October, 2006

Genetic evidence for ubiquitin-specific proteases USP24 and USP40 as candidate genes for late-onset Parkinson disease

Yonghong Li
Steven J Schrodi
Charles Rowland
Kristina Tacey
Joseph Catanese, et al.

Available at: http://works.bepress.com/steve_schrodi/7/
RESEARCH ARTICLE

Genetic Evidence for Ubiquitin-Specific Proteases USP24 and USP40 as Candidate Genes for Late-Onset Parkinson Disease

Yonghong Li, Steven Schrodi, Charles Rowland, Kristina Tacey, Joseph Catanese, and Andrew Grupe
Celera Diagnostics, Alameda, California
Communicated by Johannes Zschocke

INTRODUCTION

Genetic studies have identified several loci that play a role in conferring risk to, or affecting age at onset (AAO) of Parkinson disease (PD; MIM #168600). The regions on chromosomes 1q32 and 2q36–q37, designated as the PARK10 (MIM #606852) and PARK11 (MIM #607688) loci, respectively, have been reported by several studies [Hicks et al., 2002; Li et al., 2002; Pankratz et al., 2003]. The PARK10 locus appears to affect both disease risk and AAO, while the PARK11 locus has only been implicated in disease risk. A recent association study showed that several candidate genes in the PARK10 locus are associated with AAO and/or risk [Oliveira et al., 2005]. Another study that was based on a ~200,000 marker genome scan identified a single nucleotide polymorphism (SNP) that maps to the PARK11 locus [Maraganore et al., 2005], but this reported association remains controversial [Farrer et al., 2006; Goris et al., 2006; Li et al., 2006].

To fine-scale map within these loci, we targeted SNPs in a biological candidate gene under the chromosome 1 linkage peak and, at the same time, performed a small scale SNP positional scan preferentially targeting potential functional variants across the chromosome 2 linkage region. Here we report the results of these experiments, which allowed us to identify multiple SNPs in two members of the ubiquitin-specific protease family, USP24 and USP40, that are significantly associated with risk for late-onset PD in the analyzed sample set.

MATERIALS AND METHODS

Study Samples

Initially, we constructed a case-control sample set with 224 PD cases and 186 population/convenience controls that were available through the National Institute of Neurological Disorders and Stroke (NINDS) Human Genetics Resource at the Coriell Institute (Camden, NJ; http://ccr.coriell.org/ninds/) (Table 1). All cases and controls were Caucasians, and cases had an AAO of ≥50 years. This sample set included cases and controls sampled at the National Institute of Neurological Disorders and Stroke (NINDS) Human Genetics Resource at the Coriell Institute (Camden, NJ; http://ccr.coriell.org/ninds/).

The Supplementary Material referred to in this article can be accessed at http://www.interscience.wiley.com/jpages/1059-7794/suppmat.

Received 8 March 2006; accepted revised manuscript 15 May 2006.
*Correspondence to: Yonghong Li, Celera Diagnostics, 1401 Harbor Bay Parkway, Alameda, CA 94502.
E-mail: yonghong.li@celeradiagnostics.com
DOI 10.1002/humu.20382
Published online in Wiley InterScience (www.interscience.wiley.com).
similar ages but was not well matched for gender. When more samples became available through the Coriell Institute, we expanded our case-control sample set using a stricter matching scheme. The expanded case-control sample set consisted of 311 late-onset PD cases and 311 age- and gender-matched controls (Table 1). A total of 272 pairs had identical gender and age at sampling; 39 other pairs had identical gender but age at sampling was within a 3-year interval, with most cases younger than controls. As a result of this stricter matching scheme, 23 cases and 37 controls from the initial sample set were not included in the expanded sample set. The expanded sample set thus included 201 cases and 149 controls from the initial sample set and an additional 110 cases and 162 controls. No cases in the initial or the expanded sample set carried the p.Gly2019Ser substitution in the LRRK2 gene, which is known to cause sporadic PD [Gilks et al., 2005; Lesage et al., 2006; Ozelius et al., 2006].

Genotyping

DNA samples were individually genotyped by allele-specific real-time PCR as previously described [Li et al., 2004]. Genes investigated included USP24 (GenBank: XM_371254.3), USP40 (GenBank: NM_018218.1), and PCSK9 (GenBank: NM_174936.2).

