Human Molecular Genetics, 2011, Vol. 20, No. 2 ***345–353***
doi:10.1093/hmg/ddq469
Advance Access published on November 2, 2010

Dissection of the genetics of Parkinson's disease identifies an additional association 5′ of *SNCA* and multiple associated haplotypes at 17q21

The UK Parkinson's Disease Consortium and The Wellcome Trust Case Control Consortium 2[†]

Received July 13, 2010; Revised October 8, 2010; Accepted October 26, 2010

We performed a genome-wide association study (GWAS) in 1705 Parkinson's disease (PD) UK patients and 5175 UK controls, the largest sample size so far for a PD GWAS. Replication was attempted in an additional cohort of 1039 French PD cases and 1984 controls for the 27 regions showing the strongest evidence of association ($P < 10^{-4}$). We replicated published associations in the 4q22/*SNCA* and 17q21/*MAPT* chromosome regions ($P < 10^{-10}$) and found evidence for an additional independent association in 4q22/*SNCA*. A detailed analysis of the haplotype structure at 17q21 showed that there are three separate risk groups within this region. We found weak but consistent evidence of association for common variants located in three previously published associated regions (4p15/*BST1*, 4p16/*GAK* and 1q32/*PARK16*). We found no support for the previously reported SNP association in 12q12/*LRRK2*. We also found an association of the two SNPs in 4q22/*SNCA* with the age of onset of the disease.

INTRODUCTION

Parkinson's disease (PD) is the second most common neurodegenerative disease after Alzheimer's disease. It affects over 1% of the elderly population and despite effective symptomatic therapies is progressive and disabling. The motor phenotype is characterized by variable severity of bradykinesia, rigidity and tremor. The age at onset varies, but for the 'typical sporadic' patients it is in the seventh decade and beyond. As our population ages so the prevalence of PD is increasing. The clinical definition of the disease is based on the core features listed above and includes initial responsiveness to levodopa. This clinical phenotype correlates very highly with the pathological phenotype of Lewy body neurodegeneration.

Previous genetic studies of familial forms of PD have identified rare, highly penetrant variants in several chromosome regions, in particular 4q22/*SNCA*, 12q12/*LRRK2* 1p36/*Pink1*, 1p36/*DJ-1* and 6q26/*parkin* (1,2). However, our knowledge of the genetic factors underlying the sporadic form of PD remains poor. The advent of rapid, robust and cost-effective approaches to a systematic genome-wide association analysis has enabled appropriately powered large-scale studies to be undertaken for the first time. Recently, two groups (3,4) have reported their PD genome-wide association results. These two studies provided strong evidence of association at two chromosome regions: 4q22/*SNCA* and 17q21/*MAPT*. In addition, they present suggestive evidence of association for common variants in three chromosome regions 4p15/*BST1*, 1q32/*PARK16* and 12q12/*LRRK2*. Edwards *et al.* (5) have since confirmed the associations at 4q22/*SNCA* and 17q21/*MAPT* in a meta-analysis totalling 1752 cases and 1745 controls, but they did not find evidence for any other loci associated with PD. Identification of new associations requires additional genome-wide association study (GWAS) using larger sample sizes to provide the required statistical power to detect subtle effects on PD risk.

RESULTS

To further understand the genetic basis of PD, we undertook a PD GWAS in the UK population as part of the Wellcome

To whom correspondence should be addressed at: UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK. Tel: +44 2078373611; Fax: +44 2072785616; Email: n.wood@ucl.ion.ac.uk (N.W.W.); University of Oxford, The Wellcome Trust Center for Human Genetics Roosevelt Drive, Oxford, OX3 7BN, UK. Tel: +44 1865287725; Fax: +44 1865287501; Email: peter.donnelly@well.ox.ac.uk (P.D.)
[†]A full list of authors and affiliations is provided in the Appendix section. A full list of members of the Wellcome Trust Case Control Consortium 2 (WTCCC2) consortium is provided in the Supplementary Material.

© *The Author 2010. Published by Oxford University Press.*
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

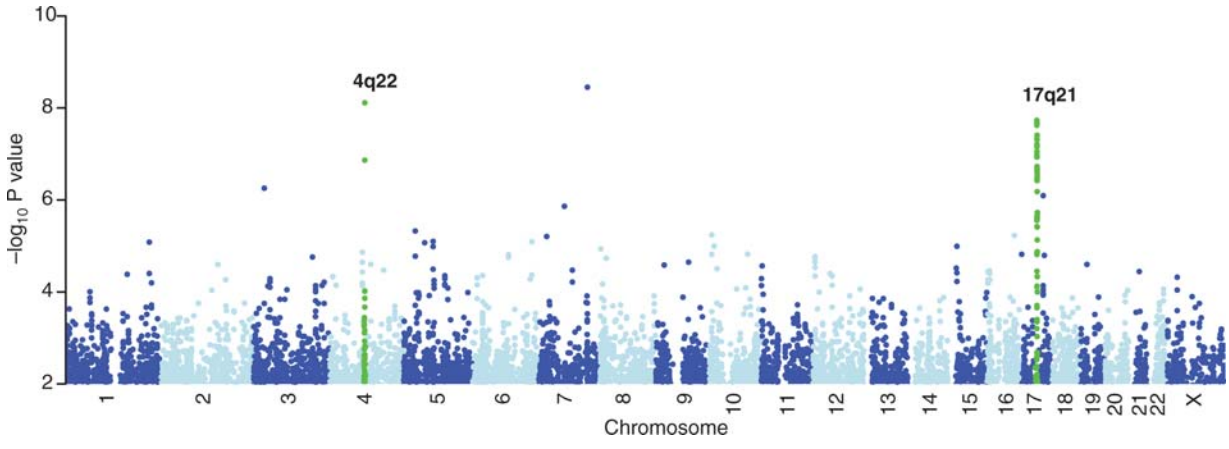

Figure 1. Genome-wide association plot. Genome-wide association results for PD at 532, 616 SNPs. Alternating chromosomes are show in shades of blue. Previously identified loci that also replicate in this study are highlighted in green.

