

Evidence for More than One Parkinson's Disease-Associated Variant within the *HLA* Region

Erin M. Hill-Burns¹, Stewart A. Factor², Cyrus P. Zabetian³, Glenys Thomson⁴, Haydeh Payami^{1*}

1 New York State Department of Health Wadsworth Center, Albany, New York, United States of America, **2** Department of Neurology, Emory University School of Medicine, Atlanta, Georgia, United States of America, **3** Veteran's Affairs Puget Sound Health Care System and Department of Neurology, University of Washington, Seattle, Washington, United States of America, **4** Department of Integrative Biology, University of California, Berkeley, California, United States of America

Abstract

Parkinson's disease (PD) was recently found to be associated with *HLA* in a genome-wide association study (GWAS). Follow-up GWAS's replicated the PD-*HLA* association but their top hits differ. Do the different hits tag the same locus or is there more than one PD-associated variant within *HLA*? We show that the top GWAS hits are not correlated with each other ($0.00 \leq r^2 \leq 0.15$). Using our GWAS (2000 cases, 1986 controls) we conducted step-wise conditional analysis on 107 SNPs with $P < 10^{-3}$ for PD-association; 103 dropped-out, four remained significant. Each SNP, when conditioned on the other three, yielded $P_{\text{SNP1}} = 5 \times 10^{-4}$, $P_{\text{SNP2}} = 5 \times 10^{-4}$, $P_{\text{SNP3}} = 4 \times 10^{-3}$ and $P_{\text{SNP4}} = 0.025$. The four SNPs were not correlated ($0.01 \leq r^2 \leq 0.20$). Haplotype analysis (excluding rare SNP2) revealed increasing PD risk with increasing risk alleles from $\text{OR} = 1.27$, $P = 5 \times 10^{-3}$ for one risk allele to $\text{OR} = 1.65$, $P = 4 \times 10^{-8}$ for three. Using additional 843 cases and 856 controls we replicated the independent effects of SNP1 ($P_{\text{conditioned-on-SNP4}} = 0.04$) and SNP4 ($P_{\text{conditioned-on-SNP1}} = 0.04$); SNP2 and SNP3 could not be replicated. In pooled GWAS and replication, SNP1 had $\text{OR}_{\text{conditioned-on-SNP4}} = 1.23$, $P_{\text{conditioned-on-SNP4}} = 6 \times 10^{-7}$; SNP4 had $\text{OR}_{\text{conditioned-on-SNP1}} = 1.18$, $P_{\text{conditioned-on-SNP1}} = 3 \times 10^{-3}$; and the haplotype with both risk alleles had $\text{OR} = 1.48$, $P = 2 \times 10^{-12}$. Genotypic OR increased with the number of risk alleles an individual possessed up to $\text{OR} = 1.94$, $P = 2 \times 10^{-11}$ for individuals who were homozygous for the risk allele at both SNP1 and SNP4. SNP1 is a variant in *HLA-DRA* and is associated with *HLA-DRA*, *DRB5* and *DQA2* gene expression. SNP4 is correlated ($r^2 = 0.95$) with variants that are associated with *HLA-DQA2* expression, and with the top *HLA* SNP from the IPDGC GWAS ($r^2 = 0.60$). Our findings suggest more than one PD-*HLA* association; either different alleles of the same gene, or separate loci.

Citation: Hill-Burns EM, Factor SA, Zabetian CP, Thomson G, Payami H (2011) Evidence for More than One Parkinson's Disease-Associated Variant within the *HLA* Region. PLoS ONE 6(11): e27109. doi:10.1371/journal.pone.0027109

Editor: Mathias Toft, Oslo University Hospital, Norway

Received: July 8, 2011; **Accepted:** October 10, 2011; **Published:** November 9, 2011

Copyright: © 2011 Hill-Burns et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Study was supported by grants from the National Institute of Neurological Disorders And Stroke (NINDS R01 NS36960 and R01 NS067469 to HP), the Edmond J. Safra Michael J. Fox Foundation Global Genetic Consortium Initiative (HP), the Department of Veterans Affairs (101BX000531 to CZ), The Close to a Cure Foundation: A Fund for Parkinson's Research of Foundation for the Carolinas (SF), and the National Institute of Allergy and Infectious Disease (AI40076, GT). Funding support for the replication dataset was provided by the National Institute of Neurological Disorders and Stroke (Foroud/Myers, PI). 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.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: hpayami@wadsworth.org

Introduction

The recent discovery of an association between PD and *HLA* was made in a hypothesis neutral genome-wide association study (GWAS) [1]. Historically, *HLA*-disease associations have been conducted with the highly polymorphic "classical" *HLA* loci; i.e., those that encode the diversity for antigen recognition; whereas, the association peak in our GWAS was in *HLA-DRA* which is practically monomorphic and hence not normally investigated for disease associations. The finding that genetic variants in immune response affect risk of developing PD, firmly grounds, at the DNA level, the long held notion that the immune system and inflammation play a significant role in PD [2].

