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

Davina J Hensman Moss*[1], MBBS, Antonio F. Pardiñas*[2], PhD, Prof Douglas Langbehn[3], PhD, Kitty Lo[4], PhD, Prof Blair R. Leavitt[5], MD,CM, Prof Raymund Roos[6], MD, Prof Alexandra Durr[7], MD, Prof Simon Mead[8], PhD, the REGISTRY investigators and the TRACK-HD investigators, Prof Peter Holmans[2], PhD, Prof Lesley Jones[§2], PhD, Prof Sarah J Tabrizi[§1], PhD.

* These authors contributed equally to this work

§ These authors contributed equally to this work

1) UCL Huntington's Disease Centre, UCL Institute of Neurology, Dept. of Neurodegenerative Disease, London, UK
2) MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK
3) University of Iowa Carver College of Medicine, Dept. of Psychiatry and Biostatistics, Iowa, USA
4) UCL Genetics Institute, Div. of Biosciences, London, UK
5) Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada
6) Department of Neurology, Leiden University Medical Centre, Leiden, Netherlands
7) ICM and APHP Department of Genetics, Inserm U 1127, CNRS UMR 7225, Sorbonne Universités, UPMC Univ Paris 06 UMR S 1127, Pitié- Salpêtrière University Hospital, Paris, France
8) MRC Prion Unit, UCL Institute of Neurology, London, UK

Corresponding authors:

Sarah J Tabrizi at s.tabrizi@ucl.ac.uk

Lesley Jones at JonesL1@cardiff.ac.uk

ABSTRACT

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

Methods We generated a progression score based on principal component analysis of prospectively acquired longitudinal changes in motor, behavioural, cognitive and imaging measures in the TRACK-HD cohort of HD gene mutation carriers (data collected 2008 – 2011). We generated a parallel progression score using 1773 previously genotyped subjects from the REGISTRY study of HD mutation carriers (data collected 2003 – 2013). 216 subjects from TRACK-HD were genotyped. Association analyses was performed using GCTA, gene-wide analysis using MAGMA and meta-analysis using METAL.

Findings Longitudinal motor, cognitive and imaging scores were correlated with each other in TRACK-HD subjects, justifying a single, cross-domain measure as a unified progression measure in both studies. The TRACK-HD and REGISTRY progression measures were correlated with each other (r=0·674), and with AAO (r=0·315, r=0.234 respectively). A meta-analysis of progression in TRACK-HD and REGISTRY gave a genome-wide significant signal ($p=1.12x10^{-10}$) on chromosome 5 spanning 3 genes, *MSH3*, *DHFR* and *MTRNR2L2*. The lead SNP in TRACK-HD (rs557874766) is genome-wide significant in the meta-analysis ($p=1.58x10^{-8}$), and encodes an amino acid change (Pro67Ala) in MSH3. In TRACK-HD, each copy of the minor allele at this SNP is associated with a 0.4 (95% CI=0.16,0.66) units per year reduction in the rate of change of the Unified Huntington's Disease Rating Scale (UHDRS) Total Motor Score, and 0.12 (95% CI=0.06,0.18) units per year in

the rate of change of UHDRS Total Functional Capacity. The associations remained significant after adjusting for AAO.

Interpretation The multi-domain progression measure in TRACK-HD is associated with a functional variant that is genome-wide significant in a meta-analysis. The strong association in only 216 subjects implies that the progression measure is a sensitive reflection of disease burden, that the effect size at this locus is large, or both. As knock out of Msh3 reduces somatic expansion in HD mouse models, this highlights somatic expansion as a potential pathogenic modulator, informing therapeutic development in this untreatable disease.

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

Research in context

Evidence before this study

Huntington's disease (HD) is universally caused by a tract of 36 or more CAG in exon 1 of *HTT*. Genetic modifiers of age at motor onset have recently been identified in HD that highlight pathways, which if modulated in people, might delay disease onset. Onset of disease is preceded by a long prodromal phase accompanied by substantial brain cell death and age at motor onset is difficult to assess accurately and is not available in disease free at risk subjects. We searched all of PubMed up to Oct 31st 2016 for articles published in English containing "Huntington* disease" AND "genetic modifier" AND "onset" which identified 13 studies, then "Huntington* disease" AND "genetic modifier" AND "progression" which identified one review article. Amongst the 13 studies of genetic modification of HD onset most were small candidate gene studies; these were superseded by the one large genome wide genetic modifiers of HD study which identified three genome-wide significant loci, and implicated DNA handling in HD 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: there is substantial correlation among these variables. We used the disease progression measure as a quantitative variable in a genome-wide association study and in only 216 people from TRACK-HD detected a locus on chromosome 5 containing three significant genes, *MTRNR2L2*, *MSH3* and *DHFR*. The index variant encodes an amino acid 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, likely attributable to the same variants. A meta-analysis of the two studies strengthened the associations. There was some correlation between the progression measures and AAO of disease but this was not responsible for the association with disease progression. We also detected a signal on chromosome 15 in the REGISTRY study at the same locus as that previously associated with AAO.

