Genetic and functional analyses point to *FAN1* as the source of multiple HD modifier effects

Kyung-Hee Kim,[1,2] Eun Pyo Hong,[1,2] Jun Wan Shin,[1,2] Michael J. Chao,[1,2] Jacob Loupe,[1,2] Tammy Gillis,[1] Jayalakshmi S. Mysore,[1] Peter Holmans,[3,^], Lesley Jones,[3,^] Michael Orth,[4,^] Darren G. Monckton,[5,^] Jeffrey D. Long,[6,^] Seung Kwak,[7,^] Ramee Lee,[7] James F. Gusella,[1,8,9,^] Marcy E. MacDonald,[1,2,8,^] and Jong-Min Lee[1,2,8,^,*]

[1] Molecular Neurogenetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA

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

[3] Medical Research Council Centre for Neuropsychiatric Genetics and Genomics, Department of Psychological Medicine and Neurology, School of Medicine, Cardiff University, Cardiff CF24 4HQ, UK

[4] Department of Neurology, University of Ulm, Ulm 89081, Germany

[5] Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK

[6] Department of Biostatistics, College of Public Health, and Department of Psychiatry, Carver College of Medicine, University of Iowa, Iowa City, Iowa 52242, USA

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

[8] Medical and Population Genetics Program, the Broad Institute of M.I.T. and Harvard, Cambridge, MA 02142, USA

[9] Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA 02115, USA

^ The GeM-HD consortium

* To whom correspondence should be addressed: Molecular Neurogenetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; Email: jlee51@mgh.harvard.edu

Abstract

A recent genome-wide association study of Huntington's disease (HD) implicated genes involved in DNA maintenance processes as modifiers of onset, including multiple genome-wide significant signals in a chr15 region containing the DNA repair gene *FAN1.* Here, we have carried out detailed genetic, molecular and cellular investigation of the modifiers at this locus. We find that missense changes within or near the DNA binding domain (*p.Arg507His* and *p.Arg377Trp*) reduce FAN1's DNA binding activity and its capacity to rescue mitomycin C-induced cytotoxicity, accounting for two infrequent onset-hastening modifier signals. We also identified a third onset-hastening modifier signal whose mechanism of action remains uncertain but does not involve an amino acid change in FAN1. We present additional evidence that a frequent onset-delaying modifier signal does not alter *FAN1* coding sequence but is associated with increased *FAN1* mRNA expression in the cerebral cortex. Consistent with these findings and other cellular overexpression/suppression studies, knock-out of *FAN1* increased CAG repeat expansion in HD induced pluripotent stem cells. Together, these studies support the process of somatic CAG repeat expansion as a therapeutic target in HD, and clearly indicate that multiple genetic variations act by different means through *FAN1* to influence HD onset in a manner that is largely additive, except in the rare circumstance that two onset-hastening alleles are present. Thus, an individual's particular combination of *FAN1* haplotypes may influence their suitability for HD clinical trials, particularly if the therapeutic agent aims to reduce CAG repeat instability.

Introduction

Huntington's disease (HD; MIM 143100) is due to an expanded (>35) CAG trinucleotide repeat in *HTT* (MIM 613004) [1] that encodes a lengthened polyglutamine segment in the large huntingtin protein. The HD mutation precipitates a constellation of neurological symptoms (e.g., involuntary movements, psychiatric disturbances, and cognitive decline)[2] and leads to premature death.[3] Both age-at-onset of characteristic motor signs and age-at-death show a strong inverse correlation with *HTT* CAG repeat length,[4-7] but there is remaining variance that shows heritability, indicating that in addition to *HTT*, other genes play a role in determining the exact course of HD in any given individual. To identify genetic factors that modify HD age-at-onset, the GeM-HD Consortium carried out an initial genome-wide association study (GWAS) of ~4,000 HD subjects that revealed three genome-wide significant onset modification signals, including 2 opposing modifier signals at 15q13.2-13.3.[8; 9] Recently, an expanded GWAS of ~9,000 HD subjects showed that the timing of HD onset is due to a length-dependent property of the expanded CAG repeat rather than to polyglutamine size, and that the effects of this property are modified by at least 6 loci that harbor genes involved in DNA repair/maintenance processes.[10] Together, these findings point to somatic CAG repeat expansion as the likely mechanism that determines age-at-onset, suggesting this process as a target for therapeutic intervention to delay or prevent HD onset. A detailed understanding of the modification effects could also underpin the use of genetic modifier genotype in clinical trial inclusion criteria and/or outcome analysis.

Among other genes, the 15q13.2-13.3 modifier locus contains *FAN1* (MIM 613534), which encodes a protein that participates in DNA interstrand cross-link (ICL) repair, and there is some support from model systems for reduced FAN1 levels resulting in trinucleotide repeat instability.[11; 12] The recent GWAS produced evidence of additional onset modifier signals in this region and suggested that onset-hastening and onset-delaying signals might be respectively associated with *FAN1* missense variants that are potentially deleterious and with eQTLs that increased *FAN1* mRNA levels in the cerebral cortex.[10] To test these predictions and to better inform potential HD intervention strategies, we have explored this chromosome (chr) 15 locus and its modifier signals in greater detail, using molecular, cellular, and genetic strategies.

Material and Methods

Subjects and phenotype

HD subjects for our initial onset modifier GWA analysis (namely GeM-HD GWA123)[9] and additional HD subjects for second GWA analysis (namely GeM-HD GWA12345) were described previously.[10] Age-at-onset of motor symptoms was based mainly on rater's estimation; if such data were not available, we then used age-at-onset data in the clinical records, or provided by family members or subjects. A stringent and robust phenotyping model for calculating individual deviation of age-at-onset from the expected age-at-onset based on one's expanded CAG repeat length was described previously.[6] This residual age-at-onset, representing the age-at-onset that was not explained by the CAG repeat size, was the primary phenotype used for genetic analysis. HD subjects with positive and negative residual age-at-onset developed motor symptoms later or earlier than expected, respectively. For example, a HD participant with residual age-at-onset of 10 means that the subject developed motor symptoms 10 years later than expected based on one's expanded CAG repeat length. The distribution of residual age-at-onset of motor symptoms of study subjects was very similar to a theoretical normal distribution, allowing standard statistical analyses.[10] Subject consents and the overall study design were reviewed and approved by the Partners HealthCare Institutional Review Board.[10]

Genotype imputation and association analysis using the quantitative residual age-at-onset phenotype

Subjects were genotyped in 5 HD modifier GWA studies as described previously.[10] In order to obtain genotype data with the highest SNP density for the current study, we performed two independent genotype imputations using 1000 Genomes Project (KGP) data (phase 3, mixed population) and Haplotype Reference Consortium (HRC, European) as the reference panels using the Michigan Imputation Server. Each typed GWA data set was subjected to extensive quality control analyses implemented by the Michigan Imputation Server, with genotype imputation being carried out using EAGLE for the phasing and MINIMAC3 for the imputation. The two independent sets of imputed genotype data were then merged by supplementing HRC imputation data with KGP imputation-specific data. In the end, 9,058 samples[10] were imputed for analysis in this study (4,414 males and 4,644 females). Since mixed effect models generated similar results to fixed effect models (data not shown), all SNPs (except 32 SNPs with imputation call rate < 100%) were analyzed to determine the level of

association between the minor allele count of a test SNP (additive model) and residual age-at-onset of subjects in a fixed effect model. Each fixed effect model for a given test SNP also included a set of covariates such as GWA study, sex, and the 4 top ancestry characteristics obtained from MDS (multi-dimensional scaling) analysis using genome-wide typed data by the PLINK program (version 1.07). After considering the patterns of association signals and recombination rate, our analysis focused on the chr15: 30900000-31500000 region. Low performing SNPs (5,206) were flagged (e.g., minor allele frequency < 0.01%, Hardy-Weinberg equilibrium p-value < 1E-6, or mean imputation R square values < 0.5). Minor allele frequency data were based on imputed genotype data of the study subjects. Details of the original chr15 modifier haplotype-tagging SNPs are described elsewhere.[10] Genomic coordinates were based on the GRCh37/hg19 assembly.

CNV analysis

Genotype intensity files were processed by the PENNCNV program to determine copy number variation (CNV) genotypes. We performed QC analysis by taking CNV segments that 1) were greater than 10 KB in size and 2) were determined by more than 10 sites. Frequencies of CNV (duplication and deletion) were calculated based on study subjects. Overall CNV frequencies are very low in the region; 124 subjects carry CNVs in the chr15 region, so we performed association analysis using the residual age-at-onset phenotype excluding these subjects in order to judge the impact of CNVs on the observed SNP association signals. To test whether increased dosage of *FAN1* affects age-at-onset, the residual age at onset of HD subjects carrying 3 copies of *FAN1* were compared to that of HD subjects with two copies of *FAN1*. Statistical significance of increased *FAN1* copy number was evaluated by an ANOVA model with the same covariates that were used in SNP association analysis.

Nomenclature of modifier haplotypes

To clarify each modifier haplotype on chr15, we used a naming system that was used previously.[10] Briefly, the first number represents chromosome number, the letter 'M' stands for modifier, the next letter indicates the order of discovery (A represents first modifier locus in that chromosome), and the last number is given to distinguish apparent independent modifier effects at the locus. Although the 15AM4 haplotype was originally thought to represent a single independent modifier effect, further analysis here indicates that this haplotype

captures multiple separate modifier effects, leading us in this study to focus on 15AM1, 15AM2, 15AM3, and 15AM5, each of which appears to capture a single independent modifier effect.

Construction of nucleotide-resolution modifier haplotypes using KGP data

We constructed full length modifier haplotypes based on haplotype-tagging SNPs and KGP data (phase 3). KGP chromosomes carrying the modifier haplotypes were identified based on refined modifier haplotype-tagging SNPs (15AM1, 3; 15AM2, 46; 15AM3, 3; and 15AM5, 2; refer to Table S2). Then, the most frequent alleles at each site in the KGP chromosomes were taken as the consensus alleles of each modifier haplotype. Subsequent analysis was focused on modifier haplotype boundaries defined by the haplotype-tagging SNPs. Next, we split the variation sites in the reconstructed haplotypes into two groups depending on whether or not SNPs were tested in our association analysis. For SNPs that we tested and are also annotated in the KGP, we evaluated whether modification signals were contributed to by functional changes in proteins. Similarly, we examined annotations of SNPs that were not tested in our analysis but described in the KGP data set to judge a possibility of modification of HD due to functional changes in proteins.

Functional annotation of SNPs

SNPs were annotated based on the ExAC database to identify variations with functional impact; we focused on 1) protein-altering variations and 2) deleterious mutations. Protein-altering variants included SNPs annotated as frameshift variant, in-frame deletion, in-frame insertion, missense variant, protein altering variant, splice acceptor variant, splice donor variant, start lost, stop gained, or stop lost.

Supplemental targeted capture sequencing analysis and genotype-based analysis

In order 1) to confirm functional variations on the modifier haplotypes identified by KGP haplotype analysis, 2) to detect variations that were not annotated in KGP, and 3) to investigate their roles in modifying HD, we performed targeted capture sequencing analysis of selected representative HD samples carrying modifier haplotypes. Previously described haplotype-tagging SNPs were used to select samples for capture sequencing. Capture probes were designed to enrich the target region (chr15: 31050000-31314317) using the Agilent SureDesign online tool; tiling probes were designed to cover each base pair position by 5 independent probes.