Our original GWAS that uncovered an association between PD and *HLA* was performed with the NeuroGenetic Research Consortium (NGRC) data. NGRC is a single data set (2000 persons with PD, 1986 control volunteers) which was collected using uniform protocols for all study procedures including subject selection and diagnosis, data collection, genotyping and data analysis [1]. The NGRC GWAS revealed a spike at *HLA* for association with PD that

reached genome-wide significance. The association peak was at rs3129882, a SNP in intron 1 of *HLA-DRA* which had previously been shown to associate with variation in expression of *HLA-DRA*, *DRB5* and *DQA2* [3,4]. Association of rs3129882 with PD had an odds ratio (OR) = 1.31, and $P = 3 \times 10^{-8}$ in discovery and was replicated in independent datasets in the same study [1]. The *HLA* region spike in the NGRC data included 107 SNPs that achieved P values of 10^{-3} to 3×10^{-8} . A subsequent GWAS conducted in the Dutch population (772 cases, 2024 controls) confirmed the association of PD with *HLA* [5]. Their most significant SNP was rs4248166; $\text{OR} = 1.36$, $P = 4 \times 10^{-5}$ which also maps to the *HLA* class II region. The involvement of the *HLA* region in PD was also confirmed by the International Parkinson Disease Genomics Consortium (IPDGC) meta-analysis [6], which identified chr6:32588205 in the *HLA* class II region as the most significant SNP in their discovery sample (5333 cases, 12019 controls) with $\text{OR} = 0.70$, $P = 3 \times 10^{-8}$ and in their replication sample (7053 cases, 9007 controls) $\text{OR} = 0.80$, $P = 9 \times 10^{-8}$.

It is not unexpected that different GWAS's are identifying different *HLA* SNPs; arrays with different SNPs were used.

Table 1. HLA SNPs that have shown the most significant associations with PD in three GWAS's.

GWAS	N cases	N controls	SNP	Gene	BP	MAF cases	MAF controls	OR	P
NGRC [1]	2000	1986	rs3129882	<i>DRA</i>	32517508	0.46	0.40	1.31	3×10^{-8}
Dutch [5]	772	2024	rs4248166	<i>BTNL2</i>	32474399	0.21	0.17	1.36	4×10^{-5}
IPDGC [6]	5333	12019	chr6:32588205	3' of <i>DRB5</i>	32588205	(0.15 unspecified)	0.70	3×10^{-8}	

doi:10.1371/journal.pone.0027109.t001

However, do the different top SNPs from various studies all tag the same locus or could there be more than one PD-associated susceptibility variant in *HLA*? The aim of this study was to explore this question.

Results

HLA hits in three GWAS's

The most significant *HLA* SNPs from the three GWAS's including and subsequent to our original report are shown in **Table 1**. They span a ~100 kb region in the *HLA* class II region. We expected to find strong linkage disequilibrium (LD) (measured by r^2 , where r is the correlation coefficient) among them, assuming all three are tagging the same PD-susceptibility locus. Surprisingly, there was very little correlation between the SNPs as evidenced by pair wise $r^2 = 0.00, 0.09$ and 0.15 (**Figure 1**).

Conditional analysis in NGRC GWAS

To gain a better understanding of the *HLA* association with PD, we performed step-wise conditional analysis in the NGRC GWAS. Conditional analysis allows testing all the SNPs in the region (or a chosen subset, here according to statistical significance), to identify the most significant one, and repeating the analysis conditioned on the most significant SNP to see if others are significant in addition to the top SNP [7,8,9]. The process is repeated, each time conditioning on all SNPs that emerged as most significant in prior rounds, until all SNPs whose significance is dependent on other SNPs are identified and removed. We performed conditional analysis on 107 *HLA* SNPs that had achieved $P < 0.001$ for PD-

association in the NGRC GWAS [1], using PLINK v1.07 software [10] (**Table S1**). In round 1, when analysis was performed conditioned on the single most significant SNP (SNP1, rs3129882), 90 of 106 SNPs lost significance while 16 SNPs remained significant at $P < 0.05$. The SNP with the lowest P value (rs3993757, $P = 0.002$) was marked SNP2. When the analysis was repeated (round two) conditioned on both SNP1 and SNP2 for the 15 SNPs that survived round one, 13 were associated with PD with $P < 0.05$, and the most significant SNP of this analysis (rs2844505, $P = 0.006$) was marked SNP3. When the analysis was repeated with the 12 remaining SNPs conditioning now on SNP1 and SNP2 and SNP3 (round three), only one SNP had $P < 0.05$ (rs9268515, $P = 0.025$) and it was marked SNP4. Summary of the results are shown in **Table 2**. For the full analysis see **Table S1**.

Step-wise conditional analysis revealed four *HLA* SNPs with seemingly independent effects on PD. We re-tested association of each of the four SNPs with PD conditioning on the other three. We obtained $P = 5 \times 10^{-4}$ for SNP1 conditioned on SNP2 and SNP3 and SNP4, $P = 5 \times 10^{-4}$ for SNP2 conditioned on SNP1 and SNP3 and SNP4, $P = 4 \times 10^{-3}$ for SNP3 conditioned on SNP1 and SNP2 and SNP4, and $P = 0.025$ for SNP4 conditioned on SNP1 and SNP2 and SNP3.

Interaction among SNPs 1–4 in NGRC

We tested for and did not find significant evidence for interaction among the four SNPs. The full model testing all pair-wise interactions among the four SNPs compared with a model with no interactions yielded $P = 0.9$. Testing each pair-wise interaction, with all other SNPs in the model as covariates, yielded $P = 0.2$ – 0.7 . Lack of evidence for interaction could have been due to insufficient power (discussed further under Replication).

Linkage Disequilibrium

We examined LD between the four SNPs that withstood conditional analysis. LD was measured as D' and r^2 (correlation coefficient) [11]. In the context of disease association, r^2 is commonly used to assess correlation among SNPs (see for example [8]). Using the NGRC data for estimating LD, pair-wise D' for the four NGRC SNPs ranged from 0.17 to 0.75 and r^2 ranged from 0.00 to 0.09. Using the 1000 Genomes Project data for estimating LD (to allow inclusion of IPDGC SNP) the NGRC SNPs 1–4 had $D' = 0$ to 0.88 and $r^2 = 0.01$ to 0.20 (**Figure 2**). In relation to the top SNPs from other GWAS's (**Figure 2**), SNP1, SNP2, and SNP3 showed little or no correlation with them ($0.00 \leq r^2 \leq 0.15$), whereas SNP4 was moderately correlated with the top SNP from IPDGC ($r^2 = 0.60$).

