6-19-2012

Pooling/bootstrap-based GWAS (pbGWAS) identifies new loci modifying the age of onset in PSEN1 p.Glu280Ala Alzheimer's disease

J. I. Velez
Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health

S. C. Chandrasekharappa
Genome Technology Branch, National Human GenomeResearch Institute, National Institutes of Health

E. Henao
Universidad de Antioquia

A. F. Martinez
Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health

U. Harper
Genome Technology Branch, National Human GenomeResearch Institute, National Institutes of Health

See next page for additional authors

Recommended Citation

Velez, J. I.; Chandrasekharappa, S. C.; Henao, E.; Martinez, A. F.; Harper, U.; Jones, M.; Solomon, B. D.; Lopez, L.; Garcia, G.; Aguirre-Acevedo, D. C.; Acosta-Baena, N.; Correa, J. C.; Lopera-Gomez, C. M.; Jaramillo-Elorza, M. C.; Rivera, D.; Kosik, K. S.; Schork, N. J.; and Swanson, J. M., "Pooling/bootstrap-based GWAS (pbGWAS) identifies new loci modifying the age of onset in PSEN1 p.Glu280Ala Alzheimer's disease" (2012). *All Faculty*. 37.
https://digitalcommons.fiu.edu/all_faculty/37

Authors

J. I. Velez, S. C. Chandrasekharappa, E. Henao, A. F. Martinez, U. Harper, M. Jones, B. D. Solomon, L. Lopez, G. Garcia, D. C. Aguirre-Acevedo, N. Acosta-Baena, J. C. Correa, C. M. Lopera-Gomez, M. C. Jaramillo-Elorza, D. Rivera, K. S. Kosik, N. J. Schork, and J. M. Swanson

Open

ORIGINAL ARTICLE

Pooling/bootstrap-based GWAS (*pb*GWAS) identifies new loci modifying the age of onset in *PSEN1* p.Glu280Ala Alzheimer's disease

JI Vélez[1,10], SC Chandrasekharappa[2], E Henao[3], AF Martinez[1], U Harper[2], M Jones[2], BD Solomon[1], L Lopez[3], G Garcia[3], DC Aguirre-Acevedo[3], N Acosta-Baena[3], JC Correa[4], CM Lopera-Gómez[4], MC Jaramillo-Elorza[4], D Rivera[3], KS Kosik[5], NJ Schork[6], JM Swanson[7,8], F Lopera[3,10] and M Arcos-Burgos[1,3,9]

The literature on GWAS (genome-wide association studies) data suggests that very large sample sizes (for example, 50,000 cases and 50,000 controls) may be required to detect significant associations of genomic regions for complex disorders such as Alzheimer's disease (AD). Because of the challenges of obtaining such large cohorts, we describe here a novel sequential strategy that combines pooling of DNA and bootstrapping (*pb*GWAS) in order to significantly increase the statistical power and exponentially reduce expenses. We applied this method to a very homogeneous sample of patients belonging to a unique and clinically well-characterized multigenerational pedigree with one of the most severe forms of early onset AD, carrying the *PSEN1* p.Glu280Ala mutation (often referred to as E280A mutation), which originated as a consequence of a founder effect. In this cohort, we identified novel loci genome-wide significantly associated as modifiers of the age of onset of AD (*CD44*, rs187116, $P = 1.29 \times 10^{-12}$; *NPHP1*, rs10173717, $P = 1.74 \times 10^{-12}$; *CADPS2*, rs3757536, $P = 1.54 \times 10^{-10}$; *GREM2*, rs12129547, $P = 1.69 \times 10^{-13}$, among others) as well as other loci known to be associated with AD. Regions identified by *pb*GWAS were confirmed by subsequent individual genotyping. The *pb*GWAS methodology and the genes it targeted could provide important insights in determining the genetic causes of AD and other complex conditions.

INTRODUCTION

Large-scale individual genotyping is typically used for genome-wide association studies (GWAS).[1] This approach is frequently laborious, expensive, and time consuming, especially in studies requiring large numbers of participants.[2] Furthermore, initial theoretical expectations of sample sizes required to detect small effects were not met, and now large sample requirements are proposed.[3] DNA pooling and group genotyping can dramatically reduce the cost of large-scale GWAS aimed at mapping disease susceptibility loci.[4]

In DNA pooling, as opposed to individual genotyping, individual samples from n_1 cases and n_2 controls are combined into two independently pooled samples.[4] In general, cases and controls are defined based on either the presence or absence of a disease trait, or as related to a quantitative trait (for example, by dichotomizing the continuous trait by selecting the two extremes of the distribution).[4,5] Disease-associated single-nucleotide polymorphisms (SNPs) are identified by a statistical test[4] after correction by multiple testing.[6] To corroborate results, SNPs shown to be disease associated in the DNA pools (as well as other SNPs in the

vicinity) are genotyped, preferably in an independent sample for strong replication.[4]

The pooling strategy described above is very efficient in terms of genotyping costs and for the initial discovery phase. However, for replication, DNA samples from new patients are necessary, which constitutes a crucial limitation, especially in the case of rare diseases.

Here, we describe a new pooling/bootstrap-based GWAS strategy (*pb*GWAS), aimed at the identification of disease-associated SNPs; this strategy makes use of DNA pooling and resampling (bootstrapping)[7] to randomly sort samples from cases and controls in order to generate multiple pairs of new comparable DNA pools. Statistical evidence from multiple comparisons of the pairs is combined in a sequential way using formal meta-analysis (Figure 1).