This generated ultra-long 120-mer biotinylated complementary RNA baits, and these capture probes were used for solution-based SureSelect target enrichment.[13] Briefly, genomic DNAs were sheared to produce smaller fragments, and libraries were prepared with sequence specific adaptors and indexes for multiplexing. DNA libraries were hybridized with biotinylated cRNA baits, which were complementary to regions to capture. The bait-library complexes were pulled down by magnetic beads. After the beads were washed, the RNA bait was digested to obtain the target DNA and subjected to sequencing using Illumina HiSeq 100bp paired-end sequencing at the Broad Institute. Samples were independently captured / indexed, and all capture materials were pooled together for sequencing in a single lane. Sequence reads were aligned (BWA and PICARD) for variant calling using the Genome Analysis Toolkit Best Practices workflow. At least 2.5 million sequence reads were produced for each sample, generating mean coverage of ~167X. Called variations were compared to KGP data. For shared variation sites between KGP data and our capture sequencing data, we evaluated the allele frequency of functional variations identified from KGP haplotype analysis to confirm the presence of functional variations with expected frequencies. For capture sequencing-specific variations, we focused on exon variations with expected allele frequency of the causal variation.

eQTL analysis using the CommonMind Consortium data and co-localization analysis

We obtained genotype and expression data from the CommonMind Consortium (CMC) in order to compare significances of SNPs in HD modifier association analysis with those in gene level eQTL analysis results. A standard eQTL analysis was performed by modeling expression levels of *FAN1* as a function of a test SNP (additive model), MDS covariates, and sex in a fixed effect model. SNPs in 1MB flanking regions of *FAN1* were tested (2,732 SNPs in 451 samples including 269 males). Then, SNPs were matched between HD GWA analysis and CMC eQTL analysis based on chromosome, genomic location (GRCh37/hg19), reference allele, and alternative allele. 1,840 SNPs were shared by HD GWA data and CMC eQTL data, and subsequently compared. Co-localization analysis was based on 'coloc' R package (version 3.2-1) in the R environment. Posterior probability of 1) causal variant for HD modifier trait only, 2) causal variant for CMC eQTL trait only, and 3) one common causal variant were calculated by 'coloc.abf' function using summary data for HD GWAS and CMC eQTL.

Cell culture

HEK293T cells and *FAN1* knock-out HEK293T cells were cultured in DMEM (Invitrogen) supplemented with 10% fetal bovine serum (Hyclone) and Penicillin-Streptomycin (Invitrogen) at 37°C and 5% CO_2. Lymphoblastoid cell lines (LCLs) derived from HD subjects were cultured in RPMI (Sigma-Aldrich) supplemented with 5% fetal bovine serum and Penicillin-Streptomycin at 37°C and 5% CO_2.

***FAN1* knock-out HEK293T cells**

To obtain modifier FAN1 protein extracts without endogenous FAN1, we inactivated endogenous *FAN1* in the HEK293T cells by the CRISPR/Cas9 approach. A *FAN1*-specific gRNA (target site, chr15:31206161-31206180) was selected from the GeCKO library,[14; 15] and potential off-target sites were checked in the Optimized CRISPR design website. Annealed oligos were inserted into the lentiCRISPR V2 vector (Addgene, 52961) as recommended in the GeCKO website.[14; 15] HEK293T cells were then transfected with *FAN1*-targeting CRISPR/Cas9 plasmid using Lipofectamine 2000 (Invitrogen). After selection with 1 µg/ml of Puromycin (Invitrogen) for 48 hours, single cells were picked and independently expanded to establish clonal lines. Two lines were developed and further analyzed by Sanger sequencing. The levels of FAN1 protein were evaluated by immunoblot analysis (FAN1 antibody, Abcam, ab95171) using whole cell lysate (50 µg). The location of FAN1 protein band in the immunoblot analysis was confirmed by FAN1-overexpressed samples.

Cloning of *FAN1* modifier haplotypes and generation of full length cDNA expression vectors

We selected a panel of HD LCLs to generate full-length *FAN1* expression vectors; cells were selected based on the genotypes of haplotype-tagging SNPs (rs150393409 for 15AM1; rs35811129 for 15AM2). Total RNA samples were prepared by RNeasy Plus mini kit (Qiagen), and cDNA was subsequently prepared by SuperScript III first-strand kit (Invitrogen). The concentration and purity of DNA and RNA samples were determined using NanoDrop (ND-1000). We used cDNA to amplify full length 15AM1, and 15AM2 *FAN1* (primers, 5'-GGCGGCCGCAATGATGTCAGAAGGGAAACCTCCTGAC-3' and 5'-CAGATCTTTAGCTAAGGCTTTGGCTCTTAGCTCCAAC-3'). Amplified DNA was isolated from the agarose gel and extracted for digestion by NotI and BglII. Digested DNA was inserted into a linearized FLAG-tagged

mammalian expression vector (pFLAG-CMV-4 expression vector; Sigma-Aldrich, E7158). The vector for 15AM3 was generated by site directed mutagenesis using 15AM2 plasmid. For all plasmids, full sequence was verified by Sanger sequencing analysis. Our *FAN1* plasmids confirmed the sequences reconstructed from KGP haplotype analysis and supplemental capture sequencing analysis.

Qualitative detection of interaction between FAN1 and 3' flap DNA substrate

To qualitatively assess whether FAN1 protein interacts with 3' flap DNA substrate, we performed EMSA (electrophoretic mobility shift assay)-type assays. Our 3' flap DNA substrate consisted of oligonucleotides that have been used previously in a crystal structure study, and shows strong binding affinity to FAN1.[16] The constituents of this complex are: 1) a 40-nucleotide long oligo with biotin at the 5' end (5'-CCCGTCCAGGTCTCGTCCGCGCCACTCGTGTCCAGCGTCG-3'; namely Biotin-A1), 2) a post-nick probe with 5'-phospho (5'-TGCGGACGAGACCTGGACGGG-3'; namely B2), and 3) a pre-nick oligo (5'-CGACGCTGGACACGAGTGGCTTTTTTTT-3'; namely C3).[16] These three oligos were annealed by incubating at 95°C for 5 min and gradually cooling to 4°C in annealing buffer (10 mM Tris, 1 mM EDTA, 50 mM NaCl, pH 8.0).[16] Annealing reaction included 5 femtomol of A1, 10 femtomole of B2, and 10 femtomole of C3 in 50 µl of annealing buffer. Fig. S7A shows sequences and a conceptual illustration of the annealed oligos. To detect interaction between FAN1 and DNA substrate, nuclear extracts obtained from HEK293T cells or *FAN1* knock-out HEK293T cells (prepared by the NE-PER Nuclear and Cytoplasmic Extraction Reagent; Pierce, 78833) were incubated with annealed DNA oligos; 10 µg of nuclear extract were incubated with annealed DNA oligos in 25 mM Tris-HCl pH 7.5, 150 mM NaCl, 50 ng/µl Poly (dI•dC), 5% glycerol, 0.1 mg/ml BSA, 10 mM $CaCl_2$ and 0.1 mM DTT at 4°C for 30 min. The binding mixtures were then resolved on 5% TBE gels (Bio-Rad), and FAN1-DNA substrate interaction was detected by Biotin-Streptavidin Detection Kit (Pierce, 20148). FAN1-DNA substrate interaction was also confirmed by excess amount (1000X) of non-labeled DNA oligo complex (competitor comprising non-biotinylated A1, B2, and C3). Due to the low performance of a commercially available FAN1 antibody, super shift assay was not performed.

Quantitative evaluation of DNA binding activities of modifier FAN1

To determine the levels of DNA binding activity of specific modifier haplotype FAN1 protein, we performed quantitative pull-down assays. The same oligo annealing procedures were performed to generate 3' flap structure except for the amounts of oligos; we incubated 10 picomole of biotinylated A1 (Biotin-A1), 20 picomole of B2, and 20 picomole of C3 in a 30 µl reaction volume. 10 µl of annealed oligos (i.e., 3' flap DNA substrate) was used for each reaction with nuclear lysate. To generate modifier FAN1 protein, *FAN1* knock-out HEK293T cells were transfected with *FAN1* expression vectors for 72 hours (Lipofectamine 2000), and nuclear extracts were prepared (NE-PER Nuclear and Cytoplasmic Extraction Reagent; Pierce, 78833). One hundred µg of individual nuclear extract was used for binding with DNA substrate in 300 µl of binding buffer (PBS, pH 7.4, protease inhibitors; on ice). FAN1-DNA complex was then incubated with 30 µl of Dynabeads MyOne Streptavidin (Invitrogen) at room temperature for 90 min.[17] After washing with PBS (3 times), the FAN1-DNA complex was dissociated with 0.1% SDS in NuPAGE LDS Sample Buffer (Invitogen), and resolved on NuPAGE gels (Thermo; 4-12% Bis-Tris Protein Gel). The amount of FAN1 pulled down by 3' flap DNA was determined by Immunoblot assay by anti-FAN1 antibody (Abcam ab95171). Densitometric quantification of bands was performed by Multi Gauge version 2.3 (Fujifilm). The quality of nuclear extracts was evaluated by Lamin A/C in the input (Cell Signaling, #4777; data not shown). FAN1 DNA binding activity data were normalized by the amount of FAN1 in the input. In addition, DNA binding activity of different modifier FAN1 was determined using annealed oligos hypothesized to form a DNA hairpin/loop structure. To generate CAG repeat-induced DNA hairpin/loop structures, we used CAG repeat oligos (5'-CGACGCTGGACACGAGTGGC-[CAG]*n*-GCGGACGAGACCTGGACGGG-3'), where *n*= 0, 2, 4, 6, 8, or 10 CAGs inserted into the reverse complementary sequence of Biotin-A1. Annealing these oligos to Biotin-A1 (using the same procedures as above) was used to generate a complex of Biotin-A1 and CAG repeat oligos with a CAG loop-outs with the potential to form hairpins. To check the annealed oligos, we incubated 5 femtomol of Biotin-A1 and 10 femtomol of CAG repeat oligo in 50 µl of annealing buffer. The annealed oligos were resolved on 5% TBE gels and detected by Biotin-Streptavidin Detection Kit. For quantitative DNA binding assays, we annealed 10 picomole of Biotin-A1 and 20 picomole of CAG repeat probe in a 36 µl reaction volume. Ten µl of annealed oligos was incubated with overexpressed FAN1 nuclear lysate (100 µg). Procedures for binding, streptavidin pull down, and data normalization were same as those in assays using 3' flap DNA substrate. Nuclease activity was not determined in these assays because these assay conditions were not optimal for measuring FAN1

catalytic activity, and the genetic investigation suggested a role for altered DNA binding or expression levels rather than FAN1 catalytic activity.[12]

Effects of modifier *FAN1* on cytotoxicity induced by mitomycin C (MMC)

We determined whether our modifier *FAN1* expression vectors could rescue MMC-mediated cytotoxicity in the absence of endogenous *FAN1* in HEK293T cells. Briefly, *FAN1* knock-out HEK293T cells were transfected with *FAN1* overexpression vectors (Invitrogen, Lipofectamine 2000, 72 hours). Then, *FAN1* knock-out HEK293T cells and FAN1-overexpressing knock-out HEK293T cells were treated with either DMSO or MMC (10 µM for 24 hours; Abcam). Cell viability of FAN1 deficient cells was determined by CellTiter-Glo luminescent cell viability assay (Promega). Each condition was tested in 5 wells, and the entire experiments were repeated 4 times. We also determined whether HD LCLs with different modifier *FAN1* haplotypes show different sensitivity to this DNA damaging agent. Ten different HD LCLs were used: 1) 5 lines carrying one copy of 15AM1 *FAN1* (*HTT* CAG: 37, 41, 41, 41, 53) and 2) 5 LCLs carrying at least one copy of 15AM2 (*HTT* CAG: 41, 42, 49, 52, 53). HD LCLs were plated in 96-well plates at $2x10^4$ cells/well and treated with MMC for 48 hours. Cell viability was determined by CellTiter-Glo Luminescent Cell Viability Assay. For each concentration, 6 wells were used, and 4 independent experiments were performed.

