A multifaceted approach to repeat-associated hereditary ataxia

Fulya Akçimen

Department of Human Genetics

Faculty of Medicine and Health Sciences

McGill University, Montréal

December 2021

A thesis submitted to McGill University in partial fulfillment

of the requirements of the degree of Doctor of Philosophy

Abstract

Hereditary ataxias form a heterogeneous group of disorders characterized by an incoordination of gait, speech, and hand movements. Classification of each subtype is based on the mode of inheritance (i.e., autosomal dominant, autosomal recessive or X-linked) and the associated genetic risk factor. Known ataxia-associated risk factors encompass a range of genomic variations: single nucleotide variants and structural variants (e.g., deletions and insertions), as well as expansions of repetitive sequences (e.g., CAG trinucleotides repeats). The work presented in this thesis focuses on the prevalence, and polyvalent implication, of expanded repeats in different cohorts of ataxia cases and control individuals. We performed an *in silico* tandem repeat genotyping approach to examine CAG repeat expansions associated with spinocerebellar ataxia types 1, 2, 3 (SCA1, 2, 3) using the whole-genome sequencing data of 2,504 samples and the 1000 Genomes Project. This study identified positive and/or asymptomatic individuals and provided the distribution of repeat length across different populations. In addition, we examined the prevalence of a recently identified *RFC1* repeat expansion in Brazilian and Canadian cohorts of ataxia cases. To the best of our knowledge, it was the first follow-up analysis of *RFC1* repeat expansion in unrelated cohorts of ataxia cases. *RFC1* repeat expansions are the recently identified cause of recessive Cerebellar Ataxia, Neuropathy, Vestibular Areflexia Syndrome (CANVAS). Contrary to the original report, the pathogenic *RFC1* repeat expansion explained only a few cases of the Brazilian and Canadian cohorts. However, we also identified two previously unreported *RFC1* repeat motifs. Furthermore, we carried out a genome-wide association study to search for possible genetic modifiers of age at onset in SCA3. Using patients from five different geographical origins, we demonstrated that along with the associated repeat length, there are additional genetic variants that could explain the variability in age at onset. Overall, this thesis comprises different approaches

to understand the nature and prevalence of some of the hereditary ataxia-associated repeats and to examine the implications of additional genetic variants in SCA3, one of the most common repeat-associated subtypes of hereditary ataxia.

Résumé

Les ataxies héréditaires forment un groupe hétérogène de troubles caractérisés par un manque de coordination de la marche, de la parole ainsi que des mouvements de la main. La classification de chaque sous-type d'ataxie se fait en fonction du mode de transmission (c.-à-d. autosomique dominant, autosomique récessif ou bien une transmission liée au chromosome X) et des facteurs de risque génétique associés. À ce jour, différents facteurs génétiques sont associés aux ataxies: variations d'un seul nucléotide et variations structurelles (ex. délétions et insertions), ainsi que des expansions de séquences répétées (ex. trinucléotides de type CAG). Les travaux menés dans cette thèse portent sur la prévalence et sur l'implication polyvalente des répétitions élargies dans diverses cohortes d'ataxie. À l'aide d'une approche de génotypage de répétitions en tandem *in silico*, nous examinons les expansions CAG associées aux ataxies spinocérébelleuse de type 1, 2, et 3 (SCA1, 2, 3) en utilisant des données de séquençage du génome entier de 2,504 échantillons faisant partie du projet 1000 génomes. D'une part cette étude a identifié des individus positifs et/ou asymptomatique, ainsi que la distribution de la taille des répétitions dans différentes populations. Nous avons également évalué la prévalence d'une variation d'expansion répétée de *RFC1*, récemment identifiée, dans une cohorte brésilienne et canadienne de cas d'ataxie. Des expansions de séquences répétées ont récemment été identifié comme le facteur de risque génétique pour une forme d'ataxie récéssive (CANVAS : *Cerebellar Ataxia, Neuropathy, Vestibular Areflexia Syndrome*). À notre connaissance, il s'agissait de la première étude de réplication pour cette expansion de *RFC1* pour donner suite à la publication originale à ce sujet. Nous avons établi qu'en dépit des observations initiales, l'expansion pathogénique de *RFC1* n'expliquerait que quelques cas dans les cohortes brésiliennes et canadiennes que nous avons examinées. Nous avons également documenté pour la première fois, deux motifs supplémentaires

de répétition au niveau de *RFC1*. Enfin, nous avons mené une étude d'association pangénomique pour identifier de modificateurs génétiques liés à l'âge d'apparition de l'ataxie spinocérébelleuse de type 3. En utilisant des patients de cinq origines géographiques différentes, nous avons démontré qu'en plus de la taille des répétitions associées, il existe des variantes génétiques supplémentaires qui pourraient expliquer la variabilité de l'âge d'apparition. Dans l'ensemble, cette thèse comprend des différentes approches pour comprendre la nature ainsi que la prévalence de certaines répétitions associées à l'ataxie héréditaire et a permis d'examiner les implications de variantes génétiques supplémentaires dans l'un des sous-types d'ataxie le plus courant, l'ataxie spinocérébelleuse de type 3.

Table of Contents

List of Abbreviations

Abbreviation	Definition
1KGP	1000 Genomes Project
ACB	African Caribbean in Barbados
ADCA	Autosomal dominant cerebellar ataxia
AFR	Africans
ALS	Amyotrophic lateral sclerosis
AMR	Americans
AO	Ages at onset
AOA-1	Ataxia with oculomotor apraxia type 1
AOA-2	Ataxia with oculomotor apraxia type 2
APOE	Apolipoprotein E
ARCA-1	Autosomal recessive cerebellar ataxia type 1
ARCA-2	Autosomal recessive cerebellar ataxia type 2
ATN1	Atrophin-1
ATXN1	Ataxin-1
ATXN2	Ataxin-2
ATXN3	Ataxin-3
ATXN7	Ataxin-7
BEB	Bengali in Bangladesh
C9orf72	Chromosome 9 open reading frame 72
CACNA1A	Calcium voltage-gated channel subunit alpha1 A
CACNB4	Calcium voltage-gated channel auxiliary subunit beta 4
CADD	Combined Annotation Dependent Depletion
$(CAG)_{exp}$	Expanded CAG repeat
$(CAG)_{nor}$	Normal CAG repeat
CANVAS	Cerebellar ataxia, neuropathy and vestibular areflexia syndrome
CDK7	Cyclin-Dependent Kinase 7
CEU	Utah residents with Northern and Western European ancestry
CHB	Han Chinese in Beijing China

Chr	Chromosome
CIHR	Canadian Institutes of Health Research
cM	Centimorgan
CTCF	CCCTC-Binding Factor
DRPLA	Dentatorubral-pallidoluysian atrophy
EAS	East Asians
EAS	Episodic ataxia
ERCC6	ERCC excision repair 6, chromatin remodeling factor
ESN	Esan in Nigeria
EUR	Europeans
FAN1	Fanconi anemia complementation group D2 and Fanconi anemia complementation group I associated nuclease 1
FDR	False discovery rate
FRDA	Friedreich ataxia
FTD	Frontotemporal dementia
FUMA	Functional mapping and annotation of genetic associations
FXN	Frataxin
FXTAS	Fragile X tremor ataxia syndrome
GBR	British from England and Scotland
GCTA	Genome-wide complex trait analysis
GeCIP	Genomics England Clinical Interpretation Partnership
GeM-HD	Genetic Modifiers of Huntington's Disease
GO	Gene ontology
GSEA	Gene set enrichment analysis
GWAS	Genome-wide association study
HD	Huntington's disease
hg19	Human genome 19
HRC	Haplotype Reference Consortium
HSP40	Heat shock protein family
HTT	Huntingtin
HWE	Hardy-Weinberg Equilibrium

IBS	Iberian populations in Spain
IC	Incomplete penetrance
IGSR	The International Genome Sample Resource
ITU	Indian Telugu in the United Kingdom
KCNA1	Potassium Voltage-Gated Channel Subfamily A Member 1
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LD	Linkage disequilibrium
LWK	Luhya in Webuye Kenya
MAF	Minor allele frequency
Mb	Megabase
MIM	Mendelian inheritance in man
MJD	Machado-Joseph disease
MLH1	MutL homolog 1
MSL	Mende in Sierra Leone
NGS	Next generation sequencing
NOP56	Nucleolar protein 56
PASCAL	Pathway Scoring Algorithm
PC	Principal components
PCR	Polymerase chain reaction
PEL	Peruvian in Lima Peru
PJL	Punjabi in Lahore Pakistan
QC	Quality control
RAG	Recombination-activating gene
RAN	Repeat associated non-ATG translation
RFC1	Replication factor C subunit 1
RPA	Replication protein A
RPPCR	Repeat-primed PCR
SACS	Sacsin
SARA	Scale for the Assessment and Rating of Ataxia
SAS	South Asians
SBMA	Spinal and bulbar muscular atrophy

SCA	Spinocerebellar ataxia
SD	Standard deviation
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SPAX	Spastic ataxia
SPG7	Spastic paraplegia type 7
STR	Short-tandem repeat
TRIM29	Tripartite Motif Containing 29
TSI	Toscani in Italia
VEGAS	Versatile Gene-based Association Study
WGS	Whole-genome sequencing
YRI	Yoruba in Ibadan Nigeria

List of Figures

List of Supplementary Figures

Chapter 3

Chapter 4

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Supplementary Tables

Chapter 2

Chapter 3

Chapter 4

Acknowledgements

I would like to start with thanking my supervisors Drs. Guy Rouleau and Patrick Dion. Thank you for giving me the chance to join your team, for your guidance, support and providing me with a tremendous amount of independence. It has been always an honor to work with you.

Next, I would like to thank Jay, the best lab mate ever! Thanks to you, I have never felt alone since the day I joined the lab. You have always been a great example not only as a scientist but also as a wonderful person. I would also like to thank Cal for always being a super motivating friend. You have pushed me to do much better in a lot of projects. Both of you have always been very supportive. Our CFJ group means a lot to me, and I hope we will stay in touch forever.

Cynthia, I would like to thank you for your friendship and kindness. Thank you again for being very patient and nice as well as for making my adaptation smoother when I first joined the lab.

Faezeh and Zoe, thank you for being such good friends and colleagues. I have really enjoyed all the moments and laughter we had!

I would also like to thank past and present Rouleau lab members. Patrick, thank you again for not only being a great mentor but also a friend. I am grateful to all of you, Hélène, Vessela, Gabby, Rachel, Paria, Parizad, and Mehrdad. Your enthusiasm and support made the lab a very peaceful and motivating environment to work in.

I also would like to thank our student affairs administrator and advisor at the Department of Human Genetics; Ross MacKay and Rimi Joshi for making all the administrative procedures fast and smooth with their extraordinary support.

Son olarak aileme, anneme, ablam Funda ve kardeşim Can’a, anneannem ve dedeme, doktoram süresince gösterdikleri sınırsız destekleri, anlayışları, sağladıkları moral ve motivasyon için çok teşekkür ederim.

Cancığım, her anını gülümseyerek hatırlayacağım dört yıllık Montreal maceramız yakında bitiyor. Bir paragrafla anlatmak ne kadar yetersiz olacaksa da, paylaştığımız her şey için sana çok teşekkür ederim. Senin desteğin olmasa sanırım tamamladığımız hiçbir şeye başlamazdım bile. Yakında başka bir şehirde, başka bir maceraya başlıyoruz. O zaman bu sadece bir teşekkür değil, yeni başlangıçlarımıza bir merhaba olsun.

Contribution to Original Knowledge

The work presented in this thesis provides the following distinct and original contributions to knowledge:

Chapter 2 provides the distribution of CAG repeat length that are associated with four neurological diseases across different populations in a public dataset. It presents evidence that 1000 Genomes Project (1KGP) representing self-declared healthy individuals from a general population may contain positive individuals for adult-onset diseases. Through providing samples that are positive for disease associated repeats, this study serves as cautionary tale for the usage of 1KGP dataset.

To our knowledge, Chapter 3 is the first follow-up analysis of the *RFC1* repeat expansions that was identified as a common cause of late-onset ataxia recently. We reported a lower prevalence of disease-causing biallelic *RFC1* repeat expansion in Brazilian and Canadian cohorts. In addition, we identified two novel repeat configurations, expanding our knowledge on the dynamic structure of *RFC1* repeats.

Chapter 4 presents a genome-wide association study suggesting genetic factors that can modify age at onset in spinocerebellar ataxia type 3. It describes how we analysed the relationship between disease-associated repeat length and age at onset, and identified additional modifiers associated with age at onset variability.

Format of the Thesis

The work described in this thesis was performed under the co-supervision of Dr. Guy Rouleau and Dr. Patrick Dion. It is a manuscript-based thesis and follows the Thesis Preparation Guidelines by the Department of Graduate and Postdoctoral Studies. This thesis contains seven chapters. Chapter 1 consists of a general introduction with an overall rationale, objectives, and hypotheses. Chapter 2, 3, and 4 have been published in *Movement Disorders*, *Frontiers in Genetics*, and *Aging*. A bridging statement is included between the manuscripts. Chapter 5 presents a general discussion, Chapter 6 consists of future directions and conclusion. For the manuscripts, references are included at the end of each chapter (Chapters 2, 3, and 4). The master reference list contains references for citations in Chapters 1, 5 and 6.

Contribution of Authors

Chapter 2 is a manuscript authored by Fulya Akçimen, Jay P. Ross, Calwing Liao, Dan Spiegelman, Patrick A. Dion, and Guy A. Rouleau. It was published in *Movement Disorders* on November 11th, 2020. FA performed all analyses and wrote the first draft of the manuscript. DS installed the software for the analyses. JPR and CL revised the manuscript for intellectual content. PAD and GAR oversaw the manuscript.

Chapter 3 is a manuscript authored by Fulya Akçimen, Jay P. Ross, Cynthia V. Bourassa, Calwing Liao, Daniel Rochefort, Maria Thereza Drumond Gama, Marie-Josée Dicaire, Orlando G. Barsottini, Bernard Brais, José Luiz Pedroso, Patrick A. Dion, and Guy A. Rouleau. It was published in Frontiers in Genetics on November 22nd, 2019. FA performed all analyses and wrote the first draft of the manuscript. CVB and DR contributed to analysis and interpretation of the data and revised the manuscript. JPR and CL revised the manuscript for intellectual content. MG, M-JD, OB, BB, and JP contributed to the acquisition of data and revising the manuscript for intellectual content. PAD and GAR contributed to the design of the study and writing and revising the manuscript.

Chapter 4 is a manuscript authored by Fulya Akçimen, Sandra Martins, Calwing Liao, Cynthia V. Bourassa, Hélène Catoire, Garth A. Nicholson, Olaf Riess, Mafalda Raposo, Marcondes C. França Jr., João Vasconcelos, Manuela Lima, Iscia Lopes-Cendes, Maria Luiza Saraiva-Pereira, Laura B. Jardim, Jorge Sequeiros, Patrick A. Dion, and Guy A. Rouleau. It was published in Aging on March 23rd, 2020. FA performed all analyses and wrote the first draft of the manuscript. SM, CVB, and HC contributed to the analysis and interpretation of the data. CL and JPR revised the manuscript for intellectual content. SM, GAN, OR, MR, MCF, JV, ML, IL, MLS, LBJ, and JS contributed to the acquisition of data and revising the manuscript for intellectual

content. PAD and GAR oversaw the manuscript, contributed to the design and conceptualized the study; interpretation of the data; and drafting the manuscript.

I, Fulya Akçimen, have read, understood, and abided by all norms and regulations of academic integrity of McGill University.

CHAPTER 1: GENERAL INTRODUCTION

1.1. Overview of hereditary ataxias

Ataxias are a group of neurodegenerative disorders characterized by loss of balance, incoordination of gait, and slurred speech. These symptoms and signs are often associated with damage in the cerebellum, the brain region where the coordination of movement is initiated (Jayadev *et al*., 2013; Bird, 1998). Hereditary forms of ataxia should be distinguished from non-genetic (acquired) causes of ataxia by a positive family history, molecular genetic testing, and clinical examination. Acquired ataxias can be immune-mediated or associated with alcoholism, infections, or brain tumors (Jayadev *et al*., 2013).

Diagnosis of hereditary ataxia is established on the basis of a clinical presentation (lack of gait and hand coordination, usually associated with dysarthria and nystagmus), a positive family history for ataxia, and absence of any evidence to support an acquired cause (Bird, 1998, Klockgether *et al*., 2019). Different impairments in ataxia are examined by the Scale for the Assessment and Rating of Ataxia (SARA) which includes eight items reflecting neurologic manifestations of cerebellar ataxia with a total score of 0 in case of no ataxia and up to 40 for the most severe ataxia cases. These items are related to gait, stance, sitting, speech disturbance, finger chase test, nose finger test, fast alternating hand movements, and heel-shin slide (Schmitz-Hübsch *et al*., 2006). Although SARA was originally developed for only a single group of ataxias, which was spinocerebellar ataxia (SCA), it was later validated as a reliable measure in non-SCA ataxia patients (Weyer *et al*., 2007).

Hereditary ataxias can be classified by the mode of inheritance as autosomal dominant, autosomal recessive, X-linked, and mitochondrial. Although each main group has several subtypes

with a unique underlying genetic background, most subtypes of ataxias have overlapping symptoms.

1.1.1. Autosomal dominant cerebellar ataxias

Autosomal dominant cerebellar ataxias (ADCAs) include SCAs, episodic ataxias (EAs), and spastic ataxia type 1, all of which are transmitted vertically from one generation to the next within a family. The prevalence of ADCAs is estimated to be approximately 1-5:100,000, whereas it can be higher in isolated populations due to possible founder effects (van de Warrenburg *et al*., 2002; Ruano *et al*., 2014; Schols *et al*., 2004). Although age at onset can be variable between different types of ADCAs, progressive gait ataxia and dysarthria in adulthood are the most common signs (Bird, 1998). Physical findings of the ADCAs highly overlap; nevertheless, there are key distinguishing features specific for each subtype (Table 1).

1.1.1.1. Spinocerebellar ataxia

There are more than 40 different subtypes of SCAs. They are classified by either causal genes or chromosomal locations if the related genes are unknown (Muller, 2021). Each subtype is named “SCA” followed by a number that represents the chronological order in which the disease locus was identified, except for a more complicated form, dentatorubral-pallidoluysian atrophy (DRPLA) (Klockgether *et al.*, 2019).

The prevalence of individual subtypes of SCAs varies across regions, usually due to founder effects. The average worldwide prevalence of SCA is estimated to be 2.7:100,000, ranging from 0 to 5.6 cases per 100,000 individuals (Ruano *et al*., 2014). A founder population is due to geographic isolation from outside populations and occurs when a subgroup of a larger population

is prevented from reproducing with the larger population. Therefore, it can explain the increased prevalence of some hereditary diseases in some populations (Kivisild *et al*., 2013). For example, SCA3, which is the most common ADCA worldwide, compassing 20-50 % of families with SCA, has a higher incidence in a small area of the Tagus River Valley (1:1,000) and has the highest worldwide prevalence (1:239) in Flores Island, Portugal (Klockgether *et al*., 2019; Bettencourt *et al*., 2008; Bettencourt *et al*., 2011). Similarly, SCA2 is the most common ataxia subtype in Cuba, especially in Holguin province, where a frequency of 40 cases per 100,000 individuals was estimated for people of Spanish ancestry due to a possible founder effect (Orozco Diaz *et al*., 1990). Furthermore, an irregular countrywide distribution of SCA1 pedigrees, a haplotype association with a specific *ATXN1* variant as well as high SCA1 concentration in Central Poland suggested a possible founder effect (Krysa *et al*., 2016).

SCAs can be categorized into two major groups based on their genetic background: those associated with repeat expansion variants or those caused by conventional variants including single nucleotide variants (SNVs), small insertions, or deletions. On the other hand, there are still a number of SCA subtypes in which the potential loci were identified but the causative genetic variants in those loci have yet to be identified (Jayadev *et al*., 2013; Marelli *et al*., 2011).

Repeat-associated SCAs

Most SCAs are caused by repeat expansions in either coding or noncoding regions of the genome (Hersheson *et al*., 2012). Among these, CAG trinucleotide repeat expansions in various genes are the most common causes of SCAs (Klockgether *et al*., 2019). For SCA types 1, 2, 3, 6, 7, 17, and DRPLA, CAG repeat expansions occur in a protein-coding region of the gene. As these expansions encode a polyglutamine repeat, these SCAs are categorized into polyglutamine diseases, together with Huntington's disease (HD) and spinal and bulbar muscular atrophy (SBMA). In addition, several expanded repeats in noncoding regions cause SCA8[$(CTG)_{exp}$], SCA10[$(ATTCT)_{exp}$], SCA12[$(CAG)_{exp}$], SCA31[$(TGGAA)_{exp}$], SCA36[$(GGCCTG)_{exp}$] and SCA37[$(ATTTC)_{exp}$] (Table 1) (Figure 1) (Hersheson *et al*., 2012; Paulson, 2018; Seixas *et al*., 2017).

The length of the normal repeat allele and disease-causing expanded allele vary among diseases, which usually falls into four ranges in each SCA subtypes: wild type, mutable normal (intermediate), reduced-penetrance, and full-penetrance alleles. Mutable normal alleles are longer than wild type alleles and not disease-causative. However, they can expand on transmission and turn into disease-causing expansions, which increases the disease risk in subsequent generations. Reduced-penetrance alleles are longer than mutable normal alleles and may or may not cause the disease in an individual. Therefore, the causality of mutable normal and reduced-penetrance alleles should be interpreted with caution and in consultation with genetic testing centers (Jayadev *et al*., 2013; Bird, 1998).

There is an inverse correlation between expanded repeat length and the age at which pathogenesis leads to disease onset in most of the repeat-associated SCAs (SCA types 1, 2, 3, 6,

7, 10, 17, 31, 37, DRPLA) (Paulson, 2018; Seixas *et al*., 2017; Bettencourt *et al*., 2016; Teive *et al.*, 2004; Sato *et al*., 2009; Yoshida *et al*., 2017). Longer repeat tracks are usually associated with earlier ages at onset (AO), explaining up to 88% of the variability in AO of these diseases (Bettencourt *et al*.,2016). No such correlation between the related repeat expansion length and AO in SCA8 and SCA12 has been reported yet (Cleary *et al*., 1993; O'Hearn *et al*., 2012).

Table 1. List of autosomal dominant cerebellar ataxias caused by expanded repeats, adapted from Jayadev *et al*., 2013 and Bird, 1998 (last revised 2019).

Disease	**Gene**	**Variant type**	**Distinguishing clinical features**
SCA1	*ATXN1*	CAG repeat expansion	Pyramidal signs, peripheral neuropathy
SCA2	*ATXN2*	CAG repeat expansion	Slow saccadic eye movements, peripheral neuropathy, decreased deep tendon reflexes, dementia
SCA3	*ATXN3*	CAG repeat expansion	Pyramidal and extrapyramidal signs; lid retraction, nystagmus, decreased saccade velocity; amyotrophy fasciculations, sensory loss
SCA6	*CACNA1A*	CAG repeat expansion	Sometimes episodic ataxia, very slow progression
SCA7	*ATXN7*	CAG repeat expansion	Visual loss with retinopathy
SCA8	*ATXN8/ ATXN8OS*	CAG repeat expansion	Slowly progressive, sometimes brisk decreased deep tendon reflexes, decreased vibration sense; rarely, cognitive impairment
SCA10	*ATXN10*	ATTCT repeat expansion	Occasional seizures
SCA12	*PPP2R2B*	CAG repeat expansion	Slowly progressive ataxia; action tremor in the 30s; hyperreflexia; subtle parkinsonism; cognitive/psychiatric disorders including dementia
SCA17	*TBP*	CAG repeat expansion	Mental deterioration; occasional chorea, dystonia, myoclonus, epilepsy; Purkinje cell loss, intranuclear inclusions with expanded polyglutamine
SCA31	*BEAN1*	insertion of TGGAA repeat	Normal sensation
SCA36	*NOP56*	GGCCTG repeat expansion	Muscle fasciculations, tongue atrophy, hyperreflexia
SCA37	*DAB1*	insertion of ATTTC repeat	Abnormal vertical eye movements
DRPLA	*ATN1*	CAG repeat expansion	Chorea, seizures, dementia, myoclonus
Pure cerebellar ataxia	*C9orf72*	GGGGCC repeat expansion	-
EA2	*CACNA1A*	CAG repeat expansion	Gait ataxia, nystagmus, attacks lasting minutes to hours; posture change induces, vertigo, permanent ataxia later

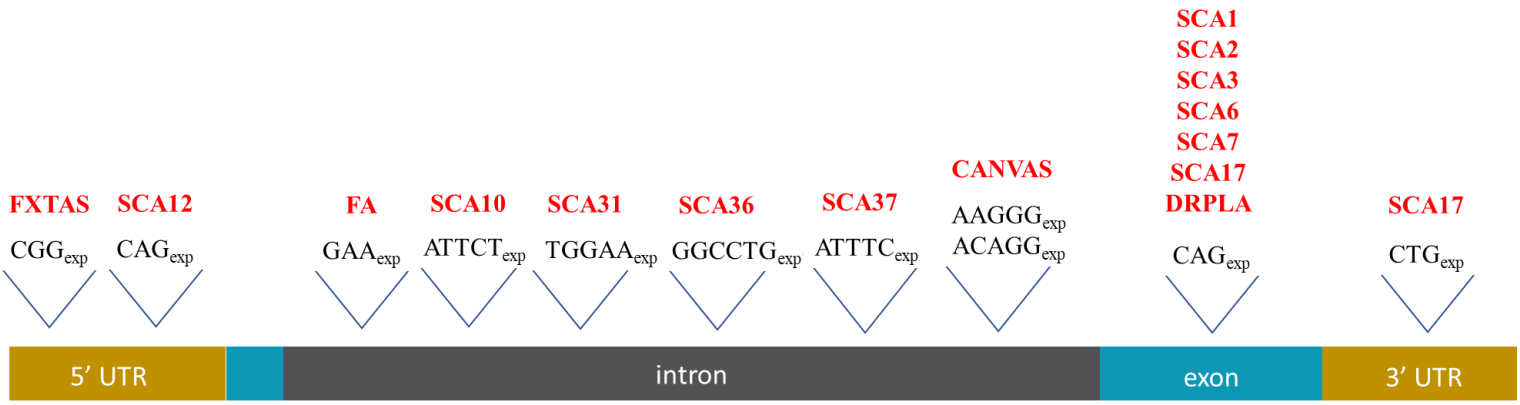


Figure 1. Known repeat expansions in hereditary ataxia (Adapted from Klockgether, 2019).

