Genomewide Linkage Scan Reveals Novel Loci Modifying Age of Onset of Huntington’s Disease in the Venezuelan HD Kindreds

Javier Gayán,1,2 Denise Brocklebank,1 J. Michael Andresen,3 Gorka Alkorta-Aranburu,3 The US-Venezuela Collaborative Research Group, M. Zameel Cader,1 Simone A. Roberts,4,5,7 Stacey S. Cherny,6 Nancy S. Wexler,1 Lon R. Cardon,1 and David E. Housman3

1Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom
2Neocodex, Seville, Spain
3Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts
4Department of Neurology, College of Physicians and Surgeons, Columbia University, New York
5Department of Psychiatry, College of Physicians and Surgeons, Columbia University, New York
6Department of Psychiatry, Genome Research Centre, and The State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong, Pokfulam, Hong Kong
7Hereditary Disease Foundation, New York, New York

The age of onset of Huntington’s disease (HD) is inversely correlated with the CAG length in the HD gene. The CAG repeat length accounts for 70% of the variability in HD age of onset. However, 90% of individuals worldwide with expanded alleles possess between 40 and 50 CAG repeat lengths in their HD gene. For these people, the size of their repeat only determines 44% of the variability in their age of onset. Once the effect of the CAG repeat has been accounted for, the residual variance in age of onset is a heritable trait. Targeted candidate gene studies and a genome scan have suggested some loci as potential modifiers of the age of onset of HD. We analyzed the large Venezuelan kindreds in which the HD gene was originally identified. These kindreds offer greater analytic power than standard sib-pair designs. We developed novel pedigree-member selection procedures to maximize power. Using a 5,858-single-nucleotide-polymorphism marker panel, we performed a genomewide linkage analysis. We discovered two novel loci on chromosome 2. Chromosome 2p25 (logarithm of the odds ratio (LOD) = 4.29) and 2q35 (LOD = 3.39) may contain genes that modify age of onset. A third linkage peak on chromosome 6q22 (LOD = 2.48) may confirm the most promising locus from a previous genome scan. Two other candidate loci are suggestive on chromosome 5 (LOD = 3.31 at 5p14 and LOD = 3.14 at 5q32). All these regions harbor candidate genes that are potential HD modifier genes. Finding these modifier genes can reveal accessible and promising new therapeutic pathways and targets to ameliorate and cure HD. Genet. Epidemiol. 32:445–453, 2008. © 2008 Wiley-Liss, Inc.

Key words: Huntington’s disease; modifier genes; genomewide; linkage; large pedigrees; age of onset

INTRODUCTION

Huntington’s disease (HD) is a fatal neurological disorder, characterized by diverse motor, cognitive and psychiatric abnormalities. Severe, uncontrolled movement of all parts of the body, known as chorea, is the most frequent motor disturbance. Decline in intellectual capacities, particularly in memory and executive functioning, and psychiatric symptoms, especially depression, irritability and sometimes schizophrenic-like symptoms, are progressive over time. The disease has a 15–20 year relentless, unremitting course to death.

HD is inherited as an autosomal dominant trait, caused by expansion of an unstable CAG repeat in the HD gene on chromosome 4p16.3 [Gusella et al., 1983; The Huntington’s Disease Collaborative Research Group, 1993]. Fewer than 36 CAG repeats usually produce no observable symptoms, while 36–39 repeats lead to incomplete penetrance. People with 40 or more CAG repeats invariably manifest HD symptoms if they live a normal lifespan. Despite this strong genetic determinism for disease status, there still exists important variability in age of onset and symptomatology, currently the focus of intense investigation [Wexler et al., 2004]. We have carried out a linkage genome scan in extended HD kindreds to discover novel genetic factors which influence the age of onset of HD.

Although HD symptoms usually manifest between 30 and 40 years of age, age of onset is quite variable, ranging...
from early childhood to late adulthood. Not only does the number of CAG repeats in an individual’s HD allele determine whether or not a person becomes ill, but the age of disease onset is also inversely correlated with the number of CAG repeats in that person’s HD allele. Considering all repeat lengths, the number of CAG repeats accounts for about 70% of the variance in age of onset [Duyao et al., 1993]. However, the majority of people with expanded HD alleles, over 90% of individuals worldwide, have between 40 and 50 CAG repeats. For these individuals, the number of CAGs only accounts for approximately 44% of the variability in age of onset [Wexler et al., 2004]. People with 44 CAGs, for example, can vary as much as 20 years or more in age of onset, even within the same family. This variability in age of onset not explained by the size of the CAG repeat is called the residual age of onset. We have demonstrated that the residual age of onset is itself highly heritable, with heritability estimates of 40–80% [Wexler et al., 2004].

Discovering novel genes that may influence HD would be of great benefit for future therapeutic research by revealing druggable targets. Although the precise mutation causing HD has been known since 1993, there remain many unanswered questions about the disease’s pathogenesis. Symptomatic treatments are negligible and cures do not exist.