Trust Case Control Consortium 2 (WTCCC2). Our initial sample size consisted of 5667 UK control samples and 2190 UK PD cases passing the aforementioned phenotype criteria (Supplementary Material, Table S1). UK PD cases were randomly selected from population and hospital-based clinics with an emphasis on sporadic cases with no family history for this disease. Half of the cases had been screened for the known, rare, highly penetrant G2019S mutation (1) in the *LRKK2* gene (see Materials and Methods). We identified 14 G2019S carriers, a frequency consistent with previous reports of G2019S frequency in UK PD cases. Owing to our focus on common rather than rare variants, we excluded these 14 individuals from the GWA scan.

Case samples were genotyped on the Illumina Human660-Quad platform and control samples were genotyped on the Illumina 1.2 M Duo platform; the overlap SNP set was used in this analysis. Following sample quality control (Supplementary Material, Table S2), our final data set consisted of 1705 cases and 5175 controls. We attempted replication of the top findings in a French PD case–control collection of 1039 cases and 1984 controls genotyped on the Illumina 610 genotyping array (Saad *et al*., manuscript in preparation).

After removal of 61 568 SNPs that did not pass our quality filters (see Materials and Methods), 532 616 SNPs remained for analysis. We used the program SNPTEST (6) to test each SNP for the association with case–control status using a score test, analogous to the Cochran–Armitage trend test, but modified to cope with uncertainty in genotype calls. Visual inspection of the cluster plots identified a small number of additional exclusions. An analysis of the distribution of association test statistics after removal of the SNPs located in two of the established PD associated regions (4q22/*SNCA* and 17q21/*MAPT*) showed an over-dispersion factor $\lambda = 1.039$. As the QQ plot (Supplementary Material, Fig. S1) and the λ value do not show widespread departure from the null model, we deemed it unnecessary to correct test statistics for inflation.

We found three independent regions associated with PD with *P*-values of less than 5×10^{-8} (Fig. 1). Two of them are previously reported findings: 4q22/*SNCA* (our most associated SNP rs356220, $P = 5.18 \times 10^{-9}$) and 17q21/*MAPT* (most associated SNP rs7215239, $P = 1.49 \times 10^{-8}$). The third is in the 7q32 chromosome region, where the top SNP, rs10447854 ($P = 3.11 \times 10^{-9}$), is the only SNP in the chromosome region showing evidence of association. It is in low linkage disequilibrium (LD) with other SNPs in the area and is neighbouring a recombination hotspot (Supplementary Material, Figure S2). In total, we identified 27 chromosome regions containing at least one SNP associated at $P < 10^{-4}$ in the UK case–control data (Supplementary Material, Table S3) and attempted replication at these loci in the French case–control data set. Among these loci, only the established associations 4q22/*SNCA* and 17q21/*MAPT* successfully replicated at $P < 10^{-4}$ (Supplementary Material, Table S3). The third strong association in our GWAS at 7q32 (rs10447854) showed marginal evidence of association in the replication set but in the opposite direction to that of our initial finding. Thorough checks did not find genotyping errors or allele switches and we therefore assume that in spite of the convincing statistical evidence in our GWAS, this result is a false positive.

Next, we examined other reported PD associations for evidence of association in our data. We observed *P*-values < 0.05 for SNPs in 4p15/*BST1* and 4p16/*GAK* (Table 1). We found no conclusive support for the published association in the 1q32/*PARK16* chromosome region (4) (Table 1) and, unlike both previously published PD GWAS of non-familial cases, we found no evidence of association in the 12q12/*LRRK2* chromosome region (Table 1).

To identify further risk variants, which may not be well correlated with any single SNP in our data, we used the program IMPUTE2 (7) to impute an additional 1 200 917 SNPs. To do so, we exploited data from both the HapMap project (www.hapmap.org) and additional WTCCC2 genotyping available for the control collections (see Materials and Methods). Association analysis of this enriched data set did not yield any compelling additional signals of association.

Next, we further investigated the signal at the loci showing strong evidence for association in our data. An association analysis in the 4q22/*SNCA* region conditional on the most associated SNP (rs356220) revealed a second independent association at rs7687945, which is located 5′ of *SNCA*. The

Table 1. Association results at previously reported loci

Chr	Locus	SNP	Position	Risk allele	RAF[a] Cases	Controls	*P*-value	OR[b] (95% CI)
1q32	*PARK16*	rs823128	203980001	A	0.975	0.969	0.059	1.25 (0.99–1.57)
4p16	*GAK*	rs1564282	842313	A	0.103	0.089	0.016	1.17 (1.03–1.34)
4p15	*BST1*	rs4698412	15346446	A	0.567	0.547	0.046	1.08 (1.00–1.17)
4q22	*SNCA*	rs2736990	90897564	G	0.501	0.449	1.36×10^{-7}	1.23 (1.14–1.33)
12q12	*LRRK2*	rs1994090	38714828	C	0.234	0.225	0.338	1.05 (0.95–1.15)
17q21	*MAPT*	rs393152	41074926	A	0.803	0.757	4.75×10^{-8}	1.31 (1.19–1.44)

Association results from the discovery data set for loci which have been associated with sporadic PD in previous genome-wide association studies.
[a]Risk allele frequency.
[b]Odds ratio shown for the risk allele.

likelihood ratio test of the one-SNP model compared with two-SNP model gives a *P*-value of 2.87×10^{-5}. This finding was replicated in the French data ($P = 0.00158$), where rs2301134 was used in place of rs7687945 ($r^2 = 0.98$, calculated in our 58C control data). The two SNPs showing signals, rs356220 and rs7687945, are in low LD with each other ($r^2 = 0.16$, calculated in the 58C data), but, interestingly, the LD is sufficient for the effect of the second SNP to be masked by that of the first SNP in single-SNP analyses ($P = 0.131$ for rs7687945 in single-SNP analysis). This is an example of a general phenomenon known in statistics as Simpson's paradox (8,9): association for one variable is seen when the analysis is stratified by another variable, whereas in a marginal analysis either no association is observed or it is in the opposite direction.

