BRIEF REPORT

Expanded CAG Repeats in *ATXN1*, *ATXN2*, *ATXN3*, and *HTT* in the 1000 Genomes Project

Fulya Akçimen, MSc,[1,2] Jay P. Ross, BSc,[1,2] Calwing Liao, BSc,[1,2] Dan Spiegelman, MSc,[2] Patrick A. Dion, PhD,[2,3] and Guy A. Rouleau, MD, PhD, FRCP(C)[1,2,3*]

[1]*Department of Human Genetics, McGill University, Montréal, Québec, Canada* [2]*Montreal Neurological Institute and Hospital, McGill University, Montréal, Québec, Canada* [3]*Department of Neurology and Neurosurgery, McGill University, Montréal, Québec, Canada*

ABSTRACT: Background: Spinocerebellar ataxia types 1, 2, 3 and Huntington disease are neurodegenerative disorders caused by expanded CAG repeats.
Methods: We performed an in-silico analysis of CAG repeats in *ATXN1*, *ATXN2*, *ATXN3*, and *HTT* using 30× whole-=genome sequencing data of 2504 samples from the 1000 Genomes Project.
Results: Seven *HTT*-positive, 3 *ATXN2*-positive, 1 *ATXN3*-positive, and 6 possibly *ATXN1*-positive samples were identified. No correlation was found between the repeat sizes of the different genes. The distribution of CAG alleles varied by ethnicity.
Conclusion: Our results suggest that there may be asymptomatic small expanded repeats in almost 0.5% of these populations. © 2020 International Parkinson and Movement Disorder Society

Key Words: ataxia; CAG-repeat diseases; *ATXN1*; *ATXN2*; *ATXN3*; *HTT*; 1KGP

***Correspondence to:** Guy A. Rouleau, Montreal Neurological Institute and Hospital, 3801 University Street, Room 636, Montréal, Québec H3A 2B4, Canada; E-mail: guy.rouleau@mcgill.ca

Relevant conflicts of interest/financial disclosures: The authors declare no conflicts of interests. None of the authors have received any funding from any institution, including personal relationships, interests, grants, employment, affiliations, patents, inventions, honoraria, consultancies, royalties, stock options/ownership, or expert testimony for the last 12 months.

Funding agencies: These data were generated at the New York Genome Center with funds provided by NHGRI grant 3UM1HG008901-03S1. F.A. and C.L. are funded by the Fonds de Recherche du Québec–Santé (FRQS). J.P.R. is funded by the Canadian Institutes of Health Research (CIHR; FRN 159279). G.A.R. holds a Canada Research Chair in Genetics of the Nervous System and the Wilder Penfield Chair in Neurosciences.

Received: 29 June 2020; **Revised:** 15 September 2020; **Accepted:** 27 September 2020

Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/mds.28341

Spinocerebellar ataxias (SCAs) and Huntington disease (HD) are rare autosomal-dominant neurodegenerative disorders. SCAs are genetically heterogeneous diseases, of which at least 6 distinct forms are caused by an expanded CAG repeat in a known gene — SCA1 (MIM 164400), SCA2 (MIM 183090), SCA3 (MIM 109150), SCA6 (MIM 183086), SCA7 (MIM 164500), and SCA17 (MIM 607136).[1] Alleles with 40 or more CAG repeats in *HTT* are fully penetrant and cause HD, whereas alleles with repeat size ranging from 36 to 39 are associated with an increasing risk of developing disease with reduced penetrance.[2] Deleterious alleles for the most common SCAs (SCA1, 2, 3) contain more than 45 repeats (or 39 uninterrupted with a CAT codon), 33, and 45 CAG repeats in *ATXN1*, *ATXN2*, and *AXTN3*, respectively.[3-5]

The International Genome Sample Resource (IGSR) curates public data resources that are created by the 1000 Genomes Project (1KGP).[6,7] The 1KGP phase 3 panel consists of 2504 unrelated samples from 26 subpopulations in Africa (AFR, n = 661), East Asia (EAS, n = 504), Europe (EUR, n = 503), South Asia (SAS, n = 489), and America (AMR, n = 347). Donors were older than 18 years and self-declared healthy at the time of collection. The project holds self-reported ethnicity and sex data. No phenotype, medical, or personal identifying information was collected.[6] Previously, various types of structural variants including insertions, deletions, duplications, and copy-number variants were mapped in 1KGP. However, known disease-related short tandem repeats (STRs) have not been reported in this data set.[8] In 2019, the New York Genome Center resequenced the samples in the final phase of 1KGP. High-coverage polymerase chain reaction (PCR)-free whole-genome sequencing (WGS) data of a total of 2504 samples from 26 populations were added.[7]

