Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study

Davina J Hensman Moss, Antonio F Pardiñas*, Douglas Langbehn, Kitty Lo, Blair R Leavitt, Raymund Roos, Alexandra Durr, Simon Mead, the TRACK-HD investigators†, the REGISTRY investigators†, Peter Holmans, Lesley Jones§, Sarah J Tabrizi§*

Summary

Background Huntington's disease is caused by a CAG repeat expansion in the huntingtin gene, *HTT*. Age at onset has been used as a quantitative phenotype in genetic analysis looking for Huntington's disease modifiers, but is hard to define and not always available. Therefore, we aimed to generate a novel measure of disease progression and to identify genetic markers associated with this progression measure.

Methods We generated a progression score on the basis of principal component analysis of prospectively acquired longitudinal changes in motor, cognitive, and imaging measures in the 218 indivduals in the TRACK-HD cohort of Huntington's disease gene mutation carriers (data collected 2008–11). We generated a parallel progression score using data from 1773 previously genotyped participants from the European Huntington's Disease Network REGISTRY study of Huntington's disease mutation carriers (data collected 2003–13). We did a genome-wide association analyses in terms of progression for 216 TRACK-HD participants and 1773 REGISTRY participants, then a meta-analysis of these results was undertaken.

Findings Longitudinal motor, cognitive, and imaging scores were correlated with each other in TRACK-HD participants, justifying use of a single, cross-domain measure of disease progression in both studies. The TRACK-HD and REGISTRY progression measures were correlated with each other (r=0·674), and with age at onset (TRACK-HD, r=0·315; REGISTRY, r=0·234). The meta-analysis of progression in TRACK-HD and REGISTRY gave a genome-wide significant signal ($p=1·12\times10^{-10}$) on chromosome 5 spanning three genes: *MSH3*, *DHFR*, and *MTRNR2L2*. The genes in this locus were associated with progression in TRACK-HD (*MSH3* $p=2·94\times10^{-8}$, *DHFR* $p=8·37\times10^{-7}$, *MTRNR2L2* $p=2·15\times10^{-9}$) and to a lesser extent in REGISTRY (*MSH3* $p=9·36\times10^{-4}$, *DHFR* $p=8·45\times10^{-4}$, *MTRNR2L2* $p=1·20\times10^{-3}$). The lead single nucleotide polymorphism (SNP) in TRACK-HD (rs557874766) was genome-wide significant in the meta-analysis ($p=1·58\times10^{-8}$), and encodes an aminoacid change (Pro67Ala) in MSH3. In TRACK-HD, each copy of the minor allele at this SNP was associated with a 0·4 units per year (95% CI 0·16–0·66) reduction in the rate of change of the Unified Huntington's Disease Rating Scale (UHDRS) Total Motor Score, and a reduction of 0·12 units per year (95% CI 0·06–0·18) in the rate of change of UHDRS Total Functional Capacity score. These associations remained significant after adjusting for age of onset.

Interpretation The multidomain progression measure in TRACK-HD was associated with a functional variant that was genome-wide significant in our meta-analysis. The association in only 216 participants implies that the progression measure is a sensitive reflection of disease burden, that the effect size at this locus is large, or both. Knockout of *Msh3* reduces somatic expansion in Huntington's disease mouse models, suggesting this mechanism as an area for future therapeutic investigation.

Funding The European Commission FP7 NeurOmics project; CHDI Foundation; the Medical Research Council UK; the Brain Research Trust; and the Guarantors of Brain.

Introduction

Huntington's disease is an autosomal dominant fatal neurodegenerative condition caused by a CAG repeat expansion in huntingtin gene, *HTT*.[1] It is a movement, cognitive, and psychiatric disorder, but symptoms, age of disease onset, and disease progression vary.[2] Age of onset reflects the trajectory of disease pathology up to the point of motor onset.[1,3] However, the transition from premanifest to manifest Huntington's disease is gradual,[4,5] making clinical definition challenging. Furthermore, psychiatric and cognitive changes might not be concurrent with motor onset.[6] Despite this imprecision in defining onset, the inverse correlation of *HTT* CAG repeat length and age at motor onset accounts for 50–70% of the observed variance in onset.[7] Part of the remaining difference in age of onset was also recently shown to be genetically encoded, and genes of the DNA damage response were identified as being likely to modify onset of Huntington's disease.[8]

The need for clinical trials close to disease onset has motivated a number of observational studies.[5,9,10] These new data provide the opportunity to investigate the association between onset and progression and whether

Lancet Neurol 2017

Published Online
June 19, 2017
http://dx.doi.org/10.1016/S1474-4422(17)30161-8

See Online/Comment
http://dx.doi.org/10.1016/S1474-4422(17)30179-5

*Joint first authors

†Investigators listed in the appendix

§Joint senior authors

UCL Huntington's Disease Centre (D J Hensman Moss MBBS, Prof S J Tabrizi PhD) **and MRC Prion Unit** (Prof S Mead PhD), **UCL Institute of Neurology, Department of Neurodegenerative Disease and UCL Genetics Institute, Division of Biosciences** (K Lo PhD), **University College London, London, UK; MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK** (A F Pardiñas PhD, Prof P Holmans PhD, Prof L Jones PhD); **Carver College of Medicine, Department of Psychiatry and Biostatistics, University of Iowa, Iowa City, IA, USA** (Prof D Langbehn PhD); **Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada** (Prof B R Leavitt MD); **Department of Neurology, Leiden University Medical Centre, Leiden, Netherlands** (Prof R Roos MD); **ICM, Inserm U 1127, CNRS UMR 7225, UPMC Univ Paris 06 UMR S 1127, Sorbonne Universités, Paris, France** (Prof A Durr MD); **and Department of Genetics, Pitié-Salpêtrière University Hospital, Paris, France** (Prof A Durr)

Correspondence to:
Prof Sarah J Tabrizi, UCL Huntington's Disease Centre, UCL Institute of Neurology, Department of Neurodegenerative Disease, University College London, London WC1N 3BG, UK
s.tabrizi@ucl.ac.uk

or

Prof Lesley Jones, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff CF24 4HQ, UK
jonesl1@cardiff.ac.uk

Research in context

Evidence before this study

Huntington's disease is caused by a tract of 36 or more CAG repeats in exon 1 of the huntingtin gene, *HTT*. Genetic modifiers of age at motor onset have been identified that highlight pathways which, if modulated in people, might delay Huntington's disease onset. Onset of disease is preceded by a long prodromal phase accompanied by substantial brain cell death; age at motor onset is difficult to assess accurately and is not available in disease-free at-risk individuals. We searched PubMed, for English language articles published until Oct 31, 2016, with the search terms "Huntington* disease" AND "genetic modifier" AND "onset", which identified 13 studies. We then searched for "Huntington* disease" AND "genetic modifier" AND "progression", which identified one review article. Among the 13 studies of genetic modification of Huntington's disease onset, most were small candidate gene studies; these were superseded by the one large study of genome-wide genetic modifiers of Huntington's disease, which identified three genome-wide significant loci, one on chromosome 8 and two on chromosome 15, these are thought likely to be associated with *RRM2B* and *FAN1*, respectively. This study also implicated DNA handling in Huntington's disease modification.

Added value of this study

We examined the prospective data from TRACK-HD and developed a measure of disease progression that reflected correlated progression in the brain imaging, motor, and cognitive symptom domains. We used the disease progression measure as a quantitative variable in a genome-wide association study and detected a locus on chromosome 5 containing three significant genes, *MTRNR2L2*, *MSH3*, and *DHFR*. The index variant encodes an aminoacid change in MSH3. We replicated this finding by generating a parallel progression measure in the less intensively phenotyped REGISTRY study and detected a similar signal on chromosome 5 that is probably attributable to the same variants. A meta-analysis of the two studies strengthened the associations. The progression measures and age of onset were correlated, but this was not responsible for the genetic association with disease progression. We also detected a signal on chromosome 15 in the REGISTRY study at the locus previously associated with age of onset.

