Genome-wide association for major depressive disorder

Citation for published version:

Digital Object Identifier (DOI):
10.1038/mp.2008.125

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Molecular Psychiatry

Publisher Rights Statement:
NIH Public Access Author Manuscript

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Genomewide Association for Major Depressive Disorder: A possible role for the presynaptic protein Piccolo

Patrick F. Sullivan, MD,
University of North Carolina, Chapel Hill pfsulliv@med.unc.edu

Eco J.C. de Geus, PhD,
VU University Amsterdam eco@psy.vu.nl

Gonneke Willemsen, PhD,
VU University Amsterdam ahm.willemsen@psy.vu.nl

Michael R. James, PhD,
Queensland Institute for Medical Research michael.james@qimr.edu.au

Jan H. Smit, PhD,
VU University Medical Center Amsterdam jh.smit@ggzba.nl

Tim Zandbelt, MSc,
VU University Medical Center Amsterdam timz@ggzba.nl

Volker Arolt, MD,
University of Münster volker.arolt@ukmuenster.de

Bernhard T. Baune, MD MPH,
James Cook University Queensland bernhard.baune@jcu.edu.au

Correspond with Dr. Sullivan, Dept. of Genetics, CB#7264, 4109D Neurosciences Research Building, Univ. of North Carolina, Chapel Hill, NC, 27599-7264, USA. Voice: +919-966-3358, FAX: +919-966-3630. E-mail: pfsulliv@med.unc.edu.

†These authors contributed equally.

Financial Disclosures (Prior 3 Years)
Dr. Baune has received honoraria for educational training of psychiatrists and general practitioners from Lundbeck, Astra Zeneca and Pfizer Pharmaceuticals and travel grants from Astra Zeneca, Bristol-Meyer 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, Loxe Pharma, Novartis, Organon, Shire, Eli Lilly & Co., Pfizer, veterinary, 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 Pharmaceuticals, Jazz Pharmaceuticals, J & J Pharmaceuticals, Knoll Pharmaceutical Company, Loxe Pharmaceuticals, Lundbeck, MedAvante, Inc., Merck, NeuroSciences, Novartis, Nutrition 21, Organon Inc., PamLab, LLC, Pfizer Inc, PharmaStar, Pharmavite, Precision Human Biobehavioral, Roche, Sanofi-Aventis, Sepracor, Solvay Pharmaceuticals Inc., Somaxin, 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 SPCCD and for a combination of zainapirone and bumipropion in MDD, copyright royalties for the MGH CEPFQ, DESS, and SAFER. Dr. Nolen has received: speaking fees from Astra Zeneca, Eli Lilly, Pfizer, Servier, Wyeth; unrestricted research funding from AstraZeneca, Eli Lilly, GlaxoSmithKline, 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.
Douglas Blackwood, MD,
University of Edinburgh d.blackwood@ed.ac.uk

Sven Cichon, PhD,
University of Bonn sven.cichon@uni-bonn.de

William L. Coventry,
University of New England wcovntr@une.edu.au

Katharina Domschke, MD MA,
University of Münster katharina.domschke@ukmuenster.de

Anne Farmer, MD PhD,
Institute of Psychiatry a.farmer@iop.kcl.ac.uk

Maurizio Fava, MD,
Harvard Medical School mfava@partners.org

Scott D. Gordon,
Queensland Institute for Medical Research scott.gordon@qimr.edu.au

Qianchuan He,
University of North Carolina, Chapel Hill heqianch@email.unc.edu

Andrew Heath, PhD,
Washington University, St. Louis heatha@psychiatry.wustl.edu

Peter Heutink, PhD,
VU University Medical Center Amsterdam p.heutink@vumc.nl

Florian Holsboer, MD PhD,
Max-Planck Institute of Psychiatry holsboer@mpipsykl.mpg.de

Witte J. Hoogendijk, MD PhD,
VU University Medical Center Amsterdam witteh@ggzba.nl

Jouke Jan Hottenga, PhD,
VU University Amsterdam jj.hottenga@psy.vu.nl

Yijuan Hu,
University of North Carolina, Chapel Hill yhu@bios.unc.edu

Martin Kohli, MSc,
Max-Planck Institute of Psychiatry kohlim@mpipsykl.mpg.de

Danyu Lin, PhD,
University of North Carolina, Chapel Hill lin@bios.unc.edu

Suzanne Lucae, MD PhD,
Max-Planck Institute of Psychiatry lucae@mpipsykl.mpg.de

Donald J. MacIntyre, MRCPsych,
Royal Edinburgh Hospital donald.macintyre@nhslothian.scot.nhs.uk

