Sparse whole-genome sequencing identifies two loci for major depressive disorder

CONVERGE consortium*

Major depressive disorder (MDD), one of the most frequently encountered forms of mental illness and a leading cause of disability worldwide, poses a major challenge to genetic analysis. To date, no robustly replicated genetic loci have been identified, despite analysis of more than 9,000 cases. Here, using low-coverage whole-genome sequencing of 5,303 Chinese women with recurrent MDD selected to reduce phenotypic heterogeneity, and 5,337 controls screened to exclude MDD, we identified, and subsequently replicated in an independent sample, two loci contributing to risk of MDD on chromosome 10: one near the SIRT1 gene \((P = 2.53 \times 10^{-10})\), the other in an intron of the LHPP gene \((P = 6.45 \times 10^{-12})\). Analysis of 4,509 cases with a severe subtype of MDD, melancholia, yielded an increased genetic signal at the SIRT1 locus. We attribute our success to the recruitment of relatively homogeneous cases with severe illness.

The existence and number of subtypes of depression have been debated over the past 100 years. The current consensus is that depression may be a collection of partly distinct diseases, with overlapping causal pathways. This aetiologic heterogeneity might therefore substantially reduce the power of genetic association studies, and hence explain the failure to find genetic risk loci. For example, there may be cases of MDD of largely environmental origin whose presence reduces the power to detect genetic effects. Also, genetic risk factors for mild depressive syndromes may not be entirely the same as those for more severe cases.

For these reasons, we investigated the genetic basis of MDD in subjects for whom known sources of phenotypic and genetic heterogeneity were minimized and known risk factors documented. The CONVERGE (China, Oxford and Virginia Commonwealth University Experimental Research on Genetic Epidemiology) consortium recruited 11,670 Han Chinese women through a collaboration involving 58 hospitals in China. We studied only women because about 45% of the genetic liability to MDD is not shared between sexes. In an attempt to obtain severe cases of MDD, we recruited only recurrent cases (mean number of episodes was 5.6).

We used low-coverage sequencing to genotype our sample. Whole-genome sequences were acquired to a mean depth of 1.7× (95% confidence intervals (CIs) 0.7–4.3) per individual, from which 32,781,340 SNP sites were identified. After applying stringent quality controls (Methods), we obtained 10,640 samples (5,303 cases of MDD, 5,337 controls) and 6,242,619 SNPs for inclusion in genome-wide association studies (GWAS). We compared genotypes from the low-coverage sequencing to genotypes called with 10× coverage sequence and to genotypes called from genotyping arrays and a mass spectrometer platform. The mean percentage concordance between genotypes from nine individuals with both low- and 10× coverage across all sites was 98.1% (Supplementary Table 1). We compared imputed genotypes to those acquired for 72 individuals using an array and to 21 SNPs genotyped on all individuals with the MassARRAY system mass spectrometer (Supplementary Notes). Overall concordance was 98.0% (Supplementary Tables 2 and 3).

Genetic association analysis was carried out with a linear mixed model with a genetic relatedness matrix (GRM) as a random effect and principal components from eigen-decomposition of the GRM as fixed effect covariates (Methods, Supplementary Notes). Fig. 1a and Extended Data Fig. 1 show the Manhattan and quantile–quantile plots, respectively, for this analysis. The genomic control inflation factor \(\lambda\) (the ratio of the observed median \(\chi^2\) to that expected by chance) for association with MDD was 1.070 (for common SNPs, minor allele frequency (MAF) >2%, \(\lambda = 1.074\)). The adjusted measure for sample size to that of 1,000 cases and 1,000 controls \(\lambda_{1000}\) was 1.013.

Two loci exceeded genome-wide significance in association with MDD: one 5′ to the sirtuin1 (SIRT1) gene on chromosome 10 \((SNP = rs12415800)\), and the other in an intron of the LHPP gene \((P = 6.45 \times 10^{-12})\). Analysis of 4,509 cases with a severe subtype of MDD, melancholia, yielded an increased genetic signal at the SIRT1 locus. We attribute our success to the recruitment of relatively homogeneous cases with severe illness.

