European Journal of Human Genetics (2014) 22, 238–242

www.nature.com/ejhg

ARTICLE

Accurate prediction of a minimal region around a genetic association signal that contains the causal variant

Zoltán Bochdanovits*[,1], Javier Simón-Sánchez[1], Marianne Jonker[2], Witte J Hoogendijk[3,4], Aad van der Vaart[2] and Peter Heutink[1]

In recent years, genome-wide association studies have been very successful in identifying loci for complex traits. However, typically these findings involve noncoding and/or intergenic SNPs without a clear functional effect that do not directly point to a gene. Hence, the challenge is to identify the causal variant responsible for the association signal. Typically, the first step is to identify all genetic variation in the locus region, usually by resequencing a large number of case chromosomes. Among all variants, the causal one needs to be identified in further functional studies. Because the experimental follow up can be very laborious, restricting the number of variants to be scrutinized can yield a great advantage. An objective method for choosing the size of the region to be followed up would be highly valuable. Here, we propose a simple method to call the minimal region around a significant association peak that is very likely to contain the causal variant. We model linkage disequilibrium (LD) in cases from the observed single SNP association signals, and predict the location of the causal variant by quantifying how well this relationship fits the data. Simulations showed that our approach identifies genomic regions of on average ~50 kb with up to 90% probability to contain the causal variant. We apply our method to two genome-wide association data sets and localize both the functional variant REP1 in the α-synuclein gene that conveys susceptibility to Parkinson's disease and the *APOE* gene responsible for the association signal in the Alzheimer's disease data set.
European Journal of Human Genetics (2014) **22,** 238–242; doi:10.1038/ejhg.2013.115; published online 5 June 2013

Keywords: complex trait; causal variant; association

INTRODUCTION

Identification of the causal variant detected by a genome-wide significant association signal remains a challenge. Both the common and rare variants can, in principle, underlie GWAS signals detected using common polymorphic markers.[1] To identify the variant responsible for the observed association, typically an extended genomic region around a replicated GWAS locus is resequenced in a large number of case chromosomes to identify all genetic variants. These are then prioritized either by analytical methods[2] or based on biological information for further follow up. Clearly, both the success and the cost of this approach strongly depend on the size of the genomic region included for resequencing. At a higher cost, a larger region can be resequenced to increase the probability that the true causal variant is included, but the total number of genetic variants that will be identified, and need further evaluation, will also increase. Therefore, an approach to accurately delineate a minimal region around a significant association signal that can be expected to harbor the causal variant is of great relevance, both to reduce cost and to make functional follow up amenable. We propose to identify a region that is likely to contain the causal variant by observing that around an association signal linkage disequilibrium (LD) between neutral polymorphic markers is known to be stronger in cases compared with controls. This property has been proposed as a test for association on its own.[3] We make use of this observation by explicitly modeling the LD between neutral markers in cases, as a function of the unknown causal variant. The most likely position of the causal variant will be estimated by considering a region of neighboring neutral SNPs, where LD in cases can be most accurately explained by the presence of the unknown causal variant.

METHODS

First we give the correlation (LD) between a neutral marker and the causal variant conditional on case status (see Supplementary Methods).

$$r_{MD|C} = r_{MC}/r_{DC} \times (p_M p_m / p_{M|C}\, p_{m|C})^{0.5} \times (p_{D|C} p_{d|C} / p_D p_d)^{0.5} \quad (1)$$

Here r_{MC} is the Pearson correlation coefficient between the marker and phenotype, that is, a measure of the observed association. p_M, p_m, $p_{M|C}$ and $p_{m|C}$ are the allele frequencies of the marker in the general population and in the cases. The terms involving subscript D are the unknown properties of the causal variant, that is, the penetrance and the frequency in general population and in cases. Further, we assume that the correlation between three adjacent loci (marker 1—causal variant—marker 2) in the order M1-D-M2 is multiplicative (see also[4]).

$$r_{M1M2|C} = r_{M1D|C} \times r_{M2D|C} \quad (2)$$

From 1 and 2 follows (in a simplified notation) that the LD between markers 1 and 2 in cases is a linear function of the observed association with