To show the high efficiency of the *pb*GWAS method, in terms of both genotyping costs and statistical power, we present the mapping of several modifying loci that influence the age of onset of Alzheimer's disease (AD), in individuals carrying the *PSEN1* p.Glu280Ala mutation (often referred to as E280A mutation), a
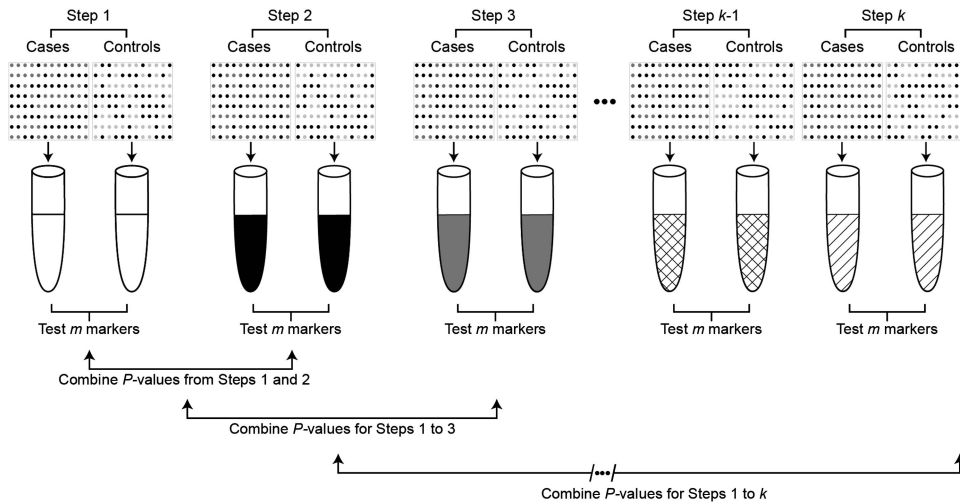
Figure 1. Pooling/bootstrap genome-wide association study (*pb*GWAS) strategy when considering $n_1 = n_2 = 96$ cases and controls. After generating a total of k pairs of randomplates (steps) by bootstrapping DNA samples, m SNPs (single-nucleotide polymorphisms) are tested on pooled samples at each step. Results from pairs 1 to k is combined using meta-analytical methods.
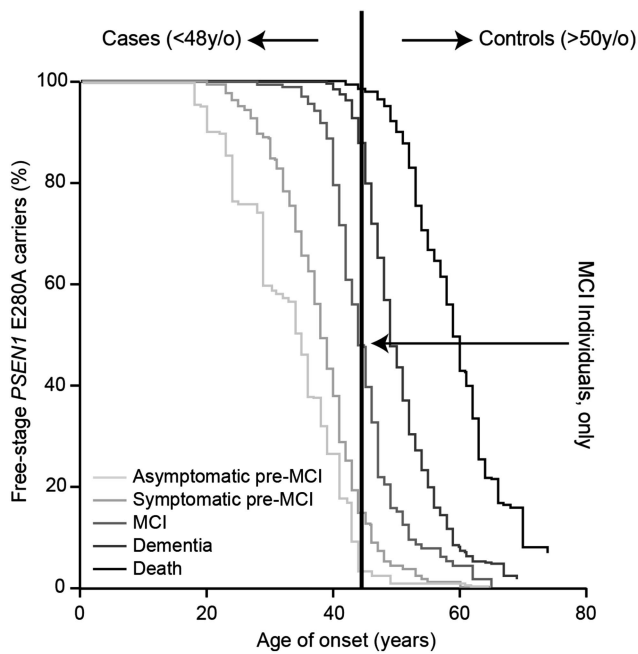

Figure 2. Definition of case and control individuals in our cohort of Alzheimer's disease patients based on a survival analysis of disease progression in *PSEN1* p.Glu280Ala carriers. Modified from Acosta-Baena et al.[8] MCI, mild cognitive impairment.

unique and fully penetrant variant that segregates in one of the largest and clinically well-characterized multigenerational pedigrees.[8,9] Along with a unique causal mutation, the pedigree inhabits a homogeneous environment, and the mean age of onset of AD falls quite tightly around a mean age of 47 years (s.d. = 6.4 years).[9] However, there are outliers with onset ages ranging from 34 to ~65 years. The signs and symptoms of the entire kindred with early-onset Alzheimer's disease (EOAD) have been rigorously followed,[8] as we have identified the vast majority of the p.Glu280Ala mutation carriers.

Using *pb*GWAS with eight pairs of resampled DNA pools from cases (p.Glu280Ala carriers with EOAD) and controls (p.Glu280Ala carriers with late-onset Alzheimer's disease (LOAD)) at a total

genotyping cost < $20,000, we successfully located new loci with associations that exceed standard thresholds for GWAS significance and also replicated other loci previously reported to be associated with AD.

MATERIALS AND METHODS

Subjects
A total of 1784 descendants of AD patients with confirmed *PSEN1* p.Glu280Ala mutations were enrolled in the E280A Antioquia cohort study at the University of Antioquia, Colombia, between January 1, 1995 and January 27, 2010. These patients were clinically followed every 2 years. Detailed description of the enrollment, clinical characteristics, and inclusion criteria are reported elsewhere.[8] The *PSEN1* mutation in this pedigree is the result of a founder effect dating back to the seventeenth century; now, after ~20 generations, the pedigree includes >3000 people with several hundred carriers of the mutation. p.Glu280Ala AD patients with known age of onset of AD symptoms were selected for this specific study. Because of the size of the pedigree, any bias of ascertainment and family structure effect may be considered as minimal. Based on a survival analysis,[8] age cutoff values were matched to the 40th and 60th percentile of the survival curve, and patients for whom AD symptoms appeared before 48 years of age (EOAD) were classified as cases; those for whom the age of appearance was after 50 years of age (LOAD) were classified as controls[8] (Figure 2). A total of 64 cases and 38 controls, all of them *PSEN1* p.Glu280Ala carriers, included in this study, were selected from the 1784 participants enrolled in the study by Acosta-Baena et al.[8]

DNA extraction and pool preparation
DNA was isolated from whole blood using the traditional phenol/chloroform extraction followed by ethanol precipitation. The Infinium HD[10] and GoldenGate[11] assays have low DNA loading requirements and offer the flexibility to support diverse experimental designs. DNA extracted from blood samples, using the solid-phase platforms from QIAGEN (QIAGEN Sciences, MD) produces high-quality genotypes and copy number variants (CNV) data. In our study, we went further and performed phenol/chloroform extraction of DNA samples to guarantee the highest purity possible (phenol/chlorhorm extraction is the preferred method of DNA isolation for more sensitive technologies like next-gen sequencing). DNA samples were verified for integrity by agarose gel electrophoresis and samples with 260/280 absorbance ratios <1.8 were again phenol/chloroform-extracted. Once the DNA samples were ready for the pooling experiments, their concentrations were measured fluorometrically to

guarantee accuracy. DNA concentration was measured using the Qubit fluorometer (Invitrogen, Carlsbad, CA, USA). To prepare the DNA pools, individual DNA samples were diluted to 75 ng μl^{-1} and transferred to two 96-well plates, one for EOAD and one for LOAD. In all, 5 μl of each of the selected samples for a particular pool were mixed together in a single tube (see Figure 1) and submitted for genotyping.

In addition, to control the inherent variability attributable to technical issues, we also did try to minimize any variability originated as a consequence of different technical and personal skills. A single senior technician was involved in the generation of the stem pools and subsequent bootstrapped subsets. Furthermore, processing of pools at each step was done simultaneously for cases and controls to avoid any batch effects.

Comparable pools were constructed by selecting half of the individuals' DNA samples available from each group using a bootstrapping procedure. Though each pair of DNA pools did not contain either duplicated or triplicated individuals, a DNA sample can result to be (randomly) selected to be part of more than one pool as a consequence of the resampling process.

pbGWAS strategy

DNA pools generation. We used the following algorithm for DNA pool generation: let h be the total number of *new* pairs of DNA pools to be generated using resampling/bootstrapping and n_1 and n_2 be the total number of individual DNA samples available for cases and controls, respectively. In order to illustrate how the strategy works, let us assume, without loss of generality, that $n_1 = n_2 = n = 96$ such that all samples can be easily organized in a 96-well format plate (Figure 1). To guarantee the highest variance in the resampling process, we randomly selected half of the individuals' DNA samples [$n_j/2$] from each group, $j = 1,2$. Although in practice it is possible to generate a very large number of bootstrap replicates, this is not always feasible and some general alternatives have been proposed.[12,13]

Allele frequency estimation. Genotyping was performed at the NHGRI Genome Technology Branch using 370CNV-Quad SNP-chips from Illumina (www.illumina.com). Genotyping was performed using Illumina 'Infinium assay' protocol.[10] In brief, the DNA was whole-genome amplified, fragmented, hybridized, fluorescently tagged, and scanned. Allele frequency values for each of the eight pairs of DNA pools were calculated, extracted, and exported in suitable formats using the Illumina's GenomeStudio Genotyping Module. Standard quality control was applied.

