University of Eastern Piedmont
"Amedeo Avogadro"
School of Medicine
Department of Health Sciences
Master's degree course in MEDICAL BIOTECHNOLOGIES

TITLE

Bioinformatic analysis and molecular validation of novel and known repeat expansions in neurodegenerative diseases using ExpansionHunter Denovo

Supervisor:

Prof.ssa Lucia Corrado

Co-supervisor:

Beatrice Piola

Candidate: Margherita Garrone

Matricula Number: 20038310

Academic year: 2025/2026

INDEX

1. SUMMARY

Repeat expansions are an important class of genetic variants involved in several neurological and neurodegenerative disorders, but their detection remains technically challenging, especially when expanded alleles are large, structurally complex or located outside predefined diagnostic panels. This study aimed to evaluate the performance and reliability of ExpansionHunter Denovo as a catalogue-free tool for detecting repeat expansion signals in patients affected by neurodegenerative diseases.

Whole-genome sequencing data from an Italian cohort of 140 patients with neurodegenerative disorders were analysed. ExpansionHunter Denovo outlier analysis was applied to a subset of 52 samples selected for homogeneous sequencing conditions. Since no matched control WGS dataset was available, an outlier-based approach was used instead of a case-control analysis. Candidate signals were prioritized according to z-score, repeat motif, genomic localization and feasibility of molecular validation. Selected loci were then investigated using conventional PCR, repeat-primed PCR, microsatellite fragment analysis and Sanger sequencing. ExpansionHunter Denovo results were also compared with a previous targeted ExpansionHunter analysis of 38 known repeat loci associated with neurodegenerative diseases. Three candidate loci outside the targeted catalogue, mapping to *CMKLR2-AS*, *PLXNA4* and *CRACR2A*, were selected for molecular validation. Molecular analyses did not support these loci as disease-associated repeat expansion candidates in the analysed patients, but showed that ExpansionHunter Denovo can identify repeat-variable regions requiring experimental confirmation and comparison with controls. The comparison with ExpansionHunter showed partial concordance at known repeat expansion loci. Some molecularly confirmed expansions were detected by ExpansionHunter Denovo, whereas others showed weak signals or were not detected, as observed for *HTT* and *ATXN1*. Validation of *TCF4*, *FGF14* and *RFC1* further showed that short-read bioinformatic predictions require orthogonal confirmation, particularly when repeat size, motif composition or allelic structure are complex. Overall, this study supports the use of ExpansionHunter Denovo as a complementary, discovery-oriented tool for identifying candidate repeat expansion signals. However, it should not be considered a first-line diagnostic test, but rather a second-level approach in genetically unresolved cases, after standard analyses and targeted evaluation of known repeat expansion loci. Future studies should integrate larger cohorts, appropriate control datasets and long-read sequencing to improve repeat sizing, motif characterization and interpretation of complex alleles.

2. INTRODUCTION

2.1 Tandem Repeats and Repeat Expansions

2.1.1 Repetitive DNA: classification and genomic organization

Repetitive elements (REs) are defined as DNA sequences that occur multiple times in the genome and constitute a substantial fraction of human genomic DNA. They include a wide variety of DNA elements with diverse structures and origins.[1,2] The study of repetitive DNA has historically been technically challenging because repeat-rich regions are difficult to amplify, map and assemble accurately, especially using PCR-based assays and short-read sequencing technologies.[1,3]

Despite this heterogeneity, REs are generally classified into two broad categories: tandem repeats (TRs) and interspersed repeats (IRs).[1,4]

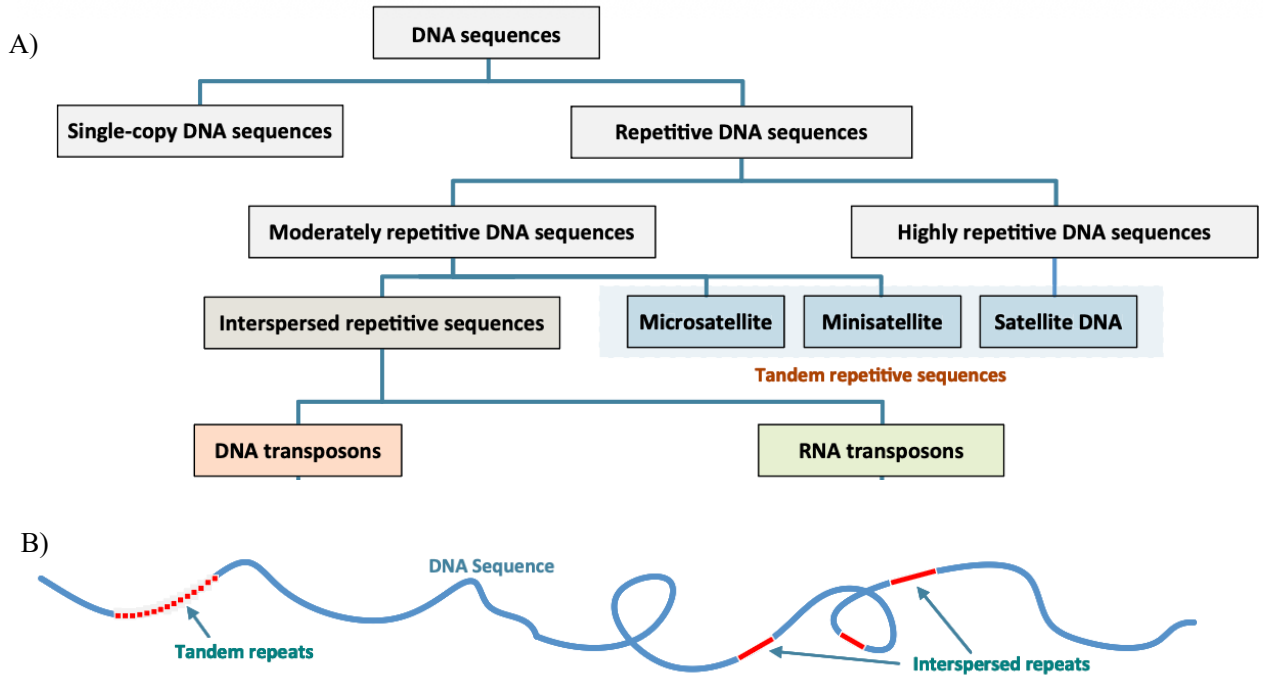

Figure 1. Classification of repetitive sequences: A) Repetitive DNA includes interspersed repeats (IRs), mainly represented by transposable elements, and tandem repeats (TRs), including microsatellites, minisatellites and satellite DNA. **B)** The lower panel illustrates the different organization of TRs, arranged consecutively at a specific locus, and IRs, dispersed across the genome. (Liao et al., 2023)

IRs are repetitive dispersed throughout the genome and are composed principally of transposable elements (TEs), such as short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs).[4–6]

Tandem repeats, in contrast, are repeated sequence units arranged consecutively in a head-to-tail orientation at a specific genomic locus.[1,5] This local organization distinguishes TRs from IRs and makes them highly variable in both repeat unit length and total array size.[2,5] TRs are

widespread throughout the genomes of many species, including humans, and are often highly conserved, suggesting that they may have important biological functions. These repetitive sequences can be found in both coding and non-coding regions.[7] According to the length of the repeated unit, tandem repeats are commonly divided into short tandem repeats (STRs), variable number of tandem repeats (VNTRs) and satellite repeats.[8]

Short tandem repeats, or microsatellites, are generally defined as tandem repeats with short motifs of 1–6 bp, even if variable motif have been identified in different disorders.[3,7,9,10]

Variable number of tandem repeats (VNTRs), or minisatellites, contain longer motifs, typically with repeat unit lengths of 10–99 bp.[3,8,10]

Lastly, satellite repeats have repeat units of at least 100 bp and are abundant in specialized chromosomal regions such as centromeres .[4,5,8]

Tandem repeats can be located in exons, introns, promoters, enhancers, untranslated regions and intergenic regions, indicating that they may influence different aspects of genome function.[5,7,11]

Among tandem repeats, STRs are highly polymorphic and represent one of the most variable types of sequence in the human genome.[3,12] Current estimates indicate that STRs occupy approximately 3–5% of the human genome, with more than one million loci identified to date.[3] In non-coding regions, STRs may influence gene expression and other regulatory processes.[2,5] Coding STRs are less frequent, but they can have direct consequences on protein sequence and structure when repeat-length variation changes the number of repeated amino acids.[2,7]

Overall, repetitive DNA is not a uniform genomic component, but a heterogeneous set of sequence classes with distinct organization, distribution and functional potential.[3,4,7]

Within this broad category, tandem repeats are especially relevant because their organization as repeated local arrays makes them both highly polymorphic under physiological conditions and susceptible to instability when repeat length increases.[5,7,10]

2.1.2 Functional roles and physiological variability of tandem repeats

Tandem repeats are dynamic genomic sequences that can influence biological processes at the DNA, RNA and protein levels, depending on their position and sequence composition.[4,7] Their potential functional relevance is closely related to their genomic distribution and organization, since tandem repeats can occur in coding sequences, introns, promoters, enhancers, untranslated regions and intergenic regions.[5,7]

The functional consequences of tandem repeats depend strongly on their genomic location, with promoter and enhancer repeats mainly affecting transcription, intronic repeats influencing

splicing and, in some cases, contributing to the formation of R-loops, UTR repeats affecting RNA metabolism and coding repeats modifying protein properties.[7,13] Repeats located in untranslated regions may also affect RNA transport, stability, translation and degradation.[7]

Tandem repeats can therefore affect a variety of biological processes, including evolution, development, brain function and behaviour.[14] The physiological relevance of tandem repeats is also linked to their high variability between individuals.[7,12]

Tandem repeat loci can display multiallelic variation because different individuals may carry different numbers of repeat units at the same locus.[7] This property enables tandem repeats to generate a broad spectrum of genetic variation.[7] Within physiological ranges, repeat-length variation is often compatible with normal function and may contribute to inter-individual differences in gene regulation, cellular traits and phenotype.[3,7]

The functional effects of tandem repeats can therefore be gradual, since changes in repeat number may produce proportional changes in gene expression, RNA processing or protein function.[7] For this reason, the same features that make tandem repeats important sources of normal variation can also predispose them to instability when repeat length increases beyond normal ranges.[5,7]

Thus, tandem repeats should not be viewed only as pathogenic elements, because their physiological variability may contribute to normal genome function, regulatory diversity and phenotypic variation.[3,7] Rather, they represent a dynamic class of genomic variation whose effects range from subtle modulation of gene expression and molecular phenotype to disease-associated consequences when instability becomes excessive.[4,5,7] Because tandem repeats are highly polymorphic and frequently multiallelic, they represent a class of variants that is particularly difficult to analyze using standard NGS pipelines.

2.1.3 From repeat instability to pathogenic expansion

A key characteristic of STRs is their intrinsic instability. This instability can lead either to repeat expansion, when the number of repeat units increases, or to repeat contraction, when repeat units are lost.[3,10]

Under physiological conditions, this mutability contributes to the high polymorphism of STR loci and represents an important source of inter-individual genetic variation.[7,12]

However, the same mutational tendency can become pathogenic when repeat length exceeds a locus-specific range compatible with normal gene or cellular function.[7,8]

Because of their repetitive structure, these loci are particularly prone to changes in repeat number during DNA replication and repair, with slipped-strand mispairing considered one of

the major mechanisms underlying STR mutation.[3,10] During slipped-strand mispairing, the template strand and the newly synthesized strand can misalign because of the repetitive nature of the sequence. This misalignment can generate looped-out repeat units, which may be incorporated or lost during subsequent DNA synthesis and repair. If the loop forms on the newly synthesized strand, the repeat tract may expand, whereas if it forms on the template strand, the repeat tract may contract.[10] Other mechanisms can also contribute to repeat length variation, including unequal crossing over, recombination-associated processes and DNA repair-dependent mechanisms.[3,5,10]

As a result, the high mutation rates of STRs make them an important source of genetic variation in human populations.[3,10,12] Within a physiological range, small differences in repeat length may be compatible with normal function and may contribute to the regulation of gene expression.[7,14] However, the same instability also predisposes these loci to further expansion. In general, longer alleles are less stable and more likely to undergo additional length changes during replication, increasing the probability that variation at these loci may become pathogenic.[15]

The biological consequences of tandem repeat variation depend on several factors, including repeat length, sequence composition, genomic localization and the secondary structures that the repeated sequence may form.[3,13] Repeat purity and the presence of interruptions within the repeat tract can also influence repeat stability, because uninterrupted alleles are generally more prone to further expansion than interrupted repeat tracts.[3,8,10] For example, some repeats can form hairpin structures that interfere with DNA synthesis, whereas GC-rich repeats may form G-quadruplexes that affect transcriptional regulation.[3,13] In several repeat expansion disorders, expanded alleles show intergenerational instability, meaning that repeat length may increase when transmitted from parent to offspring.[8,13] This intergenerational increase in repeat length can underlie genetic anticipation, in which disease tends to appear earlier or with greater severity in successive generations.[8,15]

When a repeat tract expands beyond a locus-specific threshold, it may acquire properties that disrupt normal cellular processes.[8,13] Depending on its genomic context, an expanded repeat may alter transcription, induce gene silencing, interfere with RNA metabolism, or result in the production of abnormal proteins with toxic effects.[8,13] Repeat instability can also occur somatically, generating different repeat lengths in different tissues or even in different cells of the same individual.[13,15] Somatic instability is particularly relevant in repeat expansion disorders because it can influence tissue specificity, disease severity and age at onset.[13,15]

Repeat expansions should therefore be regarded as a heterogeneous class of variants that can act at multiple levels, including DNA, RNA and protein.[13] Over the last decades, tandem repeat expansions have emerged as an important cause of human genetic disease, particularly in neurological disorders.[8,15] The transition from physiological variation to pathogenic expansion is not determined by a universal repeat number, but by thresholds that are specific to each locus and disease.[7,8,15]

Understanding the biological properties of tandem repeats and the transition from physiological variation to pathogenic expansion is therefore essential for the study of repeat expansion disorders and provides the basis for discussing their role in neurodegenerative disease.

2.2 Repeat Expansions in Neurodegenerative Diseases

Neurodegenerative diseases are a heterogeneous group of disorders characterized by progressive dysfunction and loss of selectively vulnerable neuronal populations, ultimately leading to deterioration of motor, cognitive, behavioural, and/or sensory functions.[16] They may be classified according to their predominant clinical features, anatomical distribution, or principal molecular abnormality, reflecting the marked biological and clinical heterogeneity of this disease group.[16]

Within this broad spectrum of disorders, repeat expansions have emerged as an important mutational class and are now recognized as the cause of a growing number of neurological diseases, many of which primarily affect the central nervous system.[8,15] Their relevance in neurodegeneration is reflected by the number of established repeat expansion disorders, and also by the diversity of associated phenotypes, pathogenic mechanisms, and genomic contexts in which these mutations occur.[8,15]

Although whole-genome sequencing is increasingly adopted in neurogenetics, standard variant-calling pipelines remain poorly suited for detecting repeat expansions. Consequently, a substantial fraction of repeat-mediated disorders may remain genetically unresolved despite apparently negative WGS analyse. This diagnostic gap has motivated the development of indicated computational approaches for repeat expansion detection.

2.2.1 Neurodegenerative disorders associated with repeat expansions

Repeat expansion disorders represent a clinically important group of genetic diseases caused by abnormal increases in the number of repeated DNA units at specific genomic loci.[8] Although repeat expansions can affect different tissues and organ systems, they are particularly relevant

to the nervous system, because many known repeat expansion disorders have neurological, neuromuscular or neurodegenerative manifestations.[8,15]

The first disease-associated repeat expansions were identified in the early 1990s, including CGG expansions in the *FMR1* gene causing fragile X syndrome and CAG expansions in the androgen receptor gene causing spinal and bulbar muscular atrophy.[8,13] Since these discoveries, approximately 60 repeat expansion disorders, involving more than 50 disease-associated genes, have been described, most of them affecting the nervous system.[8,13,15,17]

Neurodegenerative repeat expansion disorders can be broadly grouped according to the repeat motif, the genomic location of the expansion and the main pathogenic mechanism involved.[7,8,13] Coding CAG expansions cause polyglutamine disorders, including Huntington's disease and several spinocerebellar ataxias, whereas non-coding repeat expansions can affect promoters, untranslated regions or introns and may act through different mechanisms.[7,8,13] Examples include GAA expansions in *FXN* and *FGF14*; the GGGGCC expansion in *C9orf72*, and AAGGG repeat expansions in *RFC1*.[7,8,15,18]

Overall, repeat expansions are associated with a wide range of neurological phenotypes, including movement disorders, ataxias, motor neuron diseases, dementias, neuropathies and multisystemic syndromes.[7,8,15] This diversity shows that repeat expansions should not be interpreted as a single disease category, but should instead be considered in relation to the different neurological phenotypes they produce and the frequent overlap between clinical presentations.[8,19]