Two loci exceeded genome-wide significance in association with MDD: one 5′ to the sirtuin1 (SIRT1) gene on chromosome 10 \((SNP = rs12415800)\), and the other in an intron of the LHPP gene \((P = 6.45 \times 10^{-12})\). Analysis of 4,509 cases with a severe subtype of MDD, melancholia, yielded an increased genetic signal at the SIRT1 locus. We attribute our success to the recruitment of relatively homogeneous cases with severe illness.
and rs35936514 (LHPP) have frequencies of 3% and 8% respectively, compared to 45% and 26% in the CONVERGE cohort.

We considered whether successful mapping of MDD in the CONVERGE samples was attributable to the recruitment of a severe, more genetically determined form of the disease. We tested that hypothesis by looking within the CONVERGE cohort at a particularly severe, and more heritable form of MDD: melancholia. Prior research has suggested that MDD patients with melancholia have more impairing, recurrent episodes and that risk for MDD is higher in the co-twins of probands with the melancholic subtype than in those with non-melancholic MDD. This increase is greater in monozygotic than dizygotic twin pairs, as would be expected if the subtype were associated with greater genetic risk.

In the CONVERGE cohort, 85% of cases met the DSM-IV criteria for melancholia. We searched for a genetic association in 9,846 samples (4,509 cases and 5,337 controls) and identified the same two loci that exceeded genome-wide significance on chromosome 10. The genomic control inflation factor for melancholia was 1.069, and \( \lambda_{1000} \) was 1.014. Even though the sample for melancholia was smaller than for MDD, at the SIRT1 locus the significance of association was two orders of magnitude greater than for MDD (top SNP rs80309727, chromosome 10:69617347, MAF 45.2%, \( P = 2.95 \times 10^{-10} \)). Extended Data Fig. 3 shows the Manhattan plot, quantile–quantile plot and detailed views of the SIRT1 locus associated with melancholia. All SNPs with \( P \) values of association \( < 10^{-5} \) with melancholia are listed in Supplementary Table 5.

To determine whether the increased association might have arisen by chance, we generated an empirical distribution of odds ratios by randomly selecting 4,509 cases from the total set and re-analysing the association with each of the genome-wide significant variants. We found that the observed value lay on the 98.8th percentile at the SIRT1 locus, but at the 61.6th percentile at the LHPP locus (Extended Data Fig. 4).

Our results indicate that, as others have suggested, obtaining low-sequence coverage of a large number of individuals can be an effective...
way to screen the genome for association signals. We were able to genotype more variants than on genotyping arrays and our set is larger than publicly available sources for imputation44. Our imputation pipeline employed standard tools, and it is likely that imputation accuracy could be improved with further algorithmic research.

MDD is most probably highly polygenic45, and many additional loci remain to be discovered. We attribute the discovery and replication of two SNPs associated with MDD in the CONVERGE cohort to the recruitment of cases who were probably more homogeneous and more severely impaired than those collected in previous studies from Western cultures. In East Asia, reluctance to report MDD46 probably explains why hospital-ascertained cases are more severe, and why prevalence estimates for MDD are lower in China (3.6%47) than in the US (16.2%)48. Consistent with this interpretation, 85% of the cases of MDD in the CONVERGE cohort have melancholia, a severe subtype of MDD; mapping melancholia led to a significant increase in the genetic signal at one locus. Finally, we note that one of the replicated risk loci is located close to a gene involved in mitochondrial biogenesis (SIRT1)49, which, together with our finding that MDD is associated with increased amounts of mitochondrial DNA50, suggests an unexpected origin for at least some of the phenotypic manifestations of MDD.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 December 2014; accepted 11 June 2015. Published online 15 July 2015.