Wolfgang Maier, MD,
University of Bonn wolfgang.maier@ukb.uni-bonn.de

Kevin A. McGhee, PhD,
University of Edinburgh kevin.mcghee@ed.ac.uk

Peter McGuffin, MD PhD,
Institute of Psychiatry p.mcguffin@iop.kcl.ac.uk
Grant Montgomery, PhD,  
Queensland Institute for Medical Research grant.montgomery@qimr.edu.au  
Walter J. Muir, FRCPsych,  
University of Edinburgh walter.muir@ed.ac.uk  
Willem Nolen, MD,  
University Medical Center Groningen w.a.nolen@psy.umcg.nl  
Markus M. Nöthen, MD,  
University of Bonn markus.noethen@uni-bonn.de  
Roy H. Perlis, MD MSc,  
Harvard Medical School rperlis@chgr.mgh.harvard.edu  
Katrina Pirlo,  
Institute of Psychiatry katrina.pirlo@iop.kcl.ac.uk  
Danielle Posthuma, PhD,  
VU University Amsterdam danielle@psy.vu.nl  
Marcella Rietschel, MD PhD,  
University of Heidelberg marcella.rietschel@zi-mannheim.de  
Patrizia Rizzu, PhD,  
VU University Medical Center Amsterdam p.rizzu@vumc.nl  
Alexandra Schosser, MD,  
Institute of Psychiatry alexandra.schosser@iop.kcl.ac.uk  
August B. Smit, PhD,  
VU University Amsterdam guus.smit@falw.vu.nl  
Jordan W. Smoller, MD ScD,  
Harvard Medical School jordan_smoller@hms.harvard.edu  
Jung-Ying Tzeng, PhD,  
North Carolina State University jytzeng@stat.ncsu.edu  
Richard van Dyck, MD PhD,  
VU University Medical Center Amsterdam r.van.dyck@ggzba.nl  
Matthijs Verhage, PhD,  
VU University Amsterdam matthijs@cnrc.vu.nl  
Frans G. Zitman, MD PhD,  
Leiden University Medical Center f.g.zitman@lumc.nl  
Nicholas G. Martin, PhD,  
Queensland Institute for Medical Research nick.martin@qimr.edu.au  
Naomi R. Wray, PhD,  
Queensland Institute for Medical Research naomi.wray@qimr.edu.au  
Dorret I. Boomsma, PhD†, and  
VU University Amsterdam dorret@psy.vu.nl  
Brenda W.J.H. Penninx, PhD†  
VU University Medical Center Amsterdam b.penninx@vumc.nl

Abstract

Mol Psychiatry. Author manuscript; available in PMC 2009 October 1.
Major depressive disorder (MDD) is a common complex trait with enormous public health significance. As part of the Genetic Association Information Network (GAIN) initiative of the US Foundation for the National Institutes of Health, we conducted a genomewide association study of 435,291 SNPs genotyped in 1,738 MDD cases and 1,802 controls selected to be at low liability for MDD. Eleven of the top 200 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 plays an important role in monoaminergic neurotransmission in the brain) with p-values of $7.7 \times 10^{-7}$ for rs2715148 and $1.2 \times 10^{-6}$ for rs2522833. We undertook replication of SNPs in this region in 5 independent samples (6,079 MDD independent cases and 5,893 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 non-synonymous 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.

**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 (i.e., ≥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 ∼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 other causes (10-13), and substantial direct and indirect costs (14). A WHO 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 genomewide linkage studies (Supplemental 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 genomewide association studies (GWAS) 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.

Mol Psychiatry. Author manuscript; available in PMC 2009 October 1.
**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 via application to the dbGaP repository (20). We have attempted to follow published guidelines for GWAS (21, Box 1).

**Subjects**

The parent projects that supplied subjects for this GWAS 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 via 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 via 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 1,862 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 5,546 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 via 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, std 0.7); controls were required never to have scored highly (≥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 N=1,703 controls 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 0830-0930) 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 0700-1000) 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 re-reviewed for suitability for the deposit of de-identified phenotype and genotype data into the controlled-access dbGaP repository (20). NESDA and NTR subjects were informed of participation in GAIN via 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 multi-marker analyses. The genotyping procedures and genotyping calling algorithms are described in the Supplemental 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 3,840 (=40 plates × 96 samples per plate) or 1,860 cases + 1,860 controls + 80 parents + 20 duplicate samples + 20 HapMap samples.
Quality Control – Subjects

Of the 3,820 Dutch samples sent to Perlegen (excluding the 20 HapMap internal control samples), genotypes were delivered for 3,761 samples. Fifty-nine 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 genomewide homozygosity (~75%), 38 subjects whose genomewide IBS estimates were consistent with first- or second-degree relationships, and 57 additional subjects whose ancestry diverged from the remainder of the sample (see Supplemental Methods for details). After these exclusions (N=104) and removing duplicated and trio quality control sample, there were 3,540 subjects in the final analysis dataset including 1,738 cases and 1,802 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 dataset is included as a Supplemental File (“mddC.fam”).

Quality Control – SNPs

The unfiltered dataset 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 (≥10) than that used by Perlegen was applied. The SNP quality control process is described in detail in the Supplemental Methods. Briefly, to be included in the final analysis dataset, SNPs were required not to have any of the following features: gross mapping problem (37), ≥2 genotype disagreements in 40 duplicated samples, ≥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 (e.g., 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 dataset (included as a Supplemental File, “mddC.bim”). Additional quality control checks are described in the Supplemental Methods). Thirteen 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 females (non-PAR chrX), and those that were hemizygous in males (non-PAR chrX and chrY). All SNPs that passed quality control checks were tested for association with MDD using 1 df 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 females and males 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.
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 Supplemental 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 (i.e., 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 PCLO region and Abecasis’ MACH (v1) to impute 2,037,829 autosomal SNPs with $R^2 \geq 0.5$ (a cutoff that removes ~90% of SNPs with unreliable imputation results while sacrificing 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 $\alpha=1\times10^{-7}$ ($=0.05/500,000$), 1,738 cases and 1,802 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). Pseudo-autosomal region 1 (PAR1) is assumed to be located on chrX:1-2,692,881 and chrY:1-2,692,881 and
Results

Sample Description

Table 1 presents descriptive data for cases and controls. Controls had a higher proportion of males 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 chrX/chrY, 15 SNPs on chrY, and 0 SNPs on PAR2. The median SNP missingness was 0.00339 (inter-quartile range 0.00113-0.0105) and the median minor allele frequency was 0.2422 (inter-quartile range 0.1375-0.3646) with similar estimates in cases and controls. The average marker density over the genome was 1 SNP every 7,069 bases (=3,077,088,087 bases / 435,291 SNPs). The median inter-marker distance was 2,911 bases with interquartile range 966-7,374 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 dataset with case/control status. The estimated lambda (69) was 1.046 (similar p-value minima and lambdas 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 (i.e., if these tests were called significant, over the long-term, a minimum false discovery rate of ~28% would be incurred). As the pre-specified q-value threshold was 0.10, no SNP reached genomewide significance. The proportion of all SNPs without true effects ($p_0$) (54) was conservatively estimated to be $p_0=0.9999954$, consistent with the presence of ~2 SNPs with true effects in these GWAS data. The results of this MDD GWAS are included as a Supplemental File (“results.txt”) to facilitate comparisons with other studies.

Panel a of Figure 1 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. Panel b of Figure 1 shows a plot of $-\log_{10}(p_{trend})$ 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. Four of the top 25 associations are in the presynaptic cytomatrix protein piccolo (PCLO). Table 3 depicts the top 25 multi-SNP clusters (i.e., for an index SNP with association p<0.001, these clusters are additional SNPs within 250 kb of the index SNP with $r^2 \geq 0.50$). The full version of this table is included as a Supplemental File (“Table3_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).

Mol Psychiatry. Author manuscript; available in PMC 2009 October 1.
Focusing on Piccolo (PCLO)

Although no association met genomewide 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 non-synonymous 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 (Supplemental Figure 10), and none of the genotyped SNPs had strong LD (r^2 ≥ 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: (a) 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 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 investigation 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 CNVs with control frequencies of 0.12% and 0.89%) (75, 76 , 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 3-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 signal was from females and from subjects with reoMDD (Supplemental 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 genomewide significance, the presence of multiple possible signals in PCLO and the plausibility of a role for PCLO in the etiology of MDD led us to attempt replication in external samples. We assembled a collection of 11,972 independent subjects (6,079 MDD cases and 5,893 controls) from seven different groups and a total of six case-control replication samples (two German samples were combined, see Supplemental Methods). As with 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 (e.g., 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=9,278). 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 non-independence due to LD (ninth column in Table 4a). Similarly, haplotype analyses did not reveal significantly associated regions (Supplemental 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 ($I^2 \geq 0.4$, ninth column in Table 4a). In investigating this further (Supplemental 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 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 6.4x10^{-8} at the non-synonymous 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 dataset that were specified a priori but which should be considered exploratory.
(a) The network of proteins with which PCLO interacts in its role 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 (i.e., 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.

(b) We imputed genotypes for 2,037,829 autosomal SNPs using MACH with reference to HapMap CEU genotypes. The resulting lambda was 1.048, and the minimum p-value was $1.21 \times 10^{-7}$. 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.

(c) We assembled a list of 103 candidate genes that had been studied for association with MDD in the literature (81). Nineteen of these genes had no SNPs within its transcript and another 9 genes had inadequate coverage (>1 SNP/15 kb; Supplemental Table 17). Of the remaining 75 genes, only NOS1 (neuronal nitric oxide synthase, p=0.0006) had p<0.001. However, NOS1 (as with most genes in Supplemental Table 16) is quite large and chance is a prominent potential influence on these results.

(d) We compared the GWAS association results to a meta-analysis of gene expression data from 12 studies of post-mortem 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 1 alpha 1), 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.

(d) Alternative models, filters, and phenotypes. (i) For reoMDD, the minimum p-value over all GWAS SNPs was at the PCLO region SNP rs2715148 ($8.4 \times 10^{-8}$) 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 ($6.2 \times 10^{-6}$). (iii) Given the female predominance in MDD, we analyzed data from females and males separately. For female cases and controls, rs2715148 had the smallest p-value ($4.0 \times 10^{-7}$) and multiple other PCLO SNPs had p-values in the $10^{-5}$ – $10^{-6}$ range. For males, most PCLO SNPs had p > 0.05 and the minimum was in the SLC9A9 SNP rs4839627 ($9.1 \times 10^{-7}$). (iv) Again, given sex differences in MDD prevalence, we investigated SNPs on chrX and chrY more closely. The minimum p-value in chrX pseudo-autosomal region 1 was 0.02. For the non-PAR regions of chrX in females, the SNPs with the smallest p-values were rs11094388 (p=0.0003, intergenic), rs5971108 (p=0.0003, PITCHDI), 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 males, the minimum p-value on chrX was at rs10521594 (p=5.4×10^{-5}, intergenic) and 0.22 on chrY.
Discussion

Overview

Major depressive disorder (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 genomewide association study of 435,291 SNPs genotyped in 1,738 MDD cases and 1,802 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 genomewide genotyping designed to capture common variation in subjects of European ancestry.

According to our primary analysis plan, no SNP-MDD phenotype association reached genomewide 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 genomewide threshold after correction for multiple comparisons, 11 of the top 200 signals localized to a 167 kb region overlapping the gene piccolo (PCLO). The protein product of PCLO localizes to the presynaptic active zone and plays an important role 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 non-synonymous 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 6,079 MDD independent cases and 5,893 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 over-estimated 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 non-psychiatric 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.

Mol Psychiatry. Author manuscript; available in PMC 2009 October 1.
Joint analysis of the NESDA/NTR and QIMR samples yielded $p=6.4 \times 10^{-8}$ (uncorrected for multiple hypothesis testing) for the non-synonymous SNP rs2522833. This result suggests a specific hypothesis for future studies: an association between genetic variation in \textit{PCLO} and MDD may be detected only in population-based cases. Thus, it would be premature to exclude \textit{PCLO} from a role in the etiology of some forms of MDD.

The Heterogeneous Nature of MDD

Interpretation of the \textit{PCLO} replication efforts is consistent with two broad possibilities. The first possibility is that genetic variation in \textit{PCLO} is truly not associated with MDD. This interpretation is supported by the replication analyses (specified \textit{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 \textit{PCLO} region. The second possibility invokes a less parsimonious model involving heterogeneity, that genetic variation in \textit{PCLO} is etiologically causal to some subtypes of MDD. This interpretation is an \textit{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 \textit{a priori} results and plausible \textit{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 \textit{PCLO} results are accurate – genetic variation in \textit{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 role of such rare variants will require larger samples.

Most of the additional exploratory analyses were unrevealing, including examination of proteins known to interact with \textit{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 SLEP compendium of psychiatric genetics findings (101) to 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 \textit{CFLAR} and \textit{MARCH3} were increased and \textit{LST1} and \textit{HLA-B} were decreased in MDD post-mortem frontal cortex. These regions ranked 9, 232, 267, and 432 in the

\textit{Mol Psychiatry}. Author manuscript; available in PMC 2009 October 1.
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 WTCCC 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 genomewide significance: the minimum p-values in our study for *MAMDC1* (103) was 0.004, 0.03 for *ZNFI804A* (104), 0.002 in *ANK3* (105), and 0.03 in *CACNA1C*’(105). These overlaps are intriguing (although the play of chance cannot be excluded), and will be formally investigated as part of our participation in the Psychiatric GWAS Consortium analyses (18).

**Limitations**

(a) 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. (b) 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 copy number variation (106) that may be particularly salient for psychiatric disorders (107,108). More generally, the GWAS platform might not be sufficiently “genomewide” and 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. (c) There was an imbalance in the proportion of males 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, GWAS 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 role 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.

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**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 genomics studies (480-04-004); twin research focusing on behavior (400-05-717), Center for Medical Systems Biology (NWO Genomics); Spinozapremie (SPL56-464-14192); Centre for Neurogenomics and Cognitive Research (CNCR-VU); genomewide 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 Lentsi, 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 would like to express our thanks to: the GAIN Genotyping group (Dr. Gonçalo Abecasis, chair) for help with quality control; Drs. Gonçalo Abecasis and 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 Drs. Pam Madden (DA012854) and 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.

References


43. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, de Bakker P, Daly M, Sham P. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics. 2007; 81:559–75. [PubMed: 17701901]


102. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447(7145):661–78. [PubMed: 17554300]


Figure 1.
GWAS results for MDD in cases versus controls. Panel a shows the quantile-quantile plots and lambda 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 $\chi^2$ from all association tests divided by the expected value under the null hypothesis of no association. If λ is large (e.g., >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 (e.g., population stratification effects). The λ values in Panel a are not consistent with the presence of systematic biases in the results. Panel b depicts $-\log_{10}(p)$ by genomic location for chr1-chr22 plus chrX. Odd-numbered chromosomes are in light blue and even-numbered chromosomes in orange. The 25 smallest p-values are shown with green crosses.
Figure 2.
Plot of the 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 $-\log_{10}(p)$ for genotyped SNPs (colored diamonds) and imputed SNPs (grey diamonds), and the right y-axis the recombination rate from the HapMap EUR panel (light blue curve). The color of the genotyped SNPs corresponds to LD with the SNP with smallest p-value (rs2715148): red $0.8 \leq r^2 \leq 1.0$, orange $0.5 \leq r^2 < 0.8$, yellow $0.2 \leq r^2 < 0.5$, and white $r^2 < 0.2$. The significant and extent of all 3-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.
Figure 3. PCLO region replication results for MDD showing genomic context and forest plots for the top 12 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.
Table 1

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

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Cases</th>
<th>Controls</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects genotyped</td>
<td>1,738</td>
<td>1,802</td>
<td></td>
</tr>
<tr>
<td>Mean age in years (SD)</td>
<td>42.6 (12.6)</td>
<td>45.1 (14.1)</td>
<td>$F_{1.3538}=31.1$, $p&lt;0.001$</td>
</tr>
<tr>
<td>Female, %</td>
<td>69.6</td>
<td>62.0</td>
<td>$\chi^2=22.5$, $p&lt;0.001$</td>
</tr>
<tr>
<td>Educational level (% low/middle/high)</td>
<td>7.8 / 62.0 / 32.2</td>
<td>5.7 / 56.3 / 38.1</td>
<td>$\chi^2=16.3$, $p&lt;0.001$</td>
</tr>
<tr>
<td>Partner status, % with partner</td>
<td>68.9</td>
<td>87.0</td>
<td>$\chi^2=167.2$, $p&lt;0.001$</td>
</tr>
<tr>
<td>Smoking (current), %</td>
<td>42.0</td>
<td>20.2</td>
<td>$\chi^2=194.5$, $p&lt;0.001$</td>
</tr>
<tr>
<td>Mean neuroticism (NEO, SD)</td>
<td>39.3 (8.0)</td>
<td>28.2 (5.5)</td>
<td>$F_{1.2920}=183.1$, $p&lt;0.001$</td>
</tr>
<tr>
<td>MDD, age of onset in years (SD)</td>
<td>27.7 (12.4)</td>
<td>57.3</td>
<td></td>
</tr>
<tr>
<td>Early age of onset (&lt; 30 years), %</td>
<td>57.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family history of depression, %</td>
<td>85.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recurrent MDD</td>
<td>50.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family history, recurrent MDD, or early age of onset (&lt; 30 years)</td>
<td>94.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2

Information on the SNPs with the smallest association p-values in the GWAS.

<table>
<thead>
<tr>
<th>Basic SNP data</th>
<th>Bioinformatics</th>
<th>Results</th>
<th>MAF</th>
<th>Quality control - SNP missingness</th>
<th>Quality control - additional checks</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP</td>
<td>Chr</td>
<td>Position</td>
<td>Alleles</td>
<td>Strand</td>
<td>Gene</td>
</tr>
<tr>
<td>----------------</td>
<td>-----</td>
<td>----------</td>
<td>---------</td>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>rs12471796</td>
<td>2</td>
<td>201777830</td>
<td>A/G</td>
<td>+</td>
<td>Reg pot</td>
</tr>
<tr>
<td>rs7565124</td>
<td>2</td>
<td>20183313</td>
<td>A/G</td>
<td>+</td>
<td>ALK</td>
</tr>
<tr>
<td>rs3923028</td>
<td>2</td>
<td>29597247</td>
<td>T/C</td>
<td>-</td>
<td>SCZ</td>
</tr>
<tr>
<td>rs2094923</td>
<td>6</td>
<td>14397061</td>
<td>T/G</td>
<td>-</td>
<td>SCZ</td>
</tr>
<tr>
<td>rs1558477</td>
<td>7</td>
<td>30928587</td>
<td>T/G</td>
<td>-</td>
<td>SCZ</td>
</tr>
<tr>
<td>rs2274821</td>
<td>7</td>
<td>30993071</td>
<td>GC</td>
<td>+</td>
<td>SCZ</td>
</tr>
<tr>
<td>SNP</td>
<td>Chr</td>
<td>Position</td>
<td>Alleles</td>
<td>Strand</td>
<td>Gene</td>
</tr>
<tr>
<td>---------</td>
<td>-----</td>
<td>----------</td>
<td>--------------</td>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>rs2715148</td>
<td>7</td>
<td>82,094,696</td>
<td>A/C</td>
<td>+</td>
<td>PCLO</td>
</tr>
<tr>
<td>rs252833</td>
<td>7</td>
<td>82,098,359</td>
<td>C/A</td>
<td>+</td>
<td>PCLO</td>
</tr>
<tr>
<td>rs2528462</td>
<td>7</td>
<td>82,123,066</td>
<td>G/T</td>
<td>+</td>
<td>PCLO</td>
</tr>
<tr>
<td>rs2107849</td>
<td>7</td>
<td>82,200,330</td>
<td>A/T</td>
<td>+</td>
<td>PCLO</td>
</tr>
<tr>
<td>rs1457269</td>
<td>8</td>
<td>24825,757</td>
<td>A/G</td>
<td>−</td>
<td>Reg pot</td>
</tr>
<tr>
<td>rs7005187</td>
<td>8</td>
<td>81,663,211</td>
<td>T/C</td>
<td>+</td>
<td>Com, reg pot</td>
</tr>
<tr>
<td>rs1708489</td>
<td>10</td>
<td>34,297,638</td>
<td>A/G</td>
<td>−</td>
<td>Reg pot</td>
</tr>
<tr>
<td>rs11031682</td>
<td>11</td>
<td>32,242,721</td>
<td>T/C</td>
<td>+</td>
<td>Reg pot</td>
</tr>
<tr>
<td>rs12597382</td>
<td>12</td>
<td>44,019,689</td>
<td>T/C</td>
<td>+</td>
<td>TMEM16F</td>
</tr>
<tr>
<td>rs4765078</td>
<td>12</td>
<td>123,371,707</td>
<td>C/T</td>
<td>+</td>
<td>ZNF333</td>
</tr>
<tr>
<td>rs8023445</td>
<td>15</td>
<td>46,980,083</td>
<td>C/T</td>
<td>+</td>
<td>SHC4</td>
</tr>
<tr>
<td>rs2885179</td>
<td>19</td>
<td>14,688,830</td>
<td>A/C</td>
<td>−</td>
<td>ZNF333</td>
</tr>
<tr>
<td>rs941766</td>
<td>20</td>
<td>39,724,220</td>
<td>A/G</td>
<td>+</td>
<td>Reg pot</td>
</tr>
<tr>
<td>rs12480143</td>
<td>20</td>
<td>39,741,240</td>
<td>G/A</td>
<td>+</td>
<td>Reg pot</td>
</tr>
</tbody>
</table>
## Basic SNP data

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr</th>
<th>Position</th>
<th>Alleles</th>
<th>Strand</th>
<th>Gene</th>
<th>TAMAL (^\d)</th>
<th>SLEP (^\d)</th>
<th>Rank</th>
<th>OR (CI)</th>
<th>P-asymptotic</th>
<th>P-empirical</th>
<th>q-value</th>
<th>P-gwemp</th>
<th>All</th>
<th>Cases</th>
<th>Controls</th>
<th>HapMap_EUR</th>
<th>All</th>
<th>Cases</th>
<th>Controls</th>
<th>p-missing</th>
<th>additional checks</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs928862</td>
<td>21</td>
<td>20559590</td>
<td>G/A</td>
<td>+</td>
<td>Reg pot</td>
<td>Near CNV</td>
<td>19</td>
<td>0.78 (0.69-0.88)</td>
<td>0.000040</td>
<td>0.000044</td>
<td>0.66</td>
<td>1.00</td>
<td>0.190</td>
<td>0.170</td>
<td>0.209</td>
<td>0.167</td>
<td>0.013</td>
<td>0.012</td>
<td>0.023</td>
<td>0.77</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. q-value=see text. P-gwemp=genome-wide empirical p-value via 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.

\(^\d\) TAMAL codes. cSNP=coding SNP. 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, transcription binding site, Alu, LTR, exon, splice site, miRNA expression QTL (lymphocytes or cortex). Only positive flags are shown.