Implications of all the available evidence

The progression measures used in this study can be generated in asymptomatic and symptomatic subjects using a subset of the clinically relevant parameters gathered in TRACK-HD. We use these measures to identify genetic modifiers of disease progression in HD. We saw a signal in only 216 subjects, which replicates in a larger sample, becoming genome-wide significant, thus reducing the chance of it being a false positive. This argues for the power of better phenotypic measures in genetic studies and implies that this locus has a large effect size 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. Notably, altering levels of Msh3 in HD mice reduces somatic instability and crossing *Msh3* null mice with HD mouse models prevents somatic instability of the *HTT* CAG repeat and reduces pathological phenotypes. Polymorphism in *MSH3* has been linked to somatic instability in myotonic dystrophy type 1 patients. 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 HD and potentially in other neurodegenerative expanded repeat disorders.

INTRODUCTION

Huntington's disease (HD) is a autosomal dominant fatal neurodegenerative condition caused by a CAG repeat expansion in *HTT* (1). It is a movement, cognitive and psychiatric disorder, but symptoms, age of disease onset (AAO) and disease progression vary (2). AAO (1, 3) reflects the trajectory of disease pathology up to the point of motor onset. However, the transition from premanifest to manifest HD is gradual (4, 5), making clinical definition challenging, furthermore psychiatric and cognitive changes may 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 onset age was recently shown to be genetically encoded, identifying genes of the DNA damage response as likely to modify onset of HD (8).

The need for clinical trials close to disease onset has motivated a raft of observational studies (5, 9, 10). This provides the opportunity to investigate the relationship between onset and progression, whether they are influenced by the same biology, and permits the study of subjects before clinical onset.

TRACK-HD represents the most deeply phenotyped cohort of premanifest and symptomatic disease with annual visits involving clinical, cognitive and motor testing alongside detailed brain imaging (5, 6). We used TRACK-HD (5, 6) data to generate a novel unified Huntington's disease progression measure for use in a genetic association analysis. We developed a similar measure in subjects from the REGISTRY study to replicate our findings (9).

MATERIALS AND METHODS

Study design and participants

All experiments were performed 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 in (8). Peripheral blood samples were donated by genetically-confirmed HD gene carriers, and all subjects provided informed written consent.

TRACK-HD was a prospective observational biomarker study collecting deep phenotypic data including imaging, quantitative motor and cognitive assessments on adult subjects with early HD, premanifest HD gene carriers and controls (5, 6). It provides annually collected high quality longitudinal prospective multivariate data over three years (2008-2011) with 243 subjects at baseline (6) (**Figure 1**). Demographic details of these individuals are shown in **Supplementary Information**.

REGISTRY(9) was a multisite prospective observational study which collected phenotypic data between 2003 – 2013 on over 13,000 subjects, mostly manifest HD gene carriers. The aim is for annual assessments +/- 3 months, though this is variable. The core data include: age, CAG repeat length, UHDRS Total Motor Score (TMS) and Total Functional Capacity (TFC); some patients have further assessments such as a cognitive battery (9). 1835 adult subjects from REGISTRY were included in this study on the basis of available genotype data (8). We obtained: TMS, symbol digit modality (SDMT), verbal fluency, Stroop colour reading, word reading and interference measures, functional assessment score, and TFC.

Procedures

For both studies, atypical severity scores were derived 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, due to differences in nature of the two data sets. In TRACK-HD, 24 variables were used to stratify the cohort in terms of disease progression (**Supplementary Information)**. They were divided *a priori* into 3 broad domains: (1) brain volume measures, (2) cognitive variables, and (3) quantitative-motor variables. For each variable the input for analysis was the subject's random longitudinal slope from a mixed effects regression model with correlated random intercepts and slopes for each subject. This model regressed the observed values on clinical probability of onset statistic (CPO) derived from CAG repeat length and age, and its interaction with follow-up length. The subjects' random slope estimates thus provided a measure of atypical longitudinal change not predicted by age and CAG length. Principal Component Analyses (PCA) of the random slopes was then used to study the dimensionality of these age and CAG-length corrected longitudinal changes. Further methodological detail, including control for potential demographic confounders, is given in Supplementary Methods and a flow chart is given in **Figure 1**.

