Identification of genetic modifiers of age-at-onset for familial Parkinson's disease.

Erin M. Hill-Burns, University of Alabama at Birmingham
Owen A. Ross, Mayo Clinic
William T. Wisseemann, University of Alabama at Birmingham
Alexandra I. Soto-Ortolaza, Mayo Clinic
Sepideh Zareparsi, Oregon Health & Sciences University
Joanna Siuda, Medical University of Silesia
Timothy Lynch, University College Dublin
Zbigniew K. Wszolek, Mayo Clinic
Peter A. Silburn, Griffith University
George D. Mellick, Griffith University

Only first 10 authors above; see publication for full author list.

Journal Title: Human Molecular Genetics
Volume: Volume 25, Number 17
Publisher: Oxford University Press (OUP): Policy B - Oxford Open Option B | 2016-09-01, Pages 3849-3862
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1093/hmg/ddw206
Permanent URL: https://pid.emory.edu/ark:/25593/rwjws

Final published version: http://dx.doi.org/10.1093/hmg/ddw206

Copyright information:
© The Author 2016. Published by Oxford University Press. This is an Open Access work distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/).

Accessed December 21, 2017 4:46 PM EST
Identification of genetic modifiers of age-at-onset for familial Parkinson’s disease


Abstract

Parkinson’s disease (PD) is the most common cause of neurodegenerative movement disorder and the second most common cause of dementia. Genes are thought to have a stronger effect on age-at-onset of PD than on risk, yet there has been a phenomenal success in identifying risk loci but not age-at-onset modifiers. We conducted a genome-wide study for age-at-onset. We analysed familial and non-familial PD separately, per prior evidence for strong genetic effect on age-at-onset in familial PD. GWAS was conducted in 431 unrelated PD individuals with at least one affected relative (familial PD) and 1544 non-familial PD from the NeuroGenetics Research Consortium (NGRC); an additional 737 familial PD and 2363 non-familial PD were used for replication. In familial PD, two signals were detected and replicated robustly: one mapped to LHFPL2 on 5q14.1 (PNGRC = 3E-8, PReplication = 2E-5, PNGRC + Replication = 1E-11), the second mapped to TPM1 on 15q22.2 (PNGRC = 8E-9, PReplication = 2E-4, PNGRC + Replication = 9E-11). The variants that were associated with accelerated onset had low frequencies (<0.02). The LHFPL2...
variant was associated with earlier onset by 12.33 [95% CI: 6.2; 18.45] years in NGRC, 8.03 [2.95; 13.11] years in replication, and 8.79 [5.88; 13.70] years in the combined data. The TPM1 variant was associated with earlier onset by 15.30 [8.10; 22.49] years in NGRC, 9.29 [7.9; 16.79] years in replication, and 12.42 [7.23; 17.61] years in the combined data. Neither LHFPL2 nor TPM1 was associated with age-at-onset in non-familial PD. LHFPL2 (function unknown) is overexpressed in brain tumours. TPM1 encodes a highly conserved protein that regulates muscle contraction, and is a tumour-suppressor gene.

Introduction

Genetics plays a significant role in PD [MIM*168600], both in determining risk (if one will develop PD: cause) as well as age-at-onset (when a disease might manifest: modifier) (1). Several rare causative genes (2–11) and 28 common risk alleles (12–16) have been confirmed for PD. The known genes and risk factors account for ~5% of the heritability (17), hence much of the genetic component of PD is still missing.

Age-at-onset of PD varies by approximately 80 years (Fig. 1). The factors that contribute to the variation in age-at-onset are unknown, although genes are thought to be important. Heritability of PD has been estimated as 98% (SE = 0.25) for age-at-onset and 60% (SE = 0.10) for risk (1). Data from the most recent PD meta-genome-wide association study (GWAS) have provided significant evidence for a polygenic component to age-at-onset (18), although no specific genes were identified. Three independent complex segregation analyses have reported a significantly better fit for a genetic model than for an environmental model for PD, and found the genetic effect on age-at-onset to be significantly greater than the genetic effect on risk (19–21). In one study, the best-fit model was rare alleles with large effects on age-at-onset in familial PD (19). Another study estimated an average decrease in age-at-onset of approximately 18 years for each copy of the putative allele (21). Thus, taken collectively, the clues from complex segregation analyses were “rare variant”, “large impact on age-at-onset”, and “positive family history”.

The loci that affect risk have little effect on age-at-onset. The International PD Genetic Consortium (6,249 PD cases) (18) and studies from Denmark (1,526 cases) (22) and from Norway and Sweden (1,340 cases) (23) independently reported that the risk alleles identified to date account for <1% of the variation in age-at-onset. Thus, 99% of the 80-year variation in age-at-onset of PD remains unexplained.

Here, we report evidence for the existence of variants with low allele frequencies and large effects on age-at-onset of familial PD, which we identified via GWAS and replicated independently. We analyzed familial and non-familial PD separately because complex segregation analyses had suggested a strong genetic effect on age-at-onset of familial PD specifically (19).