Since repeat length does not account for all of the AO variability, additional factors have been suggested to be AO modifiers. In 2015, the Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium carried out a study to search for genetic modifiers of AO in HD. It was shown that the CAG repeat length explained 59.4 % of the variability in AO. To reveal genetic modifiers that may be associated with residual AO, a genome-wide association study was conducted. A total of eleven loci were found that can explain the remaining variability in AO. Interestingly, a significant association was identified in the mismatch repair gene *MLH1*, which implicates mismatch repair pathway in disease modification (Pinto *et al*., 2013; Genetic Modifiers of Huntington's Disease Consortium, 2015). Previously, it was demonstrated that *Mlh1*-knock out alters somatic CAG repeat expansion and slows the pathogenic process in mouse models of HD (Pinto *et al*., 2013). Subsequently, Bettencourt *et al*. tested the modifying effects of variants in DNA repair genes from the previous GeM-HD study and identified two genetic loci (*FAN1* and *PMS2*), that are associated with altered AO in HD as well as SCA types 1, 2, 3, 6, 7, and 17. Together with the initial findings of the GeM-HD Consortium, their results suggested possible common pathogenic mechanisms that may carry out the somatic expansion of repeats via DNA

repair defects (Teive *et al*., 2004; Genetic Modifiers of Huntington's Disease Consortium, 2015; Lee J-M *et al*., 2019).

The signs and symptoms of some hereditary diseases are likely to become more severe and appear at an earlier age in subsequent generations. This phenomenon is called anticipation. Anticipation has been previously reported in SCA types 1, 2, 3, 7, 10, 17, 31, 36, and DRPLA (Sato *et al*., 2009; Schut *et al*., 1950; Zoghbi *et al*., 1988; Figueroa *et al*., 2017; van de Warrenburg *et al*., 2001; Ansorge *et al*., 2004; Rasmussen *et al*., 2007; Matsuura *et al*., 2020; Veneziano *et al*., 1993; Grewal *et al*., 2002; Matilla *et al*., 1993). In some repeat-associated diseases, expansion of the related repeats during transmission of the gene from parent to child provides a biologic explanation for the earlier age of onset in successive generations (Bird, 1998). In contrast to most repeat-associated SCAs in which the expansion of repeat occur during parental transmission, the majority of CTG trinucleotide expansions in SCA8 occur during maternal transmission (Figueroa *et al*., 2017; Matilla *et al*., 1993; Moseley *et al*., 2000). Interestingly, despite clinically observed anticipation in SCA10, intergenerational contraction of repeat allele was demonstrated (Matsuura *et al*., 2004). Anticipation has not been observed in SCA6, 12, or 37 (O'Hearn *et al*., 2012; Casey *et al*., 1993; Serrano-Munuera *et al*., 2013; Brkanac *et al*., 2002).

SCAs caused by conventional variants

In addition to repeat-based SCAs, there are at least 34 SCA subtypes caused by conventional variants (Table 2). Those conventional variants include SNVs, small insertions, deletions, and other copy number variations.

Table 2. List of autosomal dominant cerebellar ataxias caused by conventional variants, adapted from Jayadev *et al*., 2013 and Bird, 1998 (last revised 2019).

Disease	**Gene**	**Variant type**	**Distinguishing clinical features**
SCA5	*SPTBN2*	SNVs	Early onset, slow course
SCA11	*TTBK2*	small insertions or deletions	Mild, remain ambulatory
SCA13	*KCNC3*	SNVs	Mild intellectual disability, short stature
SCA14	*PRKCG*	SNVs or deletions	Early axial myoclonus
SCA15/SCA16	*ITPR1*	Deletion	Pure ataxia, very slow progression/head tremor
SCA19/SCA22	*KCND3*	SNVs or small deletions	Slowly progressive, rare cognitive impairment, myoclonus, hyperreflexia
SCA21	*TMEM240*	SNVs	Mild cognitive impairment
SCA23	*PDYN*	SNVs	Dysarthria, abnormal eye movements, reduced vibration and position sense
SCA26	*EEF2*	SNVs	Dysarthria, irregular visual pursuits
SCA27	*FGF14*	SNVs or deletions	Early-onset tremor; dyskinesia, cognitive deficits
SCA28	*AFG3L2*	SNVs or insertions/deletions	Nystagmus, ophthalmoparesis, ptosis, increased tendon reflexes
SCA29	*ITPR1*	SNVs	Learning deficits
SCA34	*ELOVL4*	SNVs	Skin changes disappear in adulthood
SCA35	*TGM6*	SNVs or deletions	Hyperreflexia, Babinski responses; spasmodic torticollis
SCA38	*ELOVL5*	SNVs	Axonal neuropathy
SCA40	*CCDC88C*	SNVs	Brisk reflexes, spasticity
SCA41	*TRPC3*	SNVs	Uncomplicated ataxia
SCA42	*CACNA1G*	SNVs	Mild pyramidal signs, saccadic pursuit
SCA43	*MME*	SNVs	Sensorimotor axonal neuropathy
SCA44	*GRM1*	SNVS or duplication	Spasticity
SCA45	*FAT2*	SNVs	Adult onset
SCA46	*PLD3*	SNVs	Adult onset, sensory neuropathy, mild cerebellar atrophy
SCA47	*PUM1*	SNVs	Developmental delay, intellectual disability, seizures
SCA48	*STUB1*	SNVs or deletions	Progressive cognitive disability may precede ataxia

Table 2. List of autosomal dominant cerebellar ataxias caused by conventional variants, adapted from Jayadev *et al*., 2013 and Bird, 1998 (last revised 2019) (continued).

Disease	**Gene**	**Variant type**	**Distinguishing Clinical Features**
Autosomal dominant cerebellar ataxia, deafness, and narcolepsy (ADCADN)	*DNMT1*	SNVs	Deafness, sensory loss, narcolepsy
Hypomyelinating leukoencephalopathy	*TUBB4A*	SNVs	Hypomyelination, basal ganglia atrophy, rigidity, dystonia, chorea
Cerebellar atrophy with epileptic encephalopathy	*FGF12*	SNVs	Infantile seizures, intellectual deficits, microcephaly
Rapid-onset ataxia	*ATP1A3*	SNVs	Cerebellar atrophy
SPAX1	*VAMP1*	SNVs or deletions	Initial progressive leg spasticity
EA1	*KCNA1*	SNVs	Gait ataxia, myokymia, attacks lasting seconds to minutes; startle or exercise induced, no vertigo
EA2	*CANCA1A*	SNVs or deletions	Gait ataxia, nystagmus, attacks lasting minutes to hours; posture change induces, vertigo, permanent ataxia later
EA5	*CACNB4*	SNVs	Childhood to adolescent onset
EA6	*SLC1A3*	SNVs	Seizures, migraine, childhood onset
EA9	*SNC2A*	SNVs	Neonatal epilepsy, later-onset episodic ataxia, autism, hypotonia, dystonia

Furthermore, there are a number of SCA subtypes in which the causal genes or variants have not been identified yet. For example, SCA18 was reported in 26 patients from a five-generation American family of Irish ancestry. Although the disease locus was mapped to 7q22-7q32, no casual variants have been described in this locus so far (Muller, 2021; Brkanac *et al*., 2002). Similarly, using linkage analysis, Storey *et al*. identified a candidate region on chromosome 4q34.3-q35.1 for relatively pure, slowly evolving ataxia with an autosomal dominant inheritance

(SCA30) in an Australian family of Anglo-Celtic origin. The causal variant(s) in this locus have yet to be identified (Storey *et al*., 2009).

1.1.1.2. Spastic ataxia type 1

Spastic ataxia (SPAX) is a combination of spasticity and cerebellar ataxia. SPAX can resemble both SCAs and hereditary spastic paraplegias (de Bot *et al*., 2012). SPAX1 is the only autosomal dominant subtype that is characterized by ocular movement abnormalities, dysphagia, dysarthria, gait disturbance, and lower-limb spasticity, and ataxia in the form of head jerks. A heterozygous missense variant in *VAMP1* was identified as disease-causing that segregated in four large families from Newfoundland as well as three isolated cases from Ontario, Canada (Bourassa *et al*., 2012) .

1.1.1.3. Episodic ataxias

Episodic ataxias (EAs) are characterized by attacks of movement incoordination of variable duration and frequency (Giunti *et al.*, 2020). EA1 and EA2 are the two most common subtypes that have been reported in multiple families from different ethnicities. Pathogenic variants in potassium and calcium channel genes (*KCNA1 and CACNA1A*) cause EA1 and EA2, respectively. EA5 is another subtype that is caused by a channel protein, *CACNB4*, which encodes for a calcium channel protein (Subramony, 2012). These ion channel proteins are located on the neuronal or glial membrane and play important roles in excitatory neurotransmission (Choi *et al*., 2016).

1.1.2. Autosomal recessive cerebellar ataxias

Autosomal recessive cerebellar ataxias are characterized by degeneration or abnormal development of cerebellum and spinal cord (Subramony, 2012). Unlike autosomal dominant cerebellar ataxias, AO is usually early in recessive ataxias (Hersheson *et al*., 2012). Some of the recessive ataxia subtypes are treatable, such as ataxia with vitamin E deficiency and coenzyme Q10 deficiency using vitamin E supplementation and coenzyme Q10 (Bird, 1998) (Table 3).

Table 3. List of autosomal recessive cerebellar ataxias (examples of more frequent or treatable), adapted from Jayadev *et al*., 2013 and Bird, 1998 (last revised 2019).

Disease	**Gene**	**Variant type**	**Distinguishing clinical features**
Ataxia-telangiectasia	*ATM*	SNVs or insertions/deletions	Telangiectasia, immune deficiency, cancer, chromosomal instability, increased α-fetoprotein
Ataxia with oculomotor apraxia type 1	*APTX*	SNVs or insertions/deletions	Oculomotor apraxia, choreoathetosis, mild intellectual disability, hypoalbuminemia
Ataxia with oculomotor apraxia type 2	*SETX*	SNVs or insertions/deletions	Oculomotor apraxia, cerebellar atrophy, axonal sensorimotor neuropathy
Ataxia with vitamin E deficiency	*TTPA*	SNVs or insertions/deletions	Similar to FRDA, head titubation (28%), can be treated with vitamin E
Autosomal recessive spastic ataxia of Charlevoix-Saguenay	*SACS*	SNVs or insertions/deletions	Spasticity, peripheral neuropathy, retinal striation
Cerebellar ataxia, neuropathy and vestibular areflexia syndrome (CANVAS)	*RFC1*	AAGGG or ACAGG repeat expansion	Late-onset, peripheral neuropathy, vestibular areflexia
Cerebrotendinous xanthomatosis	*CYP27A1*	SNVs or insertions/deletions	Thick tendons, cognitive decline, dystonia, white matter disease, cataract, can be treated with chenodeoxycholic acid
Coenzyme Q10 deficiency	*CABC1, COQ2, COQ9, PDSS1, PDSS2*	SNVs or insertions/deletions	Seizures, cognitive decline, pyramidal signs, myopathy, can be treated with coenzyme Q10
Friedreich ataxia (FRDA)	*FXN*	GAA repeat expansion or SNVs	Hyporeflexia, Babinski responses, sensory loss, cardiomyopathy
Refsum disease	*PHYH, PEX7*	SNVs or insertions/deletions	Neuropathy, deafness, ichthyosis, retinopathy, can be treated with phytanic acid

1.1.2.1. Friedreich ataxia

Friedreich ataxia (FRDA) is one of the oldest studied recessive ataxias which was first described by Nicholas Friedreich in 19th century (Friedreich, 1863). With the advent of linkage studies, the disease locus was mapped in 9p22 (Chamberlain *et al*., 1988), and the associated biallelic GAA trinucleotide repeat expansion in frataxin (*FXN*) was identified eight years after the discovery of the locus (Campuzano *et al*., 1996). Similar to other repeat-associated diseases, AO is inversely correlated with the repeat length (Durr *et al*., 1996). Since FRDA is not usually observed in multiple generations due to its recessive inheritance, it does not exhibit anticipation (Bidichandani *et al*., 1993).

1.1.2.2. Autosomal recessive spastic ataxias

As previously defined in section 1.1.1.2, spastic ataxias resemble a combination of spasticity and cerebellar ataxia (de Bot *et al.,* 2012). Autosomal recessive spastic ataxias are caused by biallelic or compound heterozygous genetic variants and comprise SPAX2, SPAX3, SPAX4, SPAX5, autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS), and spastic paraplegia type 7 (SPG7) (Bird, 1998).

ARSACS is a distinct form of spastic ataxia that is associated with a biallelic or compound heterozygous variation in the gene encoding the sacsin protein (Engert *et al*., 2000). The disease onset of classic ARSACS is usually in the first decade of life (Vermeer *et al*., 1993). It was first identified in the Charlevoix and Saguenay–Lac-Saint-Jean regions of Québec, Canada (Bouchard *et al.,* 1978). Although the estimated carrier frequency is 1/22 in this region which is quite common for a rare disease, it has been described in various populations since then (De Braekeleer *et al*., 1993; Dupre *et al*., 2006).

1.1.2.3. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome (CANVAS)

CANVAS is an adult-onset recessive ataxia characterized by sensory neuropathy, bilateral vestibulopathy, chronic cough, and autonomic dysfunction (Szmulewicz *et al*., 2011). Recently, a biallelic pentanucleotide AAGGG repeat expansion in the Alu element in intron 2 of the *RFC1* was shown to cause CANVAS. The wild type allele contains 11 units of AAAAG, whereas the pathogenic AAGGG repeat length varies across cases, ranging from 400 to 2,000 repeats. Along with wild type $(AAAAG)_{11}$, two additional non-pathogenic repeat expansions $(AAAAG)_{exp}$ and $(AAAGG)_{exp}$ were identified by Cortese *et al* (Cortese *et al*., 2019).

The analysis of pathogenic $(AAAGG)_{exp}$ in 150 patients with sporadic late-onset ataxia and 363 adult-onset ataxia cases revealed that biallelic $(AAGGG)_{exp}$ explains 14-22% of late-onset ataxia cases (Szmulewicz *et al*., 2011; Cortese *et al*., 2019). In addition, carrier frequency of the pathogenic $(AAGGG)_{exp}$ was found to range from 0.7% to 4% in European populations (Cortese *et al*., 2019; Cortese *et al*., 2020; Rafehi *et al*., 2019) and 2.24% in Chinese Han population (Fan *et al*., 2020).

Subsequently, another *RFC1* repeat motif $(ACAGG)_{exp}$ was identified in one Japanese and two Asia-Pacific families. It was the second disease-associated variant in the *RFC1* locus; therefore, it expanded the clinical spectrum of *RFC1*-based CANVAS with additional phenotypic features of fasciculations, elevated serum creatine kinase, and sleep apnea (Scriba *et al*., 2020; Tsuchiya *et al*., 2020).

1.1.3. X-linked hereditary ataxias

X-linked ataxias are a rare type of hereditary ataxia that affect males more often than females as they are caused by variants or genomic imbalances on the X chromosome. Fragile X tremor ataxia syndrome (FXTAS) is the most common X-linked ataxia subtype associated with $(CGG)_{exp}$ repeat expansion in the *FMR1* gene (Zanni *et al*., 2018).

1.1.4. Possible mechanisms implicated in repeat-based hereditary ataxia

The pathogenesis of repeat-associated hereditary ataxia is largely unknown. However, an increasing number of causative genetic variants has revealed various potential mechanisms related to neurodegeneration. CAG repeat expansions are one of the most common variants causing hereditary ataxia. Expansions of these polyglutamine-encoding repeats contribute to the disease pathogenesis via multiple mechanisms in different subtypes. The first suggested mechanism is that polyglutamine expansions may alter the structure and function of the disease proteins, thereby disrupting their interaction with other proteins. This alteration causes the formations of aggregates and intranuclear inclusions by sequestering protein quality control components, including proteasome and molecular chaperones. A similar proteotoxicity is seen in non-repeat-based ataxias where conventional variants result in the misfolding of disease-causing proteins (Klockgether *et al*., 2019).

Another pathological mechanism is repeat-associated non-ATG (RAN) translation, which allows the initiation of mRNA translation in the three reading frames without requiring a start codon (Klockgether *et al*., 2019). This irregular mode of translation across a CAG repeat can produce polyserine and polyalanine peptides along with polyglutamine. Although it was first

discovered in SCA8, aggregation of RAN proteins have subsequently been shown in various repeat-based diseases including myotonic dystrophy type 1, SCA3, SCA31, *C9orf72*-based amyotrophic lateral sclerosis (ALS) and frontotemporal dementia, FXTAs, and HD (Zu *et al*., 2011; Klockgether *et al*., 2019; Jazurek-Ciesiolka *et al*., 2020; Swinnen *et al*., 2020). RAN translated protein toxicity may cause disrupted function of the ubiquitin proteasome system, aberrant nucleocytoplasmic transport, or nucleolar and endoplasmic reticulum stress resulting in neurodegeneration (Jazurek-Ciesiolka *et al*., 2020).

In addition to RAN translation, one base pair ribosomal frameshifting was shown to lead to translation of polyalanine stretches in SCA3 (GCA frame), causing deleterious effects in *Drosophila melanogaster* and mammalian neuronal models. It was also demonstrated that transgenic expression of polyglutamine repeat itself was not toxic and not sufficient for the neurodegenerative phenotype. Therefore, a one base pair frameshifting event was suggested to be toxic (Stochmanski *et al*., 2012). On the other hand, another study showed that the ribosomal frameshift is not required for polyalanine expression and an ATG codon is not required for the RAN polyalanine proteins. It has been suggested that both RAN polyglutamine and polyalanine proteins are toxic to cells and that the initiation and efficiency of RAN translation in SCA3 depends on the flanking sequence of the *ATXN3* repeat region (Jazurek-Ciesiolka *et al*., 2020).

RNA toxicity is also a possible pathological mechanism in SCA10, SCA31, SCA36, and SCA37, since these diseases are caused by repeat expansions in large non-protein coding intronic regions. Repeat-containing RNA can interact with RNA-binding proteins and disrupt splicing as well as a loss of their functions through sequestration (Klockgether *et al*., 2019). Similarly,

disease-causing *RFC1* repeat expansion occurs in an intronic region. Despite its recessive mode of inheritance, preliminary studies did not show reduced expression or loss of function of RFC1 protein. Therefore, production of toxic RNA can be a possible mechanism that may contribute to disease phenotype (Cortese *et al*., 2019).

1.1.5. Rationale, objectives, and hypothesis

Hereditary ataxias are a clinically and genetically heterogeneous group of disorders. One common feature is that many of these subtypes are associated with repeat expansion variants. Repeat expansions generally arise from existing polymorphic repeats and have been shown to be implicated in diseases in a polyvalent manner. For example, the first identified SCA subtypes, SCA1, 2, 3, 6, 7, 8, 12, and 17 are associated with CAG repeat expansions when the length of these repeats exceeds a certain pathogenic threshold specific for each disease. In addition, some repeat expansions are associated with ataxia when their nucleotide sequence configuration are different from the wild-type motif, such as in CANVAS and SCA37. Furthermore, repeat expansions do not always have only a dichotomous outcome; most of them are also associated with variation in AO. The overall hypothesis of this thesis is that the nature, frequency, and implication of ataxia-associated repeat expansions are different in various cohorts and that there are additional genetic factors modifying AO.

Therefore, the three objectives of the presented doctoral thesis were to:

1. Examine the CAG repeats associated with SCA 1, 2, 3, and HD using 30X whole-genome sequencing data of 2,504 samples from the 1000 Genomes Project.
2. Assess the prevalence and nature of *RFC1* repeat expansions in Brazilian and Canadian cohorts of adult-onset ataxia as well as a control population consisting of neurologically healthy individuals.
3. a. Examine the relationship between AO and length of the expanded and normal CAG alleles in SCA3.

   b. Identify genetic modifiers of AO in SCA3.

CHAPTER 2: EXPANDED CAG REPEATS IN *ATXN1*, *ATXN2*, *ATXN3* and *HTT* in the 1000 GENOMES PROJECT

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

Affiliations

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

*Corresponding author: Guy A. Rouleau, Montreal Neurological Institute and Hospital, 3801 University Street, Room 636, Montréal, Québec H3A2B4, Canada; e-mail: guy.rouleau@mcgill.ca, phone: +1 (514) 398-6644

Published in

Akçimen F, Ross JP, Liao C, Spiegelman D, Dion PA, Rouleau GA. Expanded CAG Repeats in *ATXN1*, *ATXN2*, *ATXN3*, and *HTT* in the 1000 Genomes Project. Mov Disord. 2021 Feb;36(2):514-518. doi: 10.1002/mds.28341. Epub 2020 Nov 7. PMID: 33159825.

2.1. Abstract

Spinocerebellar ataxia types 1, 2, 3 and Huntington disease are neurodegenerative disorders caused by expanded CAG repeats. We performed an in-silico analysis of CAG repeats in *ATXN1*, *ATXN2*, *ATXN3*, and *HTT* using 30X WGS data of 2,504 samples from the 1000 Genomes Project. Seven *HTT*-positive, three *ATXN2*-positive, one *ATXN3*-positive, and six possibly *ATXN1*-positive samples were identified. No correlation was found between the repeat sizes of the different genes. The distribution of CAG alleles varied between different ethnicities. Our results suggest that there may be asymptomatic small, expanded repeats in almost 0.5% of these populations.

2.2. Introduction

Spinocerebellar ataxias (SCAs), and Huntington disease (HD) are rare autosomal dominant neurodegenerative disorders. SCAs are genetically heterogeneous diseases, of which at least six 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) (Klockgether *et al*., 2019). 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 (Bates, 2005). Deleterious alleles for the most common SCAs (SCA1, 2, 3) contain over 45 repeats (or 39 uninterrupted with a CAT codon), 33, and 45 CAG repeats in *ATXN1*, *ATXN2*, and *AXTN3*, respectively (Zuhlke *et al*., 2002; Fernandez *et al*., 2000; Bettencourt *et al*., 2011).

The International Genome Sample Resource (IGSR) curates public data resources that are created by the 1000 Genomes Project (1KGP) (Auton *et al*., 2015; Fairley *et al*., 2019). The 1KGP phase 3 panel consists of 2,504 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 over 18 years of age and self-declared healthy at the time of collection. The project holds self-reported ethnicity and gender. No phenotype, medical, or personal identifying information were collected (Auton *et al*., 2015). Previously, various types of structural variants including insertions, deletions, duplications, copy-number variants, and insertions were mapped in 1KGP. However, known disease-related short-tandem repeats (STRs) have not been reported in this dataset (Sudmant *et al*., 2015). In 2019, the New York Genome Center re-sequenced the

samples in the final phase of 1KGP. High-coverage PCR-free whole-genome sequencing (WGS) data of a total of 2,504 samples from 26 populations were added (Fairley et al., 2019).

ExpansionHunter is a software that can estimate sizes of targeted STRs from PCR-free WGS data (Dolzhenko *et al*., 2017; Dolzhenko *et al*., 2019). It identifies lengths of the repeats using either spanning, flanking or in-repeat reads. Therefore, it enabled us to employ an in-silico analysis of CAG repeat expansions in HD and the most common SCAs using high coverage WGS data among different ancestries from IGSR (Auton *et al*., 2015). We hypothesized that samples in a reference dataset 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.

2.3. 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 2,504 PCR-free WGS samples of 26 populations from five super populations (AFR: African, AMR: Ad Mixed American, EAS: East Asian, EUR: European, and SAS: South Asian) were downloaded from 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 over 18 years and declared themselves to be healthy at the time of the collection.

Alignment files were indexed using SAMtools v.1.10 (Li *et al*., 2009). 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 (Dolzhenko *et al*., 2017; Dolzhenko *et al*., 2019). Violin plots representing the distributions of CAG repeat sizes in different populations were plotted in R v.3.5.1 using ggplot2 (Wickham, 2016). 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.

2.4. Results

Using ExpansionHunter, CAG repeats lengths were successfully estimated in 2,486 samples for *HTT*, 2,390 samples for *ATXN1*, 2,408 samples for *ATXN2*, and 2,339 samples for *ATXN3*. Mean CAG repeat lengths identified in each population are shown in Supplementary Table 1. The full results for all samples are listed in Supplementary Table 2. Expanded CAG repeats associated with diseases were detected in a total of 11 samples (*HTT* in seven, *ATXN2* in three, and *ATXN3* in one). No pathogenic *ATXN1* expansions that have a repeat size higher than 45 were found. However, intermediate expansions (39-44 CAG repeats) that can be in the disease-associated range in *ATXN1* were identified in six samples. However, these could be associated with the disease only in the absence of CAT trinucleotide interruptions. Since 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, $P_{ATXN1\text{-}ATXN2}$ = 0.06, $P_{ATXN1\text{-}ATXN3}$ = 0.67, $P_{ATXN2\text{-}HTT}$ = 0.27, $P_{ATXN2\text{-}ATXN3}$ = 0.76, $P_{ATXN3\text{-}HTT}$ = 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 $R^2 = 0.16$, 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 $P = 0.18$).