Table 1 | Genetic association between MDD and 12 variants in the CONVERGE cohort and a replication sample

| Chr. | Pos. | SNP | CONVERGE (n = 10,640) | Replication (n = 6,417) | Joint (n = 17,057) |
|------|------|-----|-----------------------|-------------------------|---------------------|
|      |      |     | Info | OR | s.e. | OR | s.e. | OR | s.e. | OR | s.e. | OR | s.e. |
| 1    | 11493832 | rs29222240 | T   | C   | 0.385 | 1.018 | 1.141 | 0.028 | 2.80 | 10^-6 | 0.949 | 0.037 | 1.54 | 10^-1 | 1.070 | 0.022 | 2.14 | 10^-3 |
| 2    | 17515950 | rs3766688 | T   | C   | 0.394 | 1.003 | 0.875 | 0.028 | 1.83 | 10^-6 | 0.991 | 0.037 | 8.15 | 10^-10 | 0.918 | 0.022 | 1.34 | 10^-4 |
| 5    | 9161674 | rs5751388 | A   | G   | 0.294 | 0.970 | 1.138 | 0.031 | 4.09 | 10^-6 | 1.001 | 0.041 | 9.90 | 10^-10 | 1.088 | 0.025 | 6.57 | 10^-3 |
| 6    | 5916265 | rs1921918 | G   | A   | 0.096 | 0.891 | 1.278 | 0.050 | 6.04 | 10^-7 | 1.054 | 0.062 | 3.93 | 10^-10 | 1.042 | 0.035 | 2.08 | 10^-1 |
| 8    | 126244970 | rs11880240 | C   | G   | 0.068 | 1.019 | 1.291 | 0.055 | 8.02 | 3.44 | 10^-7 | 0.954 | 0.052 | 4.49 | 10^-1 | 0.876 | 0.031 | 1.82 | 10^-5 |
| 9    | 9161674 | rs55713588 | A   | G   | 0.294 | 0.970 | 1.138 | 0.031 | 4.09 | 10^-6 | 1.001 | 0.041 | 9.90 | 10^-10 | 1.088 | 0.025 | 6.57 | 10^-3 |
| 11   | 4386107 | rs5800092 | G   | T   | 0.151 | 1.001 | 0.824 | 0.039 | 1.35 | 10^-6 | 0.962 | 0.052 | 4.49 | 10^-1 | 0.876 | 0.031 | 1.82 | 10^-5 |
| 10   | 69624180 | rs12415800 | G   | A   | 0.452 | 0.992 | 1.164 | 0.028 | 1.92 | 10^-6 | 1.130 | 0.036 | 7.71 | 10^-10 | 1.150 | 0.022 | 2.37 | 10^-10 |
| 12   | 126244970 | rs35936514 | C   | T   | 0.260 | 0.998 | 0.839 | 0.032 | 1.27 | 10^-6 | 0.838 | 0.041 | 1.68 | 10^-10 | 0.842 | 0.025 | 6.53 | 10^-12 |
| 13   | 1117671012 | rs12415800 | G   | A   | 0.452 | 0.992 | 1.164 | 0.028 | 1.92 | 10^-6 | 1.130 | 0.036 | 7.71 | 10^-10 | 1.150 | 0.022 | 2.37 | 10^-10 |
| 14   | 66838351 | rs17827252 | G   | C   | 0.463 | 1.011 | 0.887 | 0.028 | 1.44 | 10^-5 | 0.962 | 0.041 | 3.41 | 10^-10 | 0.907 | 0.023 | 2.20 | 10^-5 |
| 15   | 34493757 | rs11880240 | G   | T   | 0.068 | 1.019 | 1.291 | 0.050 | 8.02 | 10^-6 | 1.048 | 0.072 | 5.12 | 10^-10 | 1.184 | 0.043 | 9.15 | 10^-5 |
| 16   | 24656688 | rs1921918 | G   | A   | 0.721 | 0.995 | 0.883 | 0.031 | 3.22 | 10^-3 | 0.994 | 0.047 | 9.01 | 10^-10 | 0.917 | 0.026 | 1.09 | 10^-3 |
| 17   | 25011137 | rs11493832 | G   | C   | 0.260 | 0.971 | 1.130 | 0.032 | 5.86 | 10^-6 | 1.011 | 0.047 | 8.19 | 10^-10 | 1.100 | 0.027 | 2.18 | 10^-4 |