To illustrate the pattern of association, we phased the genotypes in this region and estimated the risk and sample frequency of the haplotypes defined by the two SNPs (Table 2). Viewing the data this way makes clear that the risk allele at the second SNP rs7687945 is more commonly found with the protective allele at rs356220 than would be expected were the SNPs in linkage equilibrium. As a result, the unconditional risk of the rs7687945 A allele (1.25) relative to the G allele (1.18) is 1.07 and not significantly different from 1.0 (Table 2), which explains the lack of marginally significant association at rs7687945. It is also true that risk alleles at the first SNP (rs356220) tend to occur with protective alleles at the second SNP. While this acts to ameliorate the marginal signal at the first SNP, the combination of haplotype frequencies and effect sizes is such that there is still a significant marginal association.

An important consequence of there being two SNPs affecting disease risk is that the genetic effect at this locus is actually considerably larger than the marginal analysis suggests. Table 2 highlights the substantial increase in risk associated with carrying both of the A alleles: individuals who are homozygote for this allele at both SNPs carry over a 2.5-fold increase in risk relative to individuals homozygote for the G alleles.

Having found signals at two SNPs in the association region, it is natural to ask whether the best statistical model just has effects for these two SNPs or whether it requires an additional parameter related to the way in which the SNP alleles combine onto haplotypes. We assessed this in the phased data and found no significant improvement in model fit with the additional, haplotype, parameter ($P = 0.1$). Although we see separate signals at two SNPs in our data, this could, in principle, arise if disease risk depended on a single, untyped SNP: for example, if the risk allele at the untyped SNP occurred with increasing frequencies on the GG, GA, AG and AA haplotypes in Table 2. To pursue this possibility, we examined whether any of the SNPs in the 1000 Genomes data (March 2010 release) in the association region was a better predictor than either rs356220 or rs7687945, but none was. In addition, the analysis of the 1000 Genomes data did not identify non-synonymous SNPs in strong LD ($r^2 > 0.5$) with either rs356220 or rs7687945. It is therefore unlikely that the effect on PD risk is mediated by a protein coding change.

A previous study reported additional PD associations in the 5′ region upstream of *SNCA* (10). However, the LD between our second signal, rs7687945, and these previously reported variants is low ($r^2 = 0.24$, 0.26 with rs2736994 and rs2737026, respectively, calculated in the imputed 58C data)

Table 2. Dissection of risk at the 4q22 locus

		rs7687945 G	rs7687945 A		
rs356220	G	1 20.8%	1.16 (1.04-1.29) 41.9%	1.11	1.26 (1.16-1.37)
	A	1.31 (1.17-1.47) 27.9%	1.64 (1.41-1.90) 9.4%	1.39	
		1.18	1.25		
		1.07 (0.98-1.15)			

Shown in the centre of the table are estimates of odds ratio, 95% confidence limit (in brackets) and percentage frequency of the four haplotypes defined by the alleles at rs356220 and rs7687945. In the margins of the table is the risk of each of the alleles obtained by averaging the odds ratio of two haplotypes on which the allele can be found, weighting by the sample frequency. For example, the risk of carrying the G allele at rs356220 unconditional on the allele carried at rs7687945 is 1.11 (given in the top right) and is calculated as $(1 \times 20.8 + 1.16 \times 41.9)/(20.8 + 41.9)$. By comparing the unconditional risks of the two alleles at each SNP, we recover the odds ratio estimated from a single SNP analysis.

and these previous associations were only found in female and young onset cases. Further analysis is required to determine whether these observations represent an independent association signal.

The 17q21 hit region falls on a known polymorphic 900 kb inversion. The two forms of the inversion, often termed H1 and H2, vary in frequency globally, with the minor H2 haplotype found almost exclusively in Southwest Asian and European populations, at a frequency of 5–30% (11–13). There is some evidence that H2 is under selection in Europe, where it has been linked to higher recombination rate and greater fertility (12). Suggestive associations of the region with Crohn's disease and type 1 diabetes have recently been reported (14). There is very little evidence of recombination between the two orientations of the inversion, but within these, H1 shows relatively normal recombination and variation patterns, whereas there is very little variation within the H2 haplotype (15). Analyses of the age of the inversion event and its history have led to differing conclusions (11,12,16,17). The inversion encompasses several genes, and the most associated SNP in our GWAS, rs7215239, is located in the promoter region of the *MAPT* gene. The minor allele at rs7215239 (G) is protective for PD and is in LD with multiple SNPs over a large region tagging the H2 and H1 haplotypes, an association result consistent with a previous study (18). Within the H1 haplotype, there is genetic diversity, and many sub-haplotypes have been determined (12). To date, results for associations of these sub-haplotypes with PD have shown contradictory findings (19), with one study showing distinct H1 sub-haplotypes associated with PD and supranuclear palsy (18,19) and another study indicating that the PD association at the *MAPT* gene is not due to the different sub-haplotypes of H1 but explained by the H1/H2 inversion (18). However, it has been demonstrated that a subset of H1 haplotypes (referred to as H1c) is associated with increased *MAPT* expression (20).

In order to further investigate the association at this locus, we used a recently published software package, GENECLUSTER (21), to find evidence for causal mutation(s). GENECLUSTER looks for evidence of potential additional association signals by examining the clustering of case and control haplotypes at the tips of a genealogical tree estimated from reference-panel data (here HapMap CEU) at a fine grid of locations. Differential clustering under a particular branch of the tree suggests the possibility of a mutation on that branch which affects disease risk. GENECLUSTER assesses evidence for association in a Bayesian framework, and investigates models where there is one, and separately two, disease-predisposing mutation in the region.

