Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond repeat length: from diagnostic validation to population variability of the Huntington disease-associated repeat

Ingrid Lojova
Biomedical Research Center of the Slovak Academy of Sciences
Marcel Kucharik
Biomedical Research Center of the Slovak Academy of Sciences
Andrej Balaz
Biomedical Research Center of the Slovak Academy of Sciences
Andrej Ficek
Comenius University
Diana Rusnakova
Comenius University Science Park
Silvia Bokorova
Comenius University Science Park
Tatiana Sedlackova
Comenius University Science Park
Jaroslav Budis
Comenius University Science Park
Tomas Szemes
Comenius University Science Park
Jan Radvanszky
jradvanszky@gmail.com

Biomedical Research Center of the Slovak Academy of Sciences

Research Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11361091/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. Some of the coauthors are employees or have shares in the Geneton Ltd., GenoVisio Ltd., G2 Consulting Slovakia Ltd. (according to the affiliations of the coauthors), which are biotechnology companies that develop assays and bioinformatic solutions for genetic analyses. TSz reports a relationship with Genovisio Ltd. that includes: equity or stocks.

Beyond repeat length: from diagnostic validation to population variability of the Huntington disease-associated repeat

Ingrid Lojova[1,2], Marcel Kucharik[1,2,4], Andrej Balaz[1,4], Andrej Ficek[3], Diana Rusnakova[2,4], Silvia Bokorova[2,4], Tatiana Sedlackova[2,4], Jaroslav Budis[2,4,5,7], Tomas Szemes[2,3,4], Jan Radvanszky*[1,2,3,6]

[1] Institute of Clinical and Translational Research, Biomedical Research Center of the Slovak Academy of Sciences, Dúbravská cesta 9, 841 04 Karlova Ves, Bratislava, Slovakia

[2] Comenius University Science Park, Ilkovičova 8, 841 04 Karlova Ves, Bratislava, Slovakia

[3] Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Ilkovičova 6, 841 04 Karlova Ves, Bratislava, Slovakia

[4] Geneton Ltd., Ilkovičova 8, 841 04 Karlova Ves, Bratislava, Slovakia

[5] Genovisio Ltd., Ilkovičova 8, 841 04 Karlova Ves, Bratislava, Slovakia

[6] G2 Consulting Slovakia Ltd., Slnečnicová 559/5, 930 41 Hviezdoslavov, Slovakia

[7] Slovak Center of Scientific and Technical Information, Lamačská cesta 8A, 811 04 Bratislava, Slovakia

Corresponding author:

name: Jan Radvanszky, PhD.

address: Institute of Clinical and Translational Research, Biomedical Research Center Slovak Academy of Sciences, Dubravska cesta 9, 845 05, Bratislava, Slovakia

telephone: +421 905 917 111

e-mail: jradvanszky@gmail.com

Abstract

Background: Huntington disease (HD) is associated with a structurally complex and highly variable tandem repeat (TR) motif with a complicated sequence-structure-based genotype-phenotype association. Short-read sequencing may benefit HD research and diagnosis, but several testing considerations should be addressed during implementation. WGS data were analyzed in two datasets: a validation cohort of 50 individuals (with confirmatory analyses using conventional PCR), including five HD patients, and a population cohort of 1018 Slovak individuals. HTT repeat characterization was performed using two TR-specific tools, Dante and ExpansionHunter.

Results: We assessed the accuracy and reliability of HTT complex motif allele-size determination from WGS data with conventional methods. For the clinically relevant ANC motif, concordance reached 98.95%, with all expanded alleles correctly identified. Using TR-specific filtering criteria, we generated phased, nucleotide-resolution descriptions of the complex motif, revealing high population-level allelic variability. Repeat structures at the distal ends of larger expansions were also identified, which may aid clinical interpretation.

Conclusions: Short-read WGS, supported by dedicated TR-genotyping tools, provides a practical and reliable way to characterize HTT repeat structure for use in HD diagnostics and research. Its scalability and resolution also provide a strong foundation for future studies exploring how repeat-structure variability shapes HD phenotypes and related neurological traits.

Keywords: Huntington´s disease, tandem repeats, TR genotyping, nucleotide-resolution

Abbreviations

HD	Huntington's disease
polyQ	polyglutamin
polyP	polyproline
HGVS	Human Genome Variation Society
TREDs	tandem repeat expansion disorders
WGS	whole-genome sequencing
MPS	massively parallel sequencing

1. Background

Huntington's disease (HD; OMIM #143100) is a progressive neurodegenerative disorder primarily affecting the central nervous system causing motor, cognitive, and psychiatric symptoms, but also involving multiple other systems including heart, skeletal muscle, thyroid, liver, and gastrointestinal tract [1]. Symptom onset and severity vary considerably between individuals [2]. HD is caused by a polyglutamine (polyQ) expansion in the HTT protein [3], followed by a polyproline (polyP) stretch. According to the Human Genome Variation Society (HGVS) nomenclature rules, the repeats are described at the DNA level as AGC for polyQ and CGC for polyP (see discussion for nomenclature clarification). This repeat region is relatively complex. In the most common alleles the $(AGC)_n$ tract is followed by an $(AAC)_1(AGC)_1$ structure at its 3′ end, altogether forming the polyQ tract at the protein level. We refer to this part as the $(ANC)_n$ module, to account for possible interruptions. This is followed by a $(CGC)_1(CAC)_1$ structure and by the $(CGC)_n$ tract, with an additional 3' $(CTC)_2$ repeat altogether forming the polyP stretch (Figure 1), for which we will refer as the $(CNC)_n$ module. Notably, both the central 12-bp $(AAC)_1(AGC)_1(CGC)_1(CAC)_1$ sequence and the

downstream $(CTC)_2$ repeat show population-level variability, in addition to variation in the basic (AGC)n and (CGC)n tracts [4].

Physiological alleles of the ANC module contain ≤26 repeats and are stable in intergenerational transmission. Intermediate alleles (27-35 repeats) are clinically silent but genetically unstable. Alleles with 36-40 repeats have reduced penetrance, while expansions of ≥41 repeats are fully penetrant and cause autosomal dominant HD [5]. Expansions >60 repeats are typically associated with juvenile-onset HD [6]. Historically reported genotype-phenotype correlations, based on repeat numbers detected using conventional methods, may be imprecise due to the limited resolution of conventional repeat-sizing methods and variability in sequence interruptions and repeat boundaries. Although the ANC expansion length strongly correlates with disease phenotype, it explains only 60%-70% of the variability in age at disease onset. There is growing evidence that repeat structure, specifically the loss of AAC interruptions within the clinically relevant ANC module of expanded alleles, may further influence inheritance, somatic instability, and disease course as a *cis* modifier, including earlier onset in the reduced-penetrance range (36-39 repeats) [7,8,9]. Therefore, sequence structure knowledge of an individual's allele would enable more individualized discussions of disease risk. However, current low-resolution diagnostic assays may lead to underestimation of CNG repeat length, misdiagnosis, or inaccurate prediction of risk of symptom onset, specifically in the reduced penetrance range [10]. The biological impact of such variability within the range of normal alleles has not yet been established, but cannot be ruled out for HD or other neurological diseases as clinical phenotype modifiers. Additional phenotypic variation of HD is attributed to *trans* genetic modifiers including *MSH3, FAN1, MLH1, PMS2* and environmental factors [11,12].