[1]Department of Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands; [2]Section Stochastics, Department of Mathematics, Faculty of Sciences, Vrije Universiteit, Amsterdam, The Netherlands; [3]Department of Psychiatry, VU University Medical Center, Amsterdam, The Netherlands; [4]Department of Psychiatry, Erasmus Medical Center, Rotterdam, The Netherlands
*Correspondence: Dr Z Bochdanovits, Department of Clinical Genetics, VU University Medical Center, Van der Boechorststraat 7 1081 BT Amsterdam, The Netherlands. Tel: +31 (0)20 5982813; Fax: +31 (0)20 5983596; E-mail: z.bochdanovits@vumc.nl
Received 21 April 2011; revised 19 April 2013; accepted 23 April 2013; published online 5 June 2013

the phenotype at markers 1 and 2 and the allele frequencies.

$$r_{M1M2|C} = r_{M1C} \times r_{M2C} \times P_{M1} \times P_{M2} \times P_D^2 / r_{DC}^2 \quad (3)$$

This simple relationship between LD and single SNP associations assumes that the order of the markers is M1-D-M2. This assumption implicitly also states that there is only one causal variant, hence the unknown value of P_D^2/r_{DC}^2 (ie, the slope of the regression) is the same for all marker pairs. To call the most likely location of the causal variant, we quantify how accurately the equation 3 fits the data for the different potential locations of the causal variant following the procedure described below.

Description of the algorithm

1. Take the lowest single SNP *P*-value in the associated region.
2. Take R SNPs left and right the top SNP found in 1. If multiple SNPs have equally low *P*-values, take R SNPs to the left of the leftmost and R SNPs right of the rightmost SNP. Choose R such that the region is large enough to be certain that the causal variant is present. A value of 100 was used, that is, a region of 200 consecutive SNPs was considered.
3. Within the range defined in step 2, take sliding windows of S SNPs. Assume that the causal variant is in the middle of the window and take all SNP pairs within the window that are on either side of the designated position of the causal variant (ie, in order M1-D-M2). On the basis of all such SNP pairs, compute the Pearson correlation coefficient *r* between the product $r_{M1C} \times r_{M2C} \times P_{M1} \times P_{M2}$ (ie, the predicted LD between M1 and M2 in cases barring a constant) and $r_{M1M2|C}$ (observed LD between M1 and M2 in cases).
4. Slide the window of size S SNP by SNP through the range defined under step 2 and calculate for every potential position of the causal variant, how well the LD between neutral markers in cases can be explained by the single SNP associations, quantified as the Pearson correlation coefficient *r*.
5. Define the region that is expected to hold the causal variant, by taking the *r* value from step 4 observed in the sliding window around the most strongly associated single SNP, and move out until the correlation coefficient drops below 50% of this value. This is an arbitrary threshold that was shown in the simulations to work satisfactorily.

Simulations and application to Parkinson's disease (PD) and Alzheimer's disease (AD).

Case/control status was determined using an additive risk model and individual haplotypes were generated by resampling with replacement from the phase II CEU phased data. On a randomly chosen chromosome, SNPs of specified allele frequencies were sampled to represent the causal variant and correspondingly different relative risks (RR) were simulated (ie, higher RR for