Determining disease-associated SNPs. Without loss of generality, assume that (i) h SNP chips consisting of s diallelic markers have been genotyped in cases and controls, (ii) the allele frequency of the major allele is estimated within each group using standard methods,[4] and (iii) after experimental and statistical quality control, m ($m < s$) SNPs are available for analysis. Let $p_{1,i}$ and $p_{2,i}$ be the allele frequency for the i-th SNP in cases and controls, respectively, $i = 1,2,\ldots,m$. To test whether the allele frequency in the cases is not statistically different from that in the controls, the null hypothesis is stated as

$$H_{0,i} : p_{1,i} = p_{2,i} \quad i = 1,2,\ldots,m \quad (1)$$

with the alternative hypothesis $H_{1,i}$ being either (i) $p_{1,i} \neq p_{2,i}$, (ii) $p_{1,i} > p_{2,i}$, or (iii) $p_{1,i} < p_{2,i}$. We do not consider (i) as it does not provide information about the direction of the difference in the allele frequencies.

For two independent pools, the test statistic comparing the allele frequencies for the i-th SNP has the following form[4,14,15]

$$T_i^2 = \frac{(\hat{p}_{1,i} - \hat{p}_{2,i})^2}{\hat{\sigma}_i^2} \quad i = 1,2,\ldots,m \quad (2)$$

where $\hat{p}_1$ and $\hat{p}_2$ are the estimated allele frequencies of the major allele in the pair of pools for cases and controls, respectively, and $\hat{\sigma}_i$ is the variance of $\hat{p}_{1,i} - \hat{p}_{2,i}$. When a two-pool design is used, $\hat{\sigma}_i^2$ can be

calculated as[14]

$$\hat{\sigma}_i^2 = \hat{p}_i(1 - \hat{p}_i)(\tau^2 + 1)\left(\frac{1}{2n_1} + \frac{1}{2n_2}\right) + 2\varepsilon^2 \quad i = 1,2,\ldots,m \quad (3)$$

where $\hat{p}_i = (n_1 + n_2)^{-1}(n_1\hat{p}_{1,i} + n_2\hat{p}_{2,i})$ is the common allele frequency, τ is the coefficient of variation of the number of DNA molecules of locus A contributed by each individual, and ε^2 is the variance of the pool-measurement error.[14,16]

For testing (1), we propose a modification of (3) by introducing the estimated allele frequencies from both pools so that the estimation of $\hat{p}_i$ is no longer required, and we can focus the analysis on the difference of the allele frequencies between cases and controls for the SNPs being tested. Hence, the test statistic is

$$T_i^{*2} = \frac{(\hat{p}_{1,i} - \hat{p}_{2,i})^2}{\hat{\sigma}_i^{*2}} \quad i = 1,2,\ldots,m \quad (4)$$

with

$$\hat{\sigma}_i^{*2} = (\tau^2 + 1)\left\{\frac{\hat{p}_{1,i}(1 - \hat{p}_{1,i})}{2n_1} + \frac{\hat{p}_{2,i}(1 - \hat{p}_{2,i})}{2n_2}\right\} + 2\varepsilon^2 \quad i = 1,2,\ldots,m \quad (5)$$

With the current high-throughput genotyping technologies and DNA extraction techniques, it is possible to obtain accurate estimates of the allele frequency, so the sampling variation and random experimental errors are negligible compared with ε.[16,17] Now, if ε is assumed to be close to zero, (4) can be written as

$$T_i^{*2} = \frac{2n_1n_2(\hat{p}_{1,i} - \hat{p}_{2,i})^2}{n_2\hat{p}_{1,i}(1 - \hat{p}_{1,i}) + n_1\hat{p}_{2,i}(1 - \hat{p}_{2,i})} \quad i = 1,2,\ldots,m, \quad (6)$$

which follows a χ^2 distribution with one degree of freedom under the null hypothesis. When τ and ε^2 are known, (4) is preferred. Note that high values of T_i^{*2} indicate a stronger association between the i-th SNP and the disease.

Let P_i be the P-value calculated for the i-th SNP for testing (1). For each of the h *new* pairs of DNA pools generated, disease-associated SNPs are detected based on the P-values calculated from the T^* statistic for each of the m SNPs passing quality control. For m independent tests for each pair of pools, the family-wise error rate, defined as the probability that one or more of the significance tests results in a type I error[6] must be controlled, and standard methods can be applied, including Bonferroni,[18] false discovery rate (FDR),[19] spectral decomposition, which corrects by multiple testing when SNPs being tested are in linkage disequilibrium,[20] and a gene-based (GB) multiple testing correction.[21] Here we only focus on the FDR as it is particularly well suited for exploratory analyses.[22] Thus, for each of the h pairs of new DNA pools, the disease-associated SNPs are those such that $P_{ij}^* \leqslant \alpha$, with P_{ij}^* the FDR-corrected P-value for i-th SNP when the j-th pair of *new* DNA pools is being tested, $i = 1,2,\ldots,m$, $j = 1,2,\ldots,h$.

Combining P-values. Up to this point, we have illustrated how to generate *new* DNA pools for cases and controls by randomly selecting individuals' DNA samples and rearranging them accordingly, as well as the statistical hypothesis to be tested and how to determine whether the SNPs are disease-associated. Next, we shall describe the method selected for combining the P-values generated for each of the h pairs of *new* pools being compared.

Meta-analytical methods combine the results of multiple studies addressing a research question.[23,24] The combination of one-tailed P-values is one of the most widely used meta-analytical methods for summarizing information from k independent studies.[24] Fisher's combined probability test[25] and Stouffer's weighted Z-transform[26] are two such methods. In Fisher's method, the P-values from k tests are combined using the test statistic

$$\chi_F^2 = -2\sum_{l=1}^{k}\ln(P_i) \quad (7)$$

which follows a χ^2 distribution with $2k$ degrees of freedom under the *null* hypothesis of no significant effect in any study. Stouffer's method, on the

other hand, takes advantage of the one-to-one mapping of the standard normal distribution to the P-value and uses the test statistic[27]

$$Z_S = \frac{\sum_{l=1}^{k} w_l Z_l}{\sqrt{\sum_{l=1}^{k} w_l^2}}, \quad (8)$$

which follows a standard normal distribution under the null hypothesis (see above). In (8), w_l represents the weight and Z_l the quantile of the standard normal distribution (SND) associated with the P-value for study l, $l = 1, 2, \ldots, k$. When using either method, large values of the test statistic indicate that *at least* one of the studies being combined can reject its null hypothesis.[23,27] In our context, this means that at least in one of the *new* pair of pools the allele frequency for the i-th SNP, $i = 1, 2, \ldots, m$, is greater in cases than in controls. However, no information regarding the size of the experimental effect is available.[24] Because the Fisher's method is asymmetrically sensitive to small P-values compared with large P-values and is less powerful than Stouffer's method,[27] the latter was selected for combining P-values.