Conventional HD molecular diagnostics focus on accurate identification of alleles, including the expanded ones, and on their sizing. The most commonly used method is PCR with flanking primers and fragment analysis, but large expansions may amplify incompletely. Triplet-primed PCR (TP-PCR) can complement conventional PCR in expansion detection, although it may not provide repeat counts, especially in larger alleles [5,13,14]. PCR-based methods have several limitations, including susceptibility to sequence interruptions or their loss, which can affect primer binding and fragment analysis, potentially leading to false or misleading results [5,15,16]. In addition, their limited sequence resolution also restricts accurate characterization of repeat structure and its clinical interpretation. Moreover, these methods are primarily suited to targeted single-gene testing, requiring a clear clinical suspicion or family history.

Massively parallel sequencing (MPS) has transformed genetic diagnostics and offers a potential solution to many limitations of conventional methods for genome-wide TR genotyping [17,18,19]. Both short-read and long-read approaches have already been reported, predominantly in research settings [20–23]. Long-read sequencing currently offers the most comprehensive characterization of expanded alleles, while short-read platforms remain more widely available in routine diagnostics. However, implementing MPS for HD requires addressing challenges specific to tandem repeat expansion disorders, and the full diagnostic potential of existing short-read data remains largely unexplored. In line with this, the possible informational yield is also underexploited, for the majority of tandem repeat expansion disorders (TREDs), including HD.

The aim of our study was, therefore, to describe specific aspects of the implementation of whole-genome sequencing into HD diagnostics, using second-generation short-read MPS technologies. We focused on: i) the definition of the complex repeat motif according to a

uniformized nomenclature; ii) the determination of allele identification and sizing accuracy; iii) the distinguishing between genotyping aspects of numerical repeat counting and nucleotide-resolution allele structure characterisation; iv) on the possibilities of resolving complex repeat structures, including allele haplotyping and phasing; and v) on the description of repeat structure variability of the entire HTT complex motif.

2. Methods

2.1. Sample cohort

The study included two independent sample cohorts. The validation cohort consisted of 50 individuals from Slovakia, including five HD patients with a confirmed AGC expansion, and was analysed both using conventional laboratory methods as-well-as using whole-genome sequencing (WGS). The population cohort included 1018 individuals from the Slovak population and was analysed only using WGS. This dataset was formed from the population-control subset of the PreveLynch project (ITMS: 313011V578; https://prevelynch.sk/en/). The population cohort does not consist only of healthy individuals from the population. It represents a control group for participants from families with colorectal cancer and Lynch syndrome, however, participants may suffer from various common civilization-related diseases, such as inflammatory bowel diseases, diabetes, celiac disease, and other diseases. Consistent with the Declaration of Helsinki, signed informed consents were obtained from the analyzed individuals. Ethical committee approvals and descriptions of more detailed methods are listed in the online Supplementary Methods.

2.2. Generation and computational processing of WGS data

In brief, genomic DNA was extracted from peripheral blood and subjected to PCR-free whole-genome sequencing using BGI (BGI Europe Genome Center, Copenhagen, Denmark; DNB-

SEQ PE-150bp, 110 Gb) or Illumina (TruSeq DNA PCR-Free Library kit; Institute of Applied Biotechnologies, Olomouc, Czech Republic, NovaSeq 6000; PE-150bp, 120 Gb) for the validation cohort, and Illumina (Illumina DNA PCR-free Prep kit, Novaseq 6000, PE-150bp) for the population cohort.

For the validation cohort, sequencing reads were aligned to GRCh38 using BWA-MEM v0.7.17 [24]. The HTT repeat region was genotyped using Dante v0.11.0 (available at Dante · bio.tools [25]) and ExpansionHunter v5.0.0 [23]. Both tools used the GRCh38-based HGVS motif notation NC_000004.12:g.3074878_3074973AGC[19]AAC[1]AGC[1]CGC[1]CAC[1]CGC[7]CTC[2], simplified for analysis as AGC[21]CGC[11]. Dante provided numerical (repeat counts) and nucleotide-level genotypes for the individual modules as well as nucleotide-level and phased allele predictions for the whole complex motif. ExpansionHunter provided numerical genotypes for the individual modules. Results were manually reviewed, when necessary, using read-based visualizations using the Dante visualisation interface. Only genotypes concordant with conventional testing were included in the final analysis.

For the population cohort, the *HTT* repeat region was analyzed with Dante v0.14.0 using the same motif definition as in the validation cohort. Due to the large sample size, genotyping relied mainly on automated predictions. To assess reliability, minimum read-support thresholds of three, six, and nine reads per allele were tested, with expanded alleles retained. Manual review was performed only for expanded or unusual allele structures.

2.3. Verification analyses

Verification analyses of the ANC-CNC motif, for the validation cohort, were performed using a conventional PCR and fragment analysis on an ABI 3130 genetic analyzer (Applied

Biosystems). Allele sizes were analyzed with GeneMapper v4.1 and manually reviewed when necessary. Due to differences in *HTT* repeat nomenclature, MPS results were adjusted for comparison with the conventional method: two repeat triplets were subtracted from the ANC module and four from the CNC module to account for the central interrupting sequence.

3. Results

3.1. Determination of the accuracy and reliability of HTT motif allele size determination from WGS data

WGS genotyping accuracy was assessed against conventional methods using numerical genotypes, generated individually for both modules of the HTT-complex motif, and using both Dante and ExpansionHunter.

For the clinically relevant ANC module, numerical allele size predictions were generated and evaluated for 100 alleles. The concordance rate between automatic prediction and the conventional methods was 98.95 % for Dante and 96.84 % for ExpansionHunter (only alleles within the normal range were included, since alleles outside the detection limit are reported differently by the two methods) (Supplementary Table 1). All Dante predictions were accepted after visual inspection, including one classified as “uncertain” (classification was retained because the observed pattern was non-specific and did not provide sufficient evidence for a more definitive allele assignment). Using ExpansionHunter, three genotypes were manually corrected due to nonspecific, likely stutter-related reads.

All five expanded alleles were correctly detected by both tools. Dante predicted the presence of an expansion in four cases, without size estimation, and provided an estimated expansion size in one case (with one repeat difference from conventional methods).

ExpansionHunter estimated all five sizes, with three concordant and two differing by one repeat.

For the CNC module, concordance with conventional methods was 94% for Dante, and 86% for ExpansionHunter, increasing to 99% for Dante and 87% for ExpansionHunter after revision (Supplementary Table 2). Discrepancies were mainly associated with low coverage (1-3 reads per allele) and non-specific spanning reads occurring behind the validated allele size, which were more frequent using ExpansionHunter. They particularly affected genotyping of both modules (Figure 2).

3.2. Population variability of the alleles of the HTT motif

Numerical and nucleotide-resolution characterisation of the motif and its modules was performed using Dante for both cohorts. The quality of the results derived from the smaller, validation dataset (50 individuals), was ensured by the dual sample analysis and subsequent manual results revision. The population dataset (1018 individuals) was analyzed using filters requiring at least ≥3, ≥6, or ≥9 allele-confirming spanning reads for each allele in both modules, respectively. Increasing filter stringency, although increased reliability, substantially reduced the number of evaluable samples and observed allele diversity (Table 1; Supplementary Figure 1).

