The Genetic Modifiers of Motor Onset Age (GeM MOA) website: genome-wide association analysis for genetic modifiers of Huntington's disease

Kevin Correia[a,&], **Denise Harold**[b,&,$,^], **Kyung-Hee Kim**[a,c], **Peter Holmans**[b,^], **Lesley Jones**[b,^], **Michael Orth**[d,^], **Richard H. Myers**[e,^], **Seung Kwak**[f,^], **Vanessa C. Wheeler**[a,c,^], **Marcy E. MacDonald**[a,c,g,^], **James F. Gusella**[a,g,h,^], and **Jong-Min Lee**[a,c,g,^,*]

[a]Molecular Neurogenetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA

[b]Medical Research Council (MRC) Centre for Neuropsychiatric Genetics and Genomics, Department of Psychological Medicine and Neurology, School of Medicine, Cardiff University, Cardiff, United Kingdom

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

[d]Department of Neurology, University of Ulm, Germany

[e]Department of Neurology and Genome Science Institute, Boston University School of Medicine, Boston, MA 02118, USA

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

[g]Medical and Population Genetics Program, the Broad Institute of M.I.T. and Harvard, Cambridge, MA02142, USA

[h]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Abstract

BACKGROUND—Huntington's disease (HD) is a dominantly inherited disease caused by a CAG expansion mutation in *HTT*. The age at onset of clinical symptoms is determined primarily by the length of this CAG expansion but is also influenced by other genetic and/or environmental factors.

OBJECTIVE—Recently, through genome-wide association studies (GWAS) aimed at discovering genetic modifiers, we identified loci associated with age at onset of motor signs that are significant at the genome-wide level. However, many additional HD modifiers may exist but may not have achieved statistical significance due to limited power.

---

[*]Corresponding author: 185 Cambridge Street, Boston, MA 02114, USA. Tel: 617-643-9714; Fax: 617-726-5735; jlee51@mgh.harvard.edu.
[&]Equal contribution
[$]Present address: Neuropsychiatric Genetics Research Group, Department of Psychiatry, Institute of Molecular Medicine, Trinity College Dublin, Ireland
[^]Founding GeM-HD Consortium investigators

Conflict of Interest

The authors have no conflict of interest to report.

METHODS—In order to disseminate broadly the entire GWAS results and make them available to complement alternative approaches, we have developed the internet website "GeM MOA" where genetic association results can be searched by gene name, SNP ID, or genomic coordinates of a region of interest.

RESULTS—Users of the Genetic Modifiers of Motor Onset Age (GeM MOA) site can therefore examine support for association between any gene region and age at onset of HD motor signs. GeM MOA's interactive interface also allows users to navigate the surrounding region and to obtain association p-values for individual SNPs.

CONCLUSIONS—Our website conveys a comprehensive view of the genetic landscape of modifiers of HD from the existing GWAS, and will provide the means to evaluate the potential influence of genes of interest on the onset of HD. GeM MOA is freely available at https://www.hdinhd.org/.

Introduction

Huntington's disease (HD) is a dominantly inherited neurodegenerative disorder [1, 2]. The root cause is a CAG trinucleotide repeat expansion in *HTT* [3], the gene encoding huntingtin. CAG repeats >35 units lead to progressive motor, cognitive, and psychiatric phenotypes, including characteristic choreic movements [4, 5]. The age at onset of clinical signs in HD is inversely correlated with the size of the CAG repeat [4, 6–9], establishing CAG repeat length as the most important determinant of the rate of pathogenesis leading to initial clinical manifestations. However, stringent statistical analysis using CAG repeat length as the primary predictor of age at onset has revealed additional unexplained variance in age at onset, which suggests the actions of other genetic and potentially environmental factors [7]. Recently, genome-wide association studies (GWAS) of this residual variance discovered two highly significant genetic loci associated with modification of age at onset of motor signs [10]. This demonstrates that HD can be modified prior to disease onset by other genes, supporting the heritability in residual age at onset of HD [11, 12] and pointing to the potential of genetic modifiers as therapeutic targets. However, as HD is far less frequent than many common diseases, the power of GWAS in this disorder is limited by sample size, suggesting that the application of a p-value threshold of genome-wide significance (p-value < 5E-8) may have left many *bona fide* HD genetic modifiers undetected in the first GWAS study based on ~4,000 HD subjects. The CAG repeat encodes a glutamine tract near the amino-terminus of huntingtin, a protein that participates in a wide variety of cellular functions. Consequently, lengthening of the glutamine stretch due to CAG expansion may produce subtle yet profoundly important effects on multiple huntingtin functions, exposing the potential for modification by diverse genes that might accelerate or delay pathogenesis. From our experience with initial HD modifier study, GWAS with yet larger sample sizes to increase the power to discover additional age at onset modifier variants represents the best unbiased means to comprehensively evaluate the impact of genetic variations on HD pathogenesis. However, one potential alternative, albeit biased, approach is the functional evaluation of candidate genetic modifiers using the plethora of model systems that the research community has developed. For example, promising candidate modifiers from genetic studies of HD subjects can be tested in model systems to validate these targets and