Implications of all the available evidence

The progression measures used in this study can be generated in asymptomatic and symptomatic participants using a subset of the clinically relevant parameters gathered in TRACK-HD. We used these measures to identify genetic modifiers of disease progression in Huntington's disease. We identified a signal in only 216 participants, which was replicated in a larger sample and strengthened in the meta-analysis, reducing the chance of it being a false positive. This finding argues for the power of improving phenotypic measures in genetic studies and implies that this locus has a large effect on disease progression. The index associated genetic variant in TRACK-HD encodes a Pro67Ala change in MSH3, which implicates *MSH3* as the associated gene on chromosome 5. Altering levels of Msh3 in Huntington's disease mouse models reduces somatic instability and crossing *Msh3* null mice with Huntington's disease mouse models prevents somatic instability of the *HTT* CAG repeat and reduces pathological phenotypes. Polymorphisms in *MSH3* have been linked to somatic instability in patients with myotonic dystrophy type 1. MSH3 is a non-essential neuronally expressed member of the DNA mismatch repair pathway and these data reinforce its candidacy as a therapeutic target in Huntington's disease and potentially in other neurodegenerative expanded repeat disorders.

they are influenced by the same biology, and permit the study of individuals before clinical onset. TRACK-HD[5,6] was a prospective, observational, biomarker study that represents the most deeply phenotyped cohort of people with premanifest and symptomatic Huntington's disease, with annual visits involving motor, cognitive, psychiatric and imaging assessments. We used TRACK-HD data[5,6] to generate a novel unified Huntington's disease progression measure for use in a genetic association analysis. We developed a similar measure in participants from the REGISTRY study[9] to replicate our findings. We used these disease progression measures as quantitative variables in genome-wide association analyses of the TRACK-HD and REGISTRY data, and aimed to replicate this finding in a meta-analysis.

See Online for appendix

Methods

Study design and participants

In this genome-wide association study, we examined prospective data from TRACK-HD to develop a measure of disease progression reflecting correlated progression on brain imaging, motor, and cognitive symptom domains. We used this disease progression measure as a quantitative variable in a genome-wide association study to identify associated genetic loci, and aimed to replicate this finding by generating a parallel progression measure in the less intensively phenotyped REGISTRY study.

TRACK-HD[5,6] was a prospective, observational study collecting deep phenotypic data, including imaging, quantitative motor, and cognitive assessments, from adults with early Huntington's disease, premanifest Huntington's disease gene carriers, and controls (figure 1). It provides annually collected multivariate data for 3 years (2008–11), with 243 participants at baseline.[6] Demographic details of these individuals are in the appendix.

REGISTRY[9] was a multisite, prospective, observational study, which collected phenotypic data (2003–13) for more than 13 000 participants, mostly Huntington's disease gene carriers with manifest disease (figure 1).

The study aimed for annual assessments (every 9–15 months), although in practice assessment dates were variable. The core data included age, CAG repeat length, Unified Huntington's Disease Rating Scale (UHDRS) total motor score (TMS), and total functional capacity (TFC); some patients had further assessments, such as a cognitive battery.[9] We included 1835 adult participants from REGISTRY in this study on the basis of available genotype data.[8] We obtained TMS, symbol digit modality, verbal fluency, Stroop colour reading, word reading and interference measures, functional assessment score, and TFC.

All experiments were done in accordance with the Declaration of Helsinki and approved by the University College London (UCL)/UCL Hospitals Joint Research Ethics Committee; ethical approval for the REGISTRY analysis is outlined by the Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium.[8] Peripheral blood samples were donated by genetically confirmed Huntington's disease gene carriers, and all participants provided informed written consent before study entry.

Procedures

For both studies, we derived atypical severity scores with a combination of principal component analysis (PCA) and regression of the predictable effects of the primary gene *HTT* CAG repeat length. Details differed, however, owing to differences in nature of the two datasets. In TRACK-HD, 24 variables were used to stratify the cohort in terms of disease progression (appendix). These variables were divided a priori into three broad domains: brain volume measures, cognitive variables, and quantitative motor variables. For each variable, the input for analysis was the participant's random longitudinal slope from a mixed effects regression model with correlated random intercepts and slopes for each participant. This model regressed the observed values on a clinical probability of onset statistic derived from CAG repeat length and age, and its interaction with follow-up length. The participants' random slope estimates thus provided a measure of atypical longitudinal change not predicted by age and CAG repeat length. We then used PCA of the random slopes to study the dimensionality of the longitudinal changes corrected for age and CAG repeat length. Further detail about the methods, including control for potential demographic confounders, is in the appendix and figure 1.

For REGISTRY, by contrast with TRACK-HD, follow-up length and frequency were variable and missing