For REGISTRY, in contrast to TRACK-HD, follow-up length and frequency was variable and missing 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 performed multiple imputation to fill in missing data, derived PCA severity scores and regressed off the predictive effect of age, CAG length, and gender on the PCA severity scores derived from this data to obtain the measure of atypical severity at the last visit. This gives a single point "severity" score based on how advanced a subject is compared with expectations based on their CAG repeat and age. 1773 subjects had adequate phenotypic data to score; further detail is given in Supplementary Methods and a flow chart is given in **Figure 1.**

Statistical and genetic analysis

Data analyses were performed using SAS/STAT 14·0 and 14·1 primarily via the MIXED, FACTOR and GML procedures (11). We occasionally used a log or inverse transform of a measure, with the

goal of better approximate normality of the distribution and the avoidance of inappropriate influence of extreme scores.

218 TRACK-HD study participants with complete serial phenotype data were genotyped on Illumina Omni2.5v1·1 arrays, and quality control performed as described in Supplementary Methods. Imputation was carried out using the 1000 Genomes phase 3 data as a reference (Supplementary Methods). This yielded 9·65 million biallelic markers of 216 individuals. Genotypes for the REGISTRY subjects were obtained from the GeM-HD Consortium (8), where details of their genotyping, quality control, curation and imputation are provided.

Association analyses were performed with the mixed linear model (MLM) functions included in GCTA v1·26(12). Conditional analyses were carried out using the COJO procedure included in GCTA. Because of the relatively small sample sizes, analyses were restricted to SNPs with minor allele frequency >1%. A meta-analysis of the TRACK-HD and REGISTRY association results was performed using METAL(13). 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 co-localisation analysis was carried out using GWAS-pw v0·21 (14). Gene-wide p-values were calculated using MAGMA v1·05, a powerful alternative to SNP-based analyses which aggregates the association signal inside genes while taking linkage disequilibrium (LD) between SNPs into account (15), using a window of 35kb upstream and 10kb downstream of genes (16). 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 LD with it. To maximise comparability with the GeM GWAS, our primary pathway analyses used Setscreen (17), which sums the log p-values of all SNPs in a pathway, also correcting for LD between SNPs.

All of the methods and analyses mentioned in this section are described in more detail in Supplementary Information.

RESULTS

We performed individual PCA of each domain and found that first PC scores were highly correlated between the domains ($P < 0{\cdot}0001$ in all cases, **Supplementary Information**.) No phenotypic subtypes of symptom clusters in motor, cognitive or imaging domains were observed; rather, longitudinal change in TRACK-HD not predictable by CAG-age was distributed on a correlated continuum (**Figure 2**). We therefore repeated PCA of the measures combined across all domains. The first PC of this combined analysis accounted for 23.4% of the joint variance, and was at least moderately correlated ($r>0{\cdot}4$) with most of the variables that contributed heavily to each domain-specific first PC (**Supplementary Tables 3 and 4**). The first psychiatric PC has notably lower correlation with motor and cognitive domains and CPO variables, 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 (**Figure 1 and 2B**). 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 CPO. We found a correlation of $r=0{\cdot}448$ ($p<0{\cdot}0001$) for the residual motor slope and $r=-0{\cdot}421$ ($p<0{\cdot}0001$) for the residual TFC slope. One unit increase in unified Huntington's disease progression measure corresponded to an increase of 0·71 (95% CI=0.34,1.08) units per year in the rate of change of TMS, and an increase of approximately 0·2 (95% CI=0.12,0.30) units per year in the rate of change of TFC. The 15 fastest progressing subjects in TRACK-HD showed a mean annual rate of decline in the UHDRS TMS of 2·52 more points per year than would be expected (Standard deviation =2.47, Standard Error of Mean =0.64); the 15 slowest progressing subjects had an annual TMS decline of 0·45 points less per

year than predicted by age and CAG length (Standard deviation =1.85, Standard Error of the Mean =0.48).

Huntington's disease subjects in the early stages of the disease were significantly faster progressors on the unified HD progression measure than those still in the premanifest phase ($p < 0{\cdot}0001$). Amongst the 96 subjects who had experienced onset, the rater AAO showed the expected relation with predicted AAO based on CAG length (**Supplementary Information**), and earlier than predicted AAO was correlated with faster progression on our unified HD progression measure ($r=0{\cdot}315$; $p = 0{\cdot}002$).

The unified HD progression measure developed in TRACK-HD could not be transferred directly to REGISTRY subjects with more limited data. Individual clinical measures in REGISTRY showed correlations across the motor, cognitive, and functional domains, consistent with our finding in TRACK-HD (**Supplementary Information**). PC1 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 (**Supplementary Information**). Therefore this first principal component was chosen as a measure of severity in the REGISTRY cohort (**Figure 2C**). 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 2A**) and we used this as the unified Registry progression measure. The unified REGISTRY progression measure and earlier than predicted AAO were modestly, but significantly, correlated ($r = 0{\cdot}2338$; $p<0{\cdot}0001$) (**Supplementary Information**). Atypically rapidly or slowly progressing subjects tend to become more atypical over time: correlation between time since disease onset and REGISTRY progression ($-0{\cdot}3074$; $p<0{\cdot}0001$) is greater than that between AAO and REGISTRY progression.