Haplotype analysis

We performed haplotype analysis for the SNPs that had emerged from NGRC conditional analysis (SNP1, SNP3 and SNP4; SNP2 was not included because its minor allele was rare and none of the haplotypes carrying the minor allele of SNP2 had

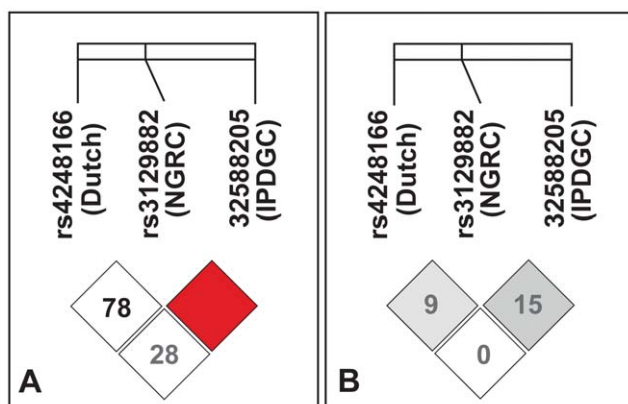


Figure 1. Linkage disequilibrium (LD) among top *HLA* associations from three PD GWAS's. LD was measured as D' (panel A) and as r^2 (panel B) among rs3129882 from the NGRC GWAS [1], rs4248166 from Dutch GWAS [5], and 6–32588205 from IPDGC meta analysis [6]. Numbers in cells are % D' (panel A) and % r^2 (panel B). The red cells with no number reflects $D' = 1.0$.

doi:10.1371/journal.pone.0027109.g001

Table 2. Step-wise conditional analysis.

SNP	BP	Minor/ Major Allele	MAF cases	MAF controls	HWE P	Unconditioned (GWAS results for SNPs that survived conditional analysis, see Table S1 for full data)		Conditioned on SNP1		Conditioned on SNP1 & SNP2		Conditioned on SNP1 & SNP2 & SNP3	
						OR	P	OR	P	OR	P	OR	P
SNP1	rs3129882	32517508	G/A	0.46	0.40	0.82	1.31						
SNP2	rs3993757	31698725	T/C	0.03	0.02	1.00	1.79	1.70	2×10^{-3}				
SNP3	rs2844505	31547042	G/A	0.29	0.25	0.01	1.23	1.15	0.011	1.16	6×10^{-3}		
SNP4	rs9268515	32487273	C/G	0.16	0.20	0.33	1.25	1.14	0.049	1.16	0.028	1.16	0.025
Number of SNPs remaining significant at the end of each round						107 with $P < 10^{-3}$ in GWAS (Table S1)		16 with $P < 0.05$		13 with $P < 0.05$		1 with $P < 0.05$	

Step-wise conditional analysis was performed for 107 SNPs in the *HLA* region that achieved $P < 10^{-3}$ in GWAS. The full analysis is shown in **Table S1**. Here, we show the summary results for the four SNPs that remained significant after conditioning on the other significant SNPs. For consistency, we show all odds ratios (OR) on the positive side (i.e., testing risk allele against the alternate allele). The risk allele at each SNP is shown in bold. All association tests were adjusted for age at enrollment, sex, and PC1 and PC2 (principal components that define Jewish/non-Jewish origin and the European country of ancestry). Once the four SNPs that retain conditioned $P < 0.05$ were identified, we re-tested association of each of the SNPs with PD conditioning on the other three. We obtained $P = 5 \times 10^{-4}$ for SNP1 conditioned on SNP2 and SNP3 and SNP4, $P = 5 \times 10^{-4}$ for SNP2 conditioned on SNP1 and SNP3 and SNP4, $P = 4 \times 10^{-3}$ for SNP3 conditioned on SNP1 and SNP2 and SNP4, and $P = 0.025$ for SNP4 conditioned on SNP1 and SNP2 and SNP3. BP = base pair position of the SNP on chromosome 6. Minor/major allele = the two alternative nucleotides at the SNP, the one with higher frequency denoted as major allele. MAF = minor allele frequency. HWE $P = P$ value for the test of Hardy-Weinberg Equilibrium.
doi:10.1371/journal.pone.0027109.t002

a frequency above 0.01) (**Table 3**). We set the haplotype that carries the lower risk allele for each SNP (AAC) as the reference and calculated relative effects of each haplotype on PD risk. (High risk alleles are shown in bold for ease of reading). The high risk allele was the minor allele for SNP1 and SNP3, and the major allele for SNP4. The haplotype with the most significant association with PD was **GGG** which has the high risk alleles at all three SNPs (OR = 1.65, 95%CI = 1.38–1.97, $P = 4 \times 10^{-8}$). The next most significant haplotype was **GAG** with two high risk alleles at SNP1 and SNP4 (OR = 1.57, 95%CI = 1.32–1.86, $P = 2 \times 10^{-7}$), followed by **AGG** with two high risk alleles at SNP3 and SNP4 (OR = 1.56, 95%CI = 1.22–1.99, $P = 4 \times 10^{-4}$), and then **AAG** with one high risk allele at SNP4 (OR = 1.27, 95%CI = 1.07–1.50, $P = 5 \times 10^{-3}$). **GAC** had the highest OR estimate despite having the risk allele only at SNP1 (OR = 1.75, 95%CI = 1.07–2.87, $P = 0.03$); note however that the confidence interval was large due to the low frequency of this haplotype.