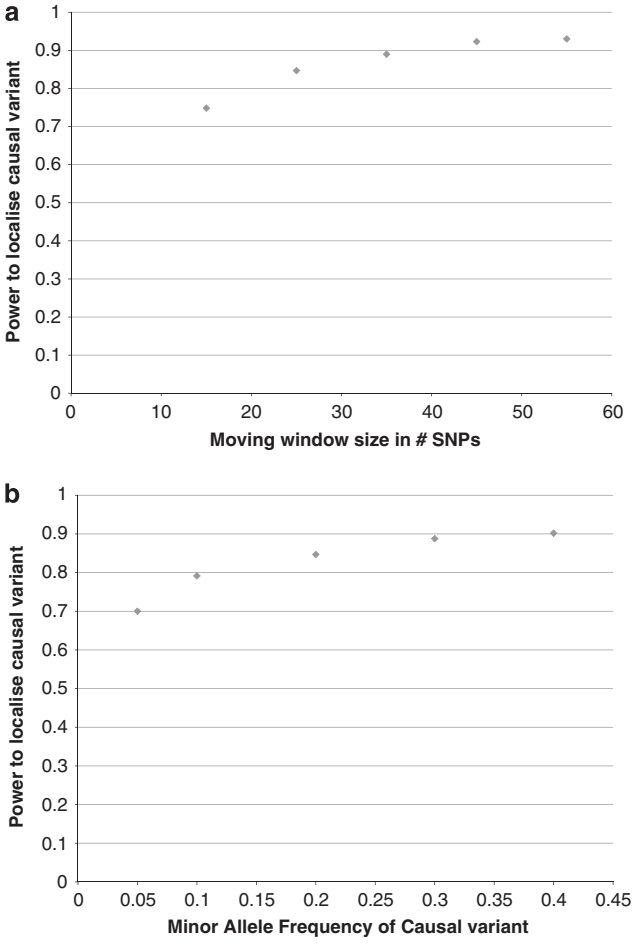

Figure 1 (**a**, **b**) Effect of (**a**) window size (MAF = 0.2) and (**b**) allele frequency (window size = 25) on power to localize the causal variant. With decreasing allele frequency, increasing RRs between 1.4 up to 2.5 were simulated to maintain a close to 100% genome-wide power to detect association.

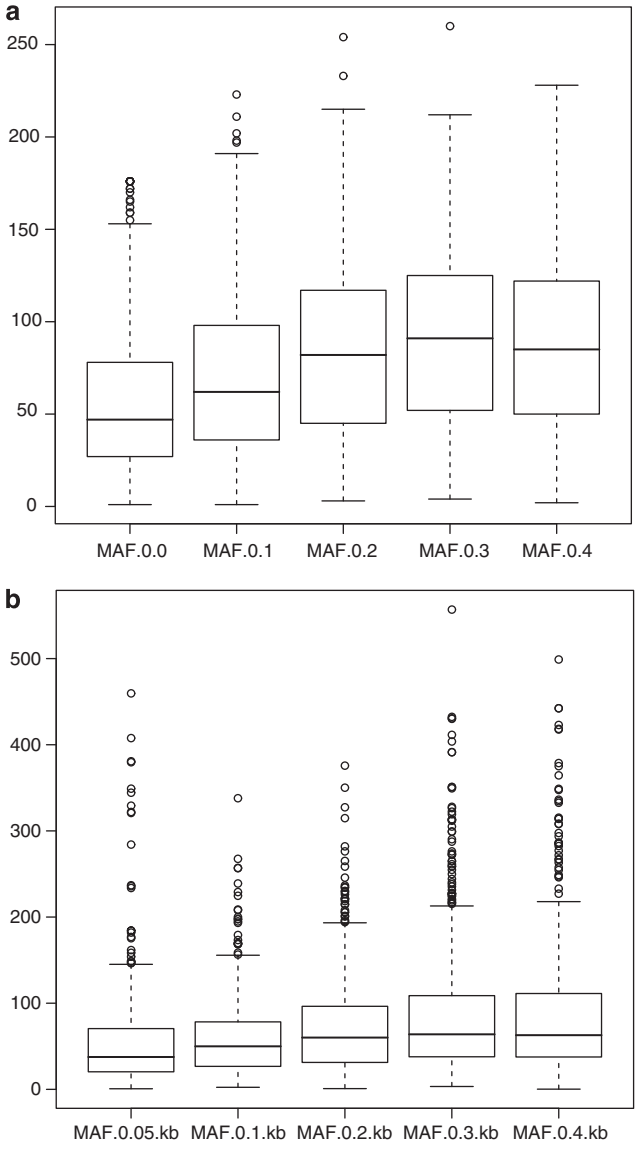

Figure 2 (**a**, **b**) Size distribution of the region called (*y* axis) to contain the causal variant in (**a**) number of SNPs and (**b**) kb for different allele frequencies (*x* axis).