Disease category	Disease	Gene	Repeat motif	Normal range	Pathogenic repeat number	Location	Inheritance
Ataxia and movement disorders	Spinocerebellar ataxia 1	*ATXN1*	CAG	6–39	>39	Exon	AD
	Spinocerebellar ataxia 2	*ATXN2*	CAG	15–29	>34	Exon	AD
	Spinocerebellar ataxia 3	*ATXN3*	CAG	13–36	55–84	Exon	AD
	Spinocerebellar ataxia 4	*ZFHX3*	GGC	<21–30	>48	Exon	AD
	Spinocerebellar ataxia 6	*CACNA1A*	CAG	4–16	21–30	Exon	AD
	Spinocerebellar ataxia 7	*ATXN7*	CAG	3–35	34 to >300	Exon	AD
	Spinocerebellar ataxia 8	*ATXN8/ATXN8OS*	CAG	6–37	~107–250	3'UTR	AD
	Spinocerebellar ataxia 10	*ATXN10*	ATTCT	10–29	280–4500	Intron	AD
	Spinocerebellar ataxia 12	*PPP2R2B*	CAG	<66	>66	5'UTR	AD
	Spinocerebellar ataxia 17	*TBP*	CAG	25–44	45–66	Exon	AD
	Spinocerebellar ataxia 27B/LOCA	*FGF14*	GAA	6/8–249	>300	Intron	AD
	Spinocerebellar ataxia 31	*BEAN1*	TGGAA	NA (as insertion)	>500	Intron	AD
	Spinocerebellar ataxia 36	*NOP56*	GGCCTG	3–8	1500–2500	intron	AD
	Spinocerebellar ataxia 37	*DAB1*	ATTTC	NA (as insertion)	Insertion	5'UTR	AD
	Spinocerebellar ataxia 51	*THAP11*	CAG	19–39	45–100	Exon	AD
	Dentatorubral-pallidoluysian atrophy	*ATN1*	CAG	7–25	>49	Exon	AD
	Cerebellar ataxia, neuropathy and vestibular areflexia syndrome	*RFC1*	AAGGG	AAAAG:11–200, AAAGG 40–1000	AAGGG 400- >2000 Different motifs	Intron	AR
	Friedreich's ataxia	*FXN*	GAA	7–22	>66	Intron	AR
	Huntington's disease	*HTT*	CAG	6–34	36–180	Exon	AD
	Huntington disease-like 2	*JPH3*	CAG/CTG	<50	>50	Exon	AD
	Kennedy's disease (spinal and bulbar muscular atrophy)	*AR*	CAG	11–24	40–62	Exon	X-linked
	Neuronal intranuclear inclusion disease	*NOTCH2NLC*	GGC	<55	>55	5'UTR	AD
ALS/FTD and dementia	C9orf72-associated ALS/FTD	*C9orf2*	GGGGCC	<30	>500	Intron	AD
	Spinocerebellar ataxia 2	*ATXN2*	CAG	15–29	Intermediate expansion	Exon	AD
	Fragile X-associated tremor/ataxia syndrome	*FMR1*	CGG	6–52	Premutation	5'UTR	X-linked
	Neuronal intranuclear inclusion disease	*NOTCH2NLC*	GGC	<55	>55	5'UTR	AD
	Mental retardation, FRA12A type	*DIP2B*	CGG	6-23	≥ 272 repeats	5'UTR	AD
Neuromuscolar/systemic diseases	Myotonic dystrophy type 1	*DMPK*	CTG	5–37	>50 to ~2000	3'UTR	AD
	Myotonic dystrophy type 2	*CNBP/ZNF9*	CCTG	<27	>75 to ~11000	Intron	AD
	Oculopharyngeal muscular dystrophy	*PABPN1*	GCG	<10	>12–17	Exon	AD
	Oculopharyngeal myopathy with leukoencephalopathy	*NUTM2B-AS1*	CGG	3–16	40–60	5'UTR	AD
	Oculopharyngodistal myopathy 1	*LRP12*	CGG	13–45	90–130	5'UTR	AD
	Oculopharyngodistal myopathy 2	*GIPC1*	CGG	12–32	97–120	5'UTR	AD
	Fragile X-associated tremor/ataxia syndrome	*FMR1*	CGG	6–52	55–200	5'UTR	X-linked
	Fragile X syndrome	*FMR1*	CGG	6–52	>200	5'UTR	X-linked
	Fragile-XE syndrome	*FMR2/AFF2*	CCG	6–25	>200	5'UTR	X-linked
	Fuchs endothelia corneal distrophy 3	*TCF4*	CTG	10-20	40-50	intronic	AD

Table 1. Main repeat expansions associated with neurological and neurodegenerative disorders: the table summarizes representative disorders according to the affected gene, repeat motif, normal and pathogenic repeat ranges, genomic region and inheritance patterns. For mode of inheritance, AD, autosomal dominant and AR, autosomal recessive. ALS, amyotrophic lateral sclerosis; FTD, frontotemporal dementia.

2.2.2 Phenotypic heterogeneity and clinical pleiotropy

One of the major features of repeat expansion disorders is their marked clinical heterogeneity. The same type of genetic variant, namely an expanded tandem repeat, can give rise to very different neurological phenotypes depending on the affected gene, repeat motif, repeat size, genomic location and disease mechanism.[8,13] This heterogeneity is particularly relevant in neurodegenerative diseases, where clinical syndromes often overlap and may include combinations of motor, cognitive, behavioural, cerebellar, extrapyramidal, peripheral nerve and autonomic manifestations.[15,16]

The same expanded locus may be associated with different phenotypes, whereas clinically similar syndromes may arise from expansions at distinct loci.[8,15,20] This is particularly evident in *C9orf72*-related disease, in which the same GGGGCC expansion may present as amyotrophic lateral sclerosis (ALS), frontotemporal dementia (FTD), or broader overlapping phenotypes.[7,15]

Another example is *FMR1*, where different expansion ranges at the same CGG repeat locus are associated with distinct clinical entities. Full mutation alleles are associated with fragile X syndrome, whereas premutation alleles are associated with fragile X-associated tremor/ataxia syndrome and fragile X-associated primary ovarian insufficiency.[7,20] Pleiotropy is also evident among non-coding CGG repeat expansion disorders. *NOTCH2NLC* expansions were initially linked to neuronal intranuclear inclusion disease, but they have also been associated with broader neurological presentations.[15,20]

Conversely, similar clinical syndromes may be caused by repeat expansions in different genes. This is evident in hereditary ataxias, where clinically overlapping cerebellar presentations can result from coding CAG expansions, non-coding repeat expansions, or recently identified intronic expansions such as those in *RFC1* and *FGF14*.[15,19] Intermediate or premutation-range expansions can further blur the boundaries between classical diagnostic categories. For example, intermediate CAG expansions in *ATXN2* are associated with increased risk of amyotrophic lateral sclerosis, whereas larger expansions in the same gene cause spinocerebellar ataxia type 2.[15,18]

Recent cohort studies support the idea that repeat expansions can contribute to phenotypic variability across broader neurodegenerative disease categories than previously expected, including amyotrophic lateral sclerosis and dementia cohorts.[18,21]

These examples show that phenotypic heterogeneity is not only a diagnostic challenge, but also a clue to the biological complexity of repeat expansion disorders.[8,15] The variable involvement of different neuronal systems suggests that expanded repeats may affect cellular function

through distinct molecular routes.[15] Therefore, understanding why repeat expansions produce such diverse clinical presentations requires examining the pathogenic mechanisms acting at the DNA, RNA and protein levels.[8,13]

2.2.3 Pathogenic mechanisms of repeat expansion disorders

Repeat expansion disorders are mechanistically heterogeneous because expanded repeats can disrupt cellular function at different molecular levels. Their pathogenic effects depend on several factors, including repeat motif, repeat length, genomic localization, transcriptional activity of the locus and the biological function of the affected gene.[8,13]

From a mechanistic perspective, the pathogenic mechanisms of repeat expansion disorders can be broadly classified into four interrelated categories.[13]

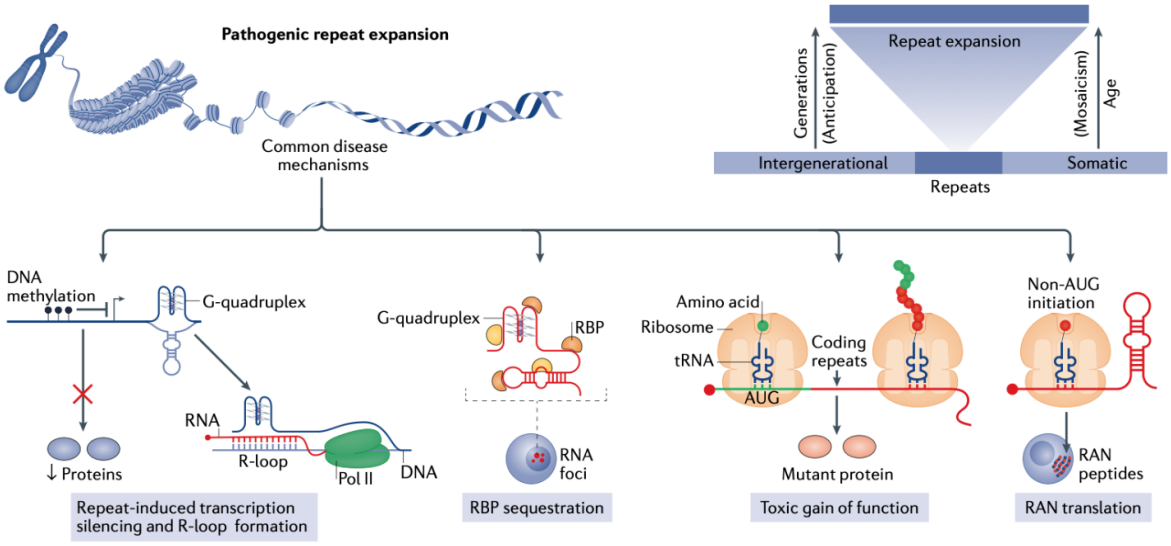

Figure 2: Main pathogenic mechanisms of repeat expansion disorders: Expanded repeats can disrupt cellular function through repeat-induced transcriptional silencing and R-loop formation, RNA-binding protein sequestration, toxic gain of function of repeat-containing proteins and repeat-associated non-AUG translation. (Malik et al., 2021)

A first category involves loss of function, in which the repeat expansion reduces or abolishes expression of the affected gene through transcriptional repression, epigenetic silencing, or defective transcription.[6,13] The classical example is fragile X syndrome, in which the CGG expansion in the 5′ untranslated region of *FMR1* leads to hypermethylation and transcriptional silencing.[6,13] A related loss-of-function mechanism has also been implicated in Friedreich ataxia, in which the GAA expansion reduces frataxin expression through transcriptional repression and heterochromatin formation. A loss-of-function component has also been proposed in *C9orf72*-associated ALS/FTD, where expanded repeats may reduce normal gene expression and thereby contribute to cellular dysfunction.[6,13,15]

A second major mechanism is RNA-mediated toxicity, in which expanded repeats are transcribed and the repeat-containing RNA acquires toxic properties.[6,13] Expanded RNAs can form abnormal secondary structures, accumulate as nuclear RNA foci and sequester RNA-binding proteins, thereby perturbing alternative splicing, RNA transport, translation, and other aspects of RNA metabolism.[6,13] Myotonic dystrophy type 1 and type 2 are well-established examples of RNA toxicity, caused respectively, by expanded CUG- and CCUG-containing transcripts that accumulates in nuclear RNA-foci and promote repeat-mediated disruption of alternative splicing.[7,8,13] In *C9orf72*-associated-ALS/FTD, bidirectional transcription of the expanded repeats generates both sense and antisense repeat-containing RNAs, which can form nuclear RNA foci and contribute to RNA-mediated toxicity.[6,13] Related mechanisms, including RNA-foci formation and alteration of RNA processing, have also been described in selected non-coding spinocerebellar ataxias; for example, in SCA8 bidirectional transcription further increases the molecular complexity of the expanded locus.[8,13] RNA-mediated toxicity is also relevant in fragile X-associated tremor/ataxia syndrome, where premutation-range CGG expansions in *FMR1* are transcribed and contribute to neurodegeneration through toxic RNA gain-of-function mechanisms.[7,13,20]

A third category involves toxic gain of function of canonically translated repeat-containing proteins. This mechanism is particularly important in polyglutamine disorders, in which CAG expansions within coding regions are translated into abnormally elongated polyglutamine tracts.[7,8] These expanded protein domains can promote protein misfolding, aggregation, abnormal protein-protein interactions and disruption of cellular homeostasis.[8,13] Huntington's disease and several spinocerebellar ataxias are representative examples of this mechanism, as they are caused by coding CAG expansions that produce expanded polyglutamine-containing proteins.[7,8]

A fourth mechanism is repeat-associated non-AUG translation. RAN translation allows expanded repeat-containing RNAs to be translated in the absence of a canonical AUG start codon, generating repetitive peptides that may be toxic to cells.[6,13] This mechanism has been described in several repeat expansion disorders, including *C9orf72*-related ALS/FTD, fragile X-associated tremor/ataxia syndrome, myotonic dystrophy and some spinocerebellar ataxias.[6,13,15] In *C9orf72*-associated disease, RAN translation of the GGGGCC expansion produces dipeptide repeat proteins that can aggregate and interfere with multiple cellular pathways, including nucleocytoplasmic transport, RNA metabolism and proteostasis.[13]

These mechanisms are not mutually exclusive, and in many repeat expansion disorders multiple pathogenic processes may coexist and interact.[13]

Expanded repeats may also affect cellular function through DNA- and chromatin-level effects. Some repeat sequences can form non-canonical nucleic acid structures, such as hairpins, G-quadruplexes, triplex DNA and R-loops, which may interfere with replication, transcription and DNA repair.[3,13] These structures can contribute to genome instability and may interact with RNA- and protein-mediated mechanisms, reinforcing the idea that repeat expansion disorders often result from combined molecular insults rather than from a single isolated pathway.[6,13]

Overall, repeat expansion disorders are best understood as multilayered diseases in which pathogenicity may arise from loss of normal gene function, toxic repeat-containing RNAs, abnormal proteins, RAN-translated peptides and repeat-induced effects on genome stability.[6,13] This mechanistic diversity helps explain why repeat expansions can produce such broad clinical variability and provides the basis for understanding their quantitative and diagnostic complexity.[8,13,15]

2.2.4 Repeat length, allelic variability and diagnostic implications

A distinctive feature of repeat expansion disorders is that pathogenicity is often quantitative rather than simply determined by the presence or absence of a variant. Repeat length has important quantitative effects on penetrance, age at onset, and disease severity, and somatic instability may further modify repeat size across tissues over time.[13,15]

The transition from normal variation to disease is therefore usually based on locus-specific repeat ranges rather than on a universal threshold. These ranges are not absolute, because some expanded alleles can be observed in unaffected individuals, and alleles below the pathogenic range may correspond to intermediate or premutation categories.[15] Normal, intermediate, premutation, reduced-penetrance and fully pathogenic alleles may coexist at the same locus, creating a continuum of biological and clinical effects rather than a strictly binary classification.[8,15]

In several repeat expansion disorders, longer repeat alleles are associated with earlier onset or more severe disease. This relationship is well established in classical disorders such as Huntington's disease and myotonic dystrophy type 1, although repeat length alone does not fully explain clinical variability, which can also be influenced by genetic background, environmental factors and tissue-specific instability.[7,15]

Intermediate and premutation-range alleles are particularly relevant in neurodegenerative disease because they may not always cause the classical phenotype but can increase susceptibility to related or overlapping disorders. An important example is represented by *ATXN2*, in which intermediate CAG expansions increase ALS risk, whereas larger

expansions cause spinocerebellar ataxia type 2. This illustrates how repeat size can influence not only disease severity, but also the clinical category associated with a given locus.[15,18]

The stability and pathogenic potential of an allele may depend not only on the total number of repeat units, but also on repeat purity, motif sequence and the presence of interruptions within the repeat tract.[3,8,15] Interruptions may stabilize some repeat alleles or modify disease risk, making the structure of the repeat tract relevant for genotype–phenotype correlations.[8,15]

Repeat length may vary between tissues or even between cells within the same individual, and the allele size measured in blood may not always fully reflect the repeat burden in disease-relevant tissues such as brain, spinal cord or muscle.[13,15] This tissue-specific variability may contribute to differences in age at onset, disease severity and selective vulnerability of neuronal populations.[13,15]

Together, these findings suggest that repeat expansions should be considered not only in classical phenotypes, but also in unresolved, atypical, or overlapping neurodegenerative presentations.[18,19,21]

Repeat expansions now occupy an important position in the molecular genetics of neurodegeneration.[3,8]

As methods for repeat detection continue to improve, tandem repeat variation may explain part of the unresolved genetic burden across ALS, FTD, ataxias, neurodegenerative dementias, and related disorders.[3,18,21] However, despite their clinical relevance, repeat expansions remain technically challenging to identify, particularly when they are large, structurally complex, or poorly resolved by standard short-read sequencing approaches.[3,15,19] For this reason, the systematic investigation of repeat expansions is essential both for molecular diagnosis and for the identification of novel disease-associated loci in unresolved neurodegenerative cohorts.[3,19]

2.3 Methods for Detecting Repeat Expansions

The detection of repeat expansions is a crucial step in the molecular diagnosis of many neurological and neuromuscular disorders. However, their analysis remains technically challenging because expanded alleles may exceed the size range of conventional assays or short-read sequencing, and repetitive or GC-rich regions can complicate amplification, mapping and accurate repeat sizing.[3,15,22]

Historically, known repeat expansions have been investigated using locus-specific methods, including Southern blotting, fluorescent fragment analysis, repeat-primed PCR and long-range PCR; although these approaches remain useful for diagnostic confirmation, they are generally

low-throughput and provide limited information on sequence context and repeat architecture.[22] The increasing use of short-read sequencing has enabled the analysis of repeat expansions from genome-wide data, but short reads often fail to span large expanded alleles and may be affected by mapping ambiguity, GC bias, stutter artefacts and allele dropout.[3,15] These limitations have promoted the development of dedicated bioinformatic tools and long-read sequencing approaches, which can improve the detection, sizing and characterization of repeat expansions, including motif composition, interruptions, methylation status and somatic mosaicism.[15,23]
Overall, the field has shifted from single-locus diagnostic assays toward more integrated strategies, in which conventional methods, short-read screening and long-read characterization provide complementary information for detecting and interpreting repeat expansions.