Research on HD in the last 20 years has advanced our knowledge immensely. The unstable CAG repeat is located in an open reading frame of exon 1 of the HD gene. It is transcribed as a longer than normal polyglutamine chain in the huntingtin protein. Gain of a new toxic function and perhaps loss or compromise of a normal function initiates a cascade of cellular and biochemical disturbances leading to cellular disability and death. It is possible that the discovery of other genetic and environmental factors influencing the age of onset of HD can improve our understanding of the pathology of HD and may lead to the discovery of new therapeutic targets.

Previous research has suggested a variety of genetic loci which may modify the onset or course of HD. For example, there have been genetic association studies of candidate genes [Squitieri et al., 2001; Andresen et al., 2007a]. In addition, a previous linkage genome scan of residual age of onset looking for modifiers in mostly small nuclear families and sibpairs, known as the HD MAPS study, has revealed potential loci [Li et al., 2003].

In this study, we carried out a genomewide linkage scan of residual age of onset of HD, using a dense single nucleotide polymorphism (SNP)-based battery of genetic markers. We performed this analysis in the large kindreds from Venezuela in which the HD gene was originally identified. Our goal was to identify novel genetic loci and provide confirmation of suggested loci that modify the age of onset of HD.

MATERIALS AND METHODS

SUBJECTS

The Venezuelan HD kindreds are the largest known group of families with HD in the world. These families have been described extensively elsewhere [Wexler et al., 2004; Andresen et al., 2007b]. The total Venezuelan HD kindreds comprise over 18,000 individuals. For the present study, we selected a subsample of 1,332 individuals from 45 independent kindreds. One very large family, originating from a single founder, was the source of 928 participants in this study. The remaining 404 participants came from 44 independent kindreds of different sizes: 3 families were of 48–59 individuals, 9 families of 10–22 individuals and 16 families of 3–9 individuals. In all, there were 29 independent kindreds selected and 16 people with HD but unconnected to any known family. Of the 1,332 individuals selected for the study, a subset of 803 people were chosen for genotyping. These included 443 people with one HD allele of 40 or more CAG repeats and a known age of onset. An additional 360 of their relatives, whose HD alleles had 35 or fewer CAGs on both chromosomes, were also genotyped as they provided critical links in the pedigree connections among individuals. For the linkage analyses, we added only familial information for 529 more individuals to create pedigree linkages. We were unable to obtain DNA samples from these people as they had either died or were unavailable, eliminating the possibility of genotyping.

PHENOTYPING

An international, interdisciplinary team has traveled annually to Venezuela since 1979, assessing these family members with an extensive and rigorous battery of neurological, psychiatric and cognitive tests [Wexler et al., 2004]. Age of onset of HD was determined by the appearance of specific motor abnormalities, characteristic of either the adult or juvenile form of the disorder. Cognitive or psychiatric abnormalities are too non-specific and variable to be useful in determining age of onset. To improve accuracy in determining age of onset, affection status was only confirmed after three diagnoses were made, either by three different neurologists in the same year, or three diagnoses in subsequent years by the same or different neurologists. The year in which a person was first diagnosed was marked as the year of onset, and age of onset was determined as year of onset minus year of birth. Age of onset is determined, in part, by the CAG repeat length, and this relationship is best described by a logarithmic curve. We, therefore, created a residual age of onset score as the residual of the linear regression of the natural log of age of onset on the repeat length. This residual age of onset phenotype is an approximately normally distributed score and uncorrelated with the subject’s repeat length.

GENOTYPING AND QUALITY CONTROL

Transformed lymphocyte lines were generated from the blood samples collected from Venezuela. The HD CAG repeat length allele was genotyped [Wexler et al., 2004], and used in this study to create the residual age of onset phenotype.

For the present study, we genotyped 803 individuals with the Illumina linkage IV_v3 SNP panel. Genotyping was carried out at the Broad Institute Center for Genotyping and Analysis. The panel includes 5,858 SNPs scattered across the autosomal and sex chromosomes which were subjected to extensive quality control analysis. The Genotyping Center withheld 70 SNPs (1.2%) as they did not pass quality control filters. An additional 31 SNPs were excluded from analysis because either they were pseudoautosomal (28 SNPs located in both X and Y chromosomes) or had doubtful chromoso-
nal locations (three SNPs). We checked Mendelian inheritance of the remaining 5,757 markers with Pedstats [Wigginton and Abecasis, 2005], Pedcheck [O’Connell and Weeks, 1998] and MERLIN [Abecasis et al., 2002], and familial relationships with ibs_pairs, a batch version of Graphical Representation of Relationships (GRR) [Abecasis et al., 2001].

We also checked the genotyping success rate of samples and SNPs. In general, genotyping was of very good quality. Eighty percent of samples had data on more than 99% of SNPs, and 97% of samples had data on 90% or more of the SNPs. These quality checks suggested 30 pedigree modifications: 23 non-paternalities and seven sample mix-ups or miscodes. We also removed 30 samples because of low-quality data or sample mix-ups, and one chromosome X SNP due to an excess of heterozygous males. Remaining Mendelian inconsistencies were eliminated by wiping out the marker genotypes within the inconsistent family. Finally, one SNP was monomorphic in our sample and 18 SNPs failed a Hardy-Weinberg equilibrium test (we considered a P-value <0.001 significant for our data set). These were also eliminated from analysis.