3.2.1 Variability of the ANC module

When considering only the numerical genotypes, in the validation dataset, four ANC allele sizes (14-31 repeats) were identified, with 19 repeats being most common (Supplementary Figure 2B). In the population dataset, the six read filter identified 24 different alleles ranging from 11-37 repeats (n=614 alleles passing the six-read filter) (Supplementary Figure 1), including six intermediate alleles (29-37 repeats) and one allele corresponding to variable

penetrance or full penetrance allele (≥38 repeats), confirmed by manual verification (Supplementary Figure 1). Using the less stringent three-read filter, 1490 alleles passed the filtering criteria, revealing three additional rare alleles (15, 33 and 38 repeats), and two additional manually confirmed expansion-range alleles (Figure 3). Overall, at least three expansion-range alleles were identified, while three other predicted expansions could not be confirmed by manual review because of not enough supporting reads (Supplementary Figure 3).

Nucleotide resolution genotyping, achievable in a total of 69 ANC alleles from the validation dataset, identified 14 different structures, with $(AGC)_{17}(AAC)_1(AGC)_1$ being the most common (Supplementary Figure 2C). Most alleles followed the typical $(AGC)_n(AAC)_1(AGC)_1$ structure, with two exceptions: one contained an additional AAC interruption, forming $(AGC)_{15}(AAC)_1(AGC)_1(AAC)_1(AGC)_1$, while the second one exhibited an ACC interruption , i.e. $(AGC)_{18}(ACC)_1(AGC)_1$. When evaluating the population dataset at a nucleotide resolution, we relied primarily on the results generated using the filter setting set on six spanning reads supporting each identified allele (Supplementary Figure 1), since this can be considered a conservative filter for this type of analysis. Among the 614 alleles the population analysis did not reveal different patterns of interruptions, when compared to those identified in the validations dataset. The length differences of alleles stemmed primarily in the repeat numbers of the uninterrupted AGC parts of the module.

Although our short-read protocol is not suitable for complete sequencing of expanded alleles, specific visualisation of partial-read alignments enabled reconstruction of the sequence structure at least of both ends of the expanded alleles. The common 3' interruption $(AGC)_n(AAC)_1(AGC)_1$ was identified in all analyzed expanded alleles, including five alleles

in the validation dataset, two in the population dataset, and one 38-repeat reduced-penetrance allele.

3.2.2 Variability of the CNC module

In the validation dataset, based on numerical genotyping, four CNC allele sizes (11, 12, 13, and 14 CNCs) were identified and were exactly recapitulated in the population dataset (Supplementary Figure 2D). No additional alleles suggested by automated analysis (both in filter setting three and six) were confirmed after manual revision, and likely represented genotyping errors.

In the validation dataset, based on nucleotide-resolution genotyping, out of the 83 concordant alleles, five distinct alleles were identified. The most frequently represented allele was $(CGC)_1(CAC)_1(CGC)_7(CTC)_2$, corresponding to the $(CNC)_{11}$ allele from numerical genotyping (Supplementary Figure 2E). The core structure of the CNC motif followed the pattern $(CGC)_1(CAC)_1(CGC)_n(CTC)_2$. However, in one allele, the CAC interruption was absent, presenting as $(CGC)_9(CTC)_2$, while in another one a deviation occurred in the terminal region, with three CTC repeats observed instead of the generally present two, forming the $(CGC)_1(CAC)_1(CGC)_7(CTC)_3$ structure.

The population analysis did not reveal novel repeat structures among the 614 alleles. 22 alleles lacked the CAC interruption, each in the form of $(CGC)_9(CTC)_2$, however, on nine different complex motif backgrounds. The $(CTC)_3$ variant of the terminal section was observed in 16 alleles across three different backgrounds.

3.2.3 Nucleotide resolution population variability of the HTT-complex locus

To resolve the overall variability of the entire HTT-complex motif, nucleotide-resolution and haplotype-phased genotyping of both modules was performed using Dante. In the validation

dataset, haplotype-phased repeat structure was determinable for 64 alleles using Dante, representing 25 distinct phased genotypes. The most frequently occurring alleles were $(AGC)_{17}(AAC)_1(AGC)_1(CGC)_1(CAC)_1(CGC)_7(CTC)_2$ and $(AGC)_{18}(AAC)_1(AGC)_1(CGC)_1(CAC)_1(CGC)_{10}(CTC)_2$ (Supplementary Figure 2A). In the population dataset, using the six-read filter, 47 unique phased allelic structures were identified. The most common alleles were $(AGC)_{17}(AAC)_1(AGC)_1(CGC)_1(CAC)_1(CGC)_7(CTC)_2$, identified 143 times, and the $(AGC)_{17}(AAC)_1(AGC)_1(CGC)_1(CAC)_1(CGC)_{10}(CTC)_2$, identified 95 times (Supplementary Figure 1) (Supplementary Tables 3 and 4).

4. Discussion

MPS now represents a tool that can overcome several limitations of conventional methods, offering not only simple expansion detection, but also more accurate characterization of repeat structures, paving the way for a more thorough understanding of the pathomechanisms of TREDs [15,26]. Additionally, analyzing other sequence variants in TREDs patients may identify genomic factors that modify the effect of repeat expansions [27–29].

Besides research and clinical potential of genomic technologies for HD molecular testing, several factors should be taken into consideration when designing and interpreting MPS-based testing results, particularly in comparison with conventional methods. Genome-scale testing, when compared to single-gene testing, elevates diagnostics of HD to highly automatized level, and even outside of laboratories having expertise specifically on HD diagnostics, creating thus strong pressure to standardize the description and nomenclature of the locus itself and the associated genomic variability. Given the complexity and clinical relevance of the HTT repeat, together with the benefits of nucleotide-resolution genotyping,

its definition should be standardized in both research and routine testing. HGVS recommendations provide the most appropriate framework. Based on the GRCh38 reference genome the basic $(CAG)_n$ module should be extended at its 3' end by the two glutamine triplets, i.e. the $(CAA)_1(CAG)_1$, from the interrupting module. Similarly, the $(CCG)_n$ module should be extended at its 5' end by the alanine-coding triplets, i.e. the $(CCG)_1(CCA)_1$, and at its 3′ end by the CCT motif, which also encodes alanine. The variability found in these marginal parts of modules, both reported in literature [10,30] and also proved by our results, supports the necessity of including them to the description of the complex motif. In line with this their inclusion into the genotyping and evaluation during HD research and diagnostics should take place. According to the 3′ rule of the HGVS recommendations, the clinically relevant repeat module previously denoted as CAG is changing to AGC at the genomic level. Thus, the expandable HD-causing module should be described as NC_000004.12:g.3074878_3074940AGC[19]AAC[1]AGC[1] (GRCh38 allele). This means 21 consecutive glutamines at the protein level with one genomic sequence interruption. Likewise, the whole complex motif of the GRCh38 reference allele should be described as NC_000004.12:g.3074878_3074973AGC[19]AAC[1]AGC[1]CGC[1]CAC[1]CGC[7]CTC[2] at the genomic level, meaning 21 consecutive glutamines followed by 11 prolines at the protein level. Given this complexity, genotype reporting should clearly specify which parts of the motif are included in the repeat count, regardless of whether conventional or MPS-based methods are used. Similarly, counting principles should be clearly defined when reporting repeat-number-based genotype-phenotype correlations.

