

 Open access • Posted Content • DOI:10.1101/834754

Genome-wide association study identifies genetic factors that modify age at onset in Machado-Joseph disease — [Source link](#)

Fulya Akçimen, Sandra Martins, Sandra Martins, Calwing Liao ...+14 more authors

Institutions: Montreal Neurological Institute and Hospital,
Institute of Molecular Pathology and Immunology of the University of Porto, University of Porto, University of Sydney ...+5 more institutions

Published on: 08 Nov 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Machado–Joseph disease and Genome-wide association study

Related papers:

- [Genome-wide association study identifies genetic factors that modify age at onset in Machado-Joseph disease.](#)
- [A Pipeline to Assess Disease-Associated Haplotypes in Repeat Expansion Disorders: The Example of MJD/SCA3 Locus.](#)
- [Segregation distortion of wild-type alleles at the Machado-Joseph disease locus: a study in normal families from the Azores islands \(Portugal\)](#)
- [A multistep mutation mechanism drives the evolution of the CAG repeat at MJD/SCA3 locus](#)
- [Association Analysis of Chromosome X to Identify Genetic Modifiers of Huntington's Disease.](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/genome-wide-association-study-identifies-genetic-factors-ej2bgtsqub>

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}, H     Catoire^{2,5}, Garth A. Nicholson⁶, Olaf Riess⁷, Mafalda Raposo⁸, Marcondes C. Fran  a Jr.⁹, Jo  o Vasconcelos¹⁰, Manuela Lima⁸, 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*}

¹Department of Human Genetics, McGill University, Montr  al, QC, Canada;

²Montreal Neurological Institute and Hospital, McGill University, Montr  al, QC, Canada;

³i3S – Instituto de Investiga  o e Inova  o em Sa  de, Universidade do Porto, Portugal;

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

⁵Department of Neurology and Neurosurgery, McGill University, Montr  al, QC, Canada;

⁶University of Sydney, Department of Medicine, Concord Hospital, Australia;

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

⁸Faculdade de Ci  ncias e Tecnologia, Universidade dos A  ores e Instituto de Biologia Molecular e Celular (IBMC), Instituto de Investiga  o e Inova  o em Sa  de (i3S), Universidade do Porto, Portugal;

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

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

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

¹²Departamento de Neurologia, Hospital do Divino Esp  rito Santo, Ponta Delgada, Portugal;

¹³Medical Genetics Service, Hospital de Cl  nicas de Porto Alegre (HCPA), Porto Alegre, Brazil;

¹⁴Depto. de Bioqu  mica – ICBS, Universidade Federal do Rio Grande do Sul (UFRGS);

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

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

¹⁷Instituto de Ci  ncias Biom  dicas Abel Salazar (ICBAS), Universidade do Porto, Portugal.

***Corresponding author:** Guy A. Rouleau

Address: Montreal Neurological Institute and Hospital,

3801 University Street, Room 636,

Montr  al, Qu  bec, Canada H3A 2B4

Email: guy.rouleau@mcgill.ca

31 Phone: +1 (514) 398-6644

32

33

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. 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.

Keywords

Machado-Joseph disease, *ATXN3*, MJD/SCA3, age at onset, modifier, GWAS

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 [1, 2]. 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 consensually contain 61 to 87 CAG repeats, whereas wild type alleles range from 12 to 44 [2].

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 [3]. 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 [3,4]. 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* [4,5,6], these were not replicated by subsequent studies [7, 8]. 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 [8].

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 [9,10]. A total of eleven [9] and fourteen loci [10] 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.

Methods

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 1.

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* [11]. The final volume for each assay was 10 μ L: 7.5 ng of gDNA, 0.2 μ M of each primer, 5 μ L of Taq PCR Master Mix Kit Qiagen®, 1 μ L of Q-Solution from Qiagen® and H₂O. Fragment length analysis was done using ABIPrism 3730xl sequencer (Applied Biosystems®, McGill University and Genome Québec Innovation Centre) and GeneMapper software [12]. 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.

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 [13]. 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 ≥ 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 [14] and PBWT [15] pipelines, implemented on the Sanger Imputation Service [16]. 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.

