Genome-wide association study identifies five new schizophrenia loci

Ripke, Stephan; Sanders, Alan R.; Kendler, Kenneth S.; Levinson, Douglas F.; Sklar, Pamela; Holmans, Peter A.; Lin, Dan-Yu; Duan, Jubao; Ophoff, Roel A.; Andreassen, Ole A.

Published in:
Nature Genetics

DOI:
10.1038/ng.940

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2011

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 22-12-2018
Genome-wide association study identifies five new schizophrenia loci

The Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium

We examined the role of common genetic variation in schizophrenia in a genome-wide association study of substantial size: a stage 1 discovery sample of 21,856 individuals of European ancestry and a stage 2 replication sample of 29,839 independent subjects. The combined stage 1 and 2 analysis yielded genome-wide significant associations with schizophrenia for seven loci, five of which are new (1p21.3, 2q32.3, 8p23.2, 8q21.3 and 10q24.32-q24.33) and two of which have been previously implicated (6p21.32-p22.1 and 18q21.2). The strongest new finding ($P = 1.6 \times 10^{-11}$) was with rs1625579 within an intron of a putative primary transcript for MIR137 (microRNA 137), a known regulator of neuronal development. Four other schizophrenia loci achieving genome-wide significance contain predicted targets of MIR137, suggesting MIR137-mediated dysregulation as a previously unknown etiologic mechanism in schizophrenia.

In a joint analysis with a bipolar disorder sample (16,374 affected individuals and 14,044 controls), three loci reached genome-wide significance: CACNA1C (rs4765905, $P = 7.0 \times 10^{-8}$), ANK3 (rs10994359, $P = 2.5 \times 10^{-9}$) and the ITIH3-ITIH4 region (rs2239547, $P = 7.8 \times 10^{-9}$).

In stage 1, we conducted a mega-analysis combining genome-wide association study (GWAS) data from 17 separate studies (with a total of 9,394 cases and 12,462 controls), three principal components as covariates to minimize inflation in significance test of imputed dosages with sample identifiers and three principal components. We imputed allelic dosages for 1,252,901 autosomal SNPs (9,394 cases and 12,462 controls), affected individuals, and 14,044 controls; Table 1 and Supplementary Tables 1.2). We imputed allelic dosages for 1,252,901 autosomal SNPs (Table 1, Supplementary Table 3 and Supplementary Note) using HapMap3 as the reference panel. We tested for association using logistic regression of imputed dosages with sample identifiers and three principal components as covariates to minimize inflation in significance testing caused by population stratification. The quantile-quantile plot (Supplementary Fig. 1) deviated from the null distribution with a population stratification inflation factor of $\lambda = 1.23$. However, $\lambda_{1000}$, a metric that standardizes the degree of inflation by sample size, was only 1.02, similar to that observed in other GWAS meta-analyses. This deviation persisted despite comprehensive quality control and inclusion of up to 20 principal components (Supplementary Fig. 1). Thus, we interpret this deviation as indicative of a large number of weakly associated SNPs consistent with polygenic inheritance. We also examined 298 ancestry-informative markers (AIMs) that reflect European-ancestry population substructure. Unadjusted analyses showed greater inflation in the test statistics than we saw for all markers (AIMs $\lambda = 2.26$ compared to all markers $\lambda = 1.56$). After inclusion of principal components, the distributions of the test statistics did not differ between AIMs ($\lambda = 1.18$) and all markers ($\lambda = 1.23$), a result inconsistent with population stratification explaining the residual inflation seen in Supplementary Figure 1. Moreover, the results of a meta-analysis using summary results generated using study specific principal components (Supplementary Note) were highly correlated with those from the mega-analysis (Pearson correlation = 0.94, with a similar $\lambda = 1.20$; Supplementary Fig. 2). Of the ten SNPs in Table 2, four increased and six decreased in significance, suggesting that the most extreme values did not result from systematic inflation artifacts. Therefore, our primary analysis used unadjusted $P$ values (nevertheless, see Table 2 for stage 1 $P$ values adjusted for $\lambda$ (ref. 6)).