A key challenge in evaluating the ANC repeat motif is accurately identifying alleles and determining their size from MPS reads. In our study, the concordance rate between MPS and PCR for sizable ANC alleles exceeded that previously reported for the same locus using

ExpansionHunter (94.8/95.5%) [31]. Precise sizing requires sufficient reads spanning both the repeat and flanking sequences. However, unlike in SNVs, different types of sequencing reads must be taken into account and only a subset of them represents an effective coverage depth, and is usable for genotyping and expansion detection [18]. Spanning reads, covering the repeat and both flanking regions, can be allele-confirming or non-specific. Only allele-confirming reads directly support repeat sizing and structure determination, while non-specific reads increase background noise and reduce genotyping accuracy. Partial reads, covering one flanking region and part or all of the repeat motif, mainly support expansion detection, but not sizing. In-repeat reads, located only in the repetitive region are used by some tools, such as ExpansionHunter, to indirect expansion size determination. Because effective coverage is assay-, tool- and locus-specific, fixed minimum-depth thresholds are difficult to define. It is important to consider, however, that the standardly achieved average genome coverage may result in significantly lower effective TR locus coverage. Manual review can improve interpretation of lower-depth data when possible, while large cohorts require stringent allele-specific read filtering and quality-control criteria to minimize false results.

When alleles exceed the sequencing read lengths, reliably detecting alleles larger than the method's size limit remains particularly important [31]. Dante and ExpansionHunter can detect larger alleles from partial reads beyond the sizing limit, similarly to patterns resembling those in RP-PCR when using conventional techniques. In our dataset, both tools reliably detected the presence of expanded alleles. However, Dante primarily indicates their presence without estimating size, whereas ExpansionHunter additionally attempts to infer repeat length from spanning, paired-end, and in-repeat reads. Accurate genotyping of expanded alleles remains more challenging than their detection. Because expanded repeats often exceed the

sequencing read length, repeat sizes must be estimated indirectly and the accuracy of these estimates decreases with increasing expansion size [23,31,32]. This was reflected in our results, where ExpansionHunter correctly identified expansions but agreed with conventional sizing in only two of five cases. Therefore, short-read sequencing tools provide reliable detection of repeat expansions, expansion length estimates for large alleles should be interpreted with caution and confirmed using conventional methods, especially if they are not supported by sufficient spanning reads. Alternatively, long-read sequencing will aid more reliable sizing and detailed characterization of large ANC alleles, however, its routine clinical implementation remains limited [33–36].

Our results support previous studies showing greater sequence-level than repeat-count variability [20,30], both within individual modules and across the whole complex motif (Supplementary Figure 2, Figure 3), supporting the need to consider both repeat length and sequence composition for reliable interpretation. We demonstrate that short-read sequencing data can effectively characterize this variability and provide complex phasing information, despite the motif exceeding the 150 bp read length. However, this relies on the availability tools supporting nucleotide-level high-resolution genotyping and specific read-alignment visualisations, and the reliability of these analyses highly depends on effective read depth. Sequence composition for sizable alleles is effective when sufficient number of error-free spanning reads are available. For expanded alleles, partial read alignments allow characterization of the 5′ and 3′ repeat ends. This can reveal repeat interruptions or sequence variants that may be overlooked by numerical-genotyping or by conventional methods.

Sequence interruptions, however, remain to represent significant technical challenges, since they may affect the results of both conventional as well as MPS based methods, leading either to wrong allele sizing, or allelic dropout, or even to missing sequence changes in the

motif structures [13–16,28,36–38]. Such effects may occur in HD too, and should be further studied.

Beyond its technical impact, the central region may also influence HD phenotype, particularly in reduced penetrance range 36-40 AGC. Loss of the central segment or the CGC-CAC (conventionally CCG-CCA) interruption is associated with earlier motor onset, cognitive impairment, and faster progression [39,40]. Other sequence variants, including loss or duplication of interruptions, have also been reported [39]. In our population cohort, we also identified the CCG[10]CTG[2] configuration in three cases. Neither this nor the previously described configurations has been associated with an earlier age at onset [41]. These findings suggest that HD age of onset may be influenced more by polyglutamine sequence structure than by tract length alone [30,39]. Loss or gain of interruptions, particularly near the reduced-penetrance range, may also affect clinical prediction by respectively underestimating or overestimating uninterrupted repeat length [10], further highlighting the need for nucleotide-resolution genotyping.

Beyond the clinically relevant module, detailed genotyping and haplotyping of the complex motif may be beneficial for several reasons. Firstly, the biological effect of its overall variability remains unclear and warrants large-scale studies to identify its potential modifying effect on HD or on other neurological traits. Secondly, allelic dropout cannot be excluded even using MPS, potentially due to variability of the repeat or flanking regions. Reducing the number of “homozygous” calls may, therefore, help to minimize uncertainty and the need for additional testing. Nucleotide-resolution genotyping of the complex motif may either help to distinguish true homozygotes from apparent homozygotes caused by undetected “null” alleles [42], or by distinguishing samples that appear homozygous by repeat count but are heterozygous at the nucleotide-resolution (Supplementary Figure 3) [10,30,39].

Conclusions

In conclusion, we demonstrate that short-read WGS combined with dedicated TR genotyping and visualization tools may improve HD research and diagnostics, extending testing beyond detection of expansions and enabling comprehensive assessment of TR complexity and other genomic variability. It may improve diagnostic accuracy in cases of unusual repeat structures, ambiguous phenotypes, unclear family medical history, overlapping clinical features, or co-occurring genetic disorders. It enables precise definition of the boundaries of the complex repetitive motif, accurate allele sizing, detection of pathogenic expansions, haplotyping and phasing, as well as nucleotide-level resolution and identification of sequence variability. This approach also offers potential for large-scale association studies investigating the effects of HTT-repeat structure on HD and other neurological traits. However, TR-specific tools and quality metrics need to be further defined, developed, validated and applied in the near future both for research as well as diagnostic applications.

Declarations

Ethics approval and consent to participate

The project complied with all relevant regulations for research involving human participants and was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all individuals for study participation and data publication. Different parts of the project were conducted with the approval of the following ethics committees: Ethical Committee of the Faculty of Natural Sciences, Comenius University in Bratislava, Slovakia (ECH19025), Ethical Committee of the National Institute of Rheumatic Diseases, Piešťany, Slovakia (04196/0029/001/001); Ethical Committee of the Bratislava Self-Governing Region

(EK BSK - 09834/2020/HF); Ethical Committee of Faculty Hospital and Policlinic Žilina (2/0114/24); Ethical Committee of the National Institute of Children's Diseases, Bratislava (EK6/3/2024); Ethical Committee of Faculty Hospital Trenčín; Ethical Committee of the Faculty of Medicine, Comenius University and Faculty Hospital Bratislava (61/2024); and Ethical Committee of the Children's Faculty Hospital in Košice (EK č. 01/2025). Participants provided written informed consent in accordance with the principles of the Declaration of Helsinki.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the online Supplementary tables or are available on request from the corresponding authors. The genetic variant alleles identified in this study are currently undergoing the submission process in the ClinVar database under Submission ID SUB16507332.

