Modification of Huntington's disease by short tandem repeats

Eun Pyo Hong,[1,2,3,*] Eliana Marisa Ramos,[1,2,*] N. Ahmad Aziz,[4,5,*] Thomas H. Massey,[6] Branduff McAllister,[6] Sergey Lobanov,[6] Lesley Jones,[6] Peter Holmans,[6] Seung Kwak,[7] Michael Orth,[8] Marc Ciosi,[9] Vilija Lomeikaite,[9] Darren G. Monckton,[9] Jeffrey D. Long,[10] Diane Lucente,[1] Vanessa C. Wheeler,[1,2] Tammy Gillis,[1] Marcy E. MacDonald,[1,2,3] Jorge Sequeiros,[11,12] James F. Gusella,[1,3,13] and Jong-Min Lee[1,2,3]

[*] **Eun Pyo Hong, Eliana Marisa Ramos, and N. Ahmad Aziz contributed equally to this work.**

Abstract

Expansions of glutamine-coding CAG trinucleotide repeats cause a number of neurodegenerative diseases, including Huntington's disease (HD) and several of the spinocerebellar ataxias (SCAs). In general, age-at-onset of the polyglutamine diseases is inversely correlated with the size of the respective inherited expanded CAG repeat. Expanded CAG repeats are also somatically unstable in certain tissues, and age-at-onset of HD corrected for individual *HTT* CAG repeat length (i.e., residual age-at-onset), is modified by repeat instability-related DNA maintenance/repair genes as demonstrated by recent genome-wide association studies (GWAS). Modification of one polyglutamine disease (e.g., HD) by the repeat length of another (e.g., *ATXN3,* CAG expansions in which cause SCA3) has also been hypothesized. Consequently, we determined whether age-at-onset in HD is modified by the CAG repeats of other polyglutamine disease genes. We found that the CAG measured repeat sizes of other polyglutamine disease genes were polymorphic in HD participants but did not influence HD age-at-onset. Additional analysis focusing specifically on *ATXN3* in a larger sample set (*n*=1,388) confirmed the lack of association between HD residual age-at-

onset and *ATXN3* CAG repeat length. Additionally, neither our HD onset modifier GWAS single nucleotide polymorphism (SNP) data nor imputed short tandem repeat (STR) data supported involvement of other polyglutamine disease genes in modifying HD. By contrast, our GWAS based on imputed STRs revealed significant modification signals for other genomic regions. Together, our STR GWAS show that modification of HD is associated with STRs that do not involve other polyglutamine disease-causing genes, refining the landscape of HD modification and highlighting the importance of rigorous data analysis, especially in genetic studies testing candidate modifiers.

Keywords: Huntington's disease; genetic modification; polyglutamine disease; *ATXN3*; short tandem repeat

Author affiliations:

[1] Molecular Neurogenetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA

[2] Department of Neurology, Harvard Medical School, Boston, MA 02115, USA

[3] Medical and Population Genetics Program, the Broad Institute of M.I.T. and Harvard, Cambridge, MA 02142, USA

[4] German Center for Neurodegenerative Diseases, Venusberg-Campus 1/99, 53127 Bonn, Germany

[5] Department of Neurology, Faculty of Medicine, University of Bonn, Bonn, Germany

[6] Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff CF24 4HQ, UK

[7] CHDI Foundation, Princeton, NJ 08540, USA

[8] University Hospital of Old Age Psychiatry and Psychotherapy, Bern University, Bern, Switzerland

[9] School of Molecular Biosciences, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK

[10] Department of Psychiatry, Carver College of Medicine and Department of Biostatistics, College of Public Health, University of Iowa, Iowa City, Iowa 52242, USA

[11] UnIGENe, IBMC - Institute for Molecular and Cell Biology, i3S - Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal

[12] ICBAS School of Medicine and Biomedical Sciences, Univ. Porto, Portugal

[13] Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA 02115, USA

Correspondence to: Jong-Min Lee, Ph.D.

Center for Genomic Medicine, Massachusetts General Hospital

185 Cambridge Street, Boston, MA 02114, USA

E-mail: jlee51@mgh.harvard.edu

Short title: Modification of HD by short tandem repeat

Introduction

Expansions of glutamine-encoding CAG repeats cause at least nine neurodegenerative diseases, including Huntington's disease (HD; MIM #143100), several spinocerebellar ataxias (SCAs), and dentatorubral-pallidoluysian atrophy (DRPLA).[1-3] The polyglutamine expansion diseases exhibit differences in pathogenesis, susceptible brain regions, and disease symptoms.[4,5] However, they share a common feature of inverse correlation between age-at-onset and the length of the causative expanded CAG repeat,[5-12] indicating that increases in CAG repeat size result in accelerated pathogenesis. However, variance in age-at-onset is not fully explained by the glutamine-encoding CAG repeat length alone. For example, the residual variance in HD age-at-onset (i.e., not due to CAG repeat size) showed heritability,[13] prompting genome-wide association studies (GWAS) to identify genetic modifiers of HD. Several genetic loci discovered to influence HD age-at-onset harbor DNA repair genes, such as *MLH1, MSH3,* and *FAN1*.[14,15] These genes have been associated with somatic instability of *HTT* CAG repeats in humans and model systems.[16-25] There is also evidence that these and other DNA repair genes may influence somatic CAG repeat expansions and impact other repeat expansion disorders.[16-19,21,26-39] The striatum, which is severely affected in HD, shows the highest levels of somatic *HTT* CAG repeat expansion; however, expanded repeats in other polyglutamine diseases can also undergo CAG expansion in this brain region.[40-44] Interestingly, candidate gene studies have reported modification of HD by normal CAG repeats in *ATXN3* (expansions in which are responsible for SCA3),[45] and conversely modification of SCA3 by the normal *HTT* CAG repeat.[46] Although *HTT* and *ATXN3* have potential roles in the DNA damage response,[37,47] neither *HTT* nor *ATXN3* are known to be directly involved in DNA repair. Therefore, this mutual modification has suggested the possibility of a novel mechanism underlying polyglutamine diseases. Consequently, using a

variety of genomic data, we set out to determine whether HD is modified by CAG repeat length in other polyglutamine disease genes or other short tandem repeats (STR).

Materials and methods

Study subjects

To identify genetic modifiers of HD motor onset, a total of 9,058 HD subjects (carrying inherited CAG 40 to 55) of European ancestry were previously analyzed in our GWAS.[15] Among those HD subjects, we analyzed participants of the COHORT study (https://clinicaltrials.gov/ct2/show/NCT00313495) to test association between HD age-at-onset and experimentally determined (i.e., genotyped) CAG repeat lengths of other polyglutamine disease-causing genes (n=606). For *ATXN3*-focused analysis, we also analyzed participants of the REGISTRY study (https://clinicaltrials.gov/ct2/show/NCT01590589) (n=885). Details of study approval, genotyping, determination of CAG repeat size, and calculation of residual age-at-onset are described elsewhere.[15]

Determination of the CAG repeats in the COHORT and REGISTRY samples

We determined the sizes of CAG repeats of *ATN1* [dentatorubral-pallidoluysian atrophy; DRPLA MIM #125370], *ATXN1* [spinocerebellar ataxia type 1; SCA1 MIM #164400], *ATXN2* [spinocerebellar ataxia type 2; SCA2 MIM #183090], *ATXN3* [Machado-Joseph disease aka spinocerebellar ataxia type 3; SCA3 MIM #109150], *CACNA1A* [spinocerebellar

ataxia type 6; SCA6 MIM #183086], and *TBP* [spinocerebellar ataxia type 17; SCA17 MIM #607136] in the participants of the COHORT study (*n*=606). After quality control, 551, 502, 604, 503, 497 and 483 COHORT samples were analyzed for *ATN1*, *ATXN1*, *ATXN2*, *ATXN3*, *CACNA1A*, and *TBP*, respectively. In addition, REGISTRY samples were analyzed to determine the CAG repeat sizes of *ATXN3* (*n*=885). CAG repeat lengths for each polyglutamine disease-causing genes were determined by polymerase chain reaction (PCR) assays, using fluorescently labelled primers with minor modifications.[48] PCR products were resolved by an ABI PRISM 3730XL automated DNA Sequencer (Applied Biosystems) and analyzed using GeneMapper version 3.7 software. A set of genomic DNA standard samples were also sequenced for each polyglutamine disease-causing repeat and used as references of CAG allele sizes. Expansions above the non-disease associated repeat range were sequenced after gel separation to further confirm the number of CAGs and the presence of CAA/CAT interruptions.