The linkage map used in these analyses was obtained from Illumina. SNP locations were confirmed with the most recent Illumina panel (4b) and discrepant locations were confirmed with the dbSNP database. All SNPs used in this study have been extensively validated and are evenly distributed across the genome (one SNP every 0.64 cM or 482 kb on average).

ANALYSIS

Linkage analysis relies on the extraction of inheritance information from genetic markers genotyped in members of a pedigree. Larger pedigrees are generally more informative and therefore more powerful to detect even subtle genetic linkage. But their analysis requires complex methodology and substantial computer resources, limiting the type of analysis that is currently feasible. Pedigrees can be subdivided into smaller groups. We used a measure of analysis called a “bit” to aid the partitioning of the families into sub-pedigrees of specified size. Each bit equals twice the number of non-founders minus the number of founders in the pedigree. Genetic software based on the Lander-Green algorithm [such as MERLIN; Abecasis et al., 2002; or Genehunter; Kruglyak et al., 1996] can obtain an exact inheritance solution in pedigrees of small to medium size (maximum size of about 22 bits). Larger pedigrees can be analyzed with software that employs a Markov Chain Monte Carlo (MCMC) algorithm [implemented, for example, in Simwalk2; Sobel and Lange, 1996; or Loki; Heath et al., 1997]. These software packages only provide approximate estimates and are limited by the number of markers that can be analyzed simultaneously.

Inheritance information can be summarized in terms of identity-by-descent (IBD) values for each genetic marker or locus for each pair of relatives in a pedigree. This IBD information can then be related to phenotypic resemblance among relatives for a test of genetic linkage between the phenotype and the genetic locus tested. Although the Venezuelan families in this study were selected because they have the HD gene segregating within them, they can be considered unselected for any segregating genes that influence the residual age of onset. Variance components (VC) linkage analysis of the residual age of onset phenotype is therefore an appropriate and powerful method of analysis.

Our goal was to perform a genomewide linkage analysis of several medium-to-large-sized families (3–59 individuals) and one very large kindred (928 individuals) using a large number of genetic markers (5,737 SNPs). The large kindreds and many markers have limited our choice of linkage analysis methods. For example, Loki, a Markov Chain Monte Carlo-based program, was unable to cope with our full kindreds even on a 4 GB RAM computer. Solar [Almasy and Blangero, 1998] can estimate single marker IBD in the full pedigrees, but the SNP markers we used are quite uninformative for single-point analysis. Unfortunately, Solar cannot compute multipoint IBDs for the full pedigree structure, because some of the relative pair relationships present in our Venezuelan kindreds were not specified in their algorithm. Simwalk2 could perform IBD analysis of the full pedigrees. But it was limited for multipoint analysis to a small number of SNPs (about 11) for the job to run in a reasonable time (40–50 days) on an AMD Opteron 2.6 GHz CPU. Actual running time is quite variable because it also depends on marker distance. Given our dense SNP map, 11 markers cover a variable and generally small genetic distance (2–11 cM). An accurate multipoint analysis may not always be provided by 11 SNPs due to a lack of marker informative-ness. It is too computer intensive, inefficient and insufficiently genetically informative to carry out the genomewide scan with these constraints.

To address these complexities of analysis, we developed a bottom-up protocol to analyze the Venezuelan kindreds at increasing levels of complexity. We began with nuclear families, increased complexity up to 50-bit pedigrees and finally analyzed full pedigrees, using the best possible analysis method available for each level of pedigree complexity. In order to generate maximally informative sub-pedigrees from the original full pedigrees, we developed a novel pedigree partitioning software that breaks large pedigrees into smaller sub-pedigrees of a pre-determined maximum size, while maximizing the inheritance information retained [Broklebank et al., in preparation]. This method is based on graph theory for pedigree representation and configuration. Overall pedigree inheritance information is estimated as a function of the kinship and fraternity coefficients between pairs of relatives. The method can then use two different combinatorial optimization techniques, such as genetic algorithm and simulated annealing, to partition the pedigree into smaller sub-pedigrees of a pre-determined maximum size. The algorithms search the space of sub-pedigree structures within the complexity constraint given, to select a pedigree configuration that maximizes the information preserved in the resulting pedigrees.

Using this software, we can separate the original pedigrees into individual nuclear families although in this case there is no need to maximize information retained because there is a single solution. A more interesting application of the software is to generate, from the original pedigrees, the largest families that can be analyzed by an exact method (about 22 bits), while maximizing the inheritance information retained in the resulting pedigrees. All the original pedigrees would then have a maximum complexity of 22 bits, and can all be analyzed simultaneously, using the same method.