Competing interests

Some of the coauthors are employees or have shares in the Geneton Ltd., GenoVisio Ltd., G2 Consulting Slovakia Ltd. (according to the affiliations of the coauthors), which are biotechnology companies that develop assays and bioinformatic solutions for genetic analyses. TSz reports a relationship with Genovisio Ltd. that includes: equity or stocks.

Funding

The research was supported by the Slovak Research and Development Agency under the Grant ID: APVV-24-0284 (GenoMicrosat2); Ministry of Education, Research, Development and Youth of the Slovak Republic within the framework of the national research infrastructure platform ESFRI–ELIXIR; Slovak Grant Agency of Science under the Grant IDs: VEGA-2/0114/24 and VEGA 2/0179/26.

Authors' contributions

IL: Writing - original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. MK: Visualization, Software, Investigation. AB: Software, Visualization. AF: Methodology, Supervision. DR: Methodology, SB: Methodology, TS: Methodology, Supervision JB: Supervision, Software, Funding acquisition. TSz: Supervision, Funding acquisition. JR: Writing - original draft, Supervision, Methodology, Funding acquisition.

References

1. Ross CA, Aylward EH, Wild EJ, Langbehn DR, Long JD, Warner JH, et al. Huntington disease: natural history, biomarkers and prospects for therapeutics. Nat Rev Neurol. 2014;10:204–16.

2. Wright GEB, Collins JA, Kay C, McDonald C, Dolzhenko E, Xia Q, et al. Length of Uninterrupted CAG, Independent of Polyglutamine Size, Results in Increased Somatic Instability, Hastening Onset of Huntington Disease. Am J Hum Genet. 2019;104:1116–26.

3. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell. 1993;72:971–83.

4. Goldberg YP, McMurray CT, Zeisler J, Almqvist E, Sillence D, Richards F, et al. Increased instability of intermediate alleles in families with sporadic Huntington disease compared to similar sized intermediate alleles in the general population. Hum Mol Genet. 1995;4:1911–8.

5. Bean L, Bayrak-Toydemir P. American College of Medical Genetics and Genomics Standards and Guidelines for Clinical Genetics Laboratories, 2014 edition: technical standards and guidelines for Huntington disease. Genet Med. 2014;16:e2.

6. Fusilli C, Migliore S, Mazza T, Consoli F, De Luca A, Barbagallo G, et al. Biological and clinical manifestations of juvenile Huntington's disease: a retrospective analysis. Lancet Neurol. 2018;17:986–93.

7. Findlay Black H, Wright GEB, Collins JA, Caron N, Kay C, Xia Q, et al. Frequency of the loss of CAA interruption in the HTT CAG tract and implications for Huntington disease in the reduced penetrance range. Genet Med. 2020;22:2108–13.

8. Stevanovski I, Chintalaphani SR, Gamaarachchi H, Ferguson JM, Pineda SS, Scriba CK, et al. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. Sci Adv. 2022;8:eabm5386.

9. Rajan-Babu I-S, Dolzhenko E, Eberle MA, Friedman JM. Sequence composition changes in short tandem repeats: heterogeneity, detection, mechanisms and clinical implications. Nat Rev Genet. 2024;25:476–99.

10. Findlay Black H, Levesley J, Kay C, Bortnick S, Javier K, Hayden MR. Clinical implications of loss of interruption variants for diagnosis, genetic counselling, and clinical trials in Huntington's disease. J Huntingtons Dis. 2026;15:516–23.

11. Mo C, Hannan AJ, Renoir T. Environmental factors as modulators of neurodegeneration: insights from gene-environment interactions in Huntington's disease. Neurosci Biobehav Rev. 2015;52:178–92.

12. Gusella JF, MacDonald ME, Lee J-M. Genetic modifiers of Huntington's disease. Mov Disord. 2014;29:1359–65.

13. Warner JP, Barron LH, Goudie D, Kelly K, Dow D, Fitzpatrick DR, et al. A general method for the detection of large CAG repeat expansions by fluorescent PCR. J Med Genet. 1996;33:1022–6.

14. Falk M, Vojtísková M, Lukás Z, Kroupová I, Froster U. Simple procedure for automatic detection of unstable alleles in the myotonic dystrophy and Huntington's disease loci. Genet Test. 2006;10:85–97.

15. Lojova I, Kucharik M, Pös Z, Balaz A, Zatkova A, Tothova Tarova E, et al. Advancing molecular diagnostics of myotonic dystrophy type 1 using short-read whole genome sequencing. Mol Cell Probes. 2025;79:102005.

16. Radvansky J, Ficek A, Minarik G, Palffy R, Kadasi L. Effect of unexpected sequence interruptions to conventional PCR and repeat primed PCR in myotonic dystrophy type 1 testing. Diagn Mol Pathol. 2011;20:48–51.

17. Dongare DB, Nishad SS, Mastoli SY, Saraf SA, Srivastava N, Dey A. High-throughput sequencing: a breakthrough in molecular diagnosis for precision medicine. Funct Integr Genomics. 2025;25:22.

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

19. Tanudisastro HA, Deveson IW, Dashnow H, MacArthur DG. Sequencing and characterizing short tandem repeats in the human genome. Nat Rev Genet. 2024;25:460–75.

20. Ciosi M, Cumming SA, Chatzi A, Larson E, Tottey W, Lomeikaite V, et al. Approaches to Sequence the HTT CAG Repeat Expansion and Quantify Repeat Length Variation. J Huntingtons Dis. 2021;10:53–74.

21. Fang L, Liu Q, Monteys AM, Gonzalez-Alegre P, Davidson BL, Wang K. DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. Genome Biol. 2022;23:108.

22. Miyatake S, Koshimizu E, Fujita A, Doi H, Okubo M, Wada T, et al. Rapid and comprehensive diagnostic method for repeat expansion diseases using nanopore sequencing. NPJ Genom Med. 2022;7:62.

23. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics. 2019;35:4754–6.

24. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

25. Budiš J, Kucharík M, Ďuriš F, Gazdarica J, Zrubcová M, Ficek A, et al. Dante: genotyping of known complex and expanded short tandem repeats. Bioinformatics. 2019;35:1310–7.

26. Lojova I, Kucharik M, Zatkova A, Balaz A, Pös Z, Tarova ET, et al. High-resolution repeat structure analysis in myotonic dystrophy type 2 diagnostics using short-read whole genome sequencing. Anal Biochem. 2025;700:115793.

27. Flower M, Lomeikaite V, Ciosi M, Cumming S, Morales F, Lo K, et al. MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. Brain. 2019;142:1876–86.

28. Wright GEB, Black HF, Collins JA, Gall-Duncan T, Caron NS, Pearson CE, et al. Interrupting sequence variants and age of onset in Huntington's disease: clinical implications and emerging therapies. Lancet Neurol. 2020;19:930–9.