In TRACK-HD, the last-visit severity scores had a correlation of 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 0·631 between the TRACK-HD and REGISTRY progression measures in the 14 subjects present in both studies.

We then performed a genome-wide association analysis using the unified TRACK-HD progression measure as a quantitative trait, which yielded a significantly associated locus on chromosome 5 spanning *DHFR*, *MSH3* and *MTRNR2L2*. The index SNP rs557874766 is a coding missense variant in *MSH3* ($p = 5{\cdot}8 \times 10^{-8}$; G=0·2179/1091 (1000 Genomes); **Figure 3A and D and Supplementary Information**). Analyses conditioning on this SNP failed to show evidence for a second independent signal in this region in TRACK-HD **(Supplementary Information).** The genes in this locus were the only ones to reach genome-wide genic significance ((15, 18) (*MTRNR2L2* $p = 2{\cdot}15 \times 10^{-9}$; *MSH3* $p = 2{\cdot}94 \times 10^{-8}$; *DHFR* $p = 8{\cdot}37 \times 10^{-7}$, **http://hdresearch.ucl.ac.uk/data-resources/**).

Performing a genome-wide association analysis in REGISTRY using the unified progression measure replicated the signal identified in TRACK-HD (lead SNP rs420522, $p = 1{\cdot}39 \times 10^{-5}$) on a narrower locus (chr5:79902336-79950781), but still tagging the same three genes (**Figure 3B and D**). No genes reach genome-wide significance, though there is evidence of association (**http://hdresearch.ucl.ac.uk/data-resources/)** at *DHFR* ($p = 8{\cdot}45 \times 10^{-4}$), *MSH3* ($p = 9{\cdot}36 \times 10^{-4}$), and *MTRNR2L2* ($p = 1{\cdot}20 \times 10^{-3}$).

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 4C and D, Supplementary Information**), and also genic associations over *MSH3*, *DHFR*, and *MTRNR2L2* (**http://hdresearch.ucl.ac.uk/data-resources/**). The most significant SNP in the meta-analysis is rs1232027, which is genome-wide significant ($p = 1.12 \times 10^{-10}$), with the p-value of rs557874766 being 1.58×10^{-8}. No other regions attained genome-wide significance **(http://hdresearch.ucl.ac.uk/data-resources/).** Rs557874766 is nominally significant in REGISTRY (p=0.010), with a direction of effect consistent with that in TRACK-HD. Analyses conditional on rs1232027 largely remove the association in this region **(Supplementary**

Information), suggesting that there is only one signal. Conditioning on rs557874766 has a similar effect **(Supplementary Information)**, so this SNP remains a plausible causal variant.

As suggested by the meta-analysis, co-localisation analyses between TRACK-HD and REGISTRY showed this locus was likely influenced by the same SNPs in both studies (posterior probability 74.33%), although conditioning REGISTRY on rs55787466 did not remove the association signal entirely **(Supplementary Information)**. Co-localisation analyses with the GTeX expression data (19) showed strong evidence (posterior probability 96-99%) that SNPs influencing progression in TRACK-HD were also eQTLs for DHFR in brain and peripheral tissues (**Supplementary Information**). Conversely, there was strong evidence (posterior probability=97·8%) that progression SNPs in REGISTRY were eQTLs for MSH3 in blood and fibroblasts (**Supplementary Information**). Despite the lack of co-localisation between the TRACK GWAS and MSH3 expression signal, several of the most significant GWAS SNPs were associated with decreased MSH3 expression and slower progression (**Supplementary Information**). 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 may operate in disease.

The second most significant association region in REGISTRY (**Supplementary Information**) tags a locus on chromosome 15 which has been previously associated to HD AAO (8). Five genes were highlighted, two of which reached genome-wide genic significance (*MTMR10* $p=2{\cdot}51x10^{-7}$; *FAN1* $p=2{\cdot}35x10^{-6}$, **http://hdresearch.ucl.ac.uk/data-resources/**). Notably, *MLH1* on chr3 contains SNPs approaching genome-wide significance ($p = 2.2 \times 10^{-7}$) in GeM-HD (8), and also shows association in the REGISTRY progression gene-wide analysis ($p = 3{\cdot}97x10^{-4}$).