Replication

We attempted to replicate the following observations: (a) SNPs 1–4 are associated with PD; (b) association of each SNP with PD remains significant when conditioned on the other three SNPs; (c) haplotype analyses will reveal similar pattern of increasing risk with increasing risk alleles as seen in our NGRC results. We performed the replication in an independent dataset (843 cases and 856 controls) that is published [12] and publicly available on the NIH database of Genotypes and Phenotypes (dbGaP, accession number phs000126.v1.p1).

We replicated the association of PD with SNP1 (rs3129882, OR = 1.20, $P = 0.006$) and with SNP4 (rs9268515, OR = 1.25, $P = 0.004$); SNP3 did not replicate (**Table 4**). SNP2 was not genotyped in that dataset and could not be imputed. SNP1 and SNP4 remained significant in replication when conditioned on each other (SNP1 OR = 1.15, $P = 0.04$; SNP4 OR = 1.19, $P = 0.04$, **Table 4**).

We performed haplotype analysis for SNP1 and SNP4. Since the NGRC haplotype analysis was performed with three SNPs, for a direct comparison, we repeated the NGRC haplotype analysis with SNP1 and SNP4, leaving out SNP3. The NGRC and replication

haplotype analyses yielded similar results (**Table 5**), with the most significant effect for the haplotype **GG** with two risk alleles (OR_{NGRC} = 1.51, P_{NGRC} = 1×10^{-9} , OR_{Replication} = 1.41, P_{Replication} = 2×10^{-4} , OR_{Pooled} = 1.48, P_{Pooled} = 2×10^{-12}), followed by **AG** (OR_{NGRC} = 1.23, P_{NGRC} = 3×10^{-3} , OR_{Replication} = 1.28, P_{Replication} = 0.01, OR_{Pooled} = 1.25, P_{Pooled} = 2×10^{-4}); the results for **GC** are less reliable due to the smaller sample size (**Table 5**).

Test of interaction between SNP1 and SNP4 in the combined NGRC and replication data yielded OR_{Interaction} = 1.12, P_{Interaction} = 0.19. If interaction exists, it is weak (OR_{interaction} ~ 1.12) and would require a larger sample size than available here to achieve significance. Our power to detect OR_{interaction} = 1.12 with $P = 0.05$ was 42%. We had 80% power to detect OR_{interaction} ≥ 1.20 , and 90% power to detect OR_{interaction} ≥ 1.23 , using the combined NGRC and replication data. Therefore, we should have been able to detect interactions with moderate or high magnitude.

The effects of SNP1 and SNP4 appeared to be additive, as suggested by the OR for genotypic combinations, increasing incrementally with the number of risk alleles an individual possessed up to OR = 1.94, $P = 2 \times 10^{-11}$ for individuals who were homozygous for the risk allele at both SNP1 and SNP4 (**Table 6**).

Discussion

It is intriguing that different GWAS's have identified different SNPs as their top PD associated SNP in the *HLA* region. We questioned whether they tag the same susceptibility locus. In this study, we demonstrated that the top hits from different studies are *not* strongly correlated with each other. Low correlation among PD-associated SNPs does not rule out the possibility that they tag the same locus, but it does raise the possibility that there may be more than one PD signal in *HLA*. Using step-wise conditional analyses on the NGRC data we uncovered four seemingly independent signals for PD within the *HLA* region. There was no correlation and no detectable interaction among the four NGRC SNPs. Two of the four NGRC SNPs, rs3129882 (SNP1) and rs9268515 (SNP4), were replicated in an independent dataset. We have therefore demonstrated that there are at least two *HLA*-PD associations that cannot be explained by LD. These signals