ExpansionHunter is software that can estimate sizes of targeted STRs from PCR-free WGS data.[9,10] It identifies lengths of the repeats using either spanning, flanking, or in-repeat reads. Therefore, it enabled us to employ in silico analysis of CAG-repeat expansions in HD and the most common SCAs using high-coverage WGS data among different ancestries from IGSR.[6] We hypothesized that samples in a reference data set such as 1KGP might carry repeat alleles associated with neurological diseases. Confirming this hypothesis would have implications for neurological studies that use these samples for genetic reference.

TABLE 1. Disease-associated CAG-repeat expansions (longest allele) in samples from the 1KGP

Sample ID	Sex	Population	Gene/disease	Associated repeat size	CAG repeat size
NA11931	F	CEU	*HTT*/HD	≥40, ≥36 (incomplete penetrance [IC])	17/52
NA20540	F	TSI			18/36 (IC)
HG02275	F	PEL			./42
HG02470	M	ACB			15/41
NA18522	M	YRI			./40
HG02727	M	PJL			10/36 (IC)
NA19466	M	LWK			17/39 (IC)
HG00148	M	GBR	*ATXN1*/SCA1	≥39 (uninterrupted) or ≥45	31/42
HG00122	F	GBR			./39
HG03575	F	MSL			./44
HG03615	M	BEB			28/39
HG03871	M	ITU			29/39
HG03352	M	ESN			33/39
HG01708	M	IBS	*ATXN2*/SCA2	≥33	22/34
HG04140	M	BEB			22/36
NA18625	F	CHB			22/34
HG02323	M	ACB	*ATXN3*/SCA3	≥45	27/45

Methods

NovaSeq (Illumina, Inc.) WGS sequencing and alignment to the GRCh38 reference genome were generated by the New York Genome Center. Alignment files (CRAM) of 2504 PCR-free WGS samples of 26 populations from 5 superpopulations (AFR, African; AMR, ad mixed American; EAS, East Asian; EUR, European; and SAS, South Asian) were downloaded from the IGSR website (https://www.internationalgenome.org/data-portal/data-collection/30x-grch38). Phenotype information was not available for the samples apart from sex and ethnicity. Individuals were older than 18 years and declared themselves to be healthy at the time of the collection.

Alignment files were indexed using SAMtools v.1.10.[11] Allele lengths of *ATXN1*, *ATXN2*, *ATXN3*, and *HTT* were estimated using ExpansionHunter v3.2.0 and its published variant catalog file containing the respective genomic loci.[9,10] Violin plots representing the distributions of CAG-repeat sizes in different populations were plotted in R v.3.5.1 using ggplot2.[12] CAG-repeat length (longest allele) for each gene was modeled by linear regression as a function of population and CAG-repeat lengths in the other genes.

Results

Using ExpansionHunter, CAG-repeat lengths were successfully estimated in 2486 samples for *HTT*, 2390 samples for *ATXN1*, 2408 samples for *ATXN2*, and 2339 samples for *ATXN3*. Mean CAG-repeat lengths identified in each population are shown in Table SS1. The full results for all samples are listed in Table SS2. Expanded CAG repeats associated with diseases were detected in a total of 11 samples (*HTT* in 7, *ATXN2* in 3, and *ATXN3* in 1). No pathogenic *ATXN1* expansions that have a repeat size greater than 45 were found. However, intermediate expansions (39–44 CAG repeats) that can be in the disease-associated range in *ATXN1* were identified in 6 samples. However, these could be associated with the disease only in the absence of CAT trinucleotide interruptions. Because interruptions were not tested in the current study, the deleterious effect of the identified *ATXN1* repeat expansions is uncertain. Detailed information of the positive samples is shown in Table 1. The CAG repeats in the examined genes were not correlated to each other ($P_{ATXN1\text{-}HTT}$ = 0.82, $P_{ATXN1–ATXN2}$ = 0.06, $P_{ATXN1–ATXN3}$ = 0.67, $P_{ATXN2–HTT}$ = 0.27, $P_{ATXN2–ATXN3}$ = 0.76, $P_{ATXN3–HTT}$ = 0.27).