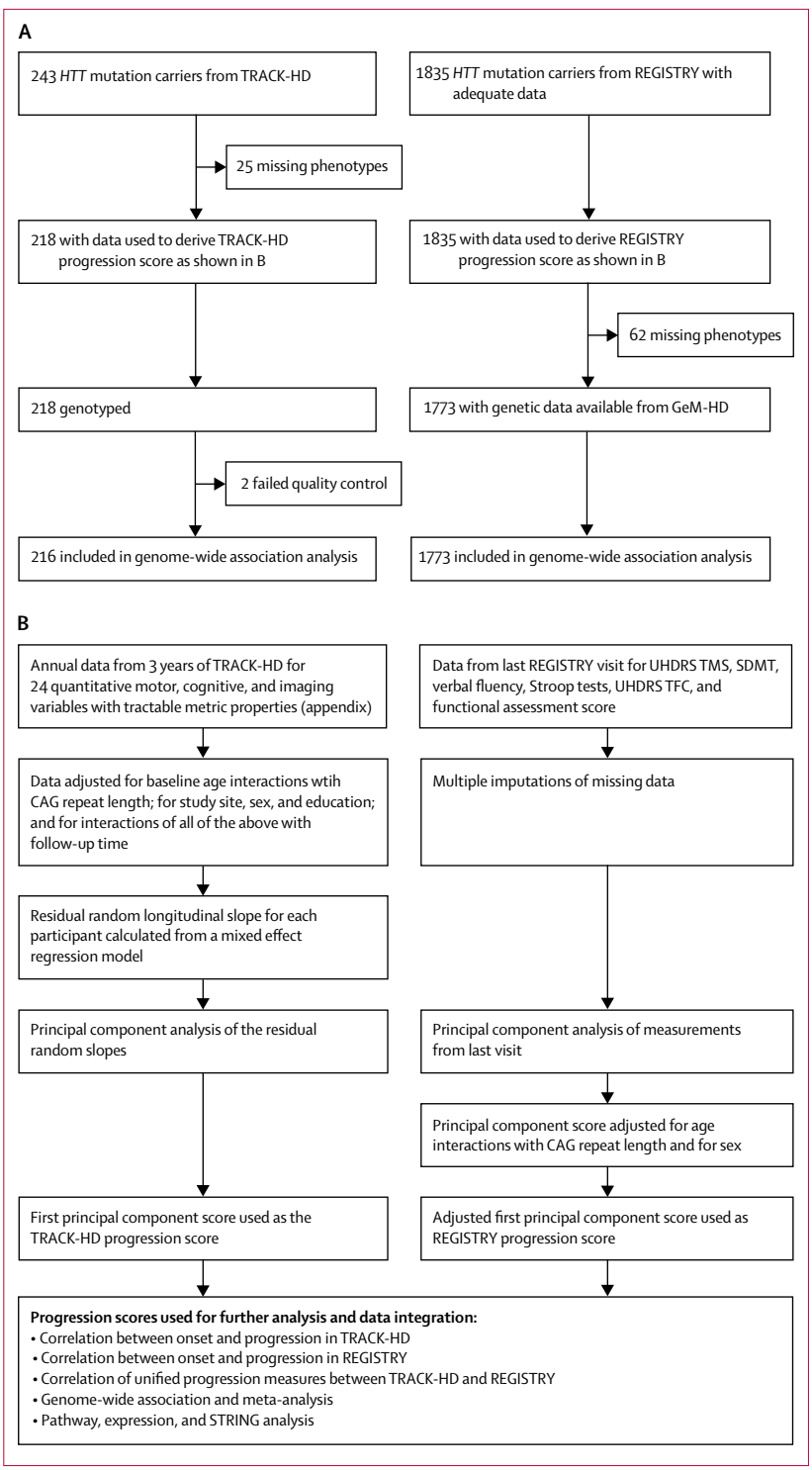

Figure 1: **TRACK-HD and EHDN REGISTRY trial designs**
(A) Numbers of participants in each part of the study from TRACK-HD (left) and REGISTRY (right). (B) Derivation and use of the progression scores from TRACK-HD (left) and REGISTRY (right). After establishing that brain imaging, quantitative motor, and cognitive variables are correlated and follow a similar trajectory, we scored the TRACK-HD participants using the first principal component as a unified progression measure, and used this measure to look for genome-wide associations with Huntington's disease progression. We replicated our findings in the EHDN REGISTRY participants by looking at how far their disease had progressed compared with expectations based on CAG repeat length or age, and used this progression measure to look for genome-wide associations. EHDN=European Huntington's Disease Network. GEM-HD= Genetic Modifiers of Huntington's Disease. SDMT=symbol digit modality test. TFC=Total Functional Capacity. UHDRS TMS=Unified Huntington's Disease Rating Scale Total Motor Score.

data were substantial, making longitudinal progression analysis problematic. We therefore examined cross-sectional status at last visit, using a single unified motor-cognitive dimension of severity. We did multiple imputation to fill in missing data, derived PCA severity scores and regressed off the predictive effect of age, CAG length, and sex on the PCA severity scores derived from these data to obtain the measure of atypical severity at the last visit. This gave a single point severity score based on how advanced a participant was compared with expectations based on their CAG repeat length and age. 1773 participants had adequate phenotypic data to score; further detail is in the appendix and figure 1.

Statistical and genetic analysis

We did data analyses using SAS/STAT version 14.0 and 14.1, mainly via the MIXED, FACTOR, and GML procedures. We occasionally used a log or inverse transform of a measure, with the goal to improve approximate normality of the distribution and avoid inappropriate influence of extreme scores.

We genotyped 218 TRACK-HD study participants with complete serial phenotype data on Illumina Omni2.5v1.1 arrays, and we did quality control measures as described in the appendix. We carried out imputation using the 1000 Genomes phase 3 data as a reference (appendix). This yielded 9·65 million biallelic markers of 216 individuals. We obtained genotypes for the REGISTRY participants from the GeM-HD Consortium (for details of their genotyping, quality control, curation, and imputation see reference 8).

We did association analyses with the mixed linear model (MLM) functions in Genome-wide Complex Trait Analysis (GCTA) version 1.26.[11] We did conditional analyses using the COJO procedure in GCTA. Because of the relatively small sample sizes, we restricted analyses to single nucleotide polymorphism (SNPs) with minor allele frequency of more than 1%. We did a meta-analysis of the TRACK-HD and REGISTRY associations using METAL.[12] To test whether the association signals in TRACK-HD and REGISTRY could have arisen from the same causal SNPs, and whether these also influenced expression, we did colocalisation analysis using GWAS-pw version 0.21.[13] We calculated gene-wide p values using Multiobjective Analyzer for Genetic Marker Acquisition (MAGMA) version 1.05, a powerful alternative to SNP-based analyses, which aggregates the association signal inside genes while taking linkage disequilibrium between SNPs into account,[14] using a window of 35 kb upstream and 10 kb downstream of genes.[15] Such an analysis can increase power over single SNP analysis when there are multiple causal SNPs in a gene, or when the causal SNP is not typed and its signal is partially captured by multiple typed SNPs in linkage disequilibrium with it. To maximise comparability with the GeM-HD genome-wide association study (GWAS), our primary pathway analyses used Setscreen,[16] which sums the log p values of all SNPs in a pathway, also correcting for linkage disequilibrium between SNPs.

All methods and analyses are described in more detail in the appendix.

Role of the funding source

The funders of this study and of the TRACK-HD and REGISTRY studies had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

In the analysis of variables from TRACK-HD, we did individual principal compenent analyses of each domain and found that the first principal component scores were highly correlated between the domains ($p<0·0001$ in all cases, appendix). We observed no phenotypic subtypes of symptom clusters in motor, cognitive, or imaging domains; rather, longitudinal change in TRACK-HD not predictable by CAG repeat length and age was distributed on a correlated continuum (figure 2). We therefore repeated PCA of the measures combined across all domains. The first principal component of this combined analysis accounted for 23·4% of the joint variance, and was at least moderately correlated ($r>0·4$) with most of the variables that contributed heavily to each domain-specific first principal component (appendix). The first psychiatric principal component had notably lower correlations with motor and cognitive domains and clinical probability of onset than the inter-correlations seen among these three measures, so was excluded from our progression measures.

The cross-domain first principal component was used as a unified Huntington's disease progression measure in the TRACK-HD cohort (figures 1, 2). To confirm that our progression measure correlated with commonly recognised measures of Huntington's disease severity not included in the progression analysis, we examined the residual change relationships between the progression score and UHDRS TMS change and TFC change after controlling for the clinical probability of onset statistic. We found a correlation of $r=0·448$ ($p<0·0001$) for the residual motor slope and $r=-0·421$ ($p<0·0001$) for the residual TFC slope. One unit increase in unified Huntington's disease progression measure corresponded to an increase of 0·71 units per year (95% CI 0·34–1·08) in the rate of change of TMS, and an increase of approximately 0·2 units per year (0·12–0·30) in the rate of change of TFC. The 15 fastest progressing participants in TRACK-HD showed a mean annual rate of decline in the UHDRS TMS of 2·52 more points per year than predicted by age and CAG length (SD 2·47, SEM 0·64); the 15 slowest progressing

participants had a mean annual TMS decline of 0·45 points less per year than would be expected (SD 1·85, SEM 0·48).

Participants in the early stages of Huntington's disease had significantly faster progression on the unified Huntington's disease progression measure than did those still in the premanifest phase (p<0·0001). Among the 96 participants who had experienced onset, the age of onset as defined by a rater showed the expected association with predicted age of onset based on CAG length (appendix), and earlier than predicted age of onset was correlated with faster progression on our unified Huntington's disease progression measure (r=0·315; p=0·002).

The unified Huntington's disease progression measure developed in TRACK-HD could not be transferred directly to REGISTRY participants, for whom more limited data were available. Individual clinical measures in REGISTRY showed correlations across the motor, cognitive, and functional domains, consistent with our finding in TRACK-HD (appendix). The first principal component accounted for 75·6% of the variance in severity; no other principal components explained any substantial amount of the common variance within the measures used (appendix). Therefore this first principal component was chosen as a measure of severity in the REGISTRY cohort (figure 2). Higher values of this measure mean greater severity than expected at a given time; we infer that this is the result of faster progression (figure 2), and we used this measure as the unified REGISTRY progression measure. This progression measure and earlier than predicted age of onset were modestly correlated (r=0·234; p<0·0001; appendix). Atypically, rapidly or slowly progressing participants tended to become more atypical over time: given that correlation between time since disease onset and REGISTRY progression (r=–0·307; p<0·0001) was greater than that between age of onset and REGISTRY progression.

