SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
NEUROGENETICS
GENETIC ASSOCIATION STUDY
SYSTEMS BIOLOGY
NEURODEGENERATION

Received
13 September 2011

Accepted
18 January 2012

Published
10 February 2012

Correspondence and requests for materials should be addressed to A.V. (andre.valente@biocant.pt) or Y.G. (garygao@gmail.com)

Rare coding SNP in DZIP1 gene associated with late-onset sporadic Parkinson's disease

André X. C. N. Valente[1,2,3], Joo H. Shin[4], Abhijit Sarkar[5] & Yuan Gao[4]

[1]Systems Biology Group, Biocant - Biotechnology Innovation Center, Cantanhede, Portugal, [2]CNC - Center for Neuroscience and Cell Biology, University of Coimbra, Coimbra, Portugal, [3]Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia, USA, [4]Lieber Institute for Brain Development, Johns Hopkins Medical Campus, 855 N. Wolfe Street, Suite 300, Baltimore, Maryland 21205, [5]Department of Physics and Vitreous State Laboratory, Catholic University of America, Washington, DC, USA.

An association between a rare, coding, non-synonymous SNP variant in the gene DZIP1 and Parkinson's disease was found, based on an analysis of the existing NGRC genome-wide association study dataset. The statistical analysis utilized the hypothesis-rich, targeted search unbiased assessment approach, rather than the hypothesis-free, genome-wide agnostic search paradigm. The association of DZIP1 with Parkinson's disease is discussed in the context of a Parkinson's disease stem-cell ageing theory.

Familial genetic linkage studies have associated six genes with Mendelian inheritable forms of Parkinson's disease (PD)[1]. However, these monogenic forms account for fewer than 10% of PD cases. Further, they lead mostly to juvenile or early onset forms of PD (before age 50). Given that no decisive environmental causative factors have been found either, the etiology of late-onset PD (comprising over 90% of all PD cases) remains essentially undetermined. A range of hypotheses are being explored[2–6]. We have proposed the theory that i) sporadic PD is best defined as a characteristic deviation from normality in the expression program of a cell (the PD-state) and ii) this PD-state can originate as a case of hematopoietic stem-cell program defect[7].

At present, considerable efforts are focused on finding differential genetic susceptibility to late-onset PD via the genome-wide association study (GWAS)[8]. In a GWAS, a set of patients and controls is genotyped at known SNP sites in the human genome. Mathematically, this assigns individuals to locations within a high-dimensional SNP space (Figure 1). Genetic susceptibilities are inferred from statistically significant differences in the placement of patients and controls in this SNP space. Large enough differential disease risks constitute practical predictive genetic markers. So far though, susceptibilities found have been typically weak (some 85% of trait associated SNPs reported have an odds ratio in the 0.5–2 range)[9]. Nonetheless, such findings can still be invaluable as indicators of the involvement of particular genes or biological processes in the disease mechanics. As of today, GWASs have reported about a dozen, modest effect (odds ratio in the 0.5–2 range), susceptibility loci for PD[10–17].

The hypothesis-free paradigm currently dominates GWAS statistical data analysis[8]. It has been previously described why this is a poor choice[18–21]. Biological knowledge and insightful hypotheses are as crucial in the analysis of a GWAS as they are in the analysis of any classical biological experiment[22–25]. The alternative hypothesis-rich mathematical theory recognizes this fact and allows biological thought to maximize statistical power[21,26]. Key in the approach is the concept of Rational Class (RC), a set of candidate laws (markers in the GWAS context) that share an underlying common rationale.

In this article, we analyze the late-onset sporadic PD GWAS NGRC dataset of Hamza et al.[10], under the hypothesis-rich framework (the late-onset, sporadic qualifier will be henceforth subsumed)[26,27]. In the *Methods* section, the focus is on describing the RCs constructed specifically for this PD GWAS analysis. Findings are summarized in the *Results* section. Finally, in the *Discussion* section, we review relevant biological information to contextualize our findings.

Results

The significant findings from the hypothesis-rich analysis of the Hamza et al. dataset are presented in Table 1. For these SNPs, the null hypothesis was that the two regions defined by the *split mode* (see Figure 3: 1-dimensional split modes) present no differential susceptibility to PD. Now, pure chance in the finite sampling of individuals

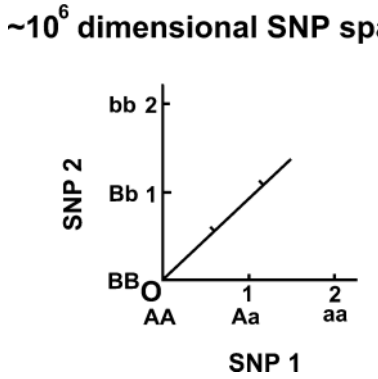

Figure 1 | In a genome-wide association study (GWAS), subjects are vectors in SNP space. Depicted is one sensible coordinate system for SNP space. Capital letters represent the major allele, lower case letters the minor allele. To each SNP therefore corresponds an axis with 3 admissible values (0, 1 and 2). At present, typical cohort sizes are in the range of 10^3 to 10^4 subjects, while the number of SNPs genotyped is on the order of 10^6.