Genet. Epidemiol.
The first and simplest complexity level was to analyze two-generation (2-gen) nuclear families. We then generated sub-pedigrees of different maximum complexities which could be analyzed by an exact method such as that implemented in MERLIN (15, 18 and 22 bits) or by an approximate method as implemented in Simwalk2 (25 and 50 bits). This analysis of increasingly more complex sub-pedigrees allowed us to compare results between different-sized kindreds and different softwares.

The smaller pedigree complexities (2gen, 15, 18 and 22 bits) were analyzed using MERLIN, which provides exact IBD estimates. In MERLIN, all markers on each chromosome can be analyzed simultaneously in a multipoint fashion, and information content (IC), IBD estimation and VC linkage tests were carried out simultaneously in a 2-cM grid. Moreover, MERLIN can model linkage disequilibrium (LD) between markers, which can result in false-positive results, if ignored. LD was modeled by creating clusters joining markers with $r^2$ values larger than 0.2 [Abecasis and Wigginton, 2005]. This option, however, was only possible for pedigrees of up to 18 bits, so the analysis of the 22-bit pedigrees was carried out without modeling LD. Chromosome X was analyzed using MERLIN in X (MIN X), the X-specific version of MERLIN.

For the larger pedigree complexities (25 and 50 bits) and for the full pedigrees (fullped), IBD probabilities were estimated using Simwalk2. They were then imported into the software QTDT [Abecasis et al., 2000] for VC linkage analysis. Simwalk2 is limited in the number of markers that can be analyzed simultaneously. Running time is also affected by the number of markers analyzed and the distance among the markers. Due to these time constraints, we did not perform a genomewide analysis of these larger pedigrees. The 25-bit sub-pedigrees were analyzed only on the chromosomes suggested by the simpler pedigree configurations.

A sliding window of 50 SNPs, with 10 SNPs overlapping between adjacent windows, was used to keep within software limitations and time constraints while minimizing the problem of IBD estimation at the window edges due to drop of information. For the overlapping 10 SNPs, we averaged the linkage test statistic obtained in each window. This strategy produced similar results and less abrupt transitions than the alternative of dropping the five most distal overlapping markers in each window. The analysis of the 50-bit sub-pedigrees was only performed on small windows of variable numbers of SNPs around the linkage peaks identified in the previous pedigree configurations. Similarly, the full pedigrees were only analyzed in small windows of 11 SNPs in selected regions.

Finally, we tested all available SNPs for association with the residual phenotype in the full pedigrees, using the total association and the orthogonal tests in the QTDT [Abecasis et al., 2000]. The original transmission disequilibrium test is a test of both linkage and association based on the observations of alleles transmitted from parents to an affected child. If all affected individuals belong to a single large pedigree descended from a single founder, then the test is just one of linkage. Although most affected family members belong to one large kindred, there are a total of 29 independent kindreds with different founders and 16 single individuals not known to be connected to any of the kindreds. These families are also informative for genetic association studies. The linkage marker panel used in this study, however, provides a very incomplete association coverage, and many potential associations will certainly have been missed.

RESULTS

Phenotypic and genotypic data were stored in a large pedigree datadate from which information was extracted as needed. Input files for the different software packages were created using our own scripts and Mega2 version 3.0 R8 [Mukhopadhyay et al., 2005]. Chromosomes were analyzed individually.

PEDIGREE PARTITIONING

The original pedigrees were subdivided into sub-pedigrees of a set maximum complexity, while maximizing the inheritance information retained. First, the full pedigrees were partitioned into 2gen pedigrees, resulting in 376 nuclear families of between 1 and 13 members, with a maximum complexity of 20 bits. There is only one way of breaking pedigrees into nuclear families. Although these families are simple to analyze they do not necessarily retain inheritance information efficiently.

Next, we separated the full pedigrees into pedigrees of a maximum complexity of 15 bits. Because there are many potential solutions, we analyzed three different configurations with close to maximal information retained to compare results among these different solutions with the same complexity. Results were consistent enough so that only one 15-bit configuration is presented. These 15-bit pedigrees resulted in about 118–125 families of 3–22 members. We also partitioned the full pedigrees into maximum complexities of 18 bits (112 families of 3–22 members), 22 bits (100 families of 3–30 members), 25 bits (86 families of 3–33 members) and 50 bits (60 families of 3–57 members).

INFORMATION CONTENT

Multipoint IC was estimated in a 2-cM grid fashion for all chromosomes using the entropy-based information measure implemented in MERLIN. This analysis was restricted to the smaller pedigree configurations (2gen, 15, 18 and 22 bits) because of MERLIN limitations. IC measures the amount of inheritance information that can be extracted from pedigree data using this particular marker set, compared to an infinitely dense set of markers. Because IC depends on the structure of the pedigrees analyzed, it is not appropriate to compare IC estimates for the different pedigree configurations created in this study. Nonetheless, the IC of each configuration reflects the amount of information extracted for that configuration and can be indirectly compared in that light.