In TRACK-HD, the last-visit severity scores had a correlation of r=0·674 with the previously calculated longitudinal unified progression measure, indicating that our progression measures for TRACK-HD and REGISTRY reflected strongly, although not perfectly, related elements of clinical phenotype. Further support for this conclusion was given by the correlation of r=0·631 between the TRACK-HD and REGISTRY progression measures in the 14 participants present in both studies.

A genome-wide association analysis using the unified TRACK-HD progression measure as a quantitative trait yielded a significantly associated locus on chromosome 5 spanning *DHFR*, *MSH3*, and *MTRNR2L2*. The index SNP rs557874766 is a coding missense variant in *MSH3* (p=$5·8\times10^{-8}$; allele frequency G=0·2179/1091 [1000 Genomes]; figure 3 and appendix). Analyses conditioning on this SNP failed to show evidence for a second independent signal in this region in TRACK-HD (appendix). The genes in this locus were the only ones to reach the commonly accepted genome-wide significance criterion (p<$2·5\times10^{-8}$)[17] for gene-wide tests in the

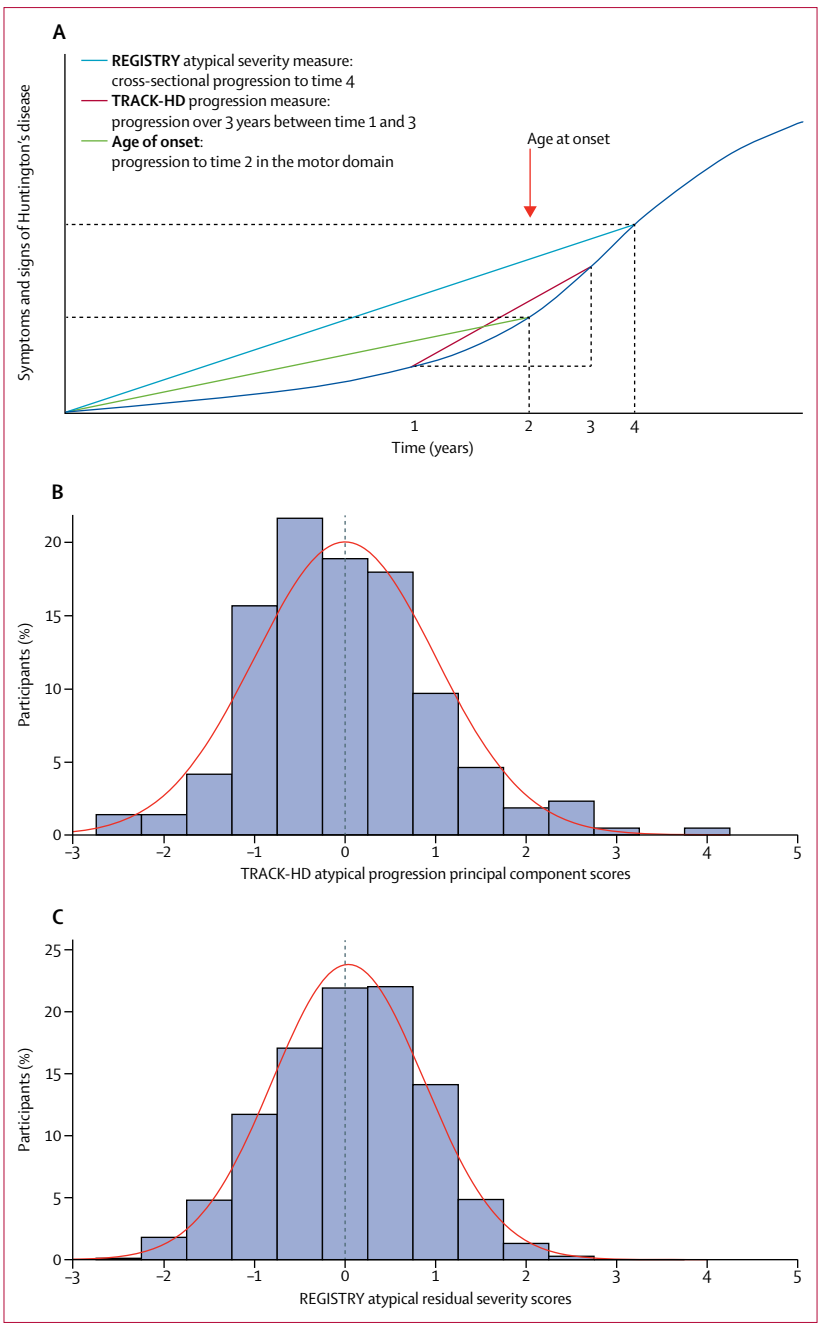

***Figure 2*: Assessment of progression in Huntington's disease**
(A) Trajectory of symptoms and signs. The TRACK-HD progression score uses longitudinal data over 3 years. Given limited longitudinal data in REGISTRY, cross-sectional severity at last visit compared with predicted severity was used as a proxy for progression. Age at onset occurs when a participant has unequivocal motor signs of Huntington's disease. 1–4 indicate points in time. (B) Distribution of the progression measure in 218 participants from the TRACK-HD cohort. (C) Distribution of atypical severity (compared with predicted severity at final visit) in 1835 members of the REGISTRY cohort. The curves in (B) and (C) are the normal distribution approximations of the severity score distributions, and absolute numbers of participants are given in the appendix.

For **the genome-wide significance criterion** see http://hdresearch.ucl.ac.uk/data-resources

MAGMA analysis significance (*MTRNR2L2* p=2·15×10^{-9}; *MSH3* p=2·94×10^{-8}; and *DHFR* p=8·37×10^{-7}).[14,18]

A genome-wide association analysis of data from REGISTRY using the REGISTRY unified progression measure replicated the signal identified in TRACK-HD (lead SNP rs420522, p=1·39×10^{-5}) on a narrower locus (chr5: 79902336–79950781), but still tagging the same three genes (figure 3). No genes reached genome-wide