***FAN1* knock-out HD iPSC and CAG repeat expansion assay**

To examine the effect of *FAN1* knock-out on CAG expansion, we used an iPSC line derived from a HD subject fibroblast (GM04723) with 72 and 15 CAG repeats and two copies of onset-delaying modifier haplotype 15AM2 [10] generated as described by the HD iPSC Consortium.[18] Cells were maintained in mTeSR1 with 5X supplements (Stemcell technologies, 85850) at 37°C and 5% CO_2. The size of normal CAG repeat is expected to be 15 in all cells but the length of expanded CAG repeat may differ from 72 in individual clones due to low level repeat mosaicism. To generate *FAN1* knock-out in these iPSC, cells were treated 10 µM of ROCK inhibitor (Millipore, 688000) 1 hour before transfection with *FAN1* specific gRNA containing pLentiCRISPR V2 plasmid and empty vector by Amaxa Human Stem Cell Nucleofector Kit 1 (Lonza, VPH-5012). After selection with Puromycin (0.2 µg/ml for 48 hrs; Invitrogen), single cells were picked and independently maintained. From these, two knock-out iPSC clonal lines were generated and respectively confirmed by Sanger sequencing (1

nucleotide insertion and 8 nucleotide deletion in knock-out clonal line #1; 1 nucleotide insertion and 2 nucleotide deletion in knock-out clonal line #2). The control and knock-out cells all showed similar growth rates (data not shown). Genomic DNAs from each iPSC clonal cell line were collected monthly for 6 months (DNeasy, Qiagen) for PCR amplification with HD CAG repeats specific primers.[19] The distribution of CAG repeats in the bulk DNA was determined by a standardized fragment-based genotyping assay using an ABI3730XL sequencer, and the CAG expansion index was calculated from GeneMapper traces as previously described, with a threshold of 10% relative to the maximum height peak.[19]

Immunoblot assay

Cells were washed three times with PBS and incubated in cell lysis buffer (Cell Signaling, #9803) containing protease inhibitors (Roche, #11697498001). Protein lysate was separated on 4-12% Bis-Tris gel (Invitrogen) and transferred to PVDF membranes. After blocking in 5% milk in TBS containing 0.1% Tween-20, the membranes were incubated overnight at 4°C with primary antibodies. The following primary antibodies were used: anti-FAN1 (Abcam, ab95171), anti-FLAG (Sigma-Aldrich, F1804), and anti-actin (Sigma-Aldrich, A4700). After washing, the blots were probed with anti-mouse, or anti-rabbit IgG-HRP secondary antibodies and visualized by ECL (PerkinElmer).

Programs and Statistics

R (version 3.3.3) was used for plotting unless otherwise specified. R 3.5.3 was used for the co-localization analysis. For human FAN1 molecular and cell experiments, the values are expressed as the mean ± standard error of independent experiments. For group comparisons, Student's t-tests were performed.

Results

Definition of modifier signals

The most recent GWAS of ~9,000 HD subjects that implicated multiple significant modifier signals (two infrequent haplotypes: 15AM1, 15AM3 and two frequent haplotypes: 15AM2, 15AM4) included 1,963 QC-passed SNPs, imputed in the chr15:30,900,000-31,500,000 region (GRCh37/hg19) using the Haplotype Research Consortium (HRC) as a reference panel.[10] To achieve a more comprehensive high density definition of these modifier signals, we have augmented the analysis to now include 13,294 variants in this region by adding short genetic variations (SGV: including both SNPs and indels) imputed using the 1000 Genome Project (KGP) as reference panel (Fig. S1). Association analysis using residual age-at-onset of motor signs (i.e., age-at-onset corrected for individual CAG repeat size) as the phenotype revealed robust onset modification signals that crossed *FAN1* and two additional protein coding genes (*MTMR10* and part of *TRPM1)* (Fig. 1). Of the 13,294 imputed SGVs, 148 were protein-altering (i.e., frameshift, in-frame deletion, in-frame insertion, missense, splice acceptor, splice donor, start lost, stop gained, or stop lost) and some of these were predicted to be deleterious (Table S1). Two, rs150393409 and rs151322829, represent the *FAN1* missense changes, *p.Arg507His* and *p.Arg377Trp*, reported to tag the 15AM1 and 15AM3 modification signals, respectively (Fig. 1, red triangles).

To test the degree of independence of all modifier signals, we carried out association analyses in which we conditioned separately on each of the modifier-tagging SNPs defining 15AM1, 15AM2, 15AM3, and 15AM4 (Fig. S2). The results indicated that significance of the tag SNPs for each of the first three signals was little affected by conditioning on each other, whereas significance of the 15AM4 tag SNP was respectively increased and decreased by conditioning on the 15AM1 and 15AM2 signals. Reciprocally, conditioning on the 15AM4 tag SNP respectively increased and decreased the significance of 15AM1 and 15AM2 tag SNPs. Thus, the very frequent 15AM4 tag SNP does not in fact capture a single discrete modifier effect but rather a combination of modifier effects; consequently we excluded it from subsequent analyses to delineate individual modifier signals. Because both 15AM1 and 15AM3 signals are captured by infrequent alleles (respectively, ~1.4% and ~0.7% frequency in the study population), we next removed the 381 subjects carrying either of these tag SNPs to simplify more detailed analyses (namely, drop-out analysis). We then repeated the

association analysis with this slightly reduced sample size (n=8,677) (Fig. 2A). After removal of the onset-hastening signals, the association analysis no longer identified rs35811129 (Fig. 2A) as best capturing the onset-delaying 15AM2 signal, but rather pointed to the indel variant rs66632437 (p-value, 1.3E-25) and SNP rs8034856 (p-value, 1.4E-25) as better tags of this modifier effect. Since indels are not imputed with the HRC reference data, we chose rs8034856 (Fig. 2A) as the new tag SNP for analyses of 15AM2. Conditioning on this new 15AM2 tag SNP revealed a previously unrecognized genome-wide significant modifier signal (Fig. 2B), which we have termed 15AM5 (Fig. 2A, cyan). This third onset-hastening modifier effect is tagged by rs79213781, with a minor allele frequency (MAF) of ~1.9% (effect size, -2.55/year/minor allele). Conditional analysis of each of the 15AM1, new 15AM2, 15AM3 and 15AM5 tag SNPs confirmed 15AM5 to be separate from the previously defined modifier signals (Fig. S3). Repeating the association analysis after removal of samples carrying any of 15AM1, 15AM3, or 15AM5 tag SNPs and conditioning on the new 15AM2 peak SNP revealed only signals with p-values < 5E-04 (data, now shown), suggesting that any additional signals which emerge as significant with a further increase in GWAS sample size will represent low frequency alleles. Finally, conditioning on all independent modifier haplotypes (15AM1, 15AM2, 15AM3, and 15AM5) eliminated all signal detected by 15AM4 (Fig. S4).

In addition to the SGV association results, Fig. 1 also shows the frequency of copy-number variation (CNV) across the region. CNVs were seen in only 124 of 9,058 HD subjects (a low overall frequency in the study population), and these did not align with the patterns of SGV association signals (Fig. 1). Exclusion of HD subjects with CNVs had a minimal impact on single SGV association analysis (Fig. S5A), indicating that CNV is not responsible for any of the individual 15AM1, 15AM2, 15AM3, or 15AM5 modifier signals. However, eight subjects who carried 3 complete copies of *FAN1* showed delayed onset (p = 0.016) compared to those without CNV (Fig. S5B), supporting the possibility that in HD, increased FAN1 levels are protective.

Construction of nucleotide-resolution modifier haplotypes and functional annotation

For each individual modifier effect, we defined the extent of the chr15 region that contains the modifier-tagging SNP and those SGVs in strong linkage disequilibrium with it (r-square > 0.8) (Table S2). From the normal chromosomes of the KGP (phase 3), we then constructed nucleotide-resolution modifier haplotypes representing the consensus alleles of those KGP chromosomes that carry the tag SNPs and delineated

protein-altering SGVs. We examined separately those SGVs that were tested in our association analysis and those not-tested (Table 1). For example, in the KGP data, 2,836 SGVs are present within the boundaries of modifier haplotype 15AM1; 1,703 of these were imputed and tested in our SNP association analysis. Among these tested variants, only two SNPs are protein-altering: 1) rs150393409 (MAF, 1.44% in the study subjects; FAN1 *p.Arg507His*; annotated as deleterious and possibly damaging in SIFT and PolyPhen, respectively),[20-22] and 2) rs3784588 (MAF, 6.72% in the study subjects; TRPM1 *p.Val1434Ile* based on NM_001252020.1; tolerated/benign). While the former produced the most significant p-value in our association analysis (p-value, 6.7E-29 in a fixed model), the latter showed non-genome-wide significance (p-value, 1.7E-07 in a fixed model). By contrast, the KGP data haplotype for the onset-delaying 15AM2 effect did not reveal any alternative alleles that change amino acid sequence at tested variation sites. The 15AM3 modifier haplotype has only one protein-altering variant, the 15AM3-tag SNP rs151322829 (MAF, 0.7% in the study subjects; FAN1 *p.Arg377Trp*; deleterious/damaging; p-value, 9.3E-9 in a fixed model). The 15AM5 modifier haplotype carries a protein-altering variant only in *MTMR10* (rs6493352; MAF, 17.4% in the study subjects; *p.Arg648His*; tolerated/benign; p-value, 1.3E-04, in a fixed model). Of the SNPs listed in the KGP data that were not imputed and thus not tested in our association analysis, none has an alternative allele that is protein-altering on any of the modifier haplotypes.

We also performed targeted DNA capture sequencing analysis [13] of 32 representative HD individuals bearing the 15AM1, 15AM2, 15AM3, and 15AM5 tag SNPs to discover functional variants that might have been missed in our imputed SGV association analysis and in the KGP data analysis (Table 1). Given the frequencies of the tag SNPs, we were able to analyze heterozygotes for 15AM1, 15AM3 and 15AM5 and homozygotes for 15AM2. Consequently, causal variants were expected to be observed at 50% frequency in the former and at 100% frequency in the latter. Consistent with the GWAS data and functional annotation of modifier haplotypes in KGP data, the capture sequencing confirmed the protein-altering SNPs at expected European ancestry allele frequencies for 15AM1 (rs150393409), 15AM3 (rs151322829), and 15AM5 (rs6493352). Interestingly, one of the subjects homozygous for the onset-delaying 15AM2 modifier haplotype was found to harbor a very rare nonsense mutation (rs184745027; FAN1 *pArg952Ter*; MAF, 0.0071% based on gnomeAD v2.1.1) in one *FAN1* allele. This subject had onset 11.7 years earlier than expected from their inherited CAG repeat length (CAG=41; observed age-at-onset=44 yrs; expected age-at-onset=55.7 yrs),

consistent with the heterozygous loss of FAN1 function producing an onset-hastening effect. Together, these detailed analysis of the chr15 HD modifier locus point to *FAN1* as the only gene common to all 4 modifier haplotypes and the only gene with evidence for causal protein sequence alterations, suggesting in particular that the 15AM1 and 15AM3 modifier effects are due respectively to the FAN1 *p.Arg507His* and *p.Arg377Trp* missense variants (Table S3).

Reduced DNA binding activity of 15AM1 and 15AM3 FAN1 proteins

These postulated causal 15AM1 (*p.Arg507His)* and 15AM3 variants (*p.Arg377Trp)* involve amino acids in and near the DNA binding domain of FAN1 (Fig. 3A), respectively, raising the possibility that they affect FAN1 DNA binding activity. To test this hypothesis, we generated full-length expression constructs for 15AM1 and 15AM2 *FAN1* from HD lymphoblastoid cells (LCLs); we then derived a 15AM3 construct from the 15AM2 construct by site directed mutagenesis. Sanger sequencing confirmed that these full-length *FAN1* expression constructs matched the KGP haplotype/sequence data, enabling us to test the effect of the selected variants on the DNA binding activity of FAN1. In qualitative *in vitro* assays using wild-type (WT) and *FAN1*-knock-out cells (Fig. S6), we confirmed that FAN1 interacts with 3' flap DNA substrate (Fig. S7).[16] Quantitative *in vitro* assays (Fig. S8) showed that the DNA binding activities of FAN1 *p.Arg507His* (15AM1) and FAN1 *p.Arg377Trp* (15AM3) are significantly reduced compared to the 15AM2 FAN1, supporting the proposed functional impact of these missense variants (Fig. 3B). Reduced binding of FAN1 *p.Arg507His* (15AM1) and FAN1 *p.Arg377Trp* (15AM3) was also observed in binding assay using CAG loop-out, although not in a structure-specific fashion (Figs. S9-S10; Fig. 3C).