Distribution of repeat expansion sizes for each gene across different ancestries within 1KGP are shown in Figure 1. Different ethnicities explained some of the variability in the CAG-repeat distributions for *ATXN3* (coefficient of determination R^2 = 0.16, analysis of variance [ANOVA] $P < 2.2 \times 10^{-16}$), *ATXN1* (coefficient of determination R^2 = 0.09, ANOVA $P < 2.2 \times 10^{-16}$), and *HTT* (coefficient of determination R^2 = 0.03, ANOVA $P = 2.77 \times 10^{-16}$). There was no difference in the means of *ATXN2* among populations (coefficient of determination R^2 = 0.0019, ANOVA P = 0.18).

Discussion

This study represents an in silico analysis of CAG-repeat expansions of the 2504 samples from 1KGP. Through leveraging public high-coverage sequencing data as well as the available STR genotyping approach, ExpansionHunter, we sought to examine the CAG repeats associated with SCA1, 2, 3, and HD in

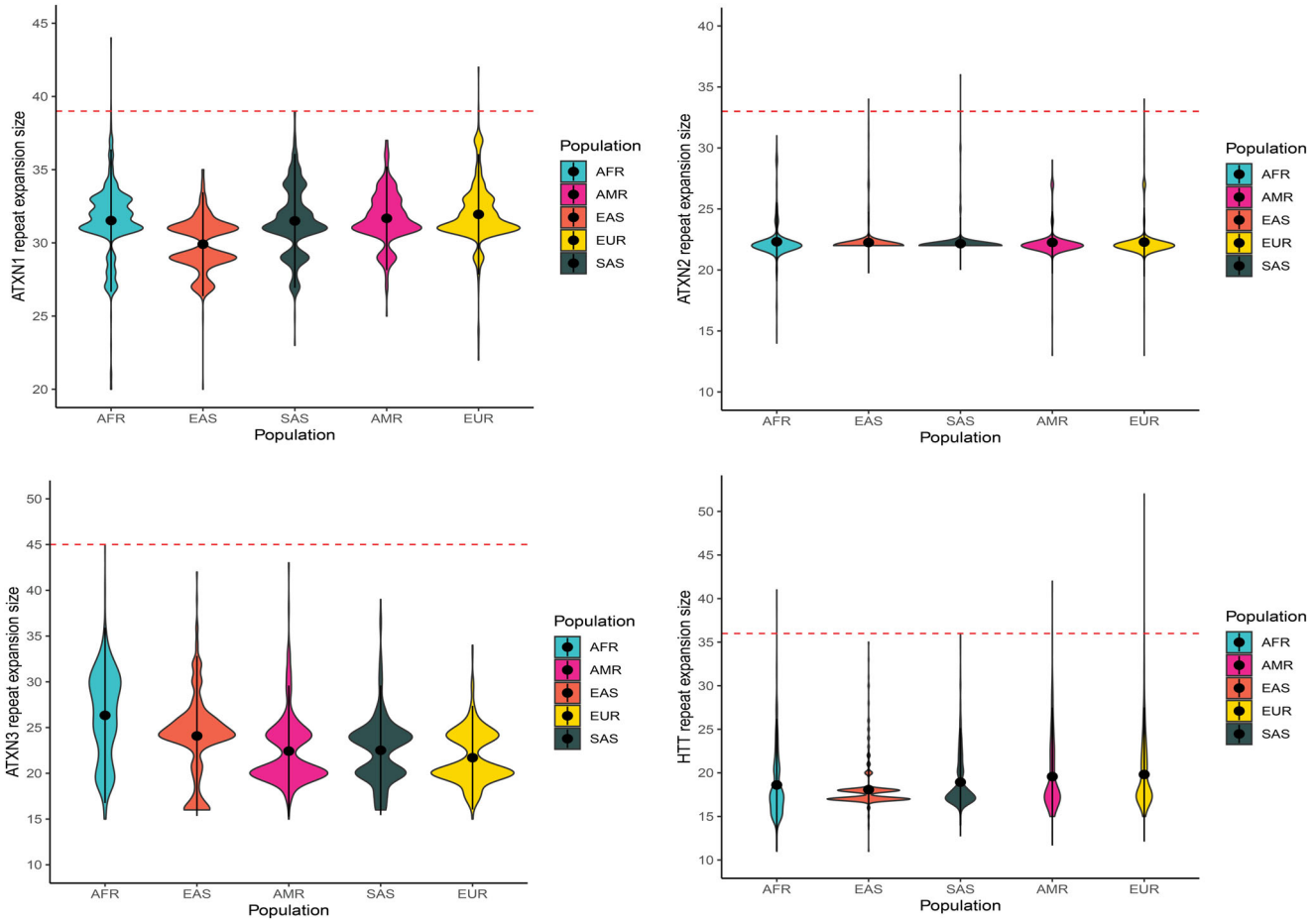

FIG. 1. Distribution of repeat expansion sizes among different ethnic groups in the 1KGP. Red line indicates the threshold for causality. [Color figure can be viewed at wileyonlinelibrary.com]