2.5. Discussion

This study represents an *in-silico* analysis of CAG repeat expansions of the 2,504 samples from 1KGP. Through leveraging public high coverage sequencing data as well as available STR genotyping approach, ExpansionHunter, we sought to examine the CAG repeats associated with SCA1, 2, 3, and HD in 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 eleven (plus six 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 are 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 (Kay *et al*., 2014), lower mean *HTT*-CAG

repeat size was observed in the samples with East Asian ancestry that is correlated with lower prevalence of HD in these populations. This pattern was also observed in the European populations that have longer *HTT*-CAG repeats and higher prevalence estimates of HD (Kay *et al*., 2014). Additionally, a skewed distribution of *ATXN3*-CAG alleles toward intermediate size repeats in African and East Asian populations was detected. Higher frequency of intermediate alleles were shown to be enriched in populations with higher prevalence of repeat expansion diseases, strengthening the hypothesis of the repeat's instability and expansion into the disease causing range from the high normal or intermediate size alleles as a cause of CAG related diseases (Kay *et al*., 2014; Budworth *et al*., 2013; Martins *et al*., 2007; Friedman, 2011).

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. Genetic variants implicated in DNA repair mechanisms that possibly influence somatic expansions were identified as candidate genetic modifiers of the diseases (Bettencourt *et al*., 2016; Akçimen *et al*., 2020; Lee *et al.,* 2019). Although common variants and mechanisms are implicated in somatic expansions of CAG repeats in respective genes, our findings suggest that germline instability occurs independently in each CAG repeat that could implicate unique mutational mechanisms.

While of interest and original, our study has some limitations. The average sample size of subpopulations is 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 dataset does not represent all populations (Fairley *et al*., 2019). 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) (Dolzhenko *et al*., 2017), 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 range between 36 to 44 repeat sizes in *ATXN1*, were not reported. Alleles with CAT interruptions in the 36 to 44 repeat range 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) (Opal *et al*., 1993). Therefore, the identified *ATXN1* repeat expansions in six samples may not be associated with SCA1. However, alleles in the mutable normal range may expand beyond the normal range during transmission to offspring which may manifest the disease (Opal *et al*., 1993). Furthermore, HD and CAG-associated SCAs usually occur in the third or fifth decade. However, the ages of onset for these diseases are highly variable (Klockgether *et al*., 2019; Budworth *et al*., 2013). Although HD is an adult-onset neurological disorder, its symptoms can appear as early as age 18 or as late as 80 (Bates, 2005; Budworth *et al*., 2013). Similarly, the average age of onset is 38 ranging between 10 and 70 in SCA3 (Akçimen et al., 2020). In a panel study for dominant cerebellar ataxias, the average age of onset was 40.9 in known CAG-associated SCAs (Coutelier *et al*., 2017). Therefore, the positive individuals might be asymptomatic at the time of collection, as these diseases usually occur in the third or fifth decade. However, due to the anonymity of the samples, no personal information including ages of the individuals were collected in 1KGP. Therefore, we were unable to infer if the positive samples were too young for the 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. This data can be useful to understand the population distribution of these repeats in different populations. Furthermore, pathogenic-length repeats in 11 samples were observed. This suggests that the datasets generated from the general populations might contain samples positive for late onset diseases, even though these samples declared themselves as 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.

2.6. Acknowledgements

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.

2.7. Tables and Figures

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

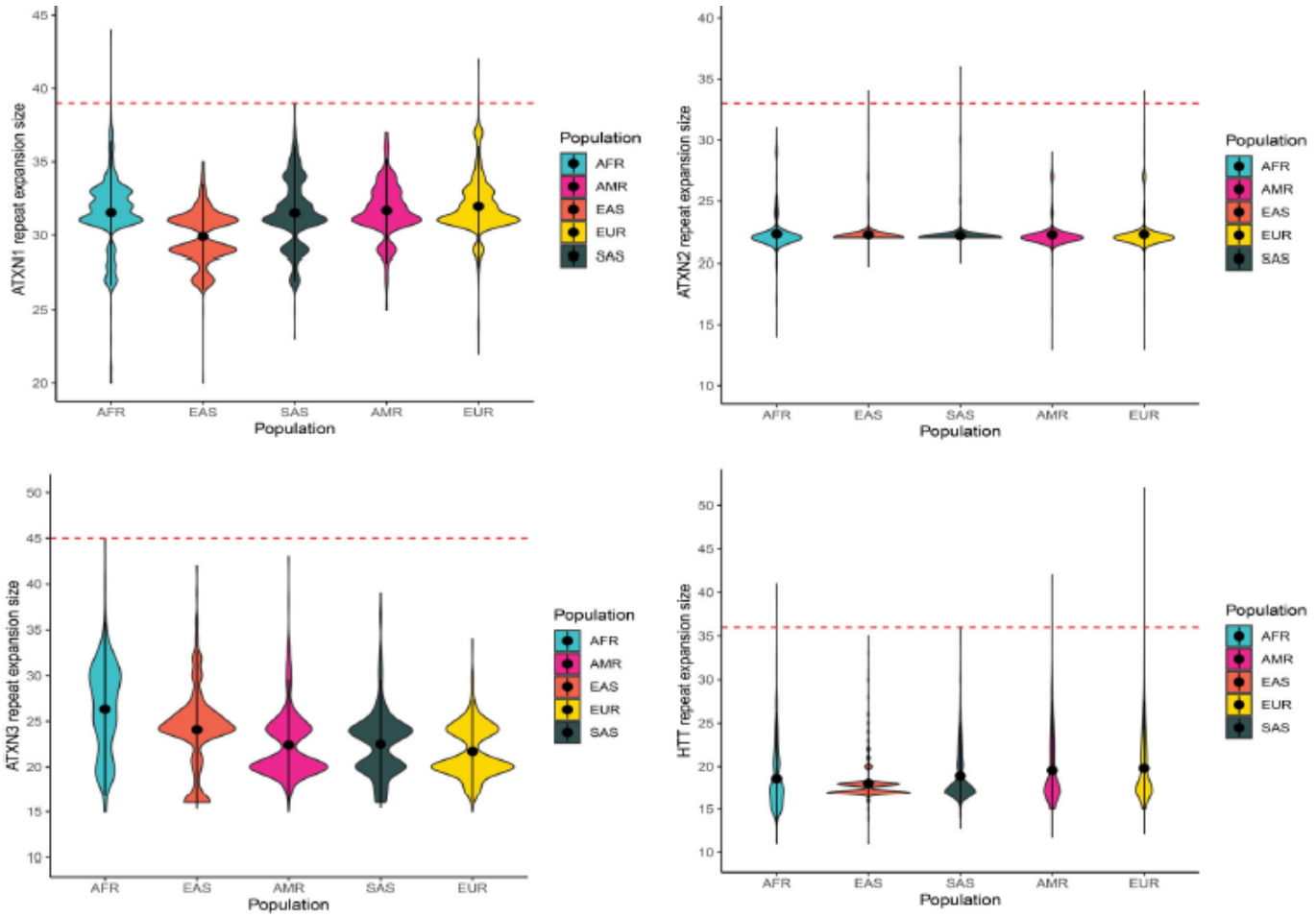


Figure 1. Distribution of repeat expansion sizes among different ethnic groups in the 1KGP. Red line indicates the threshold for causality.

2.8. Supplemental materials

Additional supporting information (Supplementary tables 1 and 2) can be downloaded from the online version of this article at the publisher's web-site: https://movementdisorders.onlinelibrary.wiley.com/doi/10.1002/mds.28341

2.9. References

Akçimen 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):4742-56. doi: 10.18632/aging.102825 [published Online First: 2020/03/25]

Auton A, Abecasis GR, Altshuler DM, *et al*. A global reference for human genetic variation. Nature 2015;526(7571):68-74. doi: 10.1038/nature15393

Bates GP. The molecular genetics of Huntington disease — a history. Nature Reviews Genetics 2005;6(10):766-73. doi: 10.1038/nrg1686

Bettencourt C, Lima M. Machado-Joseph Disease: from first descriptions to new perspectives. Orphanet J Rare Dis 2011;6:35. doi: 10.1186/1750-1172-6-35

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 2016;79(6):983-90. doi: 10.1002/ana.24656 [published Online First: 05/06]

Budworth H, McMurray CT. A brief history of triplet repeat diseases. Methods Mol Biol 2013;1010:3-17. doi: 10.1007/978-1-62703-411-1_1

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): 1579–94. doi:10.1093/brain/awx081

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):1895-903. doi: 10.1101/gr.225672.117

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):4754-56. doi: 10.1093/bioinformatics/btz431

Li H, Handsaker B, Wysoker A, *et al*. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078-9. doi: 10.1093/bioinformatics/btp352 [published Online First: 2009/06/10]

Fairley S, Lowy-Gallego E, Perry E, *et al*. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Research 2019;48(D1):D941-D47. doi: 10.1093/nar/gkz836

Fernandez M, McClain ME, Martinez RA, *et al*. Late-onset SCA2: 33 CAG repeats are sufficient to cause disease. Neurology 2000;55(4):569-72. doi: 10.1212/wnl.55.4.569 [published Online First: 2000/08/23]

Friedman JE. Anticipation in hereditary disease: the history of a biomedical concept. Human genetics 2011;130(6):705-14. doi: 10.1007/s00439-011-1022-9 [published Online First: 2011/06/15]

Kay C, Fisher E, Hayden MR. Huntington's Disease. Epidemiology: Oxford University Press 2014.

Klockgether T, Mariotti C, Paulson HL. Spinocerebellar ataxia. Nat Rev Dis Primers 2019;5(1):24. doi: 10.1038/s41572-019-0074-3

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. doi: https://doi.org/10.1016/j.cell.2019.06.036

Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/mapformat and SAMtools. Bioinformatics 2009;25(16):2078–2079

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):1502-8. doi: 10.1001/archneur.64.10.1502 [published Online First: 2007/10/10]

Opal P, Ashizawa T. Spinocerebellar Ataxia Type 1. 1998 Oct 1 [Updated 2017 Jun 22]. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2020. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1184/

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. doi: 10.1038/nature15394

Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer 2016.

Zuhlke C, Dalski A, Hellenbroich Y, *et al*. Spinocerebellar ataxia type 1 (SCA1): phenotype-genotype correlation studies in intermediate alleles. European journal of human genetics : EJHG 2002;10(3):204-9. doi: 10.1038/sj.ejhg.5200788 [published Online First: 2002/04/26]

Bridging statement to Chapter 3

CAG repeats are associated with SCA types 1, 2, 3 and HD when these repeat tracks exceed a certain threshold for each subtype. The number of CAG repeats is typically determined by a standard polymerase chain reaction and fragment length analysis. In Chapter 2, we performed an *in silico* approach to estimate CAG repeat expansions using whole-genome sequencing samples from 1000 Genomes Project dataset.

Unlike expanded CAG repeats, some repeats are associated with ataxia when they contain different sequence conformations from the wild-type motifs. In Chapter 3, we examined the prevalence of pathogenic *RFC1*-$(AAGGG)_{exp}$ expansions and other possible sequence conformations in the repeat region in various adult-onset ataxia cohorts. For this chapter, we applied repeat-primed PCR and long-range amplification followed by Sanger sequencing.

CHAPTER 3: INVESTIGATION OF THE RFC1 REPEAT EXPANSION IN A CANADIAN AND A BRAZILIAN ATAXIA COHORT: IDENTIFICATION OF NOVEL CONFORMATIONS

Fulya Akçimen MSc[1,2], Jay P. Ross[1,2], Cynthia V. Bourassa MSc[2], Calwing Liao[1,2], Daniel Rochefort MSc[2], Maria Thereza Drumond Gama MD PhD[4], Marie-Josée Dicaire[2], Orlando G. Barsottini MD PhD [4], Bernard Brais MD PhD FRCP(C)[1,2,3], José Luiz Pedroso MD PhD [4], Patrick A. Dion PhD[2,3], Guy A. Rouleau MD PhD FRCP(C)[1,2,3*]

Affiliations

[1]Department of Human Genetics, McGill University, Montréal, QC, Canada
[2]Montreal Neurological Institute and Hospital, McGill University, Montréal, QC, Canada
[3]Department of Neurology and Neurosurgery, McGill University, Montréal, QC, Canada
[4]Division of General Neurology and Ataxia Unit, Department of Neurology, Federal University of São Paulo (UNIFESP), São Paulo, Brazil

*Corresponding author: Guy A. Rouleau, Montreal Neurological Institute and Hospital, 3801 University Street, Room 636, Montréal, Québec H3A2B4, Canada; e-mail: guy.rouleau@mcgill.ca, phone: +1 (514) 398-6644

Published in

Akçimen F, Ross JP, Bourassa CV, Liao C, Rochefort D, Gama MTD, Dicaire M-J, Barsottini OG, Brais B, Pedroso JL, Dion PA and Rouleau GA (2019) Investigation of the RFC1 Repeat Expansion in a Canadian and a Brazilian Ataxia Cohort: Identification of Novel Conformations. Front. Genet. 10:1219. doi: 10.3389/fgene.2019.01219

3.1. Abstract

A biallelic pentanucleotide expansion in the *RFC1* gene has been reported to be a common cause of late-onset ataxia. In the general population, four different repeat conformations are observed: wild type sequence AAAAG (11 repeats) and longer expansions of either AAAAG, AAAGG or AAGGG sequences. However only the biallelic AAGGG expansions were reported to cause late-onset ataxia. In this study, we aimed to assess the prevalence and nature of *RFC1* repeat expansions in three cohorts of adult-onset ataxia cases: Brazilian (n=23) and Canadian (n=26) cases that are negative for the presence of variants in other known ataxia-associated genes, as well as a cohort of randomly selected Canadian cases (n = 128) without regard to a genetic diagnosis. We identified the biallelic AAGGG expansion in only one Brazilian family which presented two affected siblings, and in one Canadian case. We also observed two new repeat conformations, AAGAG and AGAGG, which suggests the pentanucleotide expansion sequence has a dynamic nature. To assess the frequency of these new repeat conformations in the general population, we screened 163 healthy individuals and observed the AAGAG expansion to be more frequent in cases than in control individuals. While additional studies will be necessary to assess the pathogenic impact of biallelic genotypes that include the novel expanded conformations, their occurrence should nonetheless be examined in future studies.

3.2. Introduction

Autosomal recessive cerebellar ataxias regroup a number of heterogenous neurodegenerative diseases. While each form of ataxia exhibits the key feature of cerebellar dysfunction, typically accompanied by gait and balance problems, some forms have distinct clinical characteristics such as, dysarthria, dysmetria or oculomotor abnormalities. Other neurological dysfunctions and/or non-neurologic phenotypes have also been reported in some cases. Among the different forms of recessive ataxia, Friedreich's ataxia (FRDA) has the highest prevalence and is the most studied (Synofzik *et al.*, 2019). In regard to prevalence FRDA is followed by autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS), ataxia with vitamin E deficiency, autosomal recessive cerebellar ataxia type 1 (ARCA-1) and type 2 (ARCA-2), and ataxia with oculomotor apraxia type 1 (AOA-1) and type 2 (AOA-2) (Noreau *et al.*, 2013).

Cortese and colleagues established the biallelic expansion of an AAGGG pentanucleotide repeat located in the second intron of the *RFC1* gene (hg19/GRCh37, chr4:39,350,045-39,350,103) to be a frequent cause of late-onset recessive ataxia; this particular expansion was reported to explain over 20% of sporadic ataxia in a cohort of Caucasian cases (Cortese *et al*., 2019). In the same study, a total of four distinct intronic repeat conformations were also identified: $(AAAAG)_{11}$, the wild-type sequence, and longer expansions of $(AAAAG)_n$, $(AAAGG)_n$ and $(AAGGG)_n$. The configuration with the AAGGG pentanucleotide was shown to be the only disease-causing conformation of the expansion, ranging in size from 600 to 2,000 repeats.

Considering that *RFC1* appears to be a novel genetic risk factor that explains a significant share of adult-onset ataxia cases, the identification of carriers in other populations may altogether expand its clinical spectrum, provide examples of variable regional prevalence, and uncover repeat

sequence differences. Therefore, we screened the *RFC1* expansion in Canadian and Brazilian ataxia patients.

3.3. Materials and methods

Two cohorts consisting of unrelated adult-onset ataxia cases were used to estimate the prevalence of the *RFC1* expansions. Detailed cohort demographics are shown in Table 1. Cohort 1 and cohort 2 comprised Brazilian (n = 23) and Canadian (n = 26) adult-onset cases, who did not carry variants in genes associated with common dominant and recessive ataxias (FRDA, DRPLA, SCA1, SCA2, SCA3, SCA6, SCA7, SCA10, SCA12, SCA17 and ARCA-1). Cohort 3 consisted of randomly selected adult-onset ataxia Canadian probands (n = 128). In addition, a cohort of 163 healthy Canadian control individuals was also examined, to estimate the frequency of the novel sequence conformations that were observed for the *RFC1* repeat expansion. All subjects provided informed consent, and the study was approved by the appropriate institutional review boards.

Screening of the *RFC1* repeat expansion was performed on genomic DNA by repeat-primed PCR (RP-PCR) as described in Cortese *et al*. using the same set of primers (Cortese *et al*., 2019). RP-PCR products were separated on an ABI3730xl DNA Analyzer (Applied Biosystems®, McGill University and Genome Québec Innovation Centre) and results were visualized using GeneMapper® v.4.0 (Applied Biosystems®). The samples that seemed biallelic for the AAGGG repeat (according to the RP-PCR results) were subjected to long-range PCR (using the same primers as Cortese *et al*. (Cortese *et al*., 2019)) and Sanger sequencing. Samples for which the allelic repeat combinations could not be determined by RP-PCR were subjected as well to a long-range PCR; the product of which was purified (QIAquick gel extraction kit, Qiagen).

The Sanger sequencing results of these long-range PCR were analyzed using Unipro UGENE version 1.31 (Okonechnikov *et al*., 2012). Finally, to compare the distribution of *RFC1* alleles in Canadian case and control groups, we performed a Chi-square test using the counts of five conformations ((AAAAG)11, (AAAAG)n, (AAAGG)n, (AAGGG)n, (AAGAG)n) in a 2×5 contingency table (Supplementary Table 2).

3.4. Results

To examine the prevalence of *RFC1*-based adult-onset ataxia, we screened the nature and size of the repeat expansions in a cohort of Brazilian cases and two cohorts of Canadian cases. The RP-PCR examination of the Brazilian cohort revealed two out of 23 individuals to be carrier of biallelic AAGGG causative expansions. However, long-range PCR and Sanger sequencing subsequently revealed one of these two individuals to actually carry a biallelic AAAGG expansion; the same biallelic expansion was observed in his sister. It therefore appears that expanded AAAGG repeat expansion can sometimes mimic the AAGGG expansion when an assessment is made only by RP-PCR, under such a context the results can lead to a misinterpretation of the true nature of the repeat expansion. The use of different RP-PCR primers (Cortese *et al*., 2019) could not resolve this mimicry of the AAGGG repeat by the AAAGG repeat.

Across the different cohorts, two cases were observed and validated to carry the causative biallelic AAGGG repeat expansion originally reported by Cortese *et al.* (Cortese *et al*., 2019). One case with two patients were Brazilian siblings, and the other one was Canadian (Figures 1A, B respectively). Clinical features of these three patients with biallelic AAGGG expansions are summarized in Supplementary Table 1. The allele count and frequency of the different repeat expansions observed in all three cohorts are shown in Table 1.

Interestingly, our cohorts of cases revealed the presence of two previously undescribed repeat expansion confirmations (AAGAG and AGAGG); both motifs were observed by long-range RP-PCR and validated by Sanger sequencing. The RP-PCR plots were characterized by a single peak, but the allele was longer than the wild type (Figures 1C, D). The novel conformations were in a heterozygous state in all 22 carrier individuals (Table 1). The average length of these novel expand configurations is 800 bp (160 repeats) ranging from 600 to 900. The approximate lengths of the repeat conformations were shown in Supplementary Figure 1.

The frequency of the expanded AAGAG and AGAGG repeat configurations was assessed in 163 Canadian control individuals showing no signs of ataxia; using a combined RP-PCR and long-range PCR approach. On the whole, a total of seven control individuals presented a heterozygous AAGAG expanded configurations. None of the control individuals tested presented an expanded AGAGG conformations. The frequency of the novel AAGAG expansion was found to be higher in cases than in controls (7.0% in cases and 2.1% in controls). The allele counts and Chi-square calculation values were shown in Supplementary Table 2. The distribution of the different conformations was found to be different in cases and controls ($P = 0.022$).

3.5. Discussion

This study represents a follow-up examination of the *RFC1* pentanucleotide repeat expansion recently found to cause adult-onset ataxia (Cortese *et al*., 2019). A total of 49 cases (26 Canadian and 23 Brazilian) for which genetic testing did not reveal the cause of the disease to be a previously identified ataxia gene and an unrelated cohort of 128 adult-onset Canadian cases for who no prior genetic test results was available. The nature and size of the conformation reported to be expanded in *RFC1* was examined using a RP-PCR and long-range PCR sequencing approach. Conversely to what was previously observed in the original study (Cortese *et al*., 2019), the AAGGG expansion explained a much smaller share of the Brazilian and Canadian cases examined here (0.6% and 4.3% of unrelated cases in Canadian and Brazilian cohorts, respectively by comparison to 22% sporadic Caucasian cases in the study by Cortese and colleagues) (Cortese *et al*., 2019).

The frequency of the four allelic repeat configurations has been described before, but only in control individuals with no history or signs of ataxia (Cortese *et al*., 2019). However, the *RFC1* repeat locus could not be assessed in 3% of this earlier examination in control individuals, an observation which led the authors to suggest the existence of additional allelic configurations.

Unlike the previous examinations of the pentanucleotide repeats of *RFC1* in the context of adult-onset ataxia (Cortese et al., 2019; Rafehi *et al*., 2019), we actually report the observation of two novel repeat conformations (AAGAG and AGAGG) in a heterozygous state. The frequency of the AAGAG repeat was observed to be 0.07 and 0.02 in Canadian cases and controls respectively. The observation of two previously unreported pentanucleotide repeats might

represents clues which will lead to a better understanding of the expansion mechanism leading to the pathogenic repeat of *RFC1*. While it is not a conclusive observation, the higher frequency of the novel AAGAG repeat in cases (by comparison to its frequency in control individuals) suggests that it could eventually be observed to also be associated with adult-onset ataxia. Hence additional work might be needed to determine the frequency of other pentanucleotide repeat conformations, and their association to adult-onset ataxia.

Given the dynamic nature of the *RFC1* repeat, multiple validations of sequences and repeat length should be performed. To prevent false positive results, the RP-PCR plots should be interpreted with caution, and each AAGGG-positive sample should be validated by Sanger sequencing to confirm its true sequence.

Disease-associated or wild type repeat interruptions have been observed across several expansion-associated diseases, such as SCA37 (Seixas *et al*., 2017; Loureiro *et al*., 2019), SCA10 (Matsuura *et al.*, 2006), and FRDA (Al-Mahdawi *et al*., 2018). Variations interrupting the pure repeat sequences of disease-causing alleles can affect their penetrance, as well as the age at onset and severity of the conditions associated with specific repeats (Al-Mahdawi *et al*., 2018) Also, interruptions in the normal alleles prevent the disease-associated expansions and provide the stability of repeats in disease-causing alleles. We did not observe the new sequence conformations along with an AAGGG expansion in any of the patients, therefore further studies will be required to determine whether they affect the function or the disease severity.

The low prevalence of the *RFC1* AAGGG expansion, as well as the identification of novel conformations might be due to the different genetic backgrounds of the Canadian and Brazilian populations (Dupré *et al*., 2006). Although we screened a control group of Canadian individuals to assess the frequency of each repeat conformation in the general population, a limitation of the current study is that the study cohort do not contain a Brazilian control group. Therefore, additional case and control cohorts should be tested for the same repeat, in order to draw a clear conclusion on its frequency in adult-onset ataxia. Further studies are warranted to confirm the structure and sequence of the repeated region, and to investigate potential biological impacts.

3.6. Acknowledgements

We would like to thank the participants of the study. We thank S. Can Akerman, Vessela Zaharieva and Hélène Catoire, for their assistance. GAR holds a Canada Research Chair and is funded by the CIHR.