2.3.1 Conventional and sequencing-based approaches

Repeat expansion detection can be broadly divided into two complementary strategies: conventional locus-specific assays and sequencing-based approaches. Conventional methods, including fluorescent fragment analysis, repeat-primed PCR, long-range PCR and Southern blotting, are designed to test known or clinically suspected repeat loci, whereas sequencing-based approaches use whole-exome sequencing (WES), whole-genome sequencing (WGS) or long-read sequencing data to broaden the analysis beyond a single preselected target.[22,24,25] This distinction is particularly relevant in neurogenetic disorders, where overlapping phenotypes, variable age at onset and incomplete penetrance may make it difficult to select one specific locus for targeted testing.[15,24]
Among conventional approaches, PCR-based fragment analysis relies on primers flanking the repeat region and uses fluorescently labelled PCR products that are separated by capillary electrophoresis, allowing estimation of fragment size and indirect calculation of repeat number.[22] This method is mainly suitable for relatively small repeat expansions, particularly when the expected pathogenic allele remains within the amplifiable size range, generally up to approximately 700–800 bp, whereas long-range PCR can extend this range up to approximately 10–15 kb.[22] However, PCR-based assays may become less reliable for very large expansions, GC-rich repeats, or alleles that amplify inefficiently, because preferential amplification of the shorter allele can lead to underestimation or missed detection of expanded alleles.[22]
Repeat-primed PCR (RP-PCR), also known as triplet-primed PCR, differs from standard flanking PCR because it uses three primers: a flanking primer, a repeat-specific primer containing a universal tail, and a universal primer.[22] After electrophoretic analysis, the presence of an expansion is indicated by a characteristic sawtooth pattern, making RP-PCR useful when

fragment analysis detects only one allele. However, RP-PCR cannot provide accurate sizing of large expansions and can only detect repeat motifs targeted by the primer design, limiting its ability to characterize alternative motifs or complex repeat structures.[22] Long-range PCR can also be used to amplify expanded repeat loci compatible with PCR amplification, using two specific primers flanking the region of interest. It is useful for selected loci and can be combined with downstream sequencing to obtain more detailed information on repeat length and motif composition. However, like other PCR-based methods, it is affected by amplification bias toward shorter alleles and is poorly suited to very large or highly GC-rich expansions.[22]

Southern blotting has historically been used for large repeat expansions that are difficult to amplify by PCR and is regarded as a gold-standard method for detecting large polynucleotide repeat expansions.[14] It involves enzymatic digestion of genomic DNA, electrophoretic separation, membrane transfer and hybridization with a locus-specific probe.[20] Although it can provide information on the approximate size of very large expansions, Southern blotting is labour-intensive, time-consuming, requires relatively large amounts of high-quality DNA, and does not provide information on the internal sequence composition of the repeat.[15,22]

Sequencing-based approaches have expanded the possibilities for repeat expansion analysis by allowing multiple loci to be interrogated from the same dataset.[24,25] WES and WGS are now widely used in research and diagnostic workflows, but standard diagnostic pipelines are primarily optimized for single-nucleotide variants and small insertions or deletions rather than repeat expansions.[24,25] As a result, repeat expansions may remain undetected unless sequencing data are analysed with specific strategies that account for repeat-containing reads, abnormal coverage patterns or reads that do not map correctly to the reference genome.[24,25]

Short-read sequencing can therefore be used as a screening approach for repeat expansions across multiple loci.[24,25] This also enables the reanalysis of existing NGS datasets for repeat expansions at known loci or other candidate regions.[25] However, short-read approaches usually provide indirect evidence of an expansion and often require orthogonal confirmation, particularly when the result is clinically relevant or when accurate sizing and sequence characterization are needed.[24,25]

In parallel, long-read sequencing offers a complementary strategy because it can span larger repeat alleles and provide more direct information on repeat size and sequence structure.[3,15] These technologies are particularly useful for characterizing complex expansions, including repeat interruptions and motif composition, which are often difficult to resolve with conventional assays or short-read sequencing.[15,23] Nevertheless, their routine use remains

limited by cost, DNA quality requirements, sequencing throughput and the complexity of data interpretation.[3,22]

Conventional and sequencing-based approaches should therefore be viewed as complementary rather than mutually exclusive. Conventional methods remain useful when a specific locus is strongly suspected or when candidate expansions require validation, whereas short-read sequencing enables broader screening and long-read sequencing supports more detailed characterization of large or structurally complex expanded alleles.[15,23]

2.3.2 Genome-wide screening and bioinformatic tools: ExpansionHunter and ExpansionHunter Denovo

The increasing availability of whole-genome sequencing (WGS) data has created new opportunities for the systematic detection of short tandem repeat expansions. However, repeat expansions are not efficiently captured by standard variant-calling pipelines, which are mainly designed to detect single nucleotide variants, small insertions/deletions and structural variants. For this reason, dedicated bioinformatic tools are required to analyse repetitive regions and infer repeat length from sequencing reads.[3]

Several computational methods have been developed to detect STR variation from short-read sequencing data, including STRetch, exSTRa, ExpansionHunter,[25,26] ExpansionHunter Denovo and STRling.[3] These tools differ in their input requirements and analytical strategy: some rely on predefined STR catalogues, whereas others can screen the genome more broadly for candidate expansions. They generally infer repeat size by integrating different types of short-read evidence, including spanning, flanking and in-repeat reads.[3,27]

ExpansionHunter is a tool for targeted genotyping of short tandem repeats and flanking variants. It performs a target search within BAM/CRAM file to identify reads that span, flank, and are fully contained in each repeat.[28] It is a sequence graph-based method that estimates repeat size from short-read WGS data using a predefined catalogue of repeat loci. For each locus, the program extracts relevant reads from a binary alignment/map file and realigns them using this graph-based model representing the locus structure. The realigned reads are then used to genotype each STR variant according to the locus specified in the variant catalogue.[26] This approach is particularly useful when the genomic position and repeat motif are already known, as in many established repeat expansion disorders. ExpansionHunter provides repeat length estimates and confidence intervals, and it has been applied in studies of neurological and neurodegenerative diseases to screen loci such as *HTT*, *FMR1*, *C9orf72* and several spinocerebellar ataxia genes.[21]

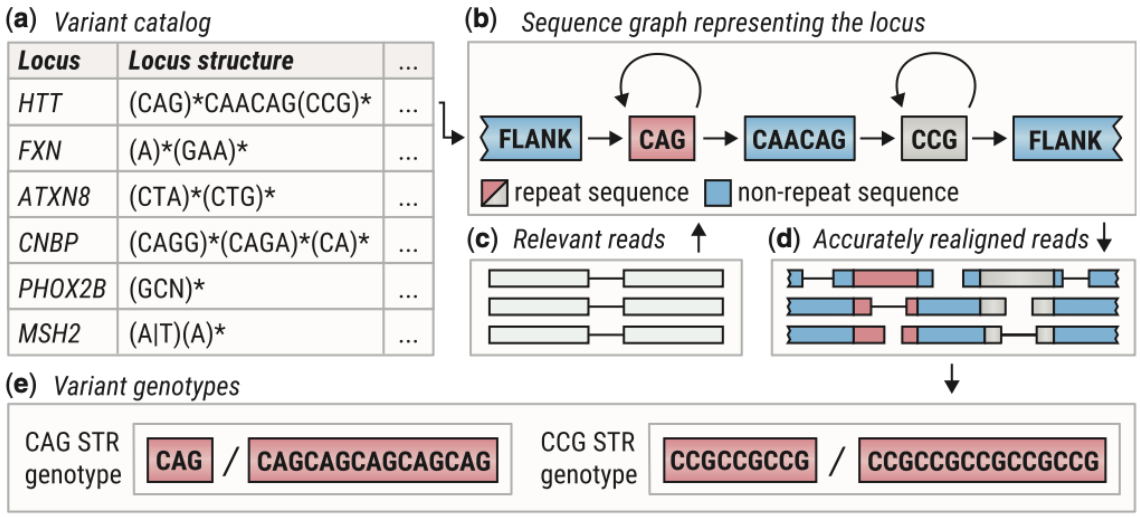

Figure 3. ExpansionHunter workflow for STR genotyping: ExpansionHunter reads locus definitions from a predefined variant catalogue and constructs a sequence graph representing each repeat locus. Relevant reads are extracted from the input alignment file, realigned to the graph, and used to infer the genotype of each STR variant. (Dolzhenko et al., 2019)

A key limitation of catalogue-based STR genotyping tools, including ExpansionHunter, is that they depend on prior knowledge of the loci to be analysed. Catalogue-based approaches are intrinsically limited to previously annotated loci and therefore cannot identify pathogenic expansions occurring at uncharacterized genomic regions or involving unexpected repeat motifs.To address this limitation, ExpansionHunter Denovo was developed as a catalogue-free method for genome-wide repeat expansion detection.[3,27]

ExpansionHunter Denovo (EHdn) is a suite of tools for detecting novel expansions of short tandem repeats. EHdn is intended for analysis of a collection of BAM/CRAM files.[29] It searches short-read sequencing alignments for in-repeat reads (IRRs), which are reads contained completely within repetitive sequence.[27] These signals can then be compared across samples to identify candidate expansions either through case-control analysis or through outlier detection. A case-control analysis is appropriate when a significant subset of cases is expected to contain expansions of the same repeat. For example, if a case-control analysis is applied to a dataset consisting of ALS patients and healthy controls, it is expected to flag the repeat expansion in the *C9orf72* gene as significant. If cases include samples from patients with diverse phenotypes, it might be appropriate to assume that there is no enrichment for any specific expansion and hence the case-control analysis is not appropriate. In this situation, an outlier analysis can be used to flag repeats that are expanded in a small proportion of cases compared to the rest of the dataset.[29]

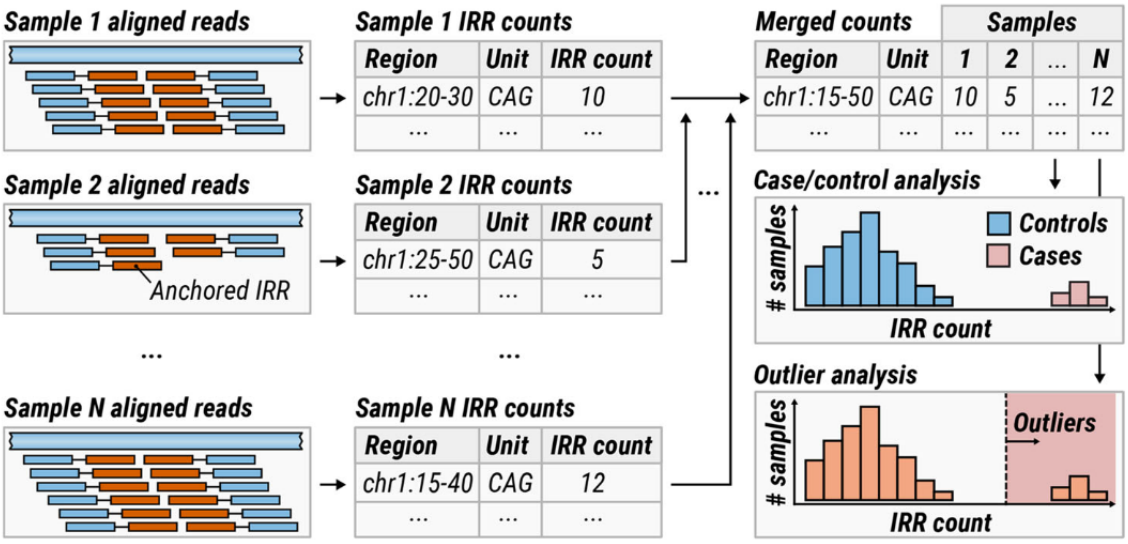

Figure 4. ExpansionHunter Denovo workflow for repeat expansion detection: EHdn identifies anchored in-repeat reads across aligned reads and summarizes them into STR profiles for each sample. The resulting profiles are merged across samples; if cases and controls can be defined, IRR counts are compared between groups, whereas outlier analysis is used when no such partition is possible. (Dolzhenko et al., 2020)

These two tools represent different but complementary strategies for analysing STR expansions from short-read WGS data. ExpansionHunter is a targeted genotyping tool: it requires a predefined variant catalogue and is therefore suited to the analysis of the annotated loci, but it is not designed as a discovery tool for unknown expansions. ExpansionHunter Denovo, in contrast, is a catalogue-free approach designed for BAM/CRAM short-read of 100-200 bp. It detects genome-wide expansion signals by identifying in-repeat reads and localizing candidate loci, but it does not provide clinical reliable sizing of expanded alleles. Therefore, EHdn results should be interpreted as screening signals requiring prioritization, comparison with appropriate controls and orthogonal validation using targeted molecular assays or long-read sequencing when clinically relevant.[3,27]

Overall, genome-wide STR screening has shifted repeat expansion analysis from repeated single-locus testing toward the parallel assessment of multiple repeat loci within the same sequencing dataset. This approach is particularly relevant for neurological disorders, in which different STR expansions can be associated with overlapping or heterogeneous clinical phenotypes.[3,15] However, despite these advantages, short-read-based screening does not fully replace confirmatory or higher-resolution approaches, because large or complex expansions may be difficult to size and characterize accurately with short reads. Therefore, clinically relevant findings may require validation or further characterization using targeted molecular assays, such as repeat-primed PCR or Southern blotting, or long-read sequencing.[3,15]

2.3.3 Technical challenges in repeat expansion detection

Despite the development of dedicated bioinformatic tools, repeat expansion detection remains technically challenging. This is mainly because STR loci are repetitive, many pathogenic alleles exceed the length of short sequencing reads, and some repeat regions have complex structures. As a result, standard short-read sequencing may have difficulties in mapping reads to repetitive regions and in accurately estimating expanded repeat length.[3,15]

A major limitation is read length. Short-read sequencing platforms usually generate reads of approximately 100–150 bp, whereas many pathogenic repeat expansions extend for hundreds or thousands of base pairs. When an expansion is longer than the read or fragment length, the allele cannot be directly spanned, making precise repeat sizing difficult.[15] This limitation becomes particularly relevant because many pathogenic repeat alleles exceed the length of short reads, and the accuracy of short-read STR genotyping decreases when repeat alleles exceed the fragment length.[3]

Read alignment represents another important source of difficulty. Repetitive reads may remain unmapped, map ambiguously, or align incorrectly to genomic regions with similar repeat motifs. This problem is particularly relevant when the repeat lies in a low-complexity region or when the flanking sequence is itself repetitive.[3] Since many computational approaches depend on read alignment, differences in both the reference genome and the alignment strategy may affect the detection of repeat expansions.[30]

Additional challenges arise from repeat composition. Highly repetitive or GC-rich regions, including CGG and GGGGCC expansions, can be difficult to amplify and sequence, making it challenging to obtain sufficient coverage in some STR regions. PCR amplification during library preparation may also introduce stutter errors, which complicate accurate repeat sizing. Moreover, standard short-read NGS does not directly detect epigenetic modifications, such as DNA methylation, which can be diagnostically relevant in some repeat expansion disorders.[15]

Repeat length alone may not be sufficient for interpretation, because pathogenic expansions can also differ in motif structure, sequence composition and repeat interruptions. These features can influence repeat stability and clinical expression, but complex repeat architecture is difficult to resolve with short-read data.[3,15] Somatic mosaicism may also affect age at onset, disease severity and progression. Recent long-read methods, such as TRGT, can help characterize repeat allele sequences, methylation levels and mosaicism from PacBio HiFi sequencing data.[23]

Long-read sequencing can overcome several limitations of short-read approaches because individual reads may span large repeat expansions and their flanking regions, allowing more direct characterization of repeat size, sequence composition and, in some cases, methylation

status.[3,15] However, long-read sequencing is not yet a complete replacement for other methods in all settings, because its implementation may still be limited by cost, lower throughput, platform-specific error profiles and the need for dedicated bioinformatic analysis.[3]

For these reasons, repeat expansion detection currently benefits from an integrated strategy, because conventional molecular assays, short-read WGS, dedicated bioinformatic tools and long-read sequencing address different aspects of the same technical problem.[3,15] Short-read WGS and computational tools are useful for broad screening and for reanalysing existing sequencing datasets, particularly because most WGS data generated so far are short-read and can be reused for STR genotyping.[3] Targeted molecular assays, such as repeat-primed PCR and Southern blotting, remain important for confirming clinically relevant findings or assessing loci that cannot be resolved confidently by short-read data.[3] Long-read sequencing provides higher-resolution characterization when repeat size, motif structure, methylation or mosaicism cannot be adequately resolved by short-read methods.[3,23] Thus, these technical challenges explain why repeat expansions remain difficult to detect and why their analysis often benefits from the combined use of computational screening, targeted molecular confirmation and long-read characterization.[3,15]

3. AIM OF THE STUDY

This study aimed to evaluate the performance and the reliability of ExpansionHunter Denovo outlier analysis for the detection of repeat expansion signals in patients affected by neurodegenerative disorders. In particular, the study focused on two main aspects: the ability of ExpansionHunter Denovo to detect candidate repeat expansion signals at novel loci, and the concordance and complementarity between ExpansionHunter Denovo and the targeted catalogue-based tool ExpansionHunter at known disease-associated repeat loci.

By integrating bioinformatic comparison with molecular validation of selected loci, this study aimed to assess whether ExpansionHunter Denovo outlier signals corresponded to detectable repeat-length variation and whether candidate loci could represent disease-associated repeat expansions or polymorphic repeat regions.

4. MATERIALS AND METHODS

4.1 Bioinformatic identification and prioritization of candidate repeat expansions

4.1.1 Study cohort and sequencing data

Whole-genome sequencing data were available from a previously characterized Italian cohort of 140 patients affected by neurodegenerative disorders. Patients were selected for whole-genome sequencing on the basis of familial cases, positive family history for other neurodegenerative disorders, or sporadic cases with comorbidity for different neurodegenerative disorders, such as amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD).

The original cohort included patients affected by ALS, ALS/FTD, FTD, spinocerebellar ataxia, Parkinson's disease (PD) and other neurodegenerative conditions. The ALS/FTD group included 69 patients who had previously tested negative for pathogenic variants in *SOD1*, *FUS*, *TARDBP,* and for the hexanucleotide mutation of *C9orf72*. The spinocerebellar ataxia cohort was composed of 40 cases negative for gene mutations or expansions in the main genes associated with ataxia (*ATXN1, ATXN2, ATXN3, etc*.). The PD cohort included 29 cases, negative for variants in the main Parkinson's disease-associated genes. Additionally, two patients with other neurodegenerative conditions (spastic paraplegia, leukodystrophy) were included.

DISEASE	CHARACTERISTICS OF PATIENTS
ALS, ALS/FTD, FTD	23 FALS
	31 SALS showing cognitive impairment or dementia
	15 SALS with positive family history for other neurodegenerative disorders
SCA	40 cases
PD	17 PD with an early onset
	12 PD with positive family history for Parkinson's Disease
LEUKODYSTROPHY	1 case
HEREDITARY SPASTIC PARAPLEGIA	1 case

Table 2. Subjects included in the study: the table summarizes the main diagnosis groups included in the original cohort of 140 patients with neurodegenerative disorders. **FALS**: familial amyotrophic lateral sclerosis; **SALS**: sporadic amyotrophic lateral sclerosis; **SCA**: spinocerebellar ataxia; **PD**: Parkinson's disease.

The study was approved by the Ethical Committee of the institution, and written informed consent was obtained from all participants.

In the present analysis, ExpansionHunter Denovo was applied to a subset of 52 samples from this cohort, numbered from #95 to #146. These samples were selected because they had been generated using the same sequencing workflow and represented the largest group with homogeneous sequencing conditions. This catalogue-free analysis was used to screen for candidate repeat expansion signals across the genome and to identify additional candidate loci not necessarily included in the previous targeted ExpansionHunter catalogue. Among the candidate loci prioritized from the ExpansionHunter Denovo output, three genes were selected for further molecular investigation.

ExpansionHunter Denovo results were also compared with data from a previous ExpansionHunter analysis performed on 76 patients from the same cohort of 140 NDD patients, including the samples analysed here. This previous analysis was based on a custom ExpansionHunter catalogue prepared for the targeted analysis of 38 repeat loci with established associations with neurodegenerative diseases (Table 3).