In stage 1 (Table 2, Supplementary Table 4 and Supplementary Figs. 3 and 4), 136 associations reached genome-wide significance ($P < 5 \times 10^{-8}$). The majority of these associations ($N = 129$) mapped to 5.5 Mb in the extended major histocompatibility complex (MHC, 6p21.32-p22.1), a region of high linkage disequilibrium (LD) previously implicated in schizophrenia in a subset of the samples used here. The other stage 1 regions included new regions (10q24.33 and 8q21.3) and previously reported regions (18q21.2 at TCF4 (encoding transcription factor 4) and 11q24.2 (ref. 8)). The signal at 11q24.2 is ~0.85 Mb from NRGN (encoding neurogranin) and is uncorrelated with the previously associated variant near this gene.

In Table 2 and Supplementary Table 4, we denote regions of association by the most significant marker. Associated SNPs with $r^2 \geq 0.2$ in HapMap3 (CEU+TSI populations) were not considered independent. However, we noticed instances where multiple SNPs within 250 kb of each other yielded evidence for association ($P < 10^{-5}$) despite weak LD ($r^2 < 0.2$) between them. For regions with $P < 10^{-6}$, we performed a conditional analysis using as covariates the dosages of the strongest associated SNP, principal components 1–4 and 6 and study indicator. We observed multiple statistically independent signals at the MHC. Although a number of SNPs within the MHC were potentially independent per HapMap $r^2$ values, only rs9272105 withstood formal conditional analysis, showing $P = 1.8 \times 10^{-6}$ conditional on association to the best SNP, rs2021722 (stage 1 $P = 4.3 \times 10^{-11}$, inter-SNP distance = 2.4 Mb, $r^2 = 0.01$ in HapMap). Excluding the MHC region, we identified six regions with at least one SNP associated at $P < 10^{-5}$ and a second SNP with a conditionally independent $P < 10^{-3}$.

1A full list of authors and affiliations appears at the end of the paper.

Received 16 February; accepted 19 August; published online 18 September 2011; doi:10.1038/ng.940
We performed 100 simulations after permuting case-control status randomly within each study. In contrast to the six regions in the real dataset, we never observed more than one region with co-localized statistically independent signals in any simulated genome-wide scan, indicating our observation is highly unlikely to have occurred by chance.

Noteworthy co-localizing independent signals occurred at three regions (Supplementary Table 5): one region with a genome-wide significant association at 10q24.32-q24.33 (Table 2), a second region that nearly met this threshold at MAD1LI (encoding mitotic arrest deficient-like 1; rs10226475, \( P = 5.06 \times 10^{-8} \), Supplementary Table 4) and a third region at CACNA1C (encoding calcium channel, voltage-dependent, L type, \( \alpha 1C \) subunit), the latter of which has previously been associated with bipolar disorder\(^{10}\) and other psychiatric phenotypes including schizophrenia\(^{11}\). The conditionally independent signal at CACNA1C was more significant than any observation made in 100 permutations of the entire experiment (both conditional \( P < 10^{-5} \)) and supports CACNA1C in schizophrenia after genome-wide correction (\( P < 0.01 \), even without considering these prior reports).

In stage 2, we evaluated in 29,839 independent subjects (8,442 cases and 21,397 controls) the most significant SNPs (\( N = 81 \)) in each LD region where at least one SNP had surpassed \( P < 2 \times 10^{-5} \).
### Table 2 Top genome-wide association results for schizophrenia