Functional consequences of reduced FAN1 DNA binding activity

We next determined whether the reduced DNA binding activity of 15AM1 and 15AM3 FAN1 is reflected in a functional impact on cells. We first evaluated whether FAN1 rescued cytotoxicity caused in FAN1-deficient HEK293T by mitomycin C (MMC).[23] 15AM2 FAN1 significantly rescued cells from this toxicity (Fig. 4A) whereas FAN1 *p.Arg507His* (15AM1) and FAN1 *p.Arg377Trp* (15AM3) did not. In addition, HD LCLs carrying the 15AM1 modifier showed increased sensitivity to MMC-mediated toxicity and consequently reduced viability compared to those with the 15AM2 modifier (Fig. 4B; Fig. S11), demonstrating that this measure of DNA repair

function of FAN1 differs by haplotype and further implicating *FAN1* variation as the source of the HD modifier effects.

The suggestion that the onset-hastening and onset-delaying modifier effects both act through an altered rate of somatic expansion of the CAG repeat has recently received mixed support from introduction of an *HTT* exon 1 transgene into FAN1-deficient osteosarcoma cells and from *FAN1* shRNA suppression in a human induced pluripotent stem (iPS) cell line with a long, inherently unstable CAG repeat (~121 CAGs from a subject with 109 CAGs).[12] To confirm that the capacity of FAN1 to influence HD onset is reflected in its ability to suppress somatic *HTT* CAG expansion at the endogenous *HTT* locus in a subject whose CAG repeat does not show dramatic modal size increase in culture, we completely inactivated *FAN1* by CRISPR/Cas9 editing in an HD iPS cell line with 73 CAGs. The result, as assessed from the calculated expansion index [19] of dividing iPSC clonal lines, was a gradual increase in the appearance of CAG repeats larger than the starting modal length over the course of continued culture (Fig. 4C; Fig. S12), indicating that FAN1 normally functions to stabilize the expanded CAG repeat.

Effects of modifier haplotypes on *FAN1* expression levels

In the recent GWAS publication, we reported correlations between the modification signals from 15AM2 (not 15AM1 or 15AM3), and *FAN1* eQTL signals in the cortex subset of the GTEx Consortium data.[10] We have not observed similar correlations for *MTMR10* or *TRPM1* eQTL signals (data not shown). To confirm in a larger data set that altered expression of *FAN1* could explain some modifier effects, we performed eQTL analysis using the CommonMind Consortium (CMC) data. As shown in Fig. 5, genome-wide significant eQTL signals for *FAN1* were generated by SNPs tagging the 15AM2 haplotype but not those tagging the 15AM1, 15AM3 or 15AM5 modifier haplotypes. In agreement with these observations, co-localization analysis of HD modifier association data and CMC eQTL data showed posterior probabilities for the effects sharing the same causal variant of 97.0% for SNPs tagging 15AM2. In contrast, 15AM1-, 15AM3-, and 15AM5-tagging SNPs revealed posterior probabilities of 8.3%, 8.5%, and 9.8%, respectively, for the HD modifier and eQTL effects sharing the same causal variant (Fig 5C). Comparison of the directions of the eQTL and onset modification signals confirmed that the onset-delaying 15AM2 modifier effect is associated with a relative increase in the level of *FAN1* mRNA expression.

Additional analyses for *FAN1* did not reveal any compelling evidence supporting a role for splicing as a genetic modifier of HD (GTEX portal). Taken together, the haplotype/sequencing data and eQTL analyses support the suspected role of *FAN1* as the gene responsible for the 4 independent modifier signals on chr15, with the causal variants on the 15AM1 and 15AM3 modifiers likely being the *FAN1* missense changes, rs150393409 and rs151322829, respectively; the 15AM2 modifier acting through alteration of *FAN1* mRNA levels. The mechanism by which the 15AM5 modifier acts is not clear and requires further investigation.

Effect of modifier genotype on age-at-onset modification

The individual SNP association analyses as performed in the GWA studies are based upon the assumption of additivity across major allele homozygotes, heterozygotes and minor allele homozygotes. Once the modifier haplotypes were defined, it became possible to use the tag SNPs to test whether the mechanisms underlying the modifier effects at this particular locus, which include both structural changes to FAN1 and differences in mRNA expression level in cortex, are indeed additive or, alternatively, show evidence of non-additive effects. Of the 9,058 GWA study subjects, 3,868 (42.7%) carried a haplotype other than 15AM1, 15AM2, 15AM3, or 15AM5 (i.e, a non-modifier or “Other” haplotype based on the 4 top tag SNPs), most frequently as a heterozygote with one of the modifier haplotypes (Table 2, top section). We first determined the effect size attributable to one copy of 15AM1, 15AM2, 15AM3 or 15AM5 by 1) selecting HD GWA subjects with one chromosome bearing the minor allele of the corresponding haplotype tag SNP and one chromosome with only major alleles at all 4 tag SNPs (i.e., a non-modifier chromosome or “Other” in Table 2) and then 2) comparing their residual age-at-onset to individuals bearing only two non-modifier chromosomes, with the difference in the respective residuals constituting the effect attributable to the modifier haplotype.

Having determined the effect size for a single copy of each modifier allele, we were then able to examine individuals carrying two modifier chromosomes. Under the additive model, for any given diplotype the expected residual age-at-onset can be predicted by simply adding the individual effect sizes of the haplotypes represented. Diplotypes that included 15AM2, which is reasonably frequent, showed no significant deviation (Student's t-test, p-value = 0.9175) from the additivity-based expectation (Table 2, middle section), indicating that each copy of 15AM2, which appears to act as an onset-delaying modifier through increased *FAN1*

expression in the cortex, has its effect independently of whether the other chromosome 15 bears another 15AM2 copy, a copy of one of the onset-hastening modifiers, or a non-modifier version of the locus.

All other combinations of modifier haplotypes (involving either homozygosity or compound heterozygosity for onset-hastening modifiers) are infrequent (Table 2, bottom section). All showed a residual age-at-onset that is more negative than expected (i.e., even earlier onset than predicted by additivity of the effects), a difference that over the entire group is statistically significant (Student's t-test, p-value = 0.03162). The fact that absence of any non-modifier FAN1 in an individual (i.e., 2 onset hastening-haplotypes) produces a worse outcome than predicted from onset-hastening modifier heterozygotes indicates that the presence of a non-modifier *FAN1* locus does act to mitigate to some degree the effect of the onset-hastening modifiers. The sample size for this group is small and therefore additional samples will be needed to replicate this finding and to examine the behavior of the individual onset-hastening modifier haplotypes.

Discussion

Our GWAS study of HD age-at-onset identified multiple modifier loci that remarkably, at or near the peak association signal, had the presence of a gene involved in DNA maintenance processes, suggesting that these genes are the sources of HD modification. At the 15q13.2-13.3 locus, the corresponding candidate gene was *FAN1* (Fanconi-Associated Nuclease 1) which encodes a protein involved in repair of ICL-induced DNA breaks, by cleaving DNA at every third nucleotide from a cut end using its 5'-3' exonuclease activity.[16] Further, while FAN1 does not participate in DNA double-strand break resection, it is required for efficient homologous recombination.[22; 24] Consequently, we set out to establish whether *FAN1* is indeed the modifier gene at this locus in order to guide strategies for therapeutic intervention and provide genetic guidance for clinical trial design. Taken together, our genetic and molecular data confirm that *FAN1* is the source of the chr15 locus modifier effect in HD and strongly point to altered FAN1 DNA binding activity (15AM1, 15AM3) and expression level (15AM2) as sources of onset-hastening and onset-delaying modification, respectively. Interestingly, the *p.Arg507His* and *p.Arg377Trp* variants associated with the 15AM1 and 15AM3 HD modifier effects were reported originally in early-onset breast cancer families, [25] but, despite a plausible argument for FAN1 function playing a role in tumorigenesis, these substitutions are not associated with increased risk of breast cancer in the general population. There is however evidence for a functional impact *in vivo* as the *p.Arg507His* substitution on 15AM1 has recently been associated in genetic studies with karyomegalic interstitial nephritis, a recessively inherited disease caused by loss of FAN1 function.[26] Our genetic data also indicate that some as yet undetermined mechanism other than a protein-altering variant can also result in earlier than expected onset (15AM5). Finally, the previously named haplotype 15AM4[10] does not represent a single independent modifier effect, but rather combinations of the effects of the 15AM1, 15AM2, 15AM3, and 15AM5.

The mirrored effects of reduced function (reduced DNA binding activity of 15AM1 and 15AM3) and increased expression (15AM2) of *FAN1* on the timing of HD onset suggest that a process involving FAN1 could represent a therapeutic target for slowing HD pathogenesis. In addition to *FAN1*, several other genes associated with DNA maintenance processes, including *MLH1, MSH3, PMS1, PMS2,* and *LIG1,* map in the vicinity of HD modifier GWAS signals on other chromosomes. The modifying effects of such DNA maintenance genes could theoretically influence HD pathogenesis by modulating accumulated DNA damage across the genome. However, the fact that the timing of HD onset is determined by a property of the CAG repeat, rather

than by continuous polyglutamine toxicity, favors a model in which these genetic modifiers act on the CAG repeat to alter its rate of somatic expansion. A role for *FAN1* in the process of CAG repeat expansion has been supported by recent studies of an HD exon 1 minigene in *FAN1* knock-out osteosarcoma cells, where CAG repeats of 30 and 70 CAGs were stable while modal repeats of 97 and 118 showed an increase in size during 40 days of culture.[12] The osteosarcoma cell assay did not distinguish the effects of wild-type FAN1, *p.Arg507His* variant FAN1 or a nuclease-inactive variant of FAN1 when these were each introduced via expression construct, suggesting that expression at endogenous levels or for a longer period of time might be required to assess a functional impact of the missense variant on expansion. Alternatively, FAN1's participation in the mechanism of repeat instability may not be precisely same in dividing cells and in non-dividing post-mitotic neurons. For example, the minor allele at *FAN1* rs3512 is associated with increased repeat expansion in the blood cells of HD subjects[27] but, as part of the 15AM2 haplotype, is associated with delayed age at onset and a suspected decrease in CAG expansion in neurons. Consequently, it remains unclear whether FAN1 acts in neurons through its ICL-resolving role or through some other function in DNA maintenance processes.

Although in our assays a difference in DNA binding for FAN1 containing either of the two missense variants was readily detectable in transfected cells, and LCLs with these mutations showed increased sensitivity to MMC, we focused our assessment of CAG expansion on the endogenous repeat in clonal HD patient-derived iPS cells from a subject with 73 CAGs. As monitored by our established expansion index measure,[19] in iPS cells with intact *FAN1* alleles, this starting repeat was relatively stable through the relatively short 6 month culture assay, as might be expected from the multi-year time-course of somatic expansion postulated to lead *in vivo* to clinical onset in HD subjects. In parallel isogenic lines with no FAN1 due to CRISPR/Cas9-mediated inactivation, we observed a slow, gradual increase in expansion index. A more exaggerated response has been seen in another human HD iPS cell line with a longer, inherently unstable repeat, where ongoing CAG expansion was increased by partial FAN1 shRNA knockdown.[12] Taken together with the increased expression level associated with the 15AM2 modifier and its onset-delaying effect, these studies argue that wild-type FAN1 functions to suppress CAG repeat expansion, a conclusion that also supports the candidacy of FAN1 as a potential modifier of other CAG repeat disorders [28] and, as suggested in a recent report that *Fan1* knock-out enhances somatic expansion of CGG repeats in a model of Fragile X disorders,[11] implies a role for FAN1 more broadly in triplet repeat diseases.