populations from different ethnicities. Although the participants declared themselves to be healthy at the time of the collection, repeats in the disease-associated range were found in at least 11 (plus 6 possibly *ATXN1*-positive) samples. We were not able to validate these findings, but if accurate, these individuals may develop the associated diseases later in life. It is interesting to note that almost all the expansions in these individuals were relatively small, close to the normal range. Most of the expansions in these 11 individuals would normally be associated with later age of onset and milder disease. This might partially explain why they were asymptomatic at the time of ascertainment. In addition, these individuals may not even have known that these diseases were in their family because relatives in their parents' generation would likely have smaller repeats, either in the disease-associated range but with a late onset or in the intermediate or high-normal range with an expansion creating a new disease allele in the individual.

The mean CAG-repeat sizes in *HTT*, *ATXN1*, and *ATXN3* varied in the populations from different ancestries. Consistent with previous studies,[13] lower mean *HTT* CAG-repeat size was observed in the samples with East Asian ancestry, which correlated with lower prevalence of HD in these populations. This pattern was also observed in European populations that have longer *HTT* CAG repeats and higher prevalence estimates of HD.[13] In addition, a skewed distribution of *ATXN3* CAG alleles toward intermediate-size repeats in African and East Asian populations was detected. Higher frequency of intermediate alleles was shown to be enriched in populations with higher prevalence of repeat expansion diseases, strengthening the hypothesis of the repeats' instability and expansion into the disease-causing range from the high-normal or intermediate-size alleles as a cause of CAG-related diseases.[13-16]

CAG-repeat lengths in one gene were not found to be correlated with repeat lengths in any another gene in this study. Various studies have been performed to identify modifiers in CAG-repeat diseases.[17–19] Genetic variants implicated in DNA repair mechanisms that possibly influence somatic expansions were identified as candidate genetic modifiers of the diseases.[17–19] Although common variants and mechanisms are implicated in somatic expansions of CAG repeats in the

respective genes, our findings suggest that germline instability occurs independently in each CAG repeat that could implicate unique mutational mechanisms.

Although of interest and original, our study has some limitations. The average sample size of subpopulations was 96. Hence, it may not be sufficient to assess the frequencies of disease-associated STRs in subpopulations. Furthermore, although the final phase of 1KGP expanded its population diversity, the current data set did not represent all populations.[7] Therefore, the addition of further samples as well as populations could improve the generalizability of our results. Although the results from ExpansionHunter were previously successfully validated by repeat-primed PCR (with overall sensitivity and specificity of 98.6% and 99.6%, respectively),[9] DNA was not available to replicate the pathogenic-size expansions identified in 1KGP samples. Another limitation is that the repeat interruptions, such as CAA interruptions in HD or CAT interruptions in SCA1, are not estimated by ExpansionHunter. The presence of interruptions, which would determine the pathogenicity of alleles in the size range between 36 and 44 repeats in *ATXN1*, were not reported. Alleles with CAT interruptions in the range of 36 to 44 repeats are considered normal. Alleles that are not interrupted by CAT repeats are associated with symptoms (≥39 repeats) or in the mutable normal range (36–38 repeats).[20] Therefore, the identified *ATXN1* repeat expansions in 6 samples may not be associated with SCA1. However, alleles in the mutable normal range may expand beyond the normal range during transmission to offspring who may manifest the disease.[20] Furthermore, HD and CAG-associated SCAs usually occur in the third or fifth decade. However, age at onset for these diseases is highly variable.[1,14] Although HD is a adult-onset neurological disorder, its symptoms can appear as early as age 18 or as late as age 80.[2,14] Similarly, the average age at onset is 38 years, ranging between 10 and 70 years in SCA3.[18] In a panel study for dominant cerebellar ataxias, the average age at onset was 40.9 years in known CAG-associated SCAs.[21] Therefore, the positive individuals might be asymptomatic at the time of collection, as these diseases usually occur in the third or fifth decade. However, because of the anonymity of the samples, no personal information including individual age was collected in 1KGP. Therefore, we were unable to infer if the positive samples were too young for symptoms.