Analysis to determine the modification of HD by the CAG repeats of other polyglutamine disease-causing genes

Residual age-at-onset of HD, representing age-at-onset corrected for individual *HTT* CAG repeat length, was based on the rater's estimation of onset age of motor symptoms and the uninterrupted *HTT* CAG repeat size.[12,15] For example, an HD subject with a positive residual age-at-onset of 5 means that the individual developed motor onset 5 years later than expected based on his or her uninterrupted CAG repeat size. To determine whether CAG repeat sizes of other polyglutamine disease-causing genes modify HD age-at-onset, we used the same residual age-at-onset phenotype used in our GWAS to identify onset modifiers of HD.[15,49] Briefly, we modeled residual age-at-onset of HD as a function of the CAG repeat of another

polyglutamine disease gene with a set of covariates including 4 genotype-based principal components, sex, and study group in a linear regression analysis. To analyze typed CAG repeats of each polyglutamine disease gene, 3 separate linear regression models were constructed to test the longer, the shorter, and the sum of both repeat lengths. To validate previous dichotomous *ATXN3* CAG repeat association analysis,[45] we also performed a Mann-Whitney *U* test to compare age-at-onset and residual age-at-onset between HD subjects carrying below vs. above the median of longer *ATXN3* CAG repeat (i.e., 23 CAG). Specifically, based on the longer of two *ATXN3* CAG repeats in a given individual, 1) HD subjects carrying 22 or shorter repeats were assigned as below the median group, and 2) HD subjects with 24 or higher repeats were assigned as above the median group for the dichotomous analysis.

SNP association analysis for other polyglutamine disease-causing genes in HD modifier GWAS data

We evaluated the levels of association between residual age-at-onset of HD and other polyglutamine disease-causing genes by checking SNP data available at the GeM Euro 9K website (cegeme.partners.org/gem.euro.9k).[49] For each gene, we took a region for the RefSeq select transcript, and identified the SNP with the highest significance in our HD subject-based GWA analysis to assess the levels of modification of HD by other polyglutamine disease genes.

Imputation of STR lengths from GWAS SNP data and association analysis

For the imputation of STR lengths in our GWAS data, we performed quality control analysis of typed genotype data by taking SNPs with call rate > 95% and minor allele frequency > 1%. Allele frequencies of SNPs and reference alleles in the typed data set were compared to those of 1000 Genomes Project data to confirm data quality using the conform-gt (https://faculty.washington.edu/browning/conform-gt.html). Then, imputation of autosomal STRs was performed by the Beagle program (https://faculty.washington.edu/browning/beagle/beagle.html; v4.1) using the 1000 Genomes Project reference panel consisting of SNPs and STRs.[50,51] Imputed STR data were further filtered by taking tandem repeats located by the 'Tandem Repeats Finder' algorithm[52] and annotated as "SimpleRepeat" in the UCSC genome browser (https://genome.ucsc.edu/index.html). These procedures generated repeat length genotypes of 66,154 tandem repeats for the 9,058 HD subjects. We finally selected 58,894 tandem repeats that were polymorphic in our data for the subsequent association analysis. The proportions of repeats of 1, 2, 3, 4, 5, and 6 nucleotide motif were 0.30, 48.89, 9.47, 29.37, 7.84, and 2.38%, accounting for 98.6% of all analyzed tandem repeats. For association analysis, the sum of two repeat sizes was used as the independent variable (which was similar to the additive model of single SNP association analysis) with the same covariates that were used in our SNP GWAS to explain residual age-at-onset.[15]

Data availability

STR GWAS summary data that support the findings of this study are available from the corresponding author, upon reasonable request.

Results

No significant association between HD and CAG repeat lengths of other polyglutamine disease-causing genes

To determine whether age-at-onset in HD is modified by the CAG repeats of other polyglutamine disease-causing genes, we determined directly the length of CAG repeats in *ATN1* (DRPLA), *ATXN1* (SCA1), *ATXN2* (SCA2), *ATXN3* (SCA3), *CACNA1A* (SCA6), and *TBP* (SCA17) in the HD individuals who participated in both the COHORT study (https://clinicaltrials.gov/ct2/show/NCT00313495) and our recent HD modifier GWA analysis.[15] The CAG repeat sizes of the polyglutamine disease-causing genes showed distinct distribution patterns. For example, *CACNA1A* and *TBP* showed the smallest and the largest median repeat sizes, while *ATXN2* and *ATXN3,* respectively showed different ranges of repeat lengths despite similar median repeat sizes (Supplementary Figure 1). Next, we performed statistical analyses of HD subjects with European ancestry to determine whether 1) the longer repeat, 2) the shorter repeat, or 3) the sum of the two repeat alleles of other polyglutamine disease-causing genes were associated with residual age-at-onset of HD. As the primary phenotype of the analysis, we used residual age-at-onset of HD motor symptoms representing age-at-onset that was corrected for individual pathogenic *HTT* CAG repeat size. In linear regression models corrected for genetic ancestry and other potential confounding factors, residual age-at-onset of HD was not significantly associated with the longer, the shorter, or the sum of the two repeat alleles of any of the tested genes (Table 1). In contrast, HD age-at-onset was significantly associated the size of expanded *HTT* CAG repeat (p-value, 2.1E-111), consistent with our previous report.[12] As expected, residual age-at-onset, representing onset

age corrected for the length of expanded *HTT* CAG repeat, was not significantly associated with *HTT* CAG repeat size (expanded repeat p-value, 0.6948; normal repeat p-value, 0.7171).

Association analysis of *ATXN3* CAG repeat

It has been proposed that the length of CAG repeat in *ATXN3* is associated with age-at-onset of HD.[45] In contrast, our initial analysis of the COHORT participants (n = 503) did not show statistically significant associations between HD residual age-at-onset and *ATXN3* CAG repeat lengths. To confirm this lack of association in a larger sample set, we also analyzed REGISTRY participants who were also part of our recent GWA study. A total of 1388 (706 males and 682 females) unique HD individuals with European ancestry (503 COHORT and 885 REGISTRY) were analyzed for *ATXN3* CAG repeats. Consistent with our initial observations, linear regression analyses to test the longer, the shorter, or the sum of the two repeats of *ATXN3* showed no statistically significant associations with residual age-at-onset of motor signs (Fig. 1). We further performed a dichotomous analysis to test whether age-at-onset or residual age-at-onset was significantly different between HD individuals carrying *ATXN3* repeats above and below the median length. As shown in Supplementary Figure 2, these two groups of HD study participants were not significantly different for age-at-onset (Supplementary Figure 2A) or residual age-at-onset (Supplementary Figure 2B), arguing against modification of HD by the *ATXN3* CAG repeats.

No SNP association signals at other polyglutamine disease-causing genes in HD onset modifier GWAS data

Previously, we discovered genetic modifiers of HD through genome-wide association analysis of SNPs,[15] and subsequently generated a website to make these GWAS results publicly accessible.[49] We used this resource to check HD modification signals at the polyglutamine disease-causing gene assessed above, plus spinocerebellar ataxia type 7 (SCA7, MIM #607640) and spinal and bulbar muscular atrophy (*AR,* MIM #313200). We also included *PPP2R2B* and *DMPK*, expansions of CAG•CTG repeats within which underlie spinocerebellar ataxia type 12 (SCA12, MIM #604326) and myotonic dystrophy type 1 (DM1, MIM #160900), respectively, but are not translated into polyglutamine. For each locus, we evaluated the RefSeq select ('ncbiRefSeqSelect' in the UCSC genome browser) as the representative transcript and identified the SNP with the smallest association *P*-value in the transcript region. The top SNPs at other CAG repeat expansion disease-causing genes were relatively infrequent except *AR.* We observed nominally significant *P*-values for association of some loci with HD residual age-at-onset (Supplementary Table 1), but when these were corrected for the gene size and number of SNPs in the region, none remained statistically significant. Together, our HD modifier SNP GWAS data did not support modification of HD by variants in these other CAG repeat expansion disease-causing genes.

Genome-wide STR association analysis

Though our GWAS, representing the largest dataset of HD individuals with genome-wide genotype and phenotype,[15] successfully identified associated SNPs, those analyses could have limited power to assess effects of polymorphic STRs on modification of HD. Recently, methods for imputation of STR lengths from genome-wide SNP data have been developed and further optimized.[51] Therefore, in order to test the association between HD onset and repeat size of polyglutamine disease-causing genes and other STRs, we imputed ~60,000

STRs for the HD subjects who participated in our HD modifier GWAS (*n*=9058). We then performed association analysis using the same residual age-at-onset phenotype as in our GWAS [15] and STR genotypes as the independent variable with a set of covariates. Specifically, we used the sum of the two repeat lengths, which is similar to the additive model in standard SNP analysis. Notably, the genomic regions containing other polyglutamine disease-causing genes showed no significant STR association signals (Fig. 2, white triangles). The imputed CAG repeat sizes of other polyglutamine disease-causing genes were also not significantly associated with HD residual age-at-onset (Table 2) based on the Bonferroni multiple correction method (*P*-value, 8.5E-7). We further evaluated the levels of association between age-at-onset of HD and imputed CAG repeat of *ATXN3*. To confirm the quality of STR imputation, we compared the genotyped and the imputed *ATXN3* CAG repeat in 1388 HD individuals where both estimates were available. The longer, the shorter, and the sum of the two alleles of the *ATXN3* repeat showed 74.6%, 87.2%, and 69.2% concordance between experimentally determined and imputed repeat lengths (Supplementary Figure 3). Moreover, more than 90% of the observed differences were fewer than 5 repeats (Supplementary Figure 3D), suggesting relatively high levels of accuracy in STR imputation. Like the analysis using typed data, association analyses to test the longer (Supplementary Figure 4A), the shorter (Supplementary Figure 4B), or the sum of the two STR alleles (data, not shown) in the imputed data did not reveal statistically significant association with HD age-at-onset. Furthermore, HD participants who carry longer *ATXN*3 repeats smaller than the median of the longer repeat (i.e., 23 CAG) showed the same age-at-onset and residual age-at-onset compared to those with longer *ATXN3* repeats larger than median (Supplementary Figure 4C and 4D). Together, our genetic analyses using both genotyped and imputed STR data strongly indicated that *ATXN3* CAG repeat length does not modify HD age-at-onset, in contrast to the previous report.