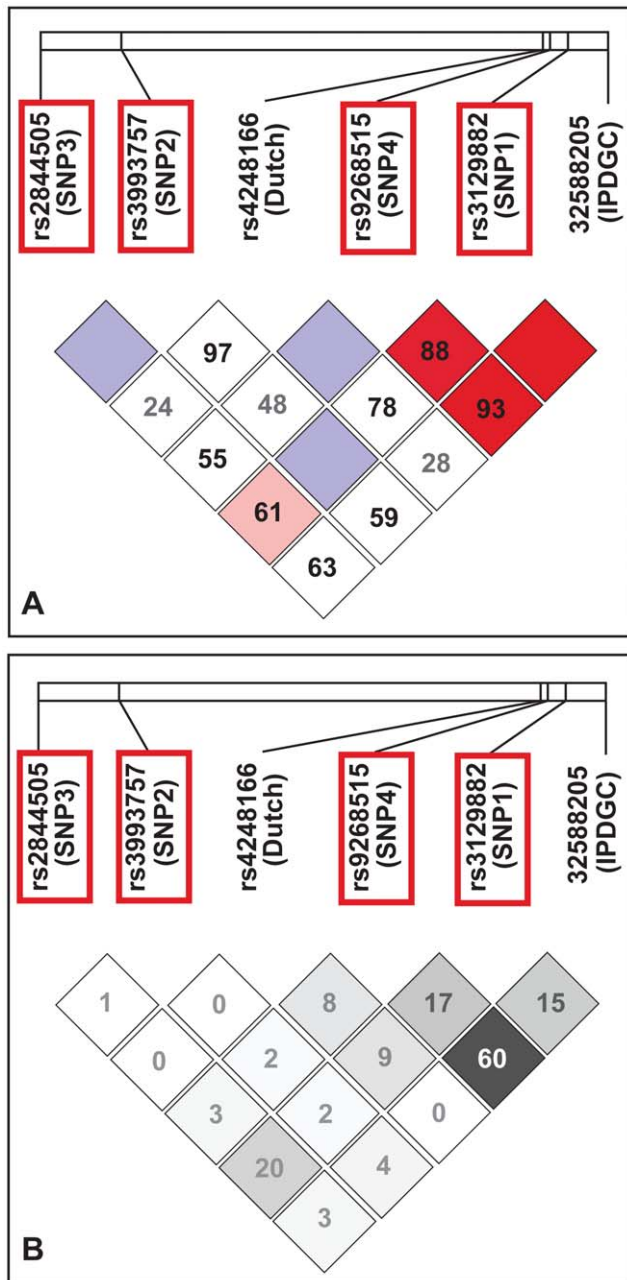


Figure 2. LD among the four significant SNPs from NGRC conditional analysis and the most significant HLA SNPs from other PD GWAS's. LD was measured as D' (panel A) and as r^2 (panel B). NGRC SNPs are numbered 1–4 and marked in red frame. The most significant HLA SNP from other GWAS's are indicated by rs number and the study name. Numbers in cells are % D' (panel A) and % r^2 (panel B). Red cell without a number reflects $D' = 1$, blue cells without a number reflect $D' = 0$.

doi:10.1371/journal.pone.0027109.g002

may indicate different PD-associated alleles at a single susceptibility gene, or different PD susceptibility loci. In support of the different alleles in the same gene hypothesis, there are known examples of multiple disease-associated alleles at the same gene for HLA-associated diseases including type1 diabetes [13] and multiple sclerosis [14], as well as multiple association signals for PD within one gene, notably, *SNCA* [8,15]. Alternatively, the signals may represent different genes. SNP1, rs3129882, the original genome-

wide significant finding in NGRC, is an expression quantitative trait locus (eQTL). It is in intron 1 of *HLA-DRA* and is associated with expression levels of *HLA-DRA*, *DRB5* and *DQA2* genes [3,4]. SNP4, rs9268515, is highly correlated ($r^2 = 0.95$) with three eQTL SNPs (rs3793127, rs3763309, rs3763312) that are associated with expression levels of *HLA-DQA2* [4]. SNP4 also correlates with the top SNP of IPDGC ($r^2 = 0.60$), which maps between *DRA* and *DRB5*, and could be tagging any of the classical HLA genes in the *DR-DQ* region due to their high LD. This raises the intriguing possibility that PD may be associated with both regulatory elements that influence HLA class II gene expression, and with a classical HLA class II allele.

The allele frequency for SNP1 varies significantly in Caucasian Americans according to the European country from which their ancestors immigrated to the US [1]. Furthermore, SNP1 is more strongly associated with sporadic PD than familial PD [1]. Therefore, depending on subject selection, a study may find a positive association (as reported in NGRC), no association, or even an inverse association between SNP1 and PD. We used principal components to correct for ethnic and geographic variability. The subpopulations that differed were the Jewish and the Irish (defined by self-report and verified by principal components [1]). SNP1 and SNP4 remained significant when we excluded the Jewish and the Irish individuals. A recent study showed an inverse association between SNP1 and PD in a mixed Irish and Polish population [16]. It is noteworthy that in the original NGRC GWAS report [1], the Americans of Irish-descent also showed an inverse association between SNP1 and PD, while all other European-descent subpopulations showed a positive association (number of individuals of Polish-descent were few). The Irish/Polish study found the association using a recessive model, as compared to additive model used in NGRC. This is also in agreement with the NGRC data, because the recessive model projects a larger effect size than additive model because it compares individuals who are homozygous for the risk allele to all others.

SNP2 and SNP3 are more than 800 kb away from SNP1 and SNP4. SNP2 had a low minor allele frequency and thus could not be studied in any detail. SNP3 indicated a potential third signal in the NGRC data, but it did not replicate. Much larger sample sizes will be required to determine if and how SNP2 and SNP3 affect susceptibility to PD.

SNP4 was the last of the four seemingly independent SNPs to show up in the conditional analysis of the NGRC data with a modest $P = 0.025$; and it was the only one of the four NGRC SNPs that showed any correlation with the top HLA hit of IPDGC. The direction of the SNP4 effect was the same as the IPDGC SNP; they reported reduced risk with the minor allele, we report increased risk with the major allele (we used the risk alleles throughout the text and the main tables to keep the ORs consistently in the positive direction). We used a liberal $P < 0.05$ given the exploratory nature of the study and that we would follow with a replication study. That our last significant HLA SNP tagged the most significant HLA SNP in IPDGC gives credence to SNP4 being a true association rather than a false positive signal due to a relaxed significance threshold. SNP4 was genotyped in NGRC and imputed in the replication study. The Impute Information score was 0.95, suggesting relatively high reliability. Cases and controls were imputed together under the same conditions, using the same method, and independently of NGRC, and the results showed significantly different allele frequencies between cases and controls for SNP4, in line with the original observation in NGRC. Thus it is unlikely that the results for replication were severely skewed by the fact that SNP4 was imputed; however, replication by other studies will help clarify further.

Table 3. Haplotype analysis.