conversely, candidates from molecular studies and model systems can be cross-checked in the HD age at onset modifier GWAS results to assess their level of relevance in humans. Consequently, dissemination of the full HD modifier GWAS results [10] to the broader HD research community could greatly facilitate rapid discovery and validation of HD modifiers. While conventional publications are effective in drawing attention to the top GWAS hits, in this format it is more difficult to discern either the entire landscape or to determine the level of significance of variants in a given gene. Therefore, to facilitate alternative genetic-based approaches to HD modifiers, we have created a user-friendly internet website where investigators can obtain association results genome-wide for all genetic variations tested in the GeM-HD Consortium HD modifier GWAS [10].

Methods

Genome-wide association analysis

Full details of genetic analysis to identify modifiers of HD residual age at onset of motor signs were described elsewhere [7, 10]. Natural log-transformed age at onset of motor symptoms was modeled by CAG repeat length using only normally distributed data points to generate a phenotyping model [7]. Subsequently, this CAG-age at onset relationship was used to calculate predicted age at onset for HD subjects with CAG lengths 40–55. Predicted age at onset was subtracted from observed age at onset to generate residual age at onset phenotype for GWA analysis. SNPs were imputed using 1000 Genomes Project Europeans as a reference panel. Then, quantitative trait loci (QTL) GWA analysis using mixed-effect model linear regression analysis was applied to GWA1+2 and GWA3 data sets. The single SNP association analysis results of GWA1+2 and GWA3 were then meta-analyzed, generating the final association analysis results for the website [10]. Quality control-passed SNPs in all 3 GWAs (7,916,833 SNPs) were used to construct a database for the website.

Calculation of sample size

For each SNP, sample size that is required to achieve 80% power, a pre-specified significance level (e.g., p-value of 0.05, 0.00001, or 0.00000005), and effect size was calculated. Effect size of a SNP was based on corresponding SNP's R square value from the GWA analysis. A power calculation function 'pwr.f2.test' in the R package 'pwr' was implemented. Sample size estimation for the SNPs with SNP R square value smaller than 0.001 was not performed, and the website will display "null" for those SNPs.

Construction of the website

The design goal for the website was to provide a simple user-friendly way to visualize the GWA analysis results and to allow the user to interactively explore the data. Genome tracks were designed to render the GWA analysis data with annotation tracks that provide gene and recombination rate information, allowing users to view results or explore regions of interest along the human genome. GWA analysis results, along with genomic annotations from the UCSC hg19 assembly and recombination rates from the HapMap project were populated into a SQLite database for fast and efficient lookups. Server-side handling of the website was managed in Python. For the visualization of the results, the D3js (www.d3js.org) visualization library for JavaScript was used to provide a framework for an interactive

display. The D3js library provides a comprehensive framework to develop dynamic and interactive charts based on queries from the user. SNPs present in all 3 GWA datasets can be queried.

Results and Discussion

The HD modifier GWA analysis results [10] were used to generate a database that supports keyword searches. In addition, gene annotations (from the UCSC genome browser, hg19 genome build) and recombination rates (from HapMap project, http://hapmap.ncbi.nlm.nih.gov/) were incorporated to provide additional annotation information. In order to use the website, a user registers with an email address and name. The registered email address is then used to log in to the main search window. The website provides 3 main search functions: search by 1) a gene symbol, 2) a SNP ID, and 3) genomic coordinates on UCSC hg19 assembly (Figure 1A). Once a search keyword is found in the database, the website will display the association results, refseq genes, and recombination rates in the region on the top, middle, and bottom panel, respectively. When a SNP or a gene is used as a search term, the browser will display association analysis results of the SNP or gene plus 250 KB flanking regions on both sides by default. A maximum of 5 MB region can be specified in the region search query. In addition, a region of interest can be navigated as a unit of 100 KB or 1MB if surrounding regions need to be examined. An example of SNP query results is shown in Figure 1A.