low-frequency variants; RR ~ 1.4, 1.45, 1.5, 1.75 and 2.5 for MAF ~ 0.4, 0.3, 0.2, 0.1, 0.05, respectively), such that the genome-wide significant power to detect association given a fixed sample size of 2500 cases and 2500 controls would be above 99%. This simulation strategy represents the scenario of already having found a genome-wide significant association, which is subsequently being followed up. Before analysis the simulated causal variant was removed from the data. Single SNP association tests were performed and 100 SNPs (see step 2, the choice of R) left and right from the most strongly associated marker SNP (see step 1) were taken as the total range to follow up. For all scenarios 800 replicates were simulated. For the analysis of the PD and AD data, the raw genotypes on 300 SNPs around the most strongly associated SNP were analyzed. For the AD data, one individual per family was randomly included in the analysis. Only subjects consented for the General Research Use were included. For a detailed description of the samples and genotyping see[5] (PD data) and dbGaP (phs000168.v1.p1; AD data).

RESULTS AND DISCUSSION

We tested the accuracy of our procedure in simulated case–control association studies based on the HapMap phase II CEU data. Eighty-five percent power was observed for the localization of the causal variant for a causal allele frequency of 0.2 when using a window size of 25 SNPs (Figure 1a). Hence, in all further simulations and for the analysis of the PD and AD data, a window size of 25 SNPs was used. Power to localize the causal variant was found to depend on the allele frequency, with more rare susceptibility alleles having a lower probability to be localized correctly (Figure 1b). Note that this is not the result of a lower power to detect the association, because the simulations were conditioned on already having found a genome-wide significant signal that is being followed up. Specifically, the less frequent causal variants were assumed to have strong effects of up to a