The table reports results for 12 SNPs in the CONVERGE and replication samples. The first five columns give the chromosome (Chr.), position (Pos.), SNP identifier (Ref.), allele frequency (Alt.) called in CONVERGE. The next five columns show the alternative allele frequency (Freq.) and results of association testing with MDD using imputed allele dosages in 10,640 CONVERGE samples (5,933 cases, 5,397 controls); information scores (Info.), odds ratio (OR) of association with MDD with respect to the alternative allele and standard error (s.e.) in the odds ratio were obtained from a logistic regression model; P-values of association (P) were obtained from a linear-mixed model with a GRM containing all samples. The next three columns present the results of association with MDD in the replication cohort of 6,417 samples (3,231 cases, 3,186 controls) from a logistic regression model. The final three columns present the results of association with MDD in a joint analysis with both CONVERGE and replication cohorts from a logistic regression model. Bold type indicates the genome-wide significant markers.
**METHODS**

No statistical methods were used to predetermine sample size.

**Sample collection.** CONVERGE collected cases of recurrent major depression from 58 provincial mental health centres and psychiatric departments of general medical hospitals, and 23 provinces of China. Controls were recruited from patients undergoing minor surgical procedures at general hospitals (37%) or from local community centres (63%). A sample size of 6,000 cases and 6,000 controls was chosen on the basis of evidence available when the study was designed (in 2007) of the likely existence of genetic loci with odds ratio of 1.2 and above. All subjects were Han Chinese women with four Han Chinese grandparents. Cases were excluded if they had a pre-existing history of bipolar disorder, psychosis or mental retardation. Cases were aged between 30 and 60 and had two or more episodes of MDD meeting DSM-IV criteria with the first episode occurring between 14 and 50 years of age, and had not abused drugs or alcohol before their first depressive episode. All subjects were interviewed using a computerized assessment system. Interviewers were postgraduate medical students, junior psychiatrists or senior nurses, trained by the CONVERGE team for a minimum of 1 week.

The diagnosis of MDD was established with the Composite International Diagnostic Interview (CIDI) (WHO lifetime version 2.1; Chinese version), which used DSM-IV criteria. The interview was originally translated into Mandarin by a team of psychiatrists at Shanghai Mental Health Centre, with the translation reviewed and modified by members of the CONVERGE team.

The replication sample was obtained from five hospitals in the north of China. Patients were diagnosed as having MDD by at least two consultant psychiatrists by DSM-IV criteria. Samples were of both sexes, and all four grandparents were Han Chinese. Cases were aged between 30 and 60 and had two or more episodes of MDD meeting DSM-IV criteria. Exclusion criteria were pregnancy, severe medical conditions, abnormal laboratory baseline values, unstable psychiatric features (for example, suicidal), a history of alcoholism or drug abuse, epilepsy, brain trauma with loss of consciousness, neurological illness, or a concomitant axis I psychiatric disorder. Control subjects were recruited from local communities and provided information about medical and family histories. Exclusion criteria were a history of major psychiatric or neurological disorders, psychiatric treatment or drug abuse, or a family history of severe forms of psychiatric disorders.

The study protocol was approved centrally by the Ethical Review Board of Oxford University (Oxford Tropical Research Ethics Committee) and the ethics committees of all participating hospitals in China. All interviewers were mental health professionals who are well able to judge decisional capacity. The study posed minimal risk (an interview and saliva sample). All participants provided written informed consent.