In the association analysis results panel (genomic coordinate on X-axis and -log10(p-value) on Y-axis) (Figure 1B), each circle represents a SNP, and circle size is proportional to the GWA analysis significance level (i.e., the bigger the circle, the more significant the SNP). This panel is interactive, and more information concerning the SNP can be obtained by clicking the corresponding circle. If a SNP circle is clicked, then, the browser will change the color of the selected SNP, and show detailed information on a separate box (e.g., SNP ID, location of the SNP, alleles, frequencies in HD with European ancestry, effect size if selected SNP is genome-wide significant, rank in the association analysis results and p-value). The right side of the inset shows sample size estimations to achieve 80% power and a pre-specified significance level (i.e., 0.05, 0.00001, or 0.00000005). The association results plot panel (Figure 1B) also shows genome-wide significance level (p-value, 5E-8; purple line) and suggestive significance level (p-value, 1E-5; green line). The middle panel shows refseq genes in the region, with each rectangle showing a representative refseq transcript (Figure 1C). When multiple refseq transcripts represent the same gene in the UCSC refseq database, then the longest transcript and coding DNA sequence (CDS) are selected to show the approximate location of the gene. In the gene track plot, once a rectangle is clicked an additional information box displaying representative refseq transcript ID and genomic coordinates will appear. Lastly, the bottom plot panel shows recombination rates in the region based on HapMap data (Figure 1D). The recombination rate plot on the bottom is scaled based on the entire HapMap recombination rate data set, and therefore, allows approximation of rates in the region compared to the other regions with highest recombination rates. Also, the website provides a route to evaluate previously published association signals with an increased power. As an example, we compiled SNPs from some of the candidate studies, and compared published results with results in the GeM MOA.

None of the SNPs from candidate studies reached the level of significance for an unbiased genome-wide approach in GeM-HD Consortium modifier GWA analysis (Table 1), suggesting that the nominal significances from small sample-sized candidate-based studies might have been overestimated due to 1) lack of power, 2) outliers, 3) population stratification, and/or 4) lack of multiple test correction as noted previously [13].

Although our GWA analysis of the largest HD genetic cohort assembled to date detected three genome-wide significant modifier effects, additional modifier loci may remain undetected due to lack of sufficient power to achieve the genome-wide significance level (i.e., p-value of 5E-8). Consequently, loci displaying higher p-values may be worthy of further investigation, if supported by additional functional evidence. However, as most SNP association signals represent indirect association, rather than direct detection of the contributing functional variant, the location of the SNP candidate does not necessarily implicate the nearest gene. Depending on the local genetic structure, the physical distance between the SNP and the actual contributing functional variation can be large, if strong linkage disequilibrium exists [14], and its functional consequences may not in any event impact on the nearest gene, if it regulates the gene at a distance. Conversely, the absence of even suggestive p-values in the vicinity of a gene may not dictate a lack of potential for it to modify HD as this may instead reflect a lack of naturally occurring variations that both impact the gene's function and are sufficiently frequent to be informative in the GWA. Given these considerations, it is strongly advised that the results presented in the GeM MOA website be interpreted with care and rigor so as to effectively facilitate modifier discovery by the broader HD research community. This web site is to provide a user-friendly graphical interface to HD modifier GWA analysis summary results, and therefore, does not support data download. However, if original data are required, genotype data will be made available by sending inquiries to info@chdifoundation.org with the words "GWAS data" in the subject line.

In summary, we have constructed an internet website where association between a SNP genotype and the residual age at HD motor onset can be obtained genome-wide. This freely available internet research resource provides a comprehensive view of genetic interactions with the *HTT* CAG expansion mutation, permitting individual SNPs or gene regions to be judged for their potential to influence HD pathogenesis. The HD human genetic data presented in our website can therefore greatly enhance ongoing HD research community efforts to develop therapeutic interventions capable of slowing or halting HD pathogenesis.

Acknowledgments

References

1. Bates GP, Dorsey R, Gusella JF, Hayden MR, Kay C, et al. Huntington disease. Nature Reviews Disease Primers. 2015:15005.