Haplotypes SNP1 (rs3129882) SNP3 (rs2844505) SNP4 (rs9268515)	Freq. Cases	Freq. Controls	OR (95% CI)	P
AAC	0.11	0.15	Reference	
AAG	0.33	0.36	1.27 (1.07–1.50)	5×10^{-3}
AGC	0.03	0.03	1.34 (0.93–1.91)	0.11
GAC	0.02	0.01	1.75 (1.07–2.87)	0.03
AGG	0.07	0.06	1.56 (1.22–1.99)	4×10^{-4}
GGC	0.006	0.007	-	-
GAG	0.26	0.23	1.57 (1.32–1.86)	2×10^{-7}
GGG	0.18	0.15	1.65 (1.38–1.97)	4×10^{-8}

Haplotypes were composed of SNP1, SNP3 and SNP4, shown in that order. These SNPs and SNP2 survived conditional analysis at $P < 0.05$. SNP2 had low frequency and any haplotype with minor allele of SNP2 was too infrequent (< 0.01) to be included in haplotype association tests. For each SNP, the allele that was associated with higher risk for PD (see **Table 2**) is shown in bold. OR and P values were calculated for each haplotype relative to AAC haplotype which has the lowest risk for PD. Analyses were adjusted for age at enrollment, sex, PC1, and PC2 (principal components that define Jewish/non-Jewish origin and the European country of ancestry). AAG consists of major alleles for each SNP; GGC consists of minor alleles for each SNP. Haplotype GGC was not tested because its frequency was < 0.01 . doi:10.1371/journal.pone.0027109.t003

We do not know if *HLA* plays a larger role for PD risk for certain individuals and perhaps little or no role for others. *HLA* may be involved in a subtype of PD; for example, some cases of PD may be due to infection [17] or autoimmunity [18]. Alternatively, *HLA* may have a ubiquitous role in all PD perhaps via inflammatory response to a variety of causes [19]. The two possibilities are not mutually exclusive; i.e., it is possible that *HLA* and the immune system affect PD pathogenesis in more than one way.

Our results illustrate the utility of conditional analyses and examination of LD structure in replication studies. These exploratory studies can be used to investigate the possibility of more than one disease association in a region. They can also help understand the differences in the results across association studies. We acknowledge the exploratory nature of this study and the difficulty of deciphering *HLA*-disease association due to the complex structure of the region. At this point, we have probably identified markers and not the true risk alleles. Not only is it critical to understand the nature of the association(s) of PD with *HLA*, it is equally important to understand the interrelationship between PD and other *HLA*-associated disorders, particularly multiple sclerosis which like PD is a neurodegenerative disorder [14,20,21,22]. Solving this evolving story will take open collaboration to amass large datasets with genotype, sequence, expression and epigenetic data.

Materials and Methods

Human Subjects

This study was approved by the Institutional Review Boards of the New York State Department of Health, Emory University and Atlanta VA Medical Center, VA Puget Sound Health Care System and University of Washington, and Albany Medical College. All participants gave consent. The majority was signed written consent. A subset of participants who preferred to remain anonymous gave verbal consent. Both written and verbal consent procedures were approved by the IRBs and documented in each participant's data file. A GWAS conducted with 2000 cases and 1986 controls from the NGRC, which had previously identified *HLA* as a PD-associated gene [1] was examined as the primary dataset here. Persons with PD had been diagnosed using standard criteria [23]. Cases and controls were unrelated, non-Hispanic Caucasians from the United States. (See below for Replication dataset.)

Genotyping

DNA was obtained from whole blood for NGRC subjects, and from blood, cell lines, or whole genome amplified DNA for the replication dataset. NGRC individuals were genotyped using the Illumina HumanOmni1-Quad_v1-0_B array, with a call rate of 99.92% and 99.99% reproducibility. Details of the GWAS genotyping and statistical quality control (QC) have been previously published [1]. 9,232 SNPs were genotyped in the *HLA* region in NGRC (from 29–33 Mb on chromosome 6; genome build 36).

Quality Control for NGRC GWAS

NGRC genome-wide genotypes had previously been filtered based on standard QC criteria [1]. Two principal components (PC1 and PC2) were found to associate significantly with PD, representing Jewish/non-Jewish ancestry and European countries of ancestral origin. Sex also significantly associated with case/control status because PD is more prevalent in men than in women, and the population used in this study reflects that disparity. Furthermore, while controls in NGRC were intentionally selected to be older than patients' age at onset, age effects are not completely corrected by the study design. Thus, all association tests with NGRC data included age at enrollment, sex, PC1, and PC2 as covariates.

Step-wise conditional Analysis

The step-wise conditioning followed a previously described and commonly used protocol [7], with the modification that here we used logistic regression instead of chi-square and adjusted for covariates (sex, age, PC1, PC2). We used 107 *HLA* SNPs that reached $P < 0.001$ in GWAS. For round 1 of the conditional analysis, all 107 SNPs were ranked by P value. The SNP with lowest P value was marked as SNP1. The remaining 106 SNPs were tested for association with PD risk conditioned on SNP1. The SNP with the lowest P value was marked SNP2, and analysis was repeated now conditioning on SNP1 and SNP2. The process was repeated until no remaining SNPs had $P < 0.05$ for association with PD. We chose to use $P < 0.05$, rather than a more stringent threshold, because the study was exploratory and would be followed with replication.

Linkage Disequilibrium

We used Haploview [24] to construct haplotypes, visualize haploblocks, and estimate LD and r^2 . Since not all of the SNPs

Table 4. Replication of conditional analysis.