**DNA sequencing.** DNA was extracted from saliva samples using the Oragene protocol. A barcoded library was constructed for each sample. All saliva samples were randomized in allocation to sequencing batches, and experimenters performing the sequencing procedure were blinded to sample allocation and outcome assessment. Sequencing reads obtained from Illumina HiSeq machines were aligned to Genome Reference Consortium Human Build 37 patch release 5 (GRCh37.p5) using default parameters after filtering of reads containing adapter sequences or consisting of more than 50% poor quality (base quality ≤5) bases. Samtools (v0.1.18) was used to index the alignments in BAM format, and Picardtools (v1.62) was used to mark PCR duplicates for downstream filtering. The Genome Analysis Toolkit’s (GATK, version 2.6)”4 BaseRecalculator was then run on the BAM files to create base quality score recalibration tables, masking known SNPs and INDELS from dbSNP (version 137, excluding all sites added after version 129). Base quality recalibration (BQSR) was then performed on the BAM files using GATK’s UnifiedGenotyper (version 2.7-2-g6bda569), and the option set ‘–genotype_likelihood_model’ to ‘BOTHS’, used default annotation outputs for variant calls, and set the ‘–dbsNP’ option in order to use dbsNP v137 redis to fill in the variant ID column of the output variant call format (VCF) files. Variant quality score recalibration was then performed on these sites using the GATK’s VariantRecalibrator (version 2.7-2-g6bda569) and the biallelic SNPs from the 1000G Phase 1 ASN panel using 572 haplotypes from the 1000 Genomes Phase 1 ASN samples as a reference panel for six iterations on chunks containing roughly 3,000 SNPs with 600 SNPs of overlap. After both rounds of imputation we removed the outer 300 SNPs of every window and ligated imputation results of adjacent chunks. A final set of allele dosages and genotype probabilities was generated from these two sets of imputed results by replacing the results in the former with those in the latter at all sites imputed in the latter. We then applied a conservative set of inclusion thresholds for SNPs for GWAS: (a) P value for violation of the Hardy–Weinberg equilibrium >10^-5; (b) information score >0.9; (c) MAF in CONVERGE >0.5%, to arrive at the final set of 6,242,619 SNPs for GWAS.

**Sample selection for GWAS.** Using both processed sequencing data and imputed dosages at SNPs that passed quality control, we assessed the sequencing and imputation quality of all 11,670 samples whose genomic variants we imputed. We first looked into both the nuclear genome and mitochondrial genome for an excess of variants called, since this would indicate cross-sample contamination due to technical issues during sequencing. We quantified the number of singleton variants called in genic regions of the nuclear genome and found a mean of 71.55 private variants per sample that were supported by more than 2 sequencing reads passing sequencing quality controls. We excluded 117 samples with a number of singletons greater than the 99th percentile. Coverage of the mitochondrial genome was, on average, 102×, allowing us to obtain high-quality sequences for this part of the genome. We found a mean of 15.70 heteroplasmic sites per sample, and 116 samples were found to have greater than the 99th percentile of the number of heteroplasmic sites. Of these 116 samples, 26 were already discarded for having excess nuclear genome singletons; and we excluded the remaining 90.

We then checked imputation quality based on the certainty of genotypes imputed (maximum genotype probability >0.9). We identified 29 individuals who had fewer than 90% of their sites with maximum genotype probabilities >0.9. We excluded these samples from further analysis.

Finally, we assessed the 11,144 remaining samples for genetic relatedness. Although being unrelated to other individuals recruited for the CONVERGE study was a clear criteria in our data collection process, there were instances when the same patient or a relative of the patient visited multiple hospitals and was thus recruited more than once. To exclude duplicates and first-degree relatives from our sample for GWAS, we estimated pairwise genome-wide identity by descent (IBD) using identity by state (IBS) information in hard-called genotypes from imputed genotype probabilities at 399,211 common tagging SNPs across all autosomes (MAF >1%, linkage disequilibrium (LD) <0.5, all known in 1000 Genomes Phase 1). We implemented this in PLINK (v1.07)” with the option ‘–memory <value> Mb’ (where <value> is the total memory (in Megabytes) available on the computer) to set the memory required by PLINK for running downstream analyses.