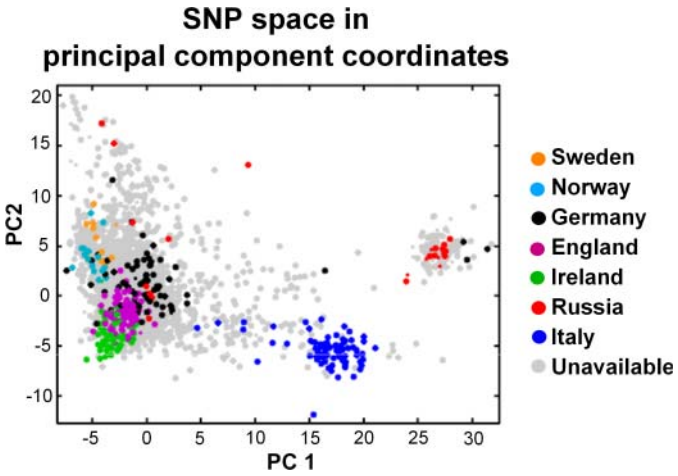

Figure 2 | The Hamza et al.[10] cohort data in SNP space, after a change from the Figure 1 coordinate system to principal component coordinates (first two principal components shown). Color indicates the country of parental origin for subjects that reported such information and for whom both parents had a common origin. The plot replicates a similar figure in Hamza et al.. Smaller circles denote individuals with a lower statistical weight, due to the process of population homogenization across SNP space regarding the PD to control subject ratio[56].

from the population could create the false impression of differential susceptibility. Assuming the null hypothesis, the *reference probability* indicates the ease of such stochasticity producing an unwarranted call (as per the hypothesis-rich framework) of differential susceptibility[27]. We emphasize that the quoted reference probabilities already take into consideration the presence of multiple-hypotheses testing. For easiness of comparison, the arbitrariness in defining *odds ratio* (given the validity of the inverse of any choice) was settled by making every odds ratio larger than unity. The *minor allele effect* entry then indicates which region carries the greater risk of PD.

The reported SNPs in the SNCA region and in the HLA-DRA region had all been noted as significant in previous GWASs[10,28,29]. The SNPs reported in the chromosome 17 q21.31 region (usually categorized as the MAPT region) validate the previous GWAS based association of this region with PD (most of these MAPT region SNPs have been specifically previously reported, though we could not confirm all)[10–17]. The novel finding is the increased susceptibility to PD conferred by a rare, coding, non-synonymous SNP variant in the DZIP1 gene (Figure 4).

Discussion

The PD working theory we put forward in previous work[7] provides a possible context for the connection of DZIP1 with PD found in this analysis. Therefore, we start by reviewing it. Firstly, PD would be defined in terms of the PD-state, a characteristic deviation from the normal expression program of a healthy cell. Singular cellular manifestations of PD would therefore be de-emphasized in favor of this systems-level definition. Crucially, the PD-state would be a generic cell state, not restricted to neurons. Secondly, the PD-state would

Table 1 | Summary of the findings from hypothesis-rich analysis of the Hamza et al. GWAS PD dataset. See the *Results* main text section for meaning of the entries

Rational Class	RC #2 coding region minor allele freq: 10%–30%	RC #7 hematopoietic coding region minor allele freq.< 10%	RC #15 generic (non-coding/non-UTR) 30% < minor allele freq.	
Gene	**IMP5, MAPT, CRHR1, KIAA1267, C17orf69, NSF**	**DZIP1**	**HLA-DRA**	**SNCA-GPRIN3, SNCA**
SNP	rs12373123, rs12185235, 17651549, rs16940665, 12185268, kgp6408681, rs1052551, rs16940674, rs36076725, rs17652121, kgp3974170, rs17574604, kgp3365508, rs10445337, rs1881193, rs3583914, rs1052553, kgp4886152, rs199533	kgp1112497	rs3129822	rs356220, rs2736990, rs356168
Location	chr. 17, q21.31 in coding regions synonymous & non-synonymous substitutions	chr. 13, q32.1 coding region non-synonymous substitution	chr. 6, p21.3 intronic	chr. 4, q22.1 intergenic (SNCA-GPRIN3) intronic (SNCA)
Minor allele frequency	19% thru 21%	0.7%	44%	40% thru 49%
Minor allele effect	protective	harmful	harmful	harmful
Split mode	minor allele dominant	minor allele dominant	extreme	minor allele dominant
Odds ratio	1.2 thru 1.3	4.4	1.5	1.3 thru 1.4
Reference probability	0.06 thru 0.09	0.03	0.04	0.04 thru 0.07