SNP	BP	Minor/Major Allele*						Pooled NGRC & Replication 2843 cases, 2842 controls													
		Replication 843 cases, 856 controls			Replication 2843 cases, 2842 controls			Unconditioned			Conditioned on SNP1			Conditioned on SNP4							
		MAF Case	MAF Control	HWE P	OR	P	Unconditioned	OR	P	MAF Case	MAF Control	HWE P	OR	P	Unconditioned	OR	P	Conditioned on SNP1	OR	P	Conditioned on SNP4
SNP1	rs3129882	32517508	G/A	0.45	0.41	0.69	1.20	0.006	-	-	1.15	0.04	0.46	0.40	0.98	1.28	3×10^{-10}	-	-	1.23	6×10^{-7}
SNP3	rs2844505	31547042	G/A	0.26	0.24	0.18	1.04	0.32	-	-	-	-	-	-	-	-	-	-	-	-	
SNP4	rs9268515	32487273	C/G	0.20	0.23	0.29	1.25	0.004	1.19	0.04	-	-	0.17	0.21	0.96	1.28	2×10^{-6}	1.18	3×10^{-3}	-	-

*Odds ratios (OR) were calculated for the risk allele (shown in bold). In the replication dataset, SNP1 and SNP3 were genotyped, SNP2 was not genotyped and did not impute, SNP4 was imputed (info score = 0.95). We used imputed genotype probabilities in the R software, adjusting for age and sex for replication, and adjusting for age, sex and study for pooled data. Replication P values are one sided [28] given the directionality of the hypothesis; P values for pooled NGRC and replication are two sided. doi:10.1371/journal.pone.0027109.t004

Table 5. Haplotype association test of SNP1 (rs3129882) and SNP4 (rs9268515) with PD in NGRC and Replication.

Haplotype SNP1 (rs3129882) SNP4 (rs9268515)	NGRC	Replication			Pooled NGRC + Replication								
		Freq in cases N = 4000	Freq in controls N = 3972	OR (95%CI)	P	Freq in cases N = 1686	Freq in controls N = 1712	OR (95%CI)	P	Freq in cases N = 5686	Freq in controls N = 5684	OR (95% CI)	P
AC	0.14	0.18	reference	reference	0.17	0.22	reference	reference	0.19	reference	reference	reference	reference
GC	0.02	0.02	1.59 (1.01–2.51)	0.05	0.02	0.02	1.71 (0.85–3.44)	0.07	0.02	0.02	1.63 (1.11–2.39)	0.01	0.01
AG	0.40	0.43	1.23 (1.07–1.42)	3×10^{-3}	0.38	0.38	1.28 (1.04–1.57)	0.01	0.41	0.41	1.25 (1.11–1.40)	2×10^{-4}	2×10^{-4}
GG	0.44	0.38	1.51 (1.32–1.72)	10^{-9}	0.43	0.39	1.41 (1.16–1.71)	2×10^{-4}	0.38	0.44	1.48 (1.33–1.65)	2×10^{-12}	2×10^{-12}

OR and P values were calculated for each haplotype relative to the low-risk haplotype (AC) for association with PD. N = number of chromosomes. The bolded alleles are associated with higher risk. Freq = haplotype frequency. P values are two sided for NGRC, one-sided for replication. doi:10.1371/journal.pone.0027109.t005

Table 6. Additive effects of SNP1 and SNP4 genotypes.*

SNP1 genotype & SNP4 genotype	N case	N control	OR (95% CI)	P
AA & C ₋ [‡]	362	535	Ref	ref
AA & GG	439	469	1.37 (1.24–1.51)	0.001
AG & C ₋	436	461	1.42 (1.28–1.56)	4 × 10 ⁻⁴
AG & GG	959	865	1.61 (1.48–1.76)	2 × 10 ⁻⁸
GG & C ₋	47	38	1.78 (1.41–2.26)	0.01
GG & GG	531	414	1.94 (1.76–2.14)	2 × 10 ⁻¹¹

*NGRC and replication are pooled.

[‡]The C allele of SNP4 is rare; therefore we combined CC and CG genotypes and denoted as C₋.

N = number of individuals with each genotype combination. Risk alleles are shown in bold. Tests were adjusted for age and sex.

doi:10.1371/journal.pone.0027109.t006

reported by other studies were genotyped or imputed in NGRC, LD was calculated using genotypes from the 1000 Genomes Project [25]. SNP genotypes were downloaded from the 1000 Genomes Project website (www.1000genomes.org; CEU low coverage, July 2010 release).

Haplotype Analysis

We used HAPSTAT-3.0 [26] to construct the haplotypes, estimate haplotype frequencies for the top associating SNPs and test haplotype association with PD, while adjusting for age, sex, PC1 and PC2. Only haplotypes with frequencies of ≥ 0.01 in cases and controls combined were included.

Replication

The dataset for replication was downloaded from dbGaP with IRB approval and Data Use Certification from the National Institute of Neurological Disorders and Stroke (NINDS) (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap>) accession number phs000126.v1.p1). The replication dataset was a single GWAS with 843 persons with PD collected by the PROGENI and GenePD studies and 856 neuro-normal controls from the NINDS Repository [12]. Cases were familial PD (one individual per family). Diagnosis was made using standard criteria. Cases and controls were Caucasian and unrelated. SNP1 (rs3129882) and SNP3 (rs2844505) were genotyped. We imputed SNP4 (rs9268515) with information score = 0.95. SNP2 was not genotyped and did not impute. IMPUTE v2 [27] was used with HapMap 3 and 1000 Genomes Project genotypes as reference data. Genotype probabilities (dose 2-0) were analyzed in R software <http://www.r-project.org/>. Haplotype analyses were performed as described above. P values were one-sided for replication given the directional hypotheses [28]. The pooled analyses of NGRC and replication had two-sided P values and were adjusted for study as well as age and sex.