A role for FAN1 in suppressing somatic CAG expansion, combined with the onset-delaying effect of the 15AM2 modifier haplotype, suggest that increasing the expression of wild-type *FAN1* could have therapeutic benefit in delaying the onset of HD. The additivity of the 15AM2 effect in the presence of either a second 15AM2 allele or a 15AM1, 15AM3 or 15AM5 allele indicates that such a treatment could be widely applicable. Increased *FAN1* expression could theoretically be achieved by upregulating an endogenous allele, but this would require that the subject expresses a wild-type coding sequence from at least one allele. It is also not certain whether this upregulation would have to be chromosome-specific, since we cannot assess from the population analysis any potential negative effect of upregulating the 15AM1, 15AM3 or 15AM5 chromosomes. An alternative strategy would be to introduce exogenous FAN1, which might be particularly effective for subjects with two onset-hastening modifiers. Even if *FAN1* expression is not the mechanism chosen for therapeutic intervention, knowledge of the *FAN1* diplotype in an HD subject can be of potential importance for designing clinical trials of other mechanisms of intervention. For any trial in which conversion from premanifest to manifest disease is used as a measure, incorporating *FAN1* genotype and its onset-modifying effect could either guide selection of trial subjects or provide greater precision in calculating the effect of treatment. Similarly, *FAN1* genotype may be useful in the premanifest phase of HD where it has been shown to influence measures such as putamen volume and symbol digits modalities test score.[29] Whether *FAN1* genotype also influences the rate of progression of individual measures in manifest disease subjects and could be used in early HD trials remains to be determined. Additionally, if the intervention being tested is aimed specifically at blocking the process of somatic CAG expansion, then *FAN1* genotype and knowledge of any potential genetic interactions with other HD modifier genes that might interfere with or enhance the ability to reliably measure outcomes will also be crucial.

HD is a devastating disorder without treatments effective in delaying onset or slowing the worsening of disease manifestations. We undertook the characterization of genetic modifiers that alter the course of the human disease as a potential route to effective mechanism-based treatments. The accumulated evidence supporting *FAN1* as the HD modifier gene at the 15q13.2-13.3 locus reinforces the view that each of the DNA maintenance genes at or near modifier signals on chromosomes 2 (*PMS1*), 3 (*MLH1*), 5 (*MSH3*), 7 (*PMS2*) and 19 (*LIG1*) will prove to be the respective sources of HD modification through influences on somatic expansion of the CAG repeat. Similarly, these findings suggest that other GWA loci without known DNA

maintenance genes may point to genes that also participate in the CAG repeat expansion process even though they are not currently recognized as being involved in DNA repair (e.g., *RRM2B* on chromosome 8, *CCDC82* on chromosome 11). The success of this GWAS strategy as applied to the timing of HD onset dictates that this approach also be applied to other landmarks in the HD disease process and to the rate of worsening disease, particularly during the period of early manifest HD most applicable to clinical trials, in order to maximize the opportunity to identify therapeutic targets for effective treatments that alleviate the suffering of families with this devastating disease.

Data availability

The original summary statistics of the GeM-HD genome-wide association study to identify genetic modifiers of HD can be obtained at https://datadryad.org/resource/doi:10.5061/dryad.5d4s2r8.

Supplemental Data

Supplemental Data include twelve figures and three tables.

Declaration of Interests

J.F.G. is a Scientific Advisory Board member and has a financial interest in Triplet Therapeutics, Inc. His NIH-funded project is using genetic and genomic approaches to uncover other genes that significantly influence when diagnosable symptoms emerge and how rapidly they worsen in Huntington Disease. The company is developing new therapeutic approaches to address triplet repeat disorders such Huntington's Disease, Myotonic Dystrophy and spinocerebellar ataxias. His interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies.

Acknowledgements

We are grateful to Huntington's disease families and clinicians who participate in HD genetic research. This research was funded by NIH NINDS grants NS091161, NS105709, and NS049206, by an Anonymous Donor and by the CHDI Foundation.

Web Resources

OMIM, http://www.omim.org

Michigan Imputation Server, https://imputationserver.sph.umich.edu/index.html

PLINK, http://zzz.bwh.harvard.edu/plink/

PENNCNV, http://penncnv.openbioinformatics.org/en/latest/

1000 Genomes Project data, http://www.internationalgenome.org/

ExAC, http://exac.broadinstitute.org/

Genome Analysis Toolkit, https://software.broadinstitute.org/gatk/

CommonMind Consortium, https://www.synapse.org/#!Synapse:syn2759792/wiki/69613

MIT CRISPR design website, http://crispr.mit.edu/

GeCKO, http://genome-engineering.org/gecko/

sQTL analysis in the GTEX portal, https://gtexportal.org/home/spliceQTLPage

gnomeAD, https://gnomad.broadinstitute.org/

References

1. The Huntington's Disease Collaborative Research Group. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72, 971-983.
2. Bates, G.P., Dorsey, R., Gusella, J.F., Hayden, M.R., Kay, C., Leavitt, B.R., Nance, M., Ross, C.A., Scahill, R.I., Wetzel, R., et al. (2015). Huntington disease. Nature Reviews Disease Primers, 15005.
3. Keum, J.W., Shin, A., Gillis, T., Mysore, J.S., Abu Elneel, K., Lucente, D., Hadzi, T., Holmans, P., Jones, L., Orth, M., et al. (2016). The HTT CAG-Expansion Mutation Determines Age at Death but Not Disease Duration in Huntington Disease. American journal of human genetics 98, 287-298.
4. Andrew, S.E., Goldberg, Y.P., Kremer, B., Telenius, H., Theilmann, J., Adam, S., Starr, E., Squitieri, F., Lin, B., Kalchman, M.A., et al. (1993). The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nature genetics 4, 398-403.
5. Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., Abbott, M., et al. (1993). Trinucleotide repeat length instability and age of onset in Huntington's disease. Nature genetics 4, 387-392.
6. Lee, J.M., Ramos, E.M., Lee, J.H., Gillis, T., Mysore, J.S., Hayden, M.R., Warby, S.C., Morrison, P., Nance, M., Ross, C.A., et al. (2012). CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. Neurology 78, 690-695.
7. Persichetti, F., Srinidhi, J., Kanaley, L., Ge, P., Myers, R.H., D'Arrigo, K., Barnes, G.T., MacDonald, M.E., Vonsattel, J.P., Gusella, J.F., et al. (1994). Huntington's disease CAG trinucleotide repeats in pathologically confirmed post-mortem brains. Neurobiol Dis 1, 159-166.
8. Correia, K., Harold, D., Kim, K.H., Holmans, P., Jones, L., Orth, M., Myers, R.H., Kwak, S., Wheeler, V.C., MacDonald, M.E., et al. (2015). The Genetic Modifiers of Motor OnsetAge (GeM MOA) Website: Genome-wide Association Analysis for Genetic Modifiers of Huntington's Disease. J Huntingtons Dis 4, 279-284.
9. Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. (2015). Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. Cell 162, 516-526.
10. Genetic Modifiers of Huntington's Disease Consortium. Electronic address, g.h.m.h.e., and Genetic Modifiers of Huntington's Disease, C. (2019). CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. Cell 178, 887-900 e814.
11. Zhao, X.N., and Usdin, K. (2018). FAN1 protects against repeat expansions in a Fragile X mouse model. DNA repair 69, 1-5.
12. Goold, R., Flower, M., Moss, D.H., Medway, C., Wood-Kaczmar, A., Andre, R., Farshim, P., Bates, G.P., Holmans, P., Jones, L., et al. (2019). FAN1 modifies Huntington's disease progression by stabilizing the expanded HTT CAG repeat. Human molecular genetics 28, 650-661.
13. Chao, M.J., Kim, K.H., Shin, J.W., Lucente, D., Wheeler, V.C., Li, H., Roach, J.C., Hood, L., Wexler, N.S., Jardim, L.B., et al. (2018). Population-specific genetic modification of Huntington's disease in Venezuela. PLoS Genet 14, e1007274.
14. Sanjana, N.E., Shalem, O., and Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods 11, 783-784.
15. Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelson, T., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 343, 84-87.
16. Wang, R., Persky, N.S., Yoo, B., Ouerfelli, O., Smogorzewska, A., Elledge, S.J., and Pavletich, N.P. (2014). DNA repair. Mechanism of DNA interstrand cross-link processing by repair nuclease FAN1. Science 346, 1127-1130.
17. Deng, W.G., Zhu, Y., Montero, A., and Wu, K.K. (2003). Quantitative analysis of binding of transcription factor complex to biotinylated DNA probe by a streptavidin-agarose pulldown assay. Anal Biochem 323, 12-18.
18. Consortium, H.D.i. (2012). Induced pluripotent stem cells from patients with Huntington's disease show CAG-repeat-expansion-associated phenotypes. Cell stem cell 11, 264-278.
19. Lee, J.M., Zhang, J., Su, A.I., Walker, J.R., Wiltshire, T., Kang, K., Dragileva, E., Gillis, T., Lopez, E.T., Boily, M.J., et al. (2010). A novel approach to investigate tissue-specific trinucleotide repeat instability. BMC systems biology 4, 29.
20. O'Donnell, L., and Durocher, D. (2010). DNA repair has a new FAN1 club. Molecular cell 39, 167-169.

21. Smogorzewska, A., Desetty, R., Saito, T.T., Schlabach, M., Lach, F.P., Sowa, M.E., Clark, A.B., Kunkel, T.A., Harper, J.W., Colaiacovo, M.P., et al. (2010). A genetic screen identifies FAN1, a Fanconi anemia-associated nuclease necessary for DNA interstrand crosslink repair. Molecular cell 39, 36-47.
22. MacKay, C., Declais, A.C., Lundin, C., Agostinho, A., Deans, A.J., MacArtney, T.J., Hofmann, K., Gartner, A., West, S.C., Helleday, T., et al. (2010). Identification of KIAA1018/FAN1, a DNA repair nuclease recruited to DNA damage by monoubiquitinated FANCD2. Cell 142, 65-76.
23. Hlavin, E.M., Smeaton, M.B., and Miller, P.S. (2010). Initiation of DNA interstrand cross-link repair in mammalian cells. Environ Mol Mutagen 51, 604-624.
24. Kratz, K., Schopf, B., Kaden, S., Sendoel, A., Eberhard, R., Lademann, C., Cannavo, E., Sartori, A.A., Hengartner, M.O., and Jiricny, J. (2010). Deficiency of FANCD2-associated nuclease KIAA1018/FAN1 sensitizes cells to interstrand crosslinking agents. Cell 142, 77-88.
25. Park, D.J., Odefrey, F.A., Hammet, F., Giles, G.G., Baglietto, L., Abcfs, Mccs, Hopper, J.L., Schmidt, D.F., Makalic, E., et al. (2011). FAN1 variants identified in multiple-case early-onset breast cancer families via exome sequencing: no evidence for association with risk for breast cancer. Breast cancer research and treatment 130, 1043-1049.
26. Bastarache, L., Hughey, J.J., Hebbring, S., Marlo, J., Zhao, W., Ho, W.T., Van Driest, S.L., McGregor, T.L., Mosley, J.D., Wells, Q.S., et al. (2018). Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science 359, 1233-1239.
27. Ciosi, M., Maxwell, A., Cumming, S.A., Hensman Moss, D.J., Alshammari, A.M., Flower, M.D., Durr, A., Leavitt, B.R., Roos, R.A.C., team, T.-H., et al. (2019). A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes. EBioMedicine 48, 568-580.
28. Bettencourt, C., Hensman-Moss, D., Flower, M., Wiethoff, S., Brice, A., Goizet, C., Stevanin, G., Koutsis, G., Karadima, G., Panas, M., et al. (2016). DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. Ann Neurol 79, 983-990.
29. Long, J.D., Lee, J.M., Aylward, E.H., Gillis, T., Mysore, J.S., Abu Elneel, K., Chao, M.J., Paulsen, J.S., MacDonald, M.E., and Gusella, J.F. (2018). Genetic Modification of Huntington Disease Acts Early in the Prediagnosis Phase. American journal of human genetics 103, 349-357.