As we randomly select individual DNA samples from n_1 cases and n_2 controls to construct the multiple *new* DNA pools, the assumption of independence across the multiple pairs does not hold. However, it is possible to use Stouffer's method to combine dependent P-values by introducing some degree of dependence (correlation) between pairs.[28] Let $-(k+1)^{-1} \leqslant \rho \leqslant 1$ be such correlation and suppose that ρ is known. Thus, (8) can be written as[28]

$$Z_S^*(\rho) = \frac{\sum_{l=1}^{k} w_l Z_l}{\sqrt{(1-\rho)A + \rho B}} \quad (9)$$

with $A = \sum_{l=1}^{k} w_l^2$ and $B = (\sum_{l=1}^{k} w_l)^2$. When ρ is not known, it can be estimated as[28]

$$\hat{\rho} = \max\{-(k-1)^{-1}, 1 - V[\Phi^{-1}(Z)]\} \quad (10)$$

where $\Phi^{-1}(\cdot)$ is the inverse of the SND, $V[\cdot]$ is the variance operator and $Z = (Z_1, Z_2, \ldots, Z_k)$ is a vector containing the quantiles of the SND for the k studies. It follows[28] that the test statistic is

$$Z_S^*(\rho^*, \kappa) = \frac{\sum_{l=1}^{k} w_l Z_l}{\sqrt{A(1-C) + BC}} \quad (11)$$

with A and B as in (9), $C = \hat{p}^* + \sqrt{2\kappa}(1-\hat{p}^*)(k+1)^{-1/2}$ and $\kappa \geqslant 0$ a parameter regulating the significance level. Note that the one-tailed combined P-value, $P_{Z_S^*(\cdot)}$, can be easily computed as under the null hypothesis of no significant effect in any study, both (9) and (11) follow a SND. If desired, $P_{Z_S^*(\cdot)}$ can easily be converted to a two-tailed test.[27]

Individual genotyping

In order to validate our DNA pooling results, and to test independence of the type of chip used, we individually genotyped 23 samples (13 EOAD, 10 LOAD) using Human610-Quad SNP-chips from Illumina, and 48 samples (27 EOAD, 21 LOAD) using Human370-Quad SNP-chips. These individuals were selected using the same bootstrap strategy described before to generate the DNA pools.

Genetic statistical analysis

For the individually genotyped sample, genotypic (using dominant, recessive, and additive models) and allelic tests were performed as implemented in Golden Helix's Genotype module (Golden Helix, Bozeman, MT, Golden Helix PBAT Software, http://www.goldenhelix.com) with the onset age status (EOAD vs LOAD) as the phenotype. For all analyses performed, significance of markers was determined using the raw P-value for the corresponding test. Further, with this individually genotyped sample, we tested the potential presence of microdifferentiation (genotype stratification) between the EOAD and the LOAD samples, as this would potentially introduce important noise at

each bootstrapping step that consequently could invalidate any association result. To evaluate that possibility, we used the principal component analysis as implemented in the Golden Helix PBAT Software.

Gene-based association

First, the P-values for the top 5% of SNPs in our discovery phase, as well as their corresponding P-values in the individual genotyping analysis, were selected and combined using (9) with $\rho = 0$. Second, a GB association test using VEGAS (versatile GB association study)[21] was performed on the latter combined P-values. VEGAS is a freely available software used for performing GB association tests, which produces a GB test statistic and then uses a simulation-based approach to calculate an empirical GB P-value. By default, patterns of linkage disequilibrium for each gene are estimated using the HapMap2 CEU population.[21] SNPs and/or genes shown to be statistically significant using either test were classified as modifiers of the age of onset in AD.

RESULTS

Discovery phase

A total 102 DNA samples ($n_1 = 64$ cases and $n_2 = 38$ controls) were considered in this study. Comparison of the age of appearance of AD symptoms in both groups disclosed statistically significant differences (EOAD: average = 45.26 years, s.d. = 2.24, range = 39–48; LOAD: average = 54.89, s.d. = 4.64, range = 51–75; $t = -11.978$, d.f. = 47.481, $P < 5 \times 10^{-15}$). Eight pairs of DNA pools were generated using an algorithm written in R[29] as previously described. Allele frequencies for 373,397 variants (CNVs and SNPs) for all DNA pools were estimated using the 370CNV-Quad SNP chip, but only 287,368 were included in the final analysis. CNVs (~ 22 K) were excluded as they were not used for analysis and ~ 64 K SNPs were excluded because at least one of the allele frequencies in at least one of the 16 SNP-chips was <1 or $>99\%$. The call rates for the 16 DNA pools ranged from 65.1 to 66.2% when clustered with the Illumina cluster file, and from 97.6 to 99% when re-clustered against each other. A replication sample, selected *ad hoc*, was used for quality control demonstrating the high reliability of the allele frequency estimation (Figure 1, see Supplementary Material online). Q-Q plots for the FDR-corrected P-values obtained at each step of comparison depict good control of the type-I error probability, with SNPs with $P < 10^{-4}$ shown in green (Figure 2, see Supplementary Material online). In addition, an empirical evaluation of the type-I error probability was implemented using a simulation procedure that considered different scenarios, for example, different number of cases and controls, markers, and number of steps. Methods used for this simulation process as well as results are presented in the Supplementary Material (see Supplementary Figures 4–8, and Supplementary Table 1 online). In general, it was demonstrated that the type-I error probability was well controlled by our *pbGWAS* strategy.