Genome-wide association analysis

A genome-wide linear mixed model based association analysis was conducted using GCTA version 1.91.7 [17]. 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). The --mlma-loco option, which takes into account the difference in allele frequency between populations, was used to control for population structure. QQ plots and Manhattan plots were generated in FUMA v.1.3.4 [18]. Regional association plots were generated using LocusZoom [19] (Supplementary Figure 3).

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 $r^2 < 0.6$ 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 [20] categories and combined annotation-dependent depletion (CADD) [21] scores were obtained from FUMA for functional annotation. Functionally annotated variants were mapped to genes based on genomic position using FUMA positional mapping tool.

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 [22]. 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) [23] 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) [24], which controls for potential bias from gene size, SNP density, as well as LD. ClueGO [25] and CluePedia [26] plugins in Cytoscape were employed to visualize identified pathways and their clustering.

Results

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 [3] 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).

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 resulting Manhattan plots and quantile-quantile (QQ) plots are shown in Figure 2. The genomic inflation factor was close to one ($\lambda = 0.98$), indicating the p-values were not inflated. No association signal was identified meeting genome-wide significance ($P < 5 \times 10^{-8}$, the genome-wide Bonferroni-corrected significance threshold); however, 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 [27].

Interaction analysis between CAG_{exp} 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, and interaction of SNP:CAG_{exp}. Association of each independent SNP with AO revealed nominally significant p-values. 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}.

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.

Pathway and gene-set enrichment analysis

A gene-set enrichment and pathway analysis was conducted using i-GSEA4GWAS. 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 and PASCAL 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.

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 [3]. It was shown previously that normal *ATXN3* allele (CAG_{nor}) had a significant influence on AO of MJD [28]; however,

several studies did not replicate this effect [6,8]. 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 [29,30]. 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 [31,32]. *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 [32]. Previously, the *MLH1* locus was identified as an AO modifier in another neurodegenerative disease caused by CAG repeat expansion, Huntington's disease [9,10,33]. 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 [34]. 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*, *RPA*, and *CDK7*) associated with different CAG instability patterns in MJD [35].

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 [36,37]. 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 [38]. Vesicle transport and synaptic pathways were also implicated in MJD, as well as in other neurodegenerative diseases [39,40]. 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 [41,42]. 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.

239 **Description of Supplemental Data**

240 Supplemental Data include three figures and seven tables

241 **Declaration of Interests**

242 The authors declare no conflict of interest.

243 **Acknowledgements**

244 The authors thank the participants for their contribution to the study. The authors would like to
 245 thank Jay P. Ross, Faezeh Sarayloo, Zoe Schmilovich and S. Can Akerman for their assistance in
 246 reviewing the manuscript and scientific content. FA was funded by the Fonds de Recherche du
 247 Québec–Santé. SM is funded by FCT (CEECIND/00684/2017) and by NORTE-01-0145-FEDER-
 248 000008, supported by Norte Portugal Regional Programme (NORTE 2020), under the
 249 PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund
 250 (ERDF). FM and LI are funded by Fundação de Amparo a Pesquisa do Estado de São Paulo
 251 (FAPESP, 2013/07559-3). MLSP and LBJ were funded by Conselho Nacional de
 252 Desenvolvimento Científico e Tecnológico, Brazil (CNPq) and by Coordenação de
 253 Aperfeiçoamento de Pessoal de Nível Superior (CAPES). GAR holds a Canada Research Chair in
 254 Genetics of the Nervous System and the Wilder Penfield Chair in Neurosciences.