In contrast, as shown in Fig. 2, we identified 3 Bonferroni significant modification signals (*P*-value < 8.5E-7) tagged by STRs on chromosome 2, chromosome 8 and chromosome 15, that potentially captured effects of modification by *PMS1*, *RRM2B*, and *FAN1*,[15] which were also implicated by SNP association. Similarly, the previously implicated *MLH1* region of chromosome 3 and the *MSH3* region at chromosome 5 showed STR association signals at suggestive significance (*P*-value, E-5). Interestingly, a near significant new signal (uncorrected *P*-value, 8.7E-7) was evident on distal chromosome 15q with a GA repeat (chr15:91711532-91711561) in *SV2B* gene, which encodes synaptic vesicle glycoprotein 2B. Still, confirmation analysis by direct repeat genotyping will be required to establish their roles in modifying HD, considering the inherent uncertainty in imputing tandem repeats from SNP data.

Discussion

The existence of multiple neurodegenerative disorders associated with lengthened polyglutamine segments in different proteins has implicated common unifying mechanisms of pathogenesis [3,4,53] and also raised the related possibility that CAG repeat disease genes show functional interaction with one another in modulating disease manifestations. For example, in early studies, SCA3 fasciculations were reported to be associated with normal *ATXN2* CAG repeat length,[54] while SCA2 onset was reported to be influenced by the normal *CACNA1A* CAG repeat which, when expanded, causes SCA6.[55] Subsequently, age-at-onset in several SCAs was reported to be influenced by CAG repeat length variation in various polyglutamine disease-causing genes.[46,56,57] There have also been interactions suggested between coding and non-coding CAG repeats, as the *ATXN1* 31 CAG repeat allele was

reported to be enriched in myotonic dystrophy, which is caused by an expanded CTG repeat in the *DM1* 3'- untranslated region of the *DMPK* gene,[58] although these findings were challenged later.[59] Specifically in HD, age-at-onset appeared to be modified by the normal CAG repeat in *ATXN3*, which is responsible for causing SCA3 when expanded.[45] Given the variability observed across these studies and the possibility that such genetic interactions could provide important insights into both underlying disease mechanisms and potential therapeutic directions, we reasoned that the possibility of modification of HD by other CAG repeat disease-causing genes was an important subject for rigorous investigation.

Using both a candidate approach based on CAG length and an unbiased SNP-based GWAS, we did not detect any significant influence of polyglutamine disease-causing genes and other CAG repeat expansion disease-causing genes on the age-at-onset of HD. The lack of replication of candidate modifiers of HD has been reported before. For example, candidate studies suggested modification of HD by *ADORA2A*,[60,61] *ATG7*,[62] *BDNF*,[63] *GRIK2*,[64] *GRIN2A*,[64-66] *GRIN2B*,[64,65] *HAP1*,[67] *HIP1*, [64] *LINC01559*,[64] *NPY2R*,[68] *PPARGC1A*,[69-72] and *UCHL1*.[73] However, none of these genes generated significant onset modification signals in our large scale unbiased genetic analysis.[14,15,49,74] Interestingly, one candidate modifier that showed a trend of association [66,75] and was replicated by GWAS is *TCERG1,* which harbors a complex coding hexamer repeat that appears to be the source of the influence on HD age-at-onset.[76] Unfortunately, the hexamer repeat with potential association was not imputed in our data because this repeat was not present in the imputation reference panel that we used. Nevertheless, the lack of replication for most candidates could be due to spurious signals from underpowered studies, confounded by ancestry differences, lack of multiple test correction, and/or outlier effects.[74,77] Outlier effects are particularly significant when using continuous variables, as we observed that a single data point could change insignificant signal into significant association.[12] This may explain the lack of replication of modifying effects of

the *ATXN3* repeat on HD age-at-onset. The overall high rate of failure to replicate reinforces the importance of rigorous data quality control and stringent statistical analysis for the association analysis of human data.

Although our data did not validate the modification of HD by other polyglutamine disease-causing genes, imputed STR data did show Bonferroni significant signals on chromosomes 2, 8, and 15. These significant STRs appear to tag previously identified modifier haplotypes of *PMS1*, *RRM2B,* and *FAN1,* which were detected in our SNP-based GWAS. Considering that this original association analysis tested more than 10 million SNPs, the detection of these 3 significant association signals from testing ~60,000 genetic polymorphisms supports the levels of power and efficiency of the STR approach and argues for its use in modifier studies of other disorders. More than sixty diseases are known to be caused by expansions of tandem repeats, and additional disease-causing repeats are being discovered with the advance of genomic technologies.[78,79] In addition to tandem repeats that cause Mendelian disorders, repeat polymorphism may contribute to the missing heritability of common polygenic disorders.[80-82] Importantly, changes in tandem repeats represent one of the major sources of *de novo* mutation with clinical significances.[83,84] For example, somatically expanded tandem repeats influence disease age of onset and tissue specificity of pathogenic features[78,79] Furthermore, significant genome-wide excess of tandem repeat mutations has been reported in the autism spectrum disorder,[85-87] implying that tandem repeats may have profound effects on human health beyond the well characterized repeat expansion disorders. Many GWAS signals are due to effects on gene expression levels, and recent findings of a role for length and motif composition of the tandem repeats including variable number tandem repeats (VNTRs) in regulating gene expression [88-92] suggest that some HD onset modification signals might be caused by altered expression of modifier genes due to polymorphic tandem repeat lengths. Therefore, investigating a potential role for tandem

repeats in regulating the expression levels of the *PMS1, RRM2B*, and *FAN1* modifier genes implicated at the chromosome 2, 8 and 15 loci may reveal an important underlying source of HD modification. Finally, the STR association signal on distal chromosome 15q suggests the possibility of an HD modifier effect due to synaptic vesicle glycoprotein 2B (*SV2B*). The *SV2B* protein localizes to synaptic vesicles, where it is believed to function in the regulation of vesicle trafficking and exocytosis. While a role at synapses makes it an attractive candidate for involvement in the HD damage mechanism(s) precipitated by the expanded CAG repeat, this locus will require replication or other confirmation to ensure that it harbors a *bona fide* genetic modifier of HD.

In summary, we expanded our approaches for identifying genetic modifiers of HD using typed and imputed repeats. We focused on STRs in this study primarily due to their clinical significance and available resources for genome-wide imputation.[51,78,79,84] However, other genetic variations (i.e., VNTR and structural variations) also generate biological consequences in humans.[93,94] Therefore, investigation of other types DNA polymorphisms may yield a more complete map of genetic modifiers of HD. Nevertheless, although our data clearly show the lack of modification of HD onset by CAG repeat size polymorphisms in other polyglutamine disease genes, they do point, along with the complex coding hexamer repeat in *TCERG1* [76] and a complex nonamer coding repeat in *MSH3*,[21] to the potential of finer delineation of other tandem repeats across the genome as a potential source of modifiers that could further refine the HD landscape and inform the development of treatments for HD.

Acknowledgements

This study would not be possible without the vital contribution of the research participants and their families. This manuscript is dedicated to the late Dr. Lesley Jones, who had made invaluable contributions to this study and countless discoveries in Huntington's disease.

Funding

This research was supported by the CHDI Foundation Inc., the U.S. National Institutes of Health (NS082079, NS091161, NS016367, NS049206, NS105709, NS119471), the Medical Research Council (UK MR/L010305/1 and fellowship MR/P001629/1), and a Cardiff University School of Medicine studentship. EMR was the recipient of a scholarship from Fundação para a Ciência e a Tecnologia (SFRH/BD/44335/2008). NAA is partly supported by an Alzheimer's Association Research Grant (Award Number: AARG-19-616534) and a European Research Council Starting Grant (Number: 101041677).