3.7. Tables and figures

Table 1. Allele frequency of *RFC1* repeat expansions in Brazilian and Canadian ataxia cohorts

	Cohort 1 n = 23	Cohort 2 n = 26	Cohort 3 n = 128	Control group n = 163
Mean age at onset	37 ± 7	57 ± 10	50 ± 13	
Geographical origin	Brazil	Canada	Canada	Canada
male:female	13/10	9/17	NA	1
Family history (familial/sporadic)	familial	both	Both	-
Prior genetic testing for other common ataxias	+	+	-	-
(AAAAG)11	29 (63%)	36 (69.2%)	193 (75.4%)	276 (84.6 %)
(AAAAG)n	6 (13%)	1 (1.9%)	20 (7.8%)	22 (6.7 %) 1 homozygous, 20 heterozygous)
(AAAGG)n	3 (6.5%)	3 (5.8%)	8 (3.1%)	8 (2.5 %) (8 heterozygous)
(AAGGG)n	5 (10.9%, 1 biallelic, 3 heterozygous)	11 (21.1%, 11 heterozygous)	16 (6.2%) (1 biallelic, 14 heterozygous)	13 (4 %) (13 heterozygous)
(AAGAG)n	2 (4.3%)	1 (1.9%)	18 (7%)	7 (2.1%)
(AGAGG)n	1 (2.2%)	0	1 (0.4%)	0
Compound heterozygotes	(AAAAG)n/ (AGAGG)n (1)	0	(AAGAG)n/ (AGAGG)n (1), (AAAAG)n/ (AAAGG)n (1), (AAAGG)n/ (AAGGG)n (1)	(AAAAG)n/ (AGAAG)n (2), (AAAGG)n/(AAGGG)n (1)

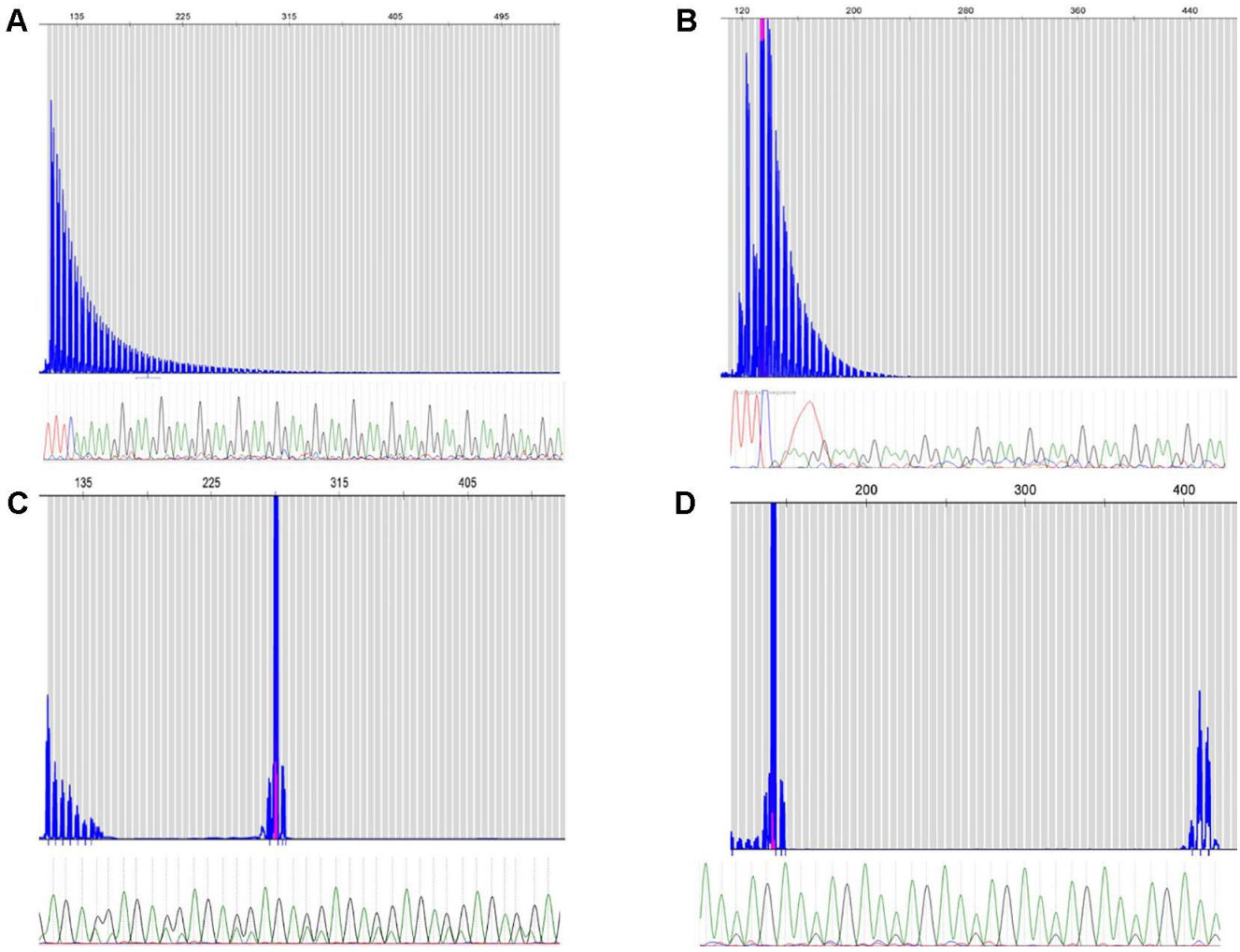


Figure 1. Repeat-primed PCR reactions targeting the AAGGG repeated conformation. Fragment plots and Sanger chromatograms of long-range PCR results are shown for biallelic $(AAGGG)_{exp}$ in Canadian (1a) and Brazilian (1b) patients. Novel AGAGG (1c) and AAGAG (1d) expansion conformations (heterozygous) required both methods for identification.

3.8. Supplemental materials

Supplementary Table 1. Clinical features of patients carrying the recessive AAGGG repeat expansion in *RFC1*

sample	origin	gender	family history	age at onset	age at examination	symptom at onset	neuropathy	cerebellar ataxia	nystagmus	cerebellar atrophy	SARA	other
Fam I-I	Brazilian	female	yes (affected sister, unaffected parents)	45	58	Dizziness, gait and balance problems	sensorimotor axonal polyneuropathy	yes	yes	yes	25	Dysarthria, brisk tendon reflexes, vestibular areflexia
Fam I-II	Brazilian	female	yes (affected sister, unaffected parents)	45	56	Dizziness, gait and balance problems	sensorimotor axonal polyneuropathy	yes	yes	yes	27	Dysarthria, brisk tendon reflexes, vestibular areflexia
Fam II-I	Italian	female	yes (affected brother)	55	58	Dizziness, gait and balance problems	none	yes	yes	yes	NA	Abnormal somatosensory evoked potentials, brisk tendon reflexes

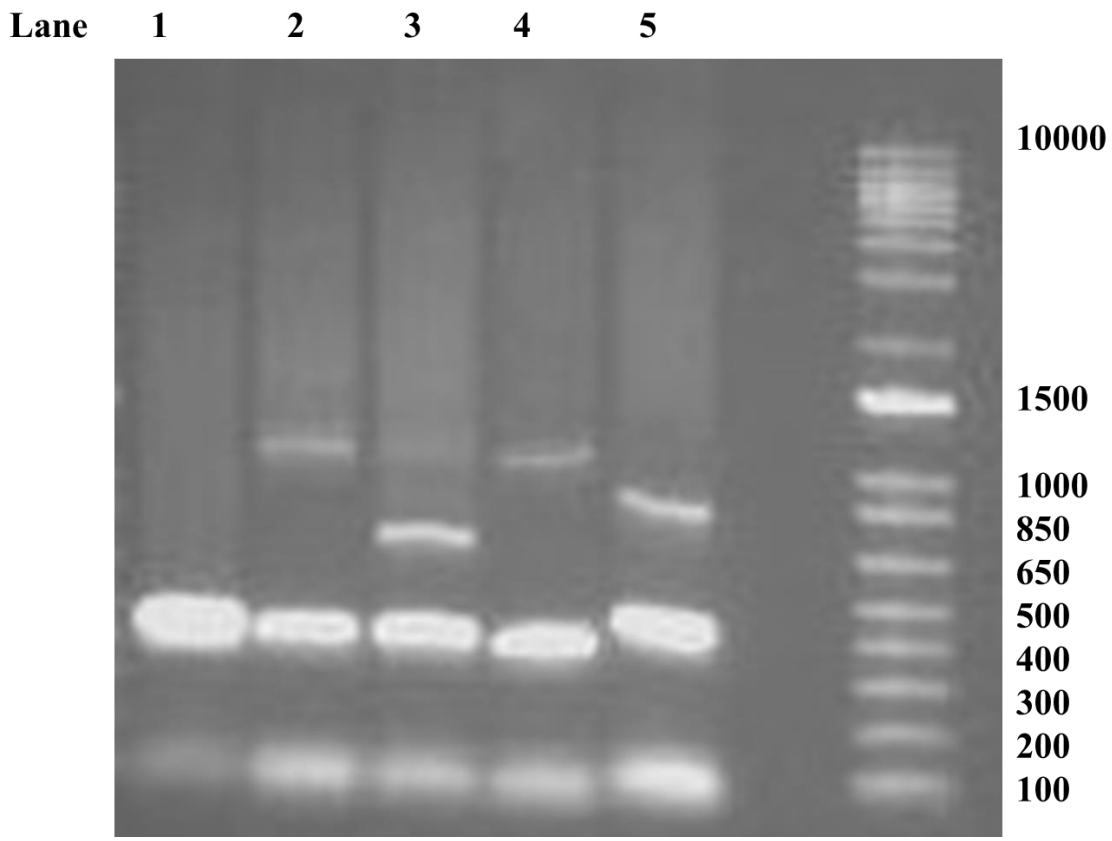
*SARA: Scale for the assessment and rating of ataxia. For vestibular areflexia, a video-head impulse test and caloric test reflex were performed.

Supplementary Table 2. The allele counts 2×5 contingency table and Chi- square calculations.

$\chi 2 = 11.429$, df = 4, $\chi 2/df = 2.86$, $P(\chi 2 > 11.429) = 0.0221$.

	Chi-square calculations for *RFC1* repeat conformations				
	$(AAAAG)_n$	$(AAAAG)_n$	$(AAGAG)_n$	$(AAAGG)_n$	$(AAGGG)_n$
Cases (256)	193 *205.84* (0.8)	20 *18.43* (0.13)	18 *10.97* (4.5)	8 *7.02* (0.14)	16 *12.73* (0.84)
Controls (326)	276 *263.16* (0.63)	22 *23.57* (0.1)	7 *14.03* (3.52)	8 *8.98* (0.11)	13 *16.27* (0.66)
	469	42	25	16	29

*Expected values are displayed in italics. Individual $\chi 2$ values are displayed in (parentheses).



Supplementary Figure 1. Long-range PCR amplification using Canadian control samples. Lane 1: wild type $(AAAAG)_{11}$, lane 2: heterozygous $(AAAAG)_{11}$ and $(AAAAG)_n$, lane 3: heterozygous $(AAAAG)_{11}$ (lower band) and $(AGAAG)_n$ (upper band), lane 4: heterozygous $(AAAAG)_{11}$ (lower band) and $(AAAGG)_n$ (upper band), lane 5: heterozygous $(AAAAG)_{11}$ (lower band) and $(AAGGG)_n$ (upper band).

3.9. References

Akçimen, F., Ross, J. P., Bourassa, C. V., Liao, C., Rochefort, D., Gama, M. T. D., et al. (2019). Investigation of the pathogenic *RFC1* repeat expansion in a Canadian and a Brazilian ataxia cohort: identification of novel conformations. bioRxiv. Cold Spring Harbor Lab. p, 593871. doi: 10.1101/593871

Al-Mahdawi, S., Ging, H., Bayot, A., Cavalcanti, F., La Cognata, V., Cavallaro, S., *et al*. (2018). Large interruptions of GAA repeat expansion mutations in Friedreich Ataxia are very rare. Front. Cell. Neurosci. 12, 443. doi: 10.3389/fncel.2018.00443

Cortese, A., Simone, R., Sullivan, R., Vandrovcova, J., Tariq, H., Yau, W., *et al*. (2019). Expansion of a recessive intronic AAGGG repeat in the *RFC1* gene is a common cause of late-onset ataxia. Nat. Genet. 51, 649–658. doi: 10.1038/s41588-019-0372-4

Dupré, N., Bouchard, J. P., Brais, B., Rouleau, G. A. (2006). Hereditary ataxia, spastic paraparesis and neuropathy in the French-Canadian population. Can. J. Neurol. Sci. 33, 149–157. doi: 10.1017/S031716710000490X

Loureiro, J. R., Oliveira, C. L., Mota, C., Castro, A. F., Costa, C., Loureiro, J. L., *et al.* (2019). Mutational mechanism for DAB1 (ATTTC)n insertion in SCA37: ATTTT repeat lengthening and nucleotide substitution. Hum. Mutat. 40, 1–9. doi: 10.1002/humu.23704

Matsuura, T., Fang, P., Pearson, C. E., Jayakar, P., Ashizawa, T., Roa, B. B., *et al*. (2006). Interruptions in the expanded ATTCT repeat of spinocerebellar ataxia type 10: repeat purity as a disease modifier? Am. J. Hum. Genet. 78, 125–129. doi: 10.1086/498654

Noreau, A., Dupre, N., Bouchard, J. P., Dion, P. A., Rouleau, G. A. (2013). "Autosomal recessive cerebellar ataxias," in Handbook of the cerebellum and cerebellar disorders. Eds. Manto, M., Gruol, D. L., Schmahmann, J. D., Koibuchi, N., Rossi, F. (New York: Springer Science+Business Media), 2177–2191. doi: 10.1007/978-94-007-1333-8_100

Okonechnikov, K., Golosova, O., Fursov, M., the UGENE team (2012). Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28, 1166–1167. doi: 10.1093/bioinformatics/bts091

Rafehi, H., Szmulewicz, D. J., Bennett, M. F., Sobreira, N. L. M., Pope, K., Smith, K. R., *et al.* (2019). Bioinformatics-based identification of expanded repeats: a non-reference intronic pentamer expansion in rfc1 causes canvas. Am. J. Hum. Genet. 105, 151–165. doi: 10.1016/j.ajhg.2019.05.016

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

Synofzik, M., Puccio, H., Mochel, F., Schöls, L. (2019). Autosomal recessive cerebellar ataxias: paving the way toward targeted molecular therapies. Neuron 101, 60–583. doi: 10.1016/j.neuron.2019.01.049

Bridging statement to Chapter 4

Within Chapter 2 and Chapter 3, we have focused on the association of different repeat expansions with ataxia through examining their length as well as sequence motifs.

In the fourth chapter, we addressed another aspect of expanded repeats which is their association with age at onset in spinocerebellar ataxia type 3, also known as Machado-Joseph disease. First, we assessed the correlation between repeat length and age at onset. Next, we performed a genome-wide association study to identify possible genetic factors that may explain the variable age at onset among patients.

CHAPTER 4: GENOME-WIDE ASSOCIATION STUDY IDENTIFIES GENETIC FACTORS THAT MODIFY AGE AT ONSET IN MACHADO-JOSEPH DISEASE

Fulya Akçimen[1,2], Sandra Martins[3,4], Calwing Liao[1,2], Cynthia V. Bourassa[2,5], Hélène Catoire[2,5], Garth A. Nicholson[6], Olaf Riess[7], Mafalda Raposo[8], Marcondes C. França Jr.[9], João Vasconcelos[10], Manuela Lima[8], Iscia Lopes-Cendes[11,12], Maria Luiza Saraiva-Pereira[13,14], Laura B. Jardim[13,15], Jorge Sequeiros[4,16,17], Patrick A. Dion[2,5], Guy A. Rouleau[1,2,5*]

Affiliations

[1]Department of Human Genetics, McGill University, Montréal, QC, Canada

[2]Montreal Neurological Institute and Hospital, McGill University, Montréal, QC, Canada

[3]i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal

[4]IPATIMUP – Institute of Molecular Pathology and Immunology of the University of Porto, Portugal

[5]Department of Neurology and Neurosurgery, McGill University, Montréal, QC, Canada

[6]University of Sydney, Department of Medicine, Concord Hospital, Australia

[7]Institute of Medical Genetics and Applied Genomics, University of Tuebingen, Tuebingen, Germany

[8]Faculdade de Ciências e Tecnologia, Universidade dos Açores e Instituto de Biologia Molecular e Celular (IBMC), Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Portugal

[9]Department of Neurology, Faculty of Medical Sciences, UNICAMP, Campinas, SP, Brazil

[10]School of Medical Sciences, Department of Medical Genetics and Genomic Medicine, University of Campinas (UNICAMP), Campinas, SP, Brazil

[11]The Brazilian Institute of Neuroscience and Neurotechnology (BRAINN), Campinas, SP, Brazil

[12]Departamento de Neurologia, Hospital do Divino Espírito Santo, Ponta Delgada, Portugal

[13]Medical Genetics Service, Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, Brazil

[14]Depto. de Bioquímica – ICBS, Universidade Federal do Rio Grande do Sul (UFRGS)

[15]Depto de Medicina Interna, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

[16]Institute for Molecular and Cell Biology (IBMC), Universidade do Porto, Porto, Portugal

[17]Instituto de Ciências Biomédicas Abel Salazar (ICBAS), Universidade do Porto, Portugal.

*Corresponding author: Guy A. Rouleau, Montreal Neurological Institute and Hospital, 3801 University Street, Room 636, Montréal, Québec H3A2B4, Canada; e-mail: guy.rouleau@mcgill.ca, phone: +1 (514) 398-6644

Published in

Akçimen F, Martins S, Liao C, Bourassa CV, Catoire H, Nicholson GA, Riess O, Raposo M, França MC, Vasconcelos J, Lima M, Lopes-Cendes I, Saraiva-Pereira ML, Jardim LB, Sequeiros J, Dion PA, Rouleau GA. (2020) Genome-wide association study identifies genetic factors that modify age at onset in Machado-Joseph disease. Aging (Albany NY). 23;12(6):4742-4756. doi: 10.18632/aging.102825. Epub 2020 Mar 23. PMID: 32205469; PMCID: PMC7138549.

4.1. Abstract

Machado-Joseph disease (MJD/SCA3) is the most common form of dominantly inherited ataxia worldwide. The disorder is caused by an expanded CAG repeat in the *ATXN3* gene. Past studies have revealed that the length of the expansion partly explains the disease age at onset (AO) variability of MJD, which is confirmed in this study (Pearson's correlation coefficient $R^2 = 0.62$). Using a total of 786 MJD patients from five different geographical origins, a genome-wide association study (GWAS) was conducted to identify additional AO modifying factors that could explain some of the residual AO variability. We identified nine suggestively associated loci ($P < 1 \times 10^{-5}$). These loci were enriched for genes involved in vesicle transport, olfactory signaling, and synaptic pathways. Furthermore, associations between AO and the *TRIM29* and *RAG* genes suggests that DNA repair mechanisms might be implicated in MJD pathogenesis. Our study demonstrates the existence of several additional genetic factors, along with CAG expansion, that may lead to a better understanding of the genotype-phenotype correlation in MJD.

4.2. Introduction

Machado-Joseph disease, also known as spinocerebellar ataxia type 3 (MJD/SCA3), is an autosomal dominant neurodegenerative disorder that is characterized by progressive cerebellar ataxia and pyramidal signs, which can be associated with a complex clinical picture and includes extrapyramidal signs or amyotrophy (Twist *et al*., 1995; Bettencourt *et al*., 2011). MJD is caused by an abnormal CAG trinucleotide repeat expansion in exon 10 of the ataxin-3 gene (*ATXN3*), located at 14q32.1. Deleterious expansions $(CAG)_{exp}$ consensually contain 61 to 87 CAG repeats, whereas wild type alleles $(CAG)_{nor}$ range from 12 to 44 (Bettencourt *et al*., 2011).

As with other diseases caused by repeat expansions, such as Huntington's disease (HD) and other spinocerebellar ataxias, there is an inverse correlation between expanded repeat size and the age at which pathogenesis leads to disease onset (Maciel *et al*., 1995). Depending on the cohort structure, the size of the repeat expansion explains 55 to 70% of the age at onset (AO) variability in MJD, suggesting the existence of additional modifying factors (Maciel *et al*., 1995; de Mattos *et al*., 2019). Although several genetic factors have been proposed as modifiers, such as CAG repeat size of normal *ATXN3* (SCA3), *HTT* (HD), *ATXN2* (SCA2) and *ATN1* (DRPLA) alleles, *APOE* status, and expression level of *HSP40* (de Mattos *et al*., 2019; Zijlstra *et al*., 2010; Tezenas du Montcel *et al*., 2014), these were not replicated by subsequent studies (Chen *et al*., 2016; Raposo *et al.*, 2015). Since CAG tract profile and allelic frequencies of the potential modifier loci can have unique characteristics in different populations, large collaborative studies are required to identify genetic modifiers in MJD, as well as replicate the findings of such studies Raposo *et al*., 2015).

Previously, Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium carried out a GWA approach of HD individuals to reveal genetic modifiers of AO in HD (Lee *et al*., 2015; Lee *et al*., 2019). A total of eleven (Lee *et al*., 2015) and fourteen loci (Lee *et al*., 2019) were found to be associated with residual age at HD onset. In the present study, we performed the first GWAS to identify some possible genetic modifiers of AO in MJD. First, we assessed the relationship between AO and size of the expanded $(CAG)_{exp}$ and normal $(CAG)_{nor}$ alleles, biological sex, and geographical origin. Next, we determined a residual AO for each subject, which is the difference between the measured AO and the predicted/estimated AO from expanded CAG repeat size alone. Using the residuals as a quantitative phenotype for a GWAS, we looked for genetic factors that modulate AO in MJD.

4.3. Results

4.3.1. The inverse correlation between $(CAG)_{exp}$ and age at onset

In the first phase of the study, the expanded *ATXN3*-CAG repeat lengths of 786 MJD patients were assessed. The mean (SD) $(CAG)_{exp}$ size were Australia: 68.2 (±3.3), Brazil: 74.3 (3.9), Germany: 72.9 (±3.6), North America: 73 (±4.3) and Portugal: 72 (±4.0). Next, the relationship between AO and $(CAG)_{exp}$ size, $(CAG)_{nor}$ size, sex and ethnicity was examined (Supplementary Table 1). The previously observed negative correlation between *ATXN3* $(CAG)_{exp}$ size and AO (Maciel *et al*., 1995) was confirmed (Pearson's correlation coefficient $R^2 = 0.62$) (Figure 1). The $(CAG)_{nor}$ size (P = 0.39), sex (P = 0.02) and geographic origin (P [Brazil] = 0.38, P [Germany] = 0.38, P [North America] = 0.33, P [Portugal] = 0.29) were not significant and their addition had little contribution to the model ($\Delta R^2 = 0.0072$). Residual AO for each sample was calculated and used as a quantitative phenotype to identify the modifiers of AO. The distribution of residual AO was close a theoretical normal distribution (Figure 1).

4.3.2. Genome-wide association study

After post-imputation quality assessments, a total of 700 individuals with genotyping information for 6,716,580 variants remained for GWAS. The Manhattan plots are shown in Figure 2. The genomic inflation factor was close to one ($\lambda = 0.98$), indicating the p-values were not inflated. Genome-wide suggestive associations ($P < 1 \times 10^{-5}$) with 204 variants across 9 loci were identified (Supplementary Table 3). The most significantly associated SNP at each locus are shown in Table 1. Positional gene mapping aligned SNPs to 17 genes by their genomic location. Fourteen of the 204 variants had a Combined Annotation Dependent Depletion (CADD)-PHRED score higher than the suggested threshold for deleterious SNPs (12.37), arguing the given loci have a functional role (Amendola et al., 2015).

4.3.3. Interaction analysis between $(CAG)_{exp}$, sex and SNP genotype

To assess a possible interaction between CAG_{exp} size and the variants identified, each of the nine variants was added to the initial linear regression, modelling AO as a function of $(CAG)_{exp}$ size, SNP, sex, the first three principal components, $(CAG)_{nor}$ size, interactions of SNP:$(CAG)_{exp}$ and SNP:sex. Association of each independent SNP with AO revealed nominally significant p-values (P [rs7480166] = 8.42×10^{-6}, P [rs62171220] = 6.33×10^{-3}, P [rs2067390] = 4.51×10^{-5}, P [rs144891322] = 1.14×10^{-5}, P [rs11529293] = 1.62×10^{-5}, P [rs585809] = 2.91×10^{-5}, P [rs72660056] = 1.66×10^{-3}, P [rs11857349] = 8.21×10^{-6}, P [rs8141510] = 1.33×10^{-3}). With the addition of the identified variants to the model, correlation coefficient R^2 increased to 0.71 ($\Delta R^2 = 0.082$). Among the nine variants, only rs585809 (mapped to *TRIM29*) had a significant interaction with CAG_{exp} ($P = 0.01$), suggesting that rs585809 might modulate AO through this epistatic

interaction on $(CAG)_{exp}$. The addition of SNP:sex interaction had little contribution to the model ($\Delta R^2 = 0.005$).

4.3.4. Association of HD-AO modifier variants in MJD

Association of previously identified HD-AO modifier loci in MJD were assessed. Among the 25 HD-AO modifier variants in 17 loci, a total of 18 variants (MAF > 0.02) in 12 loci were tested in this study (Supplementary Table 4). None of these HD-AO modifiers reached the genome-wide suggestive threshold. However, two variants rs144287831 ($P = 0.02$, effect size = - 0.98) and rs1799977 ($P = 0.02$, effect size = - 0.98) in the *MLH1* locus were found to be nominally associated with a later AO in MJD.

4.3.5. Pathway and gene-set enrichment analysis

A gene-set enrichment and pathway analysis was conducted using i-GSEA4GWAS v2 (Zhang *et al*., 2010). Various approaches and algorithms are currently in use to conduct similar analyses. To be able to make better comparisons with other studies that may use different approaches, we performed a secondary gene-set enrichment and pathway analysis using the VEGAS2 (Mishra *et al*., 2015) and PASCAL (Lamparter *et al.*, 2016) software (Supplementary Tables 5-7). We also used these results for replication purposes in our own study. A total of 13 overrepresented pathways were found, after FDR-multiple testing correction (q-value < 0.05) in the primary GSEA analysis and replicated using at least one of the secondary gene-set enrichment algorithms (Table 2). Overall, the most significantly enriched gene-sets and pathways were vesicle transport, olfactory signaling, and synaptic pathways. Visualization and clustering of pathways are shown in Figure 3.

4.4. Discussion

Using five cohorts from different geographical origins, we performed the first GWAS to examine the presence of genetic factors that could modify AO in MJD. We identified a total of nine loci that were potentially associated with either an earlier or later AO. Concomitantly, we confirmed the previously observed negative correlation between $(CAG)_{exp}$ and AO (Maciel *et al*., 1995). It was shown previously that normal *ATXN3* allele $(CAG)_{nor}$ had a significant influence on AO of MJD (França MC *et al*., 2012); however, several studies did not replicate this effect (Tezenas du Montcel *et al*., 2014; Raposo *et al*., 2015). Indeed, we did not observe an association between $(CAG)_{nor}$ and AO. However, it had little contribution to our model, with a minor difference in the correlation coefficient ($\Delta R^2 = 0.0012$).