As noted earlier, both progression measures are correlated with AAO. Thus, to test whether there is an association with progression independent of AAO, we repeated the REGISTRY progression GWAS conditioning for the AAO measure previously associated with this locus in GeM in the individuals (N=1,314) for whom we had measures of both progression and AAO. Both *MTMR10*

($p=1{\cdot}33x10^{-5}$) and *FAN1* ($p=1{\cdot}68x10^{-4}$) remained significant (**http://hdresearch.ucl.ac.uk/data-resources/**). Furthermore, the most significant SNP (rs10611148, $p=2{\cdot}84x10^{-7}$) was still significant after conditioning on AAO ($p=2{\cdot}40x10^{-5}$). Notably, the genic associations at the *MSH3* locus in the TRACK-HD sample also remain significant after correcting for AAO (**http://hdresearch.ucl.ac.uk/data-resources/),** as does the association with rs557874766 ($p=6{\cdot}30x10^{-6}$). A similar pattern is observed at the *MSH3* locus in the meta-analysis. Thus, the associations reported here are mainly due to disease progression, rather than AAO.

Gene set analysis of the 14 pathways highlighted by the GeM-HD paper (8) show that the four most significant pathways in the TRACK-HD progression GWAS are related to mismatch repair, and all show significant enrichment of signal in REGISTRY (**Table 1**). This enrichment is strengthened in the meta-analysis (**Table 1**). Notably, the top two pathways in TRACK-HD are also significant in the MAGMA competitive gene-set analysis (GO:32300 p=0·010, KEGG:3430 p=0·00697). *MSH3* ($2.94x10^{-8}$) and *POLD2* ($7{\cdot}21x10^{-4}$) show association in TRACK, with *MSH3* ($9{\cdot}52x10^{-4}$) and *MLH1* ($3{\cdot}97x10^{-4}$) showing association in REGISTRY (**Supplementary Information).** These findings are supported by analysis of DNA damage response pathways derived from Pearl *et al*. (20) (**Figure 4A, Supplementary Information)** where two mismatch repair pathways are significantly associated with the unified TRACK-HD progression measure after correction for multiple testing of pathways. Again, the meta-analysis strengthens the enrichment **(Figure 4B, Supplementary Information**). Genes from the two significant pathways in TRACK-HD are shown in the **Supplementary Information**, with the significant genes being very similar to those from the GeM pathways (**Supplementary Information).** A complete list of genes in the Pearl et al. (20) pathways is given in **http://hdresearch.ucl.ac.uk/data-resources/.**

DISCUSSION

The evidence from our study suggests that *MSH3* is likely to be a modifier of disease progression in Huntington's disease. We undertook 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 subjects from the European HD 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 (21). This SNP becomes clearly genome-wide significant at the more widely used threshold of $p=5x10^{-8}$ in a meta-analysis of TRACK-HD and REGISTRY. Furthermore, eQTL analyses show association of lower MSH3 expression with slower disease progression.

Genetic modifiers of disease in people highlight pathways for therapeutic development; any pathway containing genetic variation that ameliorates or exacerbates disease forms a pre-validated relevant target. However, while the classical case-control design in complex disease has yielded multiple genetic associations highlighting relevant biology for novel treatment design (22), studies of potential genetic modifiers in genetically simple Mendelian diseases have been difficult to conduct. The 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 may not be under negative selection pressure in the general population. Recent successful identifications of modifiers have been made in specific genetic subtypes of disease (23) or in relatively large samples with consistent clinical data (8, 24).

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 (25). 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 sub-phenotypes for genetic analysis. As psychiatric symptoms showed a different trajectory, we developed a single progression measure excluding the psychiatric data (**Figure 2A and B**). AAO 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 AAO, or a similar aspect of disease, but with greater precision. The data available in REGISTRY are less comprehensive; therefore we used a different approach by comparing cross-sectional severity at the most recent visit with that expected based on age and CAG. The unified progression measures in TRACK-HD and REGISTRY are correlated and again, the genetic associations in REGISTRY are not completely driven by AAO, demonstrating the utility of retrospective composite progression scores in genetic analysis. Prognostic indices for motor onset have been developed (26), and the development of progression scores for prospective use, for example to empower drug trials by stratifying patients by predicted rate of progression warrants further attention.

However, our study has a number of limitations. TRACK-HD has the same standardised detailed phenotypic information on nearly all participants, but in only 243 HD gene mutation carrying subjects. The REGISTRY study is much larger but the phenotypic data are less complete (**Supplementary Information**), often not collected at regular intervals and not on everyone in the study, and in multiple centres which will inevitably lead to intrinsic variation. Nevertheless, the progression measures show the expected relationship 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 subjects from ongoing large studies such as ENROLL (27), with data collected more systematically than in REGISTRY but in less detail than TRACK-HD, would be ideal.