Competing interests

J.F.G. was a Scientific Advisory Board member and had a financial interest in Triplet Therapeutics, Inc. His NIH-funded project is using genetic and genomic approaches to uncover other genes that significantly influence when diagnosable symptoms emerge and how rapidly they worsen in Huntington disease. The company is developing new therapeutic approaches to address triplet repeat disorders such Huntington's disease, myotonic dystrophy and spinocerebellar ataxias. His interests were reviewed and are managed by Massachusetts General Hospital and Mass General Brigham in accordance with their conflict of interest policies. J.F.G. has also been a consultant for Wave Life Sciences USA, Inc., Biogen, Inc.

and Pfizer, Inc. Within the last five years D.G.M. has been a scientific consultant and/or received an honoraria/stock options from AMO Pharma, Dyne, F. Hoffman-La Roche, LoQus23, Novartis, Ono Pharmaceuticals, Rgenta Therapeutics, Sanofi, Sarepta Therapeutics Inc, Script Biosciences, Triplet Therapeutics, and Vertex Pharmaceuticals and held research contracts with AMO Pharma and Vertex Pharmaceuticals. J.D.L. is a paid Advisory Board member for F. Hoffmann-La Roche Ltd and uniQure biopharma B.V., and he is a paid consultant for Vaccinex Inc, Wave Life Sciences USA Inc, Genentech Inc, Triplet Inc, and PTC Therapeutics Inc. T.H.M. is an associate member of the scientific advisory board of LoQus23 Therapeutics. L.J. was a member of the scientific advisory boards of LoQus23 Therapeutics and Triplet Therapeutics. V.C.W. was a Scientific Advisory Board member of Triplet Therapeutics, Inc., a company developing new therapeutic approaches to address triplet repeat disorders such Huntington's disease and myotonic dystrophy. Her financial interests in Triplet Therapeutics were reviewed and are managed by Massachusetts General Hospital and Mass General Brigham in accordance with their conflict of interest policies. She is a scientific advisory board member of LoQus23 Therapeutics and has provided paid consulting services to Alnylam, Acadia Pharmaceuticals Inc., Alnylam Inc., Biogen Inc. and Passage Bio. J.M.L consults for Life Edit Therapeutics and serves in the scientific advisory board of GenEdit, Inc.

References

1. Depienne C, Mandel JL. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *American journal of human genetics*. May 6 2021;108(5):764-785. doi:10.1016/j.ajhg.2021.03.011
2. Di Prospero NA, Fischbeck KH. Therapeutics development for triplet repeat expansion diseases. *Nature reviews Genetics*. Oct 2005;6(10):756-65. doi:10.1038/nrg1690
3. Ross CA. Polyglutamine pathogenesis: emergence of unifying mechanisms for Huntington's disease and related disorders. *Neuron*. Aug 29 2002;35(5):819-22. doi:10.1016/s0896-6273(02)00872-3
4. Paulson HL, Bonini NM, Roth KA. Polyglutamine disease and neuronal cell death. *Proceedings of the National Academy of Sciences of the United States of America*. Nov 21 2000;97(24):12957-8. doi:10.1073/pnas.210395797
5. Orr HT, Zoghbi HY. Trinucleotide repeat disorders. *Annual review of neuroscience*. 2007;30:575-621. doi:10.1146/annurev.neuro.29.051605.113042
6. Andrew SE, Goldberg YP, Kremer B*, et al*. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nature genetics*. Aug 1993;4(4):398-403. doi:10.1038/ng0893-398
7. Duyao M, Ambrose C, Myers R*, et al*. Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nature genetics*. Aug 1993;4(4):387-92. doi:10.1038/ng0893-387
8. Snell RG, MacMillan JC, Cheadle JP*, et al*. Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. *Nature genetics*. Aug 1993;4(4):393-7. doi:10.1038/ng0893-393
9. Persichetti F, Srinidhi J, Kanaley L*, et al*. Huntington's disease CAG trinucleotide repeats in pathologically confirmed post-mortem brains. *Neurobiology of disease*. Dec 1994;1(3):159-66. doi:10.1006/nbdi.1994.0019
10. Gusella JF, MacDonald ME. Molecular genetics: unmasking polyglutamine triggers in neurodegenerative disease. *Nature reviews Neuroscience*. Nov 2000;1(2):109-15. doi:10.1038/35039051
11. Stevanin G, Durr A, Brice A. Clinical and molecular advances in autosomal dominant cerebellar ataxias: from genotype to phenotype and physiopathology. *European journal of human genetics : EJHG*. Jan 2000;8(1):4-18. doi:10.1038/sj.ejhg.5200403
12. Lee JM, Ramos EM, Lee JH*, et al*. CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. *Neurology*. Mar 6 2012;78(10):690-5. doi:10.1212/WNL.0b013e318249f683
13. Li JL, Hayden MR, Almqvist EW*, et al*. A genome scan for modifiers of age at onset in Huntington disease: The HD MAPS study. *American journal of human genetics*. Sep 2003;73(3):682-7. doi:10.1086/378133
14. Genetic Modifiers of Huntington's Disease C. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell*. Jul 30 2015;162(3):516-26. doi:10.1016/j.cell.2015.07.003
15. Genetic Modifiers of Huntington's Disease Consortium. Electronic address ghmhe, Genetic Modifiers of Huntington's Disease C. CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell*. Aug 8 2019;178(4):887-900 e14. doi:10.1016/j.cell.2019.06.036

16. Pearson CE, Nichol Edamura K, Cleary JD. Repeat instability: mechanisms of dynamic mutations. *Nature reviews Genetics*. Oct 2005;6(10):729-42. doi:10.1038/nrg1689
17. Kovtun IV, McMurray CT. Features of trinucleotide repeat instability in vivo. *Cell research*. Jan 2008;18(1):198-213. doi:10.1038/cr.2008.5
18. Dragileva E, Hendricks A, Teed A*, et al*. Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. *Neurobiology of disease*. Jan 2009;33(1):37-47. doi:10.1016/j.nbd.2008.09.014
19. Pinto RM, Dragileva E, Kirby A*, et al*. Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. *PLoS genetics*. Oct 2013;9(10):e1003930. doi:10.1371/journal.pgen.1003930
20. Ciosi M, Maxwell A, Cumming SA*, et al*. A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes. *EBioMedicine*. Oct 2019;48:568-580. doi:10.1016/j.ebiom.2019.09.020
21. Flower M, Lomeikaite V, Ciosi M*, et al*. MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. *Brain : a journal of neurology*. Jun 19 2019;142(7):1876-86. doi:10.1093/brain/awz115
22. Goold R, Flower M, Moss DH*, et al*. FAN1 modifies Huntington's disease progression by stabilizing the expanded HTT CAG repeat. *Human molecular genetics*. Feb 15 2019;28(4):650-661. doi:10.1093/hmg/ddy375
23. Kim KH, Hong EP, Shin JW*, et al*. Genetic and Functional Analyses Point to FAN1 as the Source of Multiple Huntington Disease Modifier Effects. *American journal of human genetics*. Jul 2 2020;107(1):96-110. doi:10.1016/j.ajhg.2020.05.012
24. Goold R, Hamilton J, Menneteau T*, et al*. FAN1 controls mismatch repair complex assembly via MLH1 retention to stabilize CAG repeat expansion in Huntington's disease. *Cell reports*. Aug 31 2021;36(9):109649. doi:10.1016/j.celrep.2021.109649
25. McAllister B, Donaldson J, Binda CS*, et al*. Exome sequencing of individuals with Huntington's disease implicates FAN1 nuclease activity in slowing CAG expansion and disease onset. *Nature neuroscience*. Apr 2022;25(4):446-457. doi:10.1038/s41593-022-01033-5
26. Takano H, Onodera O, Takahashi H*, et al*. Somatic mosaicism of expanded CAG repeats in brains of patients with dentatorubral-pallidoluysian atrophy: cellular population-dependent dynamics of mitotic instability. *American journal of human genetics*. Jun 1996;58(6):1212-22.
27. Maciel P, Lopes-Cendes I, Kish S, Sequeiros J, Rouleau GA. Mosaicism of the CAG repeat in CNS tissue in relation to age at death in spinocerebellar ataxia type 1 and Machado-Joseph disease patients. *American journal of human genetics*. Apr 1997;60(4):993-6.
28. Manley K, Shirley TL, Flaherty L, Messer A. Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. *Nature genetics*. Dec 1999;23(4):471-3. doi:10.1038/70598
29. Matsuura T, Sasaki H, Yabe I*, et al*. Mosaicism of unstable CAG repeats in the brain of spinocerebellar ataxia type 2. *Journal of neurology*. Sep 1999;246(9):835-9. doi:10.1007/s004150050464
30. Kennedy L, Evans E, Chen CM*, et al*. Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Human molecular genetics*. Dec 15 2003;12(24):3359-67. doi:10.1093/hmg/ddg352
31. Wheeler VC, Lebel LA, Vrbanac V, Teed A, te Riele H, MacDonald ME. Mismatch repair gene Msh2 modifies the timing of early disease in Hdh(Q111) striatum. *Human molecular genetics*. Feb 1 2003;12(3):273-81. doi:10.1093/hmg/ddg056