About one-fourth of persons with PD report a positive family history (Table 1), but their families rarely show a Mendelian inheritance pattern and most are not caused by known PD mutations (3–11). The vast majority of familial PD remains idiopathic, and like non-familial PD, is thought to involve complex interactions between the genome and environmental exposures (24–27). It is usually assumed that the same genes operate in familial and non-familial PD; in fact, GWAS for risk have successfully uncovered numerous susceptibility loci without separating the subtypes (12–16,26–28). However, familial and non-familial PD might differ in the relative burden of genetic and non-genetic modifiers (13,29,30). If certain variants are involved predominantly in one subtype (e.g. in familial PD as segregation analysis has suggested for age-at-onset modifiers), their signal may become diluted and undetectable if familial and non-familial PD are mixed. A positive family history does not necessarily imply a genetic aetiology because non-genetic disease can also cluster in families due to a common exposure. Similarly, genetic disease may present as non-familial due to incomplete penetrance (e.g. LRKK2 mutations (29)). Moreover, a familial case may be classified as non-familial given the difficulty in recall and knowledge of family members. Despite these uncertainties, stratifying by presence/absence of family history proved to be key to identifying two genes that each affect age-at-onset by a decade.

Results

Genome-wide genotyping was conducted using Illumina HumanOmni1-Quad_v1-0_B BeadChips on 3986 subjects from NGRC (13), including 435 familial PD (one person per family), 1565 non-familial PD and 1986 controls (PD subjects were used for analysis of age-at-onset, and controls were used for ancillary tests). Subjects were unrelated (subjects with cryptic relatedness PI_HAT > 0.15 were excluded). Over 800,000 genotyped SNPs passed quality control (13). We used imputation and expanded the coverage to 7.2 million SNPs (30). Statistical testing for GWAS was conducted using Cox regression survival analysis, treating age-at-onset as a quantitative trait. Linear

Figure 1. Variation in age-at-onset of PD Age-at-onset distribution in NGRC subjects shows nearly 80 years of variation in both familial and non-familial PD. The tails (age at onset < 20 or > 89 years) were excluded from analyses.
regression was also performed which yielded similar but less significant results than Cox. Cox regression is particularly suited for the analysis of time-to-event data, such as age-at-onset, where subjects are treated as unaffected from birth until the age when they develop symptoms (event [31–34]). Using an additive genetic model, genotypes were compared for age-specific incidence of PD symptoms using Cox regression, and hazard ratios (HR) were calculated with their associated P-values. The resulting Manhattan plots and quantile-quantile (QQ) plots are shown in Figure 2. Genomic inflation factors were close to one ($\lambda_{\text{familial}} = 0.989$, $\lambda_{\text{non-familial}} = 0.996$, $\lambda_{\text{all-PD}} = 1.007$) indicating the P-values were not inflated. Genome-wide significant signals ($P < 5E-8$) were seen only in familial PD. Complete genome-wide results, including HR and P-values for 7.2 million SNPs for familial, non-familial, and all PD, are provided in the Supplementary Tables.

Familial PD

Four loci reached $P < 5E-8$ in familial PD (Fig. 2A, Table 2). They were on chromosome 5q14.1 (rs344650: minor allele frequency (MAF) = 0.016; HR = 4.77, $P = 3E-8$), chromosome 8q23.3 (rs74355301: MAF = 0.014; HR = 4.46, $P = 3E-8$), chromosome 14q21.3 (rs192855008: MAF = 0.012; HR = 7.12, $P = 4E-9$), and chromosome 15q22.2 (rs116860970: MAF = 0.013; HR = 6.52, $P = 8E-9$). Genome-wide signals for familial PD are provided in Supplementary Material, Table S1.

The signal on 5q14.1 included a variant that was directly genotyped on the GWAS array. The other three peaks were imputed. Since the fidelity of imputation for rare variants is unknown [35], we genotyped a subset of samples for the three imputed peaks (see Methods for details). Concordance between genotyped and imputed results was 98% for 15q22.2, 99% for 8q23.3 and 100% for 14q21.3. Replication samples were all genotyped. Adjusting for the first two principal components improved the association significantly ($P = 3E-8$) in familial PD in GWAS (Table 2). The discovery signal for 8q23.3 included only one SNP (down to $P = 1E-6$), which adds to the uncertainty about the original finding at this peak.

The signal from 5q14.1 mapped to the LHFPL2 (Lipoma HMGC Fusion Partner-Like 2) gene. LHFPL2 rs344650_G vs. A (5q14.1) yielded HR = 4.77 ($P = 3E-8$) in familial PD in GWAS, HR = 2.68 ($P = 2E-5$) in familial PD in replication, and HR = 3.40 ($P = 1E-11$) in a meta-analysis of familial PD in GWAS and replication (Fig. 3A).