Table 1. Reconstructed onset modifier haplotypes based on KGP data and supplemental capture sequencing data.

(A) Since the causal variations responsible for onset modification signals must exist within the region of association, we constructed the consensus haplotype for the chromosomal regions responsible for each of the 4 modifier effects. Panels of modifier haplotype-tagging SNPs (peak SNP + SNPs with R-square>0.8) were used to define the analysis regions and to identify modifier haplotypes in KGP data for subsequent analysis focusing on functional variants. For example, the 15AM1 consensus haplotype was constructed based on 12 KGP chromosomes identified by 3 15AM1-tagging SNPs. In the region defined by 15AM1-tagging SNPs (chr15: 31202961 - 31300677), 2,836 SNPs are described in the KGP data, and 1,703 of them were tested in our association analysis. Among tested variants, the 15AM1 consensus haplotype carries alternative alleles at 120 sites; 2 of those are annotated as protein-altering variations (rs150393409 and rs3784588). Of 1,133 not-tested variants, the 15AM1 consensus haplotype carries alternative alleles at 1 site (not protein-altering).

(B) Independently, representative HD samples carrying modifier haplotype-tagging SNPs were sequenced by a targeted capture sequencing method. DNA samples were heterozygous for 15AM1 (18 samples), 15AM3 (3 samples) or 15AM5 (3 samples), or were homozygous for 15AM2 (8 samples). Allele frequencies observed in the sequence data across the modifier haplotypes were compared to the expected allele frequency of the causal variation (50% for heterozygous subjects, 100% for homozygous subjects) focusing on protein-altering variants. Analysis focusing on other exon SNPs that were not annotated in KGP data did not reveal any variants with expected allele frequencies.

[a] Among 46 15AM2-tagging SNPs, 36 SNPs passed QC for capture sequencing analysis.

KGP data: Haplotype-based analysis				
Modifier haplotype	15AM1	15AM2	15AM3	15AM5
Number of tagging SNPs with R2 > 0.8	3	46	3	2
Modifier haplotype boundary (size)	31202961 - 31300677 (97.7 KB)	31184803 - 31285422 (100.6 KB)	31165091 - 31221358 (56.2 KB)	31204637 - 31282611 (77.9 KB)
Number of chromosomes carrying the haplotype	12	400	7	37
Variation sites in KGP in the region	2836	2965	1669	2292
Tested for HD onset modification	1703	1756	936	1379
Site with reference allele	1583	1652	899	1354
Site with alternative allele	120	104	37	25
Protein-altering variants	2 (rs150393409, rs3784588)	0	1 (rs151322829)	1 (rs6493352)
Not tested for HD onset modification	1133	1209	733	913
Sites with reference allele	1132	1203	727	912
Sites with alternative allele	1	6	6	1
Protein-altering variants	0	0	0	0
Capture sequencing data				
Modifier haplotype	15AM1	15AM2	15AM3	15AM5
Modifier haplotype boundary (size)	31202961 - 31300677 (97.7 KB)	31184803 - 31285422 (100.6KB)	31165091 - 31221358 (56.2 KB)	31204637 - 31282611 (77.9 KB)
Number of QC-passed tagging SNPs	3	36 [a]	3	2
Samples (chromosomes) carrying the haplotype	18 (18)	8 (16)	3 (3)	3 (3)
Expected frequency of casual variant (%)	50	100	50	50
Protein-altering SNP with expected frequency	rs150393409	0	rs151322829	rs6493352
Other exon SNP with expected frequency	0	0	0	0

Table 2. Test whether combinations of two modifier haplotypes are additive or synergistic.

To estimate the effect size of each modifier (top section), we identified HD subjects carrying only one copy of either 15AM1, 15AM2, 15AM3, or 15AM5, and compared them with HD subjects without any of these modifier haplotypes ('Other' group as the baseline). For example, the effect size of 15AM1 was calculated by subtracting the observed residual age-at-onset of the "Other" group from that of HD subjects carrying one copy of the 15AM1 modifier haplotype. N represents the number of HD subjects.

We then determined whether the onset-delaying effect of a single 15AM2 chromosome is influenced by the presence of a second modifier haplotype (middle section). Expected residual age-at-onset of each group was based on the effect sizes estimated in the top section.

Finally, a small number of HD subjects carry two copies of infrequent onset-hastening modifier haplotypes (i.e., 15AM1, 15AM3, and 15AM5; bottom section). To determine whether two onset-hastening modifier haplotypes generate non-additive effects, we compared the residual age-at-onset based upon addition of effect sizes from the top section to observed residuals. Student's t-tests were performed to determine whether observed residuals and expected residuals were significantly different.

Subjects with one modifier haplotype and effect size estimation				
1st Chr	2nd Chr	N	Observed residual (years)	Effect size (years)
Other	Other	3868	-0.55	Baseline
15AM1	Other	164	-5.02	15AM1 : -4.47
15AM2	Other	3621	0.64	15AM2 : 1.19
15AM3	Other	89	-3.90	15AM3 : -3.35
15AM5	Other	227	-2.72	15AM5 : -2.17

Effects of selected modifier haplotypes in the presence of 15AM2				
1st Chr	2nd Chr	N	Observed residual (years)	Expected residual (years)
15AM1	15AM2	85	-4.21	-3.83
15AM2	15AM2	851	1.81	1.83
15AM3	15AM2	30	-2.92	-2.71
15AM5	15AM2	105	-1.69	-1.53
			Student's t-test, p-value = 0.9175	

Subjects with two onset-hastening modifier haplotypes				
1st Chr	2nd Chr	N	Observed residual (years)	Expected residual (years)
15AM1	15AM1	1	-21.76	-9.49
15AM1	15AM3	2	-13.40	-8.37
15AM1	15AM5	6	-9.38	-7.19
15AM3	15AM3	2	-10.83	-7.25
15AM3	15AM5	1	-15.57	-6.07
15AM5	15AM5	6	-8.13	-4.89
			Student's t-test, p-value = 0.03162	

Figure Legend

Figure 1. Dense SGV marker association with residual age-at-onset at the chromosome 15 HD modifier locus.

All SGVs (dark grey, QC-passed; light grey, QC-failed), merged from two sets of imputed genotype data, were evaluated for the levels of association with residual age-at-onset. For each test variant, the residual age-at-onset as a continuous phenotype was modeled as a function of minor allele count (additive model), sex, ancestry characteristics, and study group in a fixed effect model. Y-axis and X-axis represent the levels of significance of association and genomic coordinate (GRCh37/hg19), respectively. Up-ward and down-ward triangles represent SNPs whose minor alleles are associated with delayed and hastened age-at-onset, respectively. Filled blue and red triangles represent protein-altering variants, with the latter predicted to be deleterious / damaging. Traces in blue and red represent the frequencies of duplication and deletion from the CNV analysis of the study samples (secondary Y-axis). All RefSeq transcripts for a given gene were combined to show locations of exons (vertical bars) and introns; gene symbols in red and blue represent genes on plus and minus strands, respectively. In the region of genome-wide significant signal, only *FAN1, MTMR10* and *TRPM1* are protein-coding genes. Two genome-wide significant and deleterious missense SNPs are indicated by black arrows.

Figure 2. Refinement of top tag SNPs and identification of an additional onset-hastening modifier effect.

(A) To discover additional independent modifier haplotypes, we excluded samples carrying 15AM1 or 15AM3 haplotypes, and performed association analysis. After drop-out of these infrequent onset-hastening haplotypes, the peak SNP tagging the common onset-delaying 15AM2 haplotype was rs8034856 (dark green circle) rather than the original rs35811129 (light green circle). Interestingly, a relatively infrequent SNP remained significant, suggesting the existence of an independent modifier haplotype tagged by rs79213781 (cyan circle).

(B) To confirm that the modifier effect tagged by rs79213781 (cyan circle) is independent, the new tag SNP for 15AM2 (rs8034856) was used for conditional analysis of the sample set excluding carriers of 15AM1 or 15AM3. Dotted lines indicate genome-wide significant association.

Figure 3. FAN1 missense variants on 15AM1 and 15AM3 modifier haplotypes reduce DNA binding.

(A) Through sequence analysis, we fully re-constructed *FAN1* modifier haplotypes at the levels of nucleotide and amino acid. Subsequently, LCLs from representative HD patients were used for cDNA cloning to generate full length mammalian expression constructs for the modifier haplotypes 15AM1 and 15AM2. Since 1) cell lines carrying the modifier haplotype 15AM3 were not readily available due to low allele frequency, and 2) the modifier haplotype 15AM3 differs from 15AM2 at one nucleotide, site directed mutagenesis was performed to generate the expression vector for 15AM3. The *FAN1* sequence of 15AM5 is identical to that of 15AM2, and therefore the effects of 15AM5 modifier FAN1 can be tested by using the 15AM2 modifier FAN1 construct (second diagram). Each construct has a FLAG tag at the N-terminus for biochemical assays. Subsequent sequencing analysis of expression constructs confirmed the DNA sequence reconstructed by KGP data analysis and capture sequencing. Black triangles indicate deleterious/damaging missense variations. Domains are shown: UBZ, SAP, TPR, and NUC represent ubiquitin binding Zinc finger, SAF-AIB, Acinus and PIAS (DNA binding domain), tetratricopeptide repeat, and virus type replication repair nuclease, respectively. Diagrams of haplotypes were focused on *FAN1* because modifier haplotype-specific variations do not alter protein sequence or are not associated with expression levels of other nearby genes (e.g., *MTMR10*, *TRPM1*).

(B) DNA binding activities of the 3 different FAN1 proteins were determined by *in vitro* assays using 3' flap comprising a set of 3 oligos. FLAG-tagged FAN1 protein nuclear extract and biotinylated oligos were incubated, and the protein-oligo complex was pulled down by streptavidin. The amount of FAN1 was quantified by immunoblot analysis using FAN1 antibody, and data were normalized by the amount of FAN1 in the inputs. DNA binding activity on the Y-axis represents densitometry reading of pull-down divided by that of input for each group (A.U., arbitrary unit). Five independent experiments were summarized (5 technical replicates/group).

(C) Similarly, DNA binding activity of FAN1 was determined using oligos of 10 CAG loop-out (potentially forming 5 CAG long hairpin/loop structure). Experiments were repeated three times (3 technical replicates/group). Error bars represent the standard errors. *, p-value < 0.05; **, p-value < 0.01.

Figure 4. Reduced rescue from MMC-mediated toxicity by 15AM1 and 15AM3.

(A) *FAN1* knock-out HEK293T cells were used to determine the levels of rescue afforded by over-expression of different *FAN1* haplotypes. *FAN1* knock-out HEK293T cells were treated with either vehicle (-) or MMC (+; 10 μM for 24 hrs), and then viability assays were performed to determine the levels of rescue from the MMC-induced cytotoxicity. Viability in the absence of MMC treatment was considered as 100% (i.e., control) for each experimental group. Error bars represent standard errors (12 technical replicates/condition from 3 independent experiments). NS, not significant

(B) LCLs derived from HD subjects carrying either 15AM1 (red; 5 lines, all heterozygous for 15AM1) or 15AM2 modifier FAN1 (green; 3 and 2 lines heterozygous and homozygous for the 15AM2, respectively) were treated with various concentrations of MMC to determine viability by Cell Titer-Glo assay. Vehicle treated cells were used as controls. We hypothesized that the expression levels of *FAN1* in cells were similar due to the lack of a significant *FAN1* eQTL in LCL (Fig. S11). LCLs with 15AM3 were not available due to the low frequency of the haplotype, and therefore were not tested. Error bars represent standard errors (5 biological repilcates/group, 24 technical replicates/condition for each biological replicate from 4 independent experiments). **, p-value < 0.01; ***, p-value < 0.001; ****, p-value < 0.0001 by Student's t-test.