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr.</th>
<th>Mb</th>
<th>Alleles</th>
<th>Frequency</th>
<th>( P ) (GC-adjusted ( P ))</th>
<th>OR (95% CI)</th>
<th>Consistency of direction</th>
<th>Gene</th>
<th>Distance (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1625579</td>
<td>1p21.3</td>
<td>98.3</td>
<td>TG</td>
<td>0.80</td>
<td>5.72 ( \times 10^{-7} ) (6.52 ( \times 10^{-6} ))</td>
<td>1.14 (1.08-1.19)</td>
<td>++++++ MIR137 Intragenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs17662626</td>
<td>2q32.3</td>
<td>193.7</td>
<td>AG</td>
<td>0.91</td>
<td>1.59 ( \times 10^{-11} ) (6.87 ( \times 10^{-10} ))</td>
<td>1.12 (1.09-1.16)</td>
<td>+ +++++ PCGEM1 343</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2021722</td>
<td>6p21.3-p22.1</td>
<td>30.3</td>
<td>CT</td>
<td>0.78</td>
<td>1.70 ( \times 10^{-3} ) (n.a.)</td>
<td>1.16 (1.06-1.27)</td>
<td>+ + +++ TCF4 Intragenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs10503253</td>
<td>8p23.2a</td>
<td>4.2</td>
<td>AC</td>
<td>0.19</td>
<td>1.55 ( \times 10^{-3} ) (n.a.)</td>
<td>1.10 (1.03-1.17)</td>
<td>+ + ++++ TRIM26 Intragenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7004633</td>
<td>8q21.3</td>
<td>89.8</td>
<td>GA</td>
<td>0.18</td>
<td>1.07 ( \times 10^{-3} ) (n.a.)</td>
<td>1.08 (1.01-1.14)</td>
<td>+ + ++++ CSMD1 Intragenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7914558</td>
<td>10q24.32</td>
<td>104.8</td>
<td>GA</td>
<td>0.59</td>
<td>1.82 ( \times 10^{-9} ) (3.11 ( \times 10^{-8} ))</td>
<td>1.10 (1.07-1.13)</td>
<td>++++++ NT5C2 Intragenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs11191580</td>
<td>10q24.33</td>
<td>104.9</td>
<td>TC</td>
<td>0.91</td>
<td>2.23 ( \times 10^{-8} ) (4.58 ( \times 10^{-7} ))</td>
<td>1.22 (1.15-1.29)</td>
<td>+ +++++ CNNM2 Intragenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs548181</td>
<td>11q24.2</td>
<td>125.0</td>
<td>GA</td>
<td>0.88</td>
<td>5.09 ( \times 10^{-3} ) (n.a.)</td>
<td>1.09 (1.02-1.16)</td>
<td>+ ++++++ MMP16 421</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs12966547</td>
<td>18q21.2</td>
<td>50.9</td>
<td>GA</td>
<td>0.58</td>
<td>1.82 ( \times 10^{-9} ) (3.11 ( \times 10^{-8} ))</td>
<td>1.15 (1.10-1.20)</td>
<td>+ ++++++ STT3A 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs17512836</td>
<td>18q21.2</td>
<td>51.3</td>
<td>CT</td>
<td>0.02</td>
<td>2.29 ( \times 10^{-5} ) (n.a.)</td>
<td>1.08 (1.04-1.12)</td>
<td>+ ++++++ CCDC68 126</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The SNPs listed are those with a stage 1 \( P < 5 \times 10^{-8} \) and/or a combined stage 1 and 2 \( P < 5 \times 10^{-8} \). These ten independent (\( P < 0.2 \)) SNPs represent eight physically distinct genomic loci, as there are two SNPs listed for two loci (10q24.32-24.33 and 18q21.2). For the MHC region, only one SNP is listed for clarity. The eight susceptibility loci represent (Supplementary Table 6) in the mega-analysis. Of 22 SNPs from the MHC, 5 surpassed the genome-wide significant threshold in stages 1 and 2 combined (minimum \( P = 2.2 \times 10^{-12} \) at rs2021722; Supplementary Table 6). Excluding the MHC region, a sign test for consistency between stages 1 and 2 was highly significant (\( P < 10^{-6} \)), with the same direction of effect as observed stage 1 also being observed in stage 2 for 49 of 59 SNPs. A Fisher's combined test revealed the distribution of stage 2 \( P \) values was unlikely to have occurred by chance (\( P < 10^{-12} \)).

(Figs. 1.2, Table 2, Supplementary Tables 6.7 and Supplementary Fig. 4). After adjusting for \( \lambda \) (ref. 6), four loci (1p21.3, 6p21.3-p22.1, 10q24.32-q24.33 and 18q21.2) remained significant at \( P \leq 5 \times 10^{-8} \). For the primary analyses (unadjusted for \( \lambda \)), the strongest new association was at 1p21.3 (rs1625579; \( P = 1.6 \times 10^{-12} \)), which is over 100 kb from any RefSeq protein-coding gene but is within intron 3 of AK094607, which contains the primary transcript for MIR137 (ref. 13). The next best locus, 10q24.32 (Supplementary Table 5 and Supplementary Fig. 5), has independent associations 130 kb apart at rs7914558 and rs11191580 (\( P = 1.1 \times 10^{-8} \)), implicating a 0.5-Mb region containing multiple genes (Supplementary Fig. 5). The third best locus, rs7004633 (\( P = 2.8 \times 10^{-9} \)) on 8q21.3, is 400 kb from the nearest gene (MMP16, encoding matrix metalloproteinase 16). The fourth best locus, rs10503253 (\( P = 4.4 \times 10^{-8} \)) at 8p23.2, is in an intron of CSMD1 (encoding CUB and Sushi multiple domains 1). Finally, rs17512836 (\( P = 4.7 \times 10^{-8} \)) at 2q32.3 is intergenic, mapping 300 kb from a non-coding RNA, PCGEM1 (prostate-specific transcript 1).14.
**LETTERS**