#CHROM	START	END	REF	MOTIF	*GENE*
chr2	191745598	191745646	16	GCA	*GLS*
chr3	63898360	63898390	10	GCA	*ATXN7*
chr3	63898390	63898402	4	GCC	*ATXN7*
chr3	128891419	128891499	20	CAGG	*CNBP*
chr3	128891499	128891539	10	CAGA	*CNBP_CAGA*
chr3	128891539	128891575	18	CA	*CNBP_CA*
chr4	3076603	3076660	19	CAG	*HTT*
chr4	3076666	3076693	9	CCG	*HTT*
chr4	39350044	39350099	11	AARRG	*RFC1*
chr4	41747989	41748049	20	GCN	*PHOX2B*
chr5	146258290	146258320	10	GCT	*PPP2R2B*
chr6	16327864	16327954	30	TGC	*ATXN1*
chr6	170870994	170871105	37	CGA	*TBP*
chr9	27573526	27573544	3	GGCCCC	*C9ORF72*
chr9	71652177	71652202	25	A	*FXN*
chr9	71652202	71652220	6	GAA	*FXN*
chr11	119076999	119077032	11	CGG	*CBL*
chr12	7045879	7045936	19	CAG	*ATN1*
chr12	50898784	50898805	7	GGC	*DIP2B*
chr12	112036753	112036822	23	GCT	*ATXN2*
chr13	70713485	70713515	10	CTA	*ATXN8OS_CTA*
chr13	70713515	70713560	15	CTG	*ATXN8OS*
chr13	102813925	102814074	49	GAA	*FGF14*
chr14	23790681	23790699	6	GCG	*PABPN1*
chr14	92537353	92537386	11	GCT	*ATXN3*
chr15	23086366	23086390	8	CGC	*NIPA1*
chr16	87637893	87637935	14	CTG	*JPH3*
chr18	53253386	53253458	24	CAG	*TCF4*
chr19	13318672	13318711	13	CTG	*CACNA1A*
chr19	14606853	14606886	11	CCG	*GIPC1*
chr19	46273462	46273522	20	CAG	*DMPK*
chr20	2633379	2633403	4	GGCCTG	*NOP56*
chr20	2633403	2633421	3	CGCCTG	*NOP56*
chr21	45196324	45196360	3	CGCGGGGCGGGG	*CSTB*
chr22	46191234	46191304	14	ATTCT	*ATXN10*
chrX	66765158	66765227	23	GCA	*AR*
chrX	146993568	146993628	20	CGG	*FMR1*
chrX	147582151	147582211	20	GCC	*AFF2*

Table 3. Repeat loci included in the targeted ExpansionHunter catalogue: the table lists the 38 loci selected for the targeted analysis, all with previously established associations with NDDs. Genomic coordinates refer to the hg38 human reference genome.

4.1.2 Outlier analysis with ExpansionHunter Denovo

Genome-wide screening for candidate repeat expansions was performed using ExpansionHunter Denovo v0.9.0. Aligned sequencing data were used as input files in BAM format, and all analyses were performed using the hg38 human reference genome.

ExpansionHunter Denovo was used as a catalogue-free screening approach to identify repeat expansion signals across the genome. The `profile` module was run for each sample using a command-line script that allowed batch processing of the sequencing files. This step generated an STR profile in JSON format for each analysed sample. During profile generation, the parameters `--min-anchor-mapq 50` and `--max-irr-mapq 40` were applied.

The individual STR profiles were subsequently merged into a single multi-sample file, as required by the tool to compare repeat expansion signals across the cohort using outlier analysis. This approach was used to identify sample-specific repeat expansion signals by comparing each sample with the rest of the cohort.

Candidate loci prioritization

Candidate loci were subsequently prioritized according to the ExpansionHunter Denovo signal, including the z-score, repeat motif, genomic localization and possible biological relevance of the associated gene. Higher z-scores were considered indicative of stronger sample-specific repeat expansion signals compared with the rest of the cohort.

For the newly identified candidate loci not included in the previous targeted ExpansionHunter catalogue, a z-score threshold ≥ 5 was additionally applied. This filtering strategy led to the prioritization of three genes for molecular validation.

4.1.3 Comparison between ExpansionHunter Denovo and ExpansionHunter

Some candidate loci identified by ExpansionHunter Denovo were compared with results obtained from the previous ExpansionHunter analysis at loci represented in the targeted catalogue to assess whether catalogue-free outlier detection and targeted repeat genotyping provided concordant or complementary evidence.

For each overlapping or clinically relevant locus, ExpansionHunter Denovo results were interpreted together with the corresponding ExpansionHunter output. Specifically, the comparison considered the ExpansionHunter Denovo outlier signal and z-score, the repeat motif and the genomic position of the locus, and the ExpansionHunter repeat-size estimates or predicted expansion status.

This comparison was used to evaluate differences between the two approaches in terms of candidate prioritization and to distinguish concordant, partially concordant or discordant bioinformatic signals.

Selected candidate loci were then further investigated using targeted molecular approaches, including conventional PCR followed by microsatellite fragment analysis, repeat-primed PCR and, when required, Sanger sequencing for repeat motif confirmation or to estimate repeat units in selected alleles.

4.2 Primer design

Primers for molecular validation were designed according to the genomic coordinates of the selected candidate loci, using Primer 3 Input (https://primer3.ut.ee). NCBI Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) was used to assess primer specificity and potential off-target amplification, OligoCalc (https://oligocalc.eu) was used to evaluate the propensity of the selected primers to form possibly secondary-structures, including hairpin formation and self-dimerization, and UCSC BLAT alignment (https://genome.ucsc.edu/cgi-bin/hgBlat) was used to confirm primer alignment and genomic localization against the hg38 reference genome.

For flanking PCR assays, primer pairs were designed to flank the repeat region, allowing amplification of the repeat-containing allele, as shown in Figure 5A. A universal tail, shown in red in Table 4, was added to the 5′ end of each forward primer. This sequence is complementary to the 5′-6-FAM-labelled P5 universal primer, enabling fluorescent detection of the amplified fragments by capillary electrophoresis.

For repeat-primed PCR, primer design was adapted to the structure of the target repeat. The assay included a locus-specific forward primer, a reverse primer annealing within the repeated sequence and carrying a 5′-M13 universal tail, showed in green in Table 5, and a 6-FAM-labelled M13 anchored primer complementary to this tail (Figure 5B). The fluorescently labelled anchored enable the detection of the amplification product by capillary electrophoresis. This design allowed the detection of expansion-compatible amplification patterns when full amplification of the repeat-containing allele was not achievable by conventional PCR.

Primers used for conventional PCR are reported in Table 4, whereas primers used for repeat-primed PCR are reported in Table 5.

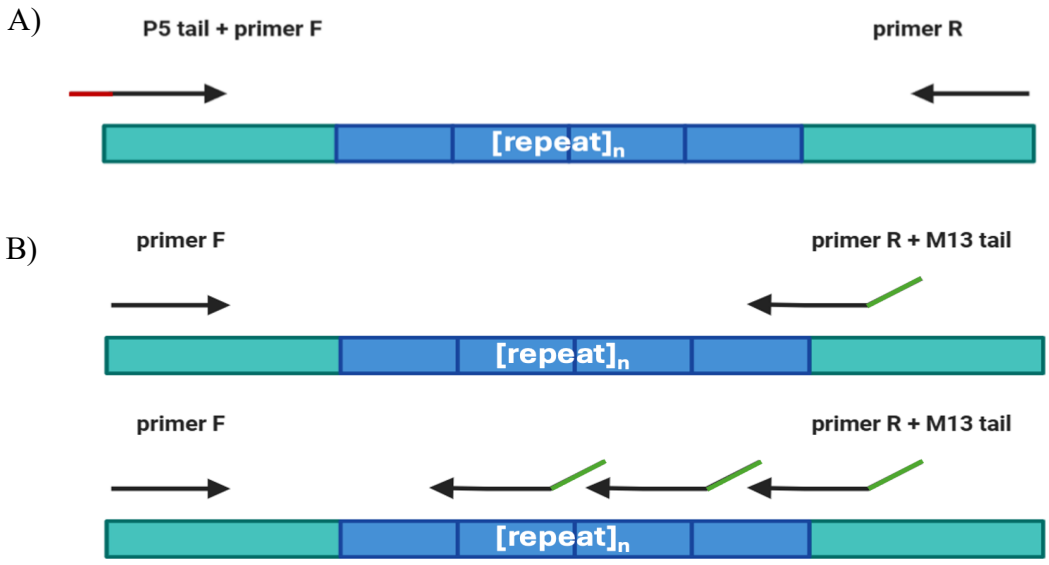

Figure 5. Schematic representation of primer design for conventional flanking PCR and for repeat-primed PCR: A) Conventional flanking PCR using a forward primer carrying the 5' P5 universal tail and a reverse primer. **B)** Repeat-primed PCR using a locus-specific forward primer, a repeat-specific reverse primer carrying a 5' M13 universal tail. Red and green segments indicate the P5 and M13, respectively.

GENE	FORWARD PRIMER	REVERSE PRIMER	Ta
TCF4	5'-*ACATACGCATCCCAGTTTGAGACGGGCTAC GTTTCCTGGCAAA*-3'	5'-*GATGAGTTTGGTGTAAGATGC*-3'	58°C
FGF14	5'-*ACATACGCATCCCAGTTTGAGACGTGCAAA TGAAGGAAAACTCTT*-3'	5'-*CAATGATGAATTAAGCAGTTCC*-3'	58°C
RFC1	5'-*ACATACGCATCCCAGTTTGAGACGTCAAGT GATACTCCAGCTACACCGTTGC*-3'	5'-*GTGGGAGACAGGCCAATCACTTCAG*-3'	64°C
CMKLR2-AS	5'-*ACATACGCATCCCAGTTTGAGACGCAGGTG AGTCTCAAACGCTT*-3'	5'-*TAGCAACTCGAGGATGAGGC*-3'	58°C
PLXNA4	5'-*ACATACGCATCCCAGTTTGAGACGATAGTG GCAATGACCCCAGG*-3'	5'-*CGGCTATGTGGTATGGCATT*-3'	59°C
CRACR2A	5'-*ACATACGCATCCCAGTTTGAGACGCCAGGA GGTGTAGGTAGGTT*-3'	5'-*GAGACACTGGCCAGAGAGAC*-3'	62°C

Table 4. Primers used for conventional PCR: The table reports the forward and the reverse primers used to amplify selected candidate loci, together with the corresponding annealing temperature (Ta).

GENE	FORWARD PRIMER	REVERSE PRIMER	ANCHORED PRIMER- M13 UNIVERSAL TAIL
FGF14	5'-*TGCCCACATAGAGCTTAGTCT*-3'	5'-*CACGACGTTGTAAAACGAC*GAAGAAGAAGAAGAAGAAGAA-3'	5'-/6-FAM/*CACGACGTTGTAAAACGAC*-3'

Table 5. Primers used for repeat-primed PCR: the table reports the forward, reverse and anchored primers used for repeat-primed PCR at the selected *FGF14* locus.

4.3 PCR

Conventional PCR was performed to amplify the selected loci using the primer pairs reported in Table 4. PCR reactions were prepared according to a general amplification protocol, with locus-specific adjustments depending on primer characteristics, expected amplicon size and sequence composition. Different reaction mixes and thermal cycling protocols were therefore optimized for each locus. The PCR reaction protocols and corresponding thermal protocol are reported in Tables 6-14.

PCR products were checked by agarose gel electrophoresis to verify successful amplification and assess the presence and intensity of the expected bands. Gels were visualized using the ChemiDoc imaging system (BioRad).

Reagent	**Initial concentration**	**Final concentration**	**Volume for the mix**
GoTaq® G2 Flexi Buffer (Promega)	5X	1X	4 µL
$MgCl_2$ (Promega)	25 mM	1.5 mM	1.2 µL
dNTPs	2.5 mM	0.2 mM	1.6 µL
Primer Forward	10 pM/µL	0.375 pM/µL	0.75 µL
Primer Reverse	10 pM/µL	0.5 pM/µL	1 µL
Fam-P5 primer	10 pM/µL	0.5 pM/µL	1 µL
GoTaq® Polymerase (Promega)	5 U/µL	0.024 U/µL	0.1 µL
H_2O			8.35 µL
Final Volume			18 µL
+2 µL DNA [25 ng/µL] tot: 50 ng			

Table 6. PCR protocols used for *CMKLR2-AS, TCF4, PLXNA4* and *CRACR2A*: the table reports the reagents, initial concentration, final concentration and final volumes used for amplification of the selected loci.

Reagent	Initial concentration	Final concentration	Volume for the mix
GoTaq® G2 Flexi Buffer (Promega)	5X	1X	4 µL
$MgCl_2$ (Promega)	25 mM	1.125 mM	0.9 µL
dNTPs	2.5 mM	0.15 mM	1.2 µL
Primer Forward	10 pM/µL	0.2 pM/µL	0.4 µL
Primer Reverse	10 pM/µL	0.36 pM/µL	0.72 µL
Fam-P5 primer	10 pM/µL	0.36 pM/µL	0.72 µL
GoTaq® Polymerase (Promega)	5 U/µL	0.018 U/µL	0.072 µL
H_2O			10 µL
Final Volume			18 µL
+ 2 µL DNA [25 ng/µL] tot: 50 ng			

Table 7. PCR protocol used for *RFC1*: the table reports the reagents, initial concentration, final concentration and final volumes used for amplification of the selected locus.

Reagent	Initial concentration	Final concentration	Volume for the mix
GoTaq® G2 Flexi Buffer (Promega)	5X	1X	3 µL
$MgCl_2$ (Promega)	25 mM	1.5 mM	0.9 µL
dNTPs	2.5 mM	0.2 mM	1.2 µL
DMSO 20%	20X	1X	1.5 µL
Primer Forward	10 pM/µL	0.5 pM/µL	0.75 µL
Primer Reverse	10 pM/µL	0.5 pM/µL	0.75 µL
GoTaq® Polymerase (Promega)	5 U/µL	0.025 U/µL	0.075 µL
H_2O			4.8 µL
Final Volume			13 µL
+ 2 µL DNA [25 ng/µL] tot: 50 ng			

Table 8. PCR protocol used for *FGF14*: the table reports the reagents, initial concentration, final concentration and final volumes used for amplification of the selected locus.

Temperature	Time	Cycles
96°C	5’	1
96°C	30’’	32
58°C	30’’	
72°C	1’	
72°C	5’	1
4°C	+∞	

Table 9. Thermal protocol used for *CMKLR2-AS*: the table reports the temperature, duration and number of cycles used for each amplification step.

Temperature	Time	Cycles
96°C	5’	1
96°C	30’’	35
58°C	30’’	
72°C	1’	
72°C	5’	1
4°C	+∞	

Table 10. Thermal protocol used for *FGF14*: the table reports the temperature, duration and number of cycles used for each amplification step.

Temperature	Time	Cycles
96°C	5’	1
96°C	30’’	33
64°C	30’’	
72°C	1’	
72°C	5’	1
4°C	+∞	

Table 11. Thermal protocol used for *RFC1*: the table reports the temperature, duration and number of cycles used for each amplification step.

Temperature	Time	Cycles
96°C	5’	1
96°C	30’’	30
58°C	30’’	
72°C	30’’	
72°C	5’	1
4°C	+∞	

Table 12. Thermal protocol used for *TCF4*: the table reports the temperature, duration and number of cycles used for each amplification step.

Temperature	Time	Cycles
96°C	5’	1
96°C	30’’	32
60°C	30’’	
72°C	1’	
72°C	5’	1
4°C	+∞	

Table 13. Thermal protocol used for *PLXNA4*: the table reports the temperature, duration and number of cycles used for each amplification step.

Temperature	Time	Cycles
96°C	5’	1
96°C	30’’	30
62°C	30’’	
72°C	30’’	
72°C	5’	1
4°C	+∞	

Table 14. Thermal protocol used for *CRACR2A*: the table reports the temperature, duration and number of cycles used for each amplification step.

4.4 Repeat-primed PCR

Repeat-primed PCR was performed for the *FGF14* locus using the primers reported in Table 5. This assay was used to evaluate the presence of an expansion-compatible amplification pattern at the selected locus.

The reaction was prepared in a final volume of 20 µL, consisting of 18 µL of reaction mix and 2 µL of genomic DNA. The repeat-primed PCR reaction protocol is reported in Table 15, whereas the thermal protocol is reported in Table 16. A ramping rate of 0.5°C/s was applied during the cycling protocol.

PCR products obtained from repeat-primed PCR were subsequently analysed by fragment analysis using capillary electrophoresis, as described in Section 4.5.

Reagent	**Initial concentration**	**Final concentration**	**Volume for the mix**
GoTaq® G2 Flexi Buffer (Promega)	5X	1X	4 µL
$MgCl_2$ (Promega)	25 mM	1.125 mM	0.9 µL
dNTPs	2.5 mM	0.15 mM	1.2 µL
Anchored primer	10 pM/µL	0.5 pM/µL	1 µL
Primer Forward	10 pM/µL	0.5 pM/µL	1 µL
Primer Reverse	5 pM/µL	0.25 pM/µL	1 µL
GoTaq® Polymerase (Promega)	5 U/µL	0.018 U/µL	0.072 µL
H_2O			8.83 µL
Final Volume			18 µL
+ 2 µL DNA [25 ng/µL] tot: 50 ng			

Table 15. Repeat-primed PCR protocol used for *FGF14*: the table reports the reagents, initial concentration, final concentration and final volumes used for repeat-primed PCR amplification of the selected locus.

Temperature	Time	Cycles
96°C	5’	1
96°C	35’’	10
65°C	35’’	
68°C	8’	
96°C	35’’	30
65°C	35’’	
68°C	8’ 20’’	
72°C	10’	1
4°C	+∞	

Table 16. Thermal protocol used for repeat-primed PCR at *FGF14* locus: the table reports the temperature, duration and number of cycles used for each amplification step.

4.5 Microsatellite fragment analysis

PCR products were analysed by fragment analysis using SeqStudio™ Genetic Analyzer (Applied Biosystems™). Before loading, PCR products were diluted according to the intensity and concentration of the amplified band observed after PCR.

For each sample, a loading mixture was prepared using 8.75 µL of formamide (Applied Biosystems™), and 0.25 µL of GeneScan™ 500 ROX™ dye Size Standard (Applied Biosystems™). Then, 9 µL of this mix was dispensed into the plate, and 1 µL of diluted PCR product was added.

Samples were denatured at 96°C for 2 minutes and then immediately placed on ice for 2 minutes. After this step, the plate was loaded onto a capillary electrophoresis analyser for fragment analysis.

Fragment analysis data were visualized using GeneMapper software v6. Detected peaks were used to estimate allele size and assess repeat-length variation at the selected loci. For repeat-primed PCR, sawtooth patterns were evaluated to identify expansion-compatible amplification patterns.

4.6 DNA extraction from agarose gel

PCR products selected for Sanger sequencing that showed two distinct bands were separated by electrophoresis on a 2% agarose gel for *FGF14* and 1.5% agarose gel for *CMKLR2-AS*. For each well, 45 µL of PCR product was mixed with 30 µL of Loading DYE and loaded onto the agarose gel. Following electrophoresis, the DNA band of interest was excised and purified using the NucleoSpin® Gel and PCR Clean-up kit (Macherery Nagel), according to the manufacturer's instructions.

The excised agarose gel band was weighed and transferred to a microcentrifuge tube. For each 100 mg of agarose gel, 200 µL of Binding Buffer NTI were added. Samples were then incubated at 50°C for 5-10 minutes, until complete gel dissolution. The dissolved samples were loaded onto a NucleoSpin® Gel and PCR clean-up column and centrifuged at 11,000 x g for 30 seconds. After centrifugation, the flow-through was discarded and the column was placed back into the collection tube.