Disease-associated SNPs for each pair of *new* DNA pools were determined using the corresponding FDR-corrected P-value of (6) with $n_1 = 64$ and $n_2 = 38$. Further, to obtain $P_{Z_S^*(\cdot)}$, FDR-corrected P-values from the eight pairs of pools were combined using (9) with $\rho = 0.95$ and $w_l = 1$, $l = 1, 2, \ldots, 8$. Figure 3 depicts Manhattan plots for loci modifying the age of onset in AD based on our discovery phase. As this figure shows, SNPs close to or within the *Nephronophthisis 1* (*NPHP1*) gene on chromosome 2, the *Arylsulfatase J* precursor (*ARSJ*) gene on chromosome 4, the *Calcium-Dependent Secretion Activator 2* (*CADPS2*) gene on chromosome 7, and the *CD44 antigen isoform 1 precursor* (*CD44*) gene on chromosome 11 were found to be statistically significant after combining the information from the generated DNA pools. Although additional loci were found to be significantly associated with the age of AD onset in our discovery phase, these four were selected as they are novel and because patterns of expression, as well as potential function, suggested them as promising candidate loci.
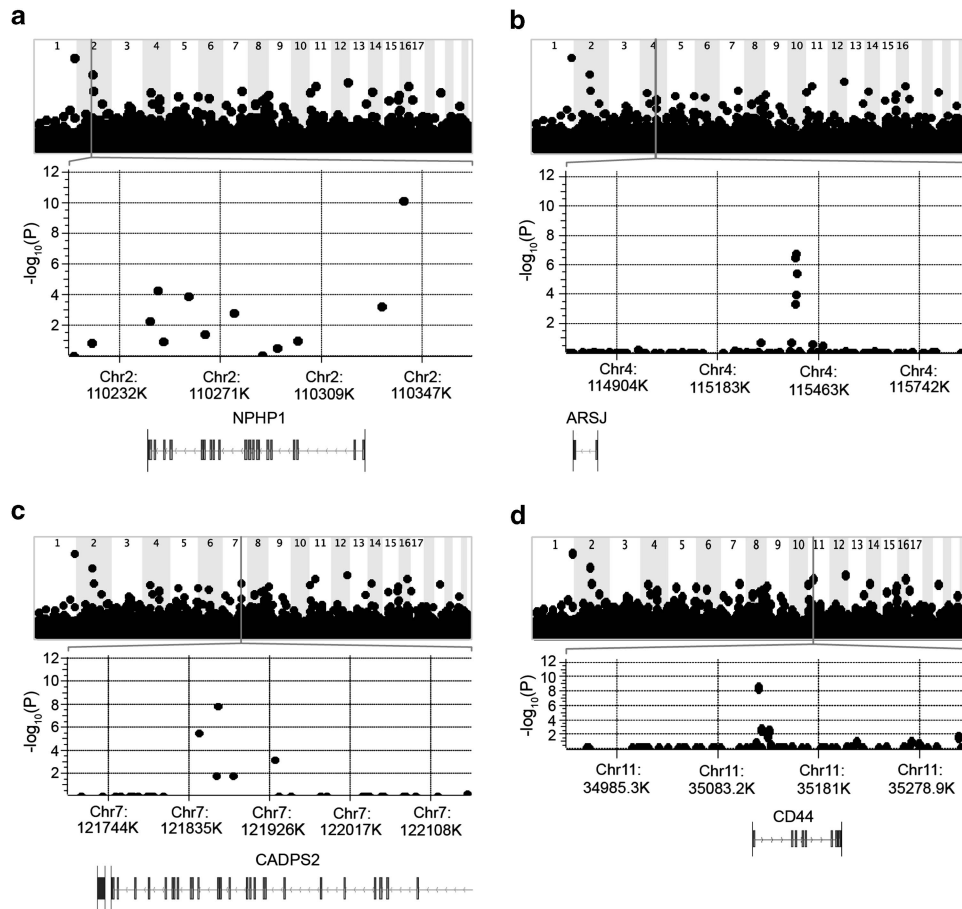
Figure 3. (a–d) Manhattan plot of the results of applying a pooling/bootstrap genome-wide association study (pbGWAS) strategy with eight pairs of DNA pools generated from patients from a unique and clinically well-characterized multigenerational pedigree affected by Alzheimer's disease with different ages of onset, all of whom carry the *PSEN1* p.Glu280Ala (E280A) mutation. At the y axis, the −log$_{10}$(P-value) for autosomal single-nucleotide polymorphisms are represented by dots; the x axis corresponds to the genomic coordinates. For display purposes, values in the y axis were smoothed by the median as implemented in Golden Helix. Abbreviations follow the same notation as in Figure 1. Though rs10173717 is ∼20 Kb downstream *NPHP1* and ∼2.5 Kb downstream of *NCRNA00116*, *NPHP1* appears to be a better candidate than *NCRNA00116* because*NPHP1* encodes a protein that interacts with *PTK2B* and *BCAR1*.[44,45] *PTK2B*, in turn, encodes a cytoplasmic protein having an important role as intermediate between neuropeptide-activated receptors or as neurotransmitters that increase calcium flux and the downstream signals that regulate neuronal activity. *BCAR1*, in turn, is involved in cellular migration, survival, transformation, and invasion.[46] Furthermore, the gene-based approach showed that a total of 10 markers inside the *NPHP1* gene (chr2: 110 238 202–110 319 928), clustered associated (statistic = 177.01, $P = 2.08 \times 10^{-4}$). This gene-based association was not present in *NCRNA00116*.

In Table 1, we present the top 30 SNPs based on the $P_{Z_5^*(\cdot)}$ values, the raw P-value for the basic genotypic test, and the combined P-value of both (see Materials and Methods section). Several SNPs shown here to act as modifiers of the age of onset in AD involve genetic loci previously reported to be associated with frank AD (that is, with the presence or absence of AD). For example, the *CADPS2* gene (marker rs3757536, Table 1) was found to be associated with AD in a voxelwise GWAS (vGWAS) aimed at detecting genes related to brain structure and function.[30] In the same vein, *CD44* gene (marker rs187116, Table 1) expression levels in lymphocytes are higher in AD subjects than in healthy subjects.[31] In addition, CD44-positive astrocytes have been shown to be involved in shaping normal neuronal mechanisms as well as in neurodegenerative process such as those involved in AD, Huntington, and Parkinson diseases.[32,33]

Individual genotyping
Call rates for the 23 samples (using the Illumina Human610-Quad SNP-chips) ranged from 37.8 to 99.8% when clustered with the Illumina cluster file, and from 56.2 to 99.6% when re-clustered against each other. For the Human370-Quad SNP-chip (48 samples), the calls ranged from 86.4 to 99.9% when clustered with the Illumina 370CNV-Quad cluster file. As implemented in GenomeStudio, genotypes for all variants were retrieved and further exported in the appropriate format to run genetic models using GoldenHelix. In Table 1, we present the corresponding raw P-values for the top 30 SNPs. Test for microdifferentiation showed no effect of stratification with two principal components extracted, excluding the possibility of type-I error inflation as a consequence of genotype stratification.