In the 17q21 hit region, the strongest GENECLUSTER evidence for association results in a log10(Bayes factor) of 6.22 under the two-mutation model. At the same position, the one-mutation model log10(Bayes factor) is much smaller at 5.23, which strongly suggests that more than one variant is needed to explain the association.

The most likely two-mutation model identified three haplotype risk backgrounds, corresponding to H2 and two subsets of H1 (Fig. 2). These three groups are well tagged by four SNPs (rs9303521, rs11079711, rs12938476, rs1880756) with the TGCC haplotype defining H2, and GATT, TGTC or TGTT haplotypes defining a subset of H1 haplotypes partitioning it into two risk groups. We phased the study genotypes across the region and labelled individuals according to the alleles carried at the four tagging SNPs. These data supported a three-group model, with H1 haplotypes separated into two risk groups, over a two-group model separating just H1 and H2 (likelihood ratio test $P = 0.003$). Relative to the risk of the H2 haplotype, we estimate one subset of H1 haplotypes (Fig. 2, coloured red) to have a risk of 1.18 (95% CI 1.07–1.30), and the other set of H1 haplotypes (Fig. 2, coloured blue) to have a risk of 1.36 (95% CI 1.23–1.50). For replication, we tested the same three-group model defined by the sets of haplotypes at these SNPs in the French data and

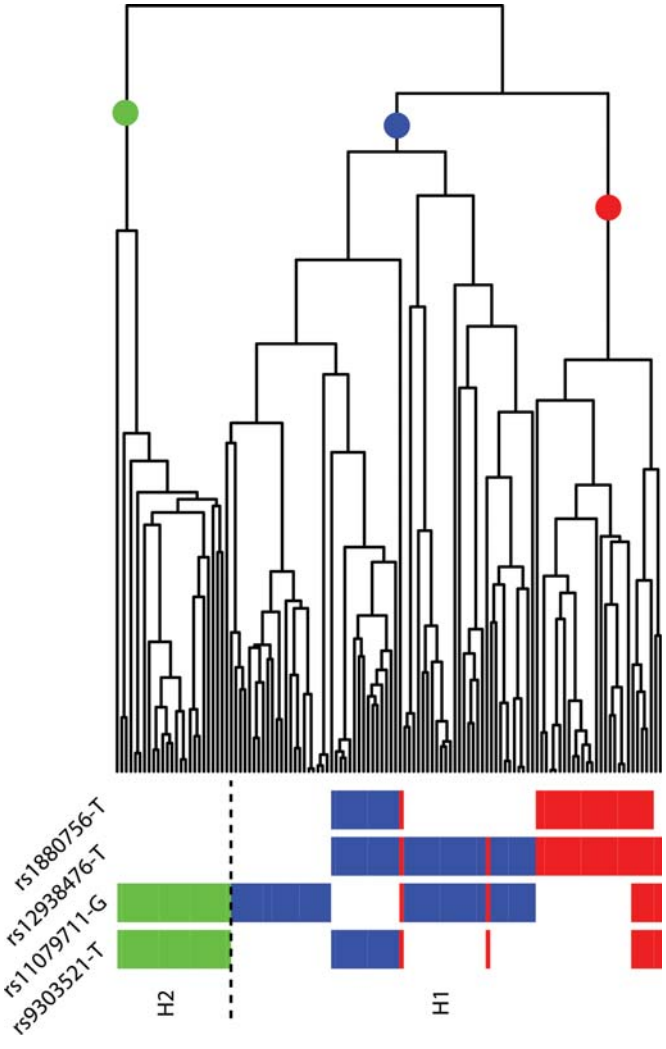

Figure 2. Haplotype tagging of three risk groups identified at the 17q21 locus. The tree shown represents the estimated ancestral relationship between HapMap 2 chromosomes at position 41.19 Mb on chromosome 17 (note that the tree is not scaled relative to time). GENECLUSTER analysis identified three risk groups defined by the braches subtending the blue, green and red circles which represent putative mutation events. Shown on the left are haplotypes delineated by four SNPs chosen to tag the three risk groups. Each chromosome carrying the allele listed by the rsID at the top the figure is coloured according to the haplotype group to which they are assigned. The green haplotypes demark the two orientations labelled H1 and H2, of the large inversion polymorphism encompassing the 17q21 region (see text).

found it significant at $P = 0.026$ (OR = 1.18, 95% CI 1.03–1.35 for the low-risk H1 haplotype and OR = 1.37, 95% CI 1.19–1.58 for the high risk H1 haplotype).

We also investigated whether the two main associations at *SNCA* and *MAPT* showed evidence of interaction, i.e. a departure from the simple model in which the risks for the main SNP at each locus combine multiplicatively. Using a one degree-of-freedom case only test for genotype correlations (22) between the *MAPT* SNP and each of the two *SNCA* SNPs, we found no evidence of interaction ($P > 0.05$).

A further association study was carried out in the PD data, treating the age of onset as a quantitative trait. No SNP passed a stringent significance threshold ($P = 10^{-7}$). However, motivated by the higher prior belief of a potential association with the age of onset for PD-associated loci, we used a less stringent significance threshold for the PD-associated SNPs in the *MAPT* and *SNCA* chromosome regions. This analysis identified an age at onset association for both PD-associated SNPs in *SNCA*. As with the case–control analysis, the evidence at these SNPs is the strongest when they are both included in the model: *F*-test of the two-SNP model gives a *P*-value of 6.37×10^{-4} compared with either rs7687945 ($P = 0.0049$) or rs356220 ($P = 0.189$) when tested separately. The direction of risk at these SNPs is the same as in the case–control analysis (Table 3). Individuals carrying homozygote A alleles at both SNPs have an average 5.48 year earlier onset of PD than those carrying homozygote G alleles at both SNPs. The association of the pair of SNPs with the age of onset was replicated in the French data ($P = 0.02$, with rs2301134 used as a proxy for rs7687945). However, while the direction of effects is the same, the individual effects and marginal significance of the two SNPs differ somewhat in the discovery and replication data sets (Table 3), so we would recommend some caution pending further replication.