Table 1. Datasets and subject characteristics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N</th>
<th>M/F</th>
<th>Age</th>
<th>Onset age</th>
<th>N</th>
<th>M/F</th>
<th>Age</th>
<th>Onset age</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Familial PD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NGRC PD</td>
<td>431</td>
<td>280/151</td>
<td>66.2 ± 10.4</td>
<td>56.9 ± 11.7</td>
<td>1554</td>
<td>1057/497</td>
<td>67.5 ± 10.6</td>
<td>58.9 ± 11.4</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1985</td>
<td>1337/648</td>
<td>67.2 ± 10.6</td>
<td>58.5 ± 11.5</td>
</tr>
<tr>
<td><strong>Replication</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUST</td>
<td>293</td>
<td>170/123</td>
<td>69.8 ± 10.3</td>
<td>57.5 ± 11.2</td>
<td>842</td>
<td>532/310</td>
<td>71.9 ± 10.1</td>
<td>60.4 ± 11.2</td>
</tr>
<tr>
<td>HBS*</td>
<td>99</td>
<td>67/32</td>
<td>63.8 ± 9.2</td>
<td>58.7 ± 9.5</td>
<td>350</td>
<td>227/123</td>
<td>66.7 ± 10.0</td>
<td>62.6 ± 10.6</td>
</tr>
<tr>
<td>MCJ</td>
<td>12</td>
<td>5/7</td>
<td>61.8 ± 9.6</td>
<td>55.1 ± 11.0</td>
<td>229</td>
<td>134/95</td>
<td>59.3 ± 10.0</td>
<td>51.3 ± 10.6</td>
</tr>
<tr>
<td>MCJ</td>
<td>142</td>
<td>90/52</td>
<td>65.9 ± 9.7</td>
<td>63.0 ± 10.7</td>
<td>182</td>
<td>113/69</td>
<td>64.9 ± 10.7</td>
<td>63.8 ± 12.0</td>
</tr>
<tr>
<td>MCJP</td>
<td>39</td>
<td>22/17</td>
<td>62.9 ± 8.6</td>
<td>55.3 ± 10.1</td>
<td>272</td>
<td>172/100</td>
<td>67.2 ± 10.3</td>
<td>59.1 ± 11.1</td>
</tr>
<tr>
<td>MCJU</td>
<td>112</td>
<td>74/38</td>
<td>66.5 ± 12.6</td>
<td>59.9 ± 12.8</td>
<td>217</td>
<td>139/78</td>
<td>70.7 ± 10.7</td>
<td>64.8 ± 12.2</td>
</tr>
<tr>
<td>UCL*A</td>
<td>40</td>
<td>21/19</td>
<td>70.8 ± 9.9</td>
<td>68.8 ± 9.7</td>
<td>271</td>
<td>156/115</td>
<td>71.4 ± 10.5</td>
<td>69.3 ± 10.6</td>
</tr>
<tr>
<td>Total</td>
<td>737</td>
<td>449/288</td>
<td>68.0 ± 10.6</td>
<td>59.6 ± 11.4</td>
<td>2363</td>
<td>1473/890</td>
<td>69.0 ± 10.9</td>
<td>61.4 ± 12.0</td>
</tr>
</tbody>
</table>

NGRC and replication datasets were tested for potential overlap; no evidence was found for overlap. Subjects with age-at-onset at the extreme tails of the distribution (<20 years and ≥89 years) were excluded from analysis. Control subjects were used to test and rule out association of SNPs with age and with disease risk. M/F – N male/N female. Age – Age-at-enrollment ± standard deviation. Onset age – age-at-onset of first motor symptom of PD (age-at-diagnosis) ± standard deviation.
The signal from 15q22.2 mapped to the TPM1 (tropomyosin) gene. TPM1 rs117267308_A vs. T (15q22.2) yielded HR = 6.47 in familial PD in GWAS (P = 2E-8), HR = 3.20 (P = 2E-4) in familial PD in replication, and HR = 4.55 (P = 9E-11) in a meta-analysis of familial PD in GWAS and replication (Fig. 4A). The presence of the rs117267308_A allele was associated with 15 years earlier onset in NGRC (β = -15.30 [-22.49; -8.10]), 9 years in replication (β = -9.29 [-16.79; -1.79]), and 12 years in combined data (β = -12.42 [-17.61; -7.23]) (Fig. 4B). Age-at-onset distribution curves generated by the Kaplan Meier method showed significant separation between rs117267308_AT and rs117267308 TT genotypes in familial PD (NGRC = 2E-10 (Fig. 5C), PReplication = 7E-3 (Fig. 5D)). rs117267308 was not associated with risk of familial PD (OR = 1.18, P = 0.67). rs117267308 was not associated with age in controls (P = 0.78) or in patients (P = 0.57 adjusted for age-at-onset). The MAPs of TPM1 were consistent with the signature pattern for an age-at-onset modifier (Fig. 6E and F).

There was no significant difference in association with age-at-onset between sexes for LHFPL2 or TPM1. In familial PD, carriers of rare alleles were heterozygous. One LHFPL2 rs344650_GG rare homozygote was observed in non-familial PD.

Non-familial PD

No signal reached P < 5E-8 in non-familial PD (Fig. 2B). Genomewide results for non-familial PD are provided in Supplementary Material, Table S2. The strongest signal in non-familial PD was at P = 6E-7 (Table 3). Note that the sample size for non-familial PD was three times larger than the sample size for familial PD.
Table 2. Signals that achieved the significance threshold in GWAS for associations with age-at-onset of familial PD