The silica membrane was washed twice with 700 µL of Wash Buffer NT3 and centrifuged at 11,000 x g for 30 seconds each time. After each centrifugation, the flow-through was discarded and the column was placed back into the collection tube. The empty column was then centrifuged at 11,000 x g for 1 minute. The columns were then incubated for 2-5 minutes at

70°C to allow evaporation of the residual Wash Buffer NT3, and Elution Buffer NE was also incubated at 70°C to pre-warm. DNA was then eluted with 15-30 µL of Elution Buffer NE. Purified DNA was quantified using a UV-Vis NanoDrop 2000 (Thermoscientific, Waltham, MA, USA) spectrophotometer and subsequently subjected to Sanger sequencing and capillary electrophoresis using a SeqStudio™ Genetic Analyzer (Applied Biosystems™).

4.7 Sanger sequencing

PCR products selected for Sanger sequencing were purified using two different approaches, depending on the number of bands observed after agarose gel electrophoresis. When a single specific amplification product was observed as a single band, the PCR product was purified using the EuroSAP - PCR enzymatic clean-up kit (Euroclone, Pero, MI, Italy), an enzymatic purification system used to remove residual primers and unincorporated dNTPs before sequencing. Conversely,

when multiple distinct amplification products were observed, most commonly two bands related to the heterozygous genotype, the band of interest was excised from the agarose gel and purified as described in Section 4.6.

For EuroSAP purification, a reaction mixture was prepared using nuclease-free water, Exonuclease I and Shrimp Alkaline Phosphatase (SAP). Exonuclease I degrades residual single-stranded primers, whereas SAP dephosphorylates unincorporated dNTPs, thereby preventing their interference with the subsequent sequencing reaction. For each sample, 2.5 µL of Exonuclease I and 2.5 µL of SAP were added to 1 µL of PCR product. Samples were incubated at 37°C for 15 minutes to allow degradation of residual primers and dephosphorylation of unincorporated dNTPs, followed by enzyme inactivation at 80°C for 15 minutes.

The purified products were directly sequenced in the forward or reverse direction, using the primers reported in Table 4, with the Z BigDye^{TM} Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). The composition of the sequencing reaction is reported in Table 17, whereas the thermal protocol is reported in table 18.

Reagent	Volume
Z BigDye™ Terminator v3.1	1 µL
Forward or Reverse primer [3.2]	1 µL
Purified PCR product	3 µL
H_2O	5 µL

Table 17. Sanger sequencing reaction protocol: the table reports the reagents and volumes used for each sequencing reaction.

Temperature	Time	Cycles
96°C	3’	1
96°C	20’’	25
50°C	5’’	
60°C	4’	
10°C	+∞	

Table 18. Thermal protocol for Sanger sequencing: the table reports the temperature, duration and number of cycles used for each sequencing step.

After the sequencing reaction, products were purified by ethanol precipitation. 1 µL of sodium acetate [3.1] (pH=5.2) and 30 µL 100% ethanol were added to each sample, followed by incubation at low temperature for 20 minutes to promote DNA precipitation. Samples were centrifuged at 4°C and 4,000 rpm for 30 minutes, and the supernatant was then carefully removed. The pellet was washed with 30 µL of 70% ethanol to remove residual salts, sodium acetate and other soluble contaminants while maintaining the DNA in a precipitated state. Samples were centrifuged at 4°C and 4,000 rpm for 15 minutes; this washing step was repeated twice. The pellet was then resuspended in 10 µL of formamide. For capillary electrophoresis loading, 10 µL of formamide (Applied Biosystem™) were dispensed into each well of the plate, and 3 µL of the resuspended sequencing product were added.

Samples were denatured at 96°C for 2 minutes and then immediately placed on ice for 2 minutes.

The plate was then loaded onto SeqStudio™ Genetic Analyzer (Applied Biosystems™) for Sanger sequencing analysis. Sequencing results were visualized and analysed using Chromas software.

5. RESULTS

5.1 Results of the bioinformatic analysis

ExpansionHunter Denovo was applied to a subset of 52 patients and identified several repeat expansion signals across the analysed subset of patients. The number of detected loci varied among samples, ranging approximately from 34 to 1800 loci detected per patient. The resulting loci were divided into two major groups: candidate novel repeat loci, and loci already included in the previous ExpansionHunter analysis, which was based on a catalogue of known repeat expansion loci associated with neurodegenerative disorders.

5.1.1 Candidate repeat loci identified by ExpansionHunter Denovo

ExpansionHunter Denovo outlier analysis was first used to identify candidate repeat expansion signals across the subset of 52 patients selected for homogeneous sequencing conditions. Since no control WGS dataset was available, ExpansionHunter Denovo was applied using an outlier-based approach rather than a case-control analysis. Candidate loci were prioritized according to the strength of the outlier signal, using a z-score threshold ≥ 5, together with repeat motif, genomic localization and potential biological relevance of the associated gene.

Following prioritization, 69 outlier signals in the 52 patients showed a z-score ≥ 5. Among these, three candidate loci located in *CMKLR2-AS*, *PLXNA4* and *CRACR2A*, identified in three different patients, were selected for further experimental investigation because of their elevated z-score. However, their presence and repeat structure required validation using targeted molecular approaches.

The main features of these loci are summarized in Table 19, while the experimental results obtained for each candidate locus are described in the dedicated sections below.

Candidate novel repeat loci				
Patient's ID	**Candidate locus**	**Genomic region (hg38)**	**Motif**	**Top case z-score (≥ 5)**
#121	*CMKLR2-AS*	Chr2: 206249133 - 206250293	AAAGG	9.97
#114	*PLXNA4*	Chr7: 132132379 - 132132889	AACAG	10.87
#104	*CRACR2A*	Chr12: 3745017 - 3746100	AAACT	5.69

Table 19. Candidate novel repeat loci selected from ExpansionHunter Denovo outlier analysis: The table summarizes the three high-priority candidate loci not included in the targeted ExpansionHunter catalogue and selected for subsequent experimental investigation. For each locus, the table reports the patient ID, candidate locus, genomic region, repeat motif and top case z-score.

5.1.2 ExpansionHunter Denovo and ExpansionHunter comparison

To evaluate the ability of ExpansionHunter Denovo to detect repeat expansions at established disease-associated loci, the outlier results were compared with those obtained from a previous targeted ExpansionHunter analysis performed at the Genetic Laboratory, using the catalogue of 38 known repeat expansion loci (Table 3).

Among the 38 loci included in the targeted ExpansionHunter catalogue, the comparison identified 19 loci of interest, including *FGF14*, *TCF4*, *RFC1*, *HTT* and *ATXN1*. Seven signals were detected by both ExpansionHunter and ExpansionHunter Denovo; however, only three of these showed a strong ExpansionHunter Denovo outlier signal. The remaining twelve loci were detected only by ExpansionHunter. Among these, patients #75 and #94 were not included in the subset of 52 patients analysed with ExpansionHunter Denovo, and therefore ExpansionHunter Denovo data were not available for these samples. All the data are summarized in Table 20.

ExpansionHunter and ExpansionHunter Denovo comparison						
Patient's ID	**Gene**	**Predicted genotype EH**		**EHDN**	**Region**	**Repeated Motif**
		Allele 1	**Allele 2**			
#75	*TCF4*	28	60	**No data**	Intron	CAG
#94	*TCF4*	23	54	**No data**	Intron	CAG
#114	*TCF4*	33	71	**PREDICTED EXP z: 6.01**	Intron	CAG
#113	*FGF14*	12	143	**x**	Intron	GAA
#114	*FGF14*	10	126	**x**	Intron	GAA
#141	*FGF14*	49	131	**PREDICTED EXP z: 2.20**	Intron	GAA
#129	*FGF14*	43	68	**PREDICTED EXP z: 5.17**	Intron	GAA
#96	*FGF14*	10	155	**PREDICTED EXP z: 1.13**	Intron	GAA
#97	*FGF14*	10	144	**x**	Intron	GAA
#98	*FGF14*	10	129	**x**	Intron	GAA
#138	*FGF14*	22	112	**x**	Intron	GAA
#108	*FGF14*	10	108	**x**	Intron	GAA
#106	*FGF14*	22	112	**x**	Intron	GAA
#110	*ATXN1*	30	36	**x**	Coding: Polyglutamine	TGC
#112	*ATXN1*	30	47	**PREDICTED EXP z: 1.92**	Coding: Polyglutamine	TGC
#111	*HTT*	20	41	**x**	Coding: Polyglutamine	CAG
#111	*RFC1*	8	34	**PREDICTED EXP z: 7.01**	Intron	ACCTC
#146	*RFC1*	10	62	**PREDICTED EXP z: 3.7**	Intron	AAAGG
#99	*RFC1*	32	47	**PREDICTED EXP z: 2.35**	Intron	AAAGG

Table 19. Comparison between ExpansionHunter and ExpansionHunter Denovo results for known repeat expansion loci associated with neurodegenerative diseases: The table summarizes the loci identified among the 38 disease-associated repeat loci included in the targeted ExpansionHunter catalogue. For each patient, the table reports the gene, the predicted genotype obtained with ExpansionHunter, the corresponding ExpansionHunter Denovo outlier signal, the genomic region and the repeated motif. “X” indicates that no signal was detected by ExpansionHunter Denovo.

5.2 Experimental validation of candidate repeat loci identified by ExpansionHunter Denovo

ExpansionHunter Denovo identified candidate repeat expansion signals at loci not included in the targeted ExpansionHunter catalogue. A subset of these candidate loci with an ExpansionHunter Denovo z-score ≥ 5 was selected for experimental investigation (Table 19). Molecular analyses were then performed to evaluate whether the computational signals corresponded to a detectable repeat-containing regions in the analysed patients.

5.2.1 *CMKLR2-AS*

The first candidate locus selected for experimental investigation was located at the *CMKLR2-AS* region and was identified by ExpansionHunter Denovo in patient #121. This locus was prioritized because it showed a high outlier signal, with a top case z-score of 9.97 (Table 19). Conventional flanking PCR was performed to evaluate amplification of the repeat-containing region. Agarose gel electrophoresis showed two amplification products in patient #121, with approximate sizes of 250 bp and 500 bp, estimated by comparison with the molecular-weight marker. The control sample showed only two lower bands of approximately 250-300 bp, and no amplification product was present in the negative control (Figure 6). An additional upper band was also visible in patient #121 but was interpreted as a non-specific amplification product.

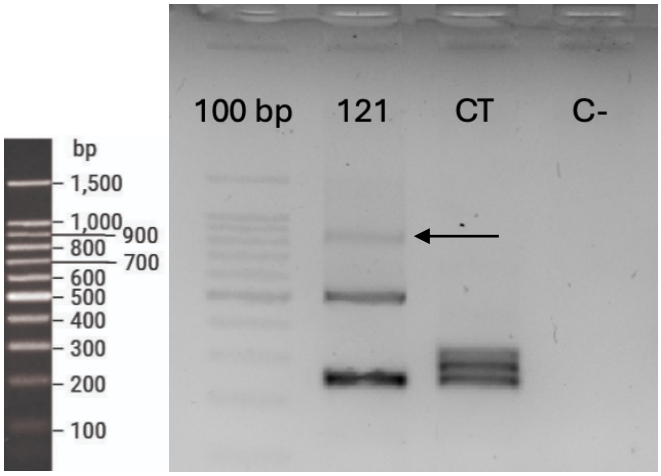

Figure 6. Conventional flanking PCR analysis of *CMKLR2-AS*: Agarose gel electrophoresis of PCR products obtained from patient #121, a control sample and a negative control. Patient #121 showed two amplification products of approximately 250 bp and 500 bp. The additional upper band, indicated by the arrow, was interpreted as a non-specific product; CT, control sample; C-, negative control; 100bp, 100 bp DNA ladder (Promega).

Microsatellite fragment analysis was then performed to obtain a more precise estimate of allele size. In patient #121, two alleles of 248.56 bp and 517.93 bp were detected, corresponding to approximately 5 and 58 repeats, respectively. No additional peak corresponding to the upper band observed on agarose gel was detected, supporting the interpretation of this band as a non-specific product. The control sample showed two peaks of 248.24 bp and 282.95 bp corresponding to approximately 5 and 12 repeat units, respectively, consistent with non-expanded repeat sizes (Figure 7).

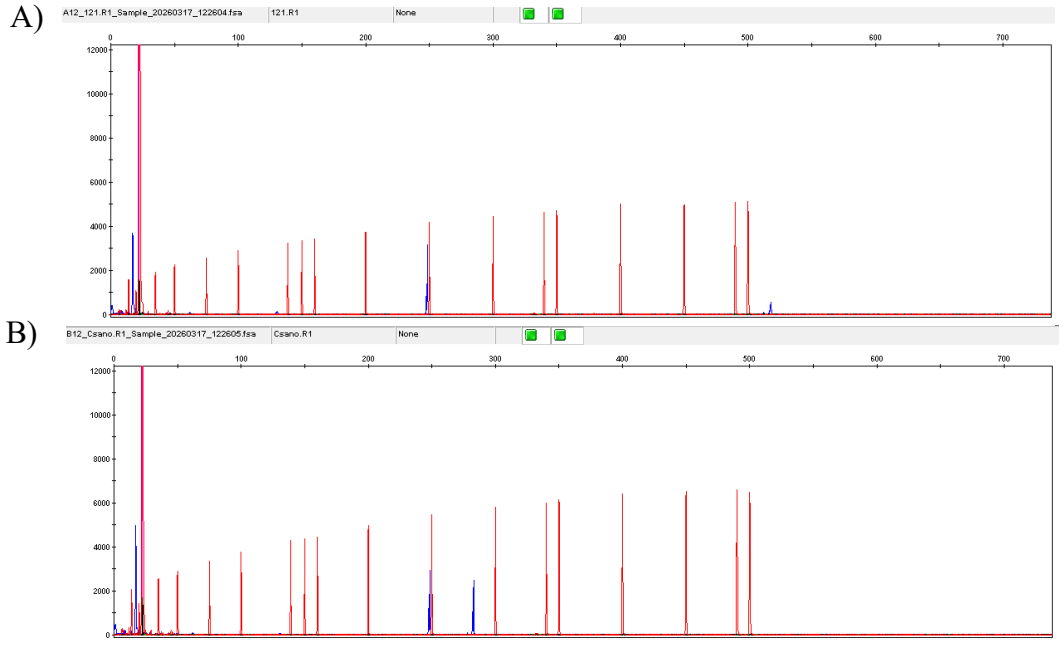

Figure 7. Microsatellite fragment analysis of *CMKLR2-AS*: fragment analysis profile of the patients. **A)** patient #121, detected allele sizes: 248.56-517.93 bp **B)** control sample, detected allele sizes: 248.48-282.95 bp.

To characterize the repeat structure of the amplified alleles, the two target-specific amplification products and the additional upper amplification products observed in patient #121 were excised from agarose gel, purified and analysed by Sanger sequencing. Sequencing of the lower band showed a short CTTTT motif repeated 5 times, underlined in red, consistent with the allele size estimated by microsatellite fragment analysis (Figure 8A). Sequencing of the highest amplification product revealed a complex pentanucleotide repeat structure composed of an initial CTTTT motif repeated 3 times, a central CTTTC motif repeated approximately 48 times, and a terminal CTTTT motif repeated 7 times. The CTTTC tract is underlined in green, whereas the CTTTT tracts are underlined in red (Figure 8B). Overall, the repeat structure was compatible with the larger allele estimated by microsatellite fragment analysis, suggesting that the motif detected in this band corresponded to the larger allele.

Figure 8. Sanger sequencing of CMKLR2-AS in patient #121: A) The lower allele showed a CTTTT motif repeated 5 times, underlined in red. B) The highest amplification product, corresponding to the larger allele, showed a complex pentanucleotide repeat structure composed of CTTTC and CTTTT repeat units. The CTTTC repeated motif is underlined in green, while the CTTTT repeat units is underlined in red. The circled "C" indicates a single-nucleotide interruption within the repeated-tract.

To evaluate whether this repeat size was specific to the index patient, the *CMKLR2-AS* locus was subsequently analysed in 45 control samples. Most controls showed estimated repeat numbers within the 5-20 repeat range. However, four control samples carried larger alleles, with repeat size compatible to or greater than the larger allele detected in patient #121. The results for these four controls are reported in Table 21.

Sample	Allele size (bp)	Estimated repeats
Patient #121	248.56 – 517.93	5 - 58
C21	292.82 – 600.68	14 – 75
C34	287.88 – 483.46	13 – 52
C35	248.61 – 494.40	5 – 54
C42	248.53 – 603.93	5 – 76

Table 21. Comparison between patient 121 and control samples carrying larger alleles at the *CMKLR2-AS* repeat locus: The table summarizes the allele sizes and the estimated repeat numbers for patient #121 and for control samples showing alleles comparable to or greater than the larger allele detected in patient #121.

Overall, molecular analysis confirmed that the *CMKLR2-AS* locus contains a polymorphic pentanucleotide repeat. Although patient #121 carried a larger allele of approximately 58 repeats, alleles of comparable or greater size were also detected in control samples. Therefore, after molecular screening, the *CMKLR2-AS* repeat was not considered a disease-associated outlier signal, in contrast with the outlier signal detected by ExpansionHunter Denovo.

5.2.2 *PLXNA4*

The second candidate locus selected for experimental investigation was located at the *PLXNA4* region and was identified by ExpansionHunter Denovo in patient #114. This locus was prioritized because it showed a high outlier signal, with a top case z-score of 10.87 (Table 19). The *PLXNA4* candidate locus was analysed by conventional flanking PCR to evaluate amplification of the repeat-containing region. Agarose gel electrophoresis showed a comparable amplification pattern between patient #114 and a control sample. Both samples showed amplification products of approximately 380 bp and 400 bp by comparison with the molecular-weight marker. An additional amplification product was visible in both patient and control sample; therefore, it was interpreted as a non-specific product. No amplification product was present in the negative control (Figure 9).

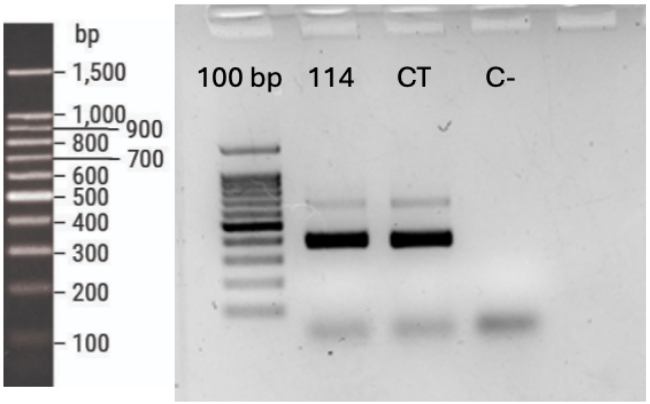

Figure 9. Conventional flanking PCR analysis of *PLXNA4*: Agarose gel electrophoresis of PCR products obtained from patient #114, a control sample and a negative control. Both patient and control sample showed amplification products of approximately 380 bp and 400 bp, the third band is an aspecific product; CT, control sample; C-, negative control; 100bp, 100 bp DNA ladder (Promega).