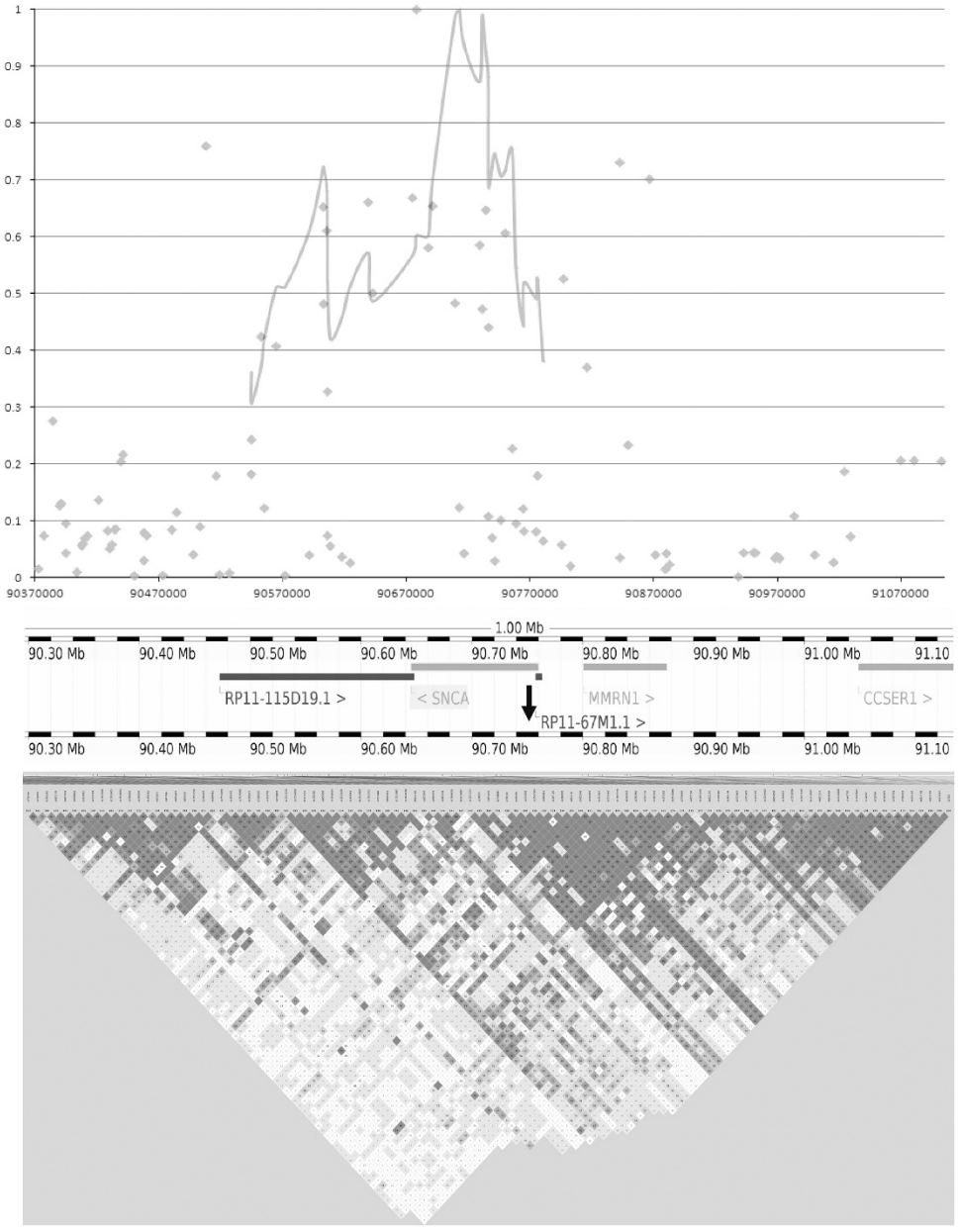

Figure 3 Local LD pattern of 100 SNPs around the SNCA association signal (bottom panel), genes located in the region (middle panel) and scaled negative log *P*-values (dots) *vs* scaled linear fit (r^2) of equation 3 (curve) (top panel). The region called to contain the causal variant is under the curve in the top panel and covers the known position of the REP1 promoter polymorphism (red arrow). A full color version of this figure is available at the *European Journal of Human Genetics* journal online.

RR of 2.5. Instead, we observed that the size of the genomic region claimed to contain the causal variant also varied with the allele frequency, with smaller regions called around the less frequent variants. The median size of the predicted regions ranged between 47 SNPs/38 kb and 91 SNPs/64 kb and less than 5% was larger than ~200 kb (Figure 2). As our approach is based on the effect of a causal variant on the local LD structure in cases, this result suggests that rare variants, even when conveying a relatively strong effect, tend to affect LD over a shorter physical distance compared with common variants. Consequently, the lower power to correctly localize more rare causal variants (Figure 1b) is probably a result of this smaller average size of the called region. It should be noted here that the converse is also expected to be true. Phenomena that might bias our method to call too large regions will not reduce and may even increase the power to correctly localize the causal variant. For example, population structure and admixture is known to affect LD patterns, probably lowering the accuracy of our model. In this case, we expect to call too large regions, because the drop to 50% relative to a lower 'best' fit would occur over larger physical distance. In conclusion, we suggest that given a significant association signal, first the relatively small genomic sequences called by our approach should be followed up in fine-mapping/resequencing studies, while maintaining up to 90% confidence that the correct region is being considered.