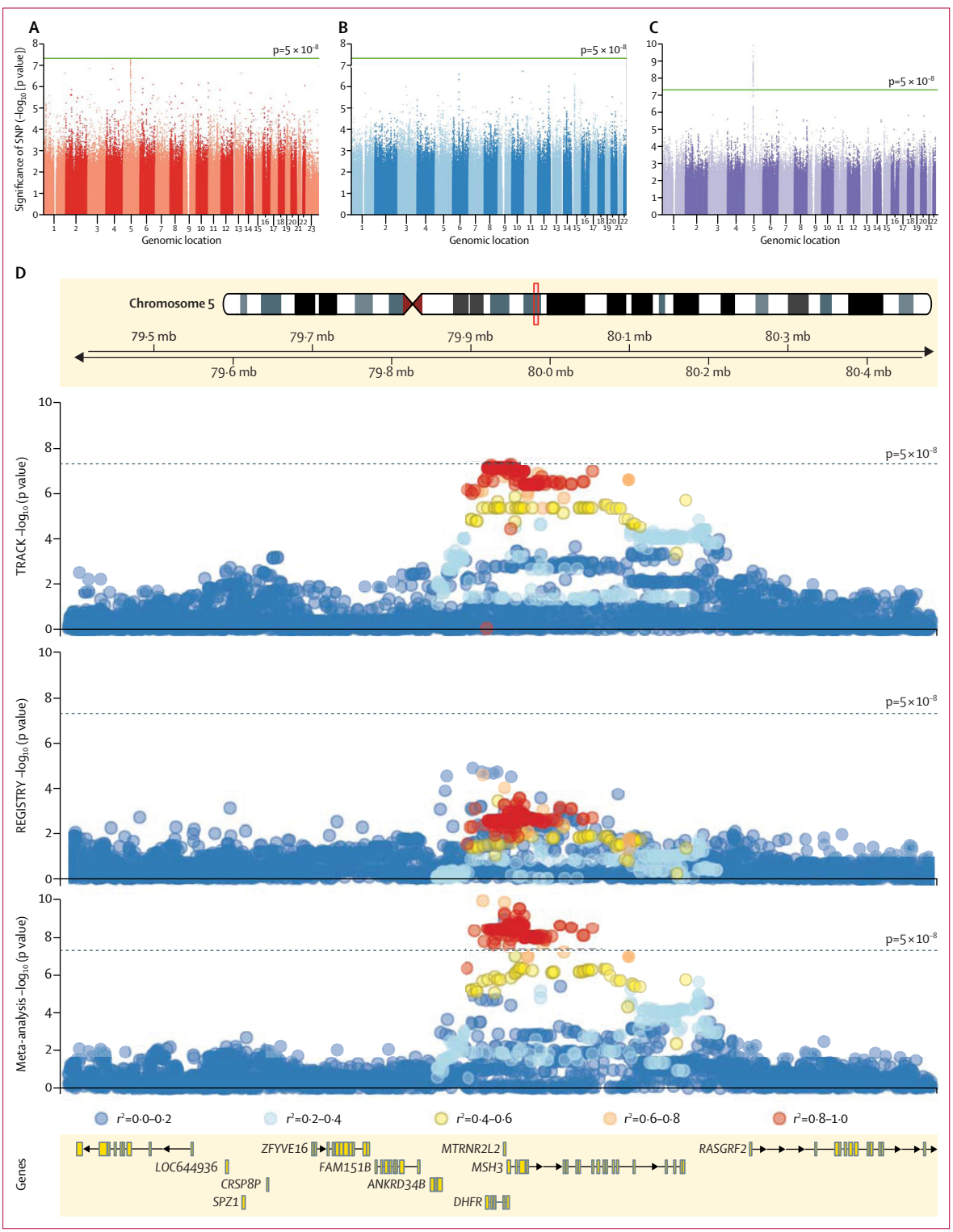

***Figure 3*: Genome-wide association analysis of progression score**
(A) Manhattan plot of TRACK-HD genome-wide association analysis yielding a locus on chromosome 5. (B) Manhattan plot of REGISTRY genome-wide association analysis showing suggestive trails on chromosome 15 in the same locus that was significant in the GeM genome-wide association study,[8] and chromosome 5 in the same area as the TRACK-HD progression genome-wide association analysis. (C) Manhattan plot of meta-analysis of the TRACK-HD and REGISTRY progression analysis. (D) Locus zoom plot of the TRACK-HD (top), REGISTRY (middle), and meta-analysis (bottom) data showing the structure of linkage disequilibrium and $-\log_{10}$ (p value) of the significant locus on chromosome 5. The top image shows the chromosome; the red square shows the region that is zoomed in on in the other panels. The colours of the circles are based on r^2 with the lead SNP in TRACK-HD as shown in the bottom of the plot; intensity of colour reflects multiple overlying SNPs. Dashed lines: p=5×10^{-8}. SNP=single nucleotide polymorphism.

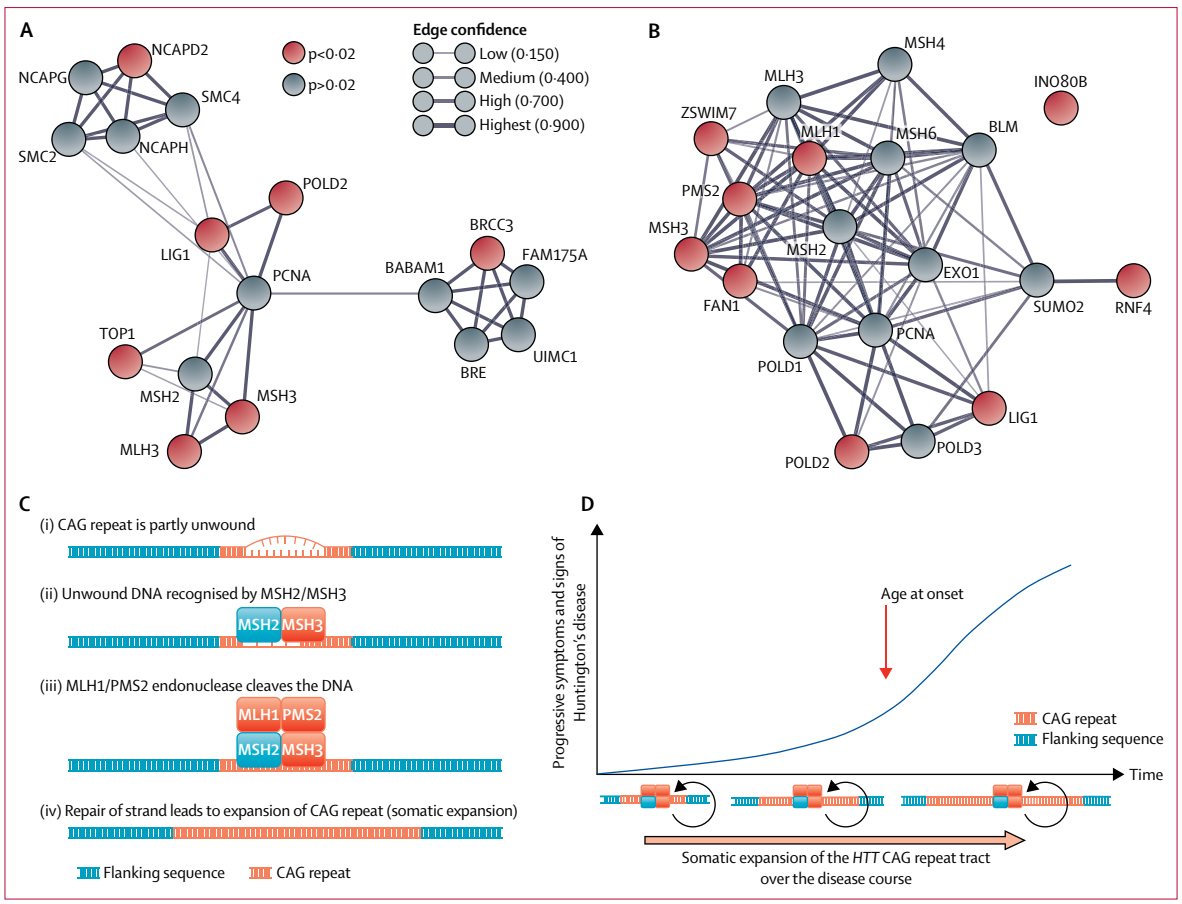

***Figure 4*: Functional linkage and possible mechanism of action of the *HTT* CAG repeat tract**

STRING diagram showing in red all proteins from the Pearl and colleagues' dataset[19] with gene-wide p<0·02 for association with Huntington's disease progression in (A) the TRACK-HD dataset and (B) the meta-analysis of TRACK-HD and REGISTRY. Ten genes which interact with these genes are shown in grey.[20] (C) How DNA mismatch repair proteins might be involved in somatic expansion of the CAG tract. Proteins with p<0·01 in the meta-analysed progression genome-wide association analysis are coloured red. (i) The CAG repeat DNA is partly unwound by lesions, constraints of the CAG tract structure, or by transcription. (ii) This unwound DNA is recognised by MutSβ (a complex of MSH2 and MSH3), (iii) which recruits the endonuclease MutLα (a complex of PMS2 and MLH1) and cleaves the DNA. (iv) Repair of the strand break leads to expansion of the CAG repeat. In neurons of the striatum, somatic expansion occurs throughout life and variants in MSH3 might promote or inhibit repeat recognition, binding, or repair. (D) Potential link between degree of somatic expansion during a patient's lifespan and rate of Huntington's disease progression.

significance, although there was evidence of association (at *DHFR* ($p=8{\cdot}45\times10^{-4}$), *MSH3* ($p=9{\cdot}36\times10^{-4}$), and *MTRNR2L2* ($p=1{\cdot}20\times10^{-3}$).