\(^\d\) 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 (15286789), breast and colon cancer mutations (17902254), MDD genomewide linkage studies (12612864, 1582139, 17427203), SCZ genomewide linkage meta-analysis (12802786), and bipolar disorder (BIP) GWAS (17554300).
Table 3
Clustering of SNPs with low p-values.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Nsnp</th>
<th>Pmin</th>
<th>N &lt;0.001</th>
<th>N &lt;0.01</th>
<th>Expressed in brain?</th>
<th>Genes</th>
<th>Gene products</th>
<th>SLEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>30,928,580</td>
<td>30,931,521</td>
<td>3</td>
<td>1.2E-06</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>Yes ADY AP1R1</td>
<td>adenylyl cyclase activating polypeptide 1 (guanylyl) receptor type 1</td>
<td>Neuroactive ligand-receptor interaction</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>82,041,596</td>
<td>82,208,167</td>
<td>10</td>
<td>1.5E-06</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>Yes PCLO</td>
<td>guaful (presynaptic cysotmine proteine)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>14,398,932</td>
<td>14,399,068</td>
<td>2</td>
<td>9.0E-06</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>[LAPTM4A]</td>
<td>lysosomal-associated protein transmembrane 4-alpha</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>20,177,820</td>
<td>20,183,313</td>
<td>2</td>
<td>1.1E-05</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>Yes [SLEP]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>46,979,618</td>
<td>46,980,083</td>
<td>2</td>
<td>1.3E-05</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Yes CI1 EzD1 RalP SHC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>201,794,446</td>
<td>201,880,818</td>
<td>2</td>
<td>2.4E-05</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Yes [SLEP]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>39,724,230</td>
<td>39,742,644</td>
<td>5</td>
<td>4.3E-05</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>14,386,148</td>
<td>14,397,061</td>
<td>3</td>
<td>4.2E-05</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>184,652,436</td>
<td>184,658,003</td>
<td>3</td>
<td>4.2E-05</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>117,174,763</td>
<td>117,282,887</td>
<td>4</td>
<td>4.8E-05</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>19</td>
<td>127,071,672</td>
<td>127,087,021</td>
<td>3</td>
<td>0.00046</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>22,752,668</td>
<td>22,792,155</td>
<td>3</td>
<td>4.6E-05</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>Yes CDH12</td>
<td>cadherin 12, type 2 (N-cadherin 2)</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>22</td>
<td>88,130,196</td>
<td>88,136,792</td>
<td>2</td>
<td>5.5E-05</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Yes ANPEP MESP2</td>
<td>alanyl (microblane) aminopeptidase (aminopeptidase A, aminopeptidase N, aminopeptidase M, microsomal aminopeptidase, CDI, 3,</td>
<td>EP3098,mesodern posterior 2 homolog (mouse)</td>
</tr>
<tr>
<td>14</td>
<td>23</td>
<td>54,098,247</td>
<td>54,102,064</td>
<td>2</td>
<td>4.3E-05</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>24</td>
<td>145,875,183</td>
<td>145,878,794</td>
<td>2</td>
<td>5.4E-05</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>27</td>
<td>32,242,721</td>
<td>32,244,520</td>
<td>2</td>
<td>4.2E-05</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>28</td>
<td>27,249,840</td>
<td>27,379,524</td>
<td>6</td>
<td>5.3E-05</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>Yes AKI26571/CHRNA2/PTK2R</td>
<td>Hypothetical protein FLJ46514/cholinergic receptor, nicotinic alpha 2 (neuroual)/PTK2R protein tyrosine kinase</td>
<td>CHRNA2 causes nocturnal frontal lobe epilepsy (OMIM 115802)</td>
</tr>
<tr>
<td>Rank</td>
<td>Chr</td>
<td>Start</td>
<td>End</td>
<td>Nsnps</td>
<td>Pmin</td>
<td>N &lt;0.001</td>
<td>N &lt;0.01</td>
<td>Expressed in brain?</td>
<td>Genes</td>
<td>Gene products</td>
<td>SLEP ³</td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
<td>---------</td>
<td>---------</td>
<td>-------</td>
<td>-------</td>
<td>----------</td>
<td>---------</td>
<td>---------------------</td>
<td>-------</td>
<td>--------------</td>
<td>--------</td>
</tr>
<tr>
<td>29</td>
<td>3</td>
<td>12,453,817</td>
<td>12,459,985</td>
<td>2</td>
<td>0.00005</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>PPARG</td>
<td>peroxisome proliferator-activated receptor gamma</td>
<td>Type 2 diabetes mellitus risk gene</td>
</tr>
<tr>
<td>32</td>
<td>3</td>
<td>99,975,821</td>
<td>100,183,009</td>
<td>2</td>
<td>7.2E-05</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>DCBLD2 ST3GAL6</td>
<td>discodin, CUB and LCCL domain containing 2/ST3 beta-galactoside alpha-2,3-sialyltransferase 6</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>3</td>
<td>70,481,852</td>
<td>70,476,913</td>
<td>2</td>
<td>8.2E-05</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Expresses</td>
<td>DCBLD2 ST3GAL6</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>2</td>
<td>7,424,098</td>
<td>7,440,754</td>
<td>3</td>
<td>8.9E-05</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Expresses</td>
<td>DCBLD2 ST3GAL6</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>5</td>
<td>54,352,638</td>
<td>54,363,712</td>
<td>3</td>
<td>0.000071</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>GZMK</td>
<td>granzyme K (granzyme 3; tryptase II)</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>13</td>
<td>111,889,281</td>
<td>111,902,203</td>
<td>2</td>
<td>7.8E-05</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Expresses</td>
<td>DCBLD2 ST3GAL6</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>1</td>
<td>211,470,329</td>
<td>211,508,991</td>
<td>4</td>
<td>0.000072</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>Expresses</td>
<td>DCBLD2 ST3GAL6</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>8</td>
<td>24,784,536</td>
<td>24,825,085</td>
<td>2</td>
<td>0.000359</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>NEF3 NEFM</td>
<td>neurofilament 3 (150kDa medium)/neurofilament, medium polypeptide 150kDa</td>
<td></td>
</tr>
</tbody>
</table>

³SLEPs:Sullivan Lab Evidence Project (http://slep.unc.edu) a compendium of genetic findings from the literature.
Table 4

**PCLO replication results.**

<table>
<thead>
<tr>
<th>Gene</th>
<th>N</th>
<th>Z</th>
<th>P</th>
<th>P_corr</th>
<th>I²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>100</td>
<td>0.5</td>
<td>0.05</td>
<td>0.02</td>
<td>0.5</td>
</tr>
<tr>
<td>Gene 2</td>
<td>200</td>
<td>1.0</td>
<td>0.01</td>
<td>0.005</td>
<td>0.2</td>
</tr>
<tr>
<td>Gene 3</td>
<td>300</td>
<td>1.5</td>
<td>0.005</td>
<td>0.001</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Abbreviations: N=total sample size for an analysis, Z=logistic regression beta divided by its standard error, P=asymptotic p-value from Wald chi-square test (1 df) uncorrected for multiple comparisons, P_corr=empirical p-value from accounting for multiple comparisons and LD structure (50K permutations), and I²=i-squared, an index of heterogeneity of logistic regression parameter estimates.