29. Tomé S, Manley K, Simard JP, Clark GW, Slean MM, Swami M, et al. MSH3 polymorphisms and protein levels affect CAG repeat instability in Huntington's disease mice. PLoS Genet. 2013;9:e1003280.

30. Heinzmann A, Petit E, Dawson J, Kay C, Davoine C-S, Méreaux J-L, et al. Sequence Variants in Small CAG Repeat Expansions of the Gene and Disease Onset and Progression in Huntington Disease. Neurology. 2026;106:e214404.

31. Ibañez K, Polke J, Hagelstrom RT, Dolzhenko E, Pasko D, Thomas ERA, et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. Lancet Neurol. 2022;21:234–45.

32. Halman A, Dolzhenko E, Oshlack A. STRipy: A graphical application for enhanced genotyping of pathogenic short tandem repeats in sequencing data. Hum Mutat. 2022;43:859–68.

33. Chiu R, Rajan-Babu I-S, Friedman JM, Birol I. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. Genome Biol. 2021;22:224.

34. Giesselmann P, Brändl B, Raimondeau E, Bowen R, Rohrandt C, Tandon R, et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. Nat Biotechnol. 2019;37:1478–81.

35. Dolzhenko E, English A, Dashnow H, De Sena Brandine G, Mokveld T, Rowell WJ, et al. Characterization and visualization of tandem repeats at genome scale. Nat Biotechnol. 2024;42:1606–14.

36. Sitarčík J, Vinař T, Brejová B, Krampl W, Budiš J, Radvánszky J, et al. WarpSTR: determining tandem repeat lengths using raw nanopore signals. Bioinformatics [Internet]. 2023;39. https://doi.org/10.1093/bioinformatics/btad388

37. Alfano M, De Antoni L, Centofanti F, Visconti VV, Maestri S, Degli Esposti C, et al. Characterization of full-length expanded alleles in myotonic dystrophy type 2 patients by Cas9-mediated enrichment and nanopore sequencing. Elife [Internet]. 2022;11. https://doi.org/10.7554/eLife.80229

38. Radvansky J, Ficek A, Kadasi L. Repeat-primed polymerase chain reaction in myotonic dystrophy type 2 testing. Genet Test Mol Biomarkers. 2011;15:133–6.

39. Dawson J, Kay C, Black HF, Bortnick S, Javier K, Xia Q, et al. The frequency and clinical impact of synonymous HTT loss-of-interruption and duplication-of-interruption variants in a diverse HD cohort. Genet Med. 2024;26:101239.

40. Dawson J, Baine-Savanhu FK, Ciosi M, Maxwell A, Monckton DG, Krause A. A probable -acting genetic modifier of Huntington disease frequent in individuals with African ancestry. HGG Adv. 2022;3:100130.

41. Arning L, Saft C, Nguyen HP. Challenges facing genetic diagnostics of Huntington's disease: an update. Expert Rev Mol Diagn. 2026;1–7.

42. Williams LC, Hegde MR, Nagappan R, Faull RL, Giles J, Winship I, et al. Null alleles at the Huntington disease locus: implications for diagnostics and CAG repeat instability. Genet Test. 2000;4:55–60.

43. Dawson J, Baine-Savanhu FK, Ciosi M, Maxwell A, Monckton DG, Krause A. A probable -acting genetic modifier of Huntington disease frequent in individuals with African ancestry. HGG Adv. 2022;3:100130.

Figures

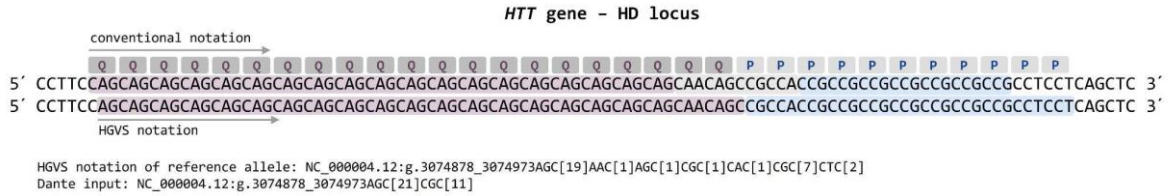


Figure 1: Schematic representation of the structure of the HTT complex locus, including both the conventional notation as well as the HGVS-based notation.

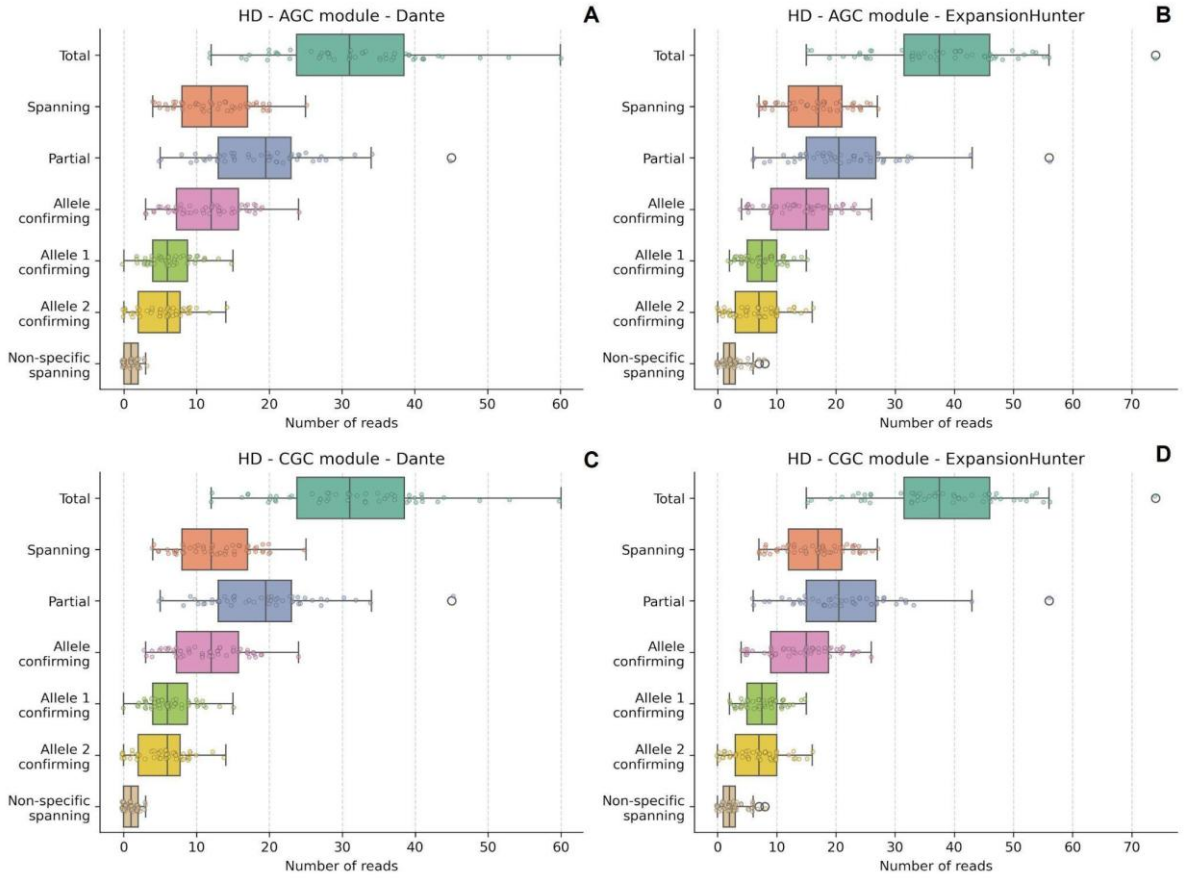


Figure 2: Number of different read types covering the A) ANC module and C) CNC module using Dante tool; and B) ANC, D) CNC modules using ExpansionHunter.