Combined statistics
A total of 14,369 SNPs were in the top 5% after combining the P-values from the discovery phase. Table 2 reports the genes shown to be statistically significantly associated with AD age of onset after performing the GB test on the combined P-value from the discovery phase and the individual genotyping using VEGAS. Q-Q plots for the FDR-corrected P-values obtained for the

Table 1. Pooling/bootstrap GWAS, individual genotyping, and combined *P*-values for the top 30 SNPs potentially modifying the age of onset in AD patients carrying the p.Glu280Ala mutation

Name	Chromosome	Position (bp)	Closest gene(s)	P-value		
				pbGWAS[a]	Individual genotyping[b]	Combined[c]
rs4341804	17	218,145	RPH3AL	1.86×10^{-9}	$1.9 \times 10^{-5\,d}$	7.10×10^{-13}
rs778296	13	104 986 574	DAOA	2.90×10^{-10}	$1.25 \times 10^{-4\,d}$	1.58×10^{-12}
rs4923709	15	34 405 797	ATPBD4/MEIS2	7.91×10^{-9}	$1.72 \times 10^{-4\,d}$	3.34×10^{-11}
rs187116	11	35 122 376	CD44	3.38×10^{-11}	$3.73 \times 10^{-4\,d}$	1.29×10^{-12}
rs203632	8	51 173 297	SNTG1	1.08×10^{-8}	$7.35 \times 10^{-4\,d}$	2.68×10^{-10}
rs1952966	14	31 508 235	NUBPL	5.93×10^{-10}	$8.22 \times 10^{-4\,d}$	3.36×10^{-11}
rs1437683	2	119 015 357	DDX18	5.68×10^{-10}	$8.77 \times 10^{-4\,d}$	3.56×10^{-11}
rs1974888	16	64 402 296	CDH11	7.76×10^{-10}	$1.39 \times 10^{-3\,d}$	8.53×10^{-11}
rs12129547	1	238 761 878	GREM2	9.07×10^{-14}	$1.69 \times 10^{-3\,d}$	1.69×10^{-13}
rs4977114	8	144 250 835	CYP11B1/LY6E	3.26×10^{-10}	$2.77 \times 10^{-3\,d}$	1.23×10^{-10}
rs4865656	5	50 695 545	ISL1	2.33×10^{-11}	$4.47 \times 10^{-3\,d}$	3.95×10^{-11}
rs6489968	12	113 623 027	TBX3	2.73×10^{-12}	$8.22 \times 10^{-3\,d}$	2.51×10^{-11}
rs533178	4	56 844 407	AASDH/PPAT/PAICS	6.41×10^{-9}	$1.00 \times 10^{-2\,d}$	7.27×10^{-9}
rs889701	16	6 074 670	A2BP1	4.89×10^{-8}	$1.06 \times 10^{-2\,d}$	3.34×10^{-8}
rs10173717	2	110 340 339	NPHP1	2.45×10^{-14}	$1.06 \times 10^{-2\,d}$	1.74×10^{-12}
rs4279283	4	115 401 739	CAMK2D	4.04×10^{-8}	$1.09 \times 10^{-2\,d}$	3.05×10^{-8}
rs3757536	7	121 867 133	CADPS2	1.21×10^{-11}	$1.31 \times 10^{-2\,d}$	1.54×10^{-10}
rs1134597	16	3 529 111	CLUAP1/NLCR3	3.80×10^{-9}	$1.65 \times 10^{-2\,d}$	1.12×10^{-8}
rs2998144	10	134 852 827	KNDC1	4.19×10^{-10}	$1.70 \times 10^{-2\,d}$	2.63×10^{-9}
rs7662084	4	115 398 766	CAMK2D/ARSJ	2.94×10^{-9}	$1.77 \times 10^{-2\,d}$	1.06×10^{-8}
rs6419191	4	115 400 783	ANK2/CAMK2D	1.64×10^{-9}	$2.28 \times 10^{-2\,d}$	1.09×10^{-8}
rs10863904	1	209 750 959	RD3	2.76×10^{-8}	$2.43 \times 10^{-2\,d}$	8.19×10^{-8}
rs2575735	8	97 603 827	SDC2	2.24×10^{-8}	$3.54 \times 10^{-2\,d}$	1.34×10^{-7}
rs9283839	6	72 149 401	OGFRL1	2.43×10^{-10}	$4.51 \times 10^{-2\,d}$	1.08×10^{-8}
rs4965279	15	98 493 374	ADAMTS17	9.38×10^{-9}	5.58×10^{-2}	1.69×10^{-7}
rs7826446	8	114 772 471	CSMD3/TRPS1	8.61×10^{-9}	6.41×10^{-2}	9.12×10^{-7}
rs12977050	19	35 520 535	ZNF536	3.01×10^{-9}	1.32×10^{-1}	4.73×10^{-7}
rs2804737	6	660 196	EXOC2	2.15×10^{-8}	1.85×10^{-1}	3.28×10^{-6}
rs2734394	21	36 936 522	CLDN14	5.74×10^{-9}	2.75×10^{-1}	4.12×10^{-6}
rs7800473	7	15 619 579	MEOX2	3.79×10^{-8}	3.97×10^{-1}	3.34×10^{-5}

AD, Alzheimer's disease; bp, base pair; *pb*GWAS, pooling/bootstrap genome-wide sssociation study; SNP, single-nucleotide polymorphism.
[a]Pooling/bootstrap GWAS strategy.
[b]Performed using Human610-Quad SNP-chips (23 individuals; 13 EOAD (early-onset Alzheimer's Disease), 10 LOAD (late-onset Alzheimer's Disease)) and Human370-Quad SNP-chips (48 individuals; 27 EOAD, 21 LOAD) from Illumina; genotype data was merged using Golden Helix's editing facilities. Individuals were selected using the same bootstrap strategy described before.
[c]Using a modified Stouffer's method with equal weight for all pair of pools as described in the Materials and Methods section.
[d]Disease-associated markers for which the association was replicated in the individual genotyping step.
SNPs were ranked based on the results of our discovery phase.
Note: SNPs located in genes previously reported as being associated with Alzheimer's disease are highlighted in gray.

Table 2. Results of a gene-based association tests (VEGAS) for the combined *P*-values from the discovery phase and individual genotyping

Chromosome	Gene	N	B ($\times 1000$)	Start (bp)	Stop (bp)	Statistic	P-value
1	GREM2	2	1000	238 719 495	238 842 085	65.66	$<10^{-6}$
3	MYLK	8	1000	124 813 832	125 085 839	97.33	$<10^{-6}$
6	HLA-A	2	1000	30 018 309	30 021 633	32.19	$<10^{-6}$
7	CADPS2	7	1000	121 745 713	122 313 790	111.90	$<10^{-6}$
10	KNDC1	7	1000	134 823 960	134 889 906	87.31	$<10^{-6}$
11	CD44	8	1000	35 116 992	35 210 525	115.13	$<10^{-6}$
16	A2BP1	19	1000	6 009 132	7 702 500	101.03	$<10^{-6}$
8	SNTG1	30	1000	50 987 149	51 867 980	430.53	$<10^{-5}$
6	EXOC2	13	1000	430 137	638 109	130.34	$<10^{-5}$
8	CLU	2	1000	27 510 367	27 528 244	35.60	$<10^{-4}$
2	NPHP1	10	1000	110 238 202	110 319 928	177.01	1.2×10^{-4}
3	CLSTN2	5	1000	141 136 896	141 769 322	29.99	0.00027
4	PAICS	9	1000	56 996 671	57 022 291	92.26	0.00061
7	MEOX2	6	1000	15 617 361	15 692 833	31.61	0.00266
1	CR1	1	100	205 736 095	205 881 733	4.12	0.04077

B, total number of simulations; bp, base pair; N, total number of SNPs; SNP, single-nucleotide polymorphism; VEGAS, versatile gene-based association study.
Note: Genes previously reported to be associated with Alzheimer's disease are highlighted in gray.

combined values demonstrate good control of the type I-error probability; SNPs with $P < 10^{-4}$ shown in green (Figure 3, see Supplementary Material online).