The genetic locus identified by the unified TRACK-HD progression measure association includes three genes, but *MSH3* is the likeliest candidate. Firstly, the lead SNP is a coding variant in exon 1 of

MSH3, MSH3 Pro67Ala, with the potential to affect function (SNiPA(28) accessed 10/11/2016). Clinically, each copy of the minor allele (G) at this SNP corresponds to a decrease of approximately 0.4 (95% CI=0.16,0.66) units per year in the rate of change of TMS, and a reduction of approximately 0.12 (95% CI=0.06,0.18) units per year in the rate of change of TFC (see Supplementary Information). Secondly, *MSH3* has been extensively implicated in the pathogenesis of HD in both mouse and cell studies, though this is the first human study to link *MSH3* to HD. MSH3 is a neuronally expressed member of a family of DNA mismatch repair proteins (29); it forms a heteromeric complex with MSH2 to form MutSβ, which recognises insertion-deletion loops of up to 13 nucleotides (30) **(Figure 4D)**. There is, however, a high level of interconnectedness between pathways involved in the DNA damage response, and MutSβ is implicated in other processes (20). Changes in CAG repeat size occur in terminally differentiated neurons in several HD mouse models and in human patient striatum, the brain area most affected in HD, and notably, somatic expansion of the CAG repeat in HD patient brain predicts onset (31). *Msh3* is required for both somatic expansion of *HTT* CAG repeats and for enhancing an early disease phenotype in mouse striatum (32), *Msh3* expression level is associated with repeat instability in mouse brain, (whereas DHFR is not) (30) and expansion of CAG and CTG repeats is prevented by msh3Δ in *Saccharomyces cerevisiae (33)*. This gives a plausible mechanism through which variation in *MSH3* could operate in HD **(Figure 4C and D)**. In patients with myotonic dystrophy type 1 (DM), 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 (34). Variants in DNA repair pathways including those in *MSH3* contribute to age of onset modification of multiple CAG repeat expansion diseases (35) implicating the CAG repeat itself as the source of modification in these diseases.

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 (21) in just 216 subjects, which replicated in a larger

independent sample and strengthened on meta-analysis. This indicates that either our progression measure developed in TRACK-HD is an excellent reflection of disease pathophysiological progression or that this is a locus with a very large effect size, or, most likely, both. While there are three genes at the locus, the most significant variant gives a coding change in MSH3, which together with the prior 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 HD which can be followed up in functional experiments in HD models. These data provide additional support for the therapeutic targeting of Huntingtin 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 as it can tolerate loss of function variation (36) and could provide a therapeutic target in HD. We note that if it does operate to alter repeat expansion it may also be a drug target in other repeat expansion disorders.

Acknowledgements and roles of funding sources

We would like to thank the people who have enabled this work through their participation in the TRACK-HD and REGISTRY studies.

We would like to thank the following organisations for their support of this project: The European Commission 7th Framework Program, (FP7/2007-2013) under grant agreement n° 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, Inc., a nonprofit biomedical research organization exclusively dedicated to developing therapeutics that will substantially improve the lives of HD-affected individuals who funded the TRACK-HD and REGISTRY studies. The Medical Research Council 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.

The funders of the 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 author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Author contributions and declarations

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

DL reports grant funding from CHDI via University College London (UCL), and personal fees from Roche Pharmaceutical, Voyager Pharmaceutical, and Teva Pharmaceuticals. BRL reports grants from 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.

Figure & Table legends

Figure 1: Study Design. After establishing that brain imaging, quantitative motor and cognitive variables are correlated and follow a similar trajectory, we scored the TRACK-HD subjects using principal component 1 as a Unified progression measure, and used this measure to look for genome-wide associations with HD progression. We replicated our findings in the EHDN Registry subjects by looking at how far their disease had progressed compared with expectations based on CAG/Age, and used this progression measure to look for genome-wide associations in REGISTRY. 1835

Registry subjects had genotype data (8). UHDRS TMS: Unified Huntington's Disease Rating Scale Total Motor Score. SDMT: symbol digit modality test. TFC: Total Functional Capacity.

Figure 2: Assessing progression in Huntington's disease (A) Graphical illustration of the trajectory of HD symptoms and signs over time, annotated to show what time period the different measures of onset and progression discussed in this paper cover. The TRACK-HD progression score uses longitudinal data over 3 years. Given limited longitudinal data in REGISTRY, cross-sectional severity at last visit compared to predicted severity was used as a proxy for progression. Age at onset occurs when a subject has unequivocal motor signs of Huntington's disease. (B) Distribution of progression measure in 218 members of TRACK-HD cohort. (C) Distribution of atypical severity (compared to predicted severity at final visit) in in 1835 members of the REGISTRY cohort. The curves in (B) and (C) are the normal distribution approximations of the severity score distributions.