255

256

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	<i>THSD7B</i>	G	C	0.13	0.11	2.71	4.45×10^{-6}
rs2067390	2	191209028	<i>HIBCH, INPP1</i>	A	T	0.04	0.06	4.74	6.39×10^{-6}
rs144891322	5	85135387	<i>RPL5P17,</i>	C	T	0.02	0.007	6.10	5.18×10^{-6}
rs11529293	11	36855388	<i>C11orf74, RAG1, RAG2</i>	T	C	0.14	0.26	-2.71	3.30×10^{-6}
rs7480166	11	42984753	<i>HNRNPKP3</i>	A	G	0.40	0.40	-1.86	4.17×10^{-6}
rs585809	11	119949979	<i>TRIM29</i>	T	C	0.06	0.17	-3.76	9.50×10^{-6}
rs72660056	13	113507543	<i>ATP11A</i>	A	G	0.08	0.05	-3.29	3.94×10^{-6}
rs11857349	15	99924857	<i>TTC23, SYNM, LRRC28</i>	G	A	0.04	0.02	-4.58	3.43×10^{-6}
rs8141510	22	42821185	<i>NFAMI, CYP2D6, NAGA, NDUFA6</i>	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}	$< 1.0 \times 10^{-6}$	$< 1.0 \times 10^{-7}$

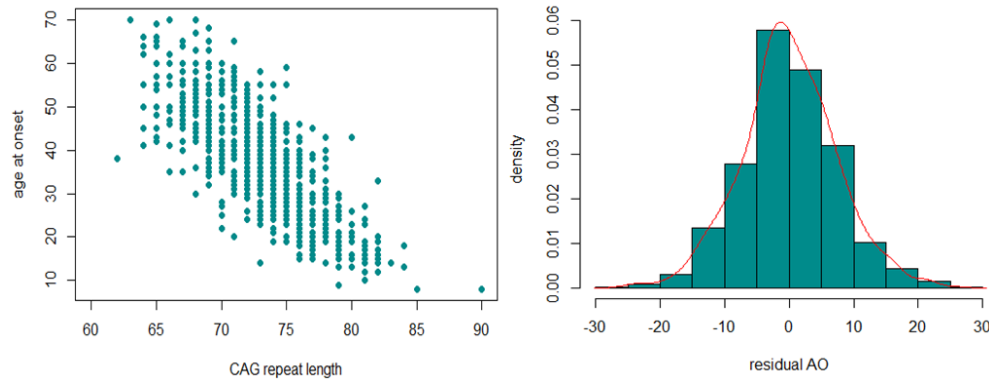


Figure 1. The inverse correlation between CAG_{exp} and AO (left) and the distribution of residual AO (right) observed in our MJD cohort.