**Figure 1** Manhattan plot for stages 1 and 2. Standard \(-\log_{10} P\) plot of the study results. For the stage 1 results, 16 regions with one or more SNP achieving \(P < 10^{-8}\) are highlighted in color and labeled with the name of the nearest gene. SNPs selected for stage 2 replication are highlighted, with the resulting combined \(P\) value after replication (that is, after incorporation of stage 2 results) indicated by the large diamonds. Blue highlighting indicates SNPs that were less significantly associated after replication, and pink highlighting indicates SNPs that were more significantly associated after replication.

*MIR137* has been implicated in regulating adult neurogenesis\(^{15,16}\) and neuronal maturation\(^{17}\), mechanisms through which variation at this locus could contribute to brain development abnormalities in schizophrenia. Of relevance, two independent schizophrenia imaging studies found *MIR137* to be one of three microRNAs with targets significantly enriched for association\(^{18}\). In stage 1, SNPs in or near 301 high-confidence predicted *MIR137* targets (with a TargetScan\(^{19}\) probability of conserved targeting 20.9) were enriched for association compared with genes matched for size and marker density: 17 predicted *MIR137* targets (Supplementary Table 8) had at least one SNP with

Figure 2 Regional association plots for five new schizophrenia loci. Regional \(P\) value plots for each of the five new schizophrenia loci: 1p21.3, 2q32.3, 8p23.2, 8q21.3 and 10q24.32-q24.33. Each plot shows the most associated SNP (key SNP) and its genomic region from the first column of Table 2: stage 1 scan results for each SNP ± 200 kb to the key SNP are shown. On the x axis is the genomic position, and on the y axis is \(-\log_{10} P\). Larger SNP symbols indicate higher LD (based on HapMap 3 data) to the key SNP than smaller SNP symbols. Color coding (from red to blue) denotes LD information; see also the legend within the plot.


**LETTERS**

P < 10^{-4}, which is more than twice as many as the control gene sets (P < 0.01). Excluding the MHC and MIR137, of the nine loci with genome-wide significant support either in stage 1 or in the combined set (six loci, 2q32.3, 8p23.2, 8q21.3, 10q24.32–q24.33, 11q24.2 and 18q21.2; Table 2 and Supplementary Tables 6.7) or in a joint analysis with bipolar disorder (three genes, CACNA1C, ANK3 and ITIH3-ITIH4, described below), four genes (TCF4, CACNA1C, CSMD1 and C10orf26) have predicted MIR137 target sites according to analyses using three different prediction programs (TargetScan\(^1\), PicTar\(^2\) and miRanda\(^3\)). In vitro overexpression and locked nucleic acid–mediated knockdown of MIR137 in neuronal cell line N2a leads to changes in expression levels of TCF4 protein, strongly supporting the prediction that TCF4 is a target of MIR137 (L.-H. Tsai, personal communication). Our observations suggest MIR137-mediated dysregulation as a new etiologic mechanism in schizophrenia.