For each pedigree configuration, IC was very stable across all autosomes (Fig. 1), only dropping substantially at the ends of some chromosomes. This drop reflects the lack of information obtained because of the low number or low informativeness of SNPs at some of these chromosome tips. Sometimes the drop is an artifact of the 2-cM grid extending beyond the location of the last SNP in a chromosome. The average IC for nuclear families was only 0.58, but this low IC estimate was due to families without parental genotypes. The average IC in nuclear families was much larger in the subsets where one parent (0.82) or both parents (0.84) were genotyped.
All the remaining pedigree configurations had a similar average IC of 0.84, suggesting that in the presence of missing parental genotypic data, larger extended families can be more efficient at extracting inheritance information than nuclear families. Moreover, the relative IC extracted from pedigrees in the range of 15–22 bits is not very different, although larger pedigrees probably provide more overall inheritance information. For chromosome X, we observed the same trend, with slightly higher IC values. We had an average IC of 0.61 for nuclear families and 0.85 for the 22-bit configuration.

The average IC of 0.84 may be considered low compared to Illumina’s specifications. This result is probably due to the number of subjects with missing DNA in our sample. Almost all of these people had died before our study began. Analyzing larger families may help overcome this problem of missing genotypic data. In conclusion, the Illumina SNP genotypes provide a good coverage of the inheritance information in these pedigrees and should result in an informative linkage analysis.

LINKAGE RESULTS

Linkage results in the smaller pedigree configurations (2gen to 18 bits) did not vary much whether the linkage analysis was performed without modeling LD or with modeling LD using different values of $r^2$ (0.1, 0.2 and 0.35). Because MERLIN could not handle LD modeling for all families in the 22-bit configuration, results for the 22-bit pedigrees ignore marker LD, a reasonable strategy as modeling LD did not seem to affect our linkage analysis significantly.

Although the residual age of onset phenotype is approximately normally distributed, small deviations from normality can affect the significance testing. We have used the permutation simulation option in MERLIN to estimate the empirical significance of our analysis of the 2gen families. In general, we observed only a small drop in significance, where the empirical logarithm of the odds ratio (LOD) score is about 90% the size of the observed LOD score (an observed LOD = 4.0 corresponds to an empirical LOD = 3.8; 3.5 to 3.4; 3.0 to 2.72; 2.5 to 2.22; 2.0 to 1.75; 1.5 to 1.34). Because we were unable to run these simulations for large pedigree complexities, we have reported observed LOD scores throughout the paper.

Figure 2 presents the genomewide linkage results for the nuclear families (2gen), 15- and 22-bit configurations. Pedigrees can be subdivided into smaller groups. We used a measure of analysis called a “bit” to subdivide the families. Each bit equals twice the number of non-founders minus the number of founders in the pedigree. To improve the figure’s clarity, we excluded the 18-bit and two of the 15-bit configurations. As expected, the location of linkage peaks is consistent for the different-sized pedigree configurations. An unexpected finding, nonetheless, was that the maximum LOD score for the different pedigree configurations was sometimes very variable. This variability may arise because families that are informative for linkage may be partitioned differently for some configurations. Or it may suggest that the LOD score is due, in part, to variable random effects.

We found LOD scores > 2 on chromosomes 2, 4, 5, 6, and 12. In addition, there were LOD scores >1.5 on chromosomes 10, 16 and 19. The most promising chromosomes were then analyzed using the larger pedigree configurations. Again, to facilitate clarity of the figures, all figures of single chromosomes display only the 2gen, 15-, 22-, 50-bit and full pedigree configurations. We excluded two of the 15-bit, as well as the 18-bit and 25-bit configurations, although they were all analyzed. All linkage figures also mark the linkage peaks found by a previous genome scan using an independent sample [Li et al., 2003]. Chromosomes 2 and 4–6 may demonstrate some overlaps in linkage peaks with HD MAPS.

The short arm of chromosome 2 harbors the largest genomewide LOD score (maximum LOD = 4.29 for the 15-bit configuration at 7.75 cM). Figure 3 shows that this linkage peak is consistent for all the different pedigree configurations. The LOD score varies, however, with different pedigree configurations (LOD range = 2.18–4.29). For the full pedigrees, the LOD score is 4.01 at 9.07 cM (marker rs1024026). Two families with extreme phenotypic values make large contributions to the LOD score in this region. This may explain, in part, the large LOD score
differences for the various pedigree configurations. Nonetheless, even if these two families potentially bias the LOD score upward, there would still be strong evidence for linkage in this region, even if these particular families had a more moderate effect. The one-LOD support interval for this locus extends approximately from 4 to 14 cM.

There is a second large linkage peak on the long arm of chromosome 2 (maximum LOD = 3.39 at 213.81 cM for the full pedigrees; LOD range = 1.27–3.39). This linkage peak is promising because the LOD score increases as larger pedigree configurations are analyzed, as would be expected theoretically. This linkage curve is narrow, especially on the distal side, with a one-LOD support interval between 206 and 216 cM.