In our GWAS, the strongest signal is for the rs11529293 variant ($P = 3.30 \times 10^{-6}$) within the *C11orf72* and *RAG* loci at 11p12. Within this locus, two *RAG* genes, recombination-activating genes *RAG1* and *RAG2*, were shown to be implicated in DNA damage response and DNA repair machineries (Lescale *et al*., 2016; Bahjat *et al*., 2017). The rs585809 variant, which was mapped to the *TRIM29* gene, was found to interact with $(CAG)_{exp}$, suggesting that it might have an effect on AO through this interaction. Both *RAG* and *TRIM29* loci were identified as AO-hastening modifiers. *TRIM29* encodes for tripartite motif protein 29, which is implicated in mismatch repair and double strand breaks pathways (Wikiniyadhanee *et al*., 2017; Masuda *et al*., 2015). TRIM29 is involved both upstream and downstream of these pathways, in the regulation of DNA repair proteins into chromatin by mediating the interaction between them. One of these DNA repair proteins is MLH1, which is implicated in mismatch repair complex (Masuda *et al*., 2015). Previously, the *MLH1* locus was identified as an AO modifier in another neurodegenerative disease caused by CAG repeat expansion, Huntington's disease (Lee *et al*., 2015; Lee *et al*., 2019;

Lee *et al.*, 2017). Additionally, in a genome-wide genetic screening study, MLH1-knock out was shown to modify the somatic expansion of the CAG repeat and slow the pathogenic process in HD mouse model (Pinto *et al*., 2013). Overall, the association of *TRIM29* and *RAG* loci suggests that DNA repair mechanisms may be implicated in the alteration of AO of MJD, as well as HD, and may have a role in the pathogenesis of other CAG repeat diseases. Interestingly, in a previous study, we found variants in three transcription-coupled repair genes (*ERCC6*, *RPA3*, and *CDK7*) associated with different CAG instability patterns in MJD (Martins *et al*., 2014).

We identified gene-sets enriched in olfactory signaling, vesicle transport, and synaptic pathways. Olfactory dysfunction is one of the main non-motor symptoms that was already described in patients with MJD (Braga-Neto *et al.*, 2011; Pedroso *et al*., 2013. In a previous study, transplantation of olfactory ensheathing cells, which are specialized glial cells of the primary olfactory system, were found to improve motor function in an MJD mice model and were suggested as a novel potential strategy for MJD treatment (Hsieh *et al*., 2017). Vesicle transport and synaptic pathways were also implicated in MJD, as well as in other neurodegenerative diseases (Wiatr *et al*., 2019; Gissen *et al*., 2007). An interruption of synaptic transmission caused by an expanded polyglutamine repeat and mutant ataxin-3 aggregates were shown in *Drosophila* and *Caenorhabditis elegans* models of MJD. Therefore, the interaction between synaptic vesicles and mutant aggregates supports the role of synaptic vesicle transport in the pathogenesis of MJD (Gunawardena *et al*., 2005; Khan *et al*., 2006). Overall, we suggest that these gene-sets and pathways might construct a larger molecular network that could modulate the AO in MJD.

In summary, our study identified nine genetic loci that may modify the AO of MJD. Identification of *TRIM29* and *RAG* genetic variants, as well as our gene-set enrichment analyses, implicated DNA repair, olfactory signaling, synaptic, and vesicle transport pathways in the

pathogenesis of MJD. Although we used different cohorts from five distinct geographical ethnicities, a replication study in similar or additional populations would add valuable evidence to support our findings.

4.5. Methods

4.5.1. Study subjects

A total of 786 MJD patients from five distinct geographical origins (Portugal, Brazil, North America, Germany, and Australia) were included in the present study. The overall average age at onset (standard deviation) was 38 (± 1.82) years, with a 1:1 male to female ratio. All subjects provided informed consent, and the study was approved by the respective institutional review boards. Detailed cohort demographics are shown in Supplementary Table 2.

4.5.2. Assessment of the *ATXN3* CAG repeat length

A singleplex polymerase chain reaction was performed to determine the length of the $(CAG)_{exp}$ and $(CAG)_{nor}$ alleles at exon 10 of *ATXN3* (Martins *et al*., 2006). The final volume for each assay was 10 µL: 7.5 ng of gDNA, 0.2 µM of each primer, 5 µL of Taq PCR Master Mix Kit Qiagen®, 1 µL of Q-Solution from Qiagen® and H_2O. Fragment length analysis was done using ABIPrism 3730xl sequencer (Applied Biosystems®, McGill University and Genome Québec Innovation Centre) and GeneMapper software (Chatterji *et al.*, 2006). A stepwise regression model was performed to assess the correlation between AO and $(CAG)_{exp}$ size, as well as gender, origin, $(CAG)_{nor}$ size, and interaction between these variables. Residual AO was calculated for each subject by subtracting individual's expected AO based upon $(CAG)_{exp}$ size from actual AO, to be used as the primary phenotype for following genetic approach.

4.5.3. Genotyping, quality control and imputation

Samples were genotyped using the Global Screening Array v.1.0 from Illumina (636,139 markers). Sample-based (missingness, relatedness, sex, and multidimensional scaling analysis) and SNP-based quality assessments (missingness, Hardy-Weinberg equilibrium, and minor allele frequency) were conducted using PLINK version 1.9 (Chang *et al*., 2015). In sample level QC, samples were excluded with one or more of the following: high missingness (missingness rate > 0.05), close relationship (pi-hat value > 0.2), discrepancy between genetically-inferred sex and reported sex, population outliers (deviation $\geqslant$ 4 SD from the population mean in multidimensional scaling analysis). All SNPs were checked for marker genotyping call rate (> 98%), minor allele frequency (MAF) > 0.05, and HWE (p-value threshold = 1.0×10^{-5}).

Phasing and imputation were performed using SHAPEIT (Delaneau *et al*., 2011) and PBWT (Durbin *et al*., 2014) pipelines, implemented on the Sanger Imputation Service (McCarthy *et al.*, 2016). Haplotype Reference Consortium (HRC) reference panel r1.1 containing 64,940 human haplotypes at 40,405,505 genetic markers were used as the reference panel. Imputed variants with an allele count of 30 (MAF > 0.02), an imputation quality score above 0.3 and an HWE p-value of > 1.0×10^{-5} were included for subsequent analysis.

4.5.4. Genome-wide association analysis

A genome-wide linear mixed model based association analysis was conducted using –mlma-loco option of GCTA version 1.91.7 (Yang *et al.*, 2011). Residual AO was modelled as a function of minor allele count of the test SNP, sex, and the first three principal components based on the scree plot (Supplementary Figure 1). Manhattan plots were generated in FUMA v.1.3.4

(Watanabe *et al*., 2017). Regional association plots were generated using LocusZoom (Pruim *et al.*, 2010) (Supplementary Figure 2).

4.5.5. Functional annotation of SNPs

Genomic risk loci were defined using SNP2GENE function implemented in FUMA. Independent suggestive SNPs ($P < 1 \times 10^{-5}$) with a threshold of r2 < 0.6 were selected within a 250 kb window. The UK Biobank release 2 European population consisting of randomly selected 10,000 subjects was used as the reference population panel. The ANNOVAR (Wang *et al*., 2010) categories and combined annotation-dependent depletion (CADD) (Rentzsch *et al*., 2019) scores were obtained from FUMA for functional annotation. Functionally annotated variants were mapped to genes based on genomic position using FUMA positional mapping tool.

4.5.6. Pathway analysis

To identify known biological pathways and gene sets at the associated loci, an enrichment approach was applied using public datasets containing Gene Ontology (GO, http://geneontology.org), the Kyoto Encyclopaedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg) and Reactome (https://reactome.org) pathways. The primary enrichment analysis was performed using the i-GSEA4GWAS v2. It uses a candidate list of a genome-wide set of genes mapped within the SNP loci and ranks them based on the strength of their association with the phenotype. Genes were mapped within 20 kb up or downstream of the SNPs with a $P < 0.05$. Gene and pathway sets meeting a false discovery rate *(FDR)*-corrected *q*-value < 0.05 were regarded as significantly associated with high confidence, and *q*-value < 0.25 was regarded to be possibly associated with the phenotype of interest. We performed a secondary gene-based association test using the Versatile Gene-based Association Study (VEGAS) algorithm that controls the number of SNPs in each gene and the linkage disequilibrium (LD) between these

SNPs using the HapMap European population. As a third algorithm to identify enriched pathways, we used Pathway Scoring Algorithm (PASCAL), which controls for potential bias from gene size, SNP density, as well as LD. ClueGO (Bindea *et al.*, 2009) and CluePedia (Bindea *et al*., 2013) plug-ins in Cytoscape were employed to visualize identified pathways and their clustering.

4.6. Acknowledgements

The authors thank the participants for their contribution to the study. The authors would like to thank Jay P. Ross, Faezeh Sarayloo, Zoe Schmilovich and S. Can Akerman for their assistance in reviewing the manuscript and scientific content.

4.7. Tables and figures

Table 1. Suggestive loci associated with residual age at onset in MJD. Chr: chromosome, MAF: minor allele frequency, 1KGP: 1000 Genomes Project.

SNP	Chr	Position (GRCh37)	Nearest gene	Minor allele	Major allele	MJD MAF	1KGP MAF	b (SNP effect)	P-value
rs62171220	2	137802855	*THSD7B*	G	C	0.13	0.11	2.71	4.45×10^{-6}
rs2067390	2	191209028	*HIBCH, INPP1*	A	T	0.04	0.06	4.74	6.39×10^{-6}
rs144891322	5	85135387	*RPL5P17,*	C	T	0.02	0.007	6.10	5.18×10^{-6}
rs11529293	11	36855388	*C11orf74,RAG1, RAG2*	T	C	0.14	0.26	-2.71	3.30×10^{-6}
rs7480166	11	42984753	*HNRNPKP3*	A	G	0.40	0.40	-1.86	4.17×10^{-6}
rs585809	11	119949979	*TRIM29*	T	C	0.06	0.17	-3.76	9.50×10^{-6}
rs72660056	13	113507543	*ATP11A*	A	G	0.08	0.05	-3.29	3.94×10^{-6}
rs11857349	15	99924857	*TTC23,SYNM, LRRC28*	G	A	0.04	0.02	-4.58	3.43×10^{-6}
rs8141510	22	42821185	*NFAM1,CYP2D6, NAGA, NDUFA6*	C	T	0.43	0.49	1.83	3.94×10^{-6}

Table 2. Pathways significant after multiple-correction ($q < 5 \times 10^{-2}$) in the primary GSEA analysis and replicated using at least one of the secondary gene-set enrichment algorithms. NA means that the pathway was not enriched by at least two significant genes in VEGAS.

Pathway	Description	p-value (GSEA)	q-value (GSEA)	p-value (VEGAS)	permuted p-value (VEGAS)	p-value (PASCAL)
GO:0030133	transport vesicle	$< 1.0 \times 10^{-3}$	8.20×10^{-3}	6.15×10^{-40}	4.46×10^{-1}	6.70×10^{-3}
KEGG:04740	olfactory transduction	$< 1.0 \times 10^{-3}$	8.30×10^{-3}	NA	NA	3.89×10^{-4}
R-HSA:381753	olfactory signaling pathway	$< 1.0 \times 10^{-3}$	8.80×10^{-3}	1.10×10^{-27}	7.71×10^{-1}	2.51×10^{-4}
GO:0044456	synapse part	$< 1.0 \times 10^{-3}$	9.30×10^{-3}	1.25×10^{-182}	$< 1.0 \times 10^{-6}$	$< 1.0 \times 10^{-7}$
R-HSA:74217	purine salvage	$< 1.0 \times 10^{-3}$	1.06×10^{-2}	1.06×10^{-2}	2.15×10^{-1}	6.48×10^{-3}
GO:0045202	Synapse	$< 1.0 \times 10^{-3}$	1.15×10^{-2}	1.15×10^{-2}	$< 1.0 \times 10^{-6}$	$< 1.0 \times 10^{-7}$
GO:0004177	aminopeptidase activity	$< 1.0 \times 10^{-3}$	1.50×10^{-2}	1.50×10^{-2}	3.41×10^{-1}	1.24×10^{-2}
GO:0008238	exopeptidase activity	$< 1.0 \times 10^{-3}$	1.80×10^{-2}	1.80×10^{-2}	2.80×10^{-2}	8.31×10^{-3}
GO:0006898	receptor mediated endocytosis	$< 1.0 \times 10^{-3}$	2.25×10^{-2}	2.25×10^{-2}	2.03×10^{-1}	6.64×10^{-3}
GO:0016917	GABA receptor activity	$< 1.0 \times 10^{-3}$	2.26×10^{-2}	2.26×10^{-2}	1.30×10^{-4}	2.30×10^{-5}
GO:0030140	trans Golgi network transport vesicle	$< 1.0 \times 10^{-3}$	2.36×10^{-2}	2.36×10^{-2}	2.80×10^{-2}	1.28×10^{-1}
GO:0009725	response to hormone stimulus	$< 1.0 \times 10^{-3}$	2.73×10^{-2}	2.73×10^{-2}	1.32×10^{-1}	1.30×10^{-4}
GO:0030425	Dendrite	$< 1.0 \times 10^{-3}$	3.86×10^{-2}	3.86×10^{-2}	$< 1.0 \times 10^{-6}$	$< 1.0 \times 10^{-7}$

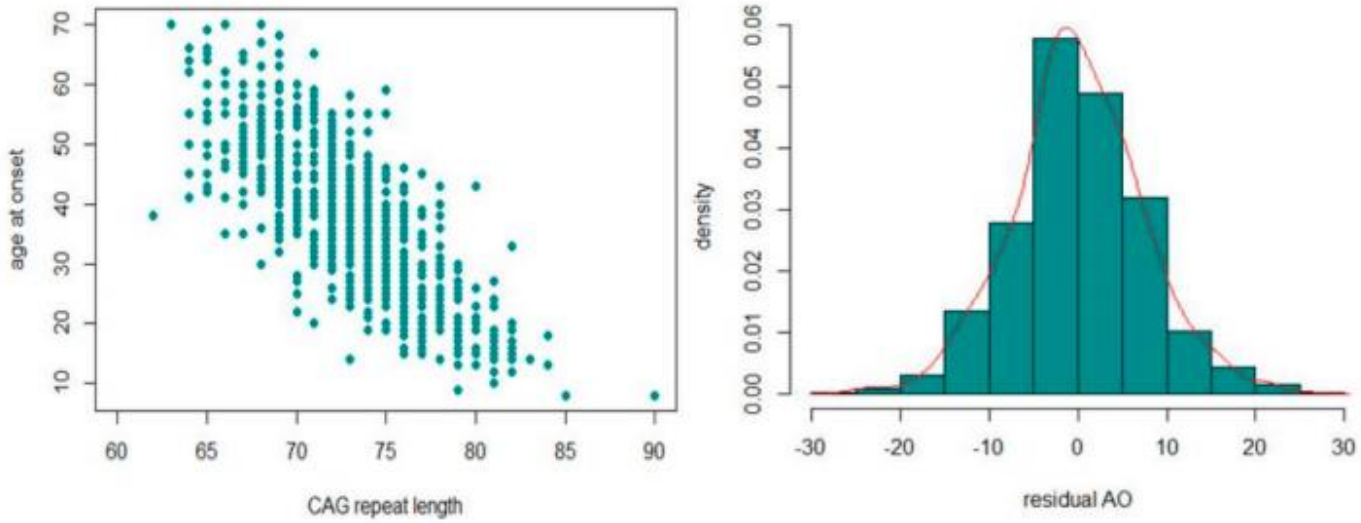


Figure 1. The inverse correlation between $(CAG)_{exp}$ and AO (left) and the distribution of residual AO (right) observed in our MJD cohort.

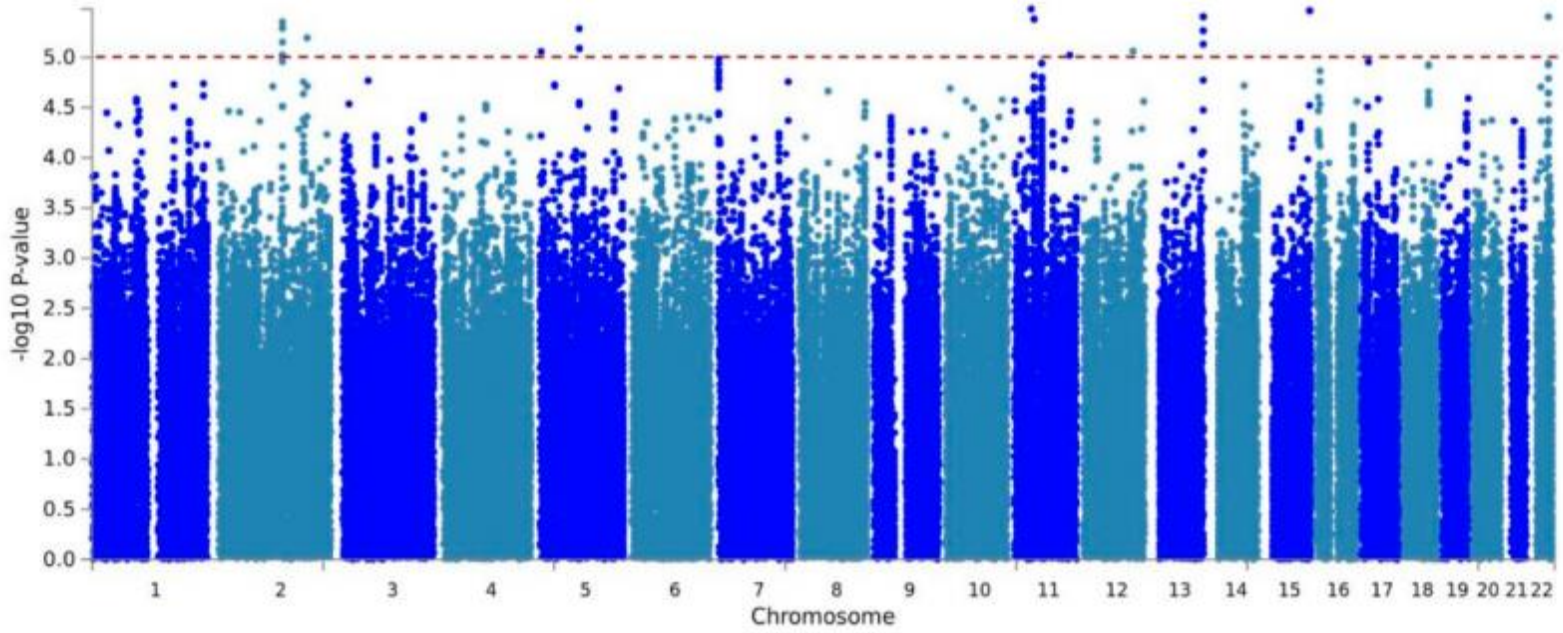


Figure 2. Manhattan plot of the GWAS for residual AO of MJD. Imputed using the HRC panel, 6,716,580 variants that passed QC are included in the plot. The x-axis shows the physical position along the genome. The y-axis shows the $-\log_{10}$(p-value) for association. The red line indicates the level of genome-wide suggestive association ($P = 1 \times 10^{-5}$).

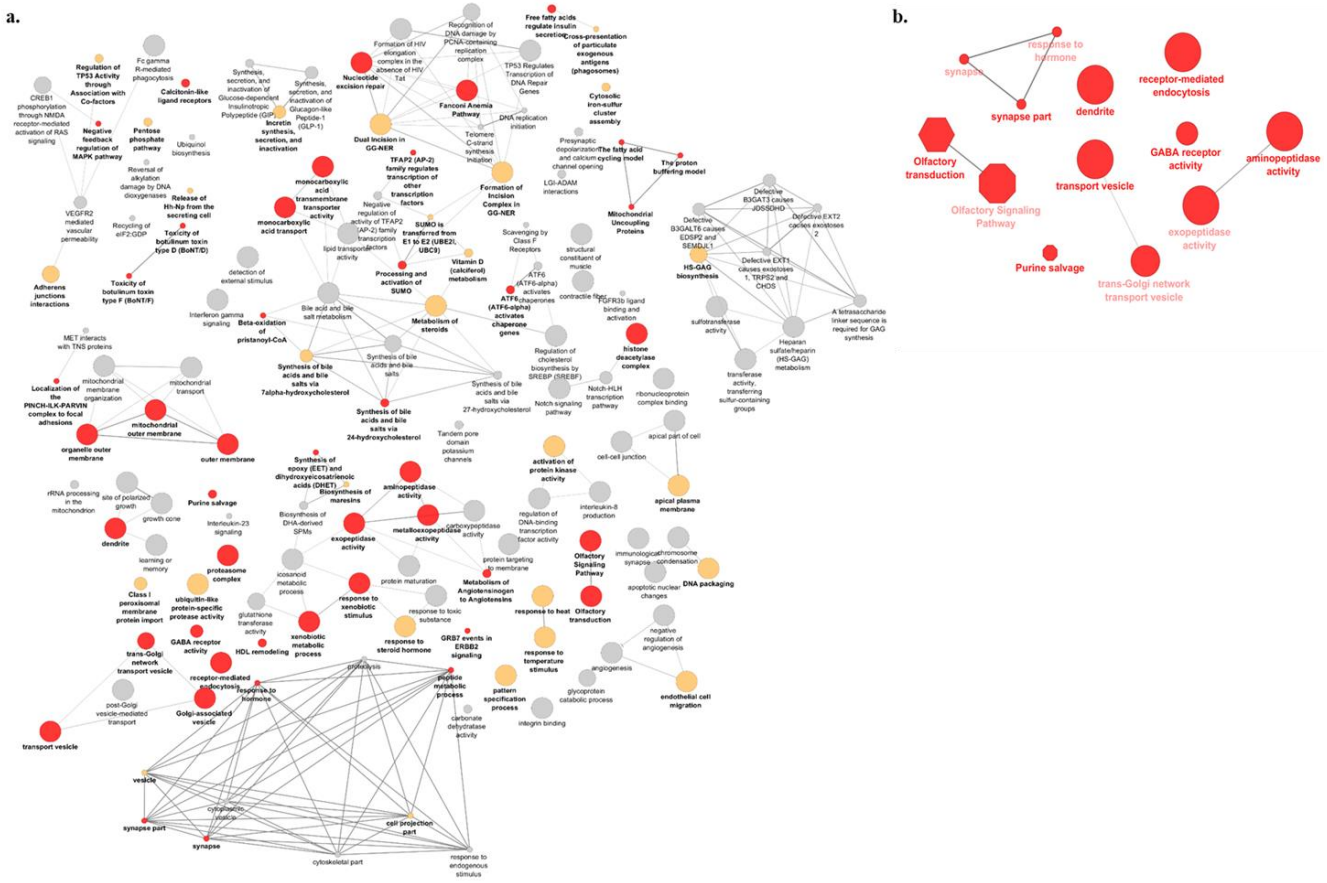


Figure 3. Visualization of the gene-sets and pathways enriched in primary GSEA analysis (a) and replicated in VEGAS and PASCAL (b). The size of the nodes corresponds to the number of the genes associated with a term. The significance is represented by the color of the nodes ($P < 0.05$, $0.05 < P < 0.1$ and $P > 0.1$ are represented by red, yellow, and gray, respectively).

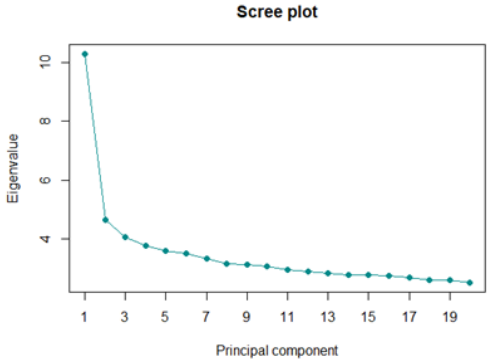
4.8. Supplemental materials

Supplementary Table 1. Linear relationship between AO and $(CAG)_{exp}$, $(CAG)_{nor}$, geographical origin, sex and pairwise interaction of the given factors. A total of 62.7 % of the variability in the AO is explained by given factors.