The study posed minimal risk (an interview and saliva sample). All participants provided their written informed consent.

**Genotype likelihood calculation and imputation.** Genotype likelihoods (GLs) were calculated at all 20,539,441 SNPs using a sample-specific binomial mixture model implemented in SNPtools (version 1.0)” and imputation was performed without a reference panel using BEAGLE (version 3.3.2)”2. We used BEAGLE to perform imputation, using ten iterations on chunks of 3,000 SNPs with 600 SNPs of overlap. A second round of imputation was performed with BEAGLE on the same GLs, from biallelic SNPs polymorphic in the 1000G Phase 1 ASN panel and using 572 haplotypes from the 1000 Genomes Phase 1 ASN samples as a reference panel for six iterations on chunks containing roughly 3,000 SNPs with 600 SNPs of overlap. After both rounds of imputation we removed the outer 300 SNPs of every window and ligated imputation results of adjacent chunks. A final set of allele dosages and genotype probabilities was generated from these two sets of imputed results by replacing the results in the former with those in the latter at all sites imputed in the latter. We then applied a conservative set of inclusion thresholds for SNPs for GWAS: (a) P value for violation of the Hardy–Weinberg equilibrium >10^-5; (b) information score >0.9; (c) MAF in CONVERGE >0.5%, to arrive at the final set of 6,242,619 SNPs for GWAS.

**Replication and joint analyses.** We genotyped the replication sample on a MassARRAY system mass spectrometer. TYPER4.0 was used to assess the reliability of genotype calls generated by SpectroREAD from the mass spectra. Default genotype call inclusion criteria were used. To perform the association analysis with MDD case–control status at these 12 sites in the replication sample, we obtained effect sizes for discovery from logistic regression with principal component (PC) correction, and then for replication from logistic regression, and then performed fixed-effects meta-analysis.
Polygenic risk profiling and binomial sign-test. Single SNP association results were obtained from the PGC study of MDD. Prior to analysis, SNPs were lifted over to GRCh37/hg19 coordinates and excluded if: (a) monomorphic in either European ($n = 379$) or East Asian ($n = 286$) populations from the 1000 Genomes Project Phase 1 reference data; or (b) absent from the filtered CONVERGE data set. To construct the PGC-trained polygenic score, we initially selected autosomal SNPs with statistical imputation information (information score) greater than 0.9 and MAF greater than 1% in both studies, and performed subsequent LD-based ‘clumping’ to remove markers from highly correlated SNP pairs (pairwise $r^2 > 0.2$ in East Asians, 500 kb window) while preferentially retaining SNPs with smaller PGC $P$ values. Using the resultant SNP set, we constructed polygene scores based on varying $P$ value thresholds ($1 \times 10^{-6}, 1 \times 10^{-4}, 1 \times 10^{-3}, 0.001, 0.01, 0.02, 0.03, 0.04, 0.05, \text{and } 1$) as previously described. We assessed the predictive value of polygenic scores in a genetically unrelated subset of the CONVERGE sample (with pairwise relatedness less than 0.1) by logistic regression, with adjustment for ancestry principal components, demonstrating significant association with MDD status. The estimated variance in MDD risk accounted for by the polygenic score is given by Nagelkerke’s $R^2$. Using the same $P$ value thresholds, we tabulated the number of independent SNPs with the same direction of allelic effect in the PGC results as observed in CONVERGE. The filtering criteria for SNPs was an information score greater than 0.9 in CONVERGE and MAF greater than 1% in both studies; and an analogous LD-clumping procedure was performed (pairwise $r^2 > 0.2$ in Europeans, 500 kb window). A one-sided binomial sign test was used to assess whether this observed fraction was significantly greater than that expected by chance. Results are given in Extended Data Table 4.