We found two distinct linkage peaks on chromosome 4. There is a linkage peak (maximum LOD = 2.07 for nuclear families; range = 1.12–2.07) at 85–87 cM. However, the LOD score for the full pedigrees was only 1.18 at 90 cM. Chromosome 4 also exhibits a second, smaller peak (maximum LOD = 1.49 for the 15-bit configuration; range = 0.00–1.49) at 1.3 cM. This peak is close to the actual HD gene (4 cM), but the linkage signal is not supported when the full pedigrees are analyzed.

The short arm of chromosome 5 (Fig. 4) displays two consecutive linkage peaks, separated by about 20 cM. A first peak is at 13–19 cM (maximum LOD = 2.67 for nuclear families; range = 1.11–2.67), and a second peak is at 40 cM (maximum LOD = 3.31 for the full pedigrees; range = 1.62–3.31). It is difficult to establish whether these two peaks are an artifact and only one true peak exists or whether there are two potentially different loci close together. The IC at this region does not vary much, so the distinct peaks do not seem to be due to lack of inheritance information. They may suggest the actual presence of multiple loci. Chromosome 5q exhibits another candidate locus (LOD = 3.14 at 147 cM for the full pedigrees).

On chromosome 6 (Fig. 5) there are again two to three consecutive peaks. The first peak is small (maximum LOD = 1.10 at 90 cM for the 18-bit configuration). The latter two peaks are so close (114–128 cM) that they may reflect the same genetic locus (maximum LOD = 2.48 at 123 cM for 50-bit configuration, range = 0.98–2.48). Even though the IC remains constant in this region, LOD scores are quite variable, changing with each pedigree configuration. Surprisingly, the full pedigree analysis only yielded a LOD = 0.98. The presence of several peaks resulted in a wide one-LOD support interval, extending from 108 to 134 cM.

We found a maximum LOD = 2.86 at 82 cM on chromosome 12 for the nuclear families configuration. The size of the LOD score in this region also varies considerably depending on the different pedigree configurations (range = 0.71–2.86). For the full pedigrees, the LOD score was 0.71 at 81 cM.

There is a double-peak linkage curve for chromosome 19, with maximum LOD scores of 1.79 at 22 cM and 1.68 at 28–30 cM. They are sufficiently close to probably represent one single genetic locus, with a maximum LOD of 0.90 when the full pedigrees are analyzed.

There are also two other linkage peaks worth noting. On chromosome 10, a maximum LOD of 1.49 was found at
1.18 cM for the 22-bit configuration, while on chromosome 16, a maximum LOD of 1.71 was obtained at 127.40 cM for one of the 15-bit configurations. Finally, all chromosome X LOD scores were smaller than 0.34.

There was some evidence for association at the linkage peaks on chromosomes 2q, 4 and 6, but these could be chance results. After correction for multiple testing, none of the association results at the SNPs within the linkage curves were statistically significant.

**DISCUSSION**

We carried out a whole genome linkage analysis to identify genomic regions containing novel modifier genes influencing HD’s age of onset. We found four chromosomal regions, two on chromosome 2, one on chromosome 5 and one on chromosome 6 where there is promising evidence to find these modifier genes (Table I). We also found intriguing loci on chromosomes 4 and 12, and perhaps on chromosomes 10, 16 and 19.

Our key discovery is a significant linkage (LOD = 4.29) on chromosome 2p25, a novel genetic locus not previously implicated in HD. We found consistent evidence for linkage in this region for the different pedigree configurations (LOD range = 2.18–4.29). The maximum LOD = 4.29 was derived from analyzing the 15-bit configurations. The full pedigree maximum LOD score is 4.01. Two families with extreme phenotypic values tended to have large contributions to this LOD score, which fits well with a major gene effect. One family is characterized by an unusually late age of onset, with onset between 56 and 69 years. This family had a large positive, residual age of onset phenotype. Family members had onset much later than would be expected by their CAG repeat length, suggesting the presence of a shared family protective factor. The other family tended to have earlier disease onset, between 26 and 34 years old. They had negative residual scores, signifying an earlier disease onset than would be predicted by their repeat lengths. This could be due to a shared family risk factor. These two families contributed 1.78 and 0.82 LOD units to the LOD score of 4.29 for the 15-bit configuration.

Even if these families had more moderate effect sizes (0.60 LOD units), there would still be evidence for linkage in this region, suggesting that this locus represents a general population effect. This genomic region is surprisingly gene poor, with only seven genes within 3 Mb of the peak at marker rs1024026, and 10 other genes in the approximate 5 Mb that cover the one-LOD support interval.

There is a second region of interest on chromosome 2, with a maximum LOD score of 3.39 at 2q35. Although the maximum LOD score is smaller than in the 2p25 region, the LOD score increases as larger pedigree configurations were analyzed. The largest LOD score was obtained for analysis of the full pedigrees, as is predicted theoretically. Although the actual gene responsible for this linkage signal cannot be located accurately in this current analysis, there are some potential candidate genes in the 1 Mb around marker rs1851328.