Microsatellite fragment analysis confirms the presence of two alleles in both patient #114 and the control sample. Patient #114 showed allele sizes of 378.40 bp and 387.14 bp (Figure 10A), while control sample showed allele sizes of 367.61 bp and 387.14 bp (Figure 10B). The allele sizes detected in patient #114 overlapped with those observed in control.

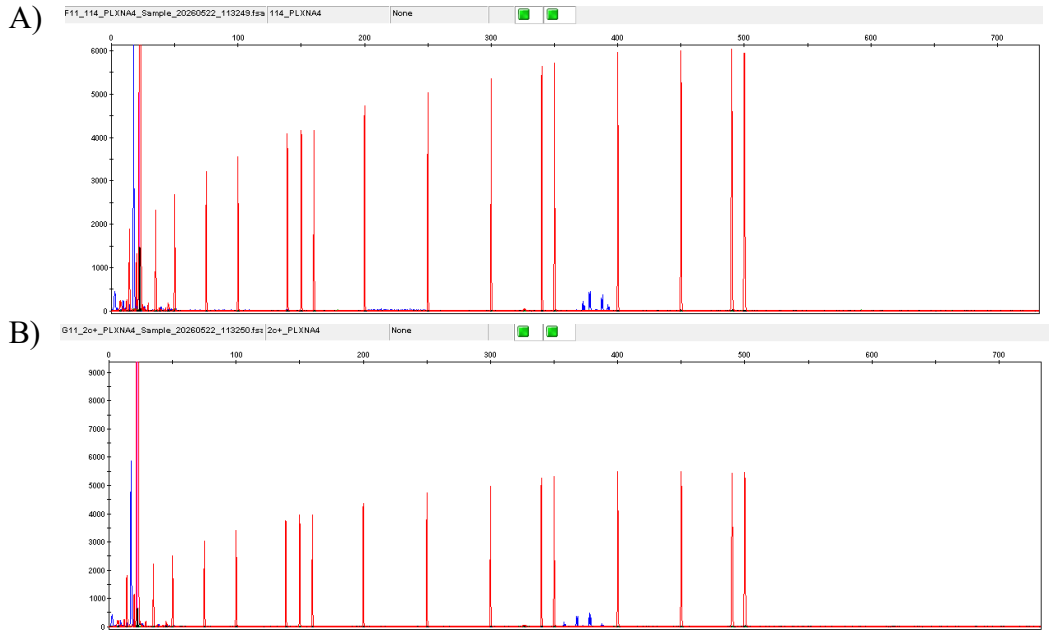

Figure 10. Microsatellite fragment analysis of *PLXNA4*: fragment analysis profile of the patients. **A)** patient #114, detected allele sizes: 378.40-387.14 bp; **B)** control sample, detected allele sizes: 367.61-387.14 bp.

Overall, molecular analysis did not provide evidence of the presence of a disease-associated repeat expansion at the *PLXNA4* locus. Although ExpansionHunter Denovo identified this locus as a strong outlier signal in patient #114, both agarose gel electrophoresis and microsatellite fragment analysis showed an allelic pattern comparable to that observed in the control sample. Therefore, based on the molecular results, the *PLXNA4* repeat was not considered a disease-associated outlier signal, despite the strong outlier signal detected by ExpansionHunter Denovo.

5.2.3 *CRACR2A*

The third candidate locus selected for experimental validation was located at the *CRACR2A* locus and was identified by ExpansionHunter Denovo in patient #104. This locus showed an AAACT repeat motif and a top z-score of 5.69 (Table 19).

The *CRACR2A* candidate locus was analysed by conventional flanking PCR to evaluate whether ExpansionHunter Denovo outlier signal was associated with a detectable increase in allele size in patient #104. Agarose gel electrophoresis showed two distinct bands in patient #104, with approximate sizes of 300 bp and 500 bp, estimated by comparison with the

molecular-weight marker. The control sample showed two amplification products of approximately 280-300 bp, whereas no amplification product was observed in negative control (Figure 11).

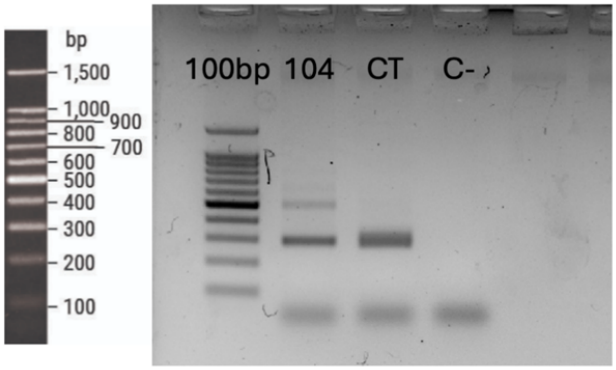

Figure 11. Conventional flanking PCR analysis of *CRACR2A*: Agarose gel electrophoresis of PCR products obtained from patient #104, a control sample and a negative control. Patient #104 showed two amplification products of approximately 300 bp and 500 bp. The control sample showed two amplification products of approximately 280-310 bp; CT, control sample; C-, negative control; 100bp, 100 bp DNA ladder (Promega).

Microsatellite fragment analysis confirmed the presence of two alleles in both patient #104 and control sample. Patient #104 showed allele sizes of 291.28 bp and 502.86 bp (Figure 12A), whereas the control sample showed allele sizes of 277.41 bp and 310.36 bp (Figure 12B). The larger allele detected in patient #104 was therefore larger than both alleles observed in the control sample.

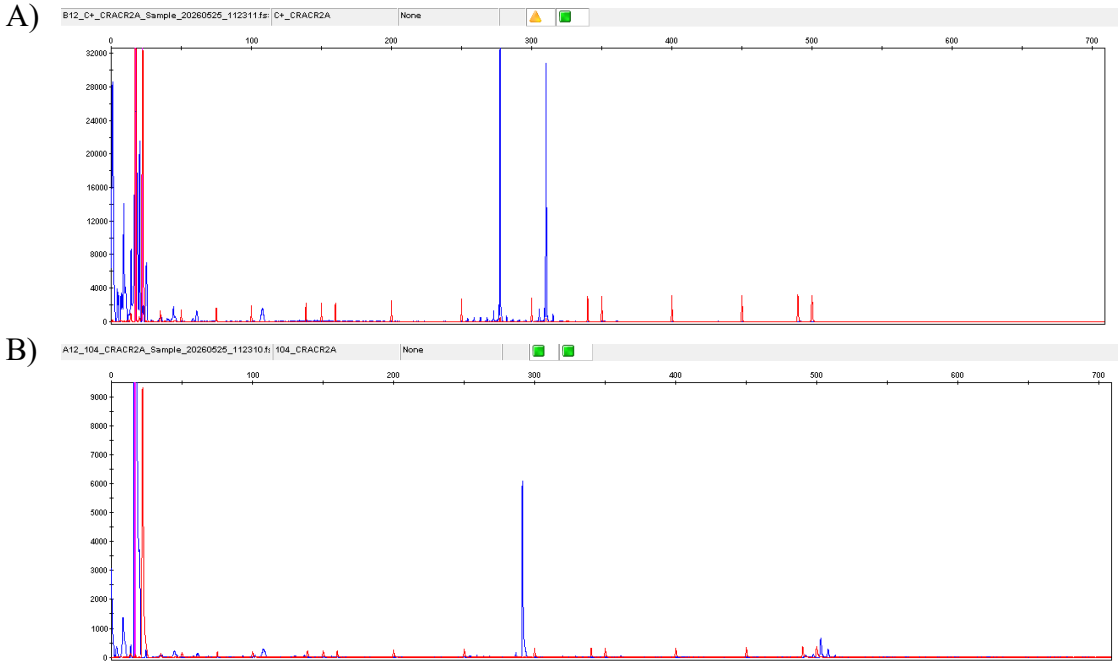

Figure 12. Microsatellite fragment analysis of *CRACR2A*: fragment analysis profile of the patients. **A)** patient #104, detected allele sizes: 291.28-502.86 bp; **B)** control sample, detected allele sizes: 277.41-310.36 bp.

To evaluate whether the larger allele detected in patient #104 was specific to the index patient, the *CRACR2A* locus was subsequently analysed in an additional set of 30 control samples. This screening showed that alleles compatible with increased repeat size were also observed in at least one control sample, indicating that the larger allele was not unique to this patient. Therefore, based on these results, the *CRACR2A* repeat observed was not interpreted as a disease-specific outlier, although it had initially emerged as an ExpansionHunter Denovo outlier.

5.3 Experimental validation of already known disease-associated repeat loci

In addition to candidate loci identified by ExpansionHunter Denovo, selected established disease-associated repeat loci were experimentally investigated. These loci were selected based on the comparison between the previous targeted ExpansionHunter analysis and the ExpansionHunter Denovo outlier results, summarized in Table 20. Molecular validation was performed to assess whether the bioinformatic predictions corresponded to detectable repeat-size variation and to evaluate the concordance between ExpansionHunter, ExpansionHunter Denovo and targeted molecular analyses. The following sections describe the results obtained for *TCF4*, *FGF14* and *RFC1*.

5.3.1 *TCF4*

ExpansionHunter identified expanded CAG repeat alleles at the intronic *TCF4* locus in patients #75, #94, #114, with predicted genotypes of 28-60, 23-54, 33-71 repeats, respectively (Table 20). ExpansionHunter Denovo results were available only for patient #114, who showed an outlier signal with a z-score of 6.01. Patients #75 and #94 were not included in the subset analysed with ExpansionHunter Denovo.

The three patients were investigated by conventional flanking PCR. Agarose gel electrophoresis showed two amplification products of different sizes in each patient. Based on comparison with the molecular-weight marker, patient #75 and #94 showed bands of approximately 300 and 450 bp, whereas patient #114 showed bands of approximately 320 and 500 bp, no amplification product was observed in the negative control (Figure 13).

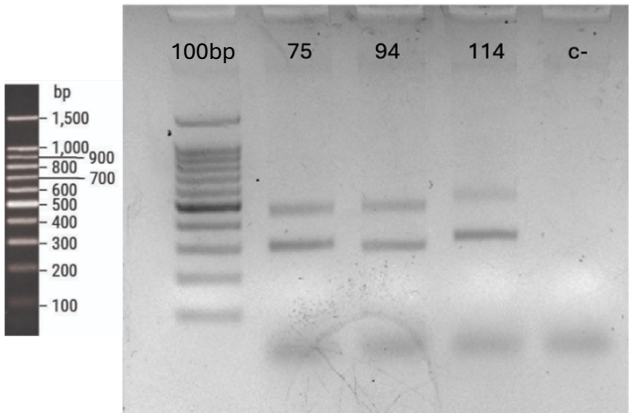

Figure 13. Conventional flanking PCR analysis of *TCF4*: Agarose gel electrophoresis of PCR products of different size showed two amplification products of approximately 300-450 bp in patients #75 and #94, and two amplification products of approximately 320-500 bp in patient #114; C-, negative control; 100bp, 100 bp DNA ladder (Promega).

Microsatellite fragment analysis allowed a more precise estimation of allele sizes confirming the presence of the two alleles in all three patients. Fragment sizes of 299.01 and 445.42 bp were detected in patient #75, 285.28 and 445.42 in patient #94 and 316.32 and 492.88 in patient #114 (Figure 14).

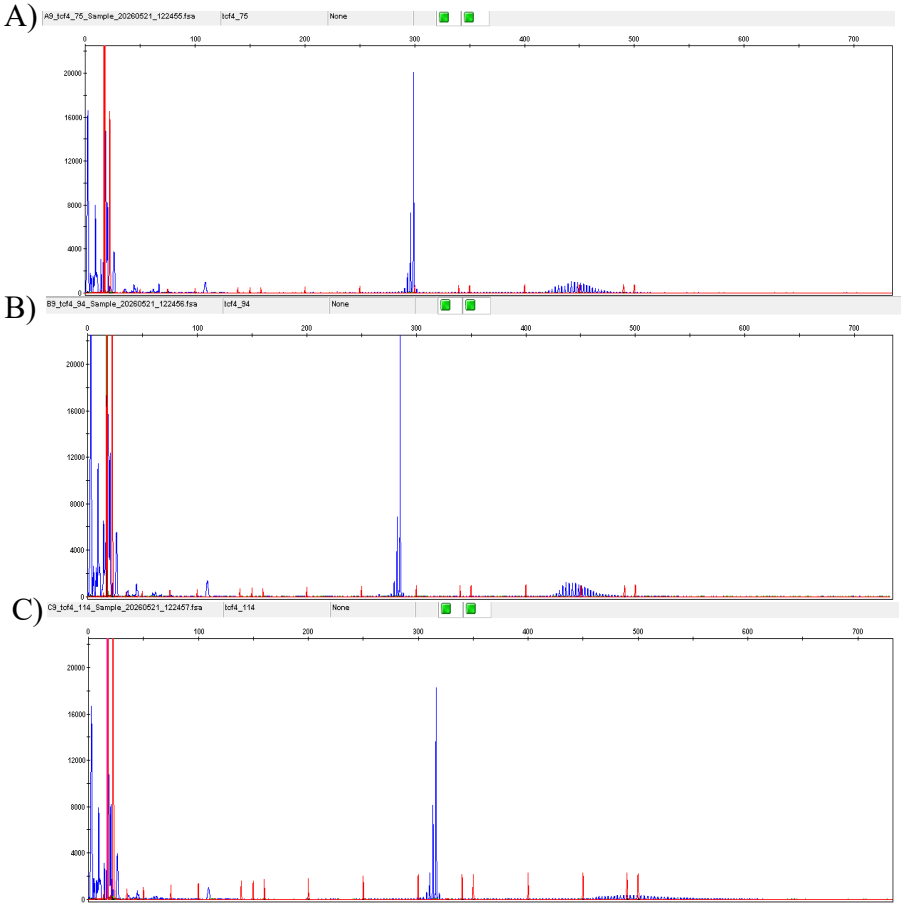

Figure 14. Microsatellite fragment analysis of *TCF4*: fragment analysis profile of the patients. **A)** patient #75, detected allele sizes: 299.01 - 445.42 bp; **B)** patient #94, detected allele sizes: 285.28 – 445.42 bp; **C)** patient #114, detected allele sizes: 316.32 – 492.88 bp.

The shorter allele of a healthy control was characterized by Sanger sequencing and used as a reference to estimate the number of CAG repeats in patient alleles. The comparison between ExpansionHunter predictions, ExpansionHunter Denovo outlier signals results and the estimated number of repeats are reported in Table 22.

Patient's ID	Gene	Predicted genotype EH		EHdn	N' of repeats estimated by validation		Pathogenic threshold
		Allele 1	Allele 2		Repeats A1	Repeats A2	
#75	*TCF4*	28	60	**No data**	28	77	≥ 50
#94	*TCF4*	23	54	**No data**	23	77	≥ 50
#114	*TCF4*	33	71	**PREDICTED EXP z: 6.01**	34	93	≥ 50

Table 22. Comparison of predicted and experimentally estimated CAG repeat sizes at the TCF4 locus: The table reports the ExpansionHunter predicted genotype, the corresponding ExpansionHunter Denovo result, the experimentally estimated number of repeats for each allele and the pathogenic threshold used for interpretation.

Molecular analysis confirmed the presence of expanded *TCF4* alleles in all three patients, with larger alleles exceeding the pathogenic threshold of 50 CAG repeats. ExpansionHunter correctly identified the expanded alleles but underestimated their size, while ExpansionHunter Denovo supported the molecular finding only in patient #114, the only one among the three included in the ExpansionHunter Denovo subset.

5.3.2 *FGF14*

The intronic GAA repeat locus in *FGF14* was investigated in patients #96, #97, #98, #106, #108, #113, #114, #129, #138, #141, selected based on the ExpansionHunter results reported in Table 20. ExpansionHunter Denovo identified outlier signals in patients #129 #96, and #141 with the strongest signal observed in patients #129 (z-score= 5.17), whereas patient #96 and #141 showed a weaker signal, with a z-score of 1.13 and 2.20 respectively. No ExpansionHunter Denovo signal was detected in the remaining patients.

Conventional flanking PCR generated two amplification products of different sizes in all analysed patients (Figure 15).

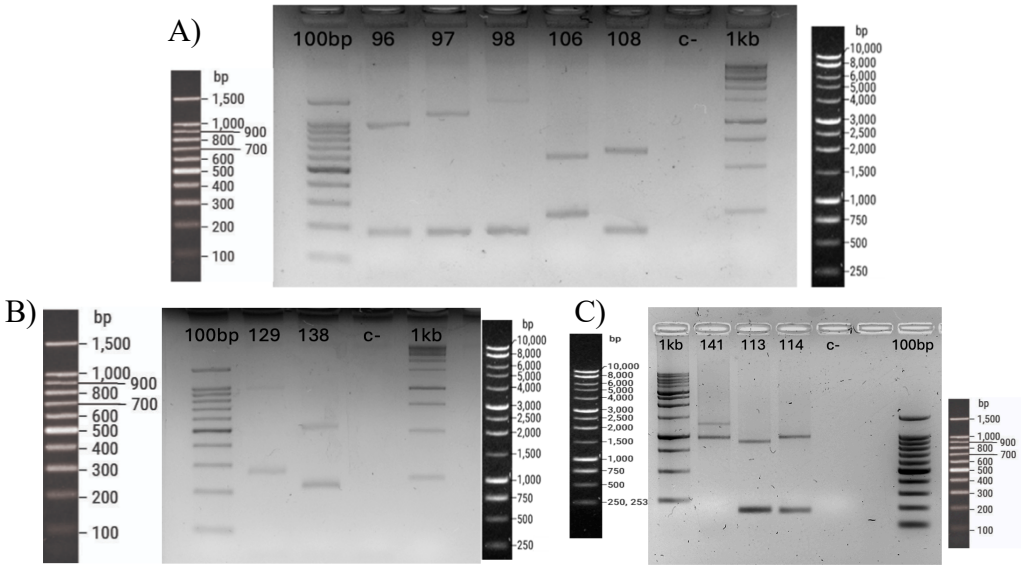

Figure 15. Conventional flanking PCR for *FGF14*: Agarose gel electrophoresis of PCR products obtained from **A)** patients #96, #97, #98, #106, #108; **B)** patients #129 and #138; **C)** patients #141, #113, #114. Two amplification products of different sizes are detected in each patient. C-, negative control; 100bp, 100 bp DNA ladder (Promega); 1kb, 1 kb DNA ladder (Promega).

The approximate allele sizes were estimated by comparison with the molecular-weight markers. Repeat number was estimated using a fixed flanking sequence length of 150 bp, corresponding to an allele with 0 GAA repeats. These values were then used to estimate the corresponding number of repeat units. The comparison between ExpansionHunter predictions, ExpansionHunter Denovo results, allele-size estimates and estimated number of repeat units is reported in Table 23.