The functional variant is only rarely known in susceptibility genes conveying risk for a common disease. As a proof of principle, we show that our method is able to accurately localize the known causal variants in the *SNCA* and *APOE* genes, conveying susceptibility to PD and AD, respectively. SNCA is the most strong risk factor for the common, nonfamilial type of PD,[6] with changes to the plasma level of the protein being associated with the affection status.[7,8] The disease associated change in SNCA expression has been shown to be caused by a functional promoter polymorphism known as REP1.[9,10] Here we aim to detect the known location of the REP1 variant in a recent PD GWAS data set.[5] In this study, the most strongly associated SNP on chromosome 4 was rs2736990 inside SNCA ($P = 8.1 \times 10^{-6}$). The pattern of the single SNP associations did not clearly point to a particular position within the *SNCA* gene, instead the top four most strongly associated SNPs were 360 kb apart and were located in several adjacent LD blocks of a total length of ~600 kb (Figure 3). Based only on the single SNP association signals and the underlying local LD pattern, up to 600 kb would need further consideration. In contrast, our method applied to these data correctly predicts that the causal variant should be within a 40 SNP and 231-kb region, including a noncoding genomic region upstream of SNCA (Figure 3, see red arrow for position of the functional promoter polymorphism REP1). Interestingly, visual inspection of the output correctly suggest that, within the claimed region, the causal variant should be located towards the transcription start site relative to the position of the most strongly associated single SNP (Figure 3). The relatively large physical size of the region compared with the number of SNPs reflects the lower SNP density of the Illumina genotyping array compared with the HapMap II data used in the simulations. Even so, a considerable gain would have been achieved when initially following up the region predicted by our method, rather than relying on single SNP associations and visual inspection of the local LD pattern. Moreover, imputation based on reference genotype data is common practice in the analysis of GWAS studies, and application of our method to imputed data easily circumvents the issue of marker density. In addition, we applied our method to the NIA Late Onset AD GWAS data, obtained from dbGaP (phs000168.v1.p1). The most well known and strongest genetic risk factor for AD is the APOE-ε4 allele on chromosome 19[11] and a genome-wide significant association signal around the known position of APOE is present in the CIDR data. The most strongly associated SNP was rs2075650 ($P = 2.2 \times 10^{-43}$) in the *TOMM40* gene. Compared with SNCA, this signal is more sharply delineated, but still the top seven SNPs cover several neighboring genes and LD blocks (Figure 4). Relying on visual inspection of the local LD pattern alone would lead to an ~200-kb region for further follow up. In contrast, our analysis predicts that the association is the result of the presence of a causal variant within a nine SNP, 37-kb region, entirely encompassing the *APOE* gene (Figure 4). Similar to the SNCA region, a considerable gain would have been obtained by applying our method to select the genomic region for further study.

Identifying the causal variant responsible for an association signal remains a challenge and the first practical consideration is to select the genomic region for follow up. Resequencing large genomic regions at the associated locus is not only costly but also results in large number of potential risk variants for functional analysis. A recently published analytical approach by Zhu *et al*,[12] addresses this problem by considering pre-genotyped reference data sets. From this dense

