

69 Seixas, A.I., Loureiro, J.R., Costa, C., Ordóñez-Ugalde, A., Marcelino, H., Oliveira, C.L. et al. (2017) A pentanucleotide ATTTC repeat insertion in the non-coding region of DAB1, mapping to SCA37, causes spinocerebellar ataxia. *Am. J. Hum. Genet.* **101**, 87–103 https://doi.org/10.1016/j.ajhg.2017.06.007

70 Depienne, C. and Mandel, J.-L. (2021) 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? *Am. J. Hum. Genet.* **108**, 764–785 https://doi.org/10.1016/j.ajhg.2021.03.011