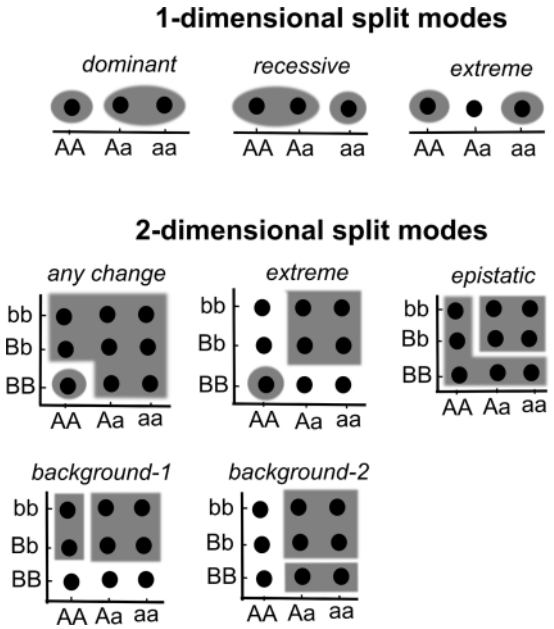

Figure 3 | Each graph shows a manner of splitting SNP space into two shaded regions. Differential risk of PD between the shaded regions is then ascertained (non-shaded regions are ignored). **1-dimensional split modes:** Utilized in RCs containing single SNPs (RCs 1 thru 15 and RC 23). **2-dimensional split modes:** Utilized in RCs containing pairs of SNPs (RCs 16 thru 19).

originate in a stem-cell program defect, associated with the ageing of stem-cells. We proposed the hematopoietic stem-cell niche as a place of origin for the PD-state, although other stem cell niches should not be ruled out from playing a part. Thirdly, the subsequent PD-state propagation to other cells would not occur evenly. Propagation would be faster to cells more amenable to reprogramming (such as other stem cells or their not yet fully differentiated progeny). Thus tissues under active regeneration would be the first to be affected. Beyond PD biology, note the validity of this theory would signal an effective degree of communication between different stem-cell niches greater then what is currently accepted.

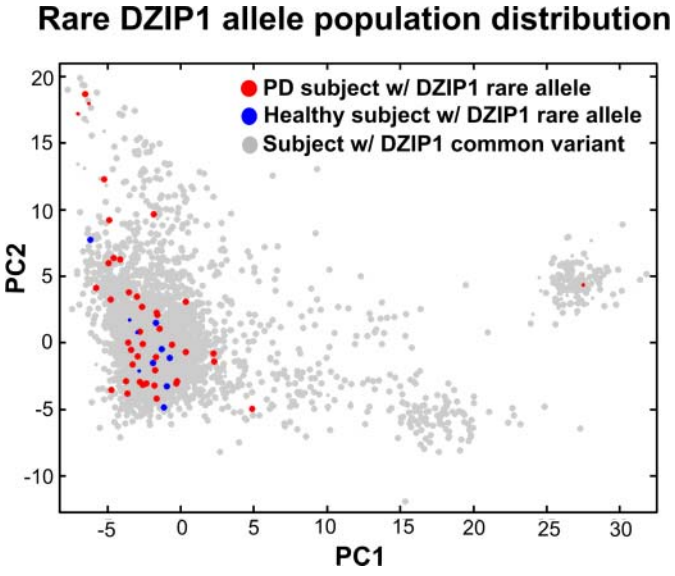

Figure 4 | Individuals in the Hamza et al. cohort carrying a copy of the rare DZIP1 allele are highlighted in SNP space, under principal component coordinates (first two principal components shown). No homozygous rare allele individuals were present in the dataset.

We now describe what is known at present about the biological role of DZIP1. The gene DZIP1 encodes a C2H2-type zinc finger protein[30]. Its acronym stands for DAZ-interacting protein 1, as DZIP1 was originally identified in a screen for protein interaction partners of the DAZ (deleted in azoospermia) protein 30. Its expression in human embryonic, stem, fetal and adult germ cells was thus well noted[30]. Zebrafish mutants in *iguana* (the DZIP1 ortholog in Zebrafish) have been invaluable in characterizing the gene. A *iguana* mutant (*fo10a*) displayed ultrastructural defects in perivascular mural cell recruitment and subsequent hemorrage, thus linking vascular stability and DZIP1[31]. Work with Zebrafish *iguana* mutants also revealed DZIP1 to be a component of the Hedgehog (Hh) signaling pathway[32,33]. Within the Hh pathway, DZIP1 acts downstream of Smoothened, modulating the activity of the Gli family of transcription factors[32,33]. DZIP1 has further been implicated in primary ciliogenesis and its role in Hh signaling may occur in this context[34–36]. Hh plays a vital part in directing embryonic pattern formation[37]. However, it continues regulating adult stem cells beyond embryogenesis[38,39]. Studies have specifically implicated Hh in the adult maintenance of hematopoietic stem cells[40], epithelial stem cells in the gastrointestinal tract[41], neuronal stem cells in the subventricular zone (SVZ) and the hippocampal dentate gyrus[42,43], hair follicle stem cells[44], mammary stem cells[45] and mesenchymal stem cells[46]. Besides its role in neurogenesis, Hh has also shown neurotrophic properties, in particular regarding dopaminergic neuron survival[47–49]. Administration of Sonic Hedgehog reduced behavioral deficits in animal models of PD[50,51]. Nonetheless, an earlier targeted genetic analysis of Sonic Hedgehog in Parkinson's patients, did not find any significant mutations in this gene[52].