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP</th>
<th>MAF</th>
<th>HR (95% CI)</th>
<th>P</th>
<th>Alpha-test (GWAS)</th>
<th>Linear regression (GWAS)</th>
<th>P</th>
<th>HR (95% CI)</th>
<th>P</th>
<th>Alpha-test (Replication)</th>
<th>Linear regression (Replication)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>LHFPL2</td>
<td>rs344650</td>
<td>0.016</td>
<td>1.18, 0.77</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td>1.18, 0.77</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LHFPL2</td>
<td>rs79503702</td>
<td>0.012</td>
<td>0.79, 0.53</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td>0.79, 0.53</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPM1</td>
<td>rs141049631</td>
<td>0.012</td>
<td>1.18, 0.77</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td>1.18, 0.77</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPM1</td>
<td>rs116860970</td>
<td>0.013</td>
<td>0.79, 0.53</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td>0.79, 0.53</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPM1</td>
<td>rs77362326</td>
<td>0.012</td>
<td>1.18, 0.77</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td>1.18, 0.77</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPM1</td>
<td>rs117484764</td>
<td>0.012</td>
<td>1.18, 0.77</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td>1.18, 0.77</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SNPs that achieved P < 5E-8 in GWAS for associations with age-at-onset of familial PD are shown. They are in four LD blocks. One SNP per block was genotyped in additional samples of familial PD for replication. GWAS was conducted using Cox regression. Replication testing was conducted using Cox regression, and datasets were combined using Meta analysis. The effect on age-at-onset was calculated using linear regression. No SNPs achieved P < 5E-8 in non-familial PD or in all PD. For the list of signals associated with age-at-onset in non-familial PD (Fig. 5A–D) and the lack of a difference in non-familial PD (Fig. 5E–H).

Discussion

The present findings provide evidence for the existence of uncommon variants with large effects on the age-at-onset of PD. Although 28 susceptibility alleles have so far been identified for PD via GWAS, much of the heritability is still unaccounted for. As a result, modifiers of age-at-onset and rare variants are now receiving increasing attention. It was recently shown that all known PD risk loci identified via GWAS account for <1% of the 80-year variation in age-at-onset (18,22,23). The loci observed in the present study would not have been detected in prior PD GWAS because they affect age-at-onset and not risk, and because the signals are undetectable unless familial and non-familial PD are separated. The present study provides proof of concept that some of the missing heritability is in age-at-onset modifiers and uncommon variants. It demonstrates that the genetic architecture of familial and non-familial PD is only partially overlapping (modifiers that operate predominantly in one and not the other subtype produce diluted undetectable signals when all PD are combined). Our study also corroborates the results of the complex segregation analyses that predicted the existence of rare genetic variants with large effects on age-at-onset of familial PD (1,19–21).

The most significant finding was the detection and replication of two signals on chromosomes 5q14.1 and 15q22.2. Each locus achieved genome-wide significance in familial PD and had no signal in non-familial PD. The minor alleles had low frequencies (0.016 and 0.012) but each locus shifted onset age by 1.18–2.25 years. The minor allele at 5q14.1 was associated with risk in non-familial PD (HR = 1.87 years) or in replication (Cox with Firth correction P = 0.02 years) and had only a weak trend in replication (Cox with Firth correction P = 0.91, β = 1.25 years) nor TPM1 (Cox with Firth correction P = 0.06, β = -1.14 years) was associated with age-at-onset in non-familial PD. Neither LHFPL2 (OR = 0.94, P = 0.77) nor TPM1 (OR = 1.18, P = 0.53) was associated with risk in non-familial PD. The Kaplan Meier curves best illustrate the contrast between the marked difference in genotype-specific age-at-onset distributions in familial PD (Fig. 5A–D) and the lack of a difference in non-familial PD (Fig. 5E–H).
10–12 years. The loci accounted for 3.5% (5q14.1) and 3.9% (15q22.2) of variation in age-at-onset.

The 5q14.1 signal maps to LHFPL2 [MIM*609718], a member of the lipoma HMGIC fusion partner (LHFP) gene family. The function of LHFPL2 is unknown. Interestingly, LHFPL2 is expressed in all normal tissues and cell lines except brain and leukocytes (40); however, while healthy brain tissue has no detectable LHFPL2 transcript, LHFPL2 protein is abundant in malignant brain tissue (41). The 15q22.2 signal maps to the tropomyosin 1 gene (TPM1 [MIM*191010]). TPM1 encodes a highly conserved actin-binding protein that plays a central role in calcium-dependent regulation of muscle contraction. TPM1 is a tumour suppressor gene (42).

Cancer and Parkinson’s disease are often likened to the two sides of a coin. Epidemiological studies have shown that the risk of developing PD is inversely associated with the risk of developing cancer (except skin cancer) (43). The pathways that lead to neuronal apoptosis, such as mitogen-activated protein kinase (MAPK) signalling, can also lead to their uncontrolled growth (44). There is also evidence from genetics for overlap, best exemplified by PARK2, which is both a tumour suppressor gene (45,46) and the most common cause of early-onset PD.
LHFPL2 and TPM1 may also be genetic links between cancer and PD. Many of the markers that associate with onset of familial PD map to sequences that are identified by the Roadmap Epigenomics Project (http://genomebrowser.wustl.edu) and ENCODE (48) as being active regulatory elements in the brain (Figs 7 and 8). The variants were not found in eQTL or mQTL databases Genevar (49), eqtl (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/), SCAN (50), or BRAINEAC (51), likely due to their low frequencies, thus we could not test their association with the expression or methylation of LHFPL2, TPM1, or adjacent genes.

We did not attempt to replicate signals that had $P > 5 \times 10^{-8}$. It is noteworthy, however, that a block of variants mapping to 9q31.1 produced similar signals in familial (HR = 1.7, $\beta = 4.45$) and non-familial PD (HR = 1.7, $\beta = 5.11$), and when combined, the signal reached $P = 4 \times 10^{-7}$. Low analytic power could have kept the 9q31.1 signal from reaching the significance threshold. The 9q31.1 signal maps to the neuronal plasticity gene LPPR1 which is highly expressed in the brain and is involved in glutamate-receptor mediated neuronal excitation (52), one of the mechanisms that is believed to cause neuronal death in PD (53).

