Genome-wide association study identifies genetic factors that modify age at onset in Machado-Joseph disease

Fulya Akçimen1,2, Sandra Martins3,4, Calwing Liao1,2, Cynthia V. Bourassa2,5, Hélène Catoire2,5, Garth A. Nicholson6, Olaf Riess7, Mafalda Raposo8, Marcondes C. França9, João Vasconcelos10, Manuela Lima8, Iscia Lopes-Cendes11,12, Maria Luiza Saraiva-Pereira13,14, Laura B. Jardim13,15, Jorge Sequeiros4,16,17, Patrick A. Dion2,5, Guy A. Rouleau1,2,5

1Department of Human Genetics, McGill University, Montréal, Québec, Canada  
2Montreal Neurological Institute and Hospital, McGill University, Montréal, Québec, Canada  
3i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal  
4IPATIMUP – Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal  
5Department of Neurology and Neurosurgery, McGill University, Montréal, Québec, Canada  
6University of Sydney, Department of Medicine, Concord Hospital, Concord, Australia  
7Institute of Medical Genetics and Applied Genomics, University of Tuebingen, Tuebingen, Germany  
8Faculdade de Ciências e Técnologia, Universidade dos Açores e Instituto de Biologia Molecular e Celular (IBMC), Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto, Portugal  
9Department of Neurology, Faculty of Medical Sciences, UNICAMP, São Paulo, Campinas, Brazil  
10School of Medical Sciences, Department of Medical Genetics and Genomic Medicine, University of Campinas (UNICAMP), São Paulo, Campinas, Brazil  
11The Brazilian Institute of Neuroscience and Neurotechnology (BRAINN), São Paulo, Campinas, Brazil  
12Departamento de Neurologia, Hospital do Divino Espírito Santo, Ponta Delgada, Portugal  
13Medical Genetics Service, Hospital de Clínica de Porto Alegre (HCPA), Porto Alegre, Brazil  
14Departamento de Bioquímica – ICBS, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil  
15Departamento 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 Ciências Biomédicas Abel Salazar (ICBAS), Universidade do Porto, Porto, Portugal

Correspondence to: Guy A. Rouleau; email: guy.rouleau@mcgill.ca  
Keywords: Machado-Joseph disease, GWAS, age at onset, ATXN3, modifier

Received: November 20, 2019 Accepted: January 27, 2020 Published: March 23, 2020

Copyright: Akçimen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

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 R² = 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 × 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.
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 (CAG$_{exp}$) consensually contain 61 to 87 CAG repeats, whereas wild type alleles (CAG$_{nor}$) 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 [4–6], 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.

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 (AR$^2 = 0.0072$). 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).

Figure 1. The inverse correlation between CAG$_{exp}$ and AO (left) and the distribution of residual AO (right) observed in our MJD cohort.
**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 [11].

**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 ($P [rs7480166] = 8.42 \times 10^{-6}$, $P [rs62171220] = 6.33 \times 10^{-3}$, $P [rs2067390] = 4.51 \times 10^{-6}$, $P [rs144891322] = 1.14 \times 10^{-5}$, $P [rs11529293] = 1.62 \times 10^{-5}$, $P [rs585809] = 2.91 \times 10^{-5}$, $P [rs72660056] = 1.66 \times 10^{-3}$, $P [rs11857349] = 8.21 \times 10^{-6}$, $P [rs8141510] = 1.33 \times 10^{-3}$). 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 ($\Delta R^2 = 0.005$).

**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 v2 [12]. 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 [13] and PASCAL [14] 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\textsubscript{exp} and AO [3]. It was shown previously that normal \textit{ATXN3} allele (CAG\textsubscript{nor}) had a significant influence on AO of MJD [15]; however, several studies did not replicate this effect [6, 8]. Indeed, we did not observe an association between CAG\textsubscript{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 [16, 17]. The rs585809 variant, which was mapped to the \textit{TRIM29} gene, was found to interact with CAG\textsubscript{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 [18, 19]. 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 [19]. Previously, the \textit{MLH1} locus was identified as an AO modifier in another neurodegenerative disease caused by CAG repeat expansion, Huntington’s disease [9, 10, 20]. 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 [21].

![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).](image-url)
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 [22].

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 [23, 24]. 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 [25]. Vesicle transport and synaptic pathways were also implicated in MJD, as well as in other neurodegenerative diseases [26, 27]. 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 [28, 29]. 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.

MATERIALS AND 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 2.