Statistical Analysis

Tests for allelic association of individual markers with disease status were carried out using the \( \chi^2 \) test. For multiple SNP analyses, SNPs within the USP24 or the USP40 region were examined using methods loosely based on those described by Cordell and Clayton [2002] to evaluate the relative importance of these SNPs. Specifically, a forward stepwise logistic regression procedure was performed, using log-additive main effects of the SNPs as potential variables in the model, requiring a model p-value < 0.1 as a condition for entry into the model.

Logic regression [Ruczinski et al., 2003], a generalized regression methodology applicable when predictor variables are binary, was used to obtain a “best” multiple SNP logistic regression model. A logic regression model can be defined as

\[
g(E \mid X) = B_0 + \sum_{i=1}^{k} B_i L_i(X),
\]

where \( g \) is an appropriate link function, \( Y \) is a phenotype trait, \( X \) are the covariates, \( B_0, B_1, \ldots, B_k \) are the parameters, and \( L_i(X) \) are Boolean combinations of the covariates (for example \[NOT X1 AND (X2 OR X3)\]). Each \( L_i \) is considered to be a logic tree. The example tree has three leaves (X1, X2, and X3). The analysis was limited to models with one tree, and 10-fold cross-validation was used to estimate the tree size (number of leaves) most likely to give a robust model when applied to data other than that upon which it was built. The best model of the chosen size was estimated, followed by a permutation test to determine if the chosen model gives a result better than could be expected by chance. The logic regression package’s simulated annealing algorithm was used throughout assigning the following program parameters: a starting temperature of 1,000, ending temperature of 0.01, and 200,000 total iterations of the annealing chain.

Population Stratification Testing

Seventy-eight autosomal SNPs were selected to test our expanded sample set (the 311 pairs of cases and controls) for the presence of population stratification. Several criteria were used to select these SNPs in an attempt to increase the likelihood that they were not associated with disease status (i.e., null SNPs): (1) SNPs could not cause amino acid residue changes (according to the current Celera Discovery System assembly), (2) SNPs were required to have existing ref SNP (rs) numbers, (3) the physical distance between any two adjacent SNPs had to be larger than 2 Mbp to reduce the chance of substantial linkage disequilibrium (LD) between sites, (4) SNPs could not be located within reported linkage regions for PD (the following chromosomal arms were excluded due to reports of linkage from OMIM: 1p, 2p, 2q, 4p, 5q, 6q, 9q, 17q, and 18p), and (5) minor allele frequencies were above 10% (internal data from 900 normal white samples) to increase the level of information and drastically reduce the probability that the SNP was monomorphic (and hence uninformative) in our sample set.

Case and control genotypes at these SNPs were used in three separate analyses. To test for the existence of population stratification, we first examined the p-value distribution of the SNPs genotypic association (Williams-corrected G-test, [Sokal and Rohlf, 1995]). Well-designed studies will exhibit p-value distributions that are approximately uniformly-distributed for null markers. However, population stratification will typically inflate the number of significant markers, thereby skewing the distribution of p-values toward smaller values. For several different significance levels \( \alpha \), we calculated the p-value for the number of significant markers at \( \alpha \) (\( S_\alpha \)) using the binomial formula in Eq. 2 to examine this potential effect.

\[
p[S_\alpha ≥ s] = \sum_{i=1}^{28} \binom{28}{i} \alpha^i (1-\alpha)^{28-i}.
\]

Substantial population stratification will inflate the number of significant markers and generate a significant p-value. The second

<table>
<thead>
<tr>
<th>TABLE 1. Sample Set Statistics*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial sample set</strong></td>
</tr>
<tr>
<td><strong>Case</strong></td>
</tr>
<tr>
<td>No. subjects</td>
</tr>
<tr>
<td>% Female</td>
</tr>
<tr>
<td>AAO (± SD)</td>
</tr>
<tr>
<td>AAO range</td>
</tr>
<tr>
<td>AAS (± SD)</td>
</tr>
<tr>
<td>AAS range</td>
</tr>
</tbody>
</table>