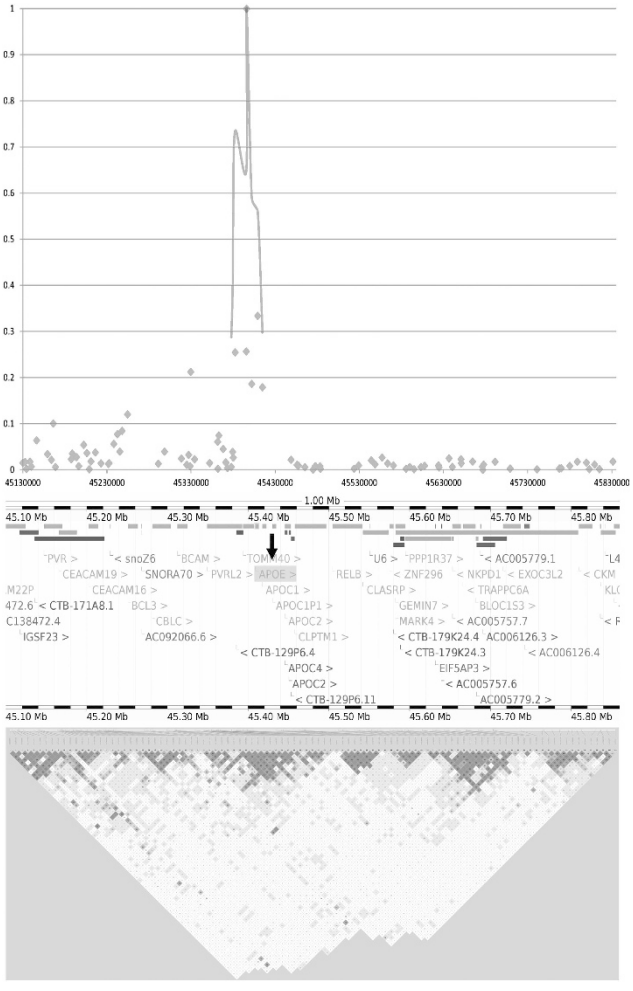

Figure 4 Local LD pattern of 100 SNPs around the APOE association signal (bottom panel), genes located in the region (middle panel) and scaled negative log *P*-values (dots) *vs* scaled linear fit (r^2) of equation 3 (curve) (top panel). The region called to contain the causal variant is under the curve in the top panel and covers the known position of APOE (red arrow). A full color version of this figure is available at the *European Journal of Human Genetics* journal online.

genotype data, which is assumed to contain the causal variant itself, individual SNPs are assigned a probability to be functionally linked to the phenotype in a fixed 1 Mb region around the most significant GWAS signal. Although the assumption that the causal variant has been observed in the reference data probably will become increasingly plausible with continuing advances in uncovering human genetic variation, the fixed large size of the region to be included in the method described by Zhu *et al* seems very conservative in light of our results. Here we propose a method to delineate a minimal region where the causal variant is located and show both in simulations and application to real data sets that our approach very accurately identifies the location of causal variants responsible for association signals, without including external information other than the original GWAS data set. We suggest that considerable gain can be achieved when designing functional follow up of genome-wide association studies by applying our approach.

A perl script is available from the corresponding author for performing the analysis.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This study was carried out within the framework of the Top Institute Pharma project: number T5-203. The Alzheimer's disease data set (phs000168.v1.p1) was obtained from dbGaP. The Parkinson's disease data set was generated within the Dutch PD consortium and was financially supported by Prinses Beatrix Fonds grant W.ORO9-25. Simulations were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org), which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003) along with a supplement from the Dutch Brain Foundation and the VU University Amsterdam.

1 Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB: Rare variants create synthetic genome-wide associations. *PLoS Biol* 2010; **8**: e1000294.
2 Udler MS, Tyrer J, Easton DF: Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genet Epidemiol* 2010; **34**: 463–468.
3 Zaykin DV, Meng Z, Ehm MG: Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. *Am J Hum Genet* 2006; **78**: 737–746.
4 Weir BS: Linkage disequilibrium and association mapping. *Annu Rev Genomics Hum Genet* 2008; **9**: 129–142.
5 Simon-Sanchez J, van Hilten JJ, van de Warrenburg B *et al*: Genome-wide association study confirms extant PD risk loci among the Dutch. *Eur J Hum Genet* 2011; **19**: 655–661.
6 Bekris LM, Mata IF, Zabetian CP: The Genetics of Parkinson disease. *J Geriatr Psychiatry Neurol* 2010; **23**: 228–242.
7 Mata IF, Shi M, Agarwal P *et al*: SNCA variant associated with Parkinson disease and plasma alpha-synuclein level. *Arch Neurol* 2010; **67**: 1350–1356.
8 Chiba-Falek O, Lopez GJ, Nussbaum RL: Levels of alpha-synuclein mRNA in sporadic Parkinson disease patients. *Mov Disord* 2006; **21**: 1703–1708.
9 Chiba-Falek O, Kowalak JA, Smulson ME, Nussbaum RL: Regulation of [alpha]-synuclein expression by Poly (ADP ribose) polymerase-1 (PARP-1) Binding to the NACP-Rep1 polymorphic site upstream of the SNCA gene. *Am J Hum Genet* 2005; **76**: 478–492.
10 Chiba-Falek O, Nussbaum RL: Effect of allelic variation at the NACP-Rep1 repeat upstream of the alpha-synuclein gene (SNCA) on transcription in a cell culture luciferase reporter system. *Hum Mol Genet* 2001; **10**: 3101–3109.
11 Bertram L, Lill CM, Tanzi RE: The genetics of Alzheimer disease: back to the future. *Neuron*, **68**: 270–281.
12 Zhu Q, Ge D, Heinzen EL *et al*: Prioritizing genetic variants for causality on the basis of preferential linkage disequilibrium. *Am J Hum Genet* 2012; **91**: 422–434.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)