The meta-analysis of TRACK-HD and REGISTRY strengthened the signal of both individual SNPs in this region, encompassing the first three exons of *MSH3* along with *DHFR* and *MTRNR2L2* (figure 4, appendix), and also gene-wide associations over *MSH3*, *DHFR*, and *MTRNR2L2* in the MAGMA analysis. The most significant SNP in the meta-analysis was rs1232027, which was genome-wide significant ($p=1{\cdot}12\times10^{-10}$); the p value of rs557874766 is $1{\cdot}58\times10^{-8}$. No other regions attained genome-wide significance. rs557874766 was nominally significant in REGISTRY (p=0·010), with a direction of effect consistent with that in TRACK-HD. Analyses conditional on rs1232027 largely removed the association in this region (appendix), suggesting that there is only one signal. Conditioning on rs557874766 had a similar effect (appendix), so this SNP remains a plausible causal variant.

As suggested by the meta-analysis, colocalisation analyses between TRACK-HD and REGISTRY showed that this locus was probably influenced by the same SNPs in both studies (posterior probability 74·33%), although conditioning REGISTRY on rs55787466 did not remove the association signal entirely (appendix). Colocalisation analyses with the GTeX (Genotype-Tissue Expression) expression data[21] showed strong evidence (posterior probability 96–99%) that SNPs influencing progression in TRACK-HD were also expression quantitative trait loci (eQTLs) for *DHFR* in brain and peripheral tissues (appendix). Conversely, there was strong evidence (posterior probability 98%) that progression SNPs in REGISTRY were eQTLs for *MSH3* in blood and fibroblasts (appendix). Despite the lack of (or low) colocalisation between the TRACK-HD GWAS

For **additional results from the meta-analysis** see http://hdresearch.ucl.ac.uk/data-resources

	TRACK-HD	REGISTRY	Meta-analysis	GeM-HD	Description
GO: 32300	$3{\cdot}46\times10^{-9}$	$8{\cdot}34\times10^{-4}$	$1{\cdot}14\times10^{-11}$	$3{\cdot}82\times10^{-5}$	Mismatch repair complex
KEGG: 3430	$2{\cdot}79\times10^{-7}$	$4{\cdot}80\times10^{-2}$	$1{\cdot}34\times10^{-11}$	$6{\cdot}65\times10^{-6}$	Mismatch repair
GO: 30983	$6{\cdot}66\times10^{-7}$	$4{\cdot}20\times10^{-2}$	$3{\cdot}17\times10^{-11}$	$7{\cdot}43\times10^{-6}$	Binding of mismatched DNA
GO: 6298	$3{\cdot}53\times10^{-6}$	$4{\cdot}59\times10^{-2}$	$6{\cdot}54\times10^{-9}$	$3{\cdot}25\times10^{-6}$	Mismatch repair
GO: 32407	$1{\cdot}82\times10^{-2}$	$1{\cdot}10\times10^{-1}$	$6{\cdot}40\times10^{-4}$	$5{\cdot}74\times10^{-5}$	Binding of the MutSα complex
GO: 32389	$2{\cdot}25\times10^{-2}$	$4{\cdot}69\times10^{-2}$	$5{\cdot}23\times10^{-4}$	$1{\cdot}66\times10^{-5}$	MutLα complex
GO: 33683	$8{\cdot}01\times10^{-2}$	$5{\cdot}87\times10^{-4}$	$6{\cdot}74\times10^{-3}$	$1{\cdot}69\times10^{-6}$	Nucleotide-excision repair, DNA incision
GO: 90141	$3{\cdot}32\times10^{-1}$	$5{\cdot}93\times10^{-2}$	$7{\cdot}87\times10^{-1}$	$2{\cdot}30\times10^{-6}$	Positive regulation of mitochondrial fission
GO: 1900063	$4{\cdot}10\times10^{-1}$	$7{\cdot}29\times10^{-1}$	$6{\cdot}93\times10^{-1}$	$8{\cdot}39\times10^{-5}$	Regulation of peroxisome organisation
GO: 90200	$4{\cdot}58\times10^{-1}$	$5{\cdot}44\times10^{-1}$	$5{\cdot}28\times10^{-1}$	$8{\cdot}89\times10^{-8}$	Positive regulation of release of cytochrome c from mitochondria
GO: 90140	$5{\cdot}39\times10^{-1}$	$3{\cdot}32\times10^{-1}$	$8{\cdot}10\times10^{-1}$	$1{\cdot}57\times10^{-5}$	Regulation of mitochondrial fission
GO: 10822	$6{\cdot}21\times10^{-1}$	$6{\cdot}28\times10^{-1}$	$8{\cdot}53\times10^{-1}$	$7{\cdot}63\times10^{-5}$	Positive regulation of mitochondrion organisation
GO: 4748	$9{\cdot}64\times10^{-1}$	$6{\cdot}97\times10^{-1}$	$9{\cdot}79\times10^{-1}$	$2{\cdot}66\times10^{-5}$	Ribonucleoside-diphosphate reductase activity, with thioredoxin disulfide as acceptor
GO: 16728	$9{\cdot}64\times10^{-1}$	$6{\cdot}97\times10^{-1}$	$9{\cdot}79\times10^{-1}$	$2{\cdot}66\times10^{-5}$	Oxidoreductase activity, acting on CH or CH_2 groups, with disulfide as acceptor

The GO and KEGG terms in the first column refer to pathways of biologically related genes in the Gene Ontology Consortium[19] and Kyoto Encyclopedia of Genes and Genomes[17] databases, respectively. The p values refer to the association between the pathway and rate of progression in the TRACK-HD, REGISTRY, and meta-analysis data in this paper, and between the pathway and age at motor onset in the Genetic Modifiers of Huntington's Disease (GeM-HD) study.[8]

***Table*: Setscreen enrichment p values for the 14 pathways highlighted in the GeM-HD study**

and *MSH3* expression signals, several of the most significant GWAS SNPs were associated with reduced MSH3 expression and slower progression (appendix). Thus, the signal on chromosome 5 could be due to the coding change in *MSH3*, or to expression changes in MSH3, DHFR, or both, and both effects might operate in disease.

For **additional results on the associations** see http://hdresearch.ucl.ac.uk/data-resources

The second most significant association region in REGISTRY (appendix) tags a locus on chromosome 15 that has been previously associated with age of onset for Huntington's disease.[8] Five genes were highlighted, two of which reached the commonly accepted genome-wide significance criterion for gene-wide tests in the MAGMA analysis (*MTMR10* $p=2{\cdot}51\times10^{-7}$; *FAN1* $p=2{\cdot}35\times10^{-6}$). Notably, *MLH1* on chromosome 3 contains SNPs approaching genome-wide significant associations with age of onset ($p=2{\cdot}2\times10^{-7}$) in GeM-HD,[8] and also shows association in the REGISTRY progression gene-wide analysis ($p=3{\cdot}97\times10^{-4}$).

Both progression measures were correlated with age of onset. Thus, to test whether there is genetic association with progression independent of age of onset, we repeated the REGISTRY progression GWAS conditioning for the age of onset measure previously associated with this locus in GeM-HD in the 1314 individuals for whom we had measures of both progression and age of onset. Both *MTMR10* ($p=1{\cdot}33\times10^{-5}$) and *FAN1* ($p=1{\cdot}68\times10^{-4}$) remained significant. Furthermore, the most significant SNP (rs10611148, $p=2{\cdot}84\times10^{-7}$) was still significant after conditioning on age of onset ($p=2{\cdot}40\times10^{-5}$). Notably, the gene-wide associations at the *MSH3* locus in the TRACK-HD sample also remained significant after correcting for age of onset, as did the association with rs557874766 ($p=6{\cdot}30\times10^{-6}$). A similar pattern was observed at the *MSH3* locus in the meta-analysis. Thus, the associations reported here are mainly due to disease progression, rather than age of onset.