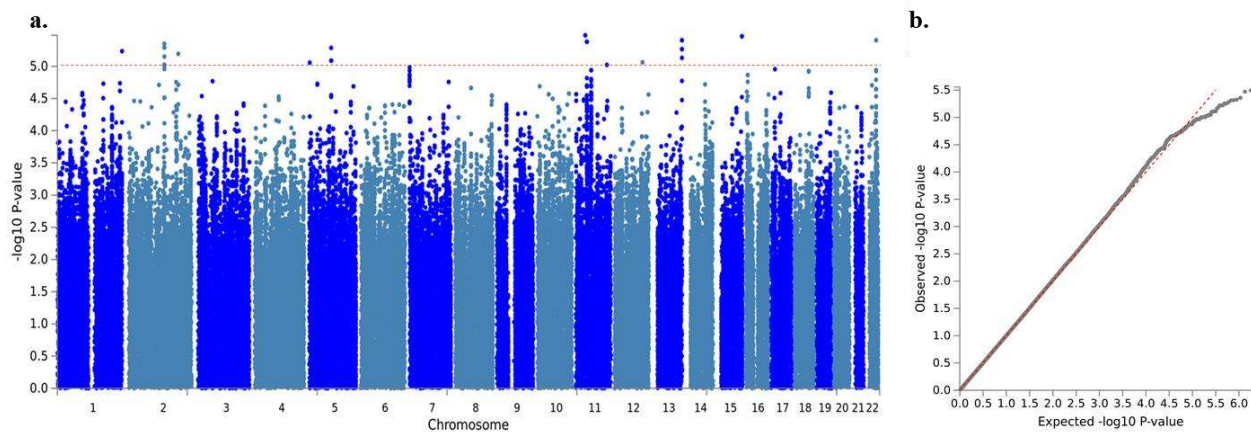
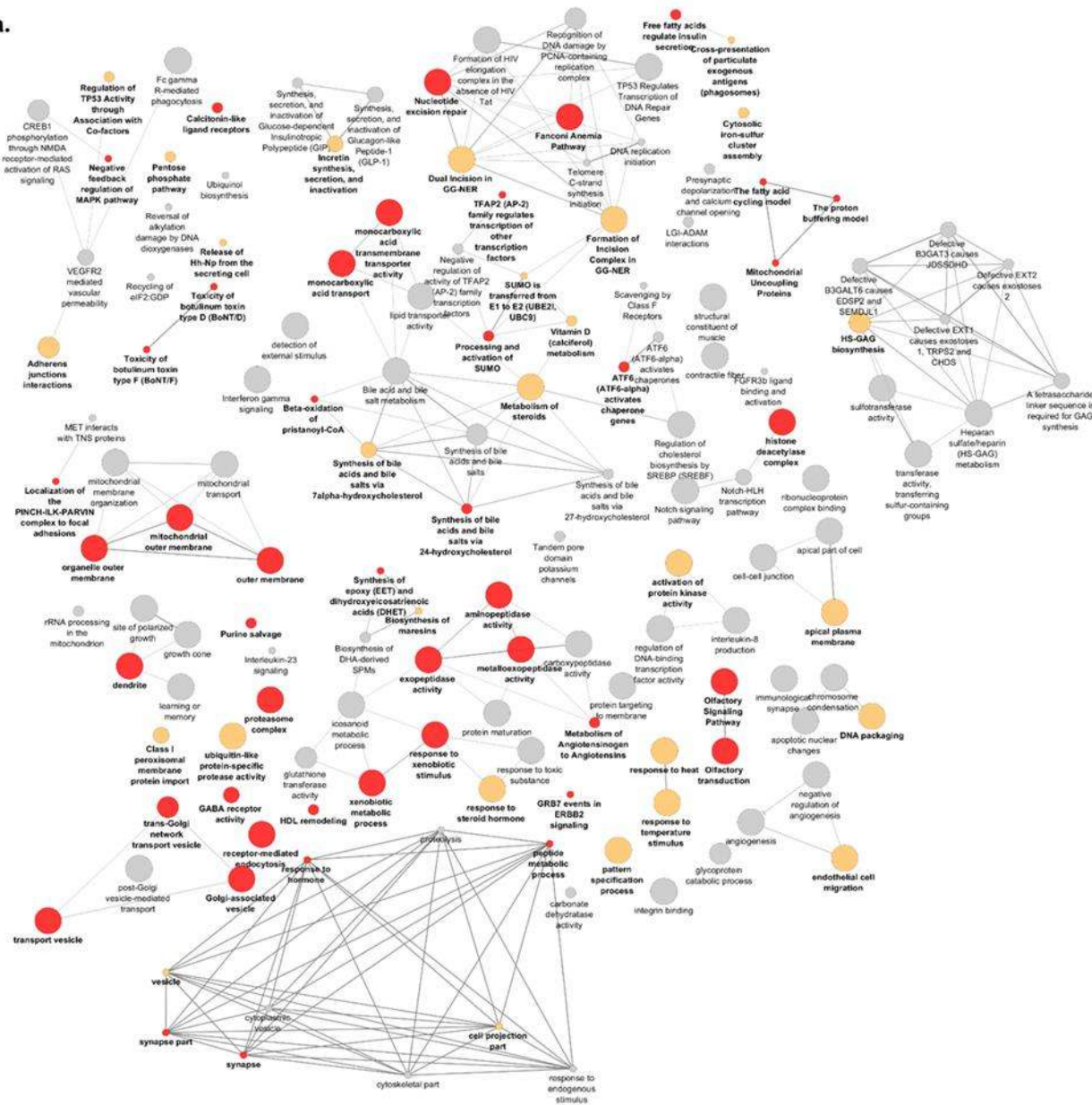
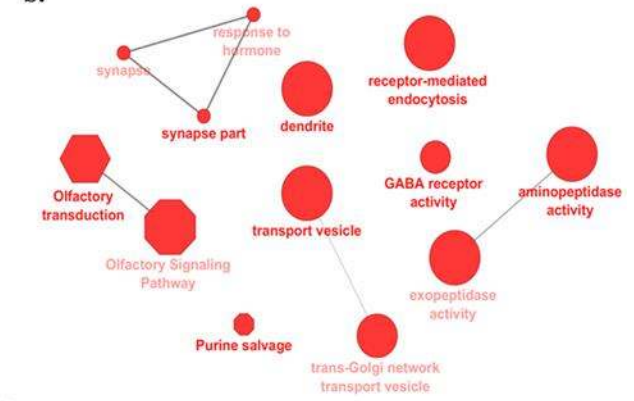