(C) *FAN1* was knocked out in iPSC derived from a HD subject (carrying 72/15 CAGs and two copies of 15AM2 modifier haplotype) by CRISPR/Cas9. Subsequently, 2 clonal lines for *FAN1* knock-out (KO; starting modal CAGs 75 and 72, respectively) and 2 control lines (empty vector treated; starting modal CAGs 74 and 73, respectively) were established. These cells were maintained in standard iPSC culture conditions, and DNA samples were collected longitudinally and analyzed by PCR sizing to calculate expansion index. Thus, these data represent the levels of CAG repeat instability in proliferating iPSC, not differentiated neurons. Each data point represent a single measurement. Fluctuations seen in KO clones (#1 at 4 months and #2 at 5 months) might be due to the possibility that some of expanded alleles were filtered out by 10% peak height threshold in our expansion index quantification procedure because of low PCR amplification efficiencies in corresponding samples.

Figure 5. Co-localization of 15AM2 modification signals and *FAN1* eQTL signals.

(A) eQTL analysis was performed using RNAseq and genotype data sets from the CMC, focusing on *FAN1*. Expression levels of *FAN1* were modeled as a function of a test SNP and covariates in a fixed effect model. Y-axis and X-axis represents significance in eQTL analysis (i.e., -log10(p-value)) and genomic location (GRCh 37/hg19), respectively. Haplotype-tagging SNPs are indicated by colored symbols (red, 15AM1, green, 15AM2; purple, 15AM3; cyan, 15AM5).

(B) Signals (i.e., -log10(p-value)) in HD modifier GWA data (X-axis) and CMC *FAN1* eQTL data (Y-axis) were compared for shared SNPs between two data sets.

(C) Co-localization analysis was performed using modifier association data and *FAN1* eQTL data of haplotype-tagging SNPs. Percent posterior probability was calculated for 1) causal variant for HD onset modification only, 2) causal variant for *FAN1* eQTL only, and 3) common causal variant for both HD onset modification and *FAN1* eQTL using haplotype-tagging SNPs.

Fig 1

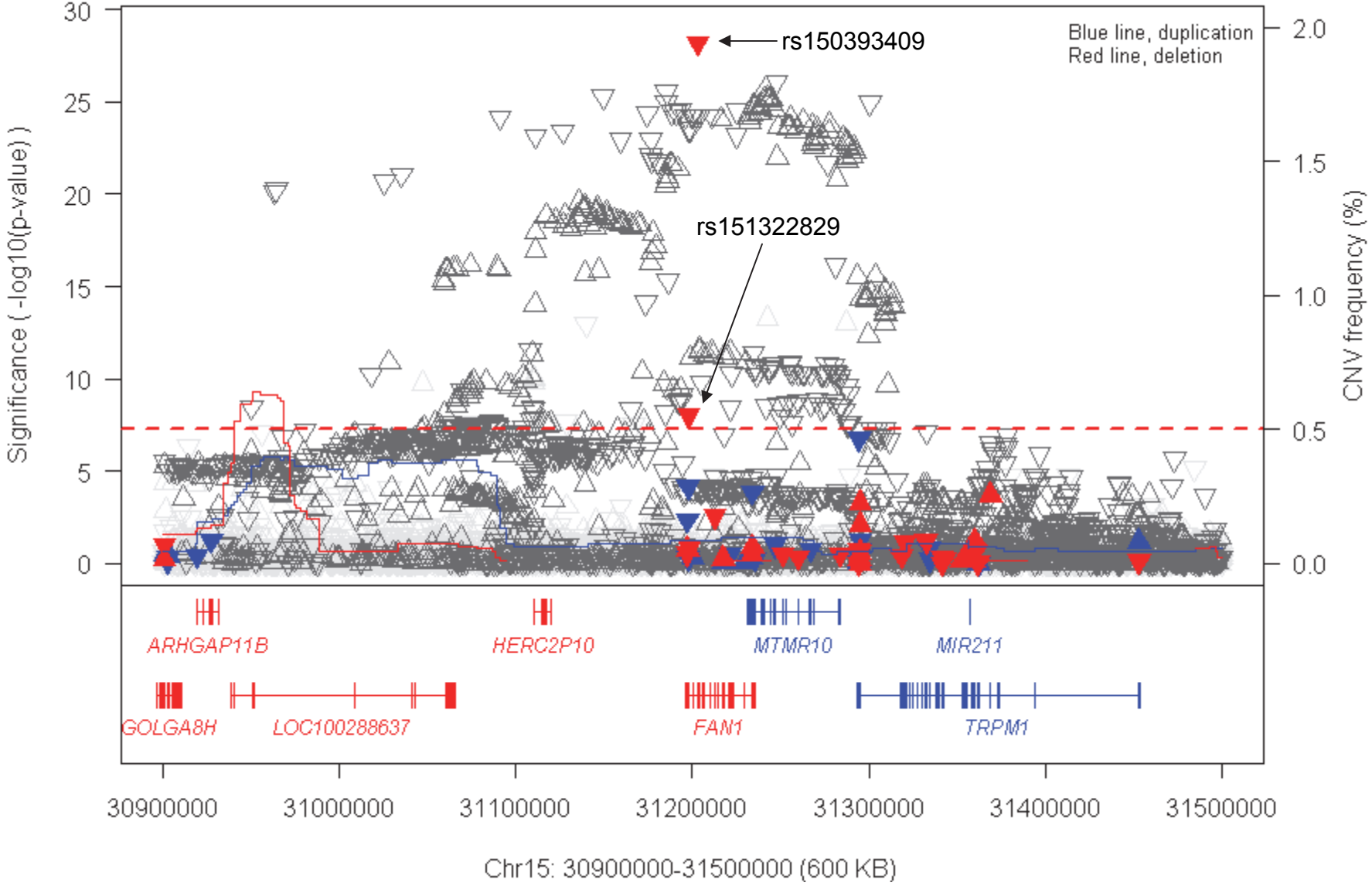

rs150393409
rs151322829
Blue line, duplication
Red line, deletion
Significance (-log10(p-value))
CNV frequency (%)
30
25
20
15
10
5
0
2.0
1.5
1.0
0.5
0.0
ARHGAP11B
HERC2P10
MTMR10
MIR211
GOLGA8H
LOC100288637
FAN1
TRPM1
30900000
31000000
31100000
31200000
31300000
31400000
31500000
Chr15: 30900000-31500000 (600 KB)

Fig 2

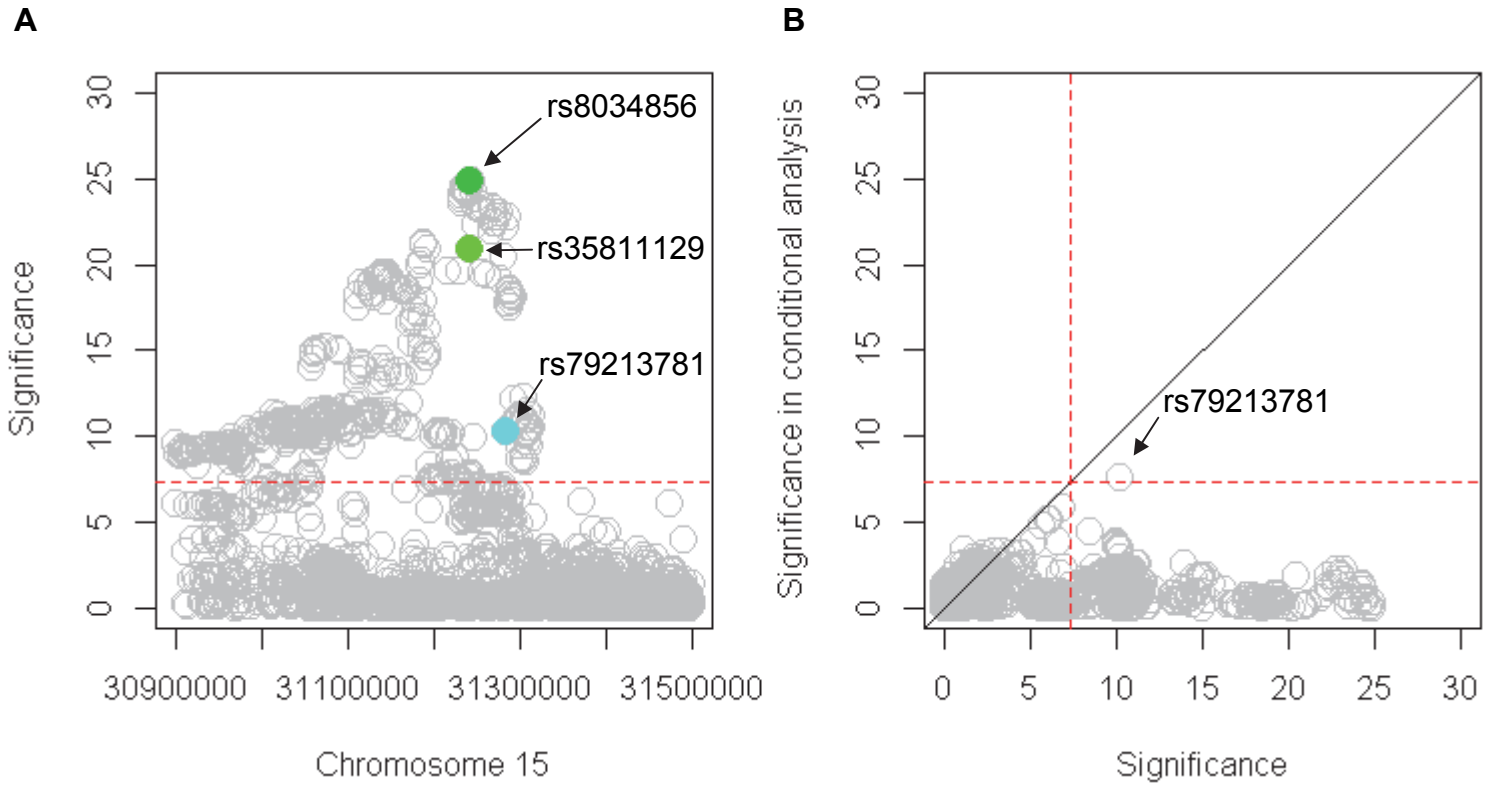

A
B
rs8034856
rs35811129
rs79213781
Significance
Chromosome 15
30900000
31100000
31300000
31500000
0
5
10
15
20
25
30
Significance in conditional analysis
rs79213781
Significance