32. Gomes-Pereira M, Fortune MT, Ingram L, McAbney JP, Monckton DG. Pms2 is a genetic enhancer of trinucleotide CAG.CTG repeat somatic mosaicism: implications for the mechanism of triplet repeat expansion. *Human molecular genetics*. Aug 15 2004;13(16):1815-25. doi:10.1093/hmg/ddh186
33. Owen BA, Yang Z, Lai M*, et al*. (CAG)(n)-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. *Nature structural & molecular biology*. Aug 2005;12(8):663-70. doi:10.1038/nsmb965
34. Tome S, Manley K, Simard JP*, et al*. MSH3 polymorphisms and protein levels affect CAG repeat instability in Huntington's disease mice. *PLoS genetics*. 2013;9(2):e1003280. doi:10.1371/journal.pgen.1003280
35. Bettencourt C, Hensman-Moss D, Flower M*, et al*. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. *Annals of neurology*. Jun 2016;79(6):983-90. doi:10.1002/ana.24656
36. Morales F, Vasquez M, Santamaria C, Cuenca P, Corrales E, Monckton DG. A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. *DNA repair*. Apr 2016;40:57-66. doi:10.1016/j.dnarep.2016.01.001
37. Massey TH, Jones L. The central role of DNA damage and repair in CAG repeat diseases. *Disease models & mechanisms*. Jan 30 2018;11(1)doi:10.1242/dmm.031930
38. Laabs BH, Klein C, Pozojevic J*, et al*. Identifying genetic modifiers of age-associated penetrance in X-linked dystonia-parkinsonism. *Nature communications*. May 28 2021;12(1):3216. doi:10.1038/s41467-021-23491-4
39. Roy JCL, Vitalo A, Andrew MA*, et al*. Somatic CAG expansion in Huntington's disease is dependent on the MLH3 endonuclease domain, which can be excluded via splice redirection. *Nucleic acids research*. Apr 19 2021;49(7):3907-3918. doi:10.1093/nar/gkab152
40. Telenius H, Kremer B, Goldberg YP*, et al*. Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. *Nature genetics*. Apr 1994;6(4):409-14. doi:10.1038/ng0494-409
41. Fortune MT, Vassilopoulos C, Coolbaugh MI, Siciliano MJ, Monckton DG. Dramatic, expansion-biased, age-dependent, tissue-specific somatic mosaicism in a transgenic mouse model of triplet repeat instability. *Human molecular genetics*. Feb 12 2000;9(3):439-45. doi:10.1093/hmg/9.3.439
42. Kennedy L, Shelbourne PF. Dramatic mutation instability in HD mouse striatum: does polyglutamine load contribute to cell-specific vulnerability in Huntington's disease? *Human molecular genetics*. Oct 12 2000;9(17):2539-44. doi:10.1093/hmg/9.17.2539
43. Watase K, Venken KJ, Sun Y, Orr HT, Zoghbi HY. Regional differences of somatic CAG repeat instability do not account for selective neuronal vulnerability in a knock-in mouse model of SCA1. *Human molecular genetics*. Nov 1 2003;12(21):2789-95. doi:10.1093/hmg/ddg300
44. Mouro Pinto R, Arning L, Giordano JV*, et al*. Patterns of CAG repeat instability in the central nervous system and periphery in Huntington's disease and in spinocerebellar ataxia type 1. *Human molecular genetics*. Aug 29 2020;29(15):2551-2567. doi:10.1093/hmg/ddaa139
45. Stuitje G, van Belzen MJ, Gardiner SL*, et al*. Age of onset in Huntington's disease is influenced by CAG repeat variations in other polyglutamine disease-associated genes. *Brain : a journal of neurology*. Jul 1 2017;140(7):e42. doi:10.1093/brain/awx122
46. Tezenas du Montcel S, Durr A, Bauer P*, et al*. Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes. *Brain : a journal of neurology*. Sep 2014;137(Pt 9):2444-55. doi:10.1093/brain/awu174

47. Gao R, Chakraborty A, Geater C*, et al*. Mutant huntingtin impairs PNKP and ATXN3, disrupting DNA repair and transcription. *eLife*. Apr 17 2019;8doi:10.7554/eLife.42988
48. Sequeiros J, Seneca S, Martindale J. Consensus and controversies in best practices for molecular genetic testing of spinocerebellar ataxias. *European journal of human genetics : EJHG*. Nov 2010;18(11):1188-95. doi:10.1038/ejhg.2010.10
49. Hong EP, MacDonald ME, Wheeler VC*, et al*. Huntington's Disease Pathogenesis: Two Sequential Components. *Journal of Huntington's disease*. 2021;10(1):35-51. doi:10.3233/JHD-200427
50. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics*. Nov 2007;81(5):1084-97. doi:10.1086/521987
51. Saini S, Mitra I, Mousavi N, Fotsing SF, Gymrek M. A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nature communications*. Oct 23 2018;9(1):4397. doi:10.1038/s41467-018-06694-0
52. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*. Jan 15 1999;27(2):573-80. doi:10.1093/nar/27.2.573
53. Gatchel JR, Zoghbi HY. Diseases of unstable repeat expansion: mechanisms and common principles. *Nature reviews Genetics*. Oct 2005;6(10):743-55. doi:10.1038/nrg1691
54. Jardim L, Silveira I, Pereira ML*, et al*. Searching for modulating effects of SCA2, SCA6 and DRPLA CAG tracts on the Machado-Joseph disease (SCA3) phenotype. *Acta neurologica Scandinavica*. Mar 2003;107(3):211-4. doi:10.1034/j.1600-0404.2003.00046.x
55. Pulst SM, Santos N, Wang D*, et al*. Spinocerebellar ataxia type 2: polyQ repeat variation in the CACNA1A calcium channel modifies age of onset. *Brain : a journal of neurology*. Oct 2005;128(Pt 10):2297-303. doi:10.1093/brain/awh586
56. Raposo M, Ramos A, Bettencourt C, Lima M. Replicating studies of genetic modifiers in spinocerebellar ataxia type 3: can homogeneous cohorts aid? *Brain : a journal of neurology*. Dec 2015;138(Pt 12):e398. doi:10.1093/brain/awv206
57. Chen Z, Zheng C, Long Z*, et al*. (CAG)n loci as genetic modifiers of age-at-onset in patients with Machado-Joseph disease from mainland China. *Brain : a journal of neurology*. Aug 2016;139(Pt 8):e41. doi:10.1093/brain/aww087
58. Savic D, Topisirovic I, Keckarevic M*, et al*. Is the 31 CAG repeat allele of the spinocerebellar ataxia 1 (SCA1) gene locus non-specifically associated with trinucleotide expansion diseases? *Psychiatric genetics*. Dec 2001;11(4):201-5. doi:10.1097/00041444-200112000-00004
59. Hellenbroich Y, Kaulich M, Opitz S, Schwinger E, Zuhlke C. No association of the SCA1 (CAG)31 allele with Huntington's disease, myotonic dystrophy type 1 and spinocerebellar ataxia type 3. *Psychiatric genetics*. Jun 2004;14(2):61-3. doi:10.1097/01.ypg.0000128763.69225.77
60. Dhaenens CM, Burnouf S, Simonin C*, et al*. A genetic variation in the ADORA2A gene modifies age at onset in Huntington's disease. *Neurobiology of disease*. Sep 2009;35(3):474-6. doi:10.1016/j.nbd.2009.06.009
61. Taherzadeh-Fard E, Saft C, Wieczorek S, Epplen JT, Arning L. Age at onset in Huntington's disease: replication study on the associations of ADORA2A, HAP1 and OGG1. *Neurogenetics*. Oct 2010;11(4):435-9. doi:10.1007/s10048-010-0248-3
62. Metzger S, Saukko M, Van Che H*, et al*. Age at onset in Huntington's disease is modified by the autophagy pathway: implication of the V471A polymorphism in Atg7. *Human genetics*. Oct 2010;128(4):453-9. doi:10.1007/s00439-010-0873-9