Patient's ID	Gene	Predicted genotype EH		EHdn	Allele size estimated by agarose gel electrophoresis (bp)		Estimated n' of repeats by validation		Pathogenic threshold
		Allele 1	Allele 2		Allele 1	Allele 2	Repeats A1	Repeats A2	≥ 250-300
#96	*FGF14*	10	155	**PREDICTED EXP z: 1.13**	180 bp	950 bp	9	267	≥ 250-300
#97	*FGF14*	10	144	**No signal**	180 bp	1200 bp	9	350	≥ 250-300
#98	*FGF14*	10	129	**No signal**	190 bp	1500 bp	13	450	≥ 250-300
#106	*FGF14*	22	112	**No signal**	250 bp	620 bp	33	157	≥ 250-300
#108	*FGF14*	10	108	**No signal**	190 bp	690 bp	13	180	≥ 250-300
#113	*FGF14*	12	143	**No signal**	200 bp	900 bp	16	250	≥ 250-300
#114	*FGF14*	10	126	**No signal**	190 bp	1020 bp	13	289	≥ 250-300
#129	*FGF14*	43	68	**PREDICTED EXP z: 5.17**	280 bp	1000 bp	43	283	≥ 250-300
#138	*FGF14*	22	112	**No signal**	220 bp	530 bp	23	127	≥ 250-300
#141	*FGF14*	49	131	**PREDICTED EXP z: 2.20**	1000 bp	1400 bp	283	416	≥ 250-300

Table 23. Comparison of predicted and experimentally estimated repeat sizes at the *FGF14* locus: The table reports ExpansionHunter predicted genotypes, ExpansionHunter Denovo results, allele sizes estimated by agarose gel electrophoresis, estimated number of repeat units and the pathogenic threshold used for interpretation.

Repeat-primed PCR was subsequently performed using a GAA-specific primer in patients #96, #97, #98, #113, #114, #129 and #141, who were selected because conventional flanking PCR showed large alleles compatible with repeat expansion at the *FGF14* locus. The aim was to assess whether the amplified alleles contained a pure GAA repeat tract, since the repeat-specific primer was designed to anneal to the GAA motif.

Saw-tooth amplification patterns compatible with the presence of a pure GAA repeat expansion were observed in patients #96, #97, #98, #113, #114, #141 (Figure 16). In contrast, no characteristic repeat-primed PCR pattern was observed in patient #129, despite the presence of a larger allele detected by conventional flanking PCR.

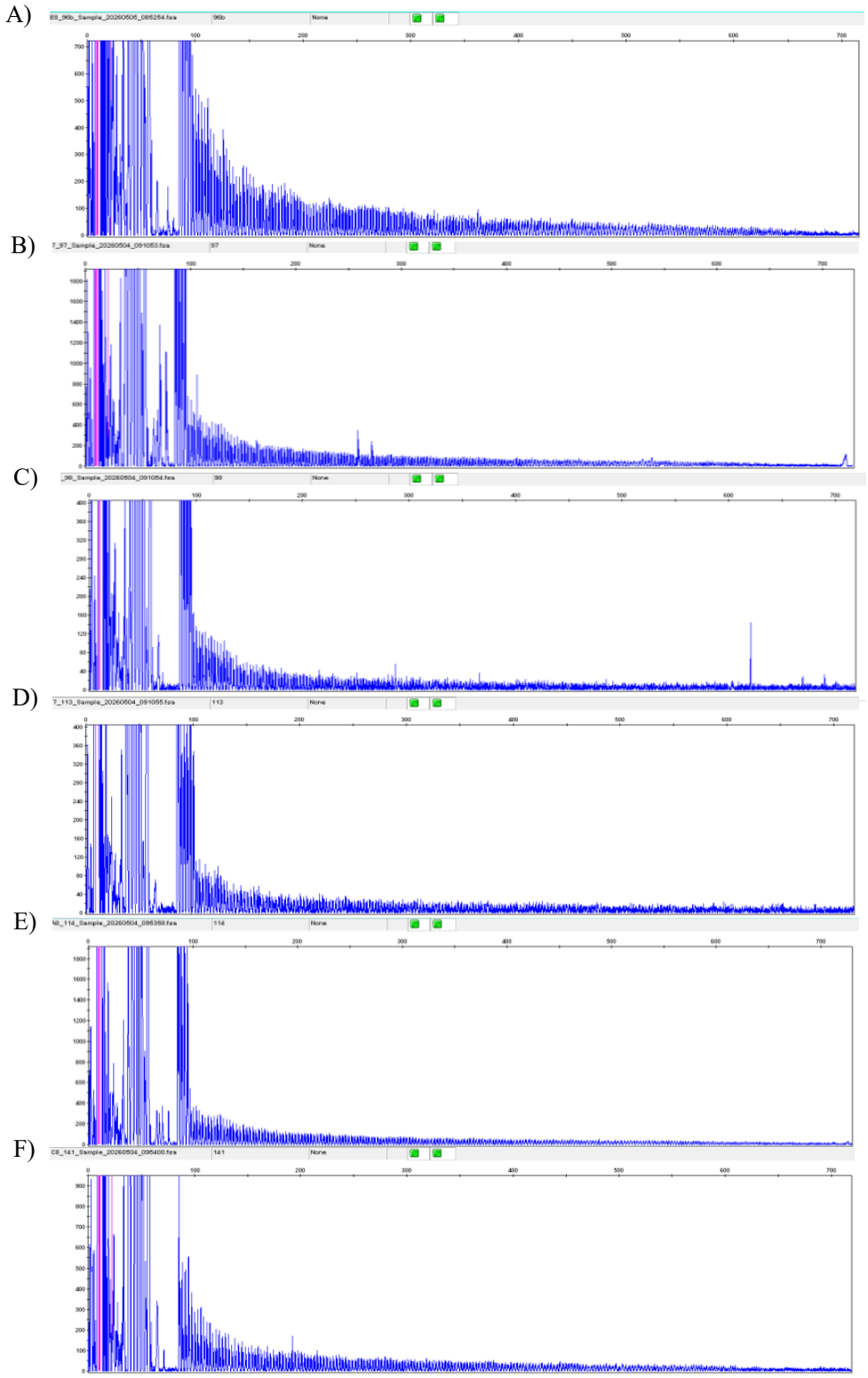

Figure 16. Microsatellite analysis of repeat-primed PCR for *FGF14*: electropherograms from patients **A)** #96, **B)** #97, **C)** #98, **D)** #113, **E)** #114, **F)** #141 showed expansion-compatible saw-tooth amplification patterns, supporting the presence of pure GAA repeat tracts.

The absence of a characteristic repeat-primed PCR in patient #129 was discordant with the larger allele detected by conventional flanking PCR. This result suggested that the repeat tract might not have a pure GAA motif composition, or that interruptions or alternative repeat motifs could have affected the annealing of the GAA-specific repeat primer.

To further investigate this hypothesis, the larger amplification product from patient 129 was excised from the agarose gel, purified and analysed by Sanger sequencing.

Sequencing analysis showed that the larger allele did not consist of a canonical uninterrupted GAA repeat tract. Instead, the sequence showed a complex repeat structure characterized by a mixed GAA/GGA motif followed by a terminal GAA repeat region (Figure 17).

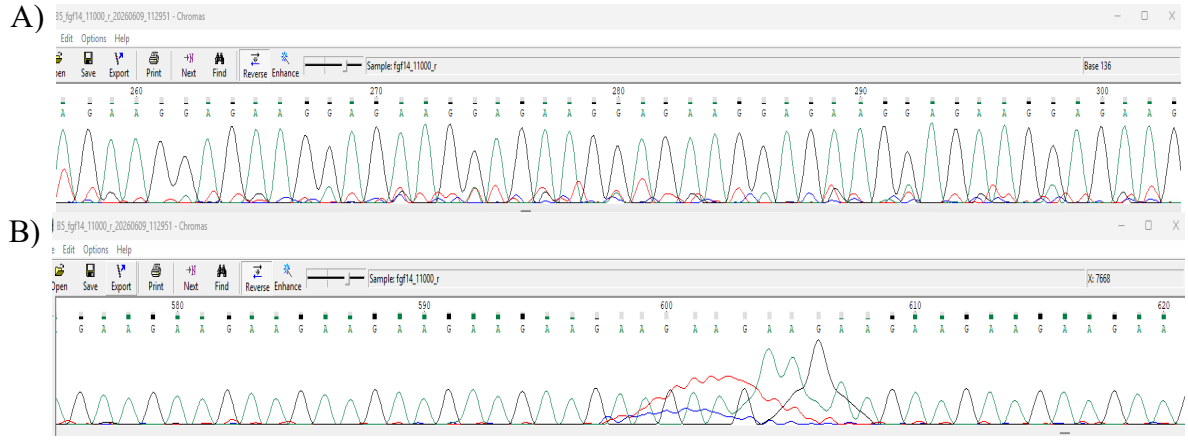

Figure 17. Sanger sequencing of the larger *FGF14* allele in patient 129: the sequence showed **A)** a complex repeat structure composed of a long-mixed GAA/GGA motif and a **B)** shorter terminal GAA-repeat region.

Overall, molecular analysis detected larger alleles compatible with repeat expansion at the *FGF14* locus in several patients. Repeat-primed PCR showed expansion-compatible saw-tooth patterns in patients #96, #97, #98, #113, #114 and #141, supporting the presence of pure GAA repeat expansions. In patient #129, Sanger sequencing showed a larger allele with a complex mixed GAA/GGA repeat structure.

ExpansionHunter identified larger alleles at the *FGF14* locus but underestimated their size, particularly for long alleles. ExpansionHunter Denovo detected outlier signals only in patients #129 #96, and #141 whereas no corresponding signal was observed in the remaining patients, despite the presence of larger alleles detected by molecular analysis.

5.3.3 *RFC1*

The intronic repeat locus in *RFC1* was investigated in patients #99, #111 and #146, for whom repeat signals were detected by both ExpansionHunter and ExpansionHunter Denovo (Table 20).

Conventional flanking PCR was performed to evaluate amplification of the repeat-containing region. Agarose gel electrophoresis showed amplification products in all three patients (Figure 18). The approximate size of each amplification product was estimated by comparison with the molecular-weight marker. Patient #99 showed approximately a lower band of 600 bp and an upper band of 1000 bp, whereas patients #111 and #146 showed one clearly detectable amplification product of approximately 350 bp.

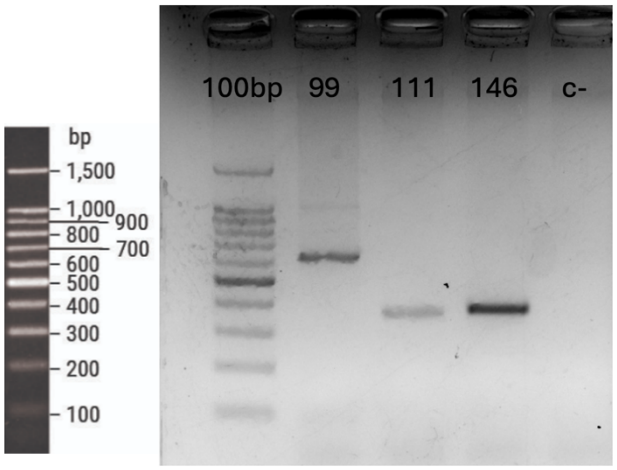

Figure 18. Conventional flanking PCR analysis of *RFC1*: Agarose gel electrophoresis of PCR products obtained from patients #99, #111 and #146. Patient #99 showed approximately an upper band of 1000 bp and a lower band of 600 bp, patients #111 and #146 showed a band of approximately 350 bp; C-, negative control; 100bp, 100 bp DNA ladder (Promega).

Microsatellite fragment analysis was then performed to obtain a precise estimate of the expansion. Patient #99 showed a fragment of 605.23 bp, while patients #111 and #146 showed fragments of 359.28 and 368.14, respectively (Figure 19). The larger band observed in patient #99 by agarose gel electrophoresis was not detected by microsatellite fragment analysis.

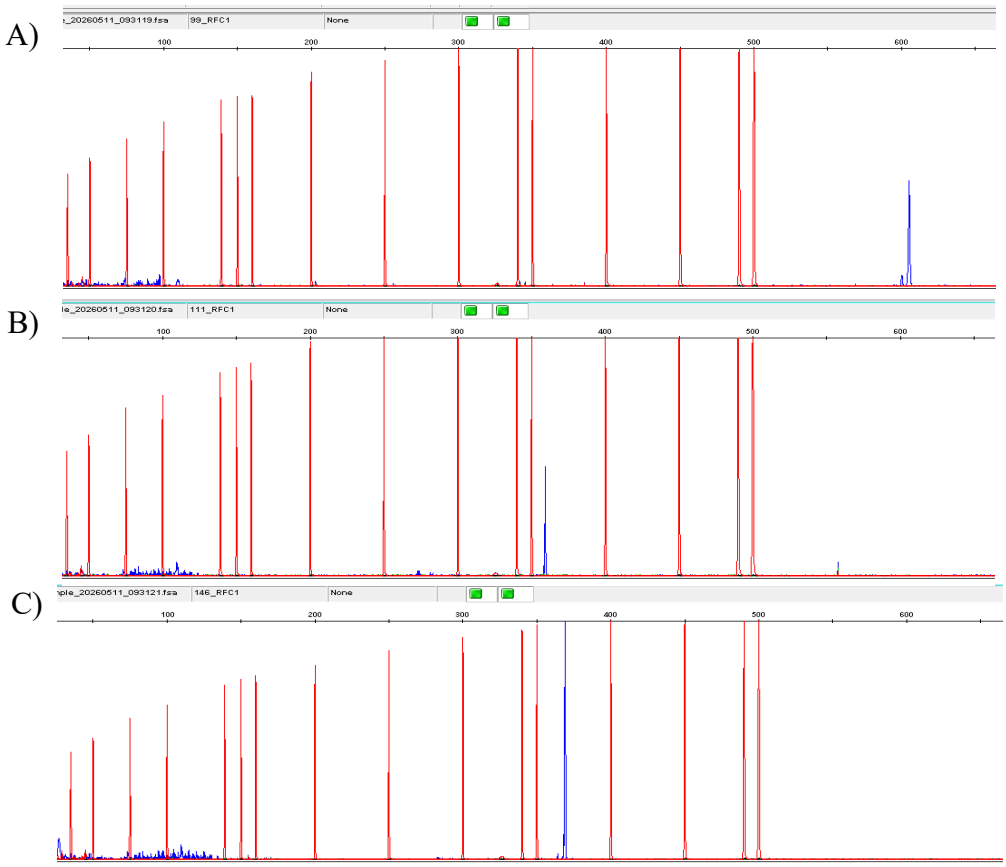

Figure 19. Microsatellite fragment analysis of *RFC1*: fragment analysis profile of the patients. **A)** patient #99, detected allele size: 605.23 bp **B)** patient #111, detected allele size: 359.28 bp; **C)** patient #146, detected allele size: 368.14 bp.

Repeat number was estimated using a fixed sequence length of 293 bp, corresponding to an allele with 0 AAAAG repeats. The comparison between ExpansionHunter predictions, ExpansionHunter Denovo results, allele-size estimates and estimated number of repeat units is reported in Table 24.

Patient's ID	Gene	Predicted genotype EH		EHdn	N' of repeats estimated by validation		Pathogenic threshold
		Allele 1	Allele 2		Repeats A1	Repeats A2	
#99	*RFC1*	32	47	**PREDICTED EXP z: 2.35**	62	Not detected	>250 or >400 depends on the motif
#111	*RFC1*	8	34	**PREDICTED EXP z: 7.01**	13	Not detected	>250 or >400 depends on the motif
#146	*RFC1*	10	62	**PREDICTED EXP z: 3.7**	15	Not detected	>250 or >400 depends on the motif

Table 24. Comparison of predicted and experimentally estimated repeat sizes at the *RFC1* locus: The table reports ExpansionHunter predicted genotypes, ExpansionHunter Denovo results, the estimated number of repeat units and the pathogenic threshold used for interpretation. Pathogenic interpretation depends on repeat motif and biallelic expansion status.

Sanger sequencing was performed in patient #111 to characterize the repeat motif, as ExpansionHunter Denovo had indicated an atypical ACCTC motif at this locus. However, sequencing analysis identified a canonical AAAAG repeat motif (Figure 20).

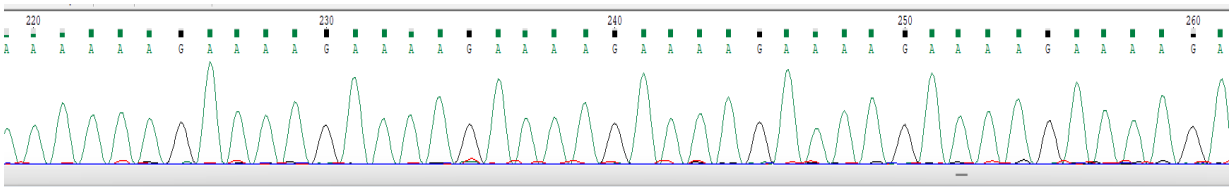

Figure 20. Sanger sequencing of *RFC1* in patient 111: the chromatogram showed an AAAAG repeat motif.

Overall, microsatellite fragment analysis detected a single allele size in each patient. In patient #99, the larger band observed by agarose gel electrophoresis was not detected by fragment analysis. In patients #111 and #146, the larger alleles predicted by ExpansionHunter were not observed in the molecular analyses performed. Based on the available molecular results, a pathogenic biallelic *RFC1* repeat expansion was not confirmed in the analysed patients.

6. DISCUSSION

Repeat expansions represent an important class of genetic variants involved in neurological and neurodegenerative disorders, but their detection remains technically challenging, especially when expanded alleles are large or structurally complex.

Although WGS is increasingly used in neurogenetics, standard short-read variant-calling pipelines are not sufficient to reliably detect and characterize these repeat expansions. To address this limitation, several dedicated bioinformatic tools have been developed, including catalogue-based approaches such as ExpansionHunter and catalogue-free approaches such as ExpansionHunter Denovo.