*A total of 272 cases/controls paired by identical gender and AAS and 39 by identical gender and AAS ± 3 years, with most cases younger than controls. AAO, age at onset in years; AAS, age at sampling in years; SD, standard deviation; n/a, not applicable.*
test assessed the departure from expected heterozygosity under Hardy-Weinberg equilibrium (HWE) across all 78 SNPs as

$$W = \sum_{i=1}^{78} \left( \frac{H_{\text{exp},i} - H_{\text{obs},i}}{H_{\text{exp},i}} \right)^2,$$

where $H_{\text{obs},i}$ is the observed proportion of heterozygotes for the $i^{th}$ SNP and $H_{\text{exp},i}$ is the expected proportion of heterozygotes under HWE such that $H_{\text{exp},i} = 2p_i(1-p_i)$, for allele frequencies $p_i$. The reader may note that $W$ is related to inbreeding coefficients from classical population genetics [Hartl and Clark, 1989]. A simple Monte Carlo simulation was used to generate a null distribution for $W$ assuming HWE where allele frequencies are drawn from a beta distribution having parameters estimated from empirical data.

Last, we wanted to attempt to correct our association results for any possible population stratification in a genomic control type of approach [e.g., Devlin and Roeder, 1999]. The approach used is similar to that described by Schrödi [2005]. With this method, a null distribution is obtained by fitting a gamma density to the empirical distribution of negative log-transformed $p$-values. That is, a null model is derived from the null markers and putative candidate SNPs are tested against this new null model to adjust for diffuse effects of population stratification. A $p$-value for a test SNP, adjusted for diffuse effects of population stratification, is calculated through integration of the gamma density.

$$P_{\text{adj}} = \int_U^{\infty} \left( \frac{\Gamma(c)}{\Gamma(b)} \right)^{-1} \frac{1}{b!} c^{-b} e^{-x} dx,$$

where $U = -\ln(p$-value test SNP), $\Gamma(\bullet)$ is the gamma function, and parameters $b$ and $c$ are obtained empirically from the distribution of negative log-transformed $p$-values for the null markers through the use of moment matching estimators, $b = \frac{c}{\mu}$ and $c = \frac{\mu^2 + s^2}{\mu}$, where $\mu$ and $s^2$ are the mean and sample variance of the negative log-transformed null marker $p$-values.

## RESULTS

### Exploratory Association Study in Linkage Regions on Chromosomes 1 and 2

As misfolding of proteins and dysfunction of the ubiquitin-proteasomal pathway are pivotal to PD pathogenesis [Dawson and Dawson, 2003; Giasson and Lee, 2003], we hypothesized that genetic variants in USP24, located within the chromosome 1 linkage peak, might be associated with PD risk. To test this hypothesis, we genotyped two SNPs in USP24 in our initial sample set; rs1165222:T$\rightarrow$G and rs13312:C$\rightarrow$T (the average $D^2$ across the region is 0.87, CEPH Caucasian samples), which also includes two other genes (USP40 and LOC388635; Fig. 2A). We genotyped 25 additional SNPs in the USP24 region and identified 16 SNPs with $p < 0.05$ (Fig. 2A; Supplementary Table S3). All of these significant SNPs, except for rs12130799:A$\rightarrow$G, were in moderate to strong LD with each other in our sample ($r^2 > 0.2$, Supplementary Table S4).

Similarly, we estimated that the significant USP40 marker was located in a 350-kbp region of strong LD between rs7563345:G$\rightarrow$A and rs13394720:T$\rightarrow$C (the average $D^2$ across the region is 0.79, CEPH Caucasian samples), which also includes three other genes (DGKD, LOC388635; Fig. 2A). We genotyped 37 SNPs within this region and found three additional SNPs to be significantly associated with PD ($p < 0.05$) (Fig. 2B, Supplementary Tables S5 and S6).

### Fine-Scale Mapping Near Disease-Associated Markers in USP24 and USP40

In an effort to identify potentially causative variants, we performed additional genotyping of SNPs within the region of strong LD that contained the original significant markers in our initial sample set. We included both tagging markers ($r^2 < 0.8$, based on the International HapMap project (http://www.hapmap.org) data) and SNPs that are likely to impact the function of these genes, such as missense and predicted TFBS variants. Based on HapMap data, we estimated that the two significant SNPs in USP24 are located in a 200-kbp region of strong LD that is flanked by rs557435:G$\rightarrow$A and rs2802860:C$\rightarrow$T (the average $D^2$ across the region is 0.91, CEPH Caucasian samples), which also includes two other genes (DGKD and LOC388635; Fig. 2A). We genotyped 25 additional SNPs in the USP24 region and identified 16 SNPs with $p < 0.05$ (Fig. 2A; Supplementary Table S3). All of these significant SNPs, except for rs12130799:A$\rightarrow$G, were in moderate to strong LD with each other in our sample ($r^2 > 0.2$, Supplementary Table S4).