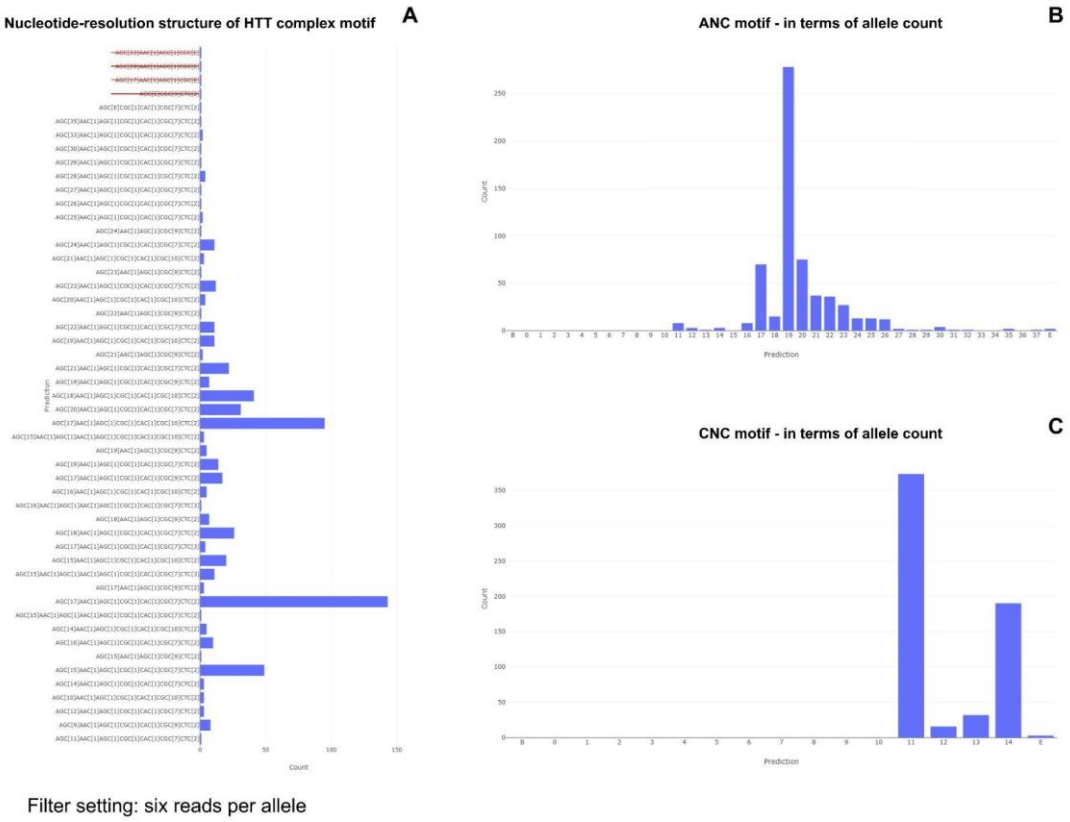


Figure 3: Distribution of ANC-CNC alleles at the nucleotide-resolution level (A); distribution of ANC (B) and CNC[39,43](C) alleles at the numerical level using a filter of six reads per allele. The filtering settings allowed alleles classified as expansions (marked as "E") to be retained. Strikethrough indicates predicted expanded alleles which were not proved by visual inspection.

Tables

Table 1: Numbers of identified ANC alleles in population cohort using the individual filters. * = genotype/phenotype correlation categories shown are defined both in the generally reported format, excluding the canonical interrupting $(AAC)_1(AGC)_1$ dual-repeat, as well as in the format considering the numerical genotype of the whole ANC motif.

Filter	Number of evaluated alleles	Number of unique alleles	27-35 (29-37)*	≥36 or "E" (≥38 or "E")*
Without applied filter	2036	257	16	4
3 reads per allele	1490	124	11	3
6 reads per allele	614	51	6	1
9 reads per allele	132	29	1	0

Supplementary Figure legends

Supplementary Figure 1: Distribution of ANC alleles at the count level and at nucleotide resolution using three minimum read-depth filters: three, six, and nine reads per allele.

Supplementary Figure 2: Numerical and nucleotide-resolution allele distributions in validation dataset. A) Distribution patterns of identified alleles of the complex HTT motif (n=64); B) Allele distribution of ANC repetitive motif (excluding expanded alleles) (n=94); C) Distribution patterns of the identified sizable alleles in our sample set of the ANC module with a full sequence resolution derived from manually revised WGS results (n=69); D) Allele distribution of CNC repetitive motif (n=94). Only the alleles that matched between the reference method in validation set and the genotypes automatically predicted by the Dante tool were included in the allele size distribution analysis. We selected Dante because it

showed a higher concordance rate than Expansion Hunter; E) Distribution patterns of the identified sizable alleles in our sample set of the CNC module with a full sequence resolution derived from manually revised WGS results (n=83). The figures are based on those alleles, which were, on the level of the simplified ANC/CNC numbers, genotyped concordantly using the WGS and conventional methods, and which were supported by at least three sequencing reads.

Supplementary Figure 3: Results of fragment analysis and Dante analysis. A) Fragment analysis indicates a homozygous genotype for the GCA (HGVS notation: AGC) motif, with a 17/17 repeat configuration. B) A heterozygous 8/10 genotype was identified for the CCG (HGVS notation: CGC) motif. C) At the nucleotide sequence level, the AGC motif is present in a heterozygous configuration. The second allele contains one additional AAC interruption compared with the first allele and, additionally, one extra repeat within the terminal CTC region.

Supplementary Table legends

Supplementary Table 1: Determined allele sizes for AGC module via conventional methods and WGS with relevant attributes (read counts, confidence)

Supplementary Table 2: Determined allele sizes for CGC module via conventional methods and WGS with relevant attributes (read counts, confidence)

Supplementary Table 3: Allele structures using filter 6 reads per allele

Supplementary Table 4: Allele structures using filter 9 reads per allele

Supplementary Material and Methods

2.1. Sample cohort

The study included two independent sample cohorts. The validation cohort consisted of 50 individuals from Slovakia, including five HD patients with a confirmed AGC expansion, and was analysed both using conventional laboratory methods as well as using whole-genome sequencing (WGS). The population cohort included 1018 individuals from the Slovak population and was analysed only using WGS. This dataset was formed from the population-control subset of the PreveLynch project (ITMS: 313011V578; https://prevelynch.sk/en/). The population cohort does not consist only of healthy individuals from the population. It represents a control group for participants from families with colorectal cancer and Lynch syndrome, however, participants may suffer from various common civilization-related diseases, such as inflammatory bowel diseases, diabetes, celiac disease, and other diseases.