The International Schizophrenia Consortium (ISC) reported evidence for a polygenic contribution to schizophrenia\(^4\). An independent family based study confirmed these results, greatly minimizing the possibility of population stratification artifact\(^12\). We reevaluated the polygenic model, dividing stage 1 samples into independent training and testing sets (Supplementary Note). The training set had 15,429 subjects (twice the size of the ISC training set), and the testing set consisted of 6,428 individuals independent of the ISC report. The proportion of variance (Nagelkerke’s \(R^2\)\(^4\)) explained in the testing set increased from 3% in the ISC to around 6% here (Supplementary Table 9 and Supplementary Fig. 6). This estimate is much lower than the true total variation in liability that is tagged by all SNPs because SNP effects are estimated with error\(^3.4.22–25\). The polygenic model appears to explain a substantial fraction of the heritability of schizophrenia\(^4\), as has been shown for other complex traits\(^2.6–28\). Some of these additional risk loci are likely contained near the most highly significant results of our stage 1 analysis. Supporting this hypothesis, of the top loci that did not reach genome-wide significance in the combined stage 1 and 2 analysis, a sign test (P < 10\(^{-4}\)) and a Fisher’s combined test (P < 10\(^{-5}\)) both showed an excess of same-direction allelic association (41 of 51 non-MHC SNPs) in the discovery and replication datasets.

Clinical, epidemiological and genetic findings suggest shared risk factors between bipolar disorder and schizophrenia\(^29\). In stage 1, three genes with strong support had prior genome-wide significant associations with bipolar disorder: CACNA1C, the region containing ITIH3-ITIH4 (encoding inter-\(\alpha\) (globulin) inhibitors H3 and H4) and ANK3 (encoding ankyrin G, 3, of Kaner (ankyrin G))\(^10,11,30\) (Supplementary Table 10). We performed a joint analysis with the Schizophrenia Psychiatric Genome-Wide Association Study (GWS) Consortium (PGC) for bipolar disorder applying identical analytical methods. After removing duplicate subjects, we analyzed 16,374 cases with schizophrenia, schizoaffectice disorder or bipolar disorder and 14,044 controls. Support for shared susceptibility was strengthened (Supplementary Table 11) at CACNA1C (rs4765905, P = 7.0 \times 10\(^{-5}\)), ANK3 (rs10994359, P = 2.5 \times 10\(^{-8}\)) and the ITIH3-ITIH4 region (rs2239547, P = 7.8 \times 10\(^{-8}\)), each of which reached genome-wide significance. A coding variant in ITIH4 (p.Pro698Thr; rs4687657) is in perfect LD with the most associated SNP. Although we included all subjects from an earlier report\(^10\), the increased support found with additional independent cases (N = 11,987) and controls (N = 7,835) provides further evidence for shared risk effects of schizophrenia and bipolar disorder.

The risk variants implicated here confer small risks (odds ratios \(\sim 1.10\)), but the polygenic analysis shows many more susceptibility variants with effects for which our sample is underpowered (Supplementary Table 12). At every stage where samples were added, we found an increase in the number of genome-wide significant loci and enhancement of signals at CACNA1C, ANK3 and ITIH3-ITIH4 when schizophrenia and bipolar disorder were jointly analyzed. Thus, gains in power offset any penalty for increased heterogeneity.

In summary, we report seven genome-wide significant schizophrenia associations (five of which are new) in a two-stage analysis of 51,695 individuals. We also report loci that confer susceptibility to both bipolar disorder and schizophrenia. The association near MIR137, associations in multiple predicted MIR137 targets and the known role of MIR137 in neuronal maturation and function together suggest an intriguing new insight into the pathogenesis of schizophrenia.

**METHODS**

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

**ACKNOWLEDGMENTS**

We thank the study participants and the research staff at the many study sites. Over 40 US National Institutes of Health grants and similar numbers of government grants from other countries, along with substantial private and foundation support, enabled this work. We greatly appreciate the sustained efforts of T. Lehner (National Institute of Mental Health) on behalf of the Schizophrenia Psychiatric Genome-Wide Association Study (GWS) Consortium (PGC). Detailed acknowledgments, including grant support, are listed in the Supplementary Note.

**AUTHOR CONTRIBUTIONS**


ONLINE METHODS

Overview. Seventeen samples representing 11 countries (Table 1 and Supplementary Table 1) comprising all known European-ancestry GWAS samples were invited to participate in stage 1. The quality of the clinical data was verified by a systematic review of data collection methods and procedures at each site (Table 1, Supplementary Table 13 and Supplementary Note).