63. Alberch J, Lopez M, Badenas C*, et al*. Association between BDNF Val66Met polymorphism and age at onset in Huntington disease. *Neurology*. Sep 27 2005;65(6):964-5. doi:10.1212/01.wnl.0000175977.57661.b1
64. Valcarcel-Ocete L, Alkorta-Aranburu G, Iriondo M*, et al*. Exploring Genetic Factors Involved in Huntington Disease Age of Onset: E2F2 as a New Potential Modifier Gene. *PloS one*. 2015;10(7):e0131573. doi:10.1371/journal.pone.0131573
65. Arning L, Kraus PH, Valentin S, Saft C, Andrich J, Epplen JT. NR2A and NR2B receptor gene variations modify age at onset in Huntington disease. *Neurogenetics*. Feb 2005;6(1):25-8. doi:10.1007/s10048-004-0198-8
66. Andresen JM, Gayan J, Cherny SS*, et al*. Replication of twelve association studies for Huntington's disease residual age of onset in large Venezuelan kindreds. *Journal of medical genetics*. Jan 2007;44(1):44-50. doi:10.1136/jmg.2006.045153
67. Metzger S, Rong J, Nguyen HP*, et al*. Huntingtin-associated protein-1 is a modifier of the age-at-onset of Huntington's disease. *Human molecular genetics*. Apr 15 2008;17(8):1137-46. doi:10.1093/hmg/ddn003
68. Kloster E, Saft C, Akkad DA, Epplen JT, Arning L. Association of age at onset in Huntington disease with functional promoter variations in NPY and NPY2R. *Journal of molecular medicine*. Feb 2014;92(2):177-84. doi:10.1007/s00109-013-1092-3
69. Taherzadeh-Fard E, Saft C, Andrich J, Wieczorek S, Arning L. PGC-1alpha as modifier of onset age in Huntington disease. *Molecular neurodegeneration*. Feb 6 2009;4:10. doi:10.1186/1750-1326-4-10
70. Weydt P, Soyal SM, Gellera C*, et al*. The gene coding for PGC-1alpha modifies age at onset in Huntington's Disease. *Molecular neurodegeneration*. Jan 8 2009;4:3. doi:10.1186/1750-1326-4-3
71. Che HV, Metzger S, Portal E, Deyle C, Riess O, Nguyen HP. Localization of sequence variations in PGC-1alpha influence their modifying effect in Huntington disease. *Molecular neurodegeneration*. Jan 6 2011;6(1):1. doi:10.1186/1750-1326-6-1
72. Weydt P, Soyal SM, Landwehrmeyer GB, Patsch W, European Huntington Disease N. A single nucleotide polymorphism in the coding region of PGC-1alpha is a male-specific modifier of Huntington disease age-at-onset in a large European cohort. *BMC neurology*. Jan 2 2014;14:1. doi:10.1186/1471-2377-14-1
73. Metzger S, Bauer P, Tomiuk J*, et al*. The S18Y polymorphism in the UCHL1 gene is a genetic modifier in Huntington's disease. *Neurogenetics*. Mar 2006;7(1):27-30. doi:10.1007/s10048-005-0023-z
74. Correia K, Harold D, Kim KH*, et al*. The Genetic Modifiers of Motor OnsetAge (GeM MOA) Website: Genome-wide Association Analysis for Genetic Modifiers of Huntington's Disease. *Journal of Huntington's disease*. 2015;4(3):279-84. doi:10.3233/JHD-150169
75. Holbert S, Denghien I, Kiechle T*, et al*. The Gln-Ala repeat transcriptional activator CA150 interacts with huntingtin: neuropathologic and genetic evidence for a role in Huntington's disease pathogenesis. *Proceedings of the National Academy of Sciences of the United States of America*. Feb 13 2001;98(4):1811-6. doi:10.1073/pnas.98.4.1811
76. Lobanov SV, McAllister B, McDade-Kumar M*, et al*. Huntington's disease age at motor onset is modified by the tandem hexamer repeat in TCERG1. *NPJ genomic medicine*. Sep 5 2022;7(1):53. doi:10.1038/s41525-022-00317-w
77. Gusella JF, MacDonald ME, Lee JM. Genetic modifiers of Huntington's disease. *Movement disorders : official journal of the Movement Disorder Society*. Sep 15 2014;29(11):1359-65. doi:10.1002/mds.26001

78. Gall-Duncan T, Sato N, Yuen RKC, Pearson CE. Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences. *Genome research*. Jan 2022;32(1):1-27. doi:10.1101/gr.269530.120
79. Malik I, Kelley CP, Wang ET, Todd PK. Molecular mechanisms underlying nucleotide repeat expansion disorders. *Nature reviews Molecular cell biology*. Sep 2021;22(9):589-607. doi:10.1038/s41580-021-00382-6
80. Hannan AJ. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends in genetics : TIG*. Feb 2010;26(2):59-65. doi:10.1016/j.tig.2009.11.008
81. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nature reviews Genetics*. May 2018;19(5):286-298. doi:10.1038/nrg.2017.115
82. Mukamel RE, Handsaker RE, Sherman MA, *et al*. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science*. Sep 24 2021;373(6562):1499-1505. doi:10.1126/science.abg8289
83. Verbiest M, Maksimov M, Jin Y, Anisimova M, Gymrek M, Bilgin Sonay T. Mutation and selection processes regulating short tandem repeats give rise to genetic and phenotypic diversity across species. *Journal of evolutionary biology*. Feb 2023;36(2):321-336. doi:10.1111/jeb.14106
84. Gymrek M. A genomic view of short tandem repeats. *Current opinion in genetics & development*. Jun 2017;44:9-16. doi:10.1016/j.gde.2017.01.012
85. Mitra I, Huang B, Mousavi N, *et al*. Patterns of de novo tandem repeat mutations and their role in autism. *Nature*. Jan 2021;589(7841):246-250. doi:10.1038/s41586-020-03078-7
86. Hannan AJ. Repeat DNA expands our understanding of autism spectrum disorder. *Nature*. Jan 2021;589(7841):200-202. doi:10.1038/d41586-020-03658-7
87. Trost B, Engchuan W, Nguyen CM, *et al*. Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature*. Oct 2020;586(7827):80-86. doi:10.1038/s41586-020-2579-z
88. Fotsing SF, Margoliash J, Wang C, *et al*. The impact of short tandem repeat variation on gene expression. *Nature genetics*. Nov 2019;51(11):1652-1659. doi:10.1038/s41588-019-0521-9
89. Bakhtiari M, Park J, Ding YC, *et al*. Variable number tandem repeats mediate the expression of proximal genes. *Nature communications*. Apr 6 2021;12(1):2075. doi:10.1038/s41467-021-22206-z
90. Eslami Rasekh M, Hernandez Y, Drinan SD, Fuxman Bass JI, Benson G. Genome-wide characterization of human minisatellite VNTRs: population-specific alleles and gene expression differences. *Nucleic acids research*. May 7 2021;49(8):4308-4324. doi:10.1093/nar/gkab224
91. Garg P, Martin-Trujillo A, Rodriguez OL, *et al*. Pervasive cis effects of variation in copy number of large tandem repeats on local DNA methylation and gene expression. *American journal of human genetics*. May 6 2021;108(5):809-824. doi:10.1016/j.ajhg.2021.03.016
92. Lu TY, Smaruj PN, Fudenberg G, Mancuso N, Chaisson MJP. The motif composition of variable number tandem repeats impacts gene expression. *Genome research*. Apr 2023;33(4):511-524. doi:10.1101/gr.276768.122
93. Hurles ME, Dermitzakis ET, Tyler-Smith C. The functional impact of structural variation in humans. *Trends in genetics : TIG*. May 2008;24(5):238-45. doi:10.1016/j.tig.2008.03.001
94. Marshall JN, Lopez AI, Pfaff AL, Koks S, Quinn JP, Bubb VJ. Variable number tandem repeats - Their emerging role in sickness and health. *Experimental biology and medicine*. Jun 2021;246(12):1368-1376. doi:10.1177/15353702211003511

Figure legends

Fig. 1. The lack of modification of HD onset by *ATXN3* CAG repeats.

To validate the COHORT data analysis of *ATXN3* results, we also analyzed the *ATXN3* repeat in the REGISTRY samples. Subsequently, we performed linear regression analysis using the combined data to determine whether the longer, shorter or sum of the two repeat alleles could explain residual age-at-onset of HD (n=1,388). Twenty-six repeat alleles were observed; the most and second most frequent repeat alleles were 23 and 14 CAGs, accounting for 53% of all repeat alleles. The Y-axis shows residual age-at-onset of HD subjects, representing age-at-onset corrected for individual *HTT* CAG repeat size (i.e., years). The X-axis represents the length of CAG repeat of *ATXN3* (i.e., [CAG]n).

Fig. 2. Genome-wide STR association analysis of residual age-at-onset of HD.

STRs were imputed based on the typed SNP data and subsequently used as the predictor variable with other covariates to explain residual age-at-onset of HD. For this analysis, we used the sum of two alleles for a given STR (i.e., additive model). Y-axis represent significance levels of association, expressed as -log10(*P*-value). A dotted red line represents Bonferroni significance (*P*-value, 8.48E-7 based on 58,894 tests). White triangles mark the polyglutamine disease-causing repeats.