DISCUSSION

We undertook a GWAS for PD in 1705 cases and 5175 controls. We found strong support for previously reported associations at 4q22 and 17q21, and support at $P < 0.05$ at 4p15 and 4p16. An earlier reported association around *LRRK2* did not replicate in our study.

We undertook additional analyses at the two major GWAS loci, 4q22/*SNCA* and 17q21/*MAPT*, and in each case uncovered additional signals. At 4q22/*SNCA*, association analysis conditioned on the top SNP revealed a second independent signal which was masked in single-SNP analyses. The presence of the additional signal substantially increases the size of the genetic effect at the locus, with each additional copy of the risk allele at both loci increasing disease risk by a factor of 1.65. *SNCA* is a likely candidate in this region: *SNCA* is a major component of Lewy bodies, the abnormal inclusions in the brain which are important in PD pathology, and previous studies have linked PD risk with *SNCA* overexpression caused by triplications of this gene (2). Our results suggest that *SNCA* risk alleles for PD may well also be associated with earlier disease onset. Previous reports have associated rare copy number variants in *SNCA* with the age of onset, whereby duplications are associated with late onset and triplications associated with early onset (23). Together these observations suggest a key role for *SNCA* in PD susceptibility and progression, which is potentially directly related to gene expression and therefore α-synuclein concentrations (2).

We found that there are at least three haplotype groups at 17q21 with differing risk. Further characterizing these haplotypes and determining the mechanisms behind the several associations reported at the 17q21 locus is an important priority for further work. While *MAPT* is a strong PD candidate gene in 17q21 and *MAPT* has differential expression in the H1 and H2 haplotypes, with higher expression in H1 compared with H2 (20,24), the PD association could well involve another gene. In particular, the H1/H2 genotype is weakly associated with type 1 diabetes and Crohn's disease risk (14) which indicates that variants on this large haplotype may affect multiple pathways.

For the *P*-value threshold of 10^{-4} that we used to select loci for replication, our sample size provides 93% power to detect an allele with a minor allele frequency (MAF) of 20% and an effect size of 1.3. Given the good genome coverage of the Illumina Human660-Quad platform that was used in this study, the absence of additional replicated associations suggests that 4q22/*SNCA* and 17q21/*MAPT* may be the only common variants with effects of this magnitude on PD risk in the UK population. However, and in addition to the well-documented contribution of highly penetrant rare variants for PD risk, our results cannot rule out the presence of a significant number of common associations but with smaller odds ratio. Future pooling of existing case–control studies into large meta-analyses is required to increase the statistical power to detect weaker associations and improve our understanding of the genetic architecture of PD.

MATERIALS AND METHODS

Case and control samples

Prior to any exclusion, the full data set comprised 2190 individuals with idiopathic PD, and 5667 population controls. Known familial cases and individuals with known Mendelian mutations (including *LRRK2* mutations) were excluded. The samples were collected through five UK-wide centres (Supplementary Material, Table S1). Case samples collected in London and Cardiff were screened for the previously reported highly penetrant and rare G0219S mutation in the *LRKK2* gene. The 14 G0219S carriers that we identified were excluded from the GWA study. The control data set was that of the previously described WTCCC2 study (25)—totalling 2930 samples from the 1958 Birth Cohort (58C) and 2737 samples from the UK Blood Services Controls (NBS).

Phenotype definition

All case subjects met the UK Brain Bank Clinical Criteria for PD (26). Of the 1705 samples that progressed through to analysis, age of onset was available for 1439 samples. The mean age of disease onset, as defined by reported age of first motor symptom, was 65.8, with the youngest at age 29 years and the oldest at age 105 years. The male-to-female

Table 3. Evidence of association at the 4q22 locus with age of onset

SNP	rs356220	rs7687945[a]
Position	90860363	90983722
Risk allele	A	A
Discovery sample		
Risk allele frequency (cases, controls)	0.41, 0.36	0.53, 0.51
F-test two-SNP *P*-value	6.62×10^{-4}	
Estimate effect in years	1.17 (0.3–2.0)	1.57 (0.71–2.42)
Replication sample		
F-test two-SNP *P*-value	0.0204	
Estimated effect in years	1.80 (0.51–3.09)	0.80 (0.47–2.06)

Association analysis at two SNPs in the 4q22/SNCA region in the original discovery data and in the French data used for replication. Analysis was performed using a linear model with age of onset as the response variable and the two SNPs as predictors. Note that we report the estimated effect of the A allele at the two SNPs in terms of the reduction in the time to onset of PD.
[a]In the replication French data rs2301134 was used as a proxy for rs7687945 ($r^2 = 0.98$).

ratio in this data was 2.7:1. Samples were excluded if they had previously been shown to have likely causative mutation in a known PD gene. All the Brain Bank cases had pathologically proven PD Braak stages (5,6).

DNA sample preparation

Genomic DNA for all cases was shipped to the Wellcome Trust Sanger Institute (WTSI), Cambridge. DNA concentrations were quantified using a PicoGreen assay (Invitrogen) and an aliquot assayed by agarose gel electrophoresis. A DNA sample was considered to pass quality control if the original DNA concentration was $\geq$50 ng/μl and the DNA was not degraded. In order to track sample identity, ~30 SNPs, including sex chromosome markers, were typed on the Sequenom platform prior to entry to the whole genome genotyping pipeline.