Assessment of the ATXN3 CAG repeat length

A singleplex polymerase chain reaction was performed to determine the length of the CAGexp and CAGnor alleles at exon 10 of ATXN3 [30]. 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 H2O. Fragment length analysis was done using ABIPrism 3730x1 sequencer (Applied Biosystems®, McGill University and Genome Québec Innovation Centre) and GeneMapper software [31]. A stepwise regression model was performed to assess the correlation between AO and CAGexp size, as well as gender, origin, CAGnor size, and interaction between these variables. Residual AO was calculated for each subject by subtracting individual’s expected AO based upon CAGexp 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 multi-dimensional scaling analysis) and SNP-based quality assessments (missingness, Hardy-Weinberg equilibrium, and minor allele frequency) were conducted using PLINK version 1.9 [32]. 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 [33] and PBWT [34] pipelines, implemented on the Sanger Imputation Service [35]. 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 –mlma-loco option of GCTA version 1.91.7 [36]. 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 [37]. Regional association plots were generated using LocusZoom [38] (Supplementary Figure 2).

**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 [39] categories and combined annotation-dependent depletions (CADD) [40] 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. 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 a $P$-value implemented in FUma [31]. Independent suggestive SNPs and ranks them based on the strength of their association 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 a $P$-value implemented in FUma [31].

**ACKNOWLEDGMENTS**

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.

**CONFLICTS OF INTEREST**

The authors declare no conflicts of interest.

**FUNDING**

FA and CL were funded by the Fonds de Recherche du Québec–Santé. SM is funded by FCT (CEECIND/00684/2017) and by NORTE-01-0145-FEEDER-000808, supported by Norte Portugal Regional Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF). FM and LI are funded by Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP, 2013/07559-3). MLSP and LBJ were funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil (CNPq) and by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). GAR holds a Canada Research Chair in Genetics of the Nervous System and the Wilder Penfield Chair in Neurosciences.

**REFERENCES**


https://doi.org/10.1038/ng.3643 PMID:27548312

https://doi.org/10.1016/j.ajhg.2010.11.011 PMID:21167468

https://doi.org/10.1038/s41467-017-01261-5 PMID:29184056

https://doi.org/10.1093/bioinformatics/btq419 PMID:20634204

https://doi.org/10.1093/nar/gkq603 PMID:20601685

https://doi.org/10.1093/nar/gky1016 PMID:30371827

https://doi.org/10.1093/bioinformatics/btp101 PMID:19237447

https://doi.org/10.1093/bioinformatics/btt019 PMID:23325622
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).
Supplementary Tables

Supplementary Table 1. Linear relationship between AO and CAG\textsuperscript{exp}, CAG\textsuperscript{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.

<table>
<thead>
<tr>
<th>Model description</th>
<th>Multiple $R^2$</th>
<th>Adjusted $R^2$</th>
<th>P-value</th>
<th>$\Delta R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AO $\sim$ CAG\textsuperscript{exp}</td>
<td>0.6200</td>
<td>0.6195</td>
<td>&lt;2.2 $\times$ 10^{-16}</td>
<td></td>
</tr>
<tr>
<td>AO $\sim$ CAG\textsuperscript{exp} + origin</td>
<td>0.6241</td>
<td>0.6216</td>
<td>&lt;2.2 $\times$ 10^{-16}</td>
<td>0.0021</td>
</tr>
<tr>
<td>AO $\sim$ CAG\textsuperscript{exp} + origin + sex</td>
<td>0.6265</td>
<td>0.6235</td>
<td>&lt;2.2 $\times$ 10^{-16}</td>
<td>0.0019</td>
</tr>
<tr>
<td>AO $\sim$ CAG\textsuperscript{exp} + origin + sex + CAG\textsuperscript{nor}</td>
<td>0.6282</td>
<td>0.6247</td>
<td>&lt;2.2 $\times$ 10^{-16}</td>
<td>0.0012</td>
</tr>
<tr>
<td>AO $\sim$ CAG\textsuperscript{exp} + origin + sex + CAG\textsuperscript{nor} + CAG\textsuperscript{exp}:CAG\textsuperscript{nor}</td>
<td>0.6301</td>
<td>0.6261</td>
<td>&lt;2.2 $\times$ 10^{-16}</td>
<td>0.0014</td>
</tr>
<tr>
<td>AO $\sim$ CAG\textsuperscript{exp} + origin + sex + CAG\textsuperscript{nor} + CAG\textsuperscript{exp}:CAG\textsuperscript{nor} + CAG\textsuperscript{exp}:origin</td>
<td>0.6328</td>
<td>0.6267</td>
<td>&lt;2.2 $\times$ 10^{-16}</td>
<td>0.0006</td>
</tr>
<tr>
<td>AO $\sim$ CAG\textsuperscript{exp} + origin + sex + CAG\textsuperscript{nor} + CAG\textsuperscript{exp}:CAG\textsuperscript{nor} + CAG\textsuperscript{exp}:origin + CAG\textsuperscript{exp}:origin</td>
<td>0.6352</td>
<td>0.6271</td>
<td>&lt;2.2 $\times$ 10^{-16}</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

Supplementary Table 2. Subjects and cohort demographics.