Figure 3: Genome-wide Association Analysis of Progression Score. Green line in A-C: $5x10^{-8}$. (A) Manhattan plot of TRACK-HD GWA analysis yielding a locus on chromosome 5. Significance of SNPs (y axis) is plotted against genomic location (x axis). (B) Manhattan plot of REGISTRY GWA analysis showing suggestive trails on chromosome 15 in the same area as the GeM GWAS significant locus (8), and chromosome 5 in the same area as the TRACK progression GWAS. (C) Manhattan plot of Meta-analysis of TRACK 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 (LD) and $-\log^{10}$(p-value) of the significant locus on chromosome. The top image shows the chromosome; the red square shows the region which 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: $5x10^{-8}$

Figure 4: Significant genes are functionally linked and may cause somatic expansion of the *HTT* CAG repeat tract. STRING diagram showing all proteins from the Pearl *et al* (20) dataset with gene-wide p-values for association with Huntington's disease progression < 0.02 in **A**: the TRACK-HD dataset and **B**, the meta-analysis of TRACK-HD and REGISTRY (**http://hdresearch.ucl.ac.uk/data-resources/**). Genes with p<0.02 coloured; 10 further interactors in grey, confidence of interaction is shown in the 'Edge confidence' box, homo sapiens protein data used: http://string-db.org/cgi/ accessed October 2016 and January 2017 (37). **C** Schematic diagram showing how DNA mismatch repair proteins may be involved in somatic expansion of the CAG tract. Proteins with p<0.01 in the meta-analysed progression GWAS are coloured red. (i) The CAG repeat DNA is partly unwound by lesions, constraints of the CAG tract structure (middle image) or by transcription. (ii) This unwound DNA is recognised by MutSbeta (MSH2/MSH3) which recruits the endonuclease MutLalpha (PMS2/MLH1) and cleaves the DNA. (iii) Repair of the strand break leads to expansion of the CAG repeat. In neurones of the striatum somatic expansion is an ongoing process that occurs throughout life and variants in MSH3 may promote or inhibit repeat recognition, binding or repair. **D** Potential link between degree of somatic expansion over a patient's lifespan and rate of Huntington's disease progression.

Table 1: Setscreen enrichment p-values for the 14 pathways highlighted in GeM-HD (8).

The GO and KEGG terms in the first column refer to pathways of biologically related genes in the Gene Ontology Consortium(1) and Kyoto Encyclopedia of Genes and Genomes (2) databases respectively. The p-values in columns 2 – 4 refer to the association between the pathway indicated and rate of progression described in this paper (TRACK- TRACK-HD study; REGISTRY- REGISTRY study; META- meta-analysis). P(GeM) refers to the association between the indicated pathway and age at motor onset in the GeM-HD study (8).

Pathway	p(TRACK)	p(REGISTRY)	P(META)	p(GeM)	Description
GO: 32300	3·46E-09	8·34E-04	1.14E-11	3·82E-05	mismatch repair complex
KEGG 3430	2·79E-07	4·80E-02	1.34E-16	6·65E-06	mismatch repair (KEGG)
GO: 30983	6·66E-07	4·20E-04	3.17E-11	7·43E-06	mismatched DNA binding
GO: 6298	3·53E-06	4·59E-02	6.54E-09	3·25E-06	mismatch repair
GO: 32407	1·82E-02	1·10E-01	6.40E-04	5·74E-05	MutSalpha complex binding
GO: 32389	2·25E-02	4·69E-02	5.23E-04	1·66E-05	MutLalpha complex
GO: 33683	8·01E-02	5·87E-04	6.74E-03	1·69E-06	nucleotide-excision repair, DNA incision
GO: 90141	3·32E-01	5·93E-02	7.87E-01	2·30E-06	positive regulation of mitochondrial fission
GO: 1900063	4·10E-01	7·29E-01	6.93E-01	8·39E-05	regulation of peroxisome organization
GO: 90200	4·58E-01	5·44E-01	5.28E-01	8·89E-08	positive regulation of release of cytochrome c from mitochondria
GO: 90140	5·39E-01	3·32E-01	8.10E-01	1·57E-05	regulation of mitochondrial fission
GO: 10822	6·21E-01	6·28E-01	8.53E-01	7·63E-05	positive regulation of mitochondrion organization
GO: 4748	9·64E-01	6·97E-01	9.79E-01	2·66E-05	ribonucleoside-diphosphate reductase activity, thioredoxin disulfide as acceptor
GO: 16728	9·64E-01	6·97E-01	9.79E-01	2·66E-05	oxidoreductase activity, acting on CH or CH2 groups, disulfide as acceptor