Genotyping methodology and quality control

Genotyping of the samples was carried out at the WTSI on the Illumina BeadArray platform. Cases were genotyped on the Illumina Human660-Quad array and the controls were genotyped on the Illumina 1.2 M Duo array. Normalized probe intensities were exported using the BeadStudio program and genotypes called separately in the 58C, NBS and PD data sets using the program Illuminus (27). For the purposes of quality control, SNPs were excluded from analysis if, in any of the data sets (58C, NBS or PD), they had a MAF less than 0.01%, a significant departure from Hardy Weinberg equilibrium ($P < 10^{-20}$) or a significant association with the plate on which the samples were assayed ($P < 10^{-6}$). We also excluded SNPs for which the observed statistical (Fisher) information about the allele frequency was less than 98% of the information contained in a hypothetical sample of the same size and expected MAF but with no missing data. An additional 39 SNPs were removed following visual inspection of cluster plots. In total, 61 636 SNPs were removed from the overlap set on the two genotyping chips, leaving 532 588 for association analysis. Sample exclusions were based on four genome-wide summary statistics of the genotyping data designed to be sensitive to possibly sources of heterogeneity: fraction of missing genotypes, autosomal heterozygosity, a measure of African and Asian ancestry (defined by a principal component analysis of the HapMap 2 data) and the average difference in the probe intensities across SNPs. By modelling the distribution of each of these summaries as a mixture, we inferred outlying individuals and excluded them from analysis. Furthermore, we exclude one of each pair of individuals showing greater than 5% identity by descent by inferring chromosomal sharing at a genome-wide subset of 11 547 SNPs. To reduce the risk of errors through sample swaps, we also removed samples for which the reported gender and genetically determined gender were discordant, or where Illumina array-based genotypes disagreed with more than 10% of the Sequenom genotypes which were typed as part of sample preparation described above. After sample quality control, 1705 cases and 5175 controls sample remained for analysis (Supplementary Material, Table S2).

Imputation and haplotype phasing

Haplotype phasing and imputation was performed using IMPUTE2 (7), which adopts a two-stage approach using both haploid and diploid reference panels. For the haploid reference panel, we used HapMap2 and HapMap3 SNP data for the 120 non-related CEU trios (see www.hapmap.org), and for the diploid reference we used 58C and NBS control data, merging genotypes from the Illumina 1.2 M Duo chip and Affymetrix Genome Wide Human SNP array 6.0. Prior to analysis with IMPUTE2, we applied standard quality control filters akin to those described above. To further protect against potential errors misleading the imputation and phasing, we checked that each genotype conformed to local patterns of LD in HapMap by employing a leave-one-out imputation strategy. Specifically, we ran IMPUTE (7) on each of the study samples in turn, both cases and controls, re-imputing known genotypes. Control individuals for which the imputed genotypes were more than 4.5% discordant with the original genotype were removed. The same rule was applied to case individuals with a discordance threshold of 6%. SNPs for which IMPUTE was confident of the imputation call but the genotyped data were discordant (and therefore indicative of genotyping error) were also removed if the difference between measure of information and error rate was greater than 0.05.

Statistical analysis

Genome-wide case–control analysis was performed using frequentist tests, under a missing data logistic regression model, as implemented in the program SNPTEST (6). Unless otherwise stated, we assumed a multiplicative model for allelic risk by encoding the genotypes at each SNP as a discrete explanatory variable with an indicator of case status as the binary response. We note that an analogous analysis using the Cochran–Armitage trend test in PLINK (29), ignoring the uncertainty in genotype calls, gave very similar results (data not shown). To look for secondary independent signals

within associated loci, we included the SNP with the lowest trend test *P*-value in the logistic regression model as a discrete covariate using PLINK (28). Likelihood ratio tests were used to compare one-SNP to two-SNP models in order to identify SNPs within the loci with independent effects on risk. For the analysis of haplotypes in the 4q22 and 17q21 regions, we employed logistic regression models to estimate the risk associated with carrying each of the haplotypes or where relevant set of haplotypes, by including a set of indicator variables denoting the haplotypes carried by each study individual. To formally compare the two-group haplotype model to the three-group model at 17q21, we re-encoded haplotypes to form a nested model and tested the need for an indicator of membership of the two risk groups within the H1 background. These analyses were carried out in the statistical package R.

Age of onset analysis was carried out by treating it as a continuous quantitave response in a linear regression model. Genome-wide analysis was performed using frequentist tests in SNPTEST calculated using missing data likelihoods. To look in detail at the combined effect of the two SNPs in 4q22, we reanalysed the data in R. Meta analysis results for both age of onset and case–control analysis were obtained assuming a standard fixed effects model to combine estimates of the odds ratio and standard errors across studies.

It has become standard practice in GWAS to refer to the odds ratio associated with a particular allele or haplotype, which we estimate to be the e^{β}, where β is the maximum likelihood estimate of the coefficient describing the effect of each predictor on the response in the assumed model. We note however that, as is of true this study and many others, where the controls are taken at random from the population without reference to case status, β is actually the log of the *relative risk* and not the log of the odds ratio.

Replication strategy

For *in silico* replication, we exchanged genotype data with a study carried out in the French population using a similar study design. The French data set consisted of a total of 1039 cases and 1984 controls (see Saad *et al*., manuscript in preparation). These samples were typed on the Illumina 610 platform, which has an overlap of 473 892 SNPs with our study. Selecting only from the subset of SNPs which passed quality control in both studies, association test data for 100 SNPs were exchanged. This SNP replication list included the 55 top hit SNPs from our study (Supplementary Material, Table S3), 20 randomly selected control SNPs and 25 SNPs from the two most recently published PD GWAS studies (3,4). Owing to the significant level of population structure in the French GWA scan, the association test included covariates for population structure computed from a principal component analysis (29).