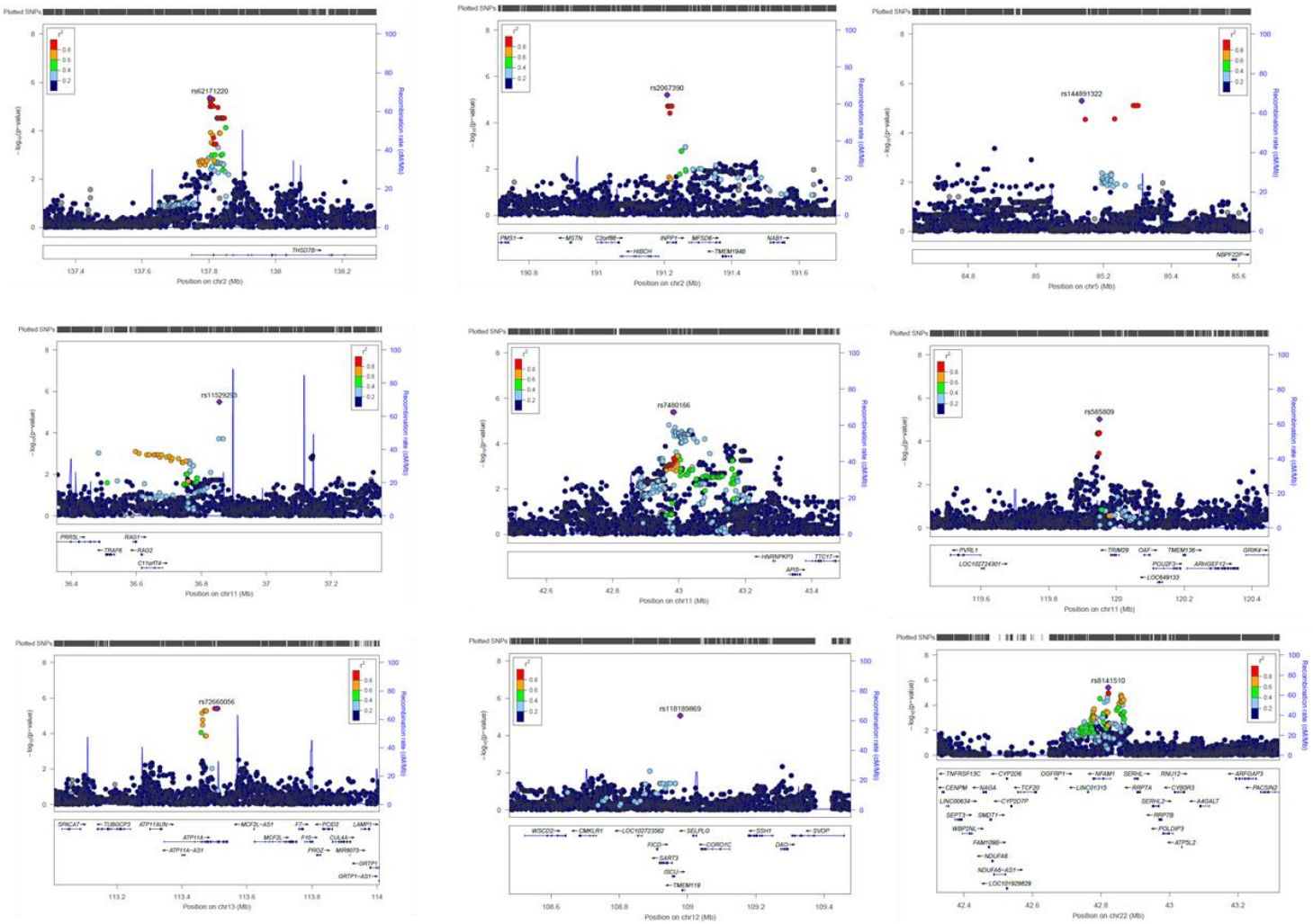
Model description	Multiple R^2	Adjusted R^2	P-value	ΔR^2
AO ~ $(CAG)_{exp}$	0.6200	0.6195	$<2.2 \times 10^{-16}$	
AO ~ $(CAG)_{exp}$ + origin	0.6241	0.6216	$<2.2 \times 10^{-16}$	0.0021
AO ~ $(CAG)_{exp}$ + origin + sex	0.6265	0.6235	$<2.2 \times 10^{-16}$	0.0019
AO ~ $(CAG)_{exp}$ + origin + sex + $(CAG)_{nor}$	0.6282	0.6247	$<2.2 \times 10^{-16}$	0.0012
AO ~ $(CAG)_{exp}$ + origin + sex + $(CAG)_{nor}$ + $(CAG)_{exp}$:$(CAG)_{nor}$	0.6301	0.6261	$<2.2 \times 10^{-16}$	0.0014
AO ~ $(CAG)_{exp}$ + origin + sex + $(CAG)_{nor}$ + $(CAG)_{exp}$:$(CAG)_{nor}$ + $(CAG)_{exp}$:origin	0.6328	0.6267	$<2.2 \times 10^{-16}$	0.0006
AO ~ $(CAG)_{exp}$ + origin + sex + $(CAG)_{nor}$ + $(CAG)_{exp}$:$(CAG)_{nor}$ + $(CAG)_{exp}$:origin + $(CAG)_{nor}$:origin	0.6352	0.6271	$<2.2 \times 10^{-16}$	0.0004

Supplementary Table 2. Subjects and cohort demographics. M:F - male-female ratio

Geographical origin	# of patients	Mean (SD) AO	M:F
Portugal	330	40.0 (±12.4)	1.0
Brazil	311	34.9 (±11.7)	1.1
North America	55	37.8 (±12.2)	0.7
Germany	51	37.6 (±9.2)	1.2
NA	34	37.1 (±11.1)	1.4
Australia	5	52.8 (±10.1)	0.3



Supplementary Figure 1. Scree plot showing the eigenvalues of the first 20 principal components (PCs). This plot indicates that the first three PCs explain the majority of the variability in data.



Supplementary Figure 2. Regional LocusZoom plots for the nine modifier loci that modify AO of MJD. Purple line indicates the genetic recombination rate (cM/Mb). SNPs in linkage disequilibrium with identified are shown in color gradient indicating r2 levels (hg19, 1KGP, Nov 2014, EUR).

Additional supporting information (Supplementary Tables 3, 4, 5, 6 and 7) can be downloaded from the online version of this article at the publisher's web-site:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7138549/

4.9. References

Amendola LM, Dorschner MO, Robertson PD, Salama JS, Hart R, Shirts BH, Murray ML, Tokita MJ, Gallego CJ, Kim DS, Bennett JT, Crosslin DR, Ranchalis J, *et al.* Actionable exomic incidental findings in 6503 participants: challenges of variant classification. Genome Res. 2015; 25:305–15. 10.1101/gr.183483.114

Bahjat M, Guikema JE. The Complex Interplay between DNA Injury and Repair in Enzymatically Induced Mutagenesis and DNA Damage in B Lymphocytes. *Int J Mol Sci*. 2017; 18:18. 10.3390/ijms18091876

Bettencourt C, Lima M. Machado-Joseph Disease: from first descriptions to new perspectives. *Orphanet J Rare Dis*. 2011; 6:35. 10.1186/1750-1172-6-35

Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pagès F, Trajanoski Z, Galon J. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 2009; 25:1091–93. 10.1093/bioinformatics/btp101

Bindea G, Galon J, Mlecnik B. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics*. 2013; 29:661–63. 10.1093/bioinformatics/btt019

Braga-Neto P, Felicio AC, Pedroso JL, Dutra LA, Bertolucci PH, Gabbai AA, Barsottini OG. Clinical correlates of olfactory dysfunction in spinocerebellar ataxia type 3. *Parkinsonism Relat Disord*. 2011; 17:353–56. 10.1016/j.parkreldis.2011.02.004

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015; 4:7. 10.1186/s13742-015-0047-8

Chatterji S, Pachter L. Reference based annotation with GeneMapper. *Genome Biol*. 2006; 7:R29–29. 10.1186/gb-2006-7-4-r29

Chen Z, Zheng C, Long Z, Cao L, Li X, Shang H, Yin X, Zhang B, Liu J, Ding D, Peng Y, Wang C, Peng H, *et al*., and Chinese Clinical Research Cooperative Group for Spinocerebellar Ataxias (CCRCG-SCA). (CAG)n loci as genetic modifiers of age-at-onset in patients with Machado-Joseph disease from mainland China. *Brain*. 2016; 139:e41–41. 10.1093/brain/aww087

de Mattos EP, Kolbe Musskopf M, Bielefeldt Leotti V, Saraiva-Pereira ML, Jardim LB. Genetic risk factors for modulation of age at onset in Machado-Joseph disease/spinocerebellar ataxia type 3: a systematic review and meta-analysis. *J Neurol Neurosurg Psychiatry*. 2019; 90:203–10. 10.1136/jnnp-2018-319200

Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2011; 9:179–81. 10.1038/nmeth.1785

Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics*. 2014; 30:1266–72. 10.1093/bioinformatics/btu014

França MC Jr, Emmel VE, D'Abreu A, Maurer-Morelli CV, Secolin R, Bonadia LC, da Silva MS, Nucci A, Jardim LB, Saraiva-Pereira ML, Marques W Jr, Paulson H, Lopes-Cendes I. Normal ATXN3 Allele but Not CHIP Polymorphisms Modulates Age at Onset in Machado-Joseph Disease. *Front Neurol*. 2012; 3:164. 10.3389/fneur.2012.00164

Gissen P, Maher ER. Cargos and genes: insights into vesicular transport from inherited human disease. *J Med Genet*. 2007; 44:545–55. 10.1136/jmg.2007.050294

Gunawardena S, Goldstein LS. Polyglutamine diseases and transport problems: deadly traffic jams on neuronal highways. *Arch Neurol*. 2005; 62:46–51. 10.1001/archneur.62.1.46

Hsieh J, Liu JW, Harn HJ, Hsueh KW, Rajamani K, Deng YC, Chia CM, Shyu WC, Lin SZ, Chiou TW. Human Olfactory Ensheathing Cell Transplantation Improves Motor Function in a Mouse Model of Type 3 Spinocerebellar Ataxia. *Cell Transplant*. 2017; 26:1611–21. 10.1177/0963689717732578

Khan LA, Bauer PO, Miyazaki H, Lindenberg KS, Landwehrmeyer BG, Nukina N. Expanded polyglutamines impair synaptic transmission and ubiquitin-proteasome system in Caenorhabditis elegans. *J Neurochem*. 2006; 98:576–87. 10.1111/j.1471-4159.2006.03895.x

Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput Biol*. 2016; 12:e1004714. 10.1371/journal.pcbi.1004714

Lee JM, Wheeler VC, Chao MJ, Vonsattel JP, Pinto RM, Lucente D, Abu-Elneel K, Ramos EM, Mysore JS, Gillis T, MacDonald ME, Gusella JF, Harold D, *et al*., and Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell*. 2015; 162:516–26. 10.1016/j.cell.2015.07.003

Lee JM, Chao MJ, Harold D, Abu Elneel K, Gillis T, Holmans P, Jones L, Orth M, Myers RH, Kwak S, Wheeler VC, MacDonald ME, Gusella JF. A modifier of Huntington's disease onset at the MLH1 locus. *Hum Mol Genet*. 2017; 26:3859–67. 10.1093/hmg/ddx286

Lee JM, Correia K, Loupe J, Kim KH, Barker D, Hong EP, Chao MJ, Long JD, Lucente D, Vonsattel JP, Pinto RM, Abu Elneel K, Ramos EM, *et al*., and Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. Electronic address: gusella@helix.mgh.harvard.edu, and Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell*. 2019; 178:887–900.e14. 10.1016/j.cell.2019.06.036

Lescale C, Deriano L. The RAG recombinase: beyond breaking. *Mech Ageing Dev*. 2017; 165:3–9. 10.1016/j.mad.2016.11.003

Maciel P, Gaspar C, DeStefano AL, Silveira I, Coutinho P, Radvany J, Dawson DM, Sudarsky L, Guimarães J, Loureiro JE, *et al*. Correlation between CAG repeat length and clinical features in Machado-Joseph disease. *Am J Hum Genet*. 1995; 57:54–61.

Masuda Y, Takahashi H, Sato S, Tomomori-Sato C, Saraf A, Washburn MP, Florens L, Conaway RC, Conaway JW, Hatakeyama S. TRIM29 regulates the assembly of DNA repair proteins into damaged chromatin. *Nat Commun*. 2015; 6:7299. 10.1038/ncomms8299

Martins S, Calafell F, Wong VC, Sequeiros J, Amorim A. A multistep mutation mechanism drives the evolution of the CAG repeat at MJD/SCA3 locus. *Eur J Hum Genet*. 2006; 14:932–40. 10.1038/sj.ejhg.5201643

Martins S, Pearson CE, Coutinho P, Provost S, Amorim A, Dubé MP, Sequeiros J, Rouleau GA. Modifiers of (CAG)(n) instability in Machado-Joseph disease (MJD/SCA3) transmissions: an association study with DNA replication, repair and recombination genes. *Hum Genet*. 2014; 133:1311–18. 10.1007/s00439-014-1467-8

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, Luo Y, Sidore C, Kwong A, *et al*., and Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016; 48:1279–83. 10.1038/ng.3643

Mishra A, Macgregor S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res Hum Genet*. 2015; 18:86–91. 10.1017/thg.2014.79

Pedroso JL, França MC Jr, Braga-Neto P, D'Abreu A, Saraiva-Pereira ML, Saute JA, Teive HA, Caramelli P, Jardim LB, Lopes-Cendes I, Barsottini OG. Nonmotor and extracerebellar features in Machado-Joseph disease: a review. *Mov Disord*. 2013; 28:1200–08. 10.1002/mds.25513

Pinto RM, Dragileva E, Kirby A, Lloret A, Lopez E, St Claire J, Panigrahi GB, Hou C, Holloway K, Gillis T, Guide JR, Cohen PE, Li GM, *et al*. Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. *PLoS Genet*. 2013; 9:e1003930. 10.1371/journal.pgen.1003930

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*. 2010; 26:2336–37. 10.1093/bioinformatics/btq419

Raposo M, Ramos A, Bettencourt C, Lima M. Replicating studies of genetic modifiers in spinocerebellar ataxia type 3: can homogeneous cohorts aid? *Brain*. 2015; 138:e398–398. 10.1093/brain/awv206

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019; 47:D886–94. 10.1093/nar/gky1016

Tezenas du Montcel S, Durr A, Bauer P, Figueroa KP, Ichikawa Y, Brussino A, Forlani S, Rakowicz M, Schöls L, Mariotti C, van de Warrenburg BP, Orsi L, Giunti P, *et al*., and Clinical Research Consortium for Spinocerebellar Ataxia (CRC-SCA), and EUROSCA network. Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes. *Brain*. 2014; 137:2444–55. 10.1093/brain/awu174

Twist EC, Casaubon LK, Ruttledge MH, Rao VS, Macleod PM, Radvany J, Zhao Z, Rosenberg RN, Farrer LA, Rouleau GA. Machado Joseph disease maps to the same region of chromosome 14 as the spinocerebellar ataxia type 3 locus. *J Med Genet*. 1995; 32:25–31. 10.1136/jmg.32.1.25

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010; 38:e164. 10.1093/nar/gkq603

Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun*. 2017; 8:1826. 10.1038/s41467-017-01261-5

Wiatr K, Piasecki P, Marczak Ł, Wojciechowski P, Kurkowiak M, Płoski R, Rydzanicz M, Handschuh L, Jungverdorben J, Brüstle O, Figlerowicz M, Figiel M. Altered Levels of Proteins and Phosphoproteins, in the Absence of Early Causative Transcriptional Changes, Shape the Molecular Pathogenesis in the Brain of Young Presymptomatic Ki91 SCA3/MJD Mouse. *Mol Neurobiol*. 2019; 56:8168–202. 10.1007/s12035-019-01643-4

Wikiniyadhanee R, Lerksuthirat T, Stitchantrakul W, Chitphuk S, Dejsuphong D. AB064. TRIM29: a novel gene involved in DNA repair mechanisms. *Ann Transl Med*. 2017; 5:AB064–064. 10.21037/atm.2017.s064

Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011; 88:76–82. 10.1016/j.ajhg.2010.11.011

Zhang K, Cui S, Chang S, Zhang L, Wang J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Res*. 2010; 38:W90–5. 10.1093/nar/gkq324

Zijlstra MP, Rujano MA, Van Waarde MA, Vis E, Brunt ER, Kampinga HH. Levels of DNAJB family members (HSP40) correlate with disease onset in patients with spinocerebellar ataxia type 3. *Eur J Neurosci*. 2010; 32:760–70. 10.1111/j.1460-9568.2010.07352.x

CHAPTER 5: GENERAL DISCUSSION

The role of tandem repeats in human diseases was established 30 years ago with the discovery of trinucleotide expansions associated with fragile X syndrome and spinal bulbar muscular atrophy (Kremer *et al*., 1991; Verkerk *et al*., 1991; La Spada *et al*., 1991). In 1993, trinucleotide CAG repeat expansions were identified in the *HTT* and *ATXN1* genes causing HD and SCA1, respectively (Orr *et al*., 1993; MacDonald *et al*., 1993). Since then, more than 60 repeat expansions have been found to cause neurological diseases, most of which are hereditary ataxia subtypes (Paulson, 2018; Depienne and Mandel, 2021).

During the first decade following the identification of pathogenic repeat expansions, linkage analysis was the most common approach to identify the genomics region containing the associated expansions. With the advent of genome-wide association studies (GWAS) as well as next-generation sequencing (NGS), rapid identification and direct observation of the disease-associated expansions became possible. These developments allowed a new wave of repeat expansion discovery which started with the identification of *NOP56* and *C9orf72* repeat expansions in SCA36 and ALS and frontotemporal dementia (FTD) (Kobayashi *et al*., 2011; DeJesus-Hernandez *et al.*, 2011; Renton *et al.*, 2011). Through a genome-wide linkage analysis, a region harboring 37 genes had been identified as a disease locus for a cohort of Japanese SCA patients. Next generation sequencing of candidate genes allowed the discovery of pathogenic hexanucleotide repeat expansion in *NOP56* causing SCA36 (Kobayashi *et al*., 2011). Similarly, the *C9orf72* locus was first identified through a linkage analysis, followed by a GWAS in the Finnish population (Vance *et al.*, 2006; Laaksovirta *et al*., 2010). A deep sequencing of the chromosome 9p21 region and repeat-primed PCR approach allowed the discovery of

hexanucleotide repeat expansion in *C9orf72* by two independent groups, that explained a remarkable proportion of familial and sporadic ALS cases (DeJesus-Hernandez *et al*., 2011; Renton *et al*., 2011). These discoveries marked a milestone in SCA and ALS research and set the precedent for the identification of additional repeats in neurological diseases via NGS. For example, *C9orf72* repeat expansion explains 40% of familial and 7% of sporadic ALS cases (Renton *et al*; 2014). The next breakthrough in ataxia repeat expansion research was almost a decade later with the identification of pentanucleotide repeat expansions in *RFC1* which are the cause of 20% of sporadic late-onset ataxia cases (Cortese *et al.*, 2019).

High coverage WGS, particularly long-read sequencing, are highly efficient approaches to estimate the length of genomic repeats. Several software have been designed to assess the repeat length from exome or genome data such as ExpansionHunter (Dolzhenko *et al*., 2017), STRetch (Dashnow *et al.*, 2018), GangSTR (Mousavi *et al*., 2019), and exSTRa (Tankard *et al*., 2018). These *in silico* approaches are catalogue-based; therefore, they are only able to detect repeat expansions at a specified locus for a given motif. In Chapter 2, we applied ExpansionHunter to high-coverage WGS data of a total of 2,504 samples from 1KGP which were re-sequenced by The International Genome Sample Resource in 2019. We estimated the distribution of CAG repeat expansions in *ATXN1*, *ATXN2*, *ATXN3*, and *HTT*. Furthermore, we reported expanded repeats in the disease-associated range in at least eleven samples in the 1KGP dataset. Our work showed that even though the participants declared themselves healthy at the time of the collection, datasets generated from general population might contain presymptomatic carriers for late-onset diseases that would develop the associated diseases later in life. Recently, ExpansionHunter has been applied to 100,000 Genomes Project participants with neurological disorders. With its high

sensitivity and specificity across all 13 disease-associated loci, ExpansionHunter was shown to be practical to diagnose neurological repeat expansion disorders (Ibanez *et al*., 2020). As a complementary approach to Chapter 2 and Ibanez *et al*., ExpansionHunter Denovo (EHdn) has been developed to identify expanded repeats without a prior knowledge of genomic regions or sequence motifs. Since EHdn estimates approximate locations and length of repeats, it may require further validation, especially for clinical purposes. Nevertheless, it detects pathogenic repeat expansions that are not discoverable via existing methods therefore enabling further novel repeat expansions in neurological diseases (Dolzhenko *et al*., 2020).

In Chapter 2, we applied an *in silico* approach to assess the presence of expanded repeats. In Chapter 3, we employed conventional PCR techniques, RPPCR, and long-range PCR amplification followed by Sanger sequencing. To the best of our knowledge, our analysis was the first follow-up study of *RFC1* repeat expansions in additional cohorts. In the original study, the pathogenic expansion explained 22% of the late-onset sporadic ataxia cases (Cortese *et al.*, 2019). In addition, the prevalence was found to be 14% in a Turkish late-onset recessive or sporadic cerebellar ataxia cohort in which common ataxias were already excluded (Traschütz *et al.*, 2021). To estimate the prevalence of *RFC1*-based CANVAS in a general late-onset ataxia cohort, we used a Canadian cohort in which no prior genetic test or selection were done. Our study showed that the prevalence was much lower in the Canadian cohort (Akçimen *et al.,* 2019). Moreover, in a large North American cohort, biallelic expansions were observed in 3.2 % of the undiagnosed ataxia patients (Syriani *et al*., 2020). In conclusion, the prevalence of *RFC1*-based CANVAS may be overestimated when it is examined in cohorts in which a prior selection of cases was applied based on clinical data, mode of inheritance and genetic tests for other ataxia subtypes.

In addition to the three sequence motifs that were identified in the original study (AAAAG, AAAGG, and AAGGG), we reported two new conformations (AAGAG and AGAGG). Furthermore, expansion of ACAGG motif was identified in two Asia-Pacific and one Japanese families (Scriba *et al*., 2020). Those patients showed additional clinical features which were not reported in genetically-defined CANVAS (Scriba *et al*., 2020; Tsuchiya *et al*., 2020). Since these new motifs have been identified in only a few samples so far, it is not sufficient to make a merit comparison between the severity of their associated phenotypes and originally identified pathogenic repeat motif. However, these new observations demonstrate the instability and potential importance of both length and repeat content.

The inverse correlation between expanded repeat length and AO was reported right after the discovery of first repeat associated neurological diseases (Andrew *et al*., 1993; Orr *et al.*, 1993). However, repeat length itself did not completely explain the AO variability itself for any of the repeat expansion diseases, suggesting the implication of additional factors. If a genetic variation was associated with a later onset or a less severe form of the disease, it would provide insight into its mechanism, and therefore, potential targets. This rationale led to the genetic studies that attempted to identify genetic modifiers of repeat-associated diseases. The GeM-HD Consortium identified variants in DNA repair proteins that may alter the somatic expansion of *HTT*-CAG repeats, which was replicated in SCA types. *MLH1* and *FAN1* loci were identified as genetic modifiers in these studies. *FAN1* variants were also replicated in additional HD and SCA cohorts (Genetic Modifiers of Huntington's Disease C., 2015; Bettencourt *et al*., 2016). In Chapter 4, we identified *TRIM29* locus as a potential genetic factor that modify AO in SCA3. TRIM29 is

involved in the regulation of DNA repair proteins such as MLH1, which was identified as an AO modifier in Huntington's disease modifier study. Therefore, we suggested that common modifier mechanisms may be implicated in CAG repeat expansion diseases.

Furthermore, a cis-acting element, CTCF binding sites upstream and/or of CAG repeats were found to be implicated in the regulation of CAG repeat instability in HD, SCA2, SCA7 and DRPLA. In SCA7, it was shown that impairment in CTCF binding promotes somatic instability of CAG repeat in *ATXN7*. Identification of additional genetics and epigenetic mechanisms, trans-factors such as CTCF protein, and cis-elements such as CTCF binding sites in the flaking regions of the repeats could be potential targets for therapy (Libby *et al*., 2008).

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS

This thesis consisted of the application of different genetic approaches to repeat-associated hereditary ataxia which included an *in silico* analysis for CAG repeat-length estimation, a conventional repeat screening method for *RFC1* repeat expansions, as well as a GWAS to identify risk variants that may modify AO in SCA3.

In Chapter 2, we showed the presence of participants that have expanded repeats in a public dataset using a targeted repeat detection software. A future approach could be to apply a similar approach in larger public datasets which include PCR-free WGS samples, such as TOPMed dataset (Taliun *et al*., 2021).

In Chapter 3, we concluded that additional studies are required to determine the pathogenicity of the novel motifs that we identified. To perform further analyses on repeat expansions, we joined 100,000 Genomes Project research community – the Genomics England Clinical Interpretation Partnership (GeCIP). As a pilot study, we have screened a total of 1,027 hereditary ataxia cases to search for new *RFC1* expansions. We identified two additional disease-associated motifs (AAGGC and AGGGC) in biallelic state in two families. This work will be continued by the validation of identified expansions using additional sequencing approaches. Identification of various distinct sequence conformations may expand the genetic and clinical spectrum of *RFC1*-based CANVAS, as well as provide insights about the pathological mechanism(s) leading to disease.

With the advent of genome-wide approaches, GWAS became one of the most applied methods to discover disease-associated genetic variants. It was applied to several neurological diseases in the last decade, including ALS, restless legs syndrome, Parkinson's disease, and

Alzheimer's disease (Laaksovirta *et al*., 2010; Winkelmann *et al*., 2007; Fung *et al*., 2006; Coon *et al*., 2007). With the availability of large-scale datasets, multiple GWAS have been performed for these diseases, that not only replicated the previous findings but also identified novel genetic variants (van Rheenen *et al.*, 2016; Wightman *et al*., 2021). Until recently, to the best of our knowledge, no GWAS has been performed using hereditary ataxia human participants. The only hereditary ataxia GWAS was performed in dog species (Gast *et al*., 2016). Currently, there are 459 cases in UK Biobank (Canela-Xandri *et al*., 2018) and 154 cases in FinnGen GWAS summary statistics (https://www.finngen.fi/en/access_results, 2020). However, even their meta-analysis may not provide enough power to identify associated variants.

Considering the fact that GWAS was one of the approaches that led to the identification of the *C9orf72* locus in ALS, the same technique could have been used to identify the *RFC1* locus in hereditary ataxia a decade ago. Therefore, one of the future directions could be to employ a GWAS approach in hereditary ataxia. In a collaborative study using 100,000 Genomes Project dataset as part of the GeCIP, we initiated a GWAS using WGS samples of 1,027 hereditary ataxia cases and non-neurological control population. Another future aim is to detect novel repeat expansions using recently developed software such as ExpansionHunter Denovo using 1,027 cases in the 100,000 Genomes Project.