Genetic mutations affecting the Hh pathway have been associated with an increased incidence of a diversity of cancers (see Merchant et al.[53] or Beachy et al.[39] for comprehensive listings). Under a cancer stem-cell hypothesis[54] interpretation, this is consistent with the role of Hh in adult stem cell homeostasis. The aberrant Hh signaling would contribute the conversion of adult stem cells (or perhaps their early progeny) into cancer stem cells, cells endowed with stem-cell properties and trapped in a pathological state of constant renewal[39,54]. Now, under our PD hypothesis, PD also originates in a stem-cell program defect. However, while in the cancer stem cell hypothesis the pathology progresses via physical replication of the cancer stem cells themselves, in PD we are proposing propagation solely of the PD characteristic expression state (the PD-state)[7]. The PD-state of a cell could possibly be physically locked in by epigenetic DNA modifications[7].

We have reported a non-synonymous SNP in the DZIP1 gene that confers increased susceptibility to PD. We emphasize that this result is based on a single population cohort of mixed European ancestry, the Hamza et al. dataset[10]. Importantly, confirmation by future cohort analyses remains to be determined. The result raises the possibility of a connection between adult stem-cell regulation and Parkinson's disease, which we explored. Again, it remains to be seen whether this PD stem-cell biology association idea will be supported or infirmed by PD research work in the next few years.

Methods

We analyzed the NGRC GWAS dataset of Hamza et al.[10], consisting of 1986 control subjects and 2000 sporadic late-onset PD patients. All individuals were Americans of self-reported European ancestry. As in any GWAS, a concern is the presence of population structure in the cohort data[55]. Likely the European population, due to historical and geographical factors, does not constitute, mating-wise, a single uniformly mixed population. Now, suppose the existing subpopulations have distinct susceptibilities to PD. This could be due to differences in genetic background, culture (e.g., diet), or physical environment. Regardless, a genetic marker of a subpopulation (e.g., a SNP variant typical of a subpopulation) would then effectively mark a distinct susceptibility to PD. This poses a problem, in that we would like to interpret markers as having a causative effect on PD susceptibility, which clearly would not be the case here. The issue may also arise merely by study recruitment centers in areas with distinct subpopulations not enrolling identical ratios of PD to control subjects. Note

that although for simplicity we allude above to discrete subpopulations, generically the mixing makeup will have a continuous character.

To analyze the dataset of Hamza et al., we used the SNP space coordinate system shown in Figure 1. The relative overall location of individuals in SNP space (Euclidean distance wise) reflects the cohort population structure. Namely, relative locations are consistent with parental country of origin for those subjects that reported such information and for whom both parents had a common origin. This is visually clear upon a change of coordinates to principal component coordinates (Figure 2).

A variety of methods exist for mitigating the population structure problem[55]. We chose to homogenize the population regarding the PD to control subject ratio, via individual weight knock-down. Briefly, this involves reducing the statistical weight of selected individuals to locally level the ratio of PD to control subjects throughout SNP space. A separate article describes in detail both the method and its application to this particular dataset[56]. The homogenization procedure reduced the dataset to a net weight of 1904 PD patients and 1802 controls (a 7% size reduction). This will be the reference dataset henceforth.

We utilized the hypothesis-rich framework to investigate the dataset[26,27]. The hypothesis-rich framework provides a *targeted search, unbiased assessment* approach to the analysis of GWAS data. The *targeted search* assertion follows from biological considerations guiding the statistical search for genetic susceptibility factors. Specifically, biological information enters the mathematical analysis via the concept of Rational Class (RC), a set of candidate genetic markers that share a common rationale. Yet, in spite of the biased search, an *unbiased assessment* is obtained from a proper mathematical treatment of multiple hypotheses testing[26,27].

We now describe the RCs constructed for the PD GWAS problem. Throughout, recall that separating markers into distinct RCs can be statistically advantageous if the resulting RCs have different True Quality Distribution and Correlation Structures (shorthand, TQDs)[26]. This can be the case whenever a biological rationale underlies the marker separation. On the other hand, RCs must be rank ordered and statistical resolution decreases with increasing rank, thus overly liberal RC creation is pointless[26,27].

A total of 23 RCs were constructed (Table 2). The first 15 RCs, containing individual SNPs, were based on the following factors:

Genomic region. we grouped SNPs by whether they fell in a coding region, in the UTR or in the remainder of the genome. Confirming the distinct biological roles of these regions, past GWASs show the incidence of trait associated SNPs in them is not uniform[8].

SNP allele frequencies. These frequencies are affected by the degree of selective pressure on the associated haplotypes. Thus, on average, the character of SNPs with different allele frequency ratios may be distinct. We divided SNPs into three broad groups, based on their minor variant frequency: <10%, 10–30% and >30%. Also, note that we are comparatively more interested in larger odds ratio markers. Given two SNP markers showing the same statistical significance (ordinarily, same p-value), the one with the lower minor variant frequency necessarily shows a larger differential trait susceptibility (larger odds ratio). Thus, as an additional benefit, the above frequency breakdown effectively protects the search for rare variant, high odds ratio markers.