AUTHOR CONTRIBUTIONS

N.W.W., P.F.C., D.B., R.A.B., A.J.L., K.B., C.E.C., K.E.M. were involved in establishing DNA collections, assembling phenotypic data and/or recruiting patients; J.H., N.W.W., N.W.W. supervised clinical and laboratory work; WTCCC2 DNA, Genotyping, Data QC and Informatics group executed GWAS sample handling, genotyping and QC; WTCCC2 Data and Analysis group, M.G., C.C.A.S., A.S., V.P. and P.D. performed statistical analyses; V.P., C.C.A.S., A.S., J.H., P.D. and N.W.W. contributed to writing the manuscript. WTCCC2 Management Committee conceived and oversaw the design and execution of the GWAS. WTCCC2 group memberships are specified in the full author list.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We kindly acknowledge the patients who participated, and the physicians who helped in recruitment. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk.

FUNDING

We acknowledge use of the British 1958 Birth Cohort DNA collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. This work was supported by the Wellcome Trust (WTCCC2, grant numbers 084747 and 083948), and Parkinson's UK (formerly The PD society) (ref no J-0804). We acknowledge funding from the Medical Research Council (G0700943). We acknowledge the work of Dr Mirhdu Wickremaratchi, Ms Victoria Newsway and Ms Dee Perera in collecting and collating patient samples and data. This work was partly undertaken at UCLH/UCL who received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. A.C.V. received financial support from the Department of Health through the award made by the National Institute for Health Research to Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology for a Specialist Biomedical Research Centre for Ophthalmology. V.P. received financial support from the JDRF (Transition award). Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

APPENDIX

The authors of this manuscript are the following.

Chris C.A. Spencer[1,†], Vincent Plagnol[2,†], Amy Strange[1], Michelle Gardner[3], Coro Paisan-Ruiz[3], Gavin Band[1], Roger A. Barker[4], Celine Bellenguez[1], Kailash Bhatia[3], Hannah Blackburn[5], Jennie M. Blackwell[6], Elvira Bramon[23], Martin A. Brown[8], Matthew A. Brown[24], David Burn[9], Juan-Pablo Casas[10], Patrick F. Chinnery[8], Carl E. Clarke[11], Aiden Corvin[12], Nicholas Craddock[13], Panos Deloukas[5], Sarah Edkins[5], Jonathan Evans[4], Colin Freeman[1], Emma Gray[5], John Hardy[3], Gavin Hudson[8], Sarah Hunt[5], Janusz Jankowski[14], Cordelia Langford[5], Andrew J. Lees[3], Hugh S. Markus[15], Christopher G. Mathew[16], Mark I. McCarthy[17,1], Karen E. Morrison[11], Colin N.A.

Palmer[18], Justin P. Pearson[7], Leena Peltonen[5], Matti Pirinen[1], Robert Plomin[19], Simon Potter[5], Anna Rautanen[1], Stephen J. Sawcer[20], Zhan Su[1], Richard C. Trembath[16], Ananth C. Viswanathan[21], Nigel W. Williams[7], Huw R. Morris[22], Peter Donnelly[1,*,†], Nicholas W. Wood[3,*,†]

1. Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK

2. UCL Genetics Institute, Gower Place, London WC1E 6BT, UK

3. UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK

4. Cambridge Centre for Brain Repair, Forvie Site, Robinson Way, Cambridge CB2 2PY, UK

5. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

6. Telethon Institute for Child Health Research, University of Western Australia, 100 Roberts Road, Subiaco, PO Box 855, West Perth, WA 6873, USA and Cambridge Institute for Medical Research, University of Cambridge School of Clinical Medicine, Cambridge CB2 0XY, UK

7. Department of Psychological Medicine and Neurology, School of Medicine, Cardiff University and MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff CF14 4XN, UK

8. Neurology M4104, The Medical School, Framlington Place, Newcastle upon Tyne NE2 4HH, UK

9. Institute for Ageing and Health, Newcastle University, Clinical Ageing Research Unit, Campus for Ageing and Vitality, Newcastle upon Tyne NE4 5PL, UK

10. Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK and Department of Epidemiology and Public Health, UCL, 1–19 Torrington Place, London WC1E 6BT, UK

11. School of Clinical and Experimental Medicine, College of Medicine and Dentistry, University of Birmingham and Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham B15 2TT, UK

12. Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin 2, Eire

13. Psychological Medicine, Cardiff University School of Medicine, Heath Park, Cardiff CF14 4XN, UK

14. Centre for Gastroenterology, Bart's and the London School of Medicine and Dentistry, London E1 2AT, UK

15. Division of Cardiac and Vascular Sciences, Department of Clinical Neurosciences, St George's Hospital, London SW17 0RE, UK

16. Department of Medical and Molecular Genetics, King's College London School of Medicine, Guy's Hospital, London SE1 9RT, UK

17. Oxford Centre for Diabetes, Endocrinology and Metabolism (OCDEM), University of Oxford, Churchill Hospital, Oxford OX3 7LJ, UK

18. Biomedical Research Institute, Ninewells Hospital and Medical School, Dundee DD1 9SY, UK

19. Social, Genetic and Developmental Psychiatry Centre, King's College London Institute of Psychiatry, Denmark Hill, London SE5 8AF, UK

20. Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK

21. NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK

22. Department of Neurology, University Hospital of Wales Cardiff, CF14 4XN, UK

23. Department of Psychosis Studies, NIHR Biomedical Research Centre for Mental Health at the Institute of Psychiatry, King's College London and The South London and Maudsley NHS Foundation Trust, Denmark Hill, London SE5 8AF, UK

24. University of Queensland Diamantina Institute, Princess Alexandra Hospital, Brisbane, Queensland, Australia

25. Department of Clinical Pharmacology, University of Oxford, OX3 3HE, UK

26. Leicester, Digestive Disease Centre, LE1 5WW, UK

*Corresponding authors. These authors jointly supervised this work.

†Contributed equally to this manuscript.