Quality control. Technical quality control was performed on genotypes generated for stage 1 by various GWAS platforms (Table 1), with quality control conducted separately using a common approach; all datasets were separately analyzed (Supplementary Table 3). The SNPs successfully genotyped in each study and common to all platforms (N = 11,310) were pruned to remove high LD and lower frequency SNPs and were then used for relatedness testing in each sample and in the combined total sample. Common quality control parameters were applied: (i) missing rate per SNP <0.05 (before sample removal below), (ii) missing rate per individual <0.02, (iii) missing rate per SNP <0.02 (after sample removal above), (iv) missing rate per SNP difference in cases and controls <0.02, (v) SNP frequency difference to HapMap <0.15 and (vi) Hardy-Weinberg equilibrium (controls) P < 10^-6. Stage 1 study sample sizes varied between 400 and 5,000 individuals (Table 1 and Supplementary Table 3). The number of SNPs per study after quality control varied between 250,000 and 680,000. On average, the quality control processes excluded 15 individuals per study (with a range of 0–100 individuals) and 38,000 SNPs per study (with a range of 5,000–160,000 SNPs). After quality control, the stage 1 GWAS datasets together comprised 21,856 individuals and, for the next steps of the ‘genetic quality control’ analysis, a set of 17,074 SNPs common to all platforms and successfully genotyped in each stage 1 sample was extracted. These SNPs were then further pruned to remove LD (leaving no pairs with r^2 > 0.05) and lower frequency SNPs (minor allele frequency <0.05), leaving 11,310 SNPs suitable for robust relatedness testing and population structure analysis.

Genetic quality control. Genetic quality control included relatedness testing and principal components analyses. Relatedness testing was done with PLINK (see URLs), reporting pairs with genome identity (pi-hat) > 0.9 as ‘identical samples’ and with pi-hat > 0.2 as being closely related. After random shuffling, one SNP from each pair was excluded from downstream analysis. From groups with multiple related pairs (for example, a family), only one individual was kept. After quality control, the PGC GWAS European-ancestry sample (stage 1) consisted of 9,394 cases with schizophrenia or schizoaffective disorder and 12,462 controls (52% of whom were psychiatrically screened). Principal component estimation was done with the same collection of SNPs on the non-related subset of individuals. We estimated the first 20 principal components and tested each of them for phenotype association (using logistic regression with study indicator variables included as covariates) and evaluated their impact on the genome-wide test statistics using λ (the genomic control inflation factor based on the median x^2) after genome-wide association of the specified principal component. Based on this we decided to include principal components 1, 2, 3, 4 and 6 for downstream analysis as associated covariates. The deviation noted in Supplementary Figure 1 was observed more strongly among SNPs with large numbers of neighboring SNPs in high LD (Supplementary Fig. 7); SNPs that tag more variants are more likely to capture true genetic effects.

Imputation. Imputation of untyped SNPs was performed within each study in batches of 300 individuals. These batches were randomly drawn in order to keep the same case-control ratio as in the total sample from that study. We used Beagle 3.0.4 (ref. 31). Imputation was performed with CEU+TSI HapMap phase 3 data (UCSC hg18/NCBI 36) with 410 phased haplotypes encompassing 1,252,901 SNPs. λ was carefully monitored before and after imputation. Each of the 17,074 SNPs present on every platform was dropped and reimputed with high correlation (Pearson r^2 = 0.99973).

Stage 1 association analyses. Stage 1 association analyses of the phenotype were performed with the imputed dosages using standard logistic regression, which intrinsically properly addresses uncertainty in the imputation results. As covariates, we used the study indicator as a categorical variable and five principal components derived from the population.

SNPs from the stage 1 association results were grouped into LD regions using a simple iterative process after all SNPs in the genome were ranked from most to least significant: if a SNP had r^2 > 0.2 (defined by Haplovie (see URLs) analysis of HapMap3 CEU+TSI data) to a more significantly associated SNP within 1 Mb on either side of it, the SNP was assigned to the region defined by the more associated SNP. Distinct regions defined by associated variants physically within ±200 kb were tested using a conditional analysis to examine the independence of the associations in the PGC schizophrenia data. LD regions with P < 2 × 10^-6 are reported in Supplementary Table 6 and were assessed in independent stage 2 (replication) samples.