Overall, in this study we provide the distribution of CAG repeats associated with SCA1, 2, 3, and HD in a large number of people in 26 different populations from 1KGP. These data can be useful for understanding the population distribution of these repeats in different populations. Furthermore, pathogenic-length repeats in 11 samples were observed. This suggests that the data sets generated from the general populations might contain samples positive for late-onset diseases, even though the individuals in these samples declared themselves healthy at the time of collection. Inclusion of 1KGP in future studies, especially in variant frequency assessments for rare diseases, should be done with caution. ■

Acknowledgments: The authors thank the 1000 Genomes Project Consortium and all the other projects that have supplied data incorporated into IGSR. These data were generated at the New York Genome Center with funds provided by NHGRI grant 3UM1HG008901-03S1. F. A. and C.L. are funded by the Fonds de Recherche du Québec–Santé (FRQS). J.P.R. is funded by the Canadian Institutes of Health Research (CIHR; FRN 159279). G.A.R. holds a Canada Research Chair in Genetics of the Nervous System and the Wilder Penfield Chair in Neurosciences.

References

1. Klockgether T, Mariotti C, Paulson HL. Spinocerebellar ataxia. Nat Rev Dis Primers 2019;5(1):24.
2. Bates GP. The molecular genetics of Huntington disease — a history. Nat Rev Genet 2005;6(10):766–773.
3. Zuhlke C, Dalski A, Hellenbroich Y, et al. Spinocerebellar ataxia type 1 (SCA1): phenotype-genotype correlation studies in intermediate alleles. Eur J Hum Genet 2002;10(3):204–209.
4. Fernandez M, McClain ME, Martinez RA, et al. Late-onset SCA2: 33 CAG repeats are sufficient to cause disease. Neurology 2000;55 (4):569–572.
5. Bettencourt C, Lima M. Machado-Joseph disease: from first descriptions to new perspectives. Orphanet J Rare Dis 2011;6:35.
6. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. Nature 2015;526(7571):68–74.
7. Fairley S, Lowy-Gallego E, Perry E, et al. The international genome sample resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res 2019;48(D1):D941–D947.
8. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. Nature 2015;526 (7571):75–81.
9. Dolzhenko E, van Vugt J, Shaw RJ, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res 2017;27(11):1895–1903.
10. Dolzhenko E, Deshpande V, Schlesinger F, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics 2019;35(22):4754–4756.
11. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25(16):2078–2079.
12. Wickham H. ggplot2: elegant graphics for data analysis. Verlag New York: Springer; 2016.
13. Kay C, Fisher E, Hayden MR. Huntington's Disease. Epidemiology. Oxford: Oxford University Press; 2014.
14. Budworth H, McMurray CT. A brief history of triplet repeat diseases. Methods Mol Biol 2013;1010:3–17.
15. Martins S, Calafell F, Gaspar C, et al. Asian origin for the worldwide-spread mutational event in Machado-Joseph disease. Arch Neurol 2007;64(10):1502–1508.
16. Friedman JE. Anticipation in hereditary disease: the history of a biomedical concept. Hum Genet 2011;130(6):705–714.
17. Bettencourt C, Hensman-Moss D, Flower M, et al. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. Ann Neurol 2016;79(6):983–990.
18. Akcimen F, Martins S, Liao C, et al. Genome-wide association study identifies genetic factors that modify age at onset in Machado-Joseph disease. Aging 2020;12(6):4742–4756.
19. Lee J-M, Correia K, Loupe J, et al. CAG repeat not Polyglutamine length determines timing of Huntington's disease onset. Cell 2019; 178(4):887–900. e14.

20. Opal P, Ashizawa T. Spinocerebellar ataxia type 1. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle, WA: University of Washington, Seattle; 1993–2020. https://www.ncbi.nlm.nih.gov/books/NBK1184/

21. Coutelier M, Coarelli G, Monin M, et al. A panel study on patients with dominant cerebellar ataxia highlights the frequency of channelopathies. Brain 2017;140(6):1579–1594.

Supporting Data

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.

SGML and CITI Use Only
DO NOT PRINT

Authors' Roles

F.A.: 1A, 1B, 1C, 2A, 2B, 2C, 3A.
J.P.R.:2C, 3B.
C.L.: 2C, 3B.
D.S.: 1B, 2B.
P.A.D.:1A, 1B, 2A, 2C, 3B.
G.A.R.: 1A, 1B, 2A, 2C, 3B.