### Validation of the Above Observed Association in an Expanded Late-Onset PD Sample Set

To further confirm the above associations, we then constructed a larger sample set after additional cases and controls became available through the NINDS Human Genetics Resource.
predicted TFBS variants (rs555687:C→B
4

USP40

markers were located in the
proteinase K subfamily. All four significant chromosome 2
markers, on average, were expected to have p-values below 0.07 in
association signals, we performed a stepwise logistic regression analysis based on the expanded
sample data, resulting in a final model consisting of rs487230 (odds
ratio [OR] = 1.55, p = 0.004) and rs12136799 (OR = 0.56, p = 0.08). No other SNPs were able to enter the model at a
significance threshold of 0.10. Results from a stepwise
procedure of the four SNPs in the USP40 region were, however,
not interpretable due to nearly complete concordance for two of
the possible SNP pairs resulting in models which suffer from severe
multicollinearity. Models that included SNP pairs where con-
cordance was not near-complete resulted in nonsignificance of
both SNPs.

We also used logic regression, a novel methodology useful
for exploring relationships between disease and multiple SNPs,
to determine if some combination of SNPs could better explain
the association with disease than any single SNP. The analysis
was limited to models containing one tree, and model sizes ranging
from one to eight leaves were considered. A 10-fold
cross-validation demonstrated that models with two leaves had
the best average score when evaluated on the test portion of the
data sets. The logic regression simulated annealing algorithm
was then used to determine the best two leaf model, which resulted in
a logic tree where subjects possessing the AA genotype for
rs838552:T>C in USP40 and not the GG genotype for rs487230
in USP24 had an OR of 0.42 vs. subjects with any other genotypes.
A permutation procedure indicated that this model was signifi-
cantly better than expected by chance (p = 0.012). The resulting
model did not represent gene–gene interaction in the statistical
sense as the risk for subjects carrying the risk genotypes for one of
the SNPs did not differ significantly between the nonrisk and risk
genotypes of the other SNP. Rather, it represented an analysis of
the combined effects of these two SNPs in predisposing individuals
to PD.

Testing the Expanded Sample Set for Population Stratification

We examined whether there was potential population stratification
in the expanded sample set by individually genotyping a set of
78 null SNPs. As an initial test to attempt to detect the presence
of population stratification, we calculated the distribution of
significant markers at various significance thresholds for these
SNPs (Table 3). Significant p-values for the number of significant
markers can indicate the presence of population stratification. The
most extreme deviation was found at a significance level of 0.07,
where six out of the 78 null markers were found to have
association p-values below 0.07. This was not different from that
expected under a model of no population stratification (5.46
markers, on average, were expected to have p-values below 0.07 in
the absence of population stratification) and hence the p-value for
the number of significant markers was not significant (p = 0.47).
Thus, there was no evidence for population stratification with this
test. Additionally, in a method analogous to that described by
Pritchard and Rosenberg [1999], Fisher’s combined p-value
[Fisher, 1954] was calculated across all null markers and yielded
a value of 0.65. This result also showed no increase in significant
markers, suggesting the absence of population stratification in our
sample set.

We then determined the deviation from expected heterozygosity
levels assuming HWE. This was done separately for cases and
controls. Observed values for the summary statistic were WCs =

Multiple SNP Analysis

To identify a subgroup of SNPs in the USP24 region that
independently explained the association signals, we performed a
forward stepwise logistic regression analysis based on the expanded
sample data, resulting in a final model consisting of rs487230 (odds
ratio [OR] = 1.55, p = 0.004) and rs12136799 (OR = 0.56, p = 0.08). No other SNPs were able to enter the model at a
significance threshold of 0.10. Results from a stepwise
procedure of the four SNPs in the USP40 region were, however,
not interpretable due to nearly complete concordance for two of
the possible SNP pairs resulting in models which suffer from severe
multicollinearity. Models that included SNP pairs where con-
cordance was not near-complete resulted in nonsignificance of
both SNPs.