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

M:F - male-female ratio.

Please browse Full Text version to see the data of Supplementary Tables 3 and 4.

Supplementary Table 3. Functional annotation of SNPs. rsID: reference SNP ID; MAF: minor allele frequency; $r^2$: the maximum $r^2$ of the SNP with one of the independent significant SNPs; nearest gene: the nearest Gene of the SNP based on ANNOVAR annotations; distance: distance to the nearest gene; func: functional consequence of the SNP on the gene obtained from ANNOVAR; CADD: CADD score which is computed based on 63 annotations.

Supplementary Table 4. Previously identified HD-AO modifier loci in MJD.
### Supplementary Table 5. Gene sets and pathways enriched in i-GSEA4GWAS.

<table>
<thead>
<tr>
<th>Pathway/Gene set name</th>
<th>Gene Set ID</th>
<th>P-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO: TRANSPORT VESICLE</td>
<td>GO:0030133</td>
<td>0.001</td>
<td>0.0082</td>
</tr>
<tr>
<td>KEGG: OLFACTORY TRANSDUCTION</td>
<td>hsa04740</td>
<td>0.001</td>
<td>0.0083</td>
</tr>
<tr>
<td>Olfactory Signaling Pathway</td>
<td>R-HSA-381753</td>
<td>0.001</td>
<td>0.0088</td>
</tr>
<tr>
<td>GO: SYNAPSE PART</td>
<td>GO:0044456</td>
<td>0.001</td>
<td>0.0093</td>
</tr>
<tr>
<td>Synthesis of bile acids and bile salts via 24-hydroxycholesterol</td>
<td>R-HSA-193775</td>
<td>0.001</td>
<td>0.0095</td>
</tr>
<tr>
<td>Negative feedback regulation of MAPK pathway</td>
<td>R-HSA-5674499</td>
<td>0.001</td>
<td>0.0101</td>
</tr>
<tr>
<td>Purine salvage</td>
<td>R-HSA-74217</td>
<td>0.001</td>
<td>0.0106</td>
</tr>
<tr>
<td>GO: METALLOEXOPEPTIDASE ACTIVITY</td>
<td>GO:008235</td>
<td>0.001</td>
<td>0.0109</td>
</tr>
<tr>
<td>GO: SYNAPSE</td>
<td>GO:0045202</td>
<td>0.001</td>
<td>0.0115</td>
</tr>
<tr>
<td>GO: MONOCARBOXYLIC ACID TRANSPORT</td>
<td>GO:0015718</td>
<td>0.001</td>
<td>0.0133</td>
</tr>
<tr>
<td>GO: OUTER MEMBRANE</td>
<td>GO:0019867</td>
<td>0.001</td>
<td>0.0141</td>
</tr>
<tr>
<td>GO: AMINOPEPTIDASE ACTIVITY</td>
<td>GO:004177</td>
<td>0.001</td>
<td>0.015</td>
</tr>
<tr>
<td>Localization of the PINCH-ILK-PARVIN complex to focal adhesions</td>
<td>R-HSA-446343</td>
<td>0.001</td>
<td>0.0159</td>
</tr>
<tr>
<td>BioCarta: PLATELETAPP PATHWAY</td>
<td>M6487</td>
<td>0.001</td>
<td>0.0161</td>
</tr>
<tr>
<td>GO: EXOPEPTIDASE ACTIVITY</td>
<td>GO:0008238</td>
<td>0.001</td>
<td>0.018</td>
</tr>
<tr>
<td>TFAP2 (AP-2) family regulates transcription of other transcription factors</td>
<td>R-HSA-8866906</td>
<td>0.001</td>
<td>0.0212</td>
</tr>
<tr>
<td>GO: MITOCHONDRIAL OUTER MEMBRANE</td>
<td>GO:0005741</td>
<td>0.001</td>
<td>0.0215</td>
</tr>
<tr>
<td>GO: ORGANELLE OUTER MEMBRANE</td>
<td>GO:0031968</td>
<td>0.003</td>
<td>0.0216</td>
</tr>
<tr>
<td>GO: HISTONE DEACETYLASE COMPLEX</td>
<td>GO:0000118</td>
<td>0.001</td>
<td>0.0218</td>
</tr>
<tr>
<td>Fanconi Anemia Pathway</td>
<td>R-HSA-6783310</td>