Interaction

The full model for testing all pair-wise interactions among the four SNPs in NGRC was [SNP1+SNP2+SNP3+SNP4+SNP1*SNP2+SNP1*SNP3+SNP1*SNP4+SNP2*SNP3+SNP2*SNP4+SNP3*SNP4+covariates] vs. [SNP1+SNP2+SNP3+SNP4+covariates]. We also tested

each pair of SNPs one at a time, while keeping all other SNPs in the model as covariates. For example, for interaction between SNP1 and SNP2 the model was [SNP1*SNP2+SNP1+SNP2+SNP3+SNP4+covariates]. All analyses included the following covariates: sex, age, PC1 and PC2. In the pooled dataset (NGRC plus replication), we tested for interaction between SNP1 and SNP4 only as [SNP1*SNP4+SNP1+SNP4+covariates]; here the covariates were age and sex. Power calculation for interaction was performed using Quanto v1.2.4 [29] with 2843 cases and 2842 controls and a two-sided $\alpha = 0.05$ to estimate power of our study to detect the observed interaction OR, and to estimate the minimum interaction OR detectable with 80% or 90% power.

Data access

NGRC data are available at www.ncbi.nlm.nih.gov/gap study accession number phs000196.v2.p1. The replication dataset was obtained from the NINDS Database found at (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap>) through dbGaP accession number phs000126.v1.p1.

Supporting Information

Table S1 Step-wise conditional analysis including 107 HLA SNPs that achieved $P < 10^{-3}$ in GWAS.
(PDF)

Acknowledgments

We thank the persons with Parkinson's disease and volunteers who participate in research. We acknowledge NINDS and thank Drs. Tatiana Foroud and Richard Myers for making their data publicly available. We thank Dr. Taye Hamza for the initial GWAS analyses [1] on which this study was founded.

Author Contributions

Conceived and designed the experiments: HP EH-B GT. Performed the experiments: EH-B. Analyzed the data: EH-B. Contributed reagents/materials/analysis tools: SF CZ HP. Wrote the paper: HP EH-B. Helped refine the final manuscript: GT SF CZ.

References

- Hamza TH, Zabetian CP, Tenesa A, Laederach A, Montimurro J, et al. (2010) Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat Genet* 42: 781–785.
- McGeer PL, McGeer EG (2008) Glial reactions in Parkinson's disease. *Mov Disord* 23: 474–483.
- Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. *Nat Genet* 39: 1217–1224.
- Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464: 773–777.

5. Simon-Sanchez J, van Hilten JJ, van de Warrenburg B, Post B, Berendse HW, et al. (2011) Genome-wide association study confirms extant PD risk loci among the Dutch. *Eur J Hum Genet* 19: 655–661.
6. Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin UM, et al. (2011) Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* 377: 641–649.
7. Payami H, Joe S, Farid NR, Stenski V, Chan SH, et al. (1989) Relative predispositional effects (RPE's) of marker alleles with disease: HLA-DR and autoimmune thyroid disease. *Am J Hum Genet* 45: 541–546.
8. Spencer CC, Plagnol V, Strange A, Gardner M, Paisan-Ruiz C, et al. (2011) Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. *Hum Mol Genet* 20: 345–353.
9. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, et al. (2010) Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 42: 436–440.
10. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559–575.
11. Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9: 477–485.
12. Pankratz N, Wilk JB, Latourelle JC, DeStefano AL, Halter C, et al. (2009) Genomewide association study for susceptibility genes contributing to familial Parkinson disease. *Hum Genet* 124: 593–605.
13. Thomson G, Valdes AM, Noble JA, Kockum I, Grote MN, et al. (2007) Relative predispositional effects of HLA class II DRB1-DQB1 haplotypes and genotypes on type 1 diabetes: a meta-analysis. *Tissue Antigens* 70: 110–127.
14. Barcellos LF, Sawcer S, Ramsay PP, Baranzini SE, Thomson G, et al. (2006) Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. *Hum Mol Genet* 15: 2813–2824.
15. Mata IF, Shi M, Agarwal P, Chung KA, Edwards KL, et al. (2010) SNCA variant associated with Parkinson disease and plasma alpha-synuclein level. *Arch Neurol* 67: 1350–1356.
16. Puschmann A, Verbeek C, Heckman MG, Soto-Ortolaza AI, Lynch T, et al. (2011) Human leukocyte antigen variation and Parkinson's disease. *Parkinsonism Relat Disord* 17: 376–378.
17. Rohn TT, Catlin LW (2010) Immunolocalization of influenza A virus and markers of inflammation in the human Parkinson's disease brain. *PLoS One* 6: e20495.
18. Benkler M, Agmon-Levin N, Shoenfeld Y (2009) Parkinson's disease, autoimmunity, and olfaction. *Int J Neurosci* 119: 2133–2143.
19. Hirsch EC, Hunot S (2009) Neuroinflammation in Parkinson's disease: a target for neuroprotection? *Lancet Neurol* 8: 382–397.
20. Payami H, Khan MA, Grennan DM, Sanders PA, Dyer PA, et al. (1987) Analysis of genetic interrelationship among HLA-associated diseases. *Am J Hum Genet* 41: 331–349.
21. Menon R, Farina C (2011) Shared molecular and functional frameworks among five complex human disorders: a comparative study on interactomes linked to susceptibility genes. *PLoS One* 6: e18660.
22. Sirota M, Schaub MA, Batzoglou S, Robinson WH, Butte AJ (2009) Autoimmune disease classification by inverse association with SNP alleles. *PLoS Genet* 5: e1000792.
23. Hughes AJ (1997) Clinicopathological aspects of Parkinson's disease. *Eur Neurol* 38: 13–20.
24. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263–265.
25. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, et al. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.
26. Lin DY, Huang BE (2007) The use of inferred haplotypes in downstream analyses. *Am J Hum Genet* 80: 577–579.
27. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5: e1000529.
28. Neter J, Kunter M, Nachtsheim C, Wasserman W (1996) *Applied Linear Statistical Models*. pp 52–53.
29. Gauderman W, Morrison J (2006) QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies, <http://hydra.usc.edu/gxe>.