21. Association, A. P. Diagnostic and statistical manual of mental disorders 4th edn (American Psychiatric Association, 1994).
22. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21, 936–939 (2011).
23. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
24. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
25. Wang, Y. et al. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. Genome Res. 23, 833–842 (2013).
26. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, 1084–1097 (2007).
27. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
28. R Development Core Team. A language and environment for statistical computing (R Foundation for Statistical Computing, 2004).
29. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 9, e1003348 (2013).
Extended Data Figure 1 | Quantile–quantile plots for major depressive disorder. Quantile–quantile plot of GWAS for MDD using the mixed linear model with exclusion of the chromosome that the marker is on (MLMe) method implemented in FastLMM on 10,640 samples (5,303 cases, 5,337 controls). Genomic inflation factor $\lambda = 1.070$, rescaled for an equivalent study of 1,000 cases and 1,000 controls ($\lambda_{1000} = 1.013$).
Extended Data Figure 2 | Forest plots of estimated SNP effects in CONVERGE and PGC studies. This figure presents the association odds ratios (OR) at 12 SNPs in CONVERGE and the best available proxy SNPs in PGC-MDD (pairwise $r^2 > 0.6$, 500 kb window; the proxy SNP is marked by an asterisk). We present the alternative allele frequency (freq), odds ratio (or) with respect to the alternative allele, standard error of odds ratio (se) and $P$ values of association (pval) for the following analyses (study): primary association analysis with a linear-mixed model using imputed allele dosages in 10,640 samples in CONVERGE (pri); validation analysis with logistic regression model with principal components (PCs) as covariates using genotypes from Sequenom on 9,921 samples in CONVERGE (sqnm); association with MDD with a logistic regression model in a replication cohort of 6,417 samples using genotypes from Sequenom (repli); joint association analysis with MDD with a logistic regression model using imputed allele dosages in CONVERGE and genotypes from Sequenom in a replication cohort (17,057 samples in total; joint).
Extended Data Figure 3 | Manhattan and quantile quantile plots for melancholia. a, Manhattan plot of GWAS for melancholia using the MLMe method implemented in FastLMM on 9,846 samples (4,509 cases, 5,337 controls). b, Quantile–quantile plot of GWAS for melancholia; $\lambda = 1.069$, $\lambda_{1000} = 1.014$. c, Regional association plot of GWAS hits on chromosome 10, focusing on top SNP rs80309727 at 5’ of SIRT1 gene, generated with LocusZoom.
Extended Data Figure 4 | Empirical estimation of the odds ratio increases due to the removal of cases not falling under the diagnostic class of melancholia from an association analysis with major depression. The figures show the empirical distributions of the odds ratios for association with each of two SNPs (rs79804696, rs35936514), after removing a random set of 796 samples, equal to the number of cases of MDD not diagnosed as being melancholic. The horizontal axis is the odds ratio for each analysis, and the vertical axis the frequency of occurrence of the odds ratio in 10,000 analyses. The vertical red line is the observed odds ratio after removing cases of MDD not diagnosed as melancholic.
Extended Data Table 1 | Comparison between association results using imputed dosages and directly genotyped markers