Figure 2. Manhattan plot (a) and QQ plot (b) of the GWAS for residual AO of MJD. Imputed using the HRC panel, 6,716,580 variants that passed QC are included in the plot. The x-axis shows the physical position along the genome. The y-axis shows the $-\log_{10}(\text{p-value})$ for association. The red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

a.



b.



271 **Figure 3.** Visualization of the gene-sets and pathways enriched in primary GSEA analysis (a) and replicated in VEGAS and PASCAL
272 (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
273 the nodes ($P < 0.05$, $0.05 < P < 0.1$ and $P > 0.1$ are represented by red, yellow and gray, respectively).

References

1. Twist, E.C. *et al.* 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 (1995).
2. Bettencourt, C. & Lima, M. Machado-Joseph Disease: from first descriptions to new perspectives. *Orphanet J Rare Dis* **6**, 35 (2011).
3. Maciel, P. *et al.* Correlation between CAG repeat length and clinical features in Machado-Joseph disease. *Am J Hum Genet* **57**, 54-61 (1995).
4. de Mattos, E.P., Kolbe Musskopf, M., Bielefeldt Leotti, V., Saraiva-Pereira, M.L. & Jardim, L.B. 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 (2019).
5. Zijlstra, M.P. *et al.* Levels of DNAJB family members (HSP40) correlate with disease onset in patients with spinocerebellar ataxia type 3. *European Journal of Neuroscience* **32**, 760-770 (2010).
6. Tezenas du Montcel, S. *et al.* Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes. *Brain* **137**, 2444-2455 (2014).
7. Chen, Z. *et al.* (CAG)_n loci as genetic modifiers of age-at-onset in patients with Machado-Joseph disease from mainland China. *Brain* **139**, e41-e41 (2016).
8. Raposo, M., Ramos, A., Bettencourt, C. & Lima, M. Replicating studies of genetic modifiers in spinocerebellar ataxia type 3: can homogeneous cohorts aid? *Brain* **138**, e398-e398 (2015).
9. Genetic Modifiers of Huntington's Disease, C. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell* **162**, 516-26 (2015).
10. Lee, J.-M. *et al.* CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell* **178**, 887-900.e14 (2019).
11. Martins, S., Calafell, F., Wong, V.C., Sequeiros, J. & Amorim, A. A multistep mutation mechanism drives the evolution of the CAG repeat at MJD/SCA3 locus. *Eur J Hum Genet* **14**, 932-40 (2006).
12. Chatterji, S. & Pachter, L. Reference based annotation with GeneMapper. *Genome biology* **7**, R29-R29 (2006).
13. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
14. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179 (2011).
15. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics (Oxford, England)* **30**, 1266-1272 (2014).
16. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279-1283 (2016).
17. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88**, 76-82 (2011).
18. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications* **8**, 1826 (2017).
19. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics (Oxford, England)* **26**, 2336-2337 (2010).
20. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
21. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* **47**, D886-D894 (2018).