For **the complete list of genes in the pathways** see http://hdresearch.ucl.ac.uk/data-resources

Gene set analysis of the 14 pathways highlighted by the GeM-HD study,[8] showed that the four biological pathways which are most significantly associated with disease progression in the TRACK-HD progression GWAS were associated with DNA mismatch repair, and all these pathways also showed significant enrichment of signal in the REGISTRY progression analysis. This enrichment was strengthened in the meta-analysis (table). Notably, the top two pathways in TRACK-HD were also significant in the MAGMA competitive gene-set analysis (GO: 32300, $p=0{\cdot}010$; KEGG: 3430, $p=0{\cdot}00697$). *MSH3* ($p=2{\cdot}94\times10^{-8}$) and *POLD2* ($p=7{\cdot}21\times10^{-4}$) show association in TRACK-HD, with *MSH3* ($p=9{\cdot}52\times10^{-4}$) and *MLH1* ($p=3{\cdot}97\times10^{-4}$) showing association in REGISTRY (appendix). These findings are supported by analysis of DNA damage response pathways derived from Pearl and colleagues[22] (figure 4, appendix), in which two mismatch repair pathways were significantly associated with the unified TRACK-HD progression measure after correction for multiple testing of pathways. Again, the meta-analysis strengthens the enrichment (figure 4, appendix). Genes from the two significant pathways in TRACK-HD are shown in the appendix, with the significant genes being very similar to those from the GeM-HD pathways (appendix). A complete list of genes in the Pearl and colleagues[22] pathways is given in the additional data on our institutional website.

Discussion

The evidence from our study suggests that *MSH3* is probably a modifier of disease progression in Huntington's disease. We did an unbiased genetic screen using a novel disease progression measure in the TRACK-HD study, and identified a significant locus on chromosome 5, which encompasses three genes: *MTRNR2L2, MSH3*, and *DHFR*. This locus replicated in an independent group of participants from the European Huntington's disease REGISTRY study using a parallel disease progression measure, and was genome-wide significant in a meta-analysis of the two studies. The lead SNP in TRACK-HD, rs557874766, is a coding variant in MSH3; it is classed of moderate impact, making it genome-wide significant given its annotation.[23] This SNP became clearly genome-wide significant at the more widely used threshold of $p=5\times10^{-8}$ in our meta-analysis of TRACK-HD and REGISTRY. Furthermore, eQTL analyses showed association of lower *MSH3* expression with slower disease progression in our analyses.

Genetic modifiers of disease in people highlight pathways for therapeutic development; any pathway containing genetic variation that ameliorates or exacerbates disease forms a prevalidated relevant target. However, although the classic case-control design in complex disease has yielded multiple genetic associations highlighting relevant biology for novel treatment design,[24] studies of potential genetic modifiers in genetically simple Mendelian diseases have been difficult. Such diseases are rare and show gene and locus heterogeneity, thus finding genuine modifying associations in such a noisy background is inherently difficult. However, variants that modify disease in the context of a Mendelian causative gene might not be under negative selection pressure in the general population. Modifiers have been identified in specific genetic subtypes of disease[25] and in relatively large samples with consistent clinical data.[8,26]

One way to increase the power of genetic studies is to obtain a more accurate measure of phenotype. Prospective multivariate longitudinal measures such as those collected in TRACK-HD are ideal.[27] Our analysis of Huntington's disease progression showed that motor, cognitive, and brain imaging variables typically progress in parallel, and that patterns of loss are not sufficiently distinct to be considered subphenotypes for genetic analysis. Because psychiatric symptoms showed a different trajectory, we developed a single progression measure excluding the psychiatric data (figure 2). Age of onset was correlated with the unified progression measure but did not explain the genetic associations observed with progression. Thus, progression seems to be measuring a different aspect of disease to age of onset, or a similar aspect of disease, but with greater precision. The data available in REGISTRY are less comprehensive than in TRACK-HD; therefore we used a different approach by comparing cross-sectional severity at the most recent visit with that expected based on age and CAG repeat length. The unified progression measures in TRACK-HD and REGISTRY are correlated and again, the genetic associations in REGISTRY are not completely driven by age of onset, showing the usefulness of retrospective composite progression scores in genetic analysis. Prognostic indices for motor onset have been developed,[28] and the development of progression scores for prospective use, for example to stratify patients by predicted rate of progression to empower drug trials, warrants further attention.

Our study has limitations. TRACK-HD has the same standardised detailed phenotypic information on nearly all participants, but in only 243 participants carrying the Huntington's disease gene mutations. The REGISTRY study is much larger but the phenotypic data are less complete (appendix), often not collected at regular intervals and not on everyone in the study, and the data were collected in multiple centres, which will inevitably lead to variation. Nevertheless, the progression measures show the expected association with change in TMS and TFC in both TRACK-HD and REGISTRY, indicating their clinical relevance. However, future development of the progression statistic and confirmation of the genetic association in participants from ongoing large studies such as ENROLL,[29] with data collected more systematically than in REGISTRY but in less detail than TRACK-HD, would be ideal.

The locus we identified by use of the unified TRACK-HD progression measure included three genes, but *MSH3* is the best candidate. First, the lead SNP is a coding variant in exon 1 of *MSH3*, MSH3 Pro67Ala, with the potential to affect function.[30] Clinically, each copy of the minor allele (G) at this SNP corresponds to a decrease of about 0·4 units per year (95% CI 0·16–0·66) in the rate of change of TMS, and a reduction of about 0·12 units per year (95% CI 0·06–0·18) in the rate of change of TFC (appendix). Second, *MSH3* has been extensively implicated in the pathogenesis of Huntington's disease in both mouse and in-vitro studies, although this is the first human study to link *MSH3* to Huntington's disease. MSH3 is a neuronally expressed member of a family of DNA mismatch repair proteins;[31] it forms a heteromeric complex with MSH2 to form MutSβ, which recognises insertion-deletion loops of up to 13 nucleotides (figure 4).[32] There is, however, a high level of interconnectedness between pathways involved in the DNA damage response, and MutSβ is implicated in other processes.[19] Changes in CAG repeat size occur in terminally differentiated neurons in several Huntington's disease mouse models and in human patient striatum, the brain area most affected in Huntington's disease, and, notably, somatic expansion of the CAG repeat in the brain of patients with Huntington's disease predicts onset.[33] *Msh3* is required both for somatic expansion of *HTT* CAG repeats and for enhancing an early disease phenotype in mouse striatum,[34] *Msh3* expression level is associated with repeat instability in mouse brain (whereas *Dhfr* is not),[32] and expansion of CAG and CTG repeats is prevented by

msh3Δ in *Saccharomyces cerevisiae*.[35] This is a plausible mechanism through which variation in *MSH3* could operate in Huntington's disease (figure 4). In patients with myotonic dystrophy type 1, somatic instability of the CTG repeat (CAG on the non-coding strand), is associated with age of onset and an *MSH3* variant was recently associated with somatic instability in blood DNA of patients.[36] Variants in DNA repair pathways, including those in *MSH3*, contribute to age of onset modification in multiple CAG repeat expansion diseases,[37] implicating the CAG repeat itself as the source of modification.