Hematopoietic fingerprints. Given our PD working theory, the set of SNPs occurring in genes with a function in the hematopoietic system acquires particular relevance. We recorded 2253 SNPs spread across 662 so called hematopoietic fingerprint genes[57]. The genes were identified by Chambers et al. via global gene expression profiling of murine hematopoietic stem cells and their major differentiated lineages (NK-cells, T-cells, B-cells, monocytes, neutrophils and nucleated erythrocytes)[57].

Combination of the above factors yielded RCs 1 thru 15 (Table 2). In these RCs, SNPs were tested for association with differential PD risk three separate times, each time based on a different mode of splitting the SNP space (Figure 3, 1-dimensional split modes). The dominant and recessive modes were motivated by their well known biological counterparts. However, a situation where phenotype is significantly more assured only under homogeneous alleles is also biologically plausible. The extreme mode accommodates these cases by excluding individuals with heterogeneous alleles from the statistical comparison. In every case, the null hypothesis was that the two regions defined by the split mode present no different susceptibility to PD. Statistical comparison between the two chosen regions was done via the Fisher exact test.

RCs 16 thru 19 were based on SNP pairs. Given there are on the order of 10^6 SNPs, potential SNP pairs are on the order of 10^{12}. A RC containing such a large number of entries is unlikely to have a favorable TQD[26]. It is therefore fundamental to prioritize SNP pairs. We generated one list of SNP pairs based on protein-protein physical interactions. For every two interacting proteins on different chromosomes, all SNP pairs with one SNP in each of the interacting proteins respective coding gene region were added to the list. The exclusion of protein pairs on the same chromosome excludes pairs of SNPs potentially in linkage disequilibrium. Protein-protein interactions were obtained from HPRD (~39000 interactions)[58]. The SNP pairs were tested for association with differential PD risk five times, each time based on a different mode of splitting the SNP space (Figure 3, 2-dimensional split modes). In every case, the null hypothesis was that the two regions defined by the split mode present no different susceptibility to PD. Statistical comparison between the two chosen regions was done via the Fisher exact test. The results of the tests were assigned to RC 16 or to RC 17 depending on whether the associated odds ratio was smaller or larger than 3. Once more, this has the benefit of safeguarding the search for high odds ratio markers.

A second list of SNP pairs was constructed based on the hematopoietic fingerprint genes. Based on the expression profiling, Chambers et al. had further divided the hematopoietic fingerprint genes into the following subclasses: hematopoietic stem cells, B-cells, naive T-cells, NK-cells, monocytes, granulocytes, nucleated erythrocytes, differentiated shared fingerprint, lymphoid shared fingerprint and myeloid shared fingerprint 57. We generated hematopoietic gene pairs by considering every possible pairing of genes within the same hematopoietic subclass, exclusive of gene pairs in the same chromosome. The procedure described above for protein pairs was then applied to the hematopoietic gene pairs, thus generating RCs 18 and 19.

RCs 20 thru 22 contained SNP tuplets generated from protein complexes. Human protein complexes were obtained from the CORUM database (~1300 complexes)[59]. Consider first RC 20, containing 2-tuplets generated from complexes of up to 4 proteins. The 2-tuplets for RC 20 were generated as follows:

Table 2 | The Rational Classes (RCs) constructed to analyze the PD GWAS data

Rational Class Rank	Rational Class Description
1	**coding** | allele freq.: < 10% | dominant, recessive and extreme
2	**coding** | allele freq.: 10% to 20% | dom., rec. extr.
3	**coding** | allele freq.: > 30% | dom., rec. extr.
4	**UTR** | allele freq.: < 10% | dom., rec. extr.
5	**UTR** | allele freq.: 10% to 20% | dom., rec. extr.
6	**UTR** | allele freq.: > 30% | dom., rec. extr.
7	**hematopoietic** | coding | allele freq.: < 10% | dom., rec. extr.
8	**hematopoietic** | coding | allele freq.: 10% to 20% | dom., rec. extr.
9	**hematopoietic** | coding | allele freq.: > 30% | dom., rec. extr.
10	**hematopoietic** | UTR | allele freq.: < 10% | dom., rec. extr.
11	**hematopoietic** | UTR | allele freq.: 10% to 20% | dom., rec. extr.
12	**hematopoietic** | UTR | allele freq.: > 30% | dom., rec. extr.
13	**non-coding/non-UTR** | allele freq.: > 30% | dom., rec. extr.
14	**non-coding/non-UTR** | allele freq.: < 10% | dom., rec. extr.
15	**non-coding/non-UTR** | allele freq.: 10% to 20% | dom., rec. extr.
16	**protein pairwise interactions** | coding | OR > 3 | 5 split modes
17	**protein pairwise interactions** | coding | OR < 3 | 5 split modes
18	**hematopoietic pairwise interactions** | coding | OR > 3 | 5 split modes
19	**hematopoietic pairwise interactions** | coding | OR < 3 | 5 split modes
20	**protein complexes** | sizes 2 thru 4 | top 2 proteins
21	**protein complexes** | sizes 3 thru 9 | top 3 proteins
22	**protein complexes** | sizes 4 thru 16 | top 4 proteins
23	**gene expression sig.** | coding and UTR | all freqs. | dom., rec., extr.