- 321 22. Zhang, K., Cui, S., Chang, S., Zhang, L. & Wang, J. i-GSEA4GWAS: a web server for identification
322 of pathways/gene sets associated with traits by applying an improved gene set enrichment
323 analysis to genome-wide association study. *Nucleic Acids Research* **38**, W90-W95 (2010).
- 324 23. Mishra, A. & Macgregor, S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin*
325 *Research and Human Genetics* **18**, 86-91 (2014).
- 326 24. Lamparter, D., Marbach, D., Rueedi, R., Kutilik, Z. & Bergmann, S. Fast and Rigorous
327 Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput*
328 *Biol* **12**, e1004714 (2016).
- 329 25. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology
330 and pathway annotation networks. *Bioinformatics (Oxford, England)* **25**, 1091-1093 (2009).
- 331 26. Bindea, G., Galon, J. & Mlecnik, B. CluePedia Cytoscape plugin: pathway insights using integrated
332 experimental and in silico data. *Bioinformatics (Oxford, England)* **29**, 661-663 (2013).
- 333 27. Amendola, L.M. *et al.* Actionable exomic incidental findings in 6503 participants: challenges of
334 variant classification. *Genome research* **25**, 305-315 (2015).
- 335 28. Franca, M.C., Jr. *et al.* Normal ATXN3 Allele but Not CHIP Polymorphisms Modulates Age at
336 Onset in Machado-Joseph Disease. *Front Neurol* **3**, 164 (2012).
- 337 29. Lescale, C. & Deriano, L. The RAG recombinase: Beyond breaking. *Mechanisms of Ageing and*
338 *Development* **165**, 3-9 (2017).
- 339 30. Bahjat, M. & Guikema, J.E.J. The Complex Interplay between DNA Injury and Repair in
340 Enzymatically Induced Mutagenesis and DNA Damage in B Lymphocytes. *Int J Mol Sci* **18**(2017).
- 341 31. Wikiniyadhanee, R., Leksuthirath, T., Stithantrakul, W., Chitphuk, S. & Dejsuphong, D. AB064.
342 TRIM29: a novel gene involved in DNA repair mechanisms. *Annals of Translational Medicine* **5**,
343 AB064-AB064 (2017).
- 344 32. Masuda, Y. *et al.* TRIM29 regulates the assembly of DNA repair proteins into damaged
345 chromatin. *Nat Commun* **6**, 7299 (2015).
- 346 33. Lee, J.-M. *et al.* A modifier of Huntington's disease onset at the MLH1 locus. *Human Molecular*
347 *Genetics* **26**, 3859-3867 (2017).
- 348 34. Pinto, R.M. *et al.* Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's
349 disease mice: genome-wide and candidate approaches. *PLoS Genet* **9**, e1003930 (2013).
- 350 35. Martins, S. *et al.* Modifiers of (CAG)(n) instability in Machado-Joseph disease (MJD/SCA3)
351 transmissions: an association study with DNA replication, repair and recombination genes. *Hum*
352 *Genet* **133**, 1311-8 (2014).
- 353 36. Braga-Neto, P. *et al.* Clinical correlates of olfactory dysfunction in spinocerebellar ataxia type 3.
354 *Parkinsonism Relat Disord* **17**, 353-6 (2011).
- 355 37. Pedroso, J.L. *et al.* Nonmotor and extracerebellar features in Machado-Joseph disease: a review.
356 *Mov Disord* **28**, 1200-8 (2013).
- 357 38. Hsieh, J. *et al.* Human Olfactory Ensheathing Cell Transplantation Improves Motor Function in a
358 Mouse Model of Type 3 Spinocerebellar Ataxia. *Cell Transplant* **26**, 1611-1621 (2017).
- 359 39. Wiatr, K. *et al.* Altered Levels of Proteins and Phosphoproteins, in the Absence of Early Causative
360 Transcriptional Changes, Shape the Molecular Pathogenesis in the Brain of Young
361 Presymptomatic Ki91 SCA3/MJD Mouse. *Mol Neurobiol* (2019).
- 362 40. Gissen, P. & Maher, E.R. Cargos and genes: insights into vesicular transport from inherited
363 human disease. *J Med Genet* **44**, 545-55 (2007).
- 364 41. Gunawardena, S. & Goldstein, L.S.B. Polyglutamine Diseases and Transport Problems: Deadly
365 Traffic Jams on Neuronal Highways. *JAMA Neurology* **62**, 46-51 (2005).
- 366 42. Khan, L.A. *et al.* Expanded polyglutamines impair synaptic transmission and ubiquitin-
367 proteasome system in *Caenorhabditis elegans*. *J Neurochem* **98**, 576-87 (2006).