To our knowledge, this is the first study to use a measure of progression to look for modifiers of a neurodegenerative Mendelian disorder. We detected association with a coding variant on chromosome 5, reaching genome-wide significance given its annotation[24] in just 216 participants, which was replicated in a larger independent sample and strengthened on meta-analysis. This indicates that our progression measure developed in TRACK-HD is an excellent reflection of disease pathophysiological progression, or that this is a locus with a large effect size, or, most likely, both. Although there are three genes at the locus, the most significant variant gives a coding change in MSH3, which together with the previous biological evidence makes it the most likely candidate. Somatic expansion of the CAG repeat through alterations in *MSH3* is a plausible mechanism for pathogenesis in Huntington's disease, which can be followed up in functional experiments in Huntington's disease models. These data provide additional support for the therapeutic targeting of HTT and the stability of its CAG repeat. Loss of, or variation in, mismatch repair complexes can cause malignancy and thus they are not regarded as ideal drug targets, but MSH3 is not essential because it can tolerate loss-of-function variation[38] and could provide a therapeutic target in Huntington's disease. We note that if MSH3 does operate to alter repeat expansion it might also be a drug target in other repeat expansion disorders.

Contributors

DJHM collected the data, did the analysis, and wrote the first draft of the manuscript. AFP did the genetic analysis and co-wrote the manuscript. DL did the statistical analysis of the phenotype, and co-wrote the manuscript. KL did the genetic analysis. BRL, RR, and AD the collected data. SM co-supervised the genetic analysis. PH co-supervised the data analyses, did the genetic analysis, and co-wrote the manuscript. LJ helped secure funding, supervised data analyses, and co-wrote the manuscript. SJT conceived the study, secured funding, recruited participants, supervised data analyses, and co-wrote the manuscript.

Declarations of interests

DL reports grant funding from the CHDI Foundation via UCL, and personal fees from Roche Pharmaceutical, Voyager Pharmaceutical, and Teva Pharmaceuticals. BRL reports grants from the CHDI Foundation via UCL, Teva Pharmaceuticals, and Lifemax Pharmaceuticals, and personal fees from Novartis, Roche, uniQure, Ionis Pharmaceuticals, and Raptor Pharmaceuticals. DJHM, KL, AD, AFP, SM, LJ, RR, PH, and SJT declare no competing interests.

Acknowledgments

We thank the people who have enabled this work through their participation in the TRACK-HD and REGISTRY studies. We also thank the following organisations for their support of this project: the European Commission 7th Framework Program (FP7/2007–2013; grant agreement number 2012–305121 "Integrated European –omics research project for diagnosis and therapy in rare neuromuscular and neurodegenerative diseases [NeurOmics]"), who provided funding for this project; CHDI Foundation, a non-profit biomedical research organisation exclusively dedicated to developing therapeutics that will substantially improve the lives of Huntington's disease-affected individuals, who funded the TRACK-HD and REGISTRY studies; the Medical Research Council (MRC) for their support of the MRC Centre for Neuropsychiatric Genetics and Genomics, MR/L010305/1; the Brain Research Trust (BRT), the Guarantors of Brain; and the Medical Research Council UK who all supported this project.

References

1 Huntington's Disease Collaborative Research Group. 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:** 971–83.

2 Ross CA, Tabrizi SJ. Huntington's disease: from molecular pathogenesis to clinical treatment. *Lancet Neurol* 2011; **10:** 83–98.

3 Hogarth P, Kayson E, Kieburtz K, et al. Interrater agreement in the assessment of motor manifestations of Huntington's disease. *Mov Disord* 2005; **20:** 293–97.

4 Long JD, Paulsen JS, Marder K, Zhang Y, Kim JI, Mills JA. Tracking motor impairments in the progression of Huntington's disease. *Mov Disord* 2013; **29:** 311–19.

5 Tabrizi SJ, Scahill RI, Owen G, et al. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *Lancet Neurol* 2013; **12:** 637–49.

6 Tabrizi SJ, Langbehn DR, Leavitt BR, et al. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *Lancet Neurol* 2009; **8:** 791–801.

7 Langbehn DR, Brinkman RR, Falush D, Paulsen JS, Hayden MR. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin Genet* 2004; **65:** 267–77.

8 Consortium GMoHsDG-H. Identification of genetic factors that modify clinical onset of Huntington's disease. *Cell* 2015; **162:** 516–26.

9 Orth M, Handley OJ, Schwenke C, et al. Observing Huntington's disease: the European Huntington's disease network's REGISTRY. *PLoS Curr* 2010; **2:** RRN1184.

10 Paulsen JS, Langbehn DR, Stout JC, et al. Detection of Huntington's disease decades before diagnosis: the Predict-HD study. *J Neurol Neurosurg Psychiatry* 2008; **79:** 874–80.

11 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.

12 Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26:** 2190–91.

13 Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genet* 2016; **48:** 709–17.

14 de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 2015; **11:** e1004219.

15 Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 2006; **7:** 29–59.

16 Moskvina V, O'Dushlaine C, Purcell S, Craddock N, Holmans P, O'Donovan MC. Evaluation of an approximation method for assessment of overall significance of multiple-dependent tests in a genomewide association study. *Genet Epidemiol* 2011; **35:** 861–66.

17 Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000; **28:** 27–30.

18 Kiezun A, Garimella K, Do R, et al. Exome sequencing and the genetic basis of complex traits. *Nat Genet* 2012; **44:** 623–30.

19 Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25:** 25–29.

20 Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015; **43:** D447–52.

21 Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015; **348:** 648–60.

22 Pearl LH, Schierz AC, Ward SE, Al-Lazikani B, Pearl FM. Therapeutic opportunities within the DNA damage response. *Nat Rev Cancer* 2015; **15:** 166–80.

23 Sveinbjornsson G, Albrechtsen A, Zink F, et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet* 2016; **48:** 314–17.

24 Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov* 2013; **12:** 581–94.

25 Trinh J, Gustavsson EK, Vilariño-Güell C, et al. DNM3 and genetic modifiers of age of onset in LRRK2 Gly2019Ser parkinsonism: a genome-wide linkage and association study. *Lancet Neurol* 2016; **15:** 1248–56.

26 Corvol H, Blackman SM, Boelle PY, et al. Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nat Commun* 2015; **6:** 8382.

27 Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 2014; **15:** 335–46.

28 Long JD, Langbehn DR, Tabrizi SJ, et al. Validation of a prognostic index for Huntington's disease. *Mov Disord* 2017; **32:** 256–63.

29 Landwehrmeyer GB, Fitzer-Attas CJ, Giuliano JD, et al. Data analytics from Enroll-HD, a global clinical research platform for Huntington's disease. *Mov Disord Clin Pract* 2016; **4:** 212–24.

30 Arnold M, Raffler J, Pfeufer A, Suhre K, Kastenmüller G. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics* 2015; **31:** 1334–36.

31 Gonitel R, Moffitt H, Sathasivam K, et al. DNA instability in postmitotic neurons. *Proc Natl Acad Sci USA* 2008; **105:** 3467–72.

32 Tome S, Manley K, Simard JP, et al. MSH3 polymorphisms and protein levels affect CAG repeat instability in Huntington's disease mice. *PLoS Genet* 2013; **9:** 16.

33 Swami M, Hendricks AE, Gillis T, et al. Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum Mol Genet* 2009; **18:** 3039–47.

34 Dragileva E, Hendricks A, Teed A, et al. Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. *Neurobiol Dis* 2009; **33:** 37–47.

35 Williams GM, Surtees JA. MSH3 Promotes dynamic behavior of trinucleotide repeat tracts in vivo. *Genetics* 2015; **200:** 737–54.

36 Morales F, Vasquez M, Santamaria C, Cuenca P, Corrales E, Monckton DG. A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. *DNA Repair* 2016; **40:** 57–66.

37 Bettencourt C, Hensman-Moss D, Flower M, et al. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. *Ann Neurol* 2016; **79:** 983–90.

38 Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016; **536:** 285–91.