1. Given a complex, consider the SNPs that fall in the coding region of the protein members of the complex. Denote them as CSNPs. Add every possible (CSNP A, CSNP B) 2-tuplet to RC 20, provided CSNP A and CSNP B are located in different chromosomes.
2. Repeat for every complex of up to 4 proteins.

Each 2-tuplet was tested for association with differential PD risk 3 separate times, as follows:

1. Under the dominant 1-dimensional split mode, assign a Fisher exact test based p-value to each CSNP in the tuplet in the standard fashion (i.e., considering the CSNP as an individual SNP, as in the RCs 1 thru 15). We formalize it by writing p-value = p(CSNP; dominant mode).
2. The p-value associated with the 2-tuplet is (max(p(CSNP A; dominant mode), p(SNP B; dominant mode)))^2 (i.e., squared).
3. Assign two more p-values to the tuplet, as above, but now utilizing the recessive and extreme 1-dimensional split modes.

RC 21 was similar to RC 20, except that:

1. It was based on complexes of sizes 3 thru 9.
2. It contained 3-tuplets (CSNP A, CSNP B, CSNP C).
3. The p-value associated with a 3-tuplet is (max(p(CSNP A; split mode), p(SNP B; split mode), p(SNP C; split mode)))^3 (i.e., cubed).

RC 22 was similar to RC 20, except that:

1. It was based on complexes of sizes 4 thru 16.
2. It contained 4-tuplets (CSNP A, CSNP B, CSNP C, CSNP D).
3. The p-value associated with a 4-tuplet is (max(p(CSNP A; split mode), p(SNP B; split mode), p(SNP C; split mode), p(SNP D; split mode)))^4 (i.e., to the fourth power).

In these complex based RCs, in every case the null hypothesis is that *none* of the SNPs in the tuplet shows differential susceptibility to PD between the two regions defined by the split mode. The anticipation is that a complex mechanistically involved in PD may produce a tuplet (or tuplets) of particular low p-value under the above tuplet p-value definition. RCs 20, 21 and 22 are kept separate to preserve potentially distinct TQDs.

Finally, RC 23 was based on genes in the blood gene expression signature for PD (involving 18 genes) we developed in earlier work 26. RC 23 contained:

1. All SNPs in the coding or UTR regions of the genes present in the expression signature.
2. All SNP pairs, exclusive of pairs in the same chromosome, with one SNP in the coding or UTR regions of one expression signature gene and the other SNP in the coding or UTR region of a second expression signature gene.

The individual SNPs were tested for association with differential risk of PD under the three 1-dimensional split modes via the Fisher exact test, as in RCs 1 thru 15. The SNP pairs were tested for association with differential risk of PD under the five 2-dimensional split modes via the Fisher exact test, as in RCs 16 thru 19. All tests were placed in a single RC, given their low number.

Quality control. At the outset, a quality control procedure was applied to the Hamza et al. dataset that excluded the following SNPs from the entire analysis:

- SNPs with a p-value less than 10^{-5} under the Hardy-Weinberg test.
- SNPs with less than a 99.9% call rate.

The quality control was implemented using the program Plink[60]. A total of 748807 SNPs passed the quality control.