With the development of new approaches and software as well as availability of WGS datasets from a wide variety of human populations, we will hopefully expand our understanding of the genetics of rare neurological diseases such as hereditary ataxias.

Master Reference List

Aboud Syriani, D., Wong, D., Andani, S., De Gusmao, C.M., Mao, Y., Sanyoura, M., Glotzer, G., Lockhart, P.J., Hassin-Baer, S., Khurana, V., et al. (2020). Prevalence of *RFC1*-mediated spinocerebellar ataxia in a North American ataxia cohort. Neurol Genet 6, e440.

Akcimen, F., Martins, S., Liao, C., Bourassa, C.V., Catoire, H., Nicholson, G.A., Riess, O., Raposo, M., Franca, M.C., Vasconcelos, J., et al. (2020). Genome-wide association study identifies genetic factors that modify age at onset in Machado-Joseph disease. Aging 12, 4742-4756.

Akcimen, F., Ross, J.P., Bourassa, C.V., Liao, C., Rochefort, D., Gama, M.T.D., Dicaire, M.J., Barsottini, O.G., Brais, B., Pedroso, J.L., et al. (2019). Investigation of the *RFC1* Repeat Expansion in a Canadian and a Brazilian Ataxia Cohort: Identification of Novel Conformations. Front Genet 10, 1219.

Amendola, L.M., Dorschner, M.O., Robertson, P.D., Salama, J.S., Hart, R., Shirts, B.H., Murray, M.L., Tokita, M.J., Gallego, C.J., Kim, D.S., et al. (2015). Actionable exomic incidental findings in 6503 participants: challenges of variant classification. Genome Res 25, 305-315.

Andrew, S.E., Goldberg, Y.P., Kremer, B., Telenius, H., Theilmann, J., Adam, S., Starr, E., Squitieri, F., Lin, B., Kalchman, M.A., et al. (1993). The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nat Genet 4, 398-403.

Ansorge, O., Giunti, P., Michalik, A., Van Broeckhoven, C., Harding, B., Wood, N., and Scaravilli, F. (2004). Ataxin-7 aggregation and ubiquitination in infantile SCA7 with 180 CAG repeats. Annals of neurology 56, 448-452.

Bettencourt, C., Hensman-Moss, D., Flower, M., Wiethoff, S., Brice, A., Goizet, C., Stevanin, G., Koutsis, G., Karadima, G., Panas, M., et al. (2016). DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. Annals of neurology 79, 983-990.

Bettencourt, C., and Lima, M. (2011). Machado-Joseph Disease: from first descriptions to new perspectives. Orphanet J Rare Dis 6, 35.

Bettencourt, C., Santos, C., Kay, T., Vasconcelos, J., and Lima, M. (2008). Analysis of segregation patterns in Machado–Joseph disease pedigrees. Journal of Human Genetics 53, 920-923.

Bidichandani, S.I., and Delatycki, M.B. (1993). Friedreich Ataxia. In GeneReviews((R)), M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, G. Mirzaa, and A. Amemiya, eds. (Seattle (WA)).

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics (Oxford, England) 25, 1091-1093.

Bird, T.D. (1998). Hereditary Ataxia Overview. In GeneReviews((R)), M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, G. Mirzaa, and A. Amemiya, eds. (Seattle (WA)).

Bouchard, J.P., Barbeau, A., Bouchard, R., and Bouchard, R.W. (1978). Autosomal recessive spastic ataxia of Charlevoix-Saguenay. Can J Neurol Sci 5, 61-69.

Bourassa, C.V., Meijer, I.A., Merner, N.D., Grewal, K.K., Stefanelli, M.G., Hodgkinson, K., Ives, E.J., Pryse-Phillips, W., Jog, M., Boycott, K., et al. (2012). *VAMP1* mutation causes dominant hereditary spastic ataxia in Newfoundland families. Am J Hum Genet 91, 548-552.

Braga-Neto, P., Felicio, A.C., Pedroso, J.L., Dutra, L.A., Bertolucci, P.H., Gabbai, A.A., and Barsottini, O.G. (2011). Clinical correlates of olfactory dysfunction in spinocerebellar ataxia type 3. Parkinsonism Relat Disord 17, 353-356.

Brkanac, Z., Fernandez, M., Matsushita, M., Lipe, H., Wolff, J., Bird, T.D., and Raskind, W.H. (2002). Autosomal dominant sensory/motor neuropathy with Ataxia (SMNA): Linkage to chromosome 7q22-q32. Am J Med Genet 114, 450-457.

Campuzano, V., Montermini, L., Molto, M.D., Pianese, L., Cossee, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A., et al. (1996). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science 271, 1423-1427.

Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2018). An atlas of genetic associations in UK Biobank. Nat Genet 50, 1593-1599.

Casey, H.L., and Gomez, C.M. (1993). Spinocerebellar Ataxia Type 6. In GeneReviews((R)), M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, G. Mirzaa, and A. Amemiya, eds. (Seattle (WA)).

Chamberlain, S., Shaw, J., Rowland, A., Wallis, J., South, S., Nakamura, Y., von Gabain, A., Farrall, M., and Williamson, R. (1988). Mapping of mutation causing Friedreich's ataxia to human chromosome 9. Nature 334, 248-250.

Chen, Z., Zheng, C., Long, Z., Cao, L., Li, X., Shang, H., Yin, X., Zhang, B., Liu, J., Ding, D., et al. (2016). (CAG)n loci as genetic modifiers of age-at-onset in patients with Machado-Joseph disease from mainland China. Brain 139, e41-e41.

Choi, K.D., and Choi, J.H. (2016). Episodic Ataxias: Clinical and Genetic Features. J Mov Disord 9, 129-135.

Cleary, J.D., Subramony, S.H., and Ranum, L.P.W. (1993). Spinocerebellar Ataxia Type 8. In GeneReviews((R)), M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, G. Mirzaa, and A. Amemiya, eds. (Seattle (WA)).

Coon, K.D., Myers, A.J., Craig, D.W., Webster, J.A., Pearson, J.V., Lince, D.H., Zismann, V.L., Beach, T.G., Leung, D., Bryden, L., et al. (2007). A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. J Clin Psychiatry 68, 613-618.

Cortese, A., Simone, R., Sullivan, R., Vandrovcova, J., Tariq, H., Yau, W.Y., Humphrey, J., Jaunmuktane, Z., Sivakumar, P., Polke, J., et al. (2019). Biallelic expansion of an intronic repeat in *RFC1* is a common cause of late-onset ataxia. Nat Genet 51, 649-658.

Cortese, A., Tozza, S., Yau, W.Y., Rossi, S., Beecroft, S.J., Jaunmuktane, Z., Dyer, Z., Ravenscroft, G., Lamont, P.J., Mossman, S., et al. (2020). Cerebellar ataxia, neuropathy, vestibular areflexia syndrome due to *RFC1* repeat expansion. Brain 143, 480-490.

Dashnow, H., Lek, M., Phipson, B., Halman, A., Sadedin, S., Lonsdale, A., Davis, M., Lamont, P., Clayton, J.S., Laing, N.G., et al. (2018). STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biol 19, 121.

de Bot, S.T., Willemsen, M.A., Vermeer, S., Kremer, H.P., and van de Warrenburg, B.P. (2012). Reviewing the genetic causes of spastic-ataxias. Neurology 79, 1507-1514.

De Braekeleer, M., Giasson, F., Mathieu, J., Roy, M., Bouchard, J.P., and Morgan, K. (1993). Genetic epidemiology of autosomal recessive spastic ataxia of Charlevoix-Saguenay in northeastern Quebec. Genet Epidemiol 10, 17-25.

de Mattos, E.P., Kolbe Musskopf, M., Bielefeldt Leotti, V., Saraiva-Pereira, M.L., and Jardim, L.B. (2019). Genetic risk factors for modulation of age at onset in Machado-Joseph disease/spinocerebellar ataxia type 3: a systematic review and meta-analysis. Journal of Neurology, Neurosurgery & Psychiatry 90, 203-210.

DeJesus-Hernandez, M., Mackenzie, I.R., Boeve, B.F., Boxer, A.L., Baker, M., Rutherford, N.J., Nicholson, A.M., Finch, N.A., Flynn, H., Adamson, J., et al. (2011). Expanded GGGGCC hexanucleotide repeat in noncoding region of *C9ORF72* causes chromosome 9p-linked FTD and ALS. Neuron 72, 245-256.

Depienne, C., and Mandel, J.L. (2021). 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? Am J Hum Genet 108, 764-785.

Dolzhenko, E., Bennett, M.F., Richmond, P.A., Trost, B., Chen, S., van Vugt, J., Nguyen, C., Narzisi, G., Gainullin, V.G., Gross, A.M., et al. (2020). ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. Genome Biol 21, 102.

Dolzhenko, E., van Vugt, J., Shaw, R.J., Bekritsky, M.A., van Blitterswijk, M., Narzisi, G., Ajay, S.S., Rajan, V., Lajoie, B.R., Johnson, N.H., et al. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res 27, 1895-1903.

Dupre, N., Bouchard, J.P., Brais, B., and Rouleau, G.A. (2006). Hereditary ataxia, spastic paraparesis and neuropathy in the French-Canadian population. Can J Neurol Sci 33, 149-157.

Durr, A., Cossee, M., Agid, Y., Campuzano, V., Mignard, C., Penet, C., Mandel, J.L., Brice, A., and Koenig, M. (1996). Clinical and genetic abnormalities in patients with Friedreich's ataxia. The New England journal of medicine 335, 1169-1175.

Engert, J.C., Berube, P., Mercier, J., Dore, C., Lepage, P., Ge, B., Bouchard, J.P., Mathieu, J., Melancon, S.B., Schalling, M., et al. (2000). ARSACS, a spastic ataxia common in northeastern Quebec, is caused by mutations in a new gene encoding an 11.5-kb ORF. Nat Genet 24, 120-125.

Fan, Y., Zhang, S., Yang, J., Mao, C.Y., Yang, Z.H., Hu, Z.W., Wang, Y.L., Liu, Y.T., Liu, H., Yuan, Y.P., et al. (2020). No biallelic intronic AAGGG repeat expansion in *RFC1* was found in patients with late-onset ataxia and MSA. Parkinsonism Relat Disord 73, 1-2.

Figueroa, K.P., Coon, H., Santos, N., Velazquez, L., Mederos, L.A., and Pulst, S.M. (2017). Genetic analysis of age at onset variation in spinocerebellar ataxia type 2. Neurol Genet 3, e155.

Franca, M.C., Jr., Emmel, V.E., D'Abreu, A., Maurer-Morelli, C.V., Secolin, R., Bonadia, L.C., da Silva, M.S., Nucci, A., Jardim, L.B., Saraiva-Pereira, M.L., et al. (2012). Normal *ATXN3* Allele but Not CHIP Polymorphisms Modulates Age at Onset in Machado-Joseph Disease. Front Neurol 3, 164.

Friedman, J.E. (2011). Anticipation in hereditary disease: the history of a biomedical concept. Human genetics 130, 705-714.

Friedreich, N. (1863). Ueber degenerative Atrophie der spinalen Hinterstränge. Archiv für pathologische Anatomie und Physiologie und für klinische Medicin 26, 391-419.

Fung, H.C., Scholz, S., Matarin, M., Simon-Sanchez, J., Hernandez, D., Britton, A., Gibbs, J.R., Langefeld, C., Stiegert, M.L., Schymick, J., et al. (2006). Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. The Lancet Neurology 5, 911-916.

Gast, A.C., Metzger, J., Tipold, A., and Distl, O. (2016). Genome-wide association study for hereditary ataxia in the Parson Russell Terrier and DNA-testing for ataxia-associated mutations in the Parson and Jack Russell Terrier. BMC Vet Res 12, 225.

Genetic Modifiers of Huntington's Disease, C. (2015). Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. Cell 162, 516-526.

Gissen, P., and Maher, E.R. (2007). Cargos and genes: insights into vesicular transport from inherited human disease. J Med Genet 44, 545-555.

Giunti, P., Mantuano, E., and Frontali, M. (2020). Episodic Ataxias: Faux or Real? International journal of molecular sciences 21.

Grewal, R.P., Achari, M., Matsuura, T., Durazo, A., Tayag, E., Zu, L., Pulst, S.M., and Ashizawa, T. (2002). Clinical features and ATTCT repeat expansion in spinocerebellar ataxia type 10. Arch Neurol 59, 1285-1290.

Gunawardena, S., and Goldstein, L.S.B. (2005). Polyglutamine Diseases and Transport Problems: Deadly Traffic Jams on Neuronal Highways. JAMA Neurology 62, 46-51.

Hersheson, J., Haworth, A., and Houlden, H. (2012). The inherited ataxias: genetic heterogeneity, mutation databases, and future directions in research and clinical diagnostics. Human mutation 33, 1324-1332.

Hsieh, J., Liu, J.W., Harn, H.J., Hsueh, K.W., Rajamani, K., Deng, Y.C., Chia, C.M., Shyu, W.C., Lin, S.Z., and Chiou, T.W. (2017). Human Olfactory Ensheathing Cell Transplantation Improves

Motor Function in a Mouse Model of Type 3 Spinocerebellar Ataxia. Cell Transplant 26, 1611-1621.

Ibanez, K., Polke, J., Hagelstrom, T., Dolzhenko, E., Pasko, D., Thomas, E., Daugherty, L., Kasperaviciute, D., McDonagh, E.M., Smith, K.R., et al. (2020). Whole genome sequencing for diagnosis of neurological repeat expansion disorders. bioRxiv, 2020.2011.2006.371716.

Jayadev, S., and Bird, T.D. (2013). Hereditary ataxias: overview. Genetics in Medicine 15, 673-683.

Jazurek-Ciesiolka, M., Ciesiolka, A., Komur, A.A., Urbanek-Trzeciak, M.O., Krzyzosiak, W.J; Fiszer, A. (2020). RAN Translation of the Expanded CAG Repeats in the SCA3 Disease Context. Journal of Molecular Biology, 432, 666993.

Kay, C., Fisher, E., and Hayden, M.R. (2014). Huntington's Disease. In Epidemiology (Oxford University Press).

Khan, L.A., Bauer, P.O., Miyazaki, H., Lindenberg, K.S., Landwehrmeyer, B.G., and Nukina, N. (2006). Expanded polyglutamines impair synaptic transmission and ubiquitin-proteasome system in Caenorhabditis elegans. J Neurochem 98, 576-587.

Kivisild, T. (2013). Founder Effect. In Brenner's Encyclopedia of Genetics (Second Edition), S. Maloy, and K. Hughes, eds. (San Diego: Academic Press), pp. 100-101.

Klockgether, T., Mariotti, C., and Paulson, H.L. (2019). Spinocerebellar ataxia. Nat Rev Dis Primers 5, 24.

Kobayashi, H., Abe, K., Matsuura, T., Ikeda, Y., Hitomi, T., Akechi, Y., Habu, T., Liu, W., Okuda, H., and Koizumi, A. (2011). Expansion of intronic GGCCTG hexanucleotide repeat in *NOP56* causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. Am J Hum Genet 89, 121-130.

Kremer, E.J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S.T., Schlessinger, D., Sutherland, G.R., and Richards, R.I. (1991). Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. Science 252, 1711-1714.

Krysa, W., Sulek, A., Rakowicz, M., Szirkowiec, W., and Zaremba, J. (2016). High relative frequency of SCA1 in Poland reflecting a potential founder effect. Neurol Sci 37, 1319-1325.

La Spada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E., and Fischbeck, K.H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature 352, 77-79.

Laaksovirta, H., Peuralinna, T., Schymick, J.C., Scholz, S.W., Lai, S.L., Myllykangas, L., Sulkava, R., Jansson, L., Hernandez, D.G., Gibbs, J.R., et al. (2010). Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. The Lancet Neurology 9, 978-985.

Lee, J.-M., Correia, K., Loupe, J., Kim, K.-H., Barker, D., Hong, E.P., Chao, M.J., Long, J.D., Lucente, D., Vonsattel, J.P.G., et al. (2019). CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. Cell 178, 887-900.e814.

Lescale, C., and Deriano, L. (2017). The RAG recombinase: Beyond breaking. Mechanisms of Ageing and Development 165, 3-9.

Libby, R.T., Hagerman, K.A., Pineda, V.V., Lau, R., Cho, D.H., Baccam, S.L., Axford, M.M., Cleary, J.D., Moore, J.M., Sopher, B.L., et al. (2008). CTCF cis-regulates trinucleotide repeat instability in an epigenetic manner: a novel basis for mutational hot spot determination. PLoS Genet 4, e1000257.

MacDonald, M.E., Ambrose, C.M., Duyao, M.P., Myers, R.H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S.A., James, M., Groot, N., et al. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72, 971-983.

Maciel, P., Gaspar, C., DeStefano, A.L., Silveira, I., Coutinho, P., Radvany, J., Dawson, D.M., Sudarsky, L., Guimaraes, J., Loureiro, J.E., et al. (1995). Correlation between CAG repeat length and clinical features in Machado-Joseph disease. Am J Hum Genet 57, 54-61.

Marelli, C., Cazeneuve, C., Brice, A., Stevanin, G., and Durr, A. (2011). Autosomal dominant cerebellar ataxias. Rev Neurol (Paris) 167, 385-400.

Martins, S., Pearson, C.E., Coutinho, P., Provost, S., Amorim, A., Dube, M.P., Sequeiros, J., and Rouleau, G.A. (2014). Modifiers of (CAG)(n) instability in Machado-Joseph disease (MJD/SCA3) transmissions: an association study with DNA replication, repair and recombination genes. Human genetics 133, 1311-1318.

Matilla, T., Volpini, V., Genis, D., Rosell, J., Corral, J., Davalos, A., Molins, A., and Estivill, X. (1993). Presymptomatic analysis of spinocerebellar ataxia type 1 (SCA1) via the expansion of the SCA1 CAG-repeat in a large pedigree displaying anticipation and parental male bias. Hum Mol Genet 2, 2123-2128.

Matsuura, T. (2020). Genetic analysis of the first SCA36 family showing clinical anticipation. J Neurol Sci 418, 117151.

Matsuura, T., Fang, P., Lin, X., Khajavi, M., Tsuji, K., Rasmussen, A., Grewal, R.P., Achari, M., Alonso, M.E., Pulst, S.M., et al. (2004). Somatic and germline instability of the ATTCT repeat in spinocerebellar ataxia type 10. Am J Hum Genet 74, 1216-1224.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nature genetics 48, 1279-1283.

Moseley, M.L., Schut, L.J., Bird, T.D., Koob, M.D., Day, J.W., and Ranum, L.P. (2000). SCA8 CTG repeat: en masse contractions in sperm and intergenerational sequence changes may play a role in reduced penetrance. Hum Mol Genet 9, 2125-2130.

Mousavi, N., Shleizer-Burko, S., Yanicky, R., and Gymrek, M. (2019). Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res 47, e90.

Muller, U. (2021). Spinocerebellar ataxias (SCAs) caused by common mutations. Neurogenetics.

O'Hearn, E., Holmes, S.E., and Margolis, R.L. (2012). Spinocerebellar ataxia type 12. Handb Clin Neurol 103, 535-547.

Orozco Diaz, G., Nodarse Fleites, A., Cordoves Sagaz, R., and Auburger, G. (1990). Autosomal dominant cerebellar ataxia: clinical analysis of 263 patients from a homogeneous population in Holguin, Cuba. Neurology 40, 1369-1375.

Orr, H.T., Chung, M.Y., Banfi, S., Kwiatkowski, T.J., Jr., Servadio, A., Beaudet, A.L., McCall, A.E., Duvick, L.A., Ranum, L.P., and Zoghbi, H.Y. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nat Genet 4, 221-226.

Paulson, H. (2018). Repeat expansion diseases. Handb Clin Neurol 147, 105-123.

Pedroso, J.L., Franca, M.C., Jr., Braga-Neto, P., D'Abreu, A., Saraiva-Pereira, M.L., Saute, J.A., Teive, H.A., Caramelli, P., Jardim, L.B., Lopes-Cendes, I., et al. (2013). Nonmotor and extracerebellar features in Machado-Joseph disease: a review. Mov Disord 28, 1200-1208.

Pinto, R.M., Dragileva, E., Kirby, A., Lloret, A., Lopez, E., St Claire, J., Panigrahi, G.B., Hou, C., Holloway, K., Gillis, T., et al. (2013). Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. PLoS Genet 9, e1003930.

Rafehi, H., Szmulewicz, D.J., Bennett, M.F., Sobreira, N.L.M., Pope, K., Smith, K.R., Gillies, G., Diakumis, P., Dolzhenko, E., Eberle, M.A., et al. (2019). Bioinformatics-Based Identification of Expanded Repeats: A Non-reference Intronic Pentamer Expansion in *RFC1* Causes CANVAS. Am J Hum Genet 105, 151-165.

Raposo, M., Ramos, A., Bettencourt, C., and Lima, M. (2015). Replicating studies of genetic modifiers in spinocerebellar ataxia type 3: can homogeneous cohorts aid? Brain 138, e398-e398.

Rasmussen, A., De Biase, I., Fragoso-Benitez, M., Macias-Flores, M.A., Yescas, P., Ochoa, A., Ashizawa, T., Alonso, M.E., and Bidichandani, S.I. (2007). Anticipation and intergenerational repeat instability in spinocerebellar ataxia type 17. Annals of neurology 61, 607-610.

Renton, A.E., Majounie, E., Waite, A., Simon-Sanchez, J., Rollinson, S., Gibbs, J.R., Schymick, J.C., Laaksovirta, H., van Swieten, J.C., Myllykangas, L., et al. (2011). A hexanucleotide repeat expansion in *C9ORF72* is the cause of chromosome 9p21-linked ALS-FTD. Neuron 72, 257-268.

Renton, A.E., Chio, A., and Traynor, B.J. (2014). State of play in amyotrophic lateral sclerosis genetics. Nat Neurosci 17 (1): 17–23

Ruano, L., Melo, C., Silva, M.C., and Coutinho, P. (2014). The global epidemiology of hereditary ataxia and spastic paraplegia: a systematic review of prevalence studies. Neuroepidemiology 42, 174-183.

Sato, N., Amino, T., Kobayashi, K., Asakawa, S., Ishiguro, T., Tsunemi, T., Takahashi, M., Matsuura, T., Flanigan, K.M., Iwasaki, S., et al. (2009). Spinocerebellar ataxia type 31 is associated with "inserted" penta-nucleotide repeats containing (TGGAA)n. Am J Hum Genet 85, 544-557.

Schmitz-Hübsch, T., du Montcel, S.T., Baliko, L., Berciano, J., Boesch, S., Depondt, C., Giunti, P., Globas, C., Infante, J., Kang, J.-S., et al. (2006). Scale for the assessment and rating of ataxia. Neurology 66, 1717-1720.

Schols, L., Bauer, P., Schmidt, T., Schulte, T., and Riess, O. (2004). Autosomal dominant cerebellar ataxias: clinical features, genetics, and pathogenesis. The Lancet Neurology 3, 291-304.

Schut, J.W. (1950). HEREDITARY ATAXIA: Clinical Study Through Six Generations. Archives of Neurology & Psychiatry 63, 535-568.

Scriba, C.K., Beecroft, S.J., Clayton, J.S., Cortese, A., Sullivan, R., Yau, W.Y., Dominik, N., Rodrigues, M., Walker, E., Dyer, Z., et al. (2020). A novel *RFC1* repeat motif (ACAGG) in two Asia-Pacific CANVAS families. Brain 143, 2904-2910.

Seixas, A.I., Loureiro, J.R., Costa, C., Ordonez-Ugalde, A., Marcelino, H., Oliveira, C.L., Loureiro, J.L., Dhingra, A., Brandao, E., Cruz, V.T., et al. (2017). A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of *DAB1*, Mapping to SCA37, Causes Spinocerebellar Ataxia. Am J Hum Genet 101, 87-103.

Serrano-Munuera, C., Corral-Juan, M., Stevanin, G., San Nicolas, H., Roig, C., Corral, J., Campos, B., de Jorge, L., Morcillo-Suarez, C., Navarro, A., et al. (2013). New subtype of spinocerebellar ataxia with altered vertical eye movements mapping to chromosome 1p32. JAMA Neurol 70, 764-771.

Stochmanski, S.J., Therrien, M., Laganière, J., Rochefort, D., Laurent, S., Karemera, L., Gaudet, R.,Vyboh, K., Van Meyel, D.J., Di Cristo, G., Dion, P.A., Gaspar, C., Rouleau, G.A. (2012). Expanded *ATXN3* frameshifting events are toxic in Drosophila and mammalian neuron models, Human Molecular Genetics 21, 2211–2218.

Storey, E., Bahlo, M., Fahey, M., Sisson, O., Lueck, C.J., and Gardner, R.J. (2009). A new dominantly inherited pure cerebellar ataxia, SCA 30. Journal of neurology, neurosurgery, and psychiatry 80, 408-411.

Subramony, S.H. (2012). Overview of autosomal dominant ataxias. Handb Clin Neurol 103, 389-398.

Swinnen, B., Robberecht, W., Van Den Bosch, L. (2020). RNA toxicity in non-coding repeat expansion disorders. EMBO J. 2;39(1):e101112.

Szmulewicz, D.J., Waterston, J.A., MacDougall, H.G., Mossman, S., Chancellor, A.M., McLean, C.A., Merchant, S., Patrikios, P., Halmagyi, G.M., and Storey, E. (2011). Cerebellar ataxia, neuropathy, vestibular areflexia syndrome (CANVAS): a review of the clinical features and video-oculographic diagnosis. Ann N Y Acad Sci 1233, 139-147.

Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290-299.