Another interesting candidate region is 6q22, where we obtained a maximum LOD = 2.48 for the 50-bit configuration. The linkage evidence at this locus is complicated by several close peaks for the different pedigree configurations. For example, there is another linkage peak (LOD = 2.45) just 4 cM distal from the main linkage peak (around 127 cM), where three consecutive SNPs are in very high LD, with $r^2$ values 0.88–0.99. Nonetheless, linkage results in this region were very similar with and without LD modeling, suggesting that this linkage peak is not an artifact. Although this linkage peak is broad, there are again some potential candidate genes in the region around marker rs941815, where the main linkage peak is located.

The short arm of chromosome 5 displays two consecutive peaks with LOD scores above 2.5. More confidence can be placed on the second peak (5p14), where a LOD score of 3.31 was obtained with the full pedigrees, and all configurations yielded LOD scores above 1.5. (Only the larger peak on chromosome 5p is listed on Table I.) Another interesting peak, this time on the q arm of chromosome 5, was revealed in the analysis of the full pedigrees (LOD = 3.14).

The linkage results on chromosomes 4 and 12 are largely driven by the smaller pedigree configurations, while the largest configuration yielded LOD scores smaller than 1.18. Finally, there are some weaker suggestions of linkage on chromosomes 10, 16 and 19. These other loci may become more interesting if they are replicated in other genome scans or are implicated by candidate gene association studies.

Some of the linkage peaks in this study are located relatively close to linkage peaks in the only other genome scan for residual age of onset of HD [Li et al., 2003]. For

<table>
<thead>
<tr>
<th>Chr.</th>
<th>cM location</th>
<th>Megabase location</th>
<th>Marker</th>
<th>Maximum LOD</th>
<th>Fullped LOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2p25</td>
<td>9 (4–14)</td>
<td>3.6 (1.5–6.3)</td>
<td>rs1024026</td>
<td>4.29</td>
<td>4.01</td>
</tr>
<tr>
<td>2q35</td>
<td>214 (206–216)</td>
<td>217.0 (208.8–218.2)</td>
<td>rs1851328</td>
<td>3.39</td>
<td>3.39</td>
</tr>
<tr>
<td>4p16</td>
<td>1 (0–9)</td>
<td>1.0 (0.0–5.5)</td>
<td>rs935971</td>
<td>1.49</td>
<td>0.00</td>
</tr>
<tr>
<td>4q21</td>
<td>87 (79–91)</td>
<td>78.9 (67.5–84.5)</td>
<td>rs1566485</td>
<td>2.07</td>
<td>1.18</td>
</tr>
<tr>
<td>5p14</td>
<td>40 (34–46)</td>
<td>20.5 (15.3–25.7)</td>
<td>rs720189</td>
<td>3.31</td>
<td>3.31</td>
</tr>
<tr>
<td>5q32</td>
<td>147 (144–154)</td>
<td>145.4 (142.5–149.6)</td>
<td>rs248771</td>
<td>3.14</td>
<td>3.14</td>
</tr>
<tr>
<td>6q22</td>
<td>123 (108–134)</td>
<td>123.9 (105.6–134.5)</td>
<td>rs941815</td>
<td>2.48</td>
<td>0.98</td>
</tr>
<tr>
<td>12q15</td>
<td>81 (72–90)</td>
<td>65.6 (55.1–75.9)</td>
<td>rs956004</td>
<td>2.86</td>
<td>0.71</td>
</tr>
<tr>
<td>19p13</td>
<td>23 (18–35)</td>
<td>7.0 (5.0–14.0)</td>
<td>rs966591</td>
<td>1.79</td>
<td>0.90</td>
</tr>
</tbody>
</table>

LOD, logarithm of the odds ratio.
some of these overlapping linkage regions, neither study has enough evidence to claim linkage independently, but the consistency of findings in the two studies makes these regions intriguing. For example, on chromosome 2q HD MAPS found a LOD = 1.6 at approximately 198 cM, 16 cM proximal from the peak found in our current study. On chromosome 6q, HD MAPS identified a region of linkage with a LOD = 2.28 at 147 cM, about 20 cM distal from our peak. In a follow-up study, the LOD score in this region has increased to over 4 [Li et al., 2006]. Furthermore, some of the linkage peaks on chromosomes 4 and 5 are in the vicinity of smaller LODs reported in the HD MAPS genome scan.

It is difficult to know if we have true replications of the previous linkage findings because of the distance between peaks across studies. The summits of the linkage peaks on chromosomes 2, 4, 5 or 6 are about 2–22 cM apart, just about close enough not to be dismissed. The linkage curves are also generally wide, allowing for some potential overlap of the linkage curves between the two studies. The use of different types of markers (microsatellites versus SNPs) in different populations may have shifted the relative location of the linkage peaks across the two studies. Although we cannot be certain of calling these consistent results true replications, it seems likely that at least some might reflect the same underlying genetic locus.

It is important to note here the relationship between age of onset and residual age of onset. It is well known that the CAG repeat length explains about 70% of the variance in age of onset, and it has been reported more recently that the heritability of the residual age of onset is in the 40–80% range [Wexler et al., 2004]. Consequently, the unexplained genetic variance influencing age of onset may be 12–24% of the total variance in age of onset [Wexler et al., 2004]. The loci identified in this study, even though they display large LOD scores for the analysis of residual age of onset, may therefore only explain a few percentage of the total variance in age of onset.