Several studies making use of DNA pooling have consistently reported a very high allele frequency estimation reliability and accuracy when it is compared with determination of allele frequencies by individual genotyping.[15,34,35] These results are even more consistent when the pool involves individuals of similar ethnic background (as it is the case of this study)[34] and remain still consistent throughout different platforms (in this study, we use a unique platform).[15,35] In order to evaluate both accuracy and reliability of allele frequencies estimated by DNA pooling in this particular study, we evaluated the correlation between gene frequencies estimated by DNA pooling with those determined by individual genotyping. Thus far, for each one of the eight pools, gene frequencies were determined for the subset of patients that were individually genotyped and making part of the specific pool. Supplementary Table 2 describes the number of DNA samples used in each step of the pbGWAS strategy that also were individually genotyped.

It is worth mentioning that numbers do not add to 100% of the samples subjected to each step of pooling because the subset—that of individuals that were subjected to individual genotyping—was obtained by bootstrap drafting. First, we compared heterozygosity values (calculated as $2pq$) of DNA pools with those of the individual genotypes represented in their respective pool. Using the heterozygosity values from both strategies for cases and controls, we further calculated the linear correlation coefficient and 95% confidence intervals associated with it (see Supplementary Table 3 online). Heterozygosity values were significantly correlated. Supplementary Figure 9 depicts the cumulative distribution function of the heterozygosity values in cases and controls for individual genotyping (red dots) and DNA pooling (black dots), demonstrating full consistency of both distributions.

DISCUSSION

In the recent era of GWAS, the identification of any significant association for complex diseases, including neurodegenerative and psychiatric disorders, meeting the rigorous criteria for statistical significance has been shown to be a very expensive and sometimes futile endeavor. For AD and other conditions, empirical data verified that initial sample size estimations from power calculations based on the assumptions of the common disease–common variant hypothesis (for example, $n \sim 1000$ for cases and controls) were not adequate to detect associations, and this extended to the meta-analyses of samples combined across studies to generate sample sizes of 10,000 cases and 10,000 controls. Now, even larger sample sizes are being used, with 50,000 cases and 50,000 controls, which may be required to detect associations using standard GWAS methods.

Based on the use of large numbers of samples, GWAS have been successful in identifying novel loci that confer susceptibility to AD. However, the costs and logistics involved in reaching such definition, resolution and power are highly demanding as well. Furthermore, because of the implicit clinical and genetic heterogeneity of neurodegenerative and psychiatric conditions, collection of sample sizes of this magnitude is almost impossible.

Here we have described a theoretical method, pbGWAS, and applied it empirically to a homogeneous sample of patients with one of the most severe forms of EOAD. These patients carrying the p.Glu208Ala mutation belong to a unique and clinically well-characterized multigenerational pedigree that segregated AD as consequence of a founder effect.[8] The pbGWAS method combines the established strategy of DNA from exquisitely well-defined homogenous cohorts, and a new a sequential strategy of resampling (bootstrapping) from small samples of cases and controls. These procedures provided increased statistical power in

order to reach the rigorous criteria for statistical significance needed by a GWAS. In contrast to the traditional GWAS approach, we have been able to identify new loci potentially implicated as modifiers of the age of onset of AD using limited and a relatively small sample.

Because DNA samples from cases and controls are repeatedly used to construct the DNA pools and test disease-associated markers, the pbGWAS strategy, as other methods that combine evidence, is more powerful.[36] In fact, by constructing eight pairs of comparable replicated DNA pools using 32 EAOD and 19 LOAD individuals, each time, and combining the information, the resulting effective sample size is 408. This sample size is comparable to 471 that, for a power of 80%, is the one required to detect a moderate odds ratio of 2.5, assuming a minimum allele frequency of 0.3, a Bonferroni genome-wide corrected level of significance of 5%, genotype frequencies in cases, controls, and in the combined group in Hardy Weinberg Equilibrium, and using 1 million of SNPs (http://bioinformatics.centrillionbio.com/sscalc/). At an average genotyping cost of US$300 per sample, the total cost would have been at least US$141,300. Therefore, savings for using our approach add up to US$120,000.

Although it is suggested that DNA pooling must be conducted as an initial screening and further individual genotyping is required for confirmation, especially for the most promising loci,[15] we argue that by genotyping only the top hits we would miss the opportunity to apply GB (as opposed to marker-based) association analysis. An additional practical argument is that prices of genotyping from thousand to million SNPs have very small differences.

Despite the herein demonstrated cost savings and intriguing results accomplished by the pbGWAS strategy in this particular case, there are some intrinsic limitations associated to the SNP-chip type and technology, inherent variability of phenotyping, sample size, and the presence of population stratification. Along with the problems of phenotype outline, population stratification is probably the more important constraint to large-scale association studies, either using individual genotyping[37] or DNA pooling.[35]

DNA pooling strategies are highly sensitive to these effects, as the real influence of genotype stratification (true microdifferentiation) cannot be discriminated from allelic stratification. Although novel statistical methods have been proposed to deal with this problem, including the use of either background markers,[38,39] genomic controls,[40,41] structured association,[42,43] or hundreds of thousands of markers,[37] for this particular type of strategy we strongly recommend the use of methodologies, as the one described by Turakulov and Easteal,[39] that not only define a minimum number of highly selected SNPs but which can also be applied under a sequential strategy.

An additional issue refers to the definition of the sample size, not only for the original stem pools, but also for the size of the bootstrapped pools and number of steps. Though in theory, the higher the size of the stem pool, the higher the potential number of bootstrapped replicated sub-samples (higher variance), in practical terms pools of 96 samples that adjust to standard plates of a broad number of current technological robotic supports, will be of choice. Similarly, we have theoretically and empirically observed that bootstrapped replicates of 50% maximize the information. With regards to the number of sequential steps of replication, we empirically found that once reached the forth step of resampling, quantitative fluctuations are small.

Though the need for replication is clear, we considered that the demonstrated technique, as well as the new findings described here, might provide important insights both into the natural history of AD, as well as a novel methodology that can be applied to a wide variety of conditions. These results will feedback from further replications in cohorts of patients with sporadic forms of AD, other type of dementia, and patients with minimal cognitive impairment. Though it is difficult to warrant that the same group of genes might be commonly causal for this wide spectrum of pathologies, the fact that some usual suspects, that is, CADPS2, are

showing up, indicates that these findings originated from the use of conspicuous and homogenous forms of a pathology might be proficient for broad applications.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

REFERENCES

1 Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010; **363**: 166–176.