The presented results form part of a long-term research project aimed at the improvement of genomic technologies and their validation, and at the genomic characterization of the Slovak population. The project complied with all relevant regulations for research involving human participants and was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all individuals for study participation and data publication. Different parts of the project were conducted with the approval of the following ethics committees: Ethical Committee of the Faculty of Natural Sciences, Comenius University in Bratislava, Slovakia (ECH19025), Ethical Committee of the National Institute of Rheumatic Diseases, Piešťany, Slovakia (04196/0029/001/001); Ethical Committee of the Bratislava Self-Governing Region (EK BSK – 09834/2020/HF); Ethical Committee of Faculty Hospital and Polyclinic Žilina (2/0114/24); Ethical Committee of the National

Institute of Children's Diseases, Bratislava (EK6/3/2024); Ethical Committee of Faculty Hospital Trenčín; Ethical Committee of the Faculty of Medicine, Comenius University and Faculty Hospital Bratislava (61/2024); and Ethical Committee of the Children's Faculty Hospital in Košice (EK č. 01/2025). Participants provided written informed consent in accordance with the principles of the Declaration of Helsinki.

Genomic DNA for analyses was extracted from peripheral blood using a standard protocol (DNA Blood Mini Kit, QIAGEN).

2.2. Genomic data generation

DNA samples were subsequently prepared according to specifications for PCR-free whole-genome sequencing. Sequencing was performed using second-generation short-read platforms: BGI (BGI Europe Genome Center, Copenhagen, Denmark; DNB-SEQ PE-150bp, 110 Gb) or Illumina (TruSeq DNA PCR-Free Library kit; Institute of Applied Biotechnologies, Olomouc, Czech Republic, NovaSeq 6000; PE-150bp, 120 Gb), for the validation cohort, and Illumina (Illumina DNA PCR-free Prep kit, Novaseq 6000, PE-150bp) for the population cohort.

2.3. Computational Processing of WGS data

For the validation cohort, sequencing reads were aligned to the human reference genome GRCh38 using BWA-MEM version 0.7.17 to create the respective BAM files 25. Numerical and nucleotide-resolution genotyping of both the individual modules, as well as phasing of the complex motif, was performed using the genotyping and visualization tool Dante (v0.11.0) (available at Dante · bio.tools) 26. In parallel, numerical TR genotyping of the individual modules was performed using ExpansionHunter (v5.0.0) 24. Both tools used the same BAM files as input, together with a TR motif definition recorded at the genomic level. The motif

definition is derived from the GRCh38 reference allele of the HD-associated HTT complex, that is NC_000004.12:g.3074878_3074973AGC[19]AAC[1]AGC[1]CGC[1]CAC[1]CGC[7]CTC[2], when using the respective HGVS guidelines (for more details see the respective parts of Introduction and Discussion). However, for the purposes of generating a simplified representation compatible with the bioinformatic tools used in our study, as a suitable input format prepared for the possible presence of other unexpected repeat structures, the 12-bp interrupting sequence (AAC)1(AGC)1(CGC)1(CAC)1, as well as the 3′ (CTC)2, were arbitrarily assigned to one of the two principal repeat modules, i.e. to the predominant repeats, based on their contribution to the encoded protein domains: NC_000004.12:g.3074878_3074973AGC[21]CGC[11]. Triplets contributing to the polyglutamine-encoding portion of the region were attributed to the ANC module, whereas those contributing to the polyproline-encoding portion were assigned to the CNC module (Figure 1). It is important to note that, however, both tools accept the simplified nomenclature as an input, besides the numerical genotypes generated by both tested tools as outputs, Dante also generates nucleotide-resolution genotypes. When considering the genotype outputs in the numerical format, the interrupting sequences are counted to their respective predominant sequence. When considering the nucleotide-resolution output, the mentioned sequence interruptions are redistributed to their correct positions in the resulting genotype determinations. Similar to general recommendations for SNVs and small indels 27, and in line with other TR publications 28, TR motif genotyping results were inspected through multiple visualization outputs of the graphical report of Dante. For both Dante and ExpansionHunter-derived numerical genotypes, histograms of spanning (covering the entire repetitive motif and both flanking sequences), partial reads (covering one flanking sequence

and the repetitive motif or part thereof) and in-repeat reads (in case of ExpansionHunter) were used to assess repeat-length distributions.

All sequence-resolved visualizations, including predicted sequence structures of modules and phasing results, were generated from Dante outputs. These included specialized visualizations of individual sequence reads and sequence logos, enabling visual inspection at nucleotide resolution. These graphical functionalities are available for all reads in a sample together, but also separated for reads supporting individual alleles, and also separately for spanning reads and partial reads, for more in depth visual inspection of results. Where necessary, manual adjustment of the automated prediction was performed and such revision was recorded. Allele distributions on numerical level from this dataset were evaluated and reported only from those alleles, which were found to be concordant between the conventional and WGS/Dante data. When determining nucleotide-resolution structures, the evaluation focused on alleles which were numerically concordantly genotyped using the conventional method and the WGS/Dante approach, and which were proved by at least two spanning reads.

For the population cohort, sequence data were processed using Illumina DRAGEN software version 4.4 to create the BAM files. Numerical and nucleotide-resolution genotyping of both the individual HTT modules, as well as phasing of the complex motif, was performed using the genotyping and visualization tool Dante (v0.14.0). The same motif definition was used as in the validation cohort. For the population cohort, manual revision was not performed due to the large number of samples. Therefore, the results rely on automated predictions of the Dante tool. To increase confidence in the automated genotyping results, three different filtering settings were applied and evaluated, based on the minimum number of reads supporting each identified allele. Specifically, minimum read-support thresholds of three, six,

and nine reads per allele were tested. In addition, the filtering settings allowed alleles classified as expansions (marked as “E”) to be retained. Manual revision was performed only in cases of specific results, such as when expanded alleles or highly unusual allele structures were detected.

2.4. Verification analyses

Verification analyses of the ANC-CNC motif, for the validation cohort, were performed using a conventional PCR. Reaction designs for individual loci were based on previously published protocols (primer sequences available upon request). Amplified fragments were separated and detected on an ABI 3130 genetic analyzer (Applied Biosystems) using 36cm capillary, POP7 polymer (Applied Biosystems), and the Dye Set G5 filter. The resulting data were analyzed and visualized using GeneMapper 4.1 software (Applied Biosystems). Allele sizes assigned by the software were manually reviewed and, if necessary, adjusted to correct the automated predictions. We adapted the interpretation of conventional method results due to the modified nomenclature of HTT locus. This reference method is based on the conventional HTT repeat description and treats the (AAC)1(AGC)1(CGC)1(CAC)1 sequence at the border of the modules as an invariant interrupting sequence, where one of the ANC specific primers are generally designed to bind. Technically, the comparison of MPS results with the results of the conventional methods was performed, therefore, as follows: for the first ANC module two repeat triplets immediately following the clinically relevant ANC region were subtracted from the results obtained by MPS. For the CNC module of the complex motif, four repetitive units were subtracted, i.e. two located before the main CGC motif and two immediately following it.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigure1.pdf
- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx