Molecular Psychiatry (2008), 1–17

www.nature.com/mp

IMMEDIATE COMMUNICATION

Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo

PF Sullivan[1], EJC de Geus[2], G Willemsen[2], MR James[3], JH Smit[4], T Zandbelt[4], V Arolt[5], BT Baune[6], D Blackwood[7], S Cichon[8], WL Coventry[9], K Domschke[5], A Farmer[10], M Fava[11], SD Gordon[3], Q He[1], AC Heath[12], P Heutink[4], F Holsboer[13], WJ Hoogendijk[4], JJ Hottenga[2], Y Hu[1], M Kohli[13], D Lin[1], S Lucae[13], DJ MacIntyre[14], W Maier[8], KA McGhee[7], P McGuffin[10], GW Montgomery[3], WJ Muir[7], WA Nolen[15], MM Nöthen[8], RH Perlis[11], K Pirlo[10], D Posthuma[2], M Rietschel[16], P Rizzu[4], A Schosser[10], AB Smit[2], JW Smoller[11], J-Y Tzeng[17], R van Dyck[4], M Verhage[2], FG Zitman[18], NG Martin[3], NR Wray[3], DI Boomsma[2,19] and BWJH Penninx[4,19]

[1]*Department of Genetics, University of North Carolina, Chapel Hill, NC, USA;* [2]*VU University Amsterdam, Amsterdam, The Netherlands;* [3]*Queensland Institute for Medical Research, Brisbane, QLD, Australia;* [4]*VU University Medical Center Amsterdam, Amsterdam, The Netherlands;* [5]*University of Münster, Münster, Germany;* [6]*James Cook University, Cairns, QLD, Australia;* [7]*University of Edinburgh, Edinburgh, UK;* [8]*University of Bonn, Bonn, Germany;* [9]*University of New England, Armidale, NSW, Australia;* [10]*Institute of Psychiatry, London, UK;* [11]*Harvard Medical School, Cambridge, MA, USA;* [12]*Washington University, St. Louis, MO, USA;* [13]*Max-Planck Institute of Psychiatry, Munich, Germany;* [14]*Royal Edinburgh Hospital, Edinburgh, UK;* [15]*University Medical Center Groningen, Groningen, The Netherlands;* [16]*University of Heidelberg, Heidelberg, Germany;* [17]*North Carolina State University, Raleigh, NC, USA and* [18]*Leiden University Medical Center, Leiden, The Netherlands*

Major depressive disorder (MDD) is a common complex trait with enormous public health significance. As part of the Genetic Association Information Network initiative of the US Foundation for the National Institutes of Health, we conducted a genome-wide association study of 435 291 single nucleotide polymorphisms (SNPs) genotyped in 1738 MDD cases and 1802 controls selected to be at low liability for MDD. Of the top 200, 11 signals localized to a 167 kb region overlapping the gene piccolo (*PCLO*, whose protein product localizes to the cytomatrix of the presynaptic active zone and is important in monoaminergic neurotransmission in the brain) with *P*-values of 7.7×10^{-7} for rs2715148 and 1.2×10^{-6} for rs2522833. We undertook replication of SNPs in this region in five independent samples (6079 MDD independent cases and 5893 controls) but no SNP exceeded the replication significance threshold when all replication samples were analyzed together. However, there was heterogeneity in the replication samples, and secondary analysis of the original sample with the sample of greatest similarity yielded $P = 6.4 \times 10^{-8}$ for the nonsynonymous SNP rs2522833 that gives rise to a serine to alanine substitution near a C2 calcium-binding domain of the PCLO protein. With the integrated replication effort, we present a specific hypothesis for further studies.
Molecular Psychiatry advance online publication, 9 December 2008; doi:10.1038/mp.2008.125

Keywords: major depressive disorder; genome-wide association; Netherlands study of depression and anxiety; Netherlands twin registry

Introduction

The defining features of major depressive disorder (MDD) are marked and persistent dysphoria plus additional cognitive signs and symptoms (anhedonia, sleep disturbance, weight/appetite changes, motor agitation/retardation, anergia, excessive guilt or worthlessness, poor concentration or indecisiveness, and recurrent thoughts of death or suicide).[1] MDD is distinct from normal sadness by its persistence (that is, $\geqslant$2 weeks), additional signs and symptoms, and substantial associated impairment. The definition of MDD excludes other conditions typified by substantial depressive symptoms (other psychiatric disorders, drug/alcohol dependence and somatic diseases). The lifetime prevalence of MDD is $\sim$15%[2–4] and is twofold higher in women[5] with a course typified by recurrence of illness.[6] It is associated with considerable morbidity,[7–9] excess mortality from suicide and

Correspondence: Dr PF Sullivan, Department of Genetics, University of North Carolina, CB No. 7264, 4109D Neurosciences Research Building, Chapel Hill, NC 27599-7264, USA.
E-mail: pfsulliv@med.unc.edu
[19]These authors contributed equally to this work.
Received 16 July 2008; revised 19 September 2008; accepted 21 October 2008

other causes,[10–13] and substantial direct and indirect costs.[14] A World Health Organization study projected MDD to be the second leading cause of disability worldwide by 2020.[15]

Although there is a considerable corpus of research on the epidemiology and biological correlates of MDD, little is known for certain about its etiology. An important etiological clue may be the familial tendency of MDD and its heritability of 31–42%.[16] This clue led to a number of genome-wide linkage studies (Supplementary Methods) and studies of >100 theoretical or positional candidate genes. As for the use of these study designs with other biomedical disorders, their application to MDD has not been as successful as had been hoped.

It is now clear that genome-wide association studies (GWASs) can be a successful tool in the genetic dissection of complex biomedical disorders.[17,18] The goal of this report is to describe a GWAS for MDD that was systematically designed to remediate a set of methodological issues common to genetic studies of MDD. Examples of these issues include small sample sizes, inhomogeneous samples in terms of ancestry and phenotyping, convenience sampling, and controls that are unaffected but not at low liability for MDD. Moreover, large-scale replication was integral to our design.

Materials and methods

This GWAS was one of the six initial Genetic Association Information Network (GAIN) studies sponsored by the Foundation for the NIH.[19] Individual phenotype and genotype data are available to researchers by application to the dbGaP repository.[20] We have attempted to follow published guidelines for GWAS (Chanock *et al.*,[21] Box 1).

Subjects

The parent projects that supplied subjects for this GWASs are longitudinal studies, the Netherlands Study of Depression and Anxiety (NESDA; http://www.nesda.nl)[22] and the Netherlands Twin Registry (NTR; http://www.tweelingenregister.org).[23] Sampling and data collection characteristics of the GAIN–MDD study have been described in detail elsewhere.[24]

MDD cases were mainly from NESDA, a longitudinal cohort study designed to be representative of individuals with depressive and/or anxiety disorders. Recruitment of participants for NESDA took place from 09/2004–02/2007, and ascertainment was from outpatient specialist mental health facilities and by primary care screening. Additional cases were from the population-based cohorts NEMESIS,[25] ARIADNE[26] and the NTR. Regardless of recruitment setting, similar inclusion and exclusion criteria were used to select MDD cases. Inclusion criteria were a lifetime diagnosis of DSM-IV MDD[1] as diagnosed by the Composite International Diagnostic Interview psychiatric interview,[27] age 18–65 years, and self-reported western European ancestry. Persons who were not fluent in Dutch and those with a primary diagnosis of schizophrenia or schizoaffective disorder, obsessive–compulsive disorder, bipolar disorder or severe substance use dependence were excluded (the etiology of MDD in these subjects may be distinct). The 1862 cases included in GAIN were recruited from mental health care organizations ($N=785$), primary care ($N=603$) and community samples (NEMESIS $N=218$, ARIADNE $N=96$ and NTR $N=160$).

Control subjects were mainly from the NTR, which has collected longitudinal data from twins and their families since 1991 (total cohort of ~22 000 participants from 5546 families). The majority of families were recruited when the twins were adolescents or young adults through city council registrations along with alternative efforts to recruit older twins. Longitudinal phenotyping includes assessment of depressive symptoms (via multiple instruments), anxiety, neuroticism and other personality measures. Inclusion required availability of both survey data and biological samples, no report of MDD at any measurement occasion, and low genetic liability for MDD. No report of MDD was determined by specific queries about medication use or whether the subject had ever sought treatment for depression symptoms and/or through the CIDI interview. Low genetic liability for MDD was determined by the use of a factor score derived from longitudinal measures of neuroticism, anxiety and depressive symptoms[28] (mean 0, s.d. 0.7); controls were required never to have scored highly ($\geqslant 0.65$) on this factor score. Finally, controls and their parents were required to have been born in the Netherlands or western Europe. Only one control per family was selected. There were controls ($N=1703$) from the NTR and additional controls from NESDA ($N=133$ from general practice, $N=24$ from ARIADNE). NESDA controls had no lifetime diagnosis of MDD or an anxiety disorder as assessed by the CIDI and reported low depressive symptoms at baseline (K-10 score <16 and inventory of depressive symptoms score <4).[29,30]

Case–control matching

If there were multiple eligible NTR controls in a family, we first matched on sex and age, and used the highest number of completed questionnaires as an additional criterion. Again, only one control per family was included.

DNA sampling

Before the start of the NESDA and NTR biological sample collection, processing, and storage protocols were harmonized and DNA extraction was conducted concurrently in the same laboratory. For NESDA, blood sampling for the NESDA participants took place during the baseline visit (between 0830 and 0930 hours) and DNA was isolated using the FlexiGene DNA AGF3000 kit (Qiagen, Valencia, CA, USA) on an AutoGenFlex 3000 workstation (Autogen,

Holliston, MA, USA). For NTR, biological samples were taken in the subject's home (between 0700 and 1000 hours) and DNA was extracted using the Puregene DNA isolation kit (Gentra, Minneapolis, MN, USA) for frozen whole blood samples. DNA concentrations were determined using the PicoGreen dsDNA Quantitation kit (Invitrogen Corporation, Carlsbad, CA, USA). All procedures were performed according to the manufacturer's protocols.

Ethical issues

The NESDA and NTR studies were approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Center, Amsterdam, an Institutional Review Board certified by the US Office of Human Research Protections (IRB number IRB-2991 under Federal-wide Assurance-3703; IRB/institute codes, NESDA 03-183; NTR 03-180). All subjects provided written informed consent. As part of the GAIN application process, consent forms were specifically rereviewed for suitability for the deposit of deidentified phenotype and genotype data into the controlled-access dbGaP repository.[20] NESDA and NTR subjects were informed of participation in GAIN by newsletters. Only 22 NESDA respondents refused informed consent for genetic research (1.7% of all respondents approached).

GWAS genotyping

Individual genotyping was conducted by Perlegen Sciences (Mountain View, CA, USA) using a set of four proprietary, high-density oligonucleotide arrays. The SNPs on these arrays were selected to tag common variation in the HapMap European and Asian panels using previously described genotype data,[31] tagging approach[32] and methodology.[33] At the beginning of GAIN, all HapMap[34] samples were genotyped with the Perlegen GWAS platform. Independent review of these data by the GAIN analysis group[19] showed 99.8% agreement with prior HapMap genotypes and the mean maximum r^2 between the Perlegen SNPs and HapMap phase II SNPs[31] was 0.89 for single and 0.96 for multimarker analyses. The genotyping procedures and genotyping calling algorithms are described in the Supplementary Methods and in prior reports.[35,36] Briefly, 40×96-well plates were sent to Perlegen for GWAS genotyping. Genotyping was conducted blind to case–control status. Cases and controls were randomly allocated to plates and to positions within plates. Each plate contained DNA samples from 93 Dutch subjects plus 3 quality control samples. The three quality control samples included: two parents of one control on that plate (40 complete trios in total); and half the plates contained the same HapMap CEU sample (used for quality control in all GAIN projects) and half had a randomly selected duplicate case sample. The total number of samples was 3840 ($=40$ plates $\times$ 96 samples per plate) or 1860 cases + 1860 controls + 80 parents + 20 duplicate samples + 20 HapMap samples.

Quality control—subjects

Of the 3820 Dutch samples sent to Perlegen (excluding the 20 HapMap internal control samples), genotypes were delivered for 3761 samples. A total of 59 samples did not have GWAS data: 39 samples with uncertain linkage between genotype and phenotype records, 7 samples with evidence of contamination, 6 samples that failed genotyping and 7 miscellaneous failures (2 of these were excluded as chrX and chrY genotyping data were consistent with the presence of XO and XXY sex chromosome status). After further analysis, 8 subjects were removed for excessive missing genotype data ($>25\%$), 1 case for high genome-wide homozygosity ($\sim 75\%$), 38 subjects whose genome-wide IBS estimates were consistent with first- or second-degree relationships and 57 additional subjects whose ancestry diverged from the remainder of the sample (see Supplementary Methods for details). After these exclusions ($N=104$) and removing duplicated and trio quality control samples, there were 3540 subjects in the final analysis data set including 1738 cases and 1802 controls. The principal reason for fewer cases than controls was the higher prevalence of substantial non-European ancestry. The list of subjects in the final analyses data set is included as a Supplementary File ('mddC.fam').

Quality control—SNPs

The unfiltered data set obtained from dbGaP contained 599 156 unique SNPs. The Perlegen genotyping algorithm yielded a quality score for each individual genotype, and a more stringent quality score cutoff ($\geqslant 10$) than that used by Perlegen was applied. The SNP quality control process is described in detail in the Supplementary Methods. Briefly, to be included in the final analysis data set, SNPs were required not to have any of the following features: gross mapping problem,[37] $\geqslant 2$ genotype disagreements in 40 duplicated samples, $\geqslant 2$ Mendelian inheritance errors in 38 complete trio samples, minor allele frequency <0.01 or >0.05 missing genotypes in either cases or controls. A Hardy–Weinberg filter was not used as lack of fit to Hardy–Weinberg expectations can occur for valid reasons (for example, a true association)[38] and given that 95.6% ($=51\,592/53\,994$) of SNPs with $P<0.00001$ from an exact test of Hardy–Weinberg equilibrium[39] in controls were already flagged for exclusion. A total of 435 291 SNPs met these criteria and were included in the final analysis data set (included as a Supplementary File, 'mddC.bim'). Additional quality control checks are described in the Supplementary Methods). A total of 13 controls were genotyped in a different study using the Illumina 317K platform and, of the 82 636 SNPs common to both platforms, the genotype agreement was 99.94%.

Single-marker statistical analyses

There were three classes of SNPs—those that could be heterozygous in all subjects (chr1-22 and chrX/PAR1), those that were heterozygous in women (non-PAR

chrX) and those that were hemizygous in men (non-PAR chrX and chrY). All SNPs that passed quality control checks were tested for association with MDD using 1 d.f. Cochran-Armitage trend tests. For complex traits, it is widely believed that the contributions of individual SNPs to disease risk are often roughly additive.[40] The Cochran-Armitage trend test can be used to detect such effects. This test is usually recommended due to its robustness to the violation of the HWE assumption:[41] *P*-values from women and men for non-PAR chrX were combined using Fisher's method.[42]

Population stratification artifacts were assessed in two ways. As described elsewhere,[36] including principal components as covariates in a logistic regression model can robustly control stratification effects. To do this, we identified a set of 127 688 SNPs in linkage equilibrium[43] and used the 'smartpca' program in EigenSoft[44] to compute 10 principal components for each subject that were included as covariates in logistic regression models (case/control status $\sim$ SNP$_i$ + PC1 + PC2 + … + PC10). We also used a stratified Cochran–Mantel–Haenszel test in PLINK[43] as a complementary approach.

For noteworthy associations, there were additional checks to ensure that an association was not due to experimental bias. These checks included: manual inspection of SNP cluster plots to ensure reasonable performance of the genotyping calling algorithm; evaluation of conformation to Hardy–Weinberg equilibrium in controls, cases and overall (discussed in the Supplementary Methods); the checks for population stratification described above; evaluation of plate-specific association results to ensure that the overall association was not driven by one or a few plates; comparison of control MAFs to the HapMap EUR panel; and evaluation of the characteristics of a SNP in high linkage disequilibrium ('proxy association') as a similar association with such a SNP decreases the chance of some forms of method artifacts.

Control of false discoveries

Given the 10^5–10^7 statistical comparisons in a GWAS, small *P*-values are expected by chance. To control the risk of false discoveries, *q*-values[45,46] were computed for all *P*-values for single-marker tests of association. A *q*-value is an estimate of the proportion of false discoveries among all significant markers, or the false discovery rate (FDR) for the corresponding *P*-value. The use of *q*-values is preferable to more traditional multiple testing controls because *q*-values provide a better balance between the competing goals of finding true positives versus controlling false discoveries, allow more similar comparisons across studies because proportions of false discoveries are much less dependent on the number of tests conducted and are relatively robust against the effects of correlated tests.[45,47–54] The *q*-value threshold for declaring significance was 0.10 (that is, the top 10% of the significant findings are, on average, allowed to be false discoveries).[50,55] FDR thresholds <0.10 result in a disproportionate drop in power to detect true effects.

Imputation

We used two imputation approaches, the SNPMStat method of Lin *et al*.[56] to impute 246 additional SNPs in the piccolo (*PCLO*) region and Abecasis' MACH (v1) to impute 2 037 829 autosomal SNPs with $R^2 \geqslant 0.5$ (a cutoff that removes $\sim$90% of SNPs with unreliable imputation results while dropping 2–3% of reliably imputed SNPs). Both SNPMStat and MACH gave similar results in the *PCLO* region. Imputed genotypes were used in secondary analyses. The HapMap2 EUR panel[31,34] was used as reference.

Statistical power

Quanto[57,58] was used to approximate statistical power given the following assumptions: two-tailed $\alpha = 1 \times 10^{-7}$ ($=0.05/500\,000$), 1738 cases and 1802 controls, lifetime morbid risk of MDD of 0.15 and a log additive genetic model. For statistical power of 0.80 ($\beta = 0.20$), the minimum detectable genotypic relative risks are 1.59, 1.40 and 1.35 for minor allele frequencies of 0.10, 0.25 and 0.40.

Software

PLINK (v1.0),[43] SAS (v9.1.3),[59] R (v2.6.1),[60] HAPSTAT (v3),[61–63] MACH1, SNPMStat,[56] HaploView,[64] and JMP (v6)[65] were used for data management, quality control, statistical analyses and graphics.

Bioinformatics

All genomic locations are per NCBI Build 35 [66] (UCSC hg17).[67] Pseudoautosomal region 1 (PAR1) is assumed to be located on chrX:1–2 692 881 and chrY:1–2 692 881 and PAR2 on chrX:154 494 747–154 824 264 and chrY:57 372 174–57 701 691.[68] SNP annotations were per TAMAL[37] based chiefly on UCSC genome browser files,[67] HapMap[34] and dbSNP.[66]

Results

Sample description

Table 1 presents descriptive data for cases and controls. Controls had a higher proportion of men and were slightly older (and thus were farther through the period of risk for MDD). Consistent with known correlates of MDD, cases had a significantly lower educational level, less often had a partner, were more often smokers and scored much higher on the NEO-FFI neuroticism scale.

SNP description

The analysis SNP set had 435 291 SNPs including 427 049 autosomal SNPs, 7 988 SNPs on the non-PAR portions of chrX, 239 SNPs on chrXY/PAR1, 15 SNPs on chrY and 0 SNPs on PAR2. The median SNP missingness was 0.00339 (interquartile range 0.00113–0.0105) and the median minor allele frequency was 0.2422 (interquartile range 0.1375–0.3646) with similar estimates in cases and controls. The average marker density over the genome

Table 1 Descriptive data for cases with MDD and controls at low liability for MDD included in the GWAS

Descriptor	*Cases*	*Controls*	*Test*
Number of subjects genotyped	1738	1802	—
Mean age in years (s.d.)	42.6 (12.6)	45.1 (14.1)	$_{1,3538}F = 31.1$, $P < 0.001$
Female (%)	69.6	62.0	$\chi^2_1 = 22.5$, $P < 0.001$
Educational level (%, low/middle/high)	7.8/62.0/32.2	5.7/56.3/38.1	$\chi^2_1 = 16.3$, $P < 0.001$
Partner status (% with partner)	68.9	87.0	$\chi^2_1 = 167.2$, $P < 0.001$
Smoking (current) (%)	42.0	20.2	$\chi^2_1 = 194.5$, $P < 0.001$
Mean neuroticism (NEO, s.d.)	39.3 (8.0)	28.2 (5.5)	$_{1,2920}F = 1831$, $P < 0.001$
MDD, age of onset in years (s.d.) early age of onset (<30 years) (%)	27.7 (12.4) 57.3	—	
Family history of depression (%)	85.5	—	
Recurrent MDD	50.9	—	
Family history, recurrent MDD or early age of onset (<30 years)	94.8	—	

Abbreviation: MDD, major depressive disorder.

was 1 SNP every 7069 bases (=3 077 088 087 bases/435 291 SNPs). The median intermarker distance was 2911 bases with interquartile range 966–7374 bases and a 99th percentile of 50.1 kb.

Single-marker association tests

We used the Cochran-Armitage trend test to test for association of the 435 291 SNPs in the GWAS data set with case/control status. The estimated λ[69] was 1.046 (similar *P*-value minima and λs were obtained using logistic regression with 10 principal components and using a stratified Cochran–Mantel–Haenszel tests based on identity-by-state clusters).[43,44] The minimum *q*-value was 0.28 (that is, if these tests were called significant, over the long term, a minimum false discovery rate of ~28% would be incurred). As the prespecified *q*-value threshold was 0.10, no SNP reached genome-wide significance. The proportion of all SNPs without true effects (P_0)[54] was conservatively estimated to be $P0 = 0.9999954$, consistent with the presence of ~2 SNPs with true effects in these GWAS data.

Figure 1a depicts the quantile–quantile plots[40] for these analyses. The observed *P*-values do not strongly depart from the *P*-value distribution expected by chance. Figure 1b shows a plot of –log10(*P*trend) by genomic location.

Table 2 presents the findings for the top 25 SNPs. The quality control metrics—SNP missingness, agreement with HWE and similarity of the control MAFs to the HapMap EUR panel—for the top 25 SNPs are generally acceptable. Of the top 25, 4 associations are in the presynaptic cytomatrix protein *PCLO*. Table 3 depicts the top 25 multi-SNP clusters (that is, for an index SNP with association $P < 0.001$, these clusters are additional SNPs within 250 kb of the index SNP with $r^2 \geqslant 0.50$). The full version of this table is included as a Supplementary File ('Table 3_full.xls'). *PCLO* is present in the top 25 clusters along with two additional multi-SNP clusters in the top 200. Other notable SNP clusters occurred in *GRM7* (rank 51), *DGKH* (rank 83, a candidate gene for bipolar disorder),[70] *DAOA* (rank 124) and *DRD2* (rank 226).

Focusing on piccolo

Although no association met genome-wide significance, there were clusters of SNPs in *PCLO* (Figure 2). Notably, 11 of the 200 smallest *P*-values localized to a 167 kb segment overlapping *PCLO*. Interest in *PCLO* was increased given its expression in brain, localization to the presynaptic active zone[71] and involvement in monoamine neurotransmission, a venerable hypothesis of the etiology of MDD.[72] Moreover, the third most significant SNP (rs2522833) codes for a nonsynonymous amino-acid change (ala-4814-ser) in *PCLO* near its C2A calcium binding domain.[73]

We investigated possible causes of spurious associations in the *PCLO* region (chr7:82 032 093–82 436 848). First, these findings were not due to plate effects as inspection of plate-specific association data for these SNPs did not show any marked outliers or systematic biases. Second, review of allelic intensity cluster plots on which genotype calls were based revealed adequate performance of the Perlegen genotype calling algorithm. Third, inspection of additional quality control metrics did not suggest systematic problems with SNPs in this region. Fourth, inspection of LD matrices excluded very high LD as the sole explanation for the results (Supplementary Figure 10), and none of the genotyped SNPs had strong LD ($r^2 \geqslant 0.8$) with rs2715148 (the SNP with the smallest *P*-value in the *PCLO* region). Fifth, population stratification can cause false-positive findings but this did not appear to explain the *PCLO* association: (1) the same 11 SNPs had *P*-values among the top 200 associations in unadjusted analyses as well as with adjustment via principal components and stratified analyses; and (b) for the 57 SNPs in the *PCLO* region, the *P*-values across these three types of analyses were consistent (the Spearman's correlations between *P*-values from trend tests, logistic regression and stratified analyses were all >0.962). Sixth, the minor allele frequencies in the control group in the *PCLO* region were usually quite similar to available EUR control groups suggesting that the *PCLO* findings were not due to an artifact of the control selection process (see below). Finally, bioinformatic investiga-

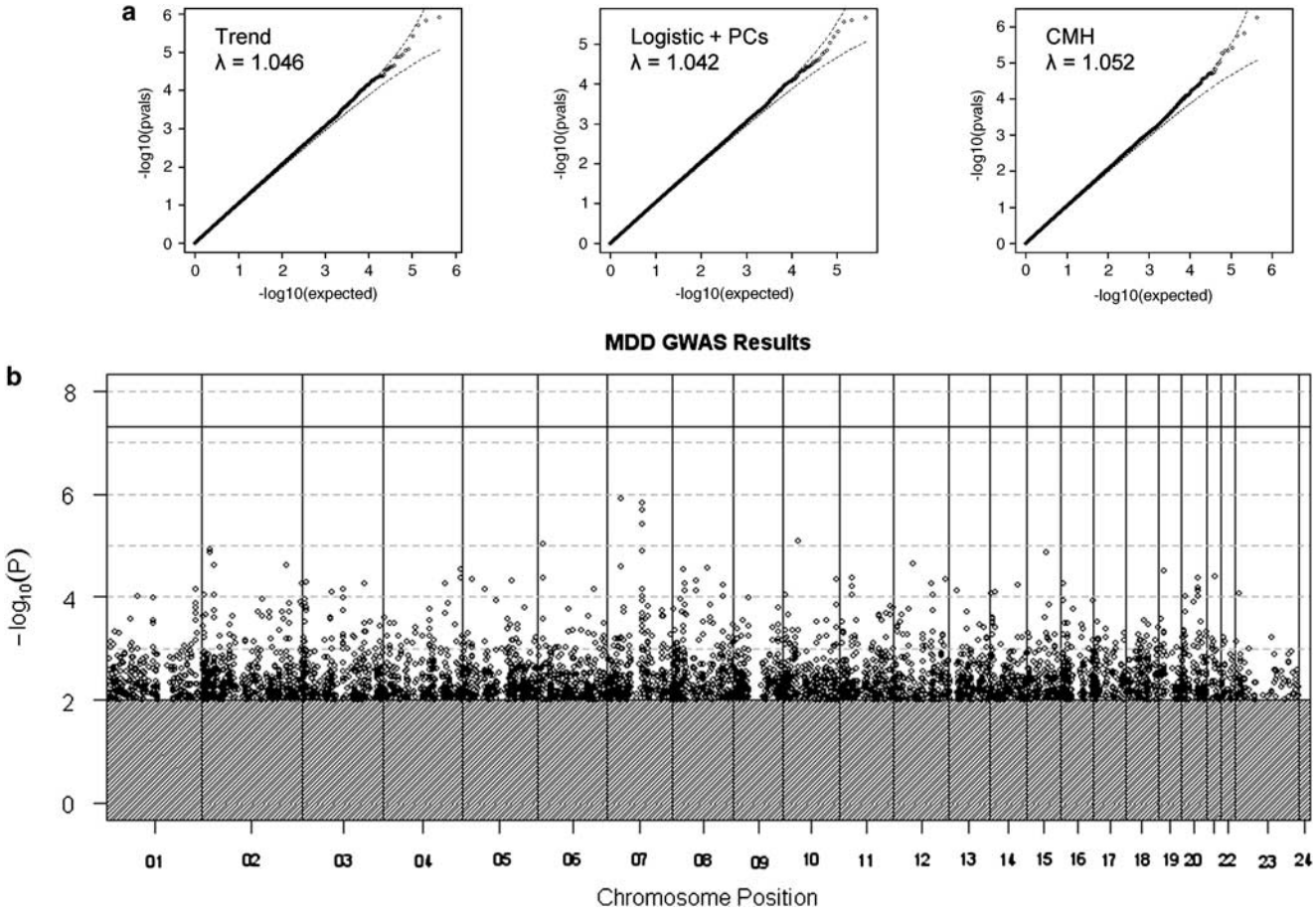

Figure 1 Genome-wide association study (GWAS) results for major depressive disorder (MDD) in cases versus controls. (**a**) Quantile–quantile plots and λ estimates for the primary analysis using the Cochran–Armitage trend test and confirmatory analyses using logistic regressions and Cochran–Mantel–Haenszel stratified tests. The dashed lines show the expected 95% probability interval for ordered *P*-values, and the circles show the observed versus expected values for all SNPs. The λ values are the median χ^2 from all association tests divided by the expected value under the null hypothesis of no association. If λ is large (for example, >1.2), there is evidence that the observed test statistics deviate from the expected. This could be due to true associations but is more likely due to a systematic bias (for example, population stratification effects). The λ values in (**a**) are not consistent with the presence of systematic biases in the results. (**b**) $-\log_{10}(P)$ by genomic location for chr1–chr22 plus chrX.

tion did not suggest that this is a problematic region to genotype as the *PCLO* region is not known to be under positive selection in humans,[74] to contain segmental duplications[67] or common copy number variants (search of the Database of Genomic Variants yielded two rare copy number variations (CNVs) with control frequencies of 0.12 and 0.89%).[75–77]

We conducted additional analyses to attempt to localize the association depicted in Figure 2. Imputation[56] supported the directly typed SNP associations but did not yield an association *P*-value markedly more significant than any directly genotyped SNP (although 22 of the 25 most significant imputed associations in the genome were in this region). Haplotype analysis using three-SNP sliding windows did not improve localization. Secondary analyses by sex, case ascertainment setting and recurrent early onset MDD (reoMDD, arguably the most heritable form of MDD)[16,78] suggested that most of the signals were from women and from subjects with reoMDD (Supplementary Table 11). The findings for reoMDD were often stronger than the primary analyses, particularly for the most significant SNP (rs2715148) where the *P*-value decreased by 1.2 orders of magnitude to 9.5×10^{-8}.

PCLO *replication*

Although no finding met genome-wide significance, the presence of multiple possible signals in *PCLO* and the plausibility of a function for *PCLO* in the etiology of MDD led us to attempt replication in external samples. We assembled a collection of 11 972 independent subjects (6079 MDD cases and 5893 controls) from seven different groups and a total of six case–control replication samples (two German samples were combined; Supplementary Methods). As with

Table 2 Information on the SNPs with the smallest association *P*-values in the GWAS

Basic SNP data					*Bioinformatics*			*Results*						*MAF*				*Quality control—SNP missingness*				*Quality control—additional checks*
SNP	*Chr*	*Position*	*Alleles*	*Strand*	*Gene*	*TAMAL*[a]	*SLEP*[b]	*Rank*	*OR (CI)*	P-*asymptotic*	P-*empirical*	q-*value*	P-*gwemp*	*All*	*Cases*	*Controls*	*HapMap_EUR*	*All*	*Cases*	*Controls*	P-*missing*	
rs12471796	2	20 177 820	A/G	+				10	1.26 (1.14–1.39)	0.000014	0.000014	0.58	0.99	0.298	0.322	0.275	0.271	0.012	0.010	0.014	0.36	
rs7565124	2	20 183 313	A/G	+		Reg pot		7	1.26 (1.14–1.40)	0.000012	0.000011	0.58	0.98	0.296	0.321	0.272	0.271	0.030	0.034	0.026	0.20	
rs3923028	2	29 597 247	T/C	−	*ALK*	CNV	CNV, mutated in colon CA	12	1.34 (1.17–1.54)	0.000024	0.000020	0.66	1.00	0.135	0.153	0.119	0.175	0.001	0.002	0.000	0.06	
rs12621441	2	201 794 446	A/G	+			Near CNV	13	1.31 (1.16–1.49)	0.000024	0.000027	0.66	1.00	0.166	0.185	0.147	0.150	0.008	0.008	0.009	0.86	
rs11132168	4	184 428 336	T/C	+			MDD linkage peak (8.6 Mb)	16	0.75 (0.65–0.86)	0.000029	0.000035	0.66	1.00	0.133	0.116	0.150	0.117	0.001	0.001	0.002	0.63	
rs17074631	4	184 652 456	G/A	+			MDD linkage peak (8.3 Mb)	23	0.75 (0.66–0.86)	0.000043	0.000040	0.66	1.00	0.137	0.120	0.154	0.076	0.003	0.005	0.002	0.26	
rs2094923	6	14 397 061	T/G	−			SCZ linkage meta-analysis (2.5 Mb)	20	0.82 (0.74–0.90)	0.000042	0.000046	0.66	1.00	0.417	0.393	0.441	0.475	0.001	0.001	0.002	1.00	
rs2274822	6	14 399 068	C/T	−			SCZ linkage meta-analysis (2.5 Mb)	6	0.79 (0.71–0.88)	0.000009	0.000007	0.58	0.96	0.268	0.245	0.291	0.283	0.003	0.002	0.003	1.00	
rs1558477	7	30 928 587	C/T	+			MDD linkage peak (3.0 Mb)	1	1.27 (1.16–1.40)	0.000001	0.000002	0.28	0.37	0.430	0.460	0.401	0.442	0.003	0.003	0.004	0.77	HWD cases
rs7791986	7	30 930 719	G/C	+			MDD linkage peak (3.0 Mb)	14	1.22 (1.12–1.35)	0.000026	0.000038	0.66	1.00	0.451	0.477	0.427	0.425	0.001	0.001	0.002	0.38	
rs2715148	7	82 094 686	A/C	+	*PCLO*	Cons, reg pot	BIP GWAS rs2715148 (P=0.03)	2	0.79 (0.72–0.87)	0.000001	0.000003	0.28	0.42	0.482	0.452	0.510	0.525	0.002	0.002	0.002	0.72	
rs2522833	7	82 098 359	C/A	+	*PCLO*	Cons, reg pot, cSNP	BIP GWAS rs7781142 (P=0.03)	3	1.26 (1.15–1.39)	0.000002	0.000002	0.28	0.52	0.455	0.485	0.427	0.425	0.002	0.000	0.003	0.03	
rs2522840	7	82 123 066	G/T	+	*PCLO*	Cons, reg pot	BIP GWAS rs7799260 (P=0.04)	4	1.25 (1.14–1.38)	0.000004	0.000003	0.40	0.74	0.456	0.484	0.428	0.425	0.004	0.002	0.006	0.18	
rs2107828	7	82 200 320	A/T	+	*PCLO*	Reg pot		8	0.81 (0.74–0.89)	0.000013	0.000007	0.58	0.99	0.460	0.433	0.486	0.500	0.037	0.036	0.038	0.79	
rs1457266	8	24 825 757	A/G	−		Reg pot	MDD linkage peak (7.4 Mb)	17	0.81 (0.73–0.89)	0.000029	0.000034	0.66	1.00	0.319	0.295	0.342	0.300	0.002	0.002	0.001	0.44	
rs7005189	8	81 663 211	T/C	+		Cons, reg pot	BIP GWAS rs11778905 (P=0.03, 9.9 kb)	15	0.76 (0.66–0.86)	0.000028	0.000036	0.66	1.00	0.153	0.134	0.170	0.150	0.001	0.000	0.002	0.25	
rs1780436	10	34 297 618	A/G	−				5	0.80 (0.73–0.88)	0.000008	0.000013	0.58	0.95	0.374	0.348	0.400	0.325	0.018	0.016	0.021	0.31	
rs11031676	11	32 242 721	T/C	+		Reg pot		21	1.26 (1.13–1.40)	0.000043	0.000035	0.66	1.00	0.232	0.253	0.212	0.139	0.006	0.007	0.004	0.28	
rs12579771	12	44 019 689	T/C	+	*TMEM16F*	Cons, reg pot	MDD linkage peak (8.8 Mb)	11	0.78 (0.69–0.87)	0.000022	0.000023	0.66	1.00	0.205	0.184	0.225	0.271	0.000	0.000	0.001	1.00	
rs4765078	12	123 171 707	C/T	+				25	0.82 (0.74–0.90)	0.000044	0.000035	0.66	1.00	0.374	0.350	0.397	0.408	0.004	0.003	0.004	1.00	
rs8023445	15	46 980 083	C/T	+	*SHC4*	Reg pot		9	0.72 (0.62–0.84)	0.000014	0.000009	0.58	0.99	0.119	0.101	0.135	0.108	0.011	0.012	0.010	0.63	
rs3885179	19	14 688 830	A/C	−	*ZNF333*	Reg pot, cSNP		18	0.61 (0.48–0.77)	0.000032	0.000031	0.66	1.00	0.046	0.035	0.056	0.033	0.021	0.024	0.018	0.24	
rs941796	20	39 724 220	A/G	+		Reg pot		22	1.22 (1.11–1.35)	0.000043	0.000037	0.66	1.00	0.398	0.422	0.374	0.408	0.013	0.013	0.012	0.88	
rs12480143	20	39 741 240	G/A	+				24	1.25 (1.13–1.39)	0.000044	0.000035	0.66	1.00	0.265	0.288	0.244	0.233	0.001	0.001	0.002	1.00	
rs928862	21	20 559 590	G/A	+		Reg pot	Near CNV	19	0.78 (0.69–0.88)	0.000040	0.000044	0.66	1.00	0.190	0.170	0.209	0.167	0.013	0.012	0.013	0.77	

Notes: Sorted by location. All locations per NCBI Build 35 (UCSC hg17). Alleles are given as minor/major. OR (CI), odds ratio (95% confidence interval). *P*-asymptotic, *P*-value from Trend test. *P*-empirical, pointwise *P*-value from adaptive permutation method in PLINK. For *q*-Value see text. *P*-gwemp, genome-wide empirical *P*-value by traditional permutation testing (5000 replicates). MAF, minor allele frequency. HapMap MAFs have been converted to the reference allele of the MDD sample. *P*-missing tests the difference in missingness between cases and controls. For noteworthy associations, the four flags refer to acceptable cluster plots, conformation to Hardy–Weinberg equilibrium, absence of plate-specific association outliers and the presence of a "proxy" SNP in high linkage disequilibrium with the primary SNP.

[a]TAMAL codes. Bioinformatic flag possibilities: coding SNP (cSNP), SNP in segmental duplication, known copy number variant (CNV), conserved base (Cons), miRNA target site, region of regulatory potential (reg pot), predicted promoter, transfactor binding site, enhancer, exon, splice site, mRNA expression QTL (lymphocytes or cortex). Only positive flags are shown.

[b]SLEP, Sullivan Lab Evidence Project (http://slep.unc.edu) a compendium of genetic findings from the literature. Sources (PubMed IDs): CNVs from Database of Genomic Variation (PMID 15286789), breast and colon cancer mutations (17932254), MDD genome-wide linkage studies (12612864, 14582139, 17427203), SCZ genome-wide linkage meta-analysis (12802786) and bipolar disorder (BIP) GWAS (17554300).

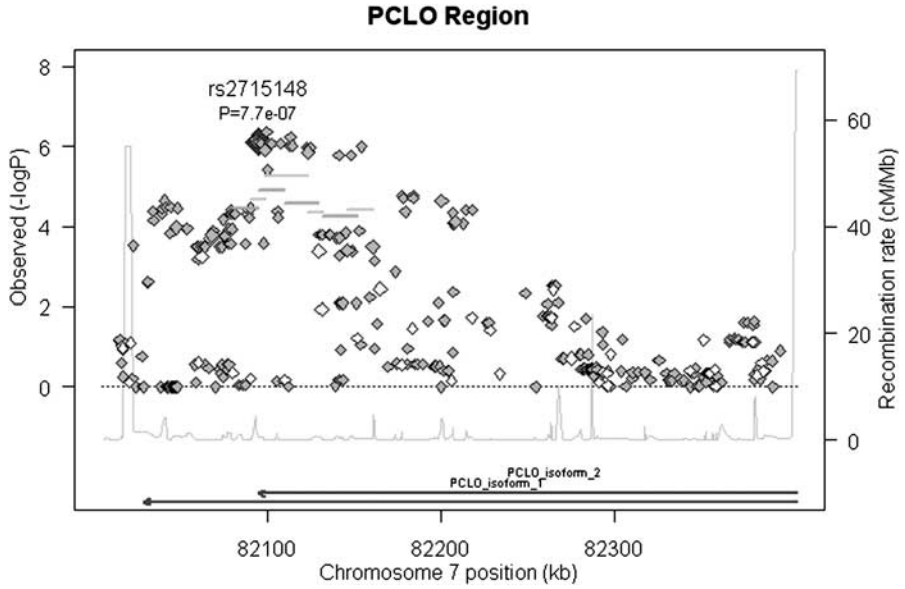

Figure 2 Plot of the piccolo (*PCLO*) region (NCBI build 35, UCSC hg17, chr7:82 000 000–82 500 000). *P*-values in this figure are all from SNPMstat. The *x* axis is chromosomal position, the left *y* axis is –log10(*P*) for genotyped SNPs (colored diamonds) and imputed SNPs (grey diamonds), and the right *y* axis is the recombination rate from the HapMap EUR panel (light blue curve). The color of the genotyped single nucleotide polymorphisms (SNPs) corresponds to LD with the SNP with smallest *P*-value (rs2715148): red $0.8 \leqslant r^2 \leqslant 1.0$, orange $0.5 \leqslant r^2 < 0.8$, yellow $0.2 \leqslant r^2 < 0.5$ and white $r^2 < 0.2$. The significant and extent of all three-SNP haplotypes with $P < 0.0001$ in this region are colored light green. The transcripts for two *PCLO* isoforms are shown in dark green at the bottom. Graph adapted from an *R* function by the Broad DGI group.

NESDA cases, all replication cases were adults of European ancestry on whom a structured clinical interview was used to substantiate the lifetime diagnosis of DSM-IV MDD,[1] and all studies excluded common MDD phenocopies (for example, depressive symptoms due to another psychiatric disorder or a general medical condition). As with NTR controls, all replication controls were adults of European ancestry ascertained from the population, and individuals reporting MDD symptoms were excluded. We estimated statistical power using Quanto[57] (assumptions: log-additive genetic model, MDD lifetime risk 0.15, MAF = 0.45 (similar to rs2522833), a genotypic relative risk of 1.14 ('shrunk' down from the observed GRR of 1.26 for rs2522833 to account for the 'Winner's Curse' phenomenon))[79] and a conservative two-tailed type 1 error rate of 0.00167 (= 0.05/30 replication SNPs). Statistical power was 97.2% for replication for the two SNPs genotyped in all samples ($N = 11\,972$) and 90.4% for the remaining SNPs ($N = 9278$). Five replication samples were genotyped for 30 SNPs using the same Sequenom iPlex SNP pool (15 SNPs were in the primary GWAS and 15 were selected to tag common variation in Europeans)[80] and one sample was successfully genotyped for two SNPs using TaqMan. The SNP selection strategy effectively cast a broad net over the region showing association in Figure 2. For the NESDA/NTR samples, agreement between the initial Perlegen genotypes in this region and independent re-genotyping was high (0.9987).

The single SNP results for MDD are depicted in Figure 3 and Table 4a. Our analytic plan dictated the combined analysis of all replication samples with the use of a one-tailed directional test. No association in the replication sample reached statistical significance after correction for multiple comparisons and SNP nonindependence due to LD (ninth column in Table 4a). Similarly, haplotype analyses did not reveal significantly associated regions (Supplementary Figure 16). There were four *P*-values < 0.05 in the replication sample but only rs10954694 also had *Z*-scores of the same sign in both samples. Table 4b shows the results for reoMDD, and no single SNP was significant after correction for multiple comparisons. When we repeated the MDD analyses restricted to female subjects, the observed significance levels did not become markedly stronger in any of the replication samples in contrast to the initial NESDA/NTR sample. Thus, results from analyses of all replication samples did not reach the *a priori* criterion for replication evidence for the involvement of *PCLO* in the etiology of MDD.

Unanticipated heterogeneity in cases

However, we observed, *a posteriori*, that there was potentially important heterogeneity in the replication samples for eight SNPs that were strongly associated in the original sample ($I^2 \geqslant 0.4$, ninth column in Table 4a). In investigating this further (Supplementary Methods), we determined that there was little evidence for genetic heterogeneity in the genotyped region for controls but, unexpectedly, there was significant heterogeneity in the cases. Principal components analysis and inspection of Table 4a and the forest plots in Figure 3 indicated that the outlier

Table 3 Clustering of SNPs with low *P*-values

Rank	*Chr*	*Start*	*End*	*Nsnps*	*Pmin*	N < *0.0001*	N < *0.001*	N < *0.01*	*Expressed in brain?*	*Genes*	*Gene products*	*SLEP*[a]
1	7	30 928 587	30 931 521	3	1.25E−06	2	0	1	Yes	ADCYAP1R1	Adenylate cyclase activating polypeptide 1 (pituitary) |receptor type I	Neuroactive ligandreceptor interaction
2	7	82 041 576	82 208 167	10	1.50E−06	6	4	0	Yes	PCLO	Piccolo (presynaptic cytomatrix protein)	
4	6	14 388 932	14 399 068	2	9.09E−06	1	1	0				
5	2	20 177 820	20 183 313	2	1.18E−05	2	0	0	Yes	LAPTM4A	Lysosomal-associated protein transmembrane 4α	
6	15	46 979 618	46 980 083	2	1.36E−05	1	1	0	Yes	CRI1/EID1/RaLP/ SHC4	CREBBP/EP300 inhibitor 1/EP300 interacting inhibitor of differentiation 1/railike protein/SHC (Src homology 2 domain containing) family, member 4	
9	2	201 794 446	201 880 818	2	2.44E−05	1	1	0	Yes	AJ487678/AJ487679/ AK125394/AY690601/ CASP10 CFLAR/ NDUFB3	Caspase 10/caspase 10/PRO3098/ caspase 10 splice variant G/caspase 10, apoptosis-related cysteine peptidase/CASP8 and FADD-like apoptosis regulator/NADH dehydrogenase (ubiquinone) 1β subcomplex, 3, 12 kDa	CASP10 causes multiple neoplasms (OMIM 601762); CFLAR upregulated in MDD in postmortem brain
14	20	39 724 220	39 742 644	5	4.27E−05	5	0	0				
15	6	14 386 148	14 397 061	3	4.23E−05	1	1	1				
16	4	184 652 456	184 658 003	3	4.28E−05	1	0	2				
17	5	117 174 763	117 282 887	4	4.84E−05	1	1	2				
19	10	127 071 672	127 087 021	3	0.000046	1	2	0				
20	5	22 752 605	22 792 155	3	4.65E−05	1	0	2	Yes	CDH12	Cadherin 12, type 2 (Ncadherin 2)	
22	15	88 130 196	88 136 792	2	5.52E−05	1	1	0	Yes	ANPEP/MESP2	Alanyl (membrane) aminopeptidase (aminopeptidase N, aminopeptidase M, microsomal aminopeptidase, CD13, p150)/ mesoderm posterior 2 homolog (mouse)	MESP2 causes spondylocostal dysostosis (OMIM 605195)
23	8	54 098 247	54 102 064	2	4.83E−05	2	0	0				
24	4	145 875 183	145 878 794	2	5.47E−05	1	0	1				
27	11	32 242 721	32 244 520	2	4.25E−05	2	0	0				
28	8	27 249 840	27 379 524	6	5.38E−05	1	2	3	Yes	AK128371/CHRNA2/ PTK2B	Hypothetical protein FLJ46514/ cholinergic receptor, nicotinic, α2 (neuronal)/ PTK2B protein tyrosine kinase 2β	CHRNA2 causes nocturnal frontal lobe epilepsy (OMIM 118502)
29	3	12 453 817	12 459 985	2	0.00005	1	1	0		PPARG	Peroxisome proliferator-activated receptorγ	Type 2 diabetes mellitus risk gene
32	3	99 975 821	100 183 009	2	7.26E−05	1	0	1	Yes	DCBLD2/ST3GAL6	Discoidin, CUB and LCCL domain containing 2/ST3β galactoside α 2,3sialyltransferase 6	
34	3	70 451 852	70 476 913	2	8.22E−05	1	0	1				
38	2	7 424 098	7 440 754	3	8.95E−05	1	1	1				
41	5	54 352 635	54 363 712	3	0.000071	1	1	1	Yes	GZMK	Granzyme K (granzyme 3; tryptase II)	
43	13	111 889 281	111 902 203	2	7.87E−05	1	1	0				
44	1	211 470 329	211 508 991	4	0.000072	1	3	0				
46	8	24 784 576	24 825 085	2	0.000359	0	2	0		NEF3/NEFM	Neurofilament 3 (150 kDa medium)/ neurofilament, medium polypeptide 150 kDa	

Abbreviation: MDD, Major depressive disorder.
[a]SLEP, Sullivan Lab Evidence Project (http://slep.unc.edu) a compendium of genetic findings from the literature.

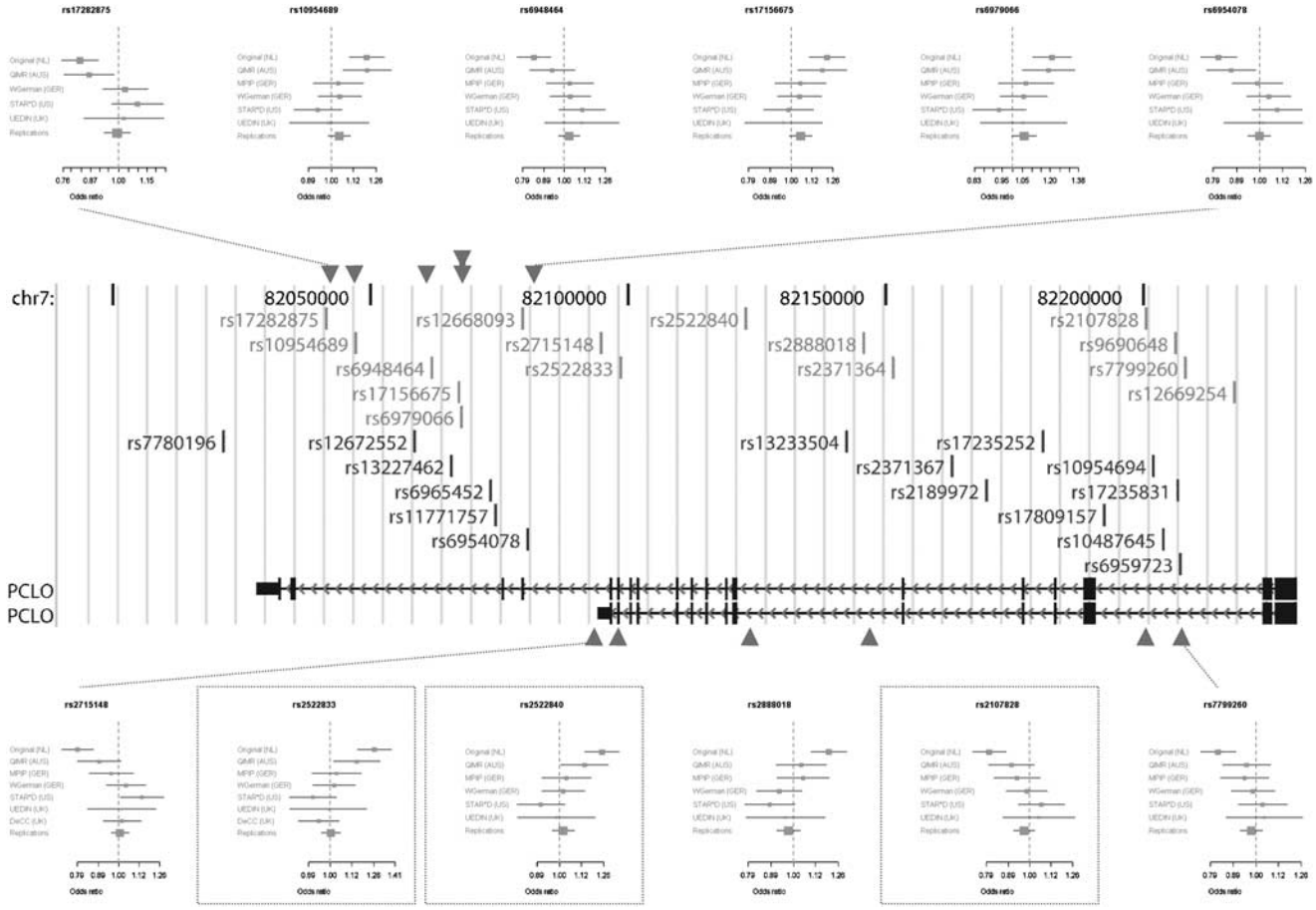

Figure 3 Piccolo (PCLO) region replication results for major depressive disorder (MDD) showing genomic context and forest plots for the top 12 single nucleotide polymorphisms (SNPs) in the original sample. The backbone of the graph is the region of *PCLO* targeted for follow-up. SNP locations are given by the grey triangles. There are 12 forest plots for the SNPs with $P<0.001$ in the original sample. Each forest plot is for one SNP and shows the odds ratio (square) and 95% confidence intervals (horizontal line) for a particular sample with the area of the square proportional to sample size.

was the Australian QIMR sample. Notably, the original and QIMR samples were particularly similar in that both studies included population-based cases and controls were selected to be at low liability for MDD based on longitudinal assessments. Of the nine SNPs with $P<0.05$ in the QIMR sample, eight had both low *P*-values and *Z*-scores with the same sign as in the NESDA/NTR sample. As an exploratory analysis, we analyzed the original and QIMR samples jointly, and the minimum *P*-value was 6.4×10^{-8} at the nonsynonymous SNP rs2522833 that gives rise to a serine to alanine substitution near the C2A calcium-binding domain of the PCLO protein.

Secondary analyses

We conducted additional analyses of the NESDA/NTR GWAS data set that were specified *a priori* but which should be considered exploratory.

(1) The network of proteins with which *PCLO* interacts in its function at the presynaptic cytoskeletal matrix is relatively well characterized, and we reasoned that genes encoding these proteins might harbor risk or protective variants. We assessed this hypothesis by testing for association conditioning on the *PCLO* nsSNP rs2522833 (that is, investigating whether controlling statistically for the effect of rs2522833 increases the salience of other SNP associations), assessing the minimum *P*-value per gene, and then comparing this list to a list of 54 genes that make proteins that interact with *PCLO*. This analysis did not reveal any SNPs or genes whose significance was markedly lower than without including rs2522833 in the logistic regression model. Moreover, no known *PCLO* interacting protein was notable on this list.

(2) We imputed genotypes for 2 037 829 autosomal SNPs using MACH with reference to HapMap CEU genotypes. The resulting λ was 1.048, and the minimum *P*-value was 1.21×10^{-7}. As noted above, 22 of the 25 most significant imputed associations were in the *PCLO* region. Investigation of SNP clustering that accounted for LD yielded results similar to those shown in Table 3.

(3) We assembled a list of 103 candidate genes that had been studied for association with MDD in the literature.[81] A total of 19 of these genes had no SNPs

within its transcript and another 9 genes had inadequate coverage (>1 SNP per 15 kb; Supplementary Table 17). Of the remaining 75 genes, only neuronal nitric oxide synthase (*NOS1*, $P=0.0006$) had $P<0.001$. However, *NOS1* (as with most genes in Supplementary Table 16) is quite large and there is a possibility of a potential influence on these results.

(4) We compared the GWAS association results to a meta-analysis of gene expression data from 12 studies of postmortem brain tissue in MDD cases compared with controls (10 frontal cortex and 2 cerebellum studies). These data are available via the Stanley Foundation (http://www.stanleygenomic.org). There were five genes with GWAS $P<0.05$ (all had gene expression changes significant at P 0.0004–0.007). The genes were: *SGCG* (sarcoglycan), *CALD1* (caldesmon 1), *EEF1A1* (eukaryotic translation elongation factor 1α1), *CFLAR* (CASP8 and FADD-like apoptosis regulator) and *TP73L* (tumor protein p73-like). There is no overlap of this list with the *PCLO* interactors or MDD candidate genes from the literature.

(5) Alternative models, filters and phenotypes: (i) For reoMDD, the minimum *P*-value over all GWAS SNPs was at the *PCLO* region SNP rs2715148 (8.4×10^{-8}) which ranked second of all SNPs using the trend test (Table 2). (ii) rs2715148 also had the smallest *P*-value under a dominant model of SNP action (6.2×10^{-6}). (iii) Given the female predominance in MDD, we analyzed data from women and men separately. For female cases and controls, rs2715148 had the smallest *P*-value (4.0×10^{-7}) and multiple other *PCLO* SNPs had *P*-values in the 10^{-5}–10^{6} range. For men, most *PCLO* SNPs had $P>0.05$ and the minimum was in the *SLC9A9* SNP rs4839627 (9.1×10^{-7}). (iv) Again, given sex differences in MDD prevalence, we investigated SNPs on chrX and chrY more closely. The minimum *P*-value in chrX pseudoautosomal region 1 was 0.02. For the non-PAR regions of chrX in women, the SNPs with the smallest *P*-values were rs11094388 ($P=0.0003$, intergenic), rs5971108 ($P=0.0003$, *PTCHD1*), rs5930667 ($P=0.0004$, intergenic), rs4618863 ($P=0.0005$, intergenic), rs2207796 ($P=0.0005$, in the very large gene *DMD*) and rs5936428 ($P=0.0009$, *FMR2*). For men, the minimum *P*-value on chrX was at rs10521594 ($P=5.4\times10^{-5}$, intergenic) and 0.22 on chrY.

Discussion

Overview

MDD is a common complex trait of enormous public health significance. As part of the GAIN initiative of the US Foundation for the NIH,[19] we conducted a GWAS of 435 291 SNPs genotyped in 1738 MDD cases and 1802 controls selected to be at low liability for MDD. Our study had numerous positive attributes including its historically large sample size, its largely population-based and longitudinal design, and relatively unbiased and dense genome-wide genotyping designed to capture common variation in subjects of European ancestry.

According to our primary analysis plan, no SNP–MDD phenotype association reached genome-wide significance as the minimum *q*-value was 0.28, greater than the pre-defined *q*-value threshold of 0.10. This result was not unexpected. For example, type 2 diabetes mellitus has arguably reaped the greatest harvest from GWAS[82] and yet two of the initial T2DM GWAS were unremarkable when analyzed independently.[83,84] One of the key lessons of the GWAS era is the importance of meta-analysis where its application to the primary GWAS can uncover positive findings that replicate well across studies.[18,85]

Is PCLO a causal risk factor for MDD?

Although no locus exceeded the genome-wide threshold after correction for multiple comparisons, 11 of the top 200 signals localized to a 167 kb region overlapping the gene *PCLO*. The protein product of *PCLO* localizes to the presynaptic active zone and is important in brain monoaminergic neurotransmission,[86] clearly intersecting with a venerable hypothesis of the etiology of mood disorders.[87] Moreover, the third most significant association was a common nonsynonymous SNP near its critical C2A binding domain in PCLO.[88,89] Although it is an obvious candidate gene, we are not aware of any prior association studies of *PCLO* and mood disorders (*PCLO* is in a region of 7q implicated by linkage in autism and one autism association study has been published).[90]

We judged the intersection of this GWAS result with prior knowledge sufficient to trigger a large-scale replication effort by genotyping *PCLO* SNPs in 6079 MDD-independent cases and 5893 controls. Statistical power to replicate exceeded 90% even after accounting for[79] the 'Winner's Curse' phenomenon (a form of regression to the mean whereby the true genotypic relative risk is overestimated in the initial study).[91,92] However, in spite of the apparent *a priori* strength of a hypothesis of genetic variation in *PCLO* in the etiology of MDD, no SNP analyzed in the replication sample met appropriately rigorous criteria for replication.[21] Therefore, unlike GWAS for many nonpsychiatric biomedical disorders, our GWAS and replication efforts did not yield 'proof beyond a reasonable doubt' level of evidence for an association between genetic variation in *PCLO* and MDD.

Investigation of the sources of heterogeneity in the replication samples indicated that controls were genetically similar to the original sample in the *PCLO* region but that cases were dissimilar. We observed, *a posteriori*, that both principal components derived from *PCLO* region genotypes in QIMR cases and effect size estimates in the QIMR replication sample tended to be similar to the original sample. This is notable because, of all the replication samples, ascertainment of QIMR subjects was most similar to the primary NESDA/NTR sample in that cases were identified from population-based sources (100% for QIMR and 60% for NESDA) rather than tertiary sources as for the other replication samples. MDD cases from clinical

samples may differ from population-based cases due to selection bias,[93] Berkson's bias,[94,95] differing referral filters[96] or even a different genetic basis[97] with respect to genetic variation in the *PCLO* region.

Joint analysis of the NESDA/NTR and QIMR samples yielded $P = 6.4 \times 10^{-8}$ (uncorrected for multiple hypothesis testing) for the nonsynonymous SNP rs2522833. This result suggests a specific hypothesis for future studies: an association between genetic variation in *PCLO* and MDD may be detected only in population-based cases. Thus, it would be premature to exclude *PCLO* from a function in the etiology of some forms of MDD.

The heterogeneous nature of MDD

Interpretation of the *PCLO* replication efforts is consistent with two broad possibilities. The first possibility is that genetic variation in *PCLO* is truly not associated with MDD. This interpretation is supported by the replication analyses (specified *a priori*) in which no SNP was significantly associated after correction for multiple comparisons and SNP dependence due to LD. This strict interpretation is generally viewed as 'best practice' in human genetics[21] but implicitly assumes etiological homogeneity for MDD in the *PCLO* region. The second possibility invokes a less parsimonious model involving heterogeneity, that genetic variation in *PCLO* is etiologically causal to some subtypes of MDD. This interpretation is an *a posteriori* hypothesis consistent with the empirical results particularly in the notable differences in associations between samples, case ascertainment strategies, and indications from principal components analysis that NESDA and QIMR cases are more similar than the clinically ascertained subjects.

It is notable that the control samples from each site were considerably more similar than cases from the same sites.

The tension between null *a priori* results and plausible *a posteriori* hypotheses is a core issue in psychiatric genetics. Important phenotypes like MDD are defined reliably and with reference to diagnostic schema developed principally for clinical purposes. Heterogeneous etiology of MDD is widely suspected but there are no proven ways to index heterogeneity (indeed, a prominent rationale for genetics studies is improve differential diagnosis).

Our results are consistent with prior observations of the heterogeneous nature of MDD, particularly with regard to ascertainment. Individuals who meet MDD criteria from community or primary care sources may have a more inclusive and less comorbid form of MDD[98] whereas tertiary ascertainment may yield subjects with greater comorbidity and perhaps distinctive etiology.[99] In particular, it is formally possible (but unproven) that the *PCLO* results are accurate—genetic variation in *PCLO* might be causal to the types of MDD seen in community samples but other loci contribute to a distinctive type of MDD seen in tertiary care samples.

Other hypotheses

There were two MDD cases who may have had unrecognized genomic disorders[100] (possible Turner's and Klinefelter's syndromes). We speculate that small numbers of cases with MDD will have CNV-related genomic disorders that are plausibly causal to MDD. Clarification of the function of such rare variants will require larger samples.

Most of the additional exploratory analyses were unrevealing, including examination of proteins known to interact with PCLO, genotype imputation, comparison of GWAS findings with MDD candidate genes from the literature and gene expression changes in the brain in cases with MDD, and alternative genetic models, phenotype definitions and sex-specific analyses.

We searched the Sullivan Lab Evidence Project (SLEP) compendium of psychiatric genetics findings[101] in an attempt to discover overlap of our findings with those reported in the literature. First, with reference to a meta-analysis of microarray studies on the Stanley brain bank MDD and control samples, expression of *CFLAR* and *MARCH3* were increased and *LST1* and *HLA-B* were decreased in MDD postmortem frontal cortex. These regions ranked 9, 232, 267 and 432 in the NESDA/NTR GWAS. Second, we looked for convergence of our findings with other GWAS of psychiatric disorders. Notable genomic locations of overlap of the top 480 regions in the present GWAS were found with GWAS for ADHD (*ITIH1*; S Faraone, personal communication), the Wellcome Trust Case-Control Consortium GWAS for bipolar disorder (*SHFM1* and *UGT2B4*)[102] and a bipolar GWAS that used DNA pooling (*GRM7* and *DGKH*).[70] Third, we looked at the minimum *P*-values in our study for genes that met or nearly achieved genome-wide significance: the minimum *P*-values in our study for *MAMDC1*[103] were 0.004, 0.03 for *ZNF804A*,[104] 0.002 for *ANK3*[105] and 0.03 for *CACNA1C*.[105] These overlaps are intriguing (although the possibility of chance cannot be excluded), and will be formally investigated as part of our participation in the Psychiatric GWAS Consortium analyses.[18]

Limitations

(1) Although statistical power has been systematically underestimated in psychiatric genetics, when we began this study in Q3 2006, it was believed that statistical power would be reasonable to detect realistic genetic effects. However, the definition of 'realistic' has shifted considerably since 2006 and it may be important to design studies that can detect genotypic relative risks < 1.10. (2) When this study began, the coverage and performance of the Perlegen GWAS platform were among the better options available.[19] The technology and pricing have evolved rapidly and superior platforms are now available. A key limitation of the Perlegen platform is its inability to assess CNV[106] that may be particularly salient for psychiatric disorders.[107,108] More generally, the GWAS platform might not be sufficiently 'genome-wide' and

Table 4 *PCLO* replication results

SNP ID	hg17 position	NESDA-NTR (NL)			All replication samples					QIMR (AUS)				M-P Inst Psych (GER)				West Germany (GER)				STAR*D (US)				U Edinburgh (UK)				DeCC (UK)			
		N	Z	P	N	Z	P	P_corr	I^2	N	Z	P	P_corr	N	Z	P	P_corr	N	Z	P	P_corr	N	Z	P	P_corr	N	Z	P	P_corr	N	Z	P	P_corr
(a) MDD (major depressive disorder)																																	
rs7780196	82 021 603	3657	2.87	0.00406	9270	1.58	0.11	0.81	0.65	2004	2.87	0.0041	0.066	1907	1.87	0.06	0.59	2482	-0.82	0.41	1	2051	−0.91	0.36	1	826	0.68	0.49	1				
rs17282875	82 041 576	3657	−4.09	4.3E−05	7235	−0.23	0.82	1	0.61	2002	−2.31	0.021	0.29					2479	0.55	0.58	1	1928	1.40	0.16	0.90	826	0.25	0.81	1				
rs10954689	82 047 024	3657	3.91	9.1E-05	9269	1.38	0.17	0.92	0.52	2004	2.88	0.0039	0.076	1899	0.54	0.59	1	2489	0.76	0.45	1	2051	−1.10	0.27	0.98	826	−0.12	0.90	1				
rs12672552	82 058 350	3658	0.58	0.56	9277	−1.38	0.17	0.92	0.00	2005	−0.30	0.77	1	1907	−0.81	0.42	1	2488	−0.02	0.99	1	2051	−0.98	0.32	0.99	826	−1.37	0.17	0.91				
rs6948464	82 061 983	3658	−3.46	0.00055	9269	0.86	0.39	1	0.00	2005	−1.05	0.29	0.99	1907	0.46	0.65	1	2480	0.56	0.58	1	2051	1.50	0.13	0.85	826	0.92	0.36	1				
rs13227462	82 065 698	3658	−0.98	0.32	9271	−0.81	0.42	1	0.69	2004	−0.24	0.81	1	1906	−1.42	0.16	0.87	2484	0.56	0.58	1	2051	1.08	0.28	0.98	826	−2.88	0.0040	0.050				
rs17156675	82 067 056	3658	3.89	0.0001	9277	1.61	0.11	0.80	0.16	2005	2.48	0.013	0.20	1907	0.70	0.48	1	2488	0.74	0.46	1	2051	−0.21	0.84	1	826	−0.39	0.69	1				
rs6979066	82 067 588	3656	3.94	8.1E−05	9237	1.81	0.07	0.66	0.39	1983	2.65	0.0082	0.13	1906	0.93	0.35	1	2475	0.91	0.36	1	2047	−0.99	0.32	0.99	826	0.48	0.63	1				
rs6965452	82 073 522	3658	−1.51	0.13	9267	−1.34	0.18	0.92	0.00	2001	−1.97	0.048	0.51	1903	−0.88	0.38	1	2486	−0.42	0.68	1	2051	0.62	0.54	1	826	−0.74	0.46	1				
rs11771757	82 074 382	3645	0.68	0.50	6774	0.48	0.63	1	0.66	1993	0.83	0.41	1	1905	0.63	0.53	1					2051	−1.85	0.06	0.59	825	1.93	0.05	0.55				
rs12668093	82 079 684	3658	0.73	0.47	9258	−2.21	0.027	0.36	0.00	2005	−0.38	0.71	1	1888	−0.99	0.32	0.99	2489	−1.53	0.13	0.82	2050	−1.25	0.21	0.95	826	−0.97	0.33	0.99				
rs6954078	82 080 529	3633	−4.44	8.8E−06	9204	−0.07	0.94	1	0.48	1990	−2.26	0.024	0.33	1890	−0.18	0.85	1	2469	0.79	0.43	1	2037	1.38	0.17	0.91	818	0.17	0.86	1				
rs2715148	82 094 686	3651	−5.05	4.4E−07	11 850	0.23	0.82	1	0.40	2003	−1.77	0.076	0.66	1905	−0.66	0.51	1	2478	0.68	0.49	1	2048	2.06	0.04	0.45	826	0.16	0.88	1	2590	0.33	0.74	0.88
rs2522833	82 098 359	3658	5.01	5.4E−07	11 934	0.12	0.90	1	0.41	2002	2.20	0.028	0.35	1892	0.49	0.62	1	2484	0.35	0.72	1	2050	−1.44	0.15	0.89	826	−0.10	0.92	1	2680	−1.10	0.27	0.37
rs2522840	82 123 066	3656	4.79	1.6E−06	9268	0.63	0.53	1	0.40	2004	2.06	0.039	0.44	1904	0.55	0.58	1	2483	0.33	0.75	1	2051	−1.52	0.13	0.84	826	−0.16	0.87	1				
rs13233504	82 142 482	3656	2.63	0.00843	9263	0.82	0.41	1	0.12	2003	1.12	0.26	0.98	1907	1.49	0.14	0.84	2476	−0.22	0.83	1.00	2051	−1.08	0.28	0.99	826	0.56	0.58	1				
rs2888018	82 145 941	3657	3.80	0.00015	9260	−0.86	0.39	1	0.19	2005	0.60	0.55	1	1903	0.73	0.47	1	2477	−1.23	0.22	0.95	2049	−1.85	0.07	0.62	826	−0.40	0.69	1				
rs2371364	82 151 525	3656	2.22	0.02611	9271	1.60	0.11	0.80	0.21	2004	2.64	0.008	0.13	1903	−0.33	0.74	1	2487	0.69	0.49	1	2051	0.09	0.93	1	826	0.24	0.81	1				
rs2371367	82 163 042	3656	2.87	0.00406	9257	0.67	0.51	1	0.00	2004	0.88	0.38	1	1893	−0.32	0.75	1	2484	0.13	0.89	1	2050	0.08	0.94	1	826	0.08	0.94	1				
rs2189972	82 169 314	3657	1.37	0.17	9272	2.20	0.028	0.35	0.00	2004	1.69	0.092	0.71	1905	0.11	0.91	1	2486	0.85	0.40	1	2051	0.80	0.42	1	826	1.16	0.25	0.98				
rs17235252	82 180 688	3658	−1.03	0.30	7363	−1.70	0.09	0.73	0.62	2004	−1.47	0.14	0.87					2487	−2.29	0.02	0.26	2046	−0.05	0.96	1	826	1.47	0.14	0.86				
rs17809157	82 192 478	3652	−2.58	0.00975	9266	−1.94	0.05	0.56	0.00	2003	−1.36	0.17	0.92	1906	−1.53	0.13	0.82	2481	−1.15	0.25	0.97	2050	0.26	0.79	1	826	0.47	0.64	1				
rs2107828	82 200 320	3658	−4.69	2.8E−06	9271	−1.05	0.29	0.99	0.09	2005	−1.59	0.11	0.79	1902	−1.09	0.27	0.98	2487	−0.30	0.76	1	2051	0.97	0.33	0.99	826	0.48	0.63	1				
rs10954694	82 201 812	3656	2.77	0.00558	9274	2.53	0.011	0.19	0.00	2005	1.24	0.22	0.95	1907	1.02	0.31	0.99	2486	1.59	0.11	0.77	2050	0.71	0.47	1	826	0.75	0.45	1				
rs10487645	82 203 942	3658	0.44	0.66	9272	−0.79	0.43	1	0.00	2005	0.17	0.87	1	1901	0.55	0.58	1	2489	−1.87	0.06	0.55	2051	−0.71	0.48	1	826	−0.15	0.88	1				
rs9690648	82 205 975	3658	0.52	0.60	9278	−1.01	0.31	0.99	0.01	2005	−0.01	0.99	1	1907.00	0.69	0.49	1	2489	−2.02	0.04	0.43	2051	−0.97	0.33	0.99	826	−0.24	0.81	1				
rs17235831	82 206 612	3652	−0.94	0.35	9271	−0.75	0.46	1	0.22	2003	−0.69	0.49	1	1904	−0.43	0.67	1	2488	−1.34	0.18	0.91	2051	0.16	0.88	1	825	1.57	0.12	0.80				
rs6959723	82 206 991	3656	−2.16	0.03084	9265	−1.94	0.05	0.55	0.00	2003	−0.70	0.48	1	1905	−1.30	0.19	0.93	2481	−1.15	0.25	1	2050	−0.42	0.67	1	826	−0.44	0.66	1				
rs7799260	82 208 167	3658	−4.19	2.8E−05	9274	−0.87	0.39	1	0.00	2005	−0.78	0.44	1	1905	−0.94	0.35	0.99	2487	−0.30	0.76	1	2051	0.51	0.61	1	826	0.41	0.68	1				
rs12669254	82 217 749	3656	2.31	0.02077	9274	−2.06	0.039	0.46	0.00	2005	−0.93	0.35	0.99	1904	−0.58	0.56	1	2489	−0.82	0.41	1	2050	−1.04	0.30	0.99	826	−1.79	0.07	0.62				
(b) reoMDD (recurrent, early onset MDD)																																	
rs7780196	82 021 603	2369	2.80	0.00514	6403	0.86	0.39	1	0.64	1255	2.16	0.03	0.36	1310	2.05	0.04	0.45	1580	−1.16	0.24	0.97	1548	−1.09	0.28	0.98	710	0.17	0.86	1				
rs17282875	82 041 576	2368	−4.48	7.5E−06	4983	0.10	0.92	1	0.00	1254	−1.00	0.32	0.99					1582	0.68	0.50	1	1437	1.06	0.29	0.99	710	0.20	0.84	1				
rs10954689	82 047 024	2369	4.06	4.9E−05	6403	1.22	0.22	0.97	0.00	1255	1.27	0.20	0.95	1305	1.09	0.28	0.98	1,85	0.11	0.91	1	1548	−0.31	0.76	1	710	0.25	0.81	1				
rs12672552	82 058 350	2369	1.19	0.24	6409	−0.50	0.61	1	0.00	1256	−1.28	0.20	0.94	1310	−0.30	0.76	1	1585	0.22	0.83	1	1548	0.18	0.86	1	710	−1.08	0.28	0.99				
rs6948464	82 061 983	2369	−2.86	0.00430	6407	0.93	0.35	1	0.00	1256	−0.58	0.56	1	1310	−0.05	0.96	1	1583	1.76	0.08	0.67	1548	1.16	0.25	0.97	710	0.86	0.39	1				
rs13227462	82 065 698	2369	−2.96	0.00306	6405	−0.91	0.36	1	0.54	1255	−0.23	0.82	1	1309	−1.93	0.05	0.51	1583	0.15	0.88	1	1,48	0.90	0.37	1	710	−2.21	0.027	0.296				
rs17156675	82 067 056	2369	3.88	1.1E−04	6409	1.15	0.25	0.98	0.27	1256	2.40	0.02	0.25	1310	0.70	0.49	1	1585	0.47	0.64	1	1548	−0.41	0.68	1	710	−0.23	0.82	1				
rs6979066	82 067 588	2367	3.64	2.7E−04	6383	1.13	0.26	0.98	0.44	1238	2.09	0.04	0.43	1309	1.37	0.17	0.91	1580	−0.24	0.81	1	1546	−1.03	0.30	0.99	710	0.58	0.56	1				
rs6965452	82 073 522	2369	−3.16	0.00155	6401	−1.14	0.26	0.98	0.00	1253	−1.15	0.25	0.97	1307	−1.12	0.26	0.97	1583	−1.04	0.30	0.98	1548	0.43	0.67	1	710	−0.69	0.49	1				
rs11771757	82 074 382	2361	−0.02	0.99	4814	0.84	0.40	1	0.60	1249	−0.33	0.74	1	1308	1.58	0.11	1					1548	−1.39	0.16	0.90	709	1.70	0.09	0.72				
rs12668093	82 079 684	2369	1.92	0.06	6395	−1.64	0.10	0.77	0.00	1256	−0.97	0.33	0.99	1296	−0.68	0.50	1	1585	−0.94	0.35	0.99	1548	−0.64	0.52	1	710	−1.28	0.20	0.94				
rs6954078	82 080 529	2355	−4.98	6.5E−07	6365	0.10	0.92	1	0.21	1251	−1.28	0.20	0.95	1300	−0.78	0.44	1	1578	1.35	0.18	0.91	1534	1.01	0.31	0.99	702	0.30	0.76	1				
rs2715148	82 094 686	2363	−5.35	9.0E−08	8416	0.32	0.75	1	0.40	1255	−0.85	0.40	1	1308	−1.17	0.24	0.96	1577	1.77	0.08	0.68	1547	1.92	0.06	0.55	710	0.19	0.85	1	2019	−0.18	0.86	0.96
rs2522833	82 098 359	2369	4.95	7.4E−07	8478	0.21	0.83	1	0.03	1254	1.09	0.28	0.98	1300	1.39	0.16	0.88	1582	−0.76	0.45	1	1547	−0.99	0.32	0.99	710	0.14	0.89	1	2085	−0.48	0.63	0.78
rs2522840	82 123 066	2368	4.89	9.9E−07	6404	0.46	0.65	1	0.18	1256	1.12	0.26	0.98	1308	1.32	0.19	0.91	1582	−0.74	0.46	1	1548	−1.08	0.28	0.98	710	0.14	0.89	1				
rs13233504	82 142 482	2367	2.26	0.02405	6405	0.50	0.62	1	0.08	1256	−0.05	0.96	1	1310	1.73	0.08	0.67	1581	−0.65	0.51	1	1548	−0.84	0.40	1	710	0.44	0.66	1				
rs2888018	82 145 941	2368	4.08	4.5E−05	6399	−1.26	0.21	0.95	0.15	1256	−1.06	0.29	0.98	1306	1.11	0.27	0.97	1579	−1.55	0.12	0.81	1548	−1.54	0.12	0.82	710	−0.62	0.54	1				
rs2371364	82 151 525	2368	1.43	0.15	6403	1.15	0.25	0.98	0.58	1256	3.23	0.001	0.02	1306	−0.68	0.49	1	1583	−0.29	0.78	1	1548	0.18	0.85	1	710	0.47	0.64	1				
rs2371367	82 163 042	2367	2.40	0.02	6397	0.14	0.89	1	0.00	1256	0.10	0.92	1	1303	−0.18	0.86	1	1581	−1.21	0.23	0.96	1547	−0.07	0.95	1	710	0.35	0.73	1				
rs2189972	82 169 314	2369	0.94	0.35	6406	1.44	0.15	0.89	0.00	1255	1.67	0.10	0.73	1310	−0.08	0.94	1	1583	−0.20	0.84	1	1548	0.21	0.83	1	710	1.00	0.32	0.99				
rs17235252	82 180 688	2369	−1.27	0.20	5096	−0.92	0.36	1	0.36	1256	−0.50	0.62	1					1585	−1.78	0.08	0.62	1545	−0.02	0.99	1	710	1.25	0.21	0.95				
rs17809157	82 192 478	2364	−1.96	0.05	6402	−0.49	0.62	1	0.00	1255	−0.68	0.49	1	1309	−0.68	0.49	1	1581	0.86	0.39	1	1547	0.54	0.59	1	710	0.59	0.55	1				
rs2107828	82 200 320	2369	−4.60	4.2E−06	6405	−0.06	0.95	1	0.00	1256	−0.56	0.58	1	1307	−1.04	0.30	0.98	1584	1.14	0.25	0.97	1548	1.18	0.24	0.97	710	0.38	0.71	1				
rs10954694	82 201 812	2367	1.91	0.06	6406	2.51	0.01	0.19	0.00	1256	1.72	0.09	0.71	1310	1.13	0.26	0.97	1583	0.70	0.48	1	1547	0.48	0.63	1	710	0.92	0.36	1				
rs10487645	82 203 942	2369	1.55	0.12	6405	−1.49	0.14	0.86	0.00	1256	−1.12	0.26	0.97	1306	0.41	0.68	1	1585	−1.01	0.31	0.99	1548	−1.16	0.25	0.97	710	−0.94	0.34	1				
rs9690648	82 205 975	2369	1.74	0.08	6409	−1.68	0.09	0.73	0.00	1256	−1.22	0.22	0.95	1310	0.43	0.67	1	1585	−1.13	0.26	0.97	1548	−1.47	0.14	0.86	710	−0.85	0.39	1				
rs17235831	82 206 612	2364	−1.03	0.30	6403	−0.25	0.80	1	0.00	1255	0.17	0.87	1	1307	0.36	0.72	1	1584	−1.43	0.15	0.87	1548	−0.04	0.97	1	709	1.21	0.23	0.96				
rs6959723	82 206 991	2368	−3.11	0.00189	6402	−1.23	0.22	0.96	0.00	1255	−0.02	0.98	1	1308	−1.08	0.28	0.98	1582	−1.38	0.17	0.90	1547	−0.12	0.90	1	710	−0.49	0.63	1				
rs7799260	82 208 167	2369	−4.30	1.7E−05	6407	−0.61	0.54	1	0.00	1256	0.19	0.85	1	1309	−1.20	0.23	0.95	1584	0.33	0.74	1	1548	0.56	0.58	1	710	0.35	0.73	1				
rs12669254	82 217 749	2368	2.50	0.01	6406	−1.83	0.07	0.63	0.04	1256	−2.17	0.03	0.35	1308	−0.15	0.88	1	1585	−0.98	0.33	0.99	1547	−0.40	0.69	1	710	−1.23	0.22	0.95				

Abbreviations: N = total sample size for an analysis, Z = logistic regression beta divided by its standard error, P = asymptotic P-value from Wald χ^2-test (1 d.f.) uncorrected for multiple comparisons, P_corr = empirical P-value from accounting for multiple comparisons and LD structure (50K permutations), and $I^2 = I^2$, an index of heterogeneity of logistic regression parameter estimates.

unbiased: the platform may have had inadequate coverage in an etiologically important region of the genome, SNPs are only one type of genetic variation, and important non-SNP genetic variation might not have been sufficiently well captured. (3) There was an imbalance in the proportion of men in cases and controls. Although it is unclear whether and how this might bias the results, it may have lead to some degree of bias. (d) Finally, GWASs are predicated upon the crucial assumption that the predominant diagnostic criteria are valid with respect to the fundamental architecture of the disorder.

Conclusions

We describe here a large effort to identify DNA sequence variation fundamental to MDD. Although our initial GWAS results for the *PCLO* region were intriguing, this highly plausible hypothesis did not find support in a large-scale replication attempt. Our hypothesis about a function of genetic variation in *PCLO* for MDD in population but not clinical settings emphasizes the importance of knowing the epidemiological sampling frame for a study. Finally, we hope that the model we used in this study—a cooperative international effort—will be adopted by groups studying other psychiatric disorders in order to maximize progress.

Acknowledgments

We acknowledge support from NWO: genetic basis of anxiety and depression (904-61-090); resolving cause and effect in the association between exercise and well-being (904-61-193); twin-family database for behavior genomic studies (480-04-004); twin research focusing on behavior (400-05-717), Center for Medical Systems Biology (NWO Genomics); Spinozapremie (SPI 56-464-14192); Centre for Neurogenomics and Cognitive Research (CNCR-VU); genome-wide analyses of European twin and population cohorts (EU/QLRT-2001-01254); genome scan for neuroticism (NIMH R01 MH059160); Geestkracht program of ZonMW (10-000-1002); matching funds from universities and mental health care institutes involved in NESDA (GGZ Buitenamstel-Geestgronden, Rivierduinen, University Medical Center Groningen, GGZ Lentis, GGZ Friesland, GGZ Drenthe). Genotyping was funded by the Genetic Association Information Network (GAIN) of the Foundation for the US National Institutes of Health, and analysis was supported by grants from GAIN and the NIMH (MH081802). Genotype data were obtained from dbGaP (http://www.ncbi.nlm.nih.gov/dbgap, accession number phs000020.v1.p1). Statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org) which is financially supported by the NWO (480-05-003). Dr Sullivan was also supported by R01s MH074027 and MH077139. Dr Schosser was supported by an Austrian Science Fund Erwin-Schrödinger-Fellowship. We express our thanks to: the GAIN Genotyping group (Dr Gonçalo Abecasis, chair) for help with quality control; Dr Gonçalo Abecasis and Dr Jun Li for assistance with MACH; Dr Shaun Purcell for PLINK; Troy Dumenil (QIMR) for expert assistance with the replication genotyping; Dr Dina Ruano (Portuguese Foundation for Science and Technology, SFRH/BPD/28725/2006); and Dr Pam Madden (DA012854) and Dr Richard Todd (AA013320) for supplying some of the phenotypes used in the Australian sample. Replication genotyping of the STAR*D samples was supported by a grant from the Bowman Family Foundation and the Sidney R Baer, Jr Foundation. We gratefully acknowledge NARSAD for funding the *PCLO* follow-up genotyping.

Conflict of interest/disclosure (past 3 years)

Dr Baune has received honoraria for educational training of psychiatrists and general practitioners from Lundbeck, AstraZeneca and Pfizer Pharmaceuticals and travel grants from AstraZeneca, Bristol-Meyrs Squibb, Janssen and Pfizer Pharmaceuticals. Dr Fava has received: research support from Abbott Laboratories, Alkermes, Aspect Medical Systems, AstraZeneca, Bristol-Myers Squibb Company, Cephalon, Eli Lilly & Company, Forest Pharmaceuticals Inc., GlaxoSmithKline, J&J Pharmaceuticals, Lichtwer Pharma GmbH, Lorex Pharmaceuticals, Novartis, Organon Inc., PamLab, LLC, Pfizer Inc., Pharmavite, Roche, Sanofi-Aventis, Solvay Pharmaceuticals Inc., Synthelabo, Wyeth-Ayerst Laboratories; advisory/consulting fees from Abbott Laboratories, Amarin, Aspect Medical Systems, AstraZeneca, Auspex Pharmaceuticals, Bayer AG, Best Practice Project Management Inc., Biovail Pharmaceuticals Inc., BrainCells Inc., Bristol-Myers Squibb Company, Cephalon, CNS Response, Compellis, Cypress Pharmaceuticals, Dov Pharmaceuticals, Eli Lilly & Company, EPIX Pharmaceuticals, Fabre-Kramer Pharmaceuticals Inc., Forest Pharmaceuticals Inc., GlaxoSmithKline, Grunenthal GmBH, Janssen Pharmaceutica, Jazz Pharmaceuticals, J&J Pharmaceuticals, Knoll Pharmaceutical Company, Lorex Pharmaceuticals, Lundbeck, MedAvante Inc., Merck, Neuronetics, Novartis, Nutrition 21, Organon Inc., PamLab, LLC, Pfizer Inc., PharmaStar, Pharmavite, Precision Human Biolaboratory, Roche, Sanofi-Aventis, Sepracor, Solvay Pharmaceuticals Inc., Somaxon, Somerset Pharmaceuticals, Synthelabo, Takeda, Tetragenex, Transcept Pharmaceuticals, Vanda Pharmaceuticals Inc., Wyeth-Ayerst Laboratories; speaking fees from AstraZeneca, Boehringer-Ingelheim, Bristol-Myers Squibb Company, Cephalon, Eli Lilly & Company, Forest Pharmaceuticals Inc., GlaxoSmithKline, Novartis, Organon Inc., Pfizer Inc., PharmaStar, Primedia, Reed-Elsevier, Wyeth-Ayerst Laboratories; has equity holdings in Compellis, MedAvante; and has royalty/patent, other income for patent applications for SPCD and for a combination of azapirones and bupropion in MDD, copyright

royalties for the MGH CPFQ, DESS and SAFER. Dr. Nolen has received: speaking fees from AstraZeneca, Eli Lilly, Pfizer, Servier, Wyeth; unrestricted research funding from AstraZeneca, Eli Lilly, GlaxoSmith Kline, Wyeth; and served on advisory boards for AstraZeneca, Cyberonics, Eli Lilly, GlaxoSmithKline, Pfizer, Servier. Dr Perlis has received consulting fees or honoraria from AstraZeneca, Bristol-Myers Squibb, Eli Lilly, GlaxoSmithKline, Pfizer and Proteus; he is a stockholder in Concordant Rater Systems, LLC, and the holder of a patent related to the monitoring of raters in clinical trials. Dr Smoller has consulted to Eli Lilly, received honoraria from Hoffman-La Roche Inc., Enterprise Analysis Corp. and MPM Capital, and has served on an advisory board for Roche Diagnostics Corporation. Dr Sullivan has received unrestricted research support from Eli Lilly.

References

1 American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association: Washington, DC, 1994.
2 Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S *et al.* Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the National Comorbidity Survey. *Arch Gen Psychiatry* 1994; **51**: 8–19.
3 Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR *et al.* The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA* 2003; **289**: 3095–3105.
4 Kessler RC, Ustun TB. The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *Int J Methods Psychiatr Res* 2004; **13**: 93–121.
5 Weissman MM, Bland R, Joyce PR, Newman S, Wells JE, Wittchen H-U. Sex differences in rates of depression: cross-national perspectives. *J Affect Disord* 1993; **29**: 77–84.
6 Piccinelli M, Wilkinson G. Outcome of depression in psychiatric settings. *Br J Psychiatry* 1994; **164**: 297–304.
7 Wells KB, Stewart A, Hays RD, Burnam MA, Rogers W, Daniels M *et al.* The functioning and well-being of depressed patients: results from the Medical Outcomes Study. *J Am Med Assoc* 1989; **262**: 914–919.
8 Broadhead WE, Blazer DG, George LK, Tse CK. Depression, disability days, and days lost from work in a prospective epidemiologic survey. *J Am Med Assoc* 1990; **264**: 2524–2528.
9 Judd LL, Paulus MP, Wells KB, Rapaport MN. Socioeconomic burden of subsyndromal depressive symptoms and major depression in a sample of the general population. *Am J Psychiatry* 1996; **153**: 1411–1417.
10 Tsuang MT, Woolson RF. Excess mortality in schizophrenia and affective disorders. *Arch Gen Psychiatry* 1978; **35**: 1181–1185.
11 Berglund M, Nilsson K. Mortality in severe depression: a prospective study including 103 suicides. *Acta Psychiatr Scand* 1987; **76**: 372–380.
12 Black DW, Winokur G, Nasrallah A. Is death from natural causes still excessive in psychiatric patients? *J Nerv Ment Dis* 1987; **175**: 674–680.
13 Zilber N, Schufman N, Lerner Y. Mortality among psychiatric patients—the groups at risk. *Acta Psychiatr Scand* 1989; **79**: 248–256.
14 Greenberg PE, Stiglin LE, Finkelstein SN, Berndt ER. The economic burden of depression in 1990. *J Clin Psychiatry* 1993; **54**: 405–418.
15 Murray CJL, Lopez AD. Evidence-based health policy: lessons from the Global Burden of Disease Study. *Science* 1996; **274**: 740–743.
16 Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: review and meta-analysis. *Am J Psychiatry* 2000; **157**: 1552–1562.
17 Altshuler D, Daly M. Guilt beyond a reasonable doubt. *Nat Genet* 2007; **39**: 813–815.
18 Psychiatric GWAS Consortium. A framework for interpreting genomewide association studies of psychiatric disorders. *Mol Psychiatry* (in press).
19 Manolio TA, Rodriguez LL, Brooks L, Abecasis G, Ballinger D, Daly M *et al.* New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet* 2007; **39**: 1045–1051.
20 Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007; **39**: 1181–1186.
21 Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G *et al.* Replicating genotype–phenotype associations. *Nature* 2007; **447**: 655–660.
22 Penninx B, Beekman A, Smit J. The Netherlands Study of Depression and Anxiety (NESDA): rationales, objectives and methods. *Int J Methods Psychiatr Res* 2008; **17**: 121–140.
23 Boomsma DI, de Geus EJ, Vink JM, Stubbe JH, Distel MA, Hottenga JJ *et al.* Netherlands Twin Register: from twins to twin families. *Twin Res Hum Genet* 2006; **9**: 849–857.
24 Boomsma DI, Willemsen G, Sullivan PF, Heutnik P, Meijer P, Sondervan D *et al.* Genome-wide association of major depression: Description of samples for the GAIN major depressive disorder study: NTR and NESDA Biobank Projects. *Eur J Hum Genet* 2008; **16**: 335–342.
25 Bijl RV, van Zessen G, Ravelli A, de Rijk C, Langendoen Y. The Netherlands Mental Health Survey and Incidence Study (NEMESIS): objectives and design. *Soc Psychiatry Psychiatr Epidemiol* 1998; **33**: 581–586.
26 Landman-Peeters KM, Hartman CA, van der Pompe G, den Boer JA, Minderaa RB, Ormel J. Gender differences in the relation between social support, problems in parent–offspring communication, and depression and anxiety. *Soc Sci Med* 2005; **60**: 2549–2559.
27 World Health Organization. *Composite International Diagnostic Interview (CIDI), Version 2.1*. World Health Organization: Geneva, Switzerland, 1997.
28 Boomsma DI, Beem AL, van den Berg M, Dolan CV, Koopmans JR, Vink JM *et al.* Netherlands twin family study of anxious depression (NETSAD). *Twin Res* 2000; **3**: 323–334.
29 Kessler RC, Barker PR, Colpe LJ, Epstein JF, Gfroerer JC, Hiripi E *et al.* Screening for serious mental illness in the general population. *Arch Gen Psychiatry* 2003; **60**: 184–189.
30 Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychol Med* 1996; **26**: 477–486.
31 Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
32 Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* 2005; **307**: 1072–1079.
33 Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74**: 106–120.
34 Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
35 Bierut LJ, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF *et al.* Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet* 2007; **16**: 24–35.
36 Sullivan PF, Lin D, Tzeng JY, van den Oord EJCG, Perkins D, Stroup TS *et al.* Genomewide association for schizophrenia in the CATIE study: results of Stage 1. *Mol Psychiatry* 2008; **13**: 570–584.
37 Hemminger BM, Saelim B, Sullivan PF. TAMAL: An integrated approach to choosing SNPs for genetic studies of human complex traits. *Bioinformatics* 2006; **22**: 626–627.

38 Wittke-Thompson JK, Pluzhnikov A, Cox NJ. Rational inferences about departures from Hardy–Weinberg equilibrium. *Am J Hum Genet* 2005; **76**: 967–986.
39 Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy–Weinberg equilibrium. *Am J Hum Genet* 2005; **76**: 887–893.
40 Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006; **7**: 781–791.
41 Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics* 1997; **53**: 1253–1261.
42 Fisher RA. *Statistical Methods for Research Workers*, 11th edn. Oliver and Boyd: London, 1950.
43 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* 2007; **81**: 559–575.
44 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.
45 Storey JD. The positive false discovery rate: a Bayesian interpretation and the *q*-value. *Ann Stat* 2003; **31**: 2013–2035.
46 Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003; **100**: 9440–9445.
47 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc (Ser B)* 1995; **57**: 289–300.
48 Brown BW, Russell K. Methods of correcting for multiple testing: operating characteristics. *Stat Med* 1997; **16**: 2511–2528.
49 Fernando RL, Nettleton D, Southey BR, Dekkers JC, Rothschild MF, Soller M. Controlling the proportion of false positives in multiple dependent tests. *Genetics* 2004; **166**: 611–619.
50 van den Oord EJ, Sullivan PF. A framework for controlling false discovery rates and minimizing the amount of genotyping in the search for disease mutations. *Hum Hered* 2003; **56**: 188–199.
51 Tsai CA, Hsueh HM, Chen JJ. Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics* 2003; **59**: 1071–1081.
52 van den Oord EJ. Controlling false discoveries in candidate gene studies. *Mol Psychiatry* 2005; **10**: 230–231.
53 Sabatti C, Service S, Freimer N. False discovery rate in linkage and association genome screens for complex disorders. *Genetics* 2003; **164**: 829–833.
54 Meinhausen N, Rice J. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann Stat* 2006; **34**: 373–393.
55 van den Oord EJ, Sullivan PF. False discoveries and models for gene discovery. *Trends Genet* 2003; **19**: 537–542.
56 Lin DY, Hu Y, Huang BE. Simple and efficient analysis of disease association with missing genotype data. *Am J Hum Genet* 2008; **82**: 444–452.
57 Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol* 2002; **155**: 478–484.
58 Gauderman WJ. Sample size requirements for matched case–control studies of gene-environment interaction. *Stat Med* 2002; **21**: 35–50.
59 SAS Institute Inc.. SAS/STAT® Software: Version 9. SAS Institute Inc.: Cary, NC, 2004.
60 R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2007.
61 Lin DY, Zeng D, Millikan R. Maximum likelihood estimation of haplotype effects and haplotype–environment interactions in association studies. *Genet Epidemiol* 2005; **29**: 299–312.
62 Zeng D, Lin DY, Avery CL, North KE, Bray MS. Efficient semiparametric estimation of haplotype–disease associations in case–cohort and nested case–control studies. *Biostatistics* 2006; **7**: 486–502.
63 Huang B, Amos C, Lin D. Detecting haplotype effects in genome-wide association studies. *Genet Epidemiol* 2007; **31**: 803–812.
64 Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263–265.
65 SAS Institute Inc. *JMP User's Guide (Version 6)*. SAS Institute Inc.: Cary, NC, 2005.
66 Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2006; **34**(Database issue): D173–D180.
67 Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* 2006; **34**(Database issue): D590–D598.
68 Blaschke RJ, Rappold G. The pseudoautosomal regions, SHOX and disease. *Curr Opin Genet Dev* 2006; **16**: 233–239.
69 Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
70 Baum AE, Akula N, Cabanero M, Cardona I, Corona W, Klemens B *et al.* A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry* 2007; **13**: 197–207.
71 Phillips GR, Huang JK, Wang Y, Tanaka H, Shapiro L, Zhang W *et al.* The presynaptic particle web: ultrastructure, composition, dissolution, and reconstitution. *Neuron* 2001; **32**: 63–77.
72 Schildkraut JJ. The catecholamine hypothesis of affective disorders: a review of supporting evidence. *Am J Psychiatry* 1965; **122**: 509–522.
73 Wang X, Kibschull M, Laue MM, Lichte B, Petrasch-Parwez E, Kilimann MW. Aczonin, a 550-kD putative scaffolding protein of presynaptic active zones, shares homology regions with Rim and Bassoon and binds profilin. *J Cell Biol* 1999; **147**: 151–162.
74 Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* 2007; **449**: 913–918.
75 Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y *et al.* Detection of large-scale variation in the human genome. *Nat Genet* 2004; **36**: 949–951.
76 Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. *Hum Mol Genet* 2007; **16**(Spec No. 2): R168–R173.
77 Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007; **17**: 1665–1674.
78 Levinson DF, Zubenko GS, Crowe RR, DePaulo RJ, Scheftner WS, Weissman MM *et al.* Genetics of recurrent early-onset depression (GenRED): design and preliminary clinical characteristics of a repository sample for genetic linkage studies. *Am J Med Genet B Neuropsychiatr Genet* 2003; **119**: 118–130.
79 Sun L, Bull S. Reduction of selection bias in genomewide genetic studies by resampling. *Genet Epidemiol* 2005; **28**: 352–367.
80 de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nat Genet* 2005; **37**: 1217–1223.
81 Lopez-Leon S, Janssens AC, Gonzalez-Zuloeta Ladd AM, Del-Favero J, Claes SJ, Oostra BA *et al.* Meta-analyses of genetic studies on major depressive disorder. *Mol Psychiatry* 2007; **13**: 772–785.
82 Frayling TM. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat Rev Genet* 2007; **8**: 657–662.
83 Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007; **316**: 1341–1345.
84 Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007; **316**: 1331–1336.
85 Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008; **40**: 638–645.
86 Sudhof TC. Neurotransmitter release. *Handb Exp Pharmacol* 2008; **184**: 1–21.
87 Shildkraut JJ. The catecholamine hypothesis of affective disorders: a review of the supporting evidence. *Am J Psychiatry* 1965; **122**: 509–522.

88 Garcia J, Gerber SH, Sugita S, Sudhof TC, Rizo J. A conformational switch in the Piccolo C2A domain regulated by alternative splicing. *Nat Struct Mol Biol* 2004; **11**: 45–53.
89 Gerber SH, Garcia J, Rizo J, Sudhof TC. An unusual C(2)-domain in the active-zone protein piccolo: implications for Ca(2+) regulation of neurotransmitter release. *EMBO J* 2001; **20**: 1605–1619.
90 Nabi R, Zhong H, Serajee FJ, Huq AH. No association between single nucleotide polymorphisms in DLX6 and Piccolo genes at 7q21-q22 and autism. *Am J Med Genet B Neuropsychiatr Genet* 2003; **119B**: 98–101.
91 Zollner S, Pritchard JK. Overcoming the winner's curse: estimating penetrance parameters from case–control data. *Am J Hum Genet* 2007; **80**: 605–615.
92 Ghosh A, Zou F, Wright FA. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *Am J Hum Genet* 2008; **82**: 1064–1074.
93 Patten SB. Selection bias in studies of major depression using clinical subjects. *J Clin Epidemiol* 2000; **53**: 351–357.
94 Galbaud du Fort G, Newman SC, Bland RC. Psychiatric comorbidity and treatment seeking. Sources of selection bias in the study of clinical populations. *J Nerv Ment Dis* 1993; **181**: 467–474.
95 Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bull* 1946; **2**: 47–53.
96 Sullivan PF, Joyce PR. Effects of exclusion criteria in depression treatment studies. *J Affect Disord* 1994; **32**: 21–26.
97 Sullivan PF, Wells JE, Joyce PR, Bushnell JA, Mulder RT, Oakley-Browne MA. Family history of depression in clinic and community samples. *J Affect Disord* 1996; **40**: 159–168.
98 Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. The lifetime history of major depression in women: reliability of diagnosis and heritability. *Arch Gen Psychiatry* 1993; **50**: 863–870.
99 McGuffin P, Katz R, Watkins S, Rutherford J. A hospital-based twin register of the heritability of DSM-IV unipolar depression. *Arch Gen Psychiatry* 1996; **53**: 129–136.
100 Shaw CJ, Lupski JR. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum Mol Genet* 2004; **13**(Spec No. 1): R57–R64.
101 Konneker T, Barnes T, Furberg H, Losh M, Bulik CM, Sullivan PF. A searchable database of genetic evidence for psychiatric disorders. *Am J Med Genet (Neuropsychiatr Genet)* 2008; **147**: 671–675.
102 WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661–678.
103 van den Oord EJ, Kuo PH, Hartmann AM, Webb BT, Moller HJ, Hettema JM *et al.* Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. *Arch Gen Psychiatry* 2008; **65**: 1062–1071.
104 O'Donovan M, Craddock N, Norton N, Williams H, Peirce T, Moskvina V *et al.* Identification of novel schizophrenia loci by genome-wide association and follow-up. *Nat Genet* 2008, Jul 30 e-pub ahead of print.
105 Ferreira M, O'Donovan M, Meng Y, Jones I, Ruderfer D, Jones L *et al.* Collaborative genome-wide association analysis of 10,596 individuals supports a role for Ankyrin-G (*ANK3*) and the alpha-1C subunit of the L-type voltage-gated calcium channel (*CACNA1C*) in bipolar disorder. *Nat Genet* 2008, Aug 17 e-pub ahead of print.
106 Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P *et al.* Large-scale copy number polymorphism in the human genome. *Science* 2004; **305**: 525–528.
107 Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *New Engl J Med* 2008; **358**: 667–675.
108 Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 2008; **320**: 539–543.

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