Figure 4. Replication results for rs117267308 in TPM1 in familial PD. In the replication datasets, excluding NGRC dataset (GWAS), the rs117267308_A allele was associated with more than three-fold higher age-specific hazard ratio (HR) and approximately 9 years earlier onset than rs117267308_T allele. (A) HR were generated using Cox regression, with Firth’s Penalized correction for datasets with 6 or fewer observations. The forest plot depicts the HR with SE for each dataset individually, and combined using Fixed and Random Effects meta-analysis. (B) Mean differences in age-at-onset were calculated using linear regression. Additive models were used (estimates are per allele). Each panel shows the replication datasets only on top, followed by NGRC plus replication datasets. W: weight of each dataset in meta-analysis under fixed or random effects model.
variants. Given these limitations, that two loci reached genome-wide significance in discovery and replicated robustly is remarkable. Our study revealed several signals for variants that achieved $P < 1E^{-6}$, which is promising enough to warrant studies that are specifically designed to detect and validate uncommon and rare variants.

Materials and Methods

Human subjects and data collection

Subjects: Institutional Review Boards and Human Subject Committees at participating institutions approved the study. Subject characteristics are shown in Table 1. For the discovery phase (GWAS) we used the subjects from NGRC (13). Uniform methods were used for diagnosis, subject selection, data collection, DNA preparation, genotyping, imputation, and analysis. Subjects included 2,000 individuals with the diagnosis of PD (54) whom we used to study age-at-onset, and 1,986 control subjects whom we used to rule out confounding due to associations with age. NGRC patients were on average 8 years past diagnosis, thus excluding early misdiagnoses which occur at a rate of 25% (55). Controls were free of neurodegenerative disease by self-report; a subset of older controls were examined and confirmed by neurologists to be unaffected (13). All patients and controls were American of European origin and unrelated to each other (PI_HAT ≤ 0.15) (13). For replication, seven datasets were used, made available by investigators at Griffith University Australia (AUST) (56), Harvard Biomarker Study (HBS) (57), University of California, Los Angeles (UCLA) (58), and Mayo Clinic Jacksonville (MCJ) (56) which included four cohorts of Irish (MCJI), Polish (MCJP), and Caucasian of European decent with mixed (MCJE) or unknown (MCJU) European countries of origin. In total, replication included DNA, age-at-onset or age-at-diagnosis, family history data, sex, and age-at-enrolment on a total of 3100 persons with PD (Table 1). All subjects were Caucasian. No overlaps: We compared all subjects across all datasets (NGRC and replications) for 74 SNP genotypes, sex, family history and age-at-onset/age-at-diagnosis. Eight pairs of individuals matched on all items. We reached out to the investigators for each dataset, obtained additional information on the 8 pairs, and were able to clear all of them as unique individuals. Additionally, we were able to confirm that there were no first-degree relatives among the carriers of LHFPL2 or TPM1 rare alleles across datasets.

Age-at-onset & Family history: NGRC subjects used for GWAS were recruited from neurology clinics sequentially and irrespective of age-at-onset or family history. Age-at-onset was defined as the age when the subject noticed the first motor symptom of PD. Age-at-onset was obtained at three independent occasions, several years apart: at the time of diagnosis by the movement disorder specialist as noted in medical records, at enrolment in our genetic study (59,60), and at enrolment in our environmental study (61). The three sources were compared, and inconsistencies that were >2 years were either resolved or the subject was designated as having unknown age-at-onset ($n = 1$). The outliers (onset ≤20 years or >89 years) were excluded from analysis ($n = 14$). Family history was obtained using a standardized self-administered questionnaire (59). Patients who reported a first or second-degree relative with PD were classified as familial PD; all others were classified as non-familial PD. Only one person per family was used. GWAS consisted of 1985 persons with PD, with known age-at-onset; 431 were familial PD and 1554 were non-familial PD. Datasets used for replication were each collected with a different study design and ascertainment method necessitating tests of heterogeneity and the use of meta-analysis. Each group had classified their samples as familial or non-familial. AUST, MCJE, MCJI, MCJP, and MCJU had collected age-at-onset. HBS and UCLA had collected age-at-diagnosis instead of age-at-onset, but age-at-diagnosis and age-at-onset are highly correlated (tested in NGRC $r^2 = 0.93$, $P < 1E^{-16}$). For HBS and UCLA we used age-at-diagnosis instead of age-at-onset. Each dataset had either age-at-onset or age-at-diagnosis, but not a mix of both. In
Figure 6. Moving average plots (MAP). Minor allele frequencies are plotted in a moving-average window across the age spectrum in NGRC controls (blue) and as a function of age-at-onset in patients (red). For the description of the MAP method see (38). Data are shown for the LHFPL2 rs344650_G allele and the TPM1 rs117267308_A allele, as well as for two well-established PD loci for the purpose of demonstration: SNCA rs356220, which is associated with risk in all PD (A), and PARK2 deletion/duplication, which is associated with risk of early-onset PD. (B) The MAP of SNCA rs356220 demonstrates the expected pattern for a variant that is associated with increased risk ubiquitously: allele frequency is higher in patients and parallels the control frequency, always staying higher, with no variation with age or age-at-onset. The plot for PARK2 is the signature pattern for variants that are associated with the risk of early-onset disease: allele frequency in patients is the highest in early-onset cases and decreases with increasing age-at-onset until it reaches the control frequency when it stops declining and remains superimposed on controls. LHFPL2 rs344650 has the signature pattern for an age-at-onset modifier in familial PD (C,D): accelerated onset in rs344650_G carriers causes the allele frequency to be highest in early-onset cases, decrease with increasing ages-at-onset, cross the control frequency and continue to drop below the control frequency – yet overall, rs344650_G frequency in all patients is the same as in controls. TPM1 rs117267308 exhibited a similar pattern consistent with an age-at-onset modifier in familial PD (E,F).

Table 3. Signals that achieved \( P < 1E-6 \) in GWAS

<table>
<thead>
<tr>
<th>CHR</th>
<th>BP</th>
<th>Gene</th>
<th>SNP</th>
<th>INFO</th>
<th>MAF</th>
<th>HR</th>
<th>P</th>
<th>MAF</th>
<th>HR</th>
<th>P</th>
<th>MAF</th>
<th>HR</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signals that Reached ( P &lt; 1E-6 ) in Familial PD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>118901768</td>
<td>SPAG17</td>
<td>rs78024109</td>
<td>0.92</td>
<td>0.17</td>
<td>4.06</td>
<td>5E-7</td>
<td>0.18</td>
<td>1.10</td>
<td>0.49</td>
<td>0.018</td>
<td>1.28</td>
<td>0.04</td>
</tr>
<tr>
<td>3</td>
<td>161715295</td>
<td>DTL1</td>
<td>rs12494760</td>
<td>0.92</td>
<td>0.19</td>
<td>3.84</td>
<td>2E-7</td>
<td>0.019</td>
<td>1.05</td>
<td>0.73</td>
<td>0.019</td>
<td>1.24</td>
<td>0.08</td>
</tr>
<tr>
<td>4</td>
<td>120114558</td>
<td>MYOZ2_USP53</td>
<td>rs116370732</td>
<td>0.95</td>
<td>0.13</td>
<td>5.14</td>
<td>4E-7</td>
<td>0.013</td>
<td>1.07</td>
<td>0.67</td>
<td>0.013</td>
<td>1.27</td>
<td>0.09</td>
</tr>
<tr>
<td>5</td>
<td>77860608</td>
<td>LHFPL2</td>
<td>rs344650</td>
<td>0.99</td>
<td>0.16</td>
<td>4.77</td>
<td>3E-8</td>
<td>0.014</td>
<td>0.96</td>
<td>0.79</td>
<td>0.014</td>
<td>1.16</td>
<td>0.25</td>
</tr>
<tr>
<td>6</td>
<td>105863402</td>
<td>PREP</td>
<td>rs6930232</td>
<td>0.98</td>
<td>0.18</td>
<td>1.65</td>
<td>4E-7</td>
<td>0.164</td>
<td>1.07</td>
<td>0.19</td>
<td>0.163</td>
<td>1.14</td>
<td>2E-3</td>
</tr>
<tr>
<td>7</td>
<td>30936024</td>
<td>AQP1</td>
<td>rs12112389</td>
<td>0.96</td>
<td>0.049</td>
<td>2.38</td>
<td>1E-7</td>
<td>0.052</td>
<td>0.96</td>
<td>0.64</td>
<td>0.051</td>
<td>1.09</td>
<td>0.24</td>
</tr>
<tr>
<td>7</td>
<td>129160558</td>
<td>SMKR1</td>
<td>rs62490863</td>
<td>0.90</td>
<td>0.013</td>
<td>5.84</td>
<td>8E-8</td>
<td>0.014</td>
<td>0.99</td>
<td>0.93</td>
<td>0.014</td>
<td>1.16</td>
<td>0.30</td>
</tr>
<tr>
<td>8</td>
<td>116638637</td>
<td>TRPS1</td>
<td>rs74355301</td>
<td>0.93</td>
<td>0.014</td>
<td>4.46</td>
<td>3E-8</td>
<td>0.016</td>
<td>1.04</td>
<td>0.77</td>
<td>0.015</td>
<td>1.23</td>
<td>0.12</td>
</tr>
<tr>
<td>10</td>
<td>129028001</td>
<td>DOCK1</td>
<td>rs149188358</td>
<td>0.97</td>
<td>0.011</td>
<td>6.02</td>
<td>2E-7</td>
<td>0.005</td>
<td>1.81</td>
<td>0.02</td>
<td>0.007</td>
<td>2.39</td>
<td>1E-5</td>
</tr>
<tr>
<td>14</td>
<td>50358528</td>
<td>KLHDC1_ARF6</td>
<td>rs192855008</td>
<td>0.95</td>
<td>0.012</td>
<td>7.12</td>
<td>4E-9</td>
<td>0.011</td>
<td>0.77</td>
<td>0.16</td>
<td>0.011</td>
<td>0.97</td>
<td>0.86</td>
</tr>
<tr>
<td>15</td>
<td>63351500</td>
<td>TPM1</td>
<td>rs116860970</td>
<td>0.95</td>
<td>0.013</td>
<td>6.52</td>
<td>8E-9</td>
<td>0.011</td>
<td>0.88</td>
<td>0.48</td>
<td>0.012</td>
<td>1.11</td>
<td>0.51</td>
</tr>
<tr>
<td>18</td>
<td>73807596</td>
<td>LOC339298</td>
<td>rs11660883</td>
<td>0.93</td>
<td>0.012</td>
<td>5.19</td>
<td>5E-7</td>
<td>0.015</td>
<td>1.02</td>
<td>0.90</td>
<td>0.015</td>
<td>1.18</td>
<td>0.24</td>
</tr>
<tr>
<td>Signals that Reached ( P &lt; 1E-6 ) in Non-Familial PD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>45356065</td>
<td>PHF21B</td>
<td>rs11630553</td>
<td>0.99</td>
<td>0.022</td>
<td>0.74</td>
<td>0.20</td>
<td>0.022</td>
<td>1.86</td>
<td>0.022</td>
<td>1.43</td>
<td>1E-3</td>
<td></td>
</tr>
<tr>
<td>Signals that Reached ( P &lt; 1E-6 ) in All PD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>103982633</td>
<td>LPPR1</td>
<td>rs62576890</td>
<td>0.92</td>
<td>0.025</td>
<td>1.72</td>
<td>0.02</td>
<td>0.024</td>
<td>1.73</td>
<td>0.024</td>
<td>1.73</td>
<td>4E-7</td>
<td></td>
</tr>
</tbody>
</table>