Fig 3

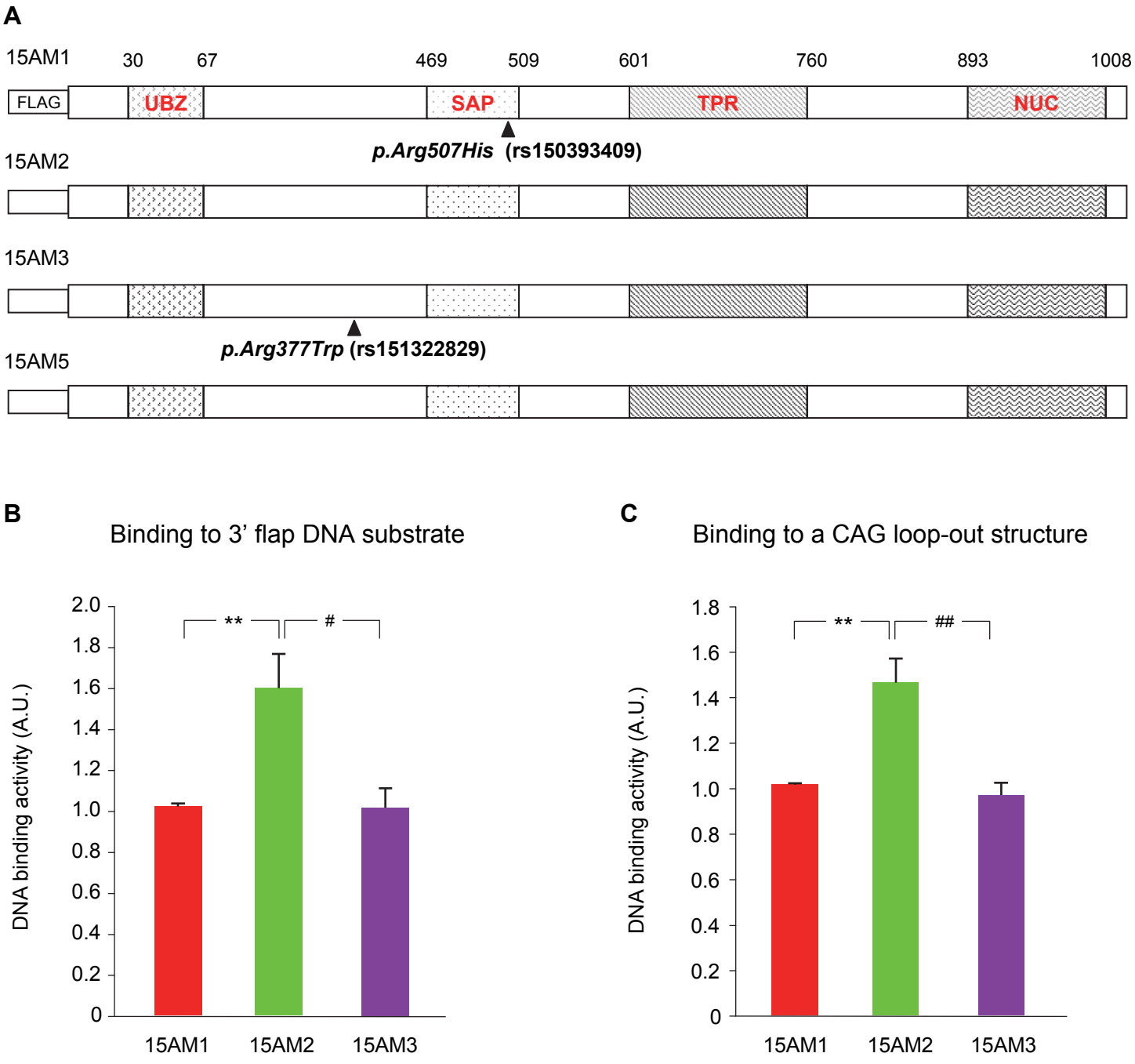

A
15AM1
30
67
469
509
601
760
893
1008
FLAG
UBZ
SAP
TPR
NUC
p.Arg507His (rs150393409)
15AM2
15AM3
p.Arg377Trp (rs151322829)
15AM5
B
Binding to 3' flap DNA substrate
DNA binding activity (A.U.)
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
**
#
15AM1
15AM2
15AM3
C
Binding to a CAG loop-out structure
DNA binding activity (A.U.)
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
**
##
15AM1
15AM2
15AM3

Fig 4

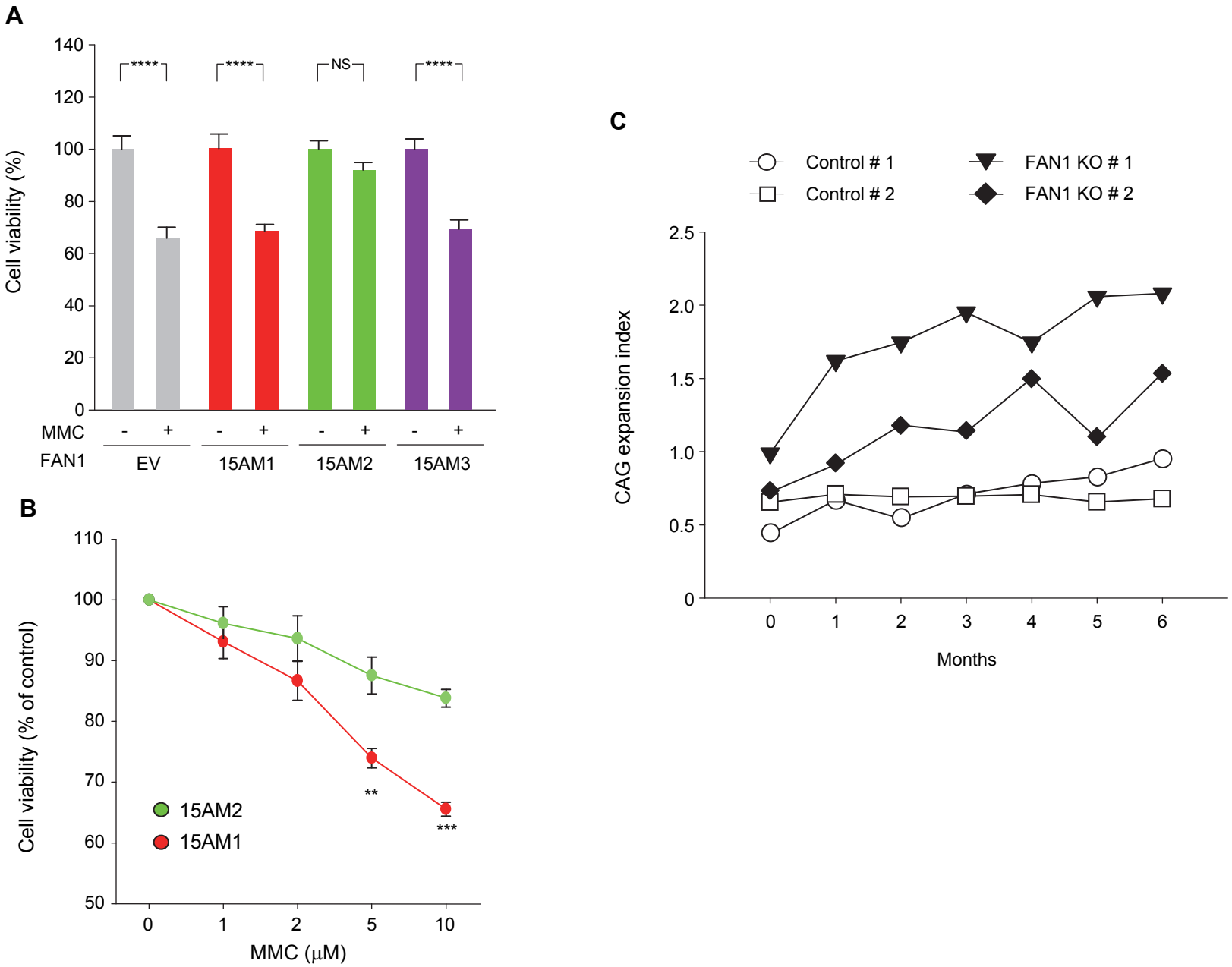
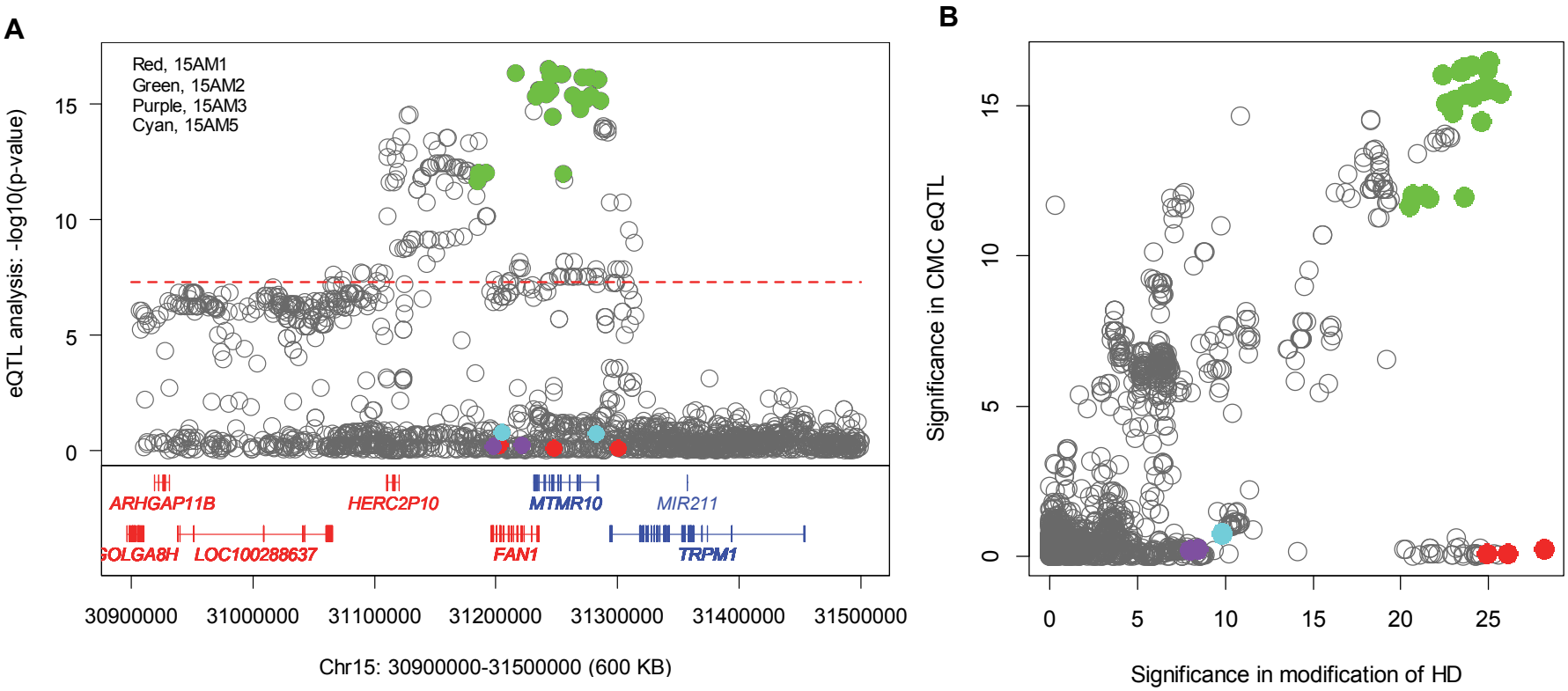

A
Cell viability (%)
140
120
100
80
60
40
20
0
****
****
NS
****
MMC
-
+
-
+
-
+
-
+
FAN1
EV
15AM1
15AM2
15AM3
B
Cell viability (% of control)
110
100
90
80
70
60
50
15AM2
15AM1
**
***
0
1
2
5
10
MMC (μM)
C
Control # 1
Control # 2
FAN1 KO # 1
FAN1 KO # 2
CAG expansion index
2.5
2.0
1.5
1.0
0.5
0
0
1
2
3
4
5
6
Months

Fig 5



C

Co-localization analysis: HD modifier association analysis data and CMC eQTL data				
Posterior probability (%)	15AM1-tagging SNPs	15AM2-tagging SNPs	15AM3-tagging SNPs	15AM5-tagging SNPs
1. Causal variant for HD modifier trait only	92	2.4E-9	89	90
2. Causal variant for CMC eQTL trait only	5.8E-21	5.1E-20	4.6E-4	3.6E-6
3. One common causal variant	8.3	97	8.5	9.8