A major asset of this study is the great analytical power contributed by the large Venezuelan extended families, in which the HD mutation was originally discovered. These large kindreds are valuable resources for genetic studies. With their rich set of familial relationships they yield a much more powerful design than the standard sib-pair study. In addition to being great resources, these large and complex kindreds pose unique analytic complications which we have resolved by developing new analytical strategies and methods. Our new pedigree-breaking algorithm allows for dividing large pedigrees into smaller sub-pedigrees of a determined complexity. This strategy can be helpful when analyzing large pedigrees with a particular method, such as an exact inheritance estimation, or simply to reduce the amount of time required for a long analysis of very large families. Especially important is that the large pedigrees are subdivided while maximizing the inheritance information retained, thereby optimizing their power for linkage analysis. This new algorithm that we have developed makes possible an economical and efficient genomewide linkage analysis of large families.

The explosion of SNP genotyping technology has made SNP-based linkage and association studies possible. A potential bias for the new wave of SNP-based linkage studies is LD among markers [Abecasis and Wigginton, 2005], as it is reasonable to expect some level of LD among the markers in the relatively dense SNP linkage panels. For example, about 0.1% of the marker-marker LD comparisons had $r^2$ values larger than 0.2. To control for this potential bias, markers in high LD can be excluded from analysis. Recently, MERLIN has implemented the modeling of marker LD by clustering markers within a specified distance or with a specified value of $r^2$. MERLIN then assumes that there is no recombination within clusters and no LD between clusters.

We carried out linkage analyses in some of our smaller pedigree configurations both without modeling LD and with modeling LD using different values of $r^2$ (0.1, 0.2 and 0.35). Interestingly, the linkage results were quite consistent across all types of analyses, with no large decreases in LOD scores. This suggests that marker LD was not markedly biasing results in our analyses. This could be perhaps because the linkage regions identified did not include markers in high LD (all $r^2$ values <0.2, except for the chromosome 6 linkage peak). It could also be that the large size of our families protects against parental missing data and diminishes the potential bias of marker LD. As results did not vary substantially, we carried out the MERLIN analyses with an $r^2$ value of 0.2. This strategy provided a good compromise between controlling for LD and not discarding too much data because of obligate recombinants.

There are several caveats about the present study. In order to carry out a linkage analysis of very large nuclear families and complete kindreds, we initially broke these large pedigrees into smaller sub-pedigrees, thereby wasting familial relationships and inheritance information. This loss of information may bias the linkage results, particularly as the pedigrees are broken into smaller units, because it can result in a loss of power. For example, large pedigrees make use of phase information to estimate IBD more accurately, which results in a more powerful linkage analysis. Nonetheless, our novel pedigree-breaking software allows the partitioning of large pedigrees while maximizing the inheritance information retained, so that this loss of power is minimized. In addition, as pedigrees are partitioned, some individuals need to be duplicated in independent sub-pedigrees, and are included in the analysis more than once. We also need to exercise caution interpreting results obtained with Simwalk2 [Sobel and Lange, 1996]. This genetic software employs a probabilistic algorithm and may not reach the correct solution, especially when analyzing large and complex families.

These concerns notwithstanding, we created a pedigree-breaking software algorithm that maximizes information retained so that the pedigrees were broken efficiently. Even more critically, we compared results obtained in the analysis of sub-pedigrees of different complexities. The location of the same linkage peaks, derived with different pedigree structures and different methods of analysis, leads us to be quite confident of the veracity of these results.

In summary, we discovered two novel potential loci on chromosomes 2p25 and 2q35 which may harbor genes modifying the age of onset of HD. We have possibly also replicated one other locus, on 6q22, which was suggested in a previous genome scan. In addition, we found suggestive evidence for novel loci on 5p14 and 5q32, and on many other chromosomes.

As well as opening up new vistas for genetic research to identify the modifiers themselves, our results emphasize and reinforce the value of studying large, extended families for linkage analyses. Just as our identification of...
the HD gene, in 1993 [The Huntington’s Disease Collaborative Research Group, 1993], in the Venezuelan HD kindreds generalized to all families with HD worldwide, it is likely that modifiers discovered by studying these families will generalize as well. We are hopeful that finding these modifiers of HD age of onset will enable us to elucidate novel pathways and drug targets. Our goal is to extend the onset of Huntington’s disease beyond the ordinary lifespan or prevent it from ever appearing, in order to extinguish it altogether.

ACKNOWLEDGMENTS

We are indebted to the Venezuelan families who have participated in this project throughout all these years. Their generous collaboration has been critical to the success of this project and many others. We thank John Broxholme for invaluable computing assistance and Daniel B. Mirel for help managing the genotyping. The Broad Institute Center for Genotyping and Analysis is supported by grant U54 RR0278-01 from the National Center for Research Resources, NIH.

REFERENCES