1. Bekris, L. M., Mata, I. F. & Zabetian, C. P. The genetics of Parkinson's disease. *Journal of Geriatric Psychiatry and Neurology* **23 (4),** 228–242 (2010).
2. Hawkes, C. H., Tredici, K. D. & Braak, H. Parkinson's disease: a dual-hit hypothesis. *Neuropathology and Applied Neurobiology* **33 (6),** 599–614 (2007).
3. Lerner, A. & Bagic, A. Olfactory pathogenesis of idiopathic Parkinson disease revisited. *Movement Disorders* **23 (8),** 1076–1084 (2008).
4. Whitton, P. Inflammation as a causative factor in the aetiology of Parkinson's disease. *British Journal of Pharmacology* **150 (8),** 963–976 (2007).
5. Halliwell, B. & Gutteridge, J. M. C. in *Free radicals in biology and medicine,* edited by Halliwell, B. & Gutteridge, J. M. C. (Oxford University Press, New York, 1999), pp. 744–788.
6. Monte, D. A. D., Lavasani, M. & Manning-Bog, A. B. Environmental factors in Parkinson's disease. *Neurotoxicology* **23 (4–5),** 487–502 (2002).
7. Valente, A. X. C. N., Sousa, J. A. B., Outeiro, T. F. & Ferreira, L. in *Science and engineering in high-throughput biology including a theory on Parkinson's disease* (Lulu Books, 2011), pp. 43–73.
8. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *The New England Journal of Medicine* **363 (2),** 166–176 (2010).
9. Hindorff, L. A. *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the USA* **106 (23),** 9362–9367 (2009).
10. Hamza, T. H. *et al*. Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nature Genetics* **42**, 781–785 (2010).
11. Simón-Sánchez, J., Schulte, C., Bras, J. M., Sharma, M. & Gibbs, J. R. Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nature Genetics* **41 (12),** 1308–1312 (2009).
12. Satake, W. *et al*. Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nature Genetics* **12**, 1303–1308 (2010).
13. Saad, M. *et al*. Genome-wide association study confirms BST1 and suggests a locus on 12q24 as the risk loci for Parkinson's disease in the european populaiton. *Human Molecular Genetics* **20 (3),** 615–627 (2011).
14. Do, C. B. *et al*. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genetics* **7 (6),** e1002141 (2011).
15. International Parkinson Disease Genomics Consortium, Inputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet* **377**, 641–649 (2011).
16. International Parkinson's Disease Genomics Consortium (IPDGC), Wellcome Trust Case Control Consortium 2 (WTCCC2), A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS Genetics* **7 (6),** e1002142 (2011).
17. The UK Parkinson's Disease Consortium and The Wellcome Trust Case Control Consortium 2, Dissection of the genetics of Parkinson's disease identifies an additional association 5′ of SNCA and multiple associated haplotypes at 17q21. *Human Molecular Genetics* **20 (2),** 345–353 (2011).
18. Roeder, K., Devlin, B. & Wasserman, L. Improving power in genome-wide association studies: weights tip the scale. *Genetic Epidemiology* **31 (7),** 741–747 (2007).
19. Chasman, D. I. On the utility of gene set methods in genomewide association studies of quantitative traits. *Genetic Epidemiology* **32 (7),** 658–668 (2008).
20. Roeder, K. & Wasserman, L. Genome-wide significance levels and weighted hypothesis testing. *Statistical Science* **24(4),** 398–413 (2009).
21. Valente, A. X. C. N. in *Science and engineering in high-throughput biology including a theory on Parkinson's disease* (Lulu Books, 2011), pp. 9–22.
22. Roeder, K., Bacanu, S.-A., Wasserman, L. & Devlin, B. Using linkage genome scans to improve power of association in genome scans. *The American Journal of Human Genetics* **78 (2),** 243–252 (2006).
23. Eskin, E. Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Research* **18**, 653–660 (2008).
24. Bakir-Gungor, B. & Sezerman, O. U. A new methodology to associate SNPs with human diseases according to their pathway related context. *PLoS ONE* **6 (10),** e26277 (2010).
25. Li, M.-X., Sham, P. C., Cherny, S. S. & Song, Y.-Q. A knowledge-based weighting framework to boost the power of genome-wide association studies. *PloS ONE* **5 (12),** e14480 (2010).
26. Valente, A. X. C. N. in *Science and engineering in high-throughput biology including a theory on Parkinson's disease* (Lulu Books, 2011), pp. 23-38.
27. Valente, A. X. C. N., Sarkar, A. & Gao, Y. Analyzing genome-wide association data through the hypothesis-rich framework. *In preparation* (2011).
28. Fung, H.-C. *et al*. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: fi rst stage analysis and public release of data. *Lancet Neurology* **5**, 911–916 (2006).
29. Edwards, T. L. *et al*. Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. *Annals of Human Genetics* **74 (2),** 97–109 (2010).
30. Moore, F. L., Jaruzelska, J., Dorfman, D. M. & Reijo-Pera, R. A. Identification of a novel gene, DZIP (DAZ-interacting protein) that encodes a protein that interacts with DAZ (deleted in azoospermia) and is expressed in embryonic stem cells and germ cells. *Genomics* **83**, 834–843 (2004).
31. Lamont, R. E. *et al*. Hedgehog signaling via angiopoietin1 is required for developmental vasculas stability. *Mechanisms of Development* **127**, 159–168 (2010).
32. Sekimizu, K. *et al*. The zebrafish iguana locus encodes Dzip1, a novel zinc-finger protein required for proper regulation of Hedgehog signaling. *Development* **131 (11),** 2521–2532 (2004).
33. Wolff, C. *et al*. iguana encodes a novel zinc-finger protein with coiled-coil domains essential for Hedgehog signal transduction in the zebrafish embryo. *Genes and Development* **18**, 1565–1576 (2004).
34. Glazer, A. M. *et al*. The Zn Finger protein Iguana impacts Hedgehog signaling by promoting ciliogenesis. *Developmental Biology* **337**, 148–156 (2010).
35. Kim, H. R., Richardson, J., Eeden, F. v. & Ingham, P. W. Gli2a protein localization reveals a role for Iguana/DZIP1 in primary ciliogenesis and a dependence of Hedgehog signal transduction on primary cilia in the zebrafish. *BMC Biology* **8: 65** (2010).
36. Tay, S. Y. *et al*. The iguana/DZIP1 protein is a novel component of the ciliogenic pathway essential for axonemal biogenesis. *Developmental Dynamics* **239**, 527–534 (2010).
37. Ingham, P. W. & McMahon, A. P. Hedgehog signaling in animal development: paradigms and principles. *Genes and Development* **15**, 3059–3087 (2001).
38. Lum, L. & Beachy, P. A. The Hedgehog response network: sensors, switches, and routers. *Science* **304**, 1755–1759 (2004).