<td>0.001</td>
<td>0.0221</td>
</tr>
<tr>
<td>GO: PROTEASOME COMPLEX</td>
<td>GO:0000502</td>
<td>0.001</td>
<td>0.0221</td>
</tr>
<tr>
<td>GO: RECEPTOR MEDIATED ENDOCYTOSIS</td>
<td>GO:0006898</td>
<td>0.001</td>
<td>0.0225</td>
</tr>
<tr>
<td>GO: GABA RECEPTOR ACTIVITY</td>
<td>GO:0016917</td>
<td>0.001</td>
<td>0.0226</td>
</tr>
<tr>
<td>Metabolism of Angiotensinogen to Angiotensins</td>
<td>R-HSA-202377</td>
<td>0.001</td>
<td>0.0231</td>
</tr>
<tr>
<td>GO: TRANS GOLGI NETWORK TRANSPORT VESICLE</td>
<td>GO:0030140</td>
<td>0.001</td>
<td>0.0236</td>
</tr>
<tr>
<td>Calcitonin-like ligand receptors</td>
<td>R-HSA-419812</td>
<td>0.001</td>
<td>0.0242</td>
</tr>
<tr>
<td>GO: MONOCARBOXYLIC ACID TRANSMEMBRANE TRANSPORTER ACTIVITY</td>
<td>GO:0008028.</td>
<td>0.001</td>
<td>0.0245</td>
</tr>
<tr>
<td>HDL remodeling</td>
<td>R-HSA-8964058</td>
<td>0.001</td>
<td>0.0262</td>
</tr>
<tr>
<td>Synthesis of epoxy (EET) and dihydroxyeicosatrienoic acids (DHET)</td>
<td>R-HSA-2142670</td>
<td>0.001</td>
<td>0.0263</td>
</tr>
<tr>
<td>GO: RESPONSE TO HORMONE STIMULUS</td>
<td>GO:0009725</td>
<td>0.001</td>
<td>0.0273</td>
</tr>
<tr>
<td>GO: PEPTIDE METABOLIC PROCESS</td>
<td>GO:0006518</td>
<td>0.004</td>
<td>0.0348</td>
</tr>
<tr>
<td>GRB7 events in ERBB2 signaling</td>
<td>R-HSA-1306955</td>
<td>0.001</td>
<td>0.0348</td>
</tr>
<tr>
<td>Mitochondrial Uncoupling Proteins</td>
<td>R-HSA-166187</td>
<td>0.008</td>
<td>0.0354</td>
</tr>
<tr>
<td>The fatty acid cycling model</td>
<td>R-HSA-167826</td>
<td>0.008</td>
<td>0.0354</td>
</tr>
<tr>
<td>The proton buffering model</td>
<td>R-HSA-167827</td>
<td>0.008</td>
<td>0.0354</td>
</tr>
<tr>
<td>ATF6 (ATF6-alpha) activates chaperone genes</td>
<td>R-HSA-381183</td>
<td>0.002</td>
<td>0.037</td>
</tr>
<tr>
<td>GO: GOLGI ASSOCIATED VESICLE</td>
<td>GO:0005798</td>
<td>0.003</td>
<td>0.0371</td>
</tr>
<tr>
<td>Free fatty acids regulate insulin secretion</td>
<td>R-HSA-400451</td>
<td>0.001</td>
<td>0.0374</td>
</tr>
<tr>
<td>GO: XENOBIOLOGICAL METABOLIC PROCESS</td>
<td>GO:0006805</td>
<td>0.001</td>
<td>0.0382</td>
</tr>
<tr>
<td>GO: DENDRITE</td>
<td>GO:0030425</td>
<td>0.001</td>
<td>0.0386</td>
</tr>
<tr>
<td>Beta-oxidation of pristanoyl-CoA</td>
<td>R-HSA-389887</td>
<td>0.001</td>
<td>0.0413</td>
</tr>
<tr>
<td>Toxicity of botulinum toxin type D (BoNT/D)</td>
<td>R-HSA-5250955</td>
<td>0.001</td>
<td>0.0414</td>
</tr>
</tbody>
</table>
Toxicity of botulinum toxin type F (BoNT/F)  R-HSA-5250981  0.001  0.0414
Processing and activation of SUMO  R-HSA-3215018  0.001  0.0416
KEGG: NUCLEOTIDE EXCISION REPAIR  hsa03420  0.001  0.0422
GO: RESPONSE TO XENOBIOTIC STIMULUS  GO:0009410  0.001  0.0462

Please browse Full Text version to see the data of Supplementary Tables 6, 7.

**Supplementary Table 6.** Gene sets and pathways enriched in i-GSEA4GWAS.

**Supplementary Table 7.** Top gene sets and pathways enriched in PASCAL.