2. Huntington G. On chorea. Med Surg Rep. 1872; 26:320–321.

3. HDCRG. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell. 1993; 72:971–983. [PubMed: 8458085]

4. Andrew SE, Goldberg YP, Kremer B, Telenius H, Theilmann J, et al. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nat Genet. 1993; 4:398–403. [PubMed: 8401589]

5. Schoenfeld M, Myers RH, Cupples LA, Berkman B, Sax DS, et al. Increased rate of suicide among patients with Huntington's disease. J Neurol Neurosurg Psychiatry. 1984; 47:1283–1287. [PubMed: 6239910]

6. Duyao M, Ambrose C, Myers R, Novelletto A, Persichetti F, et al. Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat Genet. 1993; 4:387–392. [PubMed: 8401587]

7. Lee JM, Ramos EM, Lee JH, Gillis T, Mysore JS, et al. CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. Neurology. 2012; 78:690–695. [PubMed: 22323755]

8. Persichetti F, Srinidhi J, Kanaley L, Ge P, Myers RH, et al. Huntington's disease CAG trinucleotide repeats in pathologically confirmed post-mortem brains. Neurobiol Dis. 1994; 1:159–166. [PubMed: 9173995]

9. Snell RG, MacMillan JC, Cheadle JP, Fenton I, Lazarou LP, et al. Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. Nat Genet. 1993; 4:393–397. [PubMed: 8401588]

10. Genetic Modifiers of Huntington's Disease Consortium Electronic address ghmhe, Genetic Modifiers of Huntington's Disease Ge MHDC. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. Cell. 2015; 162:516–526. [PubMed: 26232222]

11. Li JL, Hayden MR, Warby SC, Durr A, Morrison PJ, et al. Genome-wide significance for a modifier of age at neurological onset in Huntington's disease at 6q23-24: the HD MAPS study. BMC Med Genet. 2006; 7:71. [PubMed: 16914060]

12. Wexler NS, Lorimer J, Porter J, Gomez F, Moskowitz C, et al. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. Proc Natl Acad Sci U S A. 2004; 101:3498–3503. [PubMed: 14993615]

13. Gusella JF, MacDonald ME, Lee JM. Genetic modifiers of Huntington's disease. Mov Disord. 2014; 29:1359–1365. [PubMed: 25154728]

14. Lee JM, Gillis T, Mysore JS, Ramos EM, Myers RH, et al. Common SNP-based haplotype analysis of the 4p16.3 Huntington disease gene region. Am J Hum Genet. 2012; 90:434–444. [PubMed: 22387017]
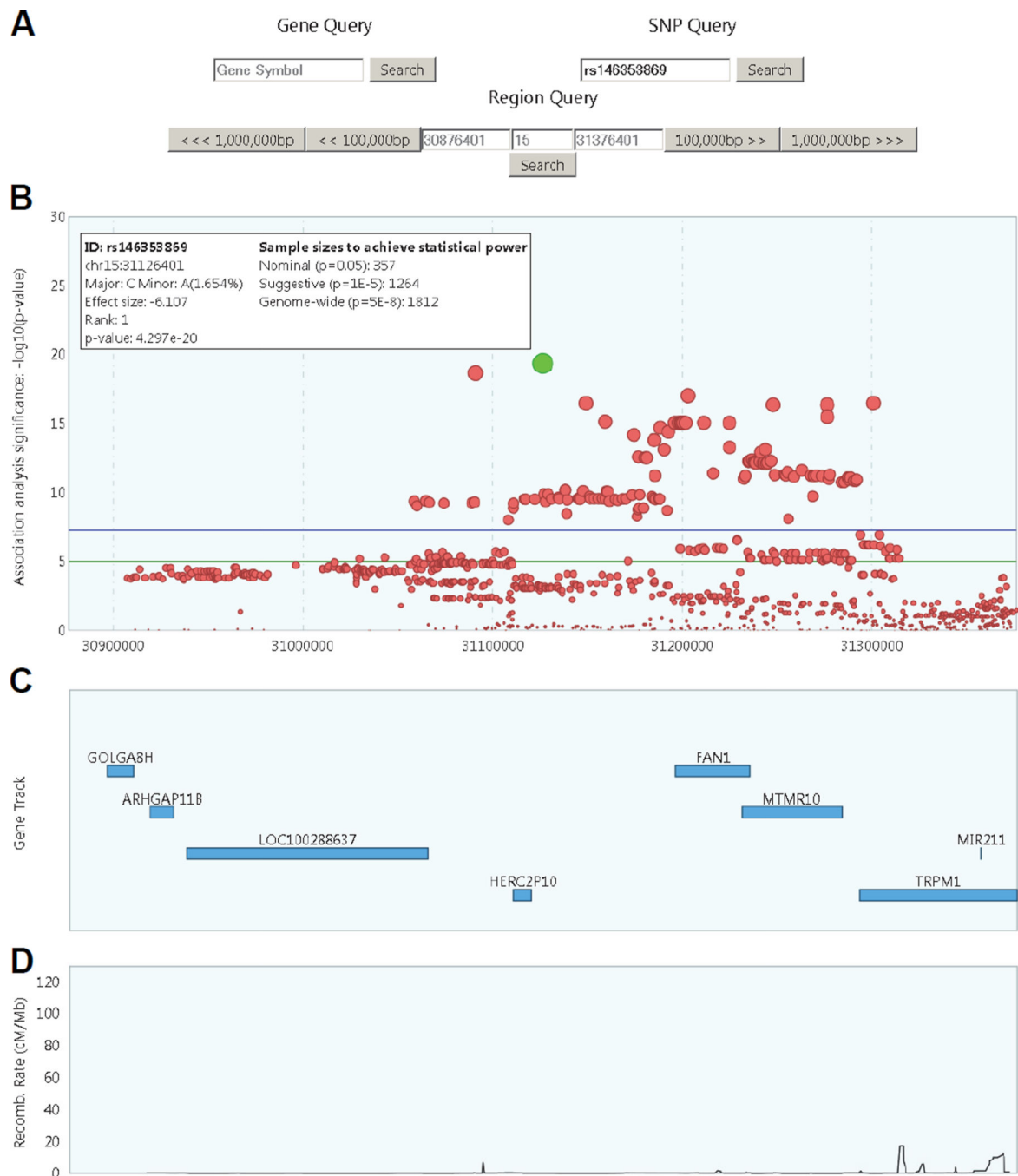
Figure 1. An example of SNP query results in the GeM MOA website