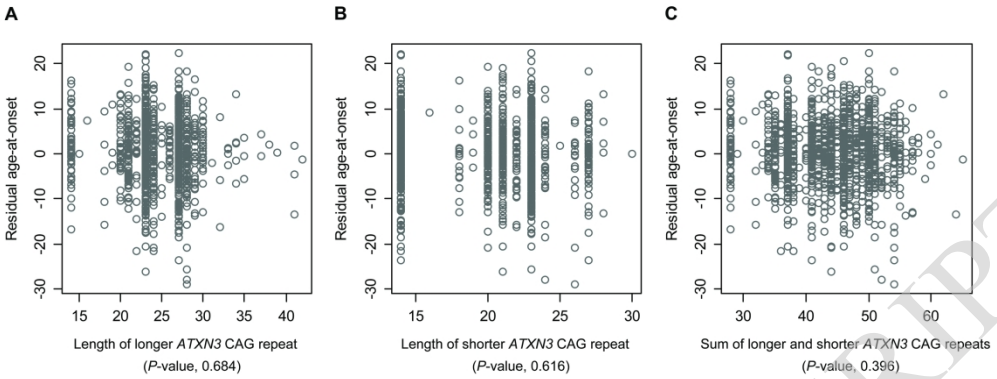
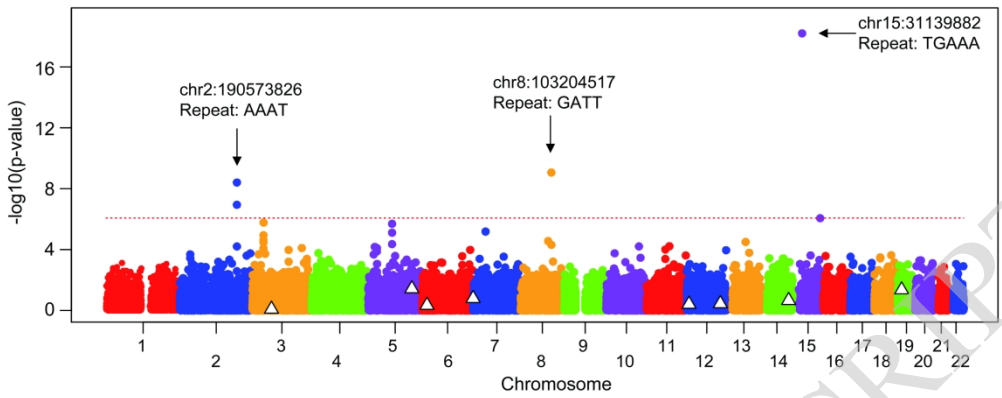


Figure 1

231x87mm (300 x 300 DPI)



Figuer 2

228x89mm (300 x 300 DPI)

Table 1. Statistical analysis to test association between residual age-at-onset of HD and CAG repeat size of other polyglutamine disease-associated genes.

Gene	Disease	Subjects (*n*)	Range of repeat length	*P*-value		
				Longer repeat	Shorter repeat	Sum of repeats
ATN1	DRPLA	551	4-19	0.4691	0.1653	0.0635
ATXN1	SCA1	502	21-37	0.3091	0.3281	0.8652
ATXN2	SCA2	604	14-33	0.7241	0.9873	0.8670
ATXN3	SCA3	503	14-41	0.5038	0.9361	0.5517
CACNA1A	SCA6	497	4-18	0.5760	0.1205	0.5201
TBP	SCA17	483	29-42	0.837	0.452	0.3904

To test whether residual age-at-onset of HD is significantly associated with CAG repeat lengths of other polyglutamine disease-associated genes, we performed linear regression analyses. For each of the test genes, either the longer, shorter, or the sum of both repeat alleles was used as a continuous predictor variable to explain residual age-at-onset of HD. Sample size and *p*-value are shown. R-squared values were smaller than 1% for all tested alleles.

Table 2. Significance of CAG repeat in imputed STR association analysis.

Polyglutamine disease	Chromosomal location (GRCh37/hg19)	Gene	*P*-value
DRPLA (Dentatorubropallidoluysian atrophy)	Chr12:7045880	*ATN1*	0.5785
SCA1 (Spinocerebellar ataxia type 1)	Chr6:16327865	*ATXN1*	0.4584
SCA2 (Spinocerebellar ataxia type 2)	Chr12:112036754	*ATXN2*	0.3602
SCA3 (Spinocerebellar ataxia type 3)	Chr14:92537353	*ATXN3*	0.2198
SCA6 (Spinocerebellar ataxia type 6)	Chr19:13318673	*CACNA1A*	0.0442
SCA7 (Spinocerebellar ataxia type 7)	Chr3:63898361	*ATXN7*	0.8243
SCA12 (Spinocerebellar ataxia type 12)	Chr5:146258291	*PPP2R2B*	0.0377
SCA17 (Spinocerebellar ataxia type 17)	Chr6:170870996	*TBP*	0.1638

Significance of CAG repeat of other polyglutamine disease genes extracted from the genome wide STR association analysis using an additive model. Since STRs were imputed only for autosomes, the SBMA CAG repeat on the X chromosome was not assessed.

Genome-wide association study (GWAS) to identify short tandem repeats that modify Huntington's disease

GWAS data of Huntington's disease (HD) patients

↓

Imputation of short tandem repeat (STR)

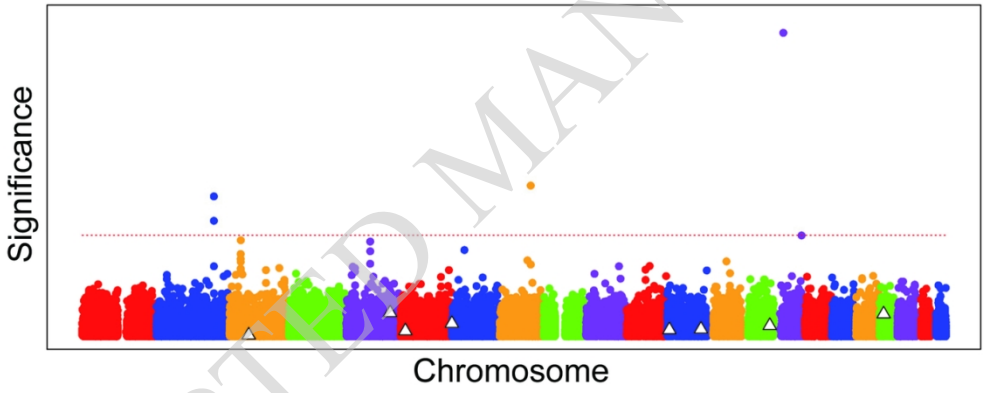
↓

Genetic association analysis using STR as the genotype and HD residual age-at-onset as the phenotype variable

↓

No impacts of CAG of polyQ disease-associated genes on HD onset

Identification of other STRs associated with HD



Graphical Abstract

118x118mm (300 x 300 DPI)

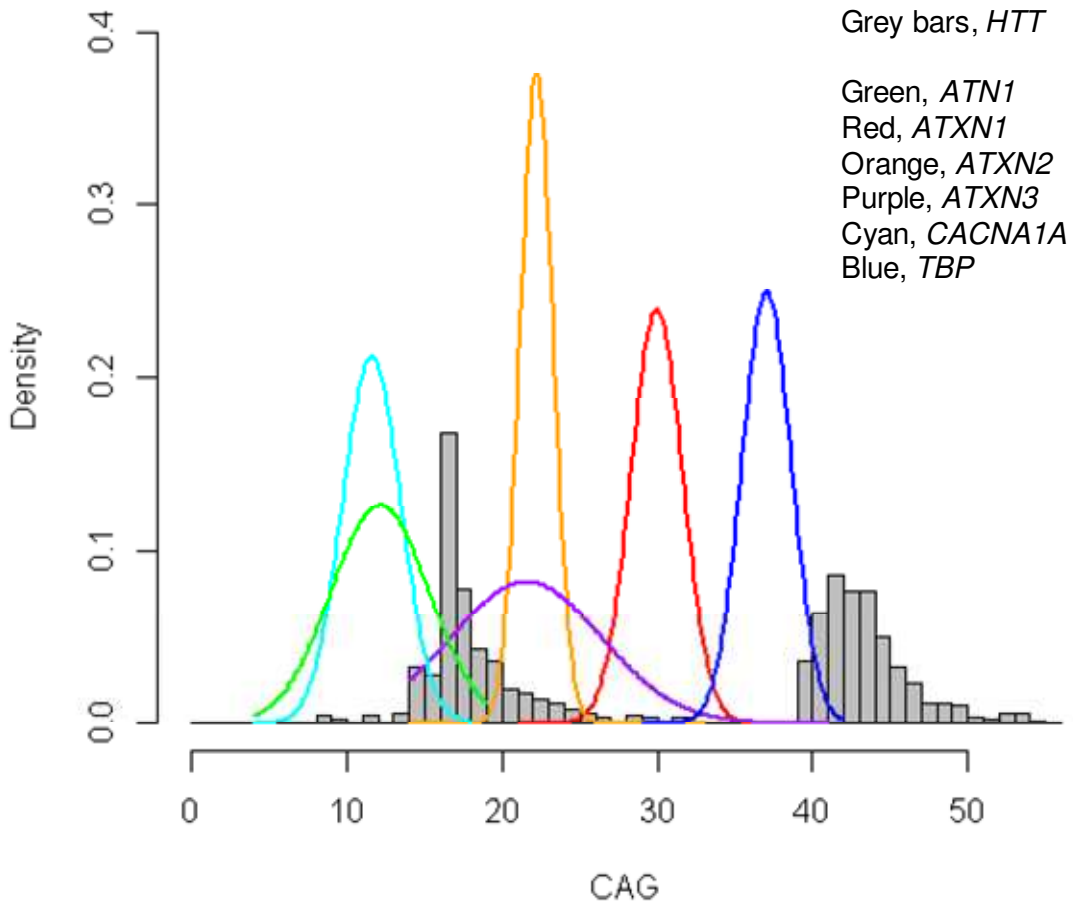
Supplementary Table 1. SNP association with residual age-at-onset in HD.