We also used logic regression, a novel methodology useful
for exploring relationships between disease and multiple SNPs,
to determine if some combination of SNPs could better explain
the association with disease than any single SNP. The analysis
was limited to models containing one tree, and model sizes ranging
from one to eight leaves were considered. A 10-fold
cross-validation demonstrated that models with two leaves had
the best average score when evaluated on the test portion of the
data sets. The logic regression simulated annealing algorithm
was then used to determine the best two leaf model, which resulted in
a logic tree where subjects possessing the AA genotype for
rs838552:T>C in USP40 and not the GG genotype for rs487230
in USP24 had an OR of 0.42 vs. subjects with any other genotypes.
A permutation procedure indicated that this model was signifi-
cantly better than expected by chance (p = 0.012). The resulting
model did not represent gene–gene interaction in the statistical
sense as the risk for subjects carrying the risk genotypes for one of
the SNPs did not differ significantly between the nonrisk and risk
genotypes of the other SNP. Rather, it represented an analysis of
the combined effects of these two SNPs in predisposing individuals
to PD.

Testing the Expanded Sample Set for Population Stratification

We examined whether there was potential population stratification
in the expanded sample set by individually genotyping a set of
78 null SNPs. As an initial test to attempt to detect the presence
of population stratification, we calculated the distribution of
significant markers at various significance thresholds for these
SNPs (Table 3). Significant p-values for the number of significant
markers can indicate the presence of population stratification. The
most extreme deviation was found at a significance level of 0.07,
where six out of the 78 null markers were found to have
association p-values below 0.07. This was not different from that
expected under a model of no population stratification (5.46
markers, on average, were expected to have p-values below 0.07 in
the absence of population stratification) and hence the p-value for
the number of significant markers was not significant (p = 0.47).
Thus, there was no evidence for population stratification with this
test. Additionally, in a method analogous to that described by
Pritchard and Rosenberg [1999], Fisher’s combined p-value
[Fisher, 1954] was calculated across all null markers and yielded
a value of 0.65. This result also showed no increase in significant
markers, suggesting the absence of population stratification in our
sample set.

We then determined the deviation from expected heterozygosity
levels assuming HWE. This was done separately for cases and
controls. Observed values for the summary statistic were WCs =
In an effort to obtain association p-values for test SNPs corrected for any existing population stratification, a gamma model was fit to the negative log-transformed p-values of the 78 null markers (b=0.77075; c=1.23606). Plotting the empirical cumulative distribution function of the negative log-transformed p-values of the 78 null markers showed a very good fit (data not shown), thereby substantiating hypothesis aside from very nonsignificant unadjusted test SNPs, the adjustment procedure slightly increased the level of significance for a given unadjusted p-value. For instance, the new null model would increase the significance of a nominal p-value of 0.05 to 0.033. Hence, to be conservative, we reported unadjusted p-values in lieu of population stratification-adjusted p-values.

**DISCUSSION**

Ubiquitin-proteosomal pathway is involved in the etiology of PD [reviewed by Dawson and Dawson, 2003; Giasson and Lee, 2003]. Thus, in addition to being positional candidate genes exhibiting interesting PD-association patterns, both USP24 and USP40 have strong a priori biological plausibility to be involved in PD. Multiple SNPs in USP24 and USP40 showed significant association with PD risk in our sample set, including nonconservative missense variants and predicted TFBS variants. The stepwise logistic regression model suggested that rs487230 and/or rs13312 showed significant association with PD risk for rs487230 with AAO but not risk and found no association with either risk or AAO with rs12130799 were the most informative markers in the USP24 region. The logic regression model also identified rs487230 together with rs838552. Both models appeared to suggest the importance of rs487230. However, its impact on USP24 function is unknown, but it is worthwhile to note that a Val/Ala polymorphism in methylenetetrahydrofolate reductase is the most frequent genetic cause of mild hyperhomocysteinemia and results in different biochemical properties of the enzyme [Yamada et al., 2001].