Tankard, R.M., Bennett, M.F., Degorski, P., Delatycki, M.B., Lockhart, P.J., and Bahlo, M. (2018). Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. Am J Hum Genet 103, 858-873.

Teive, H.A., Roa, B.B., Raskin, S., Fang, P., Arruda, W.O., Neto, Y.C., Gao, R., Werneck, L.C., and Ashizawa, T. (2004). Clinical phenotype of Brazilian families with spinocerebellar ataxia 10. Neurology 63, 1509-1512.

Tezenas du Montcel, S., Durr, A., Bauer, P., Figueroa, K.P., Ichikawa, Y., Brussino, A., Forlani, S., Rakowicz, M., Schöls, L., Mariotti, C., et al. (2014). Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes. Brain 137, 2444-2455.

Traschutz, A., Cortese, A., Reich, S., Dominik, N., Faber, J., Jacobi, H., Hartmann, A.M., Rujescu, D., Montaut, S., Echaniz-Laguna, A., et al. (2021). Natural History, Phenotypic Spectrum, and Discriminative Features of Multisystemic *RFC1* Disease. Neurology 96, e1369-e1382.

Tsuchiya, M., Nan, H., Koh, K., Ichinose, Y., Gao, L., Shimozono, K., Hata, T., Kim, Y.J., Ohtsuka, T., Cortese, A., et al. (2020). *RFC1* repeat expansion in Japanese patients with late-onset cerebellar ataxia. J Hum Genet 65, 1143-1147.

Twist, E.C., Casaubon, L.K., Ruttledge, M.H., Rao, V.S., Macleod, P.M., Radvany, J., Zhao, Z., Rosenberg, R.N., Farrer, L.A., and Rouleau, G.A. (1995). Machado Joseph disease maps to the same region of chromosome 14 as the spinocerebellar ataxia type 3 locus. Journal of Medical Genetics 32, 25-31.

van de Warrenburg, B.P., Frenken, C.W., Ausems, M.G., Kleefstra, T., Sinke, R.J., Knoers, N.V., and Kremer, H.P. (2001). Striking anticipation in spinocerebellar ataxia type 7: the infantile phenotype. Journal of neurology 248, 911-914.

van de Warrenburg, B.P., Sinke, R.J., Verschuuren-Bemelmans, C.C., Scheffer, H., Brunt, E.R., Ippel, P.F., Maat-Kievit, J.A., Dooijes, D., Notermans, N.C., Lindhout, D., et al. (2002). Spinocerebellar ataxias in the Netherlands: prevalence and age at onset variance analysis. Neurology 58, 702-708.

van Rheenen, W., Shatunov, A., Dekker, A.M., McLaughlin, R.L., Diekstra, F.P., Pulit, S.L., van der Spek, R.A., Vosa, U., de Jong, S., Robinson, M.R., et al. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nat Genet 48, 1043-1048.

Vance, C., Al-Chalabi, A., Ruddy, D., Smith, B.N., Hu, X., Sreedharan, J., Siddique, T., Schelhaas, H.J., Kusters, B., Troost, D., et al. (2006). Familial amyotrophic lateral sclerosis with frontotemporal dementia is linked to a locus on chromosome 9p13.2-21.3. Brain 129, 868-876.

Veneziano, L., and Frontali, M. (1993). Drpla. In GeneReviews((R)), M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, G. Mirzaa, and A. Amemiya, eds. (Seattle (WA)).

Verkerk, A.J., Pieretti, M., Sutcliffe, J.S., Fu, Y.H., Kuhl, D.P., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F., Zhang, F.P., et al. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 65, 905-914.

Vermeer, S., van de Warrenburg, B.P., Kamsteeg, E.J., Brais, B., and Synofzik, M. (1993). Arsacs. In GeneReviews((R)), M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, G. Mirzaa, and A. Amemiya, eds. (Seattle (WA)).

Weyer, A., Abele, M., Schmitz-Hubsch, T., Schoch, B., Frings, M., Timmann, D., and Klockgether, T. (2007). Reliability and validity of the scale for the assessment and rating of ataxia: a study in 64 ataxia patients. Mov Disord 22, 1633-1637.

Wiatr, K., Piasecki, P., Marczak, L., Wojciechowski, P., Kurkowiak, M., Ploski, R., Rydzanicz, M., Handschuh, L., Jungverdorben, J., Brustle, O., et al. (2019). Altered Levels of Proteins and Phosphoproteins, in the Absence of Early Causative Transcriptional Changes, Shape the Molecular Pathogenesis in the Brain of Young Presymptomatic Ki91 SCA3/MJD Mouse. Mol Neurobiol.

Wightman, D.P., Jansen, I.E., Savage, J.E., Shadrin, A.A., Bahrami, S., Holland, D., Rongve, A., Borte, S., Winsvold, B.S., Drange, O.K., et al. (2021). A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nat Genet 53, 1276-1282.

Winkelmann, J., Schormair, B., Lichtner, P., Ripke, S., Xiong, L., Jalilzadeh, S., Fulda, S., Putz, B., Eckstein, G., Hauk, S., et al. (2007). Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. Nat Genet 39, 1000-1006.

Yoshida, K., Matsushima, A., and Nakamura, K. (2017). Inter-generational instability of inserted repeats during transmission in spinocerebellar ataxia type 31. J Hum Genet 62, 923-925.

Zanni, G., and Bertini, E. (2018). Chapter 11 - X-linked ataxias. In Handbook of Clinical Neurology, M. Manto, and T.A.G.M. Huisman, eds. (Elsevier), pp. 175-189.

Zhang, K., Cui S Fau - Chang, S., Chang S Fau - Zhang, L., Zhang L Fau - Wang, J., and Wang, J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study.

Zijlstra, M.P., Rujano, M.A., Waarde, M.A.V., Vis, E., Brunt, E.R., and Kampinga, H.H. (2010). Levels of DNAJB family members (HSP40) correlate with disease onset in patients with spinocerebellar ataxia type 3. European Journal of Neuroscience 32, 760-770.

Zoghbi, H.Y., Pollack, M.S., Lyons, L.A., Ferrell, R.E., Daiger, S.P., and Beaudet, A.L. (1988). Spinocerebellar ataxia: variable age of onset and linkage to human leukocyte antigen in a large kindred. Annals of neurology 23, 580-584.

Zu, T., Gibbens, B., Doty, N.S., Gomes-Pereira, M., Huguet, A., Stone, M.D., Margolis, J., Peterson, M., Markowski, T.W., Ingram, M.A. et al (2011). Non-ATG-initiated translation directed by microsatellite expansions. PNAS 108: 260–265

Appendix

Appendix includes the copyright transfer agreement for Chapter 2 and ethics certificate for RouBank.

COPYRIGHT TRANSFER AGREEMENT

Date: 14-09-2020

Contributor name: Fulya Akcimen

Contributor address: 1033 Avenue des pins Montreal Quebec Canada

Manuscript number: MDS-20-0971

Expanded CAG repeats in ATXN1, ATXN2, ATXN3 and HTT in the 1000

Re: Manuscript entitled: Genomes Project (the "Contribution")

for publication in: Movement Disorders (the "Journal")

Published by Wiley on behalf of The International Parkinson and Movement Disorder Society (the "Owner")

Dear Contributor(s):

Thank you for submitting your Contribution for publication. In order to expedite the editing and publishing process and enable the Owner to disseminate your Contribution to the fullest extent, we need to have this Copyright Transfer Agreement executed. If the Contribution is not accepted for publication, or if the Contribution is subsequently rejected, this Agreement shall be null and void. **Publication cannot proceed without a signed copy of this Agreement.**

A. COPYRIGHT

The Contributor assigns to the Owner, during the full term of copyright and any extensions or renewals, all copyright in and to the Contribution, and all rights therein, including but not limited to the right to reproduce, publish, republish, transmit, sell, transfer, distribute, and otherwise use the Contribution in whole or in part in electronic and print editions of the Journal and in derivative works throughout the world, in all languages and in all media of expression now known or later developed, and to license or permit others to do so.

B. RETAINED RIGHTS

Notwithstanding the above, the Contributor or, if applicable, the Contributor's employer, retains all proprietary rights other than copyright, such as patent rights, in any process, procedure or article of manufacture described in the Contribution. This reservation of rights does not affect or limit the rights assigned to Owner in Section A.

C. PERMITTED USES BY CONTRIBUTOR

1. License. The Owner grants to Contributor a non-exclusive, non-transferable and limited license to reproduce and distribute copies of the print or electronic "preprints" of the unpublished Contribution, in the original form submitted to the Journal prior to the peer review process, solely to colleagues within the Contributor's nonprofit organization or educational institution. The Contributor shall make no more than 100 printed copies of the preprints in any calendar year. Such preprints may be posted as electronic files on the Contributor's own personal website, on the Contributor's internal intranet at Contributor's nonprofit organization or educational institution, or on a secure external website at the Contributor's nonprofit organization or educational institution, provided that access is limited to employees and/or students at Contributor's non-profit organization or educational institution. Contributor shall not charge a fee for any

preprints, and Contributor's use under this Section C shall not be for any commercial purpose, or for any systematic external distribution (e.g., posting on a listserve, public website, database connected to a public access server, or automated delivery system). The license grant in this Section does not apply to for-profit corporations, and any proposed use outside of the scope of this Section C must be pre-approved in writing by the Owner. The rights granted to Contributor under this Section C do not include reproduction, distribution or any other use of rating scales, videos or other audiovisual materials associated with the Contribution.

2. Required Citation. Prior to publication, the Contributor must provide full credit and acknowledgement of the Journal in all preprints in the following format: This is a preprint of an article accepted for publication in [Journal Title], Copyright © [year] The International Parkinson and Movement Disorder Society. After publication, the Contributor must provide a citation to the Journal in all preprints in the following format: This is a preprint of an article that was published in [Journal title]: (Title of Article, Contributor, Journal Title and Volume/ Issue, Copyright © [year] The International Parkinson and Movement Disorder Society). An electronic link must be provided to the Journal's website, located at http://www.interscience.Wiley.com. The Contributor agrees not to update the preprint or replace it with the published version of the Contribution.

3. Accepted Version. Re-use of the accepted and peer-reviewed (but not the final typeset published) version of the Contribution (the "Accepted Version") is not permitted under this Agreement. There are separate arrangements with certain funding agencies governing reuse of the Accepted Version. Additional terms apply if the Contributor receives or received funding from these agencies. The details of those relationships, and other offerings allowing open web use, are set forth at the following website: http://www.wiley.com/go/ funderstatement.

4. Additional Terms for Certain Funders. Certain funders, including the NIH, members of the Research Councils UK (RCUK) and Wellcome Trust require deposit of the Accepted Version in a public repository after an embargo period. Details of funding arrangements are set out at the following website: http://www.wiley.com/go/funderstatement. Additional terms may be applicable. Please contact the production editor for the journal at MDSprod@wiley.com if you have additional funding requirements.

If any Contributor receiving funds from applicable sources does not choose the Owner's OnlineOpen option, the Contributor will be allowed to self-archive by depositing the Accepted Version in a public repository after the following applicable embargo period has expired, subject to further conditions imposed by the RCUK:

a. 12 months from first publication online of the final published version of the Contribution for research funded by members of the Research Councils UK (RCUK) other than The Economic and Social Research Council (ESRC) and the Arts and Humanities Research Council (AHRC); or
b. 24 months from first publication online of the final published version of the Contribution for research funded by ESRC or AHRC.

5. Additional Terms for Certain Institutions. Wiley has arrangements with certain educational institutions to permit the deposit of the Accepted Version in the institutional repository after an embargo period. Details of such arrangements are set out at the following website: http://olabout.wiley.com/WileyCDA/Section/id-406074.html . Additional terms may be applicable.

If any Contributor affiliated with these applicable educational institutions does not choose the Owner's OnlineOpen option, the Contributor will be allowed to self-archive by depositing the Accepted Version in the educational institution's repository after the following applicable embargo period has expired. See the following website for details: http://olabout.wiley.com/WileyCDA/Section/id-817011.html.

D. CONTRIBUTIONS OWNED BY EMPLOYER

If the Contribution was written by the Contributor in the course of the Contributor's employment (as a "work-made-for-hire" in the course of employment), the Contribution is owned by the company/institution which must execute this Agreement (in addition to the Contributor's signature). In such case, the company/institution hereby assigns to the Owner, during the full term of copyright, all copyright in and to the Contribution for the full term of copyright throughout the world as specified in Section A above.

E. GOVERNMENT CONTRACTS

In the case of a Contribution prepared under U.S. Government contract or grant, the U.S. Government may reproduce, without charge, all or portions of the Contribution and may authorize others to do so, for official U.S. Government purposes only, if the U.S. Government contract or grant so requires. (U.S. Government, U.K. Government, and other government employees: see notes at end.)

F. CONTRIBUTOR'S REPRESENTATIONS

The Contributor represents that the Contribution is the Contributor's original work, all individuals identified as Contributors actually contributed to the Contribution, and all individuals who contributed are included. The Contribution is submitted only to this Journal and has not been published before. (If excerpts from copyrighted works owned by third parties are included, the Contributor will obtain written permission from the copyright owners for all uses as set forth in the Journal's Instructions for Contributors, and show credit to the sources in the Contribution.) The Contributor also warrants that the Contribution contains no libelous or unlawful statements, does not infringe upon the rights (including without limitation the copyright, patent or trademark rights) or the privacy of others, or contain material or instructions that might cause harm or injury. Upon request, Contributor will provide the data or will cooperating fully in obtaining and providing the data on which the Contribution is based for examination by the editors or their assignees.

G. FINANCIAL DISCLOSURES

The Contributor certifies that his/her financial and material support for this research and work, regardless of date, is clearly identified on Exhibit A to this Agreement. The Contributor has also identified on Exhibit A, all other support unrelated to this research, covering the past year from the date of submission (e.g., grants, advisory boards, employment, consultancies, contracts, honoraria, royalties, expert testimony, partnerships, or stock ownership in medically-related fields).

H. VIDEO AND PHOTOGRAPHY CONSENT

In the event that the Contribution includes, discloses or incorporates any content (including, without limitation, any video clip or photograph) which identifies any individual patient(s) ("patient identifiable content"), the Contributor obtained from such patient(s) written consent to such inclusion, disclosure or incorporation and that this consent fully complies with all legal requirements, including without limitation, all of the requirements of the laws of the jurisdiction(s) to which the patient(s) and the patient(s)' physician are subject, including the United States Health Insurance Portability and Accountability Act of 1996 ("HIPAA") if applicable. The Contributor hereby certifies that, if the patient consent form is in a language other than English, such consent form meets all of the requirements set forth in the Instructions to Authors. In addition, the Contributor hereby confirms that he/she obtained from patient(s) written consent to use the patient identifiable content in both print and online (i.e., internet/web-based) publication formats. The Contributor further certifies that the person executing any such patient consent form, to the best of his/her knowledge, had legal capacity under applicable law to execute the form on behalf of the patient.

I. ACKNOWLEDGEMENTS

14-09-2020

The Contributor should obtain written permission from all individuals named in the acknowledgement since readers may infer their endorsement of data and conclusions. The Contributor certifies that all individuals named in the acknowledgement section have provided written permission to be named.

J. MISCELLANEOUS

This Agreement may be amended or modified only in a writing executed by both parties. The waiver or failure of any party to exercise any rights under this Agreement shall not be deemed a waiver or other limitation of any other right or any future right. This Agreement shall inure to the benefit of, and shall be binding upon, the parties, their respective successors and permitted assigns. This Agreement may be executed in two (2) or more counterparts, each of which shall be an original and all of which taken together shall constitute one and the same agreement. Executed copies of this Agreement may be delivered by facsimile transmission, pdf/email or other comparable electronic means. If for any reason any provision of this Agreement shall be deemed by a court of competent jurisdiction to be legally invalid or unenforceable, the validity, legality and enforceability of the remainder of this Agreement shall not be affected and such provision shall be deemed modified to the minimum extent necessary to make such provision consistent with applicable law and, in its modified form, such provision shall then be enforceable and enforced. The parties agree to do such further acts and to execute and deliver such additional agreements and instruments from time to time as either may at any time reasonably request in order to assure and confirm unto such requesting party the rights, powers and remedies conferred in the Agreement. This Agreement, including any exhibits attached hereto, contains the entire agreement and understanding of the parties with respect to the subject matter hereof, and supersedes all prior agreements, negotiations, representations and proposals, written and oral, relating thereto.

All Contributors must sign below. Contributors must check one box except that NIH grantees should check both Contributor-owned work and the NIH grantee box. If your Contribution was written during the course of employment, your employer must also sign where indicated.

Please send your original completed and signed forms by fax or email a scanned copy to the Journal production editor. For production editor contact details please visit the Journal's online author guidelines. **Do not send in hard copies of these forms.**

[X] Contributor-owned work

Fulya Akcimen
Contributor's signature

14-09-2020
Date

Type or print name and title

Co-Contributor's signature

Date

Type or print name and title

[_____] Company/Institution-owned Work (made-for-hire in the Course of employment)

_____ Company or Institution (Employer-for-Hire) _____ Date

_____ Authorized signature of Employer _____ Date

_____ Contributor's signature _____ Date

_____ Type or print name and title

ATTACH ADDITIONAL SIGNATURE PAGES AS NECESSARY

[_____] **U.S. Government work**

Note to U.S. Government Employees

A contribution prepared by a U.S. federal government employee as part of the employee's official duties, or which is an official U.S. Government publication, is called a "U.S. Government work", and is in the public domain in the United States. In such case, Paragraph A.1 will not apply but the Contributor must type his/her name (in the Contributor's signature line) above. Contributor acknowledges that the Contribution will be published in the United States and other countries. If the Contribution was not prepared as part of the employee's duties or is not an official U.S. Government publication, it is not a U.S. Government work.

[_____] **U.K. Government work (Crown Copyright)**

Note to U.K. Government Employees

The rights in a contribution prepared by an employee of a UK government department, agency or other Crown body as part of his/her official duties, or which is an official government publication, belong to the Crown. Contributors must ensure they comply with departmental regulations and submit the appropriate authorisation to publish. If your status as a government employee legally prevents you from signing this Agreement, please contact the Journal production editor.

[_____] **Other**

Including Other Government work or Non-Governmental Organisation work

Note to Non-U.S., Non-U.K. Government Employees or Non-Governmental Organisation Employees

If your status as a government or non-governmental organisation employee legally prevents you from signing this Agreement, please contact the Journal production editor.

Exhibit A

Financial Disclosure

The Contributor has received financial and material support for this research and work regardless of date from the following sources:

Name: _____

Address: _____

Type of support: _____

This material will be printed with the published article.

In the past year from the date of submission, the Contributor has also received the following support unrelated to this research (e.g., grants, advisory boards, employment, consultancies, contracts, honoraria, royalties, expert testimony, partnerships, or stock ownership in medically-related fields):

Name: _____

Address: _____

Type of support: _____

This material will be posted on the journal website and may be printed at the Editors' discretion.

ATTACH ADDITIONAL INFORMATION AS NECESSARY

MUHC REB – NEUPSY / CÉR CUSM – NEUPSY
3801, rue University, # 686
Montréal (Québec) H3A 2B4
Tel. 514.398.1046
reb.neuro@mcgill.ca
www.leneuro.ca

Annual renewal submission

Submit date: **2021-02-10 10:11**
Project's REB approbation date: **2015-03-20**
Project number(s): **2015-164, MP-CUSM-14-051, MP-37-2015-164**
Form status: **Approved**

Submitted by: **Zaharieva, Vessela**
Nagano identifier: **ROU BANK**
Form: **F9-70479**

Administration

1. **MUHC REB Panel & Co-chair(s):**
   Neurosciences-Psychiatry (NEUPSY)
   Co-chairs: Judith Marcoux, Brigitte Pâquet

2. **REB Decision:**
   Approved - REB delegated review

3. **Comments on the decision:**

   The renewal for ethics approval applies for the following centres:
   - McGill University Health Centre
   - Centre Hospitalier de l'Université de Montréal
   - CHU - de - Quebec
   - CHU - Sainte-Justine
   - CIUSSS de l'Ouest-de-l'Ile-de-Montréal

4. **Renewal Period Granted:**
   From 2021-03-20 Until 2022-03-19

5. **Date of the REB final decision & signature**

2021-02-12

Signature

Sonia Cantini
MUHC REB Coordinator
For MUHC Co-chair mentioned above

6. FWA 00000840 - FWA 00004545

7. **Local REB number**

IRB00010120

8. **Note:**

In order to be in compliance with Good Clinical Practices, the MUHC REB (when acting as the Reviewing REB), and the PM of the MUHC does not directly communicate with sponsors. The communication channels existing between the PIs and the sponsors will continue to ensure the transmission of documents.

A. General information

1. **Indicate the full title of the research study**

Rou Bank.

2. **If relevant, indicate the full study title in French**

3. **Indicate the name of the Principal Investigator in our institution (MUHC)**

Rouleau, Guy

4. **Are there local co-investigators & collaborators involved in this project?**

No

5. **For each participating centre part of the Québec health and social services network (RSSS), indicate the name of the external investigator**

Sylvain Chouinard

What is the name of the participating center(s)?

CHU-Montréal

Nicolas Dupré

What is the name of the participating center(s)?

CHU-de-Québec

Jacques Michaud

What is the name of the participating center(s)?

CHU-SJ

Gustavo Turecki

What is the name of the participating center(s)?

CIUSSS-OMTL

Ridha Joober

What is the name of the participating center(s)?

CIUSSS-OMTL

6. **Indicate the name and the affiliation of the external collaborator(s),(if any)**

Voir la liste des sections 5 & 9

7. **Identify the study coordinator(s)**

Zaharieva, Vessela

Indicate the role of the collaborator(s)

Administrative agent

Mirarchi, Cathy

Indicate the role of the collaborator(s)

Administrative agent

B. Project development

1. **Study start date:**

2006-09-21

2. **Expected ending date of the study:**

- ☐ Determined date
- ☑ Undetermined date

3. **Date of recruitment of the first participant?**

- ☑ 1st enrollment date is...
- ☐ No participant enrolled

1st participant enrollment date:

2006-09-22

4. **Indicate the current study status at MUHC.**

Study and recruitment in progress

5. **Add a brief statement on the study status**

Study is still in progress

6. **Information about the participants at this institution, since the beginning of the project**

Number of participants who have been recruited

17071

Number of participants who have not yet completed the study (still in progress)

0

Number of participants who've completed the study

17071

Number of participants who were recruited to the study, but who were then excluded or withdrawn:

0

Number of participants who dropped out (voluntary withdrawal):

0

Number of participants who died during the study

0

7. **Information about the participants at this institution (MUHC) since the previous REB approval**

Number of participants who have been recruited

100

Number of participants who have not yet completed the study (still in progress)

0

Number of participants who've completed the study

100

Number of participants who dropped out (voluntary withdrawal):

0

Number of participants who died during the study

0

8. **Since the previous REB approval (annual renewal or initial approval):**

Were there any changes to the protocol (or to the databank management framework) ?

No

Specify the current version/date:

version 1, March 14, 2016

Date approved by the REB:

2016-03-22

Were there any changes to the information and consent form?

Yes

Specify the current version/date:

version 2, December 2020

REB approval date:

2021-01-29

Were there any reportable adverse events at this site (or, for multi-center projects, at an institution under the jurisdiction of our REB) that should be reported to the REB under section 5.2.1 of " SOP- REB-404001 " ?

https://muhc.ca/cae/page/standard-operating-procedures-sops

No

Has there has been any new information likely to affect the ethics of the project or influence the decision of a participant as to their continued participation in the project ?

No

Were there any deviations / major violations protocol (life -threatening or not meeting the inclusion / exclusion criteria)?

No

Was there a temporary interruption of the project?

No

Have the project results been submitted for publication, presented or published?

No

Has the REB been notified of a conflict of interest - (apparent , potential or actual), of one or more members of the research team - that was not known when it was last approved project?

No

Do you want to bring any other info to the REB's attention?

No

9. **For all external participating institutions, please answer the following questions:**

Please select the name of the institution concerned and attach the "Formulaire de renouvellement annuel pour les projets sites externes - Projets multicentriques":

CHU-de-Québec

Please print a copy of the "Formulaire de renouvellement annuel pour les sites externes" (see link below), have it completed by other institutions and attach it here.

Formulaire de renouvellement pour les sites externes (MP project)

This form is accessible to external researchers, via our web page.

Demande de renouvellement annuel pour les sites externes-CHUdu Quebec 2021 (002)N. Dupre.pdf

CHU-SJ

Please print a copy of the "Formulaire de renouvellement annuel pour les sites externes" (see link below), have it completed by other institutions and attach it here.

Formulaire de renouvellement pour les sites externes (MP project)

This form is accessible to external researchers, via our web page.

Demande de renouvellement annuel pour les sites externes-CHUSJ 2021 (003)-signed.pdf

CHU-Montréal

Please print a copy of the "Formulaire de renouvellement annuel pour les sites externes" (see link below), have it completed by other institutions and attach it here.

Formulaire de renouvellement pour les sites externes (MP project)

This form is accessible to external researchers, via our web page.

document03-02-2021-090642 CHUM SC.pdf

CIUSSS-OMTL

Please print a copy of the "Formulaire de renouvellement annuel pour les sites externes" (see link below), have it completed by other institutions and attach it here.

Formulaire de renouvellement pour les sites externes (MP project)

This form is accessible to external researchers, via our web page.

20210209 Demande de renouvellement annuel pour les sites externes-CIUSS-OMTL 2021 GT.pdf

Is there any institution's info (pdf form) missing?

No

10. **Is there a data safety monitoring committee analyzing data on the safety and efficacy of the treatment?**

No

C. Signature

1. **I confirm that all information is complete & accurate.**

First & last name of person who completed the submission

Vessela Zaharieva
2021-02-10 10:11