For each of other CAG repeat expansion disease genes, we evaluated the levels of association based on our recent SNP association for HD modification. A RefSeq select transcript as a representative transcript was used for a given region to identify the SNP with the highest significance in the GeM Euro 9K website. To correct the top SNP *P*-value for the gene size and number of SNPs we applied modified Sidak method to obtain corrected *P*-value.

Gene	Disease	Top SNP	MAF (%)	*P*-value	Corrected *P*-value
ATN1	DRPLA	rs181318837	1.539	0.036365	0.296653
AR	SBMA	rs5918762	15.58	0.372432	1
ATXN1	SCA1	rs80281835	1.677	0.003874	0.937191
ATXN2	SCA2	rs77838113	2.609	0.016352	0.599493
ATXN3	SCA3	rs55961283	2.499	0.103693	0.999995
ATXN7	SCA7	rs77203794	1.473	0.025268	0.952429
CACNA1A	SCA6	rs145803932	1.749	0.017769	0.99987
PPP2R2B	SCA12	rs4705448	1.114	0.000421	0.207854
TBP	SCA17	rs73256671	6.509	0.039786	0.598874
DMPK	DM1	rs183029748	1.269	0.072635	0.492709

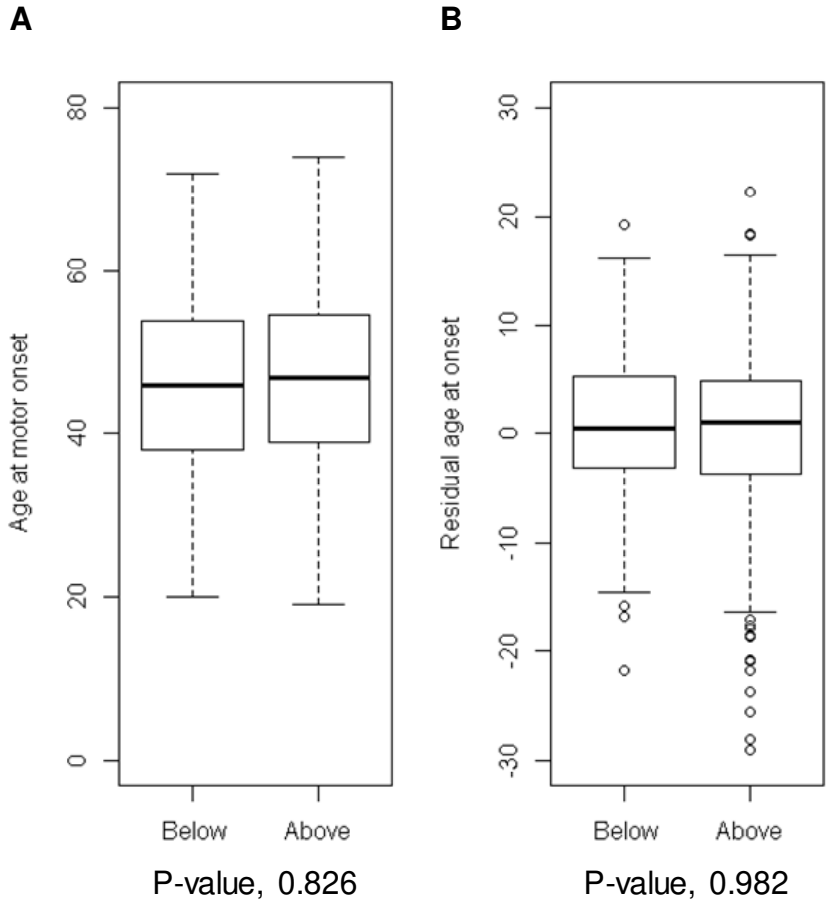
Supplementary Figure 1. Distributions of CAG repeats of polyglutamine disease-causing genes in HD subjects.

Distributions of CAG repeats at both alleles for each polyglutamine disease-causing gene are plotted. The sizes of CAG repeats of other polyglutamine disease-causing genes were determined in 483 to 604 HD individuals (Table 1). For each of other polyglutamine disease gene, mean and standard deviation was used to generate a theoretical normal distribution to plot the patterns of CAG repeat length distribution. The grey background histogram represents *HTT* CAG repeat lengths.



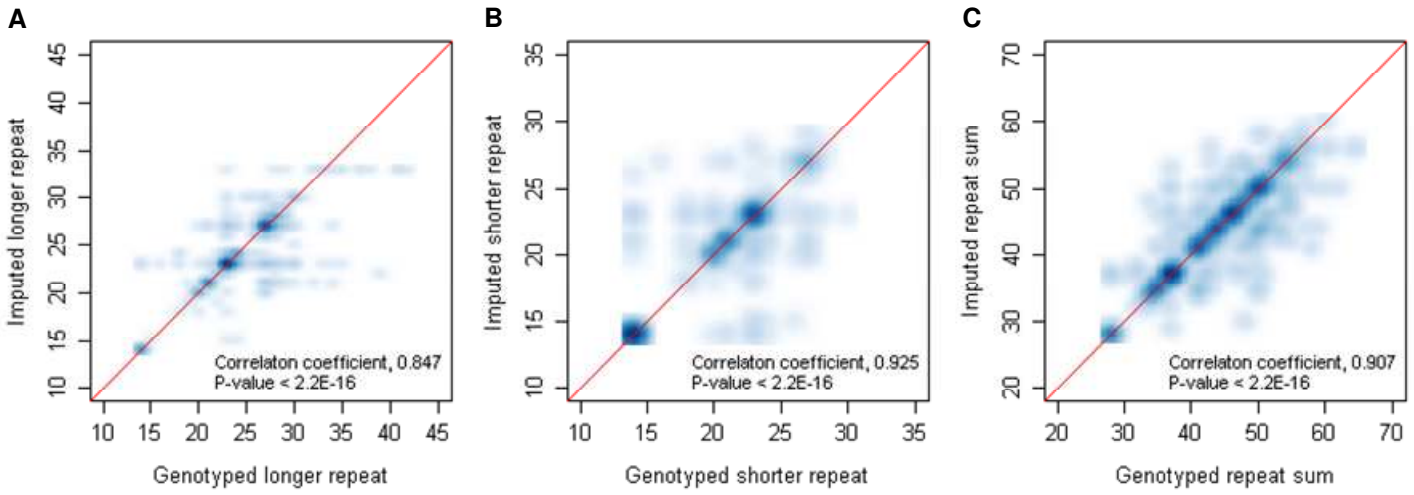
Supplementary Figure 2. Same age-at-onset and residual age-at-onset between HD subjects carrying below and above CAG repeat of *ATXN3*.

HD subjects were grouped based on the median of longer allele of the *ATXN3* CAG repeat (i.e., 23), generating a below the median group (*ATXN3* CAG < 23) and an above the median group (ATXN3 CAG > 23). Subsequently, age-at-onset (A) and residual age-at-onset (B) were compared between two groups using Mann-Whitney *U*-test. *P*-values are shown at the bottom of each plot.



Supplementary Figure 3. Correlation between genotyped and imputed *ATXN3* CAG repeats.

To evaluate the accuracy of STR imputation, we compared *ATXN3* CAG repeats that were experimentally determined and imputed. A total of 1388 samples had both typed and imputed *ATXN3* CAG repeats. We compared longer repeat (A), shorter repeat (B), and the sum of two repeats (C) in imputed data (Y-axis) and typed data (X-axis). We also calculated percentage of identical or similar (difference smaller than 5) were calculated for longer, shorter and sum (D).



D

	Longer repeat	Shorter repeat	Sum
Identical between genotyped and imputed repeats	74.6%	87.2%	69.2%
Difference between genotyped and imputed repeats < 5	93.8%	95.4%	93.8%

Supplementary Figure 4. The lack of association between HD residual age-at-onset and imputed *ATXN3* CAG repeats

Genome-wide STR association analysis used sum of the two alleles for each STR, arguing against modification of HD by *ATXN3* repeat. We further tested whether imputed longer allele (A) or imputed shorter allele (B) is associated with residual age-at-onset. Also, we compared age-at-onset (C) and residual age-at-onset (D) between HD subjects with below and above median repeat length. Respective *P*-values are shown below of each plot.