A) Three search functions can be used to obtain association analysis results. These include gene, SNP, and region query. Genomic coordinates based on UCSC hg19 assembly can be searched in region query function.

B) Association analysis results by querying the most significant SNP, rs146353869 (green circle), are shown Left side of the inset displays annotation information of the chosen SNP, alleles, frequencies in HD samples, rank of association, and association p-value. Right side

of inset shows sample sizes required to achieve 80% power and either nominal (0.05), suggestive (0.00001), or genome-wide significance (0.00000005).

C) Approximate locations of refseq genes in the selected region are shown. Clicking a rectangle representing a refseq will generate an inset in the panel, showing refseq ID, gene symbol, and coordinates.

D) Recombination rates of the region were based on HapMap data.

Table 1

Examples of previously published candidate HD modifier SNPs.

Gene, SNP	N, population (PMID)	SNP p-value	GeM-HD Consortium p-value
ADORA2A rs5751876	791, Huntington French Speaking Network (19591938)	0.019 [a]	0.2009
	419, Germans (20512606)	0.032 [a]	
ATG7 rs36117895	952, Germans, Italians, and others (20697744)	0.005	0.09299
BDNF rs6265	122 Spanish (16186551)	< 0.001	0.7388
GRIK2 rs2782901	253 Southern or Northern Europeans (26148071)	0.018 [c]	0.08693
GRIN2A rs1969060	167, Germans (15742215)	0.039	0.2558
	421, Venezuelan Kindreds (17018562)	0.04	
GRIN2A rs8049651	253 Southern or Northern Europeans (26148071)	0.015 [b]	0.5217
GRIN2B rs1806201	167 Germans (15742215)	0.001	0.04216
GRIN2B rs890	167 Germans (15742215)	0.030	0.02996
GRIN2B rs10744030	253 Southern or Northern Europeans (26148071)	0.044 [c]	0.2622
GRIN2B rs4764011	253 Southern or Northern Europeans (26148071)	0.036 [c]	0.5601
HAP1 rs4523977	878 (18192679)	0.015	0.3887
HIP1 rs2240133	253 Southern or Northern Europeans (26148071)	0.022 [b]	0.3568
LINC01559 rs10845757	253 Southern or Northern Europeans (26148071)	0.048 [b]	0.8947
NPY2R rs2234759	472, Germans (24121255)	0.0004 [b]	0.2498
PPARGC1A rs7665116	>400, Germans (19200361)	0.025 [b]	0.7712
	449, Registry (19133136)	0.0016 [b]	
	854, Germans and Italians (21211002)	0.0389 [b]	
PPARGC1A rs3736265	869, Male Registry (24383721)	0.0164 [b]	0.9202
UCHL1 rs5030732	946, Caucasians (16369839)	0.008	0.8468

Association analysis results of SNPs from some previous candidate studies are compared to results in the GeM MOA website. A statistical model is indicated next to the p-value by superscript if the model is clearly specified in the original publication. N represents sample number.

[a] Comparison of homozygotes

[b] Additive model

[c] Recessive model