39. Beachy, P. A., Karhadkar, S. S. & Berman, D. M. Tissue repair and stem cell renewal in carcinogenesis. *Nature* **432**, 324–331 (2004).
40. Bhardwaj, G., Murdoch, B., Wu, D., Baker, D. P. & Williams, K. P. Sonic hedgehog induces the proliferation of primitive human hematopoietic cells via BMP regulation. *Nature Immunology* **2**, 172 – 180 (2001).
41. Katoh, Y. & Katoh, M. Hedgehog signaling pathway and gastrointestinal stem cell signalling network (review). *International journal of molecular medicine* **18**, 1019–1023 (2006).
42. Palma, V., Lim, D. A., Dahmane, N., Sánchez, P. & Brionne, T. C. Sonic hedgehog controls stem cell behavior in the postnatal and adult brain. *Development* **132 (2),** 335–344 (2004).
43. Lai, K., Kaspar, B. K., Gage, F. H. & Schaffer, D. V. Sonic hedgehog regulates adult neural progenitor proliferation in vitro and in vivo. *Nature Neuroscience* **6**, 21–27 (2003).
44. Mill, P. *et al.* Sonic hedgehog-dependent activation of Gli2 is essential for embryonic hair follicle development. *Genes & Development* **17**, 282–294 (2003).
45. Li, N. *et al.* Reciprocal intraepithelial interactions between TP63 and hedgehog signaling regulate quiescence and activation of progenitor elaboration by mammary stem cells. *Stem Cells* **26**, 1253–1264.
46. Plaisant, M. *et al.* Inhibition of Hedgehog Signaling Decreases Proliferation and Clonogenicity of Human Mesenchymal Stem Cells. *PLoS One* **6 (2),** e16798 (2011).
47. Miao, N. *et al.* Sonic hedgehog promotes the survival of specific CNS neuron populations and protects these cells from toxic insult in vitro. *The Journal of Neuroscience* **17 (15),** 5891–5899 (1997).
48. Rafuse, V. F., Soundararajan, P., Leopold, C. & Robertson, H. A. Neuroprotective properties of cultured neural progenitor cells are associated with the production of sonic hedgehog. *Neuroscience* **131**, 899–916 (2005).
49. Bragina, O. *et al.* Smoothened agonist augments proliferation and survival of neural cells. *Neuroscience Letters* **482**, 81–85 (2010).
50. Dass, B. *et al.* Behavioural and immunohistochemical changes following supranigral administration of sonic hedgehog in 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine-treated common marmosets. *Neuroscience* **114 (1),** 99–109 (2002).
51. Tsuboi, K. & Shults, C. W. Intrastriatal injection of Sonic Hedgehog reduces behavioral impairment in a rat model of Parkinson's disease. *Experimental Neurology* **173**, 95–104 (2002).
52. Bak, M. *et al.* Mutation analysis of the Sonic hedgehog promoter and putative enhancer elements in Parkinson's disease patients. *Molecular Brain Research* **126**, 207–211 (2004).
53. Merchant, A. A. & Matsui, W. Targeting hedgehog - a cancer stem cell pathway. *Clinical Cancer Research* **16 (12),** 3130–3140 (2010).
54. Subramaniam, D., Ramalingam, S. Houchen, C. W. & Anant, S. Cancer stem cells: a novel paradigm for cancer prevention and treatment. *Mini-Reviews in Medicinal Chemistry* **10**, 359–371 (2010).
55. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nature Genetics* **36**, 512 – 517 (2004).
56. Valente, A. X. C. N., Zischkau, J., Gao, Y. & Sarkar, A. GWAS heterogeneous population normalization via subject weight knock-down. *In preparation* (2011).
57. Chambers, S. M. *et al.* Hematopoietic Fingerprints: An Expression Database of Stem Cells and Their Progeny. *Cell Stem Cell* **1**, 578–591 (2007).
58. Keshava Prasad, T. S. *et al.* Human Protein Reference Database - 2009 Update. *Nucleic Acids Research* **37 (database issue),** D767–72 (2009).
59. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Research* **38** (database issue) D497–501 (2010).
60. Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81 (3),** 559–575 (2007).

Author contributions

A.X.C.N.V. and Y.G. conceived the study; A.X.C.N.V. wrote the manuscript; J.H.S. and A.S. gave technical support and conceptual advice.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Valente, A.X.C.N., Shin, J.H., Sarkar, A. & Gao, Y. Rare coding SNP in DZIP1 gene associated with late-onset sporadic Parkinson's disease. *Sci. Rep.* **2**, 256; DOI:10.1038/srep00256 (2012).