The first relevant aspect of this study was the interpretation of candidate repeat loci identified by ExpansionHunter Denovo. ExpansionHunter Denovo detects repeat expansion signals by identifying in-repeat reads in short-read WGS data and comparing these signals across samples. In an outlier analysis, candidate loci are prioritized when a single sample exhibits a repeated signal that deviates from the background distribution of the analysed cohort. Therefore, even if a high z-score indicates that a locus is an outlier within the analysed dataset, it does not by itself demonstrate that the repeat is pathogenic or disease-associated.[27] The use of a z-score threshold is consistent with previous short-read tandem repeat analysis protocols, in which z-scores are used to identify repeat alleles that deviate from the distribution of the reference population. In particular, Halman et al. defined population outliers as alleles with a z-score ≥ 3.72.[31]

In this study, ExpansionHunter Denovo was applied as a catalogue-free outlier analysis in a clinically heterogeneous cohort for which no matched control WGS dataset was available. The aim was to evaluate its performance and reliability in patients affected by neurodegenerative diseases. Candidate loci were prioritized using the ExpansionHunter Denovo z-score. Using a z-score ≥ 5, 69 loci were selected. Among these, three regions located in *CMKLR2-AS*, *PLXNA4* and *CRACR2A,* identified in patients #121, #114 and #104 respectively, were selected for molecular investigation based on repeat motif, genomic localization and feasibility of molecular validation for experimental validation by fragment analysis. These loci were also chosen because DNA was available for the corresponding patients and the repeat-containing regions were suitable for validation using flanking PCR and microsatellite fragment analysis. Therefore, in the present study, a z-score threshold of 5 was selected as a conservative operational cut-off to prioritize strong outlier signals for experimental validation.

Molecular validation confirmed that some ExpansionHunter Denovo signals corresponded to detectable repeat-length variation. At the *CMKLR2-AS* locus, patient #121 carried a larger

pentanucleotide repeat allele, compatible with the bioinformatic outlier signal. However, screening of control samples showed that alleles of comparable or greater size were also present in controls. Similarly, at the *CRACR2A* locus, the larger allele detected in patient #104 was not unique to the patient, since alleles compatible with increased repeat size were also observed in at least one control sample. These findings suggest that both loci are polymorphic in the general population, but do not support their interpretation as disease-associated repeat expansion candidates in the analysed patients.

At the *PLXNA4* locus, ExpansionHunter Denovo identified a strong outlier signal, with a z-score of 10.87, in patient #114. However, conventional PCR and microsatellite fragment analysis did not detect a large expanded allele, but showed two alleles of small sizes comparable to those present in control population. Therefore, the molecular results suggest that the *PLXNA4* outlier signal detected by ExpansionHunter Denovo may represent a false-positive signal. Taken together, these findings show that ExpansionHunter Denovo is able to highlight repeat-variable regions outside predefined catalogues, but that, in absence of controls, the clinical interpretation of these signals requires additional evidence.

Importantly, the present analysis does not exclude the relevance of other candidate loci detected by ExpansionHunter Denovo. Only a subset of signals was selected for validation, based on z-score, availability of DNA and feasibility of rapid molecular testing. Additional loci remain to be investigated. Future analyses should include larger cohorts and appropriate control datasets, which would improve the definition of the background repeat distribution and help reduce the number of false-positive outlier signals. Overall, this first validation phase indicates that ExpansionHunter Denovo outlier analysis, even when applied to a small cohort, can detect large expanded alleles, although the lack of matched WGS control data limits the ability to distinguish potentially disease-associated expansions from polymorphic repeat variation.

To further evaluate the reliability of ExpansionHunter Denovo in detecting expanded alleles, the signals identified by ExpansionHunter Denovo were compared with the genotypes obtained using ExpansionHunter in the same patient cohort and at loci known to be associated with disease (Table 20). ExpansionHunter Denovo detected an outlier signal in 8 out of 17 loci, although only three of these showed a z-score above 5: *TCF4* in patient #114, *FGF14* in patient #129 and *RFC1* in patient #111. In the remaining cases, only weak signals were detected.

In 9 cases, ExpansionHunter Denovo did not detect an expanded allele that had been identified by ExpansionHunter, particularly at *FGF14*, *HTT* and *ATXN1* loci.

At the *HTT* locus, ExpansionHunter identified a CAG-expanded allele of 41 repeats in patient #111, previously confirmed by molecular validation. This allele size is above the pathogenic

threshold for Huntington disease.[32] However, no corresponding signal was detected by ExpansionHunter Denovo. Similarly, at the *ATXN1* locus, the expanded alleles identified by ExpansionHunter in patients #110 and #112, corresponding respectively to 36 and 47 repeats, were not detected (36 repeats) or detected with a relatively low z-score (47 repeats).

These results show that some repeat expansions may not emerge as strong outlier signals in ExpansionHunter Denovo analysis. This may occur when the repeat signal does not clearly separate from the background allelic distribution of the analysed cohort, as in the case of *ATXN1* and *HTT*, for which a broad allelic distribution in the general population is known. Moreover, both loci are located in coding regions and are generally associated with smaller pathogenic expansions compared with many non-coding repeat expansion disorders. These smaller sizes may make detection by ExpansionHunter Denovo more difficult.

In addition, the presence of more than one expanded or intermediate allele at the same locus within the cohort may reduce the relative deviation of the index sample from the background distribution. These findings indicate that ExpansionHunter Denovo outlier analysis may have reduced sensitivity for smaller or intermediate repeat expansions, particularly when larger or comparable repeat alleles are present in the cohort.

At the *TCF4* locus, ExpansionHunter Denovo data were available only for patient #114, in whom a strong outlier signal was detected, supporting partial concordance between the two approaches at this locus. The allele size was experimentally confirmed to be above the pathogenic threshold. *TCF4* alleles exceeding 50 repeats are associated with increased risk of Fuchs endothelial corneal dystrophy; therefore, *TCF4* repeat expansions are mainly reported as a risk factor for this disease, rather than as a direct cause of neurodegenerative disease.[33]

A more complex pattern was observed at the *FGF14* locus. ExpansionHunter Denovo detected outlier signals only in three patients (#96, #129 and #141) out of seven already known to carry expanded allele by ExpansionHunter.

Patient #129 showed the highest score, with a z score of 5.17. Experimental validation showed that this allele was not larger than those observed in the other two patients (131 repeats in patient #141 and 55 repeats in patients #96) but its repeat motif was different, being characterized by a mixed GAA/GGA structure rather than a pure GAA repeat.

At the *RFC1* locus, ExpansionHunter and ExpansionHunter Denovo identified signals in patients #99, #111 and #146. However, in all these patients, the predicted expanded allele was largely below the pathogenic threshold.

Experimental validation by conventional flanking PCR detected only one small allele and therefore did not confirm the presence of an expanded allele.

However, in patient #99 agarose gel electrophoresis of the conventional flanking PCR showed an additional larger band not detectable by capillary electrophoresis fluorescence fragment analysis.

This result does not exclude the presence of a second larger allele which could be not amplifiable by conventional flanking PCR. Additional analyses should include repeat-primed PCR with motif-specific primers, and methods suitable for sizing large alleles, such as Southern blotting or long-read sequencing.

Overall, discordance between ExpansionHunter and ExpansionHunter Denovo involved both repeat-size estimation and signal detection. ExpansionHunter was generally able to identify repeat-size variation at known loci included in the catalogue, but molecular validation showed that it often underestimated the size of longer alleles. Conversely, ExpansionHunter Denovo did not consistently detect previously confirmed expansions as strong outlier signals. The small size of the patient cohort and the lack of WGS data from healthy controls may have contributed to the absence of strong ExpansionHunter Denovo signals for some loci. For known disease-associated loci, ExpansionHunter represents the preferred tool because it is designed for targeted analysis of predefined repeat loci. ExpansionHunter Denovo, by contrast, is not restricted to predefined disease loci and can provide additional information through catalogue-free outlier detection. However, its performance depends strongly on whether the repeat signal is sufficiently distinct from the rest of the analysed cohort. Therefore, integration of both bioinformatic approaches with targeted molecular validation is necessary for reliable interpretation of repeat expansion signals.

This study has some limitations that should be considered when interpreting the results. The first limitation concerns the size and structure of the cohort used for ExpansionHunter Denovo outlier analysis. The analysis was performed on a subset of 52 patients selected from the original cohort of 140 patients because they had been generated using homogeneous sequencing conditions. Although this reduced technical variability, it also limited the number of samples included in the outlier analysis. Since outlier detection depends on the distribution of repeat signals across the analysed cohort, a larger sample size could improve the distinction between true outlier signals and background repeat variability.

A related limitation concerns the use of the outlier-based approach rather than the case-control analysis. Repeat expansion signals were prioritized according to their deviation from the rest of the analysed cohort, rather than according to enrichment in a disease group compared with controls. As a consequence, repeat expansions that do not generate a sufficiently distinct outlier signal may not be detected or prioritized by this approach, as observed for known repeat loci

such as *HTT* and *ATXN1* locus. This limitation is particularly relevant when two or more patients carry expansions of similar size at the same locus, because these signals may not stand out from the cohort distribution even if they reach a pathological threshold.

Another limitation concerns the molecular validation of candidate loci. Although targeted molecular assays allowed experimental validation of selected signals, only a subset of ExpansionHunter Denovo candidate loci was tested. Extending molecular screening to larger numbers of control samples and additional candidate loci would be useful, but it is technically demanding and time-consuming. This is especially true for tandem repeats located in complex genomic regions, where the candidate repeat may be embedded within other repetitive sequences or adjacent tandem repeat tracts, making primer design, amplification and interpretation more difficult.

Future analyses should also integrate molecular findings with detailed clinical and familial information. This would be especially important for candidate repeat loci that are not currently established disease genes, since biological plausibility, segregation with disease and comparison with population controls are all necessary to assess potential pathogenic relevance.

In this context, from a diagnostic workflow perspective, ExpansionHunter Denovo should not be considered a first-line diagnostic tool. Rather, it may have a role as a second-level, discovery-oriented screening approach in genetically unresolved cases, after standard analyses for known disease-associated variants have been performed, including single-nucleotide variants, small insertions/deletions, copy number variants and known repeat expansion loci. Candidate signals identified by ExpansionHunter Denovo require prioritization, comparison with available controls and orthogonal molecular validation to distinguish possible disease-associated repeat expansions from polymorphic or non-confirmed signals.

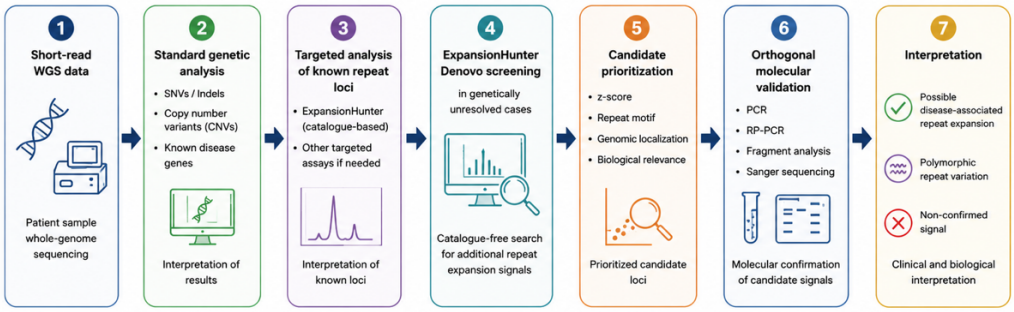

Figure 21. Proposed role of ExpansionHunter Denovo in the diagnostic workflow: ExpansionHunter Denovo may be applied after standard genetic analyses and targeted evaluation of known repeat expansion loci to identify additional candidate repeat expansion signals in genetically unresolved cases.

Overall, this study supports the use of ExpansionHunter Denovo as a complementary catalogue-free tool for identifying candidate repeat expansion signals. However, its results should always be interpreted in combination with targeted molecular validation, clinical information and appropriate control data before any disease-related conclusion is made.

BIBLIOGRAPHY

1. Padeken, J., Zeller, P., and Gasser, S.M. (2015). Repeat DNA in genome organization and stability. Curr. Opin. Genet. Dev. *31*, 12–19. https://doi.org/10.1016/j.gde.2015.03.009.

2. Liu, Y., and Xia, K. (2025). Aberrant Short Tandem Repeats: Pathogenicity, Mechanisms, Detection, and Roles in Neuropsychiatric Disorders. Genes *16*, 406. https://doi.org/10.3390/genes16040406.

3. Tanudisastro, H.A., Deveson, I.W., Dashnow, H., and MacArthur, D.G. (2024). Sequencing and characterizing short tandem repeats in the human genome. Nat. Rev. Genet. *25*, 460–475. https://doi.org/10.1038/s41576-024-00692-3.

4. Liao, X., Zhu, W., Zhou, J., Li, H., Xu, X., Zhang, B., and Gao, X. (2023). Repetitive DNA sequence detection and its role in the human genome. Commun. Biol. *6*, 954. https://doi.org/10.1038/s42003-023-05322-y.

5. Balzano, E., Pelliccia, F., and Giunta, S. (2021). Genome (in)stability at tandem repeats. Semin. Cell Dev. Biol. *113*, 97–112. https://doi.org/10.1016/j.semcdb.2020.10.003.

6. An, Z., Jiang, A., and Chen, J. (2024). Toward understanding the role of genomic repeat elements in neurodegenerative diseases. Neural Regen. Res. *20*, 646–659. https://doi.org/10.4103/NRR.NRR-D-23-01568.

7. Hannan, A.J. (2018). Tandem repeats mediating genetic plasticity in health and disease. Nat. Rev. Genet. *19*, 286–298. https://doi.org/10.1038/nrg.2017.115.

8. 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.

9. Delforge, V., Tard, C., Davion, J.-B., Dujardin, K., Wissocq, A., Dhaenens, C.-M., Mutez, E., and Huin, V. (2024). RFC1: Motifs and phenotypes. Rev. Neurol. (Paris) *180*, 393–409. https://doi.org/10.1016/j.neurol.2024.03.006.

10. Fan, H., and Chu, J.-Y. (2007). A Brief Review of Short Tandem Repeat Mutation. Genomics Proteomics Bioinformatics *5*, 7–14. https://doi.org/10.1016/S1672-0229(07)60009-6.

11. Subramanian, S., Mishra, R.K., and Singh, L. (2003). Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol. *4*, R13. https://doi.org/10.1186/gb-2003-4-2-r13.

12. Willems, T., Gymrek, M., Highnam, G., Mittelman, D., and Erlich, Y. (2014). The landscape of human STR variation. Genome Res. *24*, 1894–1904. https://doi.org/10.1101/gr.177774.114.

13. Malik, I., Kelley, C.P., Wang, E.T., and Todd, P.K. (2021). Molecular mechanisms underlying nucleotide repeat expansion disorders. Nat. Rev. Mol. Cell Biol. *22*, 589–607. https://doi.org/10.1038/s41580-021-00382-6.

14. Hannan, A.J. (2010). Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability.' Trends Genet. *26*, 59–65. https://doi.org/10.1016/j.tig.2009.11.008.

15. Chintalaphani, S.R., Pineda, S.S., Deveson, I.W., and Kumar, K.R. (2021). An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. Acta Neuropathol. Commun. *9*, 98. https://doi.org/10.1186/s40478-021-01201-x.

16. Dugger, B.N., and Dickson, D.W. (2017). Pathology of Neurodegenerative Diseases. Cold Spring Harb. Perspect. Biol. *9*, a028035. https://doi.org/10.1101/cshperspect.a028035.

17. Shi, Q., Dai, M., Ma, Y., Liu, J., Liu, X., and Wang, X.-J. (2024). DRED: A Comprehensive Database of Genes Related to Repeat Expansion Diseases. Genomics Proteomics Bioinformatics *22*, qzae068. https://doi.org/10.1093/gpbjnl/qzae068.

18. van Vugt, J.J.F.A., Zwamborn, R.A.J., Dolzhenko, E., Eberle, M.A., Weisburd, B., Bekema, E., Kooyman, M., Wang, B., Project MinE ALS Sequencing Consortium, Kamsteeg, E.-J., et al. (2025). The role of disease-associated short tandem repeats in amyotrophic lateral sclerosis. Brain Commun. *7*, fcaf482. https://doi.org/10.1093/braincomms/fcaf482.

19. Rafehi, H., Bennett, M.F., and Bahlo, M. (2023). Detection and discovery of repeat expansions in ataxia enabled by next-generation sequencing: present and future. Emerg. Top. Life Sci. *7*, 349–359. https://doi.org/10.1042/ETLS20230018.

20. Zhou, Z.-D., Jankovic, J., Ashizawa, T., and Tan, E.-K. (2022). Neurodegenerative diseases associated with non-coding CGG tandem repeat expansions. Nat. Rev. Neurol. *18*, 145–157. https://doi.org/10.1038/s41582-021-00612-7.

21. Zhu, Y., Xiao, X., Liu, Y., Wang, Z., Luo, T., Xu, T., Yang, Q., Hao, X., Zhang, C., Zhang, S., et al. (2026). Short tandem repeat expansions in patients with neurodegenerative dementia. eBioMedicine *125*. https://doi.org/10.1016/j.ebiom.2026.106190.

22. Leitão, E., Schröder, C., and Depienne, C. (2024). Identification and characterization of repeat expansions in neurological disorders: Methodologies, tools, and strategies. Rev. Neurol. (Paris) *180*, 383–392. https://doi.org/10.1016/j.neurol.2024.03.005.

23. Dolzhenko, E., English, A., Dashnow, H., De Sena Brandine, G., Mokveld, T., Rowell, W.J., Karniski, C., Kronenberg, Z., Danzi, M.C., Cheung, W.A., et al. (2024). Characterization and visualization of tandem repeats at genome scale. Nat. Biotechnol. *42*, 1606–1614. https://doi.org/10.1038/s41587-023-02057-3.

24. Bahlo, M., Bennett, M.F., Degorski, P., Tankard, R.M., Delatycki, M.B., and Lockhart, P.J. (2018). Recent advances in the detection of repeat expansions with short-read next-generation sequencing. F1000Research *7*, F1000 Faculty Rev-736. https://doi.org/10.12688/f1000research.13980.1.

25. 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.

26. Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., 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.

27. Dolzhenko, E., Bennett, M.F., Richmond, P.A., Trost, B., Chen, S., van Vugt, J.J.F.A., Nguyen, C., Narzisi, G., Gainullin, V.G., Gross, A.M., 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.

28. Illumina ExpansionHunter/docs at master · Illumina/ExpansionHunter. GitHub. https://github.com/Illumina/ExpansionHunter/tree/master/docs.

29. Illumina ExpansionHunterDenovo/documentation/images at master · Illumina/ExpansionHunterDenovo. GitHub. https://github.com/Illumina/ExpansionHunterDenovo/tree/master/documentation/images.

30. 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.

31. Halman, A., Lonsdale, A., and Oshlack, A. (2024). Analysis of Tandem Repeats in Short-Read Sequencing Data: From Genotyping Known Pathogenic Repeats to Discovering Novel Expansions. Curr. Protoc. *4*, e70010. https://doi.org/10.1002/cpz1.70010.

32. Brás, I.C., Dawson, J., Kay, C., Caron, N.S., and Hayden, M.R. Huntington Disease.

33. Wieben, E.D., Aleff, R.A., Tosakulwong, N., Butz, M.L., Highsmith, W.E., Edwards, A.O., and Baratz, K.H. (2012). A Common Trinucleotide Repeat Expansion within the Transcription Factor 4 (TCF4, E2-2) Gene Predicts Fuchs Corneal Dystrophy. PLoS ONE *7*, e49083. https://doi.org/10.1371/journal.pone.0049083.