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(6):971-83.
2. Ross CA, Tabrizi SJ. Huntington's disease: from molecular pathogenesis to clinical treatment. The Lancet Neurology. 2011;10(1):83-98.
3. Hogarth P, Kayson E, Kieburtz K, Marder K, Oakes D, Rosas D, et al. Interrater agreement in the assessment of motor manifestations of Huntington's disease. Movement disorders : official journal of the Movement Disorder Society. 2005;20(3):293-7.
4. Long JD, Paulsen JS, Marder K, Zhang Y, Kim JI, Mills JA. Tracking motor impairments in the progression of Huntington's disease. Movement disorders : official journal of the Movement Disorder Society. 2013;29(3):311-9.
5. Tabrizi SJ, Scahill RI, Owen G, Durr A, Leavitt BR, Roos RA, 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. The Lancet Neurology. 2013;12(7):637-49.
6. Tabrizi SJ, Langbehn DR, Leavitt BR, Roos RA, Durr A, Craufurd D, et al. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. Lancet neurology. 2009;8(9):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. Clinical genetics. 2004;65(4):267-77.
8. Consortium GMoHsDG-H. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. Cell. 2015;162(3):516-26.
9. Orth M, Handley OJ, Schwenke C, Dunnett SB, Craufurd D, Ho AK, et al. Observing Huntington's Disease: the European Huntington's Disease Network's REGISTRY. PLoS currents. 2010;2:RRN1184.
10. Paulsen JS, Langbehn DR, Stout JC, Aylward E, Ross CA, Nance M, et al. Detection of Huntington's disease decades before diagnosis: the Predict-HD study. Journal of neurology, neurosurgery, and psychiatry. 2008;79(8):874-80.
11. Copyright (c) 2002-2012 by SAS Institute Inc. C, NC, USA. SAS. SAS Institute Inc. Cary, NC, USA.: SAS Institute Inc; 2002-2012.
12. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011;88(1):76-82.
13. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1.
14. Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nature genetics. 2016;48(7):709-17.
15. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS computational biology. 2015;11(4):e1004219.
16. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annual review of genomics and human genetics. 2006;7:29-59.
17. 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. Genetic epidemiology. 2011;35(8):861-6.
18. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nature genetics. 2012;44(6):623-30.
19. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science (New York, NY). 2015;348(6235):648-60.

20. Pearl LH, Schierz AC, Ward SE, Al-Lazikani B, Pearl FM. Therapeutic opportunities within the DNA damage response. Nat Rev Cancer. 2015;15(3):166-80.
21. Sveinbjornsson G, Albrechtsen A, Zink F, Gudjonsson SA, Oddson A, Masson G, et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. Nature genetics. 2016;48(3):314-7.
22. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. Nat Rev Drug Discov. 2013;12(8):581-94.
23. Trinh J, Gustavsson EK, Vilariño-Güell C, Bortnick S, Latourelle J, McKenzie MB, et al. DNM3 and genetic modifiers of age of onset in LRRK2 Gly2019Ser parkinsonism: a genome-wide linkage and association study. The Lancet Neurology. 2016;15(12):1248-56.
24. Corvol H, Blackman SM, Boelle PY, Gallins PJ, Pace RG, Stonebraker JR, et al. Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. Nature communications. 2015;6:8382.
25. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nature reviews Genetics. 2014;15(5):335-46.
26. Long JD, Langbehn DR, Tabrizi SJ, Landwehrmeyer BG, Paulsen JS, Warner J, et al. Validation of a prognostic index for Huntington's disease. Movement disorders : official journal of the Movement Disorder Society. 2017;32(2):256-63.
27. Landwehrmeyer GB, Fitzer-Attas CJ, Giuliano JD, Gonçalves N, Anderson KE, Cardoso F, et al. Data Analytics from Enroll-HD, a Global Clinical Research Platform for Huntington's Disease. Movement Disorders Clinical Practice. 2016.
28. Arnold M, Raffler J, Pfeufer A, Suhre K, Kastenmüller G. SNiPA: an interactive, genetic variant-centered annotation browser. Bioinformatics. 2015;31(8):1334-6.
29. Gonitel R, Moffitt H, Sathasivam K, Woodman B, Detloff PJ, Faull RL, et al. DNA instability in postmitotic neurons. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(9):3467-72.
30. Tome S, Manley K, Simard JP, Clark GW, Slean MM, Swami M, et al. MSH3 Polymorphisms and Protein Levels Affect CAG Repeat Instability in Huntington's Disease Mice. Plos Genetics. 2013;9(2):16.
31. Swami M, Hendricks AE, Gillis T, Massood T, Mysore J, Myers RH, 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(16):3039-47.
32. Dragileva E, Hendricks A, Teed A, Gillis T, Lopez ET, Friedberg EC, et al. Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. Neurobiol Dis. 2009;33(1):37-47.
33. Williams GM, Surtees JA. MSH3 Promotes Dynamic Behavior of Trinucleotide Repeat Tracts In Vivo. Genetics. 2015;200(3):737-+.
34. 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.
35. Bettencourt C, Hensman-Moss D, Flower M, Wiethoff S, Brice A, Goizet C, et al. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. Ann Neurol. 2016;79(6):983-90.
36. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.
37. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic acids research. 2015;43(Database issue):D447-52.