Signals that reached \( P < 1E-6 \) in either of the three groups (familial, non-familial, all PD) are shown with the corresponding results for that signal in the other groups. Only one SNP is shown for each peak. CHR – chromosome, BP – base pair position of the top SNP (genome build 37), INFO – info score for imputed SNPs, MAF – minor allele frequency, HR – age-specific Hazard Ratio.
total, replication included 3100 persons with PD with known age-at-onset or age-at-diagnosis; 737 were familial PD and 2363 were non-familial PD.

Genotyping and imputation

NGRC subjects were genotyped using Illumina HumanOmni1-Quad_v1-0_B BeadChips (Illumina, San Diego, CA, USA) and the Illumina Infinium II assay protocol (13). Technical genotyping quality-control criteria have been described in detail (13). The array genotyping call rate was 99.92% and reproducibility rate was ≥99.99%. Subjects who were inadvertently enrolled twice, or had cryptic relatedness (PI-HAT > 0.15) were excluded. SNPs were excluded if MAF < 0.01, call-rate < 99%, HWE P < 1E-6, MAF difference in males vs. females >0.15, or missing rate in PD vs. control P < 1E-5. 811,597 SNPs passed quality-control measures (genotype and phenotype data for NGRC are available on dbGaP; http://www.ncbi.nlm.nih.gov/gap, accession number phs000196.v2.p1).

Principal component analysis (PCA) was conducted with HelixTree (http://www.goldenhelix.com) using a pruned subset of 104,064 SNPs, as described previously (13). No association was detected between PC 1-4 and age-at-onset in all PD (P-values for PC 1-4 = 0.09, 0.15, 0.81, 0.99), in familial PD (P = 0.21, 0.57, 0.73, 0.66), or in non-familial PD (P = 0.21, 0.19, 0.80, 0.95). Thus GWAS was carried out without adjustment for PC. However, we did reexamine the significant findings by including PC1 and PC2 in the model, and found the results to be similar and slightly more significant when corrected for PCs. Imputation was conducted using the IMPUTEv2.2.2 software (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html) (62) and the 1000 Genomes Phase I integrated variant set release v3. Imputed SNPs with info score <0.9 or MAF <0.01 were excluded. 6.4 million imputed SNPs passed quality control. In sum, GWAS included 7.2 million SNPs (0.8 million genotyped and 6.4 million imputed). Three of the four signals that reached P < 5E-8 were imputed. We genotyped a subset of the samples because the variants had low frequencies and the
quality of imputation for uncommon variants is unclear. For
TPM1: 29 heterozygotes and 53 common homozygotes (no rare
homozygotes were observed) as predicted by imputation were
genotyped. Genotyping results were 98% concordant with im-
puted genotypes. For TRPS1: 1 rare homozygote, 28 heterozygotes,
and 53 common homozygotes as predicted by imputation were
genotyped. Genotyping results were 99% concordant with im-
puted genotypes. For KLHDC1: 29 heterozygotes and 53 common
homozygotes (no rare homozygotes were observed) as predicted
by imputation were genotyped. Replication samples were
all directly genotyped using genomic DNA on Sequenom iPLEX
(Sequenom, San Diego, CA, USA) and TaqMan assays (Life
Technologies, Grand Island, NY, USA). None were imputed.
Primers are available on request.

Statistical analyses

Discovery: GWAS was conducted using the Cox regression survival
analysis, where age-at-onset was treated as a quantitative trait,
and an additive genetic model was used for SNP genotypes:

Survival(Age-at-onset, PD status) ~ SNP. Using the Cox method,
dosages (from 0 to 2 copies) of the minor allele of each SNP were
compared, age-for-age, for the hazard of developing PD. Survival
was measured as disease-free lifespan, from birth to age-at-onset.
A hazard ratio (HR) and P-value was calculated for each SNP under
the additive model. Significance was set at P = 5E-8. The “survival”
package in R software (63) was used for Cox regression (http://
www.r-project.org/). Manhattan plots were generated using
Haploviev v 4.2 (64). QQ plots were generated using R. Genomic
inflation factors (λ) were calculated using the “GenABEL” package
version 1.8-0 in R. Effect size on age-at-onset was estimated as the
difference in mean age-at-onset (β) using linear regression: [Age-
at-onset ~ SNP]. Linear regression was performed in ProbABEL v.
0.1-9d software (http://www.genabel.org/packages/ProbABEL) (65).