2 Cao P, Wang QJ, Zhu XT, Zhou H, Li R, Wang WP. Quantitative determination of allele frequency in pooled DNA by using sequencing method. *J Chromatogr B Analyt Technol Biomed Life Sci* 2011; **879**: 527–532.

3 Manolio TA, Collins R. Enhancing the feasibility of large cohort studies. *JAMA* 2010; **304**: 2290–2291.

4 Sham P, Bader JS, Craig I, O'Donovan M, Owen M. DNA Pooling: a tool for large-scale association studies. *Nat Rev Genet* 2002; **3**: 862–871.

5 Bader JS, Bansal A, Sham P. Efficient SNP-based tests of association for quantitative phenotypes using pooled DNA. *Gene Screen* 2001; **1**: 143–150.

6 Shaffer JP. Multiple hypothesis testing. *Ann Rev Psychol* 1995; **46**: 561–584.

7 Efron B. Bootstrap methods: another look at the Jacknife. *Ann Stat* 1979; **7**: 1–26.

8 Acosta-Baena N, Sepulveda-Falla D, Lopera-Gomez CM, Jaramillo-Elorza MC, Moreno S, Aguirre-Acevedo DC et al. Pre-dementia clinical stages in presenilin 1 E280A familial early-onset Alzheimer's disease: a retrospective cohort study. *Lancet Neurol* 2011; **10**: 213–220.

9 Lopera F, Ardilla A, Martinez A, Madrigal L, Arango-Viana JC, Lemere CA et al. Clinical features of early-onset Alzheimer disease in a large kindred with an E280A presenilin-1 mutation. *JAMA* 1997; **277**: 793–799.

10 Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* 2005; **37**: 549–554.

11 Fan JB, Chee MS, Gunderson KL. Highly parallel genomic assays. *Nat Rev Genet* 2006; **7**: 632–644.

12 Dmitrienko A, Govindarajulu Z. Sequential determination of the number of bootstrap samples. *J Stat Plann Inference* 2002; **100**: 349–363.

13 Guo W, Peddada S. Adaptative choice of the number of samples in large scale multiple testing. *Stat Appl Gen Mol Biol* 2008; **7**: Article 13.

14 Jawaid A, Bader JS, Purcell S, Cherny SS, Sham P. Optimal selection strategies for QTL mapping using pooled DNA samples. *Eur J Hum Genet* 2002; **10**: 125–132.

15 Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res* 1998; **8**: 1273–1288.

16 Le Hellard S, Ballereau SJ, Visscher PM, Torrance HS, Pinson J, Morris SW et al. SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Res* 2002; **30**: e74.

17 Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG. Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann Hum Genet* 2002; **66**(Part 5–6): 393–405.

18 Bonferroni CE. Il calcolo delle assicurazioni su gruppi di teste. *In Studi in Onore del Professore statore Salvatore Ortu Carboni* 1935; 13–60.

19 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B (Methodological)* 1995; **57**: 289–300.

20 Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 2004; **74**: 765–769.

21 Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM et al. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet* 2010; **87**: 139–145.

22 Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003; **100**: 9440–9445.

23 Hedges LV, Cooper H, Bushman BJ. Testing the null hypothesis in meta-analysis: a comparison of combined probability and confidence interval procedures. *Psychol Bull* 1992; **111**: 188–184.

24 Hoop JJ. *Multilevel Analysis: Techniques and Applications,*, 1st edn. Lawrence Erlbaum Associates, 2002.

25 Fisher RA. *Satistical Methods for Research Workers*, 4 edn. Oliver and Boyd: London, 1932.

26 Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RMJ. *Adjustment During Army Life*. Princeton University Press: Princeton, NJ, 1949.

27 Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J Evol Biol* 2005; **18**: 1368–1373.

28 Hartung J. A note on combining dependent tests of significance. *Biometrical J* 1999; **41**: 849–855.

29 R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org, 2011.

30 Stein JL, Hua X, Lee S, Ho AJ, Leow AD, Toga AW et al. Voxelwise genome-wide association study (vGWAS). *Neuroimage* 2010; **53**: 1160–1174.

31 Uberti D, Cenini G, Bonini SA, Barcikowska M, Styczynska M, Szybinska A et al. Increased CD44 gene expression in lymphocytes derived from Alzheimer disease patients. *Neurodegener Dis* 2010; **7**: 143–147.

32 Akiyama H, Tooyama I, Kawamata T, Ikeda K, McGeer PL. Morphological diversities of CD44 positive astrocytes in the cerebral cortex of normal subjects and patients with Alzheimer's disease. *Brain Res* 1993; **632**: 249–259.

33 Maragakis NJ, Rothstein JD. Mechanisms of disease: astrocytes in neurodegenerative disease. *Nat Clin Pract Neurol* 2006; **2**: 679–689.

34 Anantharaman R, Chew FT. Validation of pooled genotyping on the Affymetrix 500 k and SNP6.0 genotyping platforms using the polynomial-based probe-specific correction. *BMC Genet* 2009; **10**: 82.

35 Sham P, Bader JS, Craig I, O'Donovan M, Owen M. DNA Pooling: a tool for large-scale association studies. *Nat Rev Genet* 2002; **3**: 862–871.

36 Westfall PH. Combining P-values. *Encyclopedia of Biostatistics* 2005; 1–5.

37 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904–909.

38 Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 2002; **12**: 1805–1814.

39 Turakulov R, Easteal S. Number of SNPS loci needed to detect population structure. *Hum Hered* 2003; **55**: 37–45.

40 Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.

41 Devlin B, Bacanu SA, Roeder K. Genomic control to the extreme. *Nat Genet* 2004; **36**: 1129–1130; author reply 1131.

42 Setakis E, Stirnadel H, Balding DJ. Logistic regression protects against population structure in genetic association studies. *Genome Res* 2006; **16**: 290–296.

43 Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 2001; **68**: 466–477.

44 Benzing T, Gerke P, Hopker K, Hildebrandt F, Kim E, Walz G. Nephrocystin interacts with Pyk2, p130(Cas), and tensin and triggers phosphorylation of Pyk2. *Proc Natl Acad Sci USA* 2001; **98**: 9784–9789.

45 Donaldson JC, Dempsey PJ, Reddy S, Bouton AH, Coffey RJ, Hanks SK. Crk-associated substrate p130(Cas) interacts with nephrocystin and both proteins localize to cell-cell contacts of polarized epithelial cells. *Exp Cell Res* 2000; **256**: 168–178.

46 Sawada Y, Tamada M, Dubin-Thaler BJ, Cherniavskaya O, Sakai R, Tanaka S et al. Force sensing by mechanical extension of the Src family kinase substrate p130Cas. *Cell* 2006; **127**: 1015–1026.

Supplementary Information accompanies the paper on the Molecular Psychiatry website (http://www.nature.com/mp)