REFERENCES

1. Gilks, W., Abou-Sleiman, P., Gandhi, S., Jain, S., Singleton, A., Lees, A., Shaw, K., Bhatia, K., Bonifati, V., Quinn, N *et al.* (2005) A common LRRK2 mutation in idiopathic Parkinson's disease. *Lancet*, **365**, 415–416.
2. Singleton, A.B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R. *et al.* (2003) alpha-Synuclein locus triplication causes Parkinson's disease. *Science*, **302**, 415–416.
3. Satake, W., Nakabayashi, Y., Mizuta, I., Hirota, Y., Ito, C., Kubo, M., Kawaguchi, T., Tsunoda, T., Watanabe, M., Takeda, A. *et al.* (2009) Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat. Genet.*, **41**, 1303–1307.
4. Simon-Sanchez, J., Schulte, C., Bras, J., Sharma, M., Gibbs, R., Berg, D., Paisan-Ruiz, C., Lichtner, P., Scholz, S., Hernandez, D. *et al.* (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.*, **41**, 1308–1312.
5. Edwards, T.L., Scott, W.K., Almonte, C., Burt, A., Powell, E.H., Beecham, G.W., Wang, L., Zuchner, S., Konidari, I., Wang, G. *et al.* (2010) Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. *Ann. Hum. Genet.*, **74**, 97–109.
6. Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.
7. Howie, B.N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.
8. Blyth, C.R. (1972) On Simpson's paradox and the sure thing principle. *J. Am. Stat. Assoc.*, **67**, 364–366.
9. Simpson, E.H. (1951) The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. (Ser. B)*, **13**, 238–241.
10. Mueller, J., Fuchs, J., Hofer, A., Zimprich, A., Lichtner, P., Illig, T., Berg, D., Wüllner, U., Meitinger, T. and Gasser, T. (2005) Multiple regions of alpha-synuclein are associated with Parkinson's disease. *Ann. Neurol.*, **57**, 535–541.
11. Donnelly, M.P., Paschou, P., Grigorenko, E., Gurwitz, D., Mehdi, S.Q., Kajuna, S.L., Barta, C., Kungulilo, S., Karoma, N.J., Lu, R.B. *et al.* (2010) The distribution and most recent common ancestor of the 17q21 inversion in humans. *Am. J. Hum. Genet.*, **86**, 161–71.
12. Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. *et al.* (2005) A common inversion under selection in Europeans. *Nat. Genet.*, **37**, 129–137.
13. Evans, W., Fung, H.C., Steele, J., Eerola, J., Tienari, P., Pittman, A., Silva, R., Myers, A., Vrieze, F.W.-D., Singleton, A. *et al.* (2004) The tau H2 haplotype is almost exclusively Caucasian in origin. *Neurosci. Lett.*, **369**, 183–185.
14. Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., Giannoulatou, E. *et al.* (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**, 713–720.

15. Pittman, A.M., Myers, A.J., Duckworth, J., Bryden, L., Hanson, M., Abou-Sleiman, P., Wood, N.W., Hardy, J., Lees, A. and de Silva, R. (2004) The structure of the tau haplotype in controls and in progressive supranuclear palsy. *Hum. Mol. Genet.*, **13**, 1267–1274.
16. Cruts, M., Rademakers, R., Gijselinck, I., van der Zee, J., Dermaut, B., de Pooter, T., de Rijk, P., Del-Favero, J. and van Broeckhoven, C. (2005) Genomic architecture of human 17q21 linked to frontotemporal dementia uncovers a highly homologous family of low-copy repeats in the tau region. *Hum. Mol. Genet.*, **14**, 1753–1762.
17. Zody, M.C., Jiang, Z., Fung, H.C., Antonacci, F., Hillier, L.W., Cardone, M.F., Graves, T.A., Kidd, J.M., Cheng, Z., Abouelleil, A. *et al.* (2008) Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.*, **40**, 1076–1083.
18. Vandrovcova, J., Pittman, A., Malzer, E., Abou-Sleiman, P., Lees, A., Wood, N. and de Silva, R. (2009) Association of MAPT haplotype-tagging SNPs with sporadic Parkinson's disease. *Neurobiol. Aging*, **30**, 1477–1482.
19. Ezquerra, M., Pastor, P., Gaig, C., Vidal-Taboada, J.M., Cruchaga, C., Munoz, E., Marti, M.J., Valldeoriola, F., Aguilar, M., Calopa, M. *et al.* (2009) Different MAPT haplotypes are associated with Parkinson's disease and progressive supranuclear palsy. *Neurobiol. Aging* (in press).
20. Myers, A., Pittman, A., Zhao, A., Rohrer, K., Kaleem, M., Marlowe, L., Lees, A., Leung, D., McKeith, I., Perry, R. *et al.* (2007) The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts. *Neurobiol. Dis.*, **25**, 561–570.
21. Zhan Su, N.C., the Wellcome Trust Case Control Consortium, Donnelly, P. and Marchini,, J. (2009) A Bayesian method for detecting and characterizing allelic heterogeneity and boosting signals in genome-wide association studies. *Stat. Sci.*, **24**, 430–450.
22. Cordell, H. (2009) Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.
23. Fuchs, J., Nilsson, C., Kachergus, J., Munz, M., Larsson, E.M., Schule, B., Langston, J.W., Middleton, F.A., Ross, O.A., Hulihan, M. *et al.* (2007) Phenotypic variation in a large Swedish pedigree due to SNCA duplication and triplication. *Neurology*, **68**, 916–922.
24. Pittman, A.M., Myers, A.J., Abou-Sleiman, P., Fung, H.C., Kaleem, M., Marlowe, L., Duckworth, J., Leung, D., Williams, D., Kilford, L. *et al.* (2005) Linkage disequilibrium fine mapping and haplotype association analysis of the tau gene in progressive supranuclear palsy and corticobasal degeneration. *J. Med. Genet.*, **42**, 837–846.
25. Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
26. Hughes, A.J., Daniel, S.E., Kilford, L. and Lees, A.J. (1992) Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiatry*, **55**, 181–184.
27. Teo, Y., Inouye, M., Small, K., Gwilliam, R., Deloukas, P., Kwiatkowski, D. and Clark, T. (2007) A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*, **23**, 2741–2746.
28. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
29. Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.