Replication testing: SNPs that generated P < 5E-8 in discovery were
genotyped in all replication samples (familial and non-familial).
Replication samples were stratified by family history for statistical
testing. For each SNP, we tested the following hypotheses in repli-
cation; (a) SNP is associated with age-at-onset in familial PD, with
the minor allele being associated with earlier onset, and (b) SNP is
not associated with age-at-onset in non-familial PD. Each SNP was

Figure 8. Alignment of TPM1 variants with regulatory markers. Shown is a 100 kb segment of DNA surrounding the variants that associate with age-at-onset of PD in the
TPM1 region (rs116860970 ~ 50kb, chr15: 63,301,500–63,401,500, genome build 37). The box on top was generated using LocusZoom and shows the SNPs with their
associated P-values (left Y-axis) and their positions on the chromosome (X-axis). rs116860970 is shown in purple. LD (r^2) was calculated in relation to rs116860970. The
colors denote the strength of LD. The top SNPs shown in purple and red span from Intr1 3 to 3’ of TPM1. The next section is from the Roadmap Epigenomics Project
and shows regulatory marks (orange=enhancers, red=transcription start sites, and green=transcribed regions) predicted by ChromHMM, with each line representing a
different brain tissue that was analyzed (BAG=brain angular gyrus; BAC=brain anterior caudate; BCG=brain cingulate gyrus; BGM=brain germinal matrix; BHM=brain
hippocampus middle; BTL=brain inferior temporal lobe; BMFL=brain mid frontal lobe; BSN=brain substantia nigra). The bottom panel is from ENCODE and shows his-
tone acetylation and methylation marks (black) in brain cells (NH-A cell line).
tested in each of the replication datasets individually, using Cox regression in R, followed by meta-analyses of replication datasets using the “meta” package version 3.2-1 in R. For datasets that had 6 or fewer observations, Firth’s Penalized estimation was used to improve precision of Cox estimates (36,37). Datasets with zero observations (lacking rare allele) were not included in the Cox or linear regression, but were included in Kaplan Meier analysis. The effect size on age-at-onset was calculated for each dataset separately using linear regression in R, and then for all datasets combined using “meta” package in R. Meta-analysis forest plots were generated using the “meta” package in R. Moving Average Plots (MAP) of allele frequencies were generated using the algorithm described previously (38) and implemented in the “freqMAP” package in R. Moving Average Plots were generated, and log-rank tests were performed using “survival” package in R. Power: The study was designed as a GWAS for common variants. Discovery of uncommon variants was a surprise. Post-hoc power calculation for GWAS suggested we had only ~1% power to detect variants with frequencies and effect sizes that we actually detected. The replication datasets had >80% power to detect the signals from the discovery at P = 0.05 assuming no heterogeneity across datasets. PS program was used for power calculation (http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize).

Functional annotation
We used LocusZoom Version 1.1 (http://locuszoom.sph.umich.edu/locuszoom) (66) to visualize the location and LD of the top association peaks. We examined Epigenomics Roadmap (via http://genomewebbrowser.wustl.edu) and ENCODE (via http://genomewebbrowser.ucsc.edu/index.html) (48) annotations of putative regulatory elements in the regions of our associated signals. We searched eQTL and mQTL databases Genearvar (https://www.sanger.ac.uk/resources/software/genevar/) (49), etql (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/), SCAN (http://www.scandb.org/newinterface/about.html) (50) and BRAINEAC (http://www.braineac.org) (51) for eQTL or mQTL association results for the associated variants, but the variants were not found in any of the databases, likely due to their low frequencies.

Supplementary Material
Supplementary Material is available at HMG online.

Acknowledgements
We thank the persons with PD and volunteers who participated in this study. We thank Ryan J. Donahue for assistance with data management and double-checking.

Conflict of Interest statement: None declared.

Funding
This work was supported by a grant from the National Institute of Neurological Disorders And Stroke [grant number R01NS036960]. Additional support was provided by National Institutes of Health [grant number P30AG08017]; a Merit Review Award from the Department of Veterans Affairs [grant number 1101BX000531]; Office of Research & Development, Clinical Sciences Research & Development Service, Department of Veteran Affairs; and the Close to the Cure Foundation. Genome-wide array genotyping was conducted by the Center for Inherited Disease Research, which is funded by the National Institutes of Health [grant number HHSN268200782096C]. Studies providing samples and data for replication were supported by National Institutes of Health [grant numbers U01NS082157, P50AG005134, R01ES010544, U54ES012078, R01NS078086 and P50NS72187]; a gift from Carl Edward Bolch, Jr. and Susan Bass Bolch; the American Parkinson’s Disease Association; the Stowarzyszenie na Rzecz Rozwoju Neurologii Wiekui Podeszlego grant; the Harvard NeuroDiscovery Center (HND); the Parkinson’s Disease Biomarkers Program (PDBP); the U.S. Department of Defense; and the M.E.M.O. Hoffman Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Funding to pay the Open Access publication charges for this article was provided by the University of Alabama at Birmingham.

References