Two of the 17 significant USP24 SNPs we identified, rs487230 and rs13312, were described in the study by Oliveira et al [2005]. They observed a significant association of rs487230 with AAO but not risk and found no association with either risk or AAO with rs13312. The whole genome scan study by Maraganore et al. [2005] found no significant association for rs487230 with PD risk. Such seemingly inconsistent results could stem from differences in the composition of the sample sets used in these and our studies (discussed below) or stochastic effects. In addition, other significant SNPs with potential functional significance have not been examined in the studies by Oliveira et al. [2005] or Maraganore et al. [2005], and thus it remains to be determined whether they can be replicated in these or other sample sets. Nevertheless, our observations along with Oliveira et al. [2005]
suggest that genetic variations in USP24 may play a role in the etiology of PD. As for the PARK11 locus, the disease-associated variants in USP40 are ~5.5Mbp away from another disease-associated marker (rs10200894) that was reported by Maraganore et al. [2005] and replicated in our sample [Li et al., 2006]. However, rs10200894:C→G was not associated with PD in several other, large sample sets [Farrer et al., 2006; Goris et al., 2006] and is not in LD with the USP40 variants.

Unlike previously published reports that often included cases with both early- and late-onset PD (as in Oliveira et al. [2005] and Maraganore et al. [2005]), we have exclusively selected cases of later onset and age- and gender-matched controls to specifically identify genetic factors for sporadic PD, which has a typical onset after age 50. This design may be advantageous vs. a mixed sample set of younger and older individuals if there are specific genetic variants that contribute to late-onset PD but not early-onset PD (conversely, there may be genetic variants specifically associated with early-onset PD). Thus, our findings should be ideally validated in late-onset PD sample sets in future follow-up studies. As our samples were contributed by many investigators, it was possible that population stratification confounded our association study. However, several tests did not reveal evidence of significant population stratification. Indeed, there was strong statistical support for no population stratification. Thus, our analysis is the first validation of the utility of the NINDS Human Genetics Resources PD discovery sample collections, at least for the samples in this study.

A limitation of our study was the absence of an independent validation sample, although approximately one third of the cases and over half of the controls were not in our initial sample set that led us to follow up on the USP24 and USP40 genes. Recently, Skol et al. [2006] proposed to evaluate markers in two-stage association studies by joint analysis rather than replication. Although studies in other PD sample sets are necessary to confirm our findings, it is worth noting that six USP24 markers (rs1165228, rs287235:C→G, rs487230, rs555687, rs6671533>G>C, rs7512894:A>G) remained significant in the expanded sample set after conservatively correcting for multiple testing for all 27 tested SNPs ($\chi^2_{	ext{Bonferroni}} = 0.017–0.049$). The four USP40 markers did not remain significant after applying the conservative Bonferroni correction for the 80 markers tested in this region in the initial sample set and thus would require validation in future studies. However, it is encouraging that the ORs for the four USP40 markers tested in the expanded sample were consistent with the ORs estimated from the initial sample set, since if the original findings were due to random chance, the OR would be expected to move towards 1 as additional samples were added. (ORs among the set of 272 subjects that were in the expanded sample set but not included in the initial sample set were 1.42, 1.55, 1.39, and 1.35 for rs1048603, rs838552, rs838547, and rs4047199, respectively.) In addition, this study identified SNPs in two members of the same protein family, belonging to a pathway that is considered relevant to the disease etiology, which might be interpreted as biological confirmation of our results. Those mutations/variants in homologous proteins contribute to human disorders has been well documented. For example, mutations in presenilin genes PSEN1 (MIM# 104311) and PSEN2 (MIM# 600759) cause rare, early-onset Alzheimer’s disease [Levy-Lahad et al., 1995a,b; Rogaev et al., 1995; Sherrington et al., 1995]. More recently, we have reported that genetic variants in several members of the GAPD (MIM# 138400) family are associated with the more common, late-onset Alzheimer’s disease [Li et al., 2004]. Our identification of USP24 and USP40 variants, located in two distinct PD linkage regions, warrants further biological studies of these genes to assess their role in the disease process.

ACKNOWLEDGMENTS

We thank the contributors and the organizers, particularly Dr. Katrina Gwinn Hardy and Jeanne Beck of the NINDS Human Genetics Resource, for making the clinical samples available to the Parkinson disease research community, and the families and individuals for their participation. We also thank our colleagues at Celera Diagnostics for providing expert technical support and Drs. John Sninsky and Thomas J. White for their guidance and for critical review of the manuscript.

REFERENCES