Power analysis. The power analysis was performed using the Genetic Power Calculator (GPC)22 to calculate the power of the discovery sample (all 17 stage 1 samples together). Supplementary Table 12 shows the power of this sample to detect association at three significance thresholds: a canonical genome-wide significance threshold (P < 5 × 10^-8), the threshold we used for follow up (P < 2 × 10^-5) and a more modest threshold of P < 0.001. The middle of these is most relevant to the power of this study because this threshold represents a level that would have been used to advance to follow up in the larger replication study (stage 2) and reported in this manuscript. Presented are power figures at representative and relevant odds ratios (ORs) (OR = 1.10, OR = 1.15 and OR = 1.20) and allele frequencies (0.05, 0.10, 0.20 and 0.40). Notably, although the power of the sample sizes used in this study to identify loci at the significance threshold used for follow up at ORs greater than 1.15 was high, the power noticeably dropped at an OR of 1.10, where the majority of our confirmed findings were made.

Stage 2 association analyses. Stage 2 association analyses were performed using all accessible replication European-ancestry samples (stage 2), which consisted of 8,442 cases and 21,397 controls (37% of which were screened) in 19 samples representing 14 countries (Table 1 and Supplementary Table 2). We tested the top SNP in each LD region in the replication samples either by looking up the data in GWAS datasets not involved in the PGC analysis or by focused genotyping in unscanned samples (Supplementary Note). Where new genotyping was required, data were provided for the top 15–30 SNPs (depending on available resources) by collaborators. After receiving P values, ORs, standard errors and allele frequencies for the replication datasets, the results were aligned, and a standard error–weighted meta-analysis was performed of (i) the replication experiment (stage 2) alone and (ii) the combination of the PGC discovery and replication data (stages 1 and 2). An analysis of the combination of stages 1 and 2 was also performed with the genomic control–adjusted version of the stage 1 results by converting the genomic control–adjusted P value for stage 1 back into an adjusted z score, which was combined with the stage 2 meta-analysis result (Table 2).

Conditional analysis. A conditional analysis was performed for all regions with P < 10^-6 in a logistic regression framework using as covariates the dosages of the target SNP of the region, multi-dimensional scaling scores 1–4 and 6 and the study indicator. Any P < 10^-4 was then assumed to be independent from the main signal in this region. A plot of these regions after conditioning is presented in Supplementary Figure 5. Within a narrow range of ±200 kb (which is similar to distances used in the analysis of other phenotypes manifesting complex genetics, for example, height) of the original target SNP, we found three regions with a clear sign (permutation–corrected P < 0.05) for one or more independent signals (Supplementary Table 5).

Combined analysis of top-associated SNPs with PGC bipolar disorder. This combined analysis was performed because after the main analysis of the PGC schizophrenia and the PGC bipolar disorder datasets was finalized, we noticed several regions associated at high significance levels in both disorder groups. We therefore applied the same methods as already described for the stage 1 PGC schizophrenia analyses. This included the removal of one individual of each overlapping subject pair (see the description of ‘genetic quality control’). After creating principal components from the SNPs common to each dataset (see ‘genetic quality control’), we tested for association 16,374 joint cases (cases with schizophrenia, schizoaffective disorder or bipolar disorder) and 14,044 joint controls, including principal components 1–7 and...
and the study indicator as covariates (see 'stage 1 association analyses'). We thereafter focused on regions reported by bipolar disorder GWAS. After LD pruning, we report in Supplementary Table 11 the three genome-wide significant regions.

**Additional analyses.** Additional analyses performed included an interaction analysis (see below and Supplementary Table 14), stage 1 association analyses with stricter control for population stratification (Supplementary Figs. 2, 3, 8 and Supplementary Note) and a score analysis to test for a polygenic model of inheritance (Supplementary Table 9, Supplementary Fig. 6 and Supplementary Note). An interaction analysis of associated SNPs was performed between each pair of loci in Table 2 in a logistic regression model that included the main effect of each SNP. No interaction terms were significant when correcting for the number of pairs tested (Supplementary Table 14). Besides the overall stage 1 quantile-quantile and Manhattan plots (Supplementary Figs. 1, 4), we present individual stage 1 sample quantile-quantile and Manhattan plots (Supplementary Figs. 10, 11). To complement the principal component analysis plots for individual stage 1 samples (Supplementary Fig. 8), we also present multi-dimensional scaling for all stage 1 samples compared to HapMap3 anchors (Supplementary Fig. 12).