| SNP | Imputed Dosages (N=9,921) | Sequenom genotypes |
|-----|--------------------------|--------------------|
| CHR | POS | RSID | REF | ALT | OR | SE | P     | N     | r²  | OR | SE | P     |
| 1   | 11493832 | rs2922240 | C | T  | 1.141 | 0.029 | 5.82E-06 | 9,864 | 0.991 | 1.141 | 0.029 | 5.72E-06 |
| 1   | 175151950 | rs3766688 | C | T  | 0.870 | 0.029 | 1.93E-06 | 9,901 | 0.995 | 0.871 | 0.029 | 2.32E-06 |
| 5   | 228052027 | rs57047840 | G | A  | 1.141 | 0.032 | 4.13E-05 | 9,724 | 0.974 | 1.141 | 0.032 | 3.91E-05 |
| 5   | 9161674 | rs55713588 | G | A  | 1.302 | 0.052 | 3.34E-07 | 9,636 | 0.925 | 1.263 | 0.050 | 2.87E-06 |
| 6   | 4386107 | rs55800092 | T | C  | 0.819 | 0.040 | 6.35E-07 | 9,881 | 0.992 | 0.817 | 0.040 | 5.52E-07 |
| 10  | 69624180 | rs12415800 | A | G  | 1.167 | 0.029 | 8.44E-08 | 9,689 | 0.993 | 1.167 | 0.029 | 9.30E-08 |
| 10  | 126244970 | rs35936514 | T | C  | 0.845 | 0.033 | 2.91E-07 | 9,915 | 0.993 | 0.842 | 0.033 | 1.40E-07 |
| 13  | 107659212 | rs61967003 | T | C  | 1.765 | 0.116 | 9.64E-07 | 9,914 | 0.997 | 1.748 | 0.116 | 1.53E-06 |
| 14  | 66833851 | rs17827252 | G | C  | 0.896 | 0.029 | 1.12E-04 | 8,562 | 0.999 | 0.897 | 0.031 | 3.94E-04 |
| 19  | 34493757 | rs11880240 | G | C  | 1.281 | 0.056 | 1.08E-05 | 9,912 | 0.996 | 1.281 | 0.056 | 1.14E-05 |
| X   | 24656658 | rs1921918 | A | G  | 0.877 | 0.032 | 3.59E-05 | 9,899 | 0.994 | 0.880 | 0.032 | 6.39E-05 |
| X   | 25011374 | rs11573525 | T | C  | 1.158 | 0.033 | 9.60E-06 | 9,912 | 0.969 | 1.144 | 0.032 | 3.21E-05 |

The table reports results for association between MDD and 12 SNPs. The first five columns give the chromosome (CHR), genomic position (POS), SNP identifier (RSID), reference allele (REF) on Human Genome Reference GRCh37.p5, and alternative allele (ALT) called in CONVERGE. The next three columns show results for imputed allele dosages at 12 SNPs (odds ratio (OR) of association with MDD with respect to the alternative allele and standard error (SE); P values of association (P)). The next two columns present the number of samples (N) successfully genotyped using the Sequenom platform (a high-sensitivity and specificity assay), and the Pearson correlation (r²) between the imputed allele dosages and the genotypes from Sequenom. The final three columns present results from analyses of association with MDD using genotypes from the Sequenom genotyping platform. Bold type indicates the genome-wide significant markers; Extended Data Table 2 gives further information on the results for these markers.
Extended Data Table 2 | Genotype distribution and $P$ values for violation of the Hardy-Weinberg equilibrium in CONVERGE and replication cohorts

| MDD Disease State | SNP       | CONVERGE            | Replication Cohort |
|-------------------|-----------|---------------------|--------------------|
|                   |           | HomRef/Het/HomAlt  | HWE P-value        | HomRef/Het/HomAlt  | HWE P-value |
| All               | rs12415800 | 2151/5301/3169     | 0.445              | 1219/3037/1920     | 0.857       |
|                   | rs35936514 | 705/4054/5794      | 0.919              | 422/2400/3398      | 0.974       |
| Cases             | rs12415800 | 1178/2626/1490     | 0.741              | 654/1538/918       | 0.829       |
|                   | rs35936514 | 318/1919/3027      | 0.549              | 190/1130/1783      | 0.627       |
| Controls          | rs12415800 | 973/2675/1679      | 0.106              | 558/1499/1002      | 0.971       |
|                   | rs35936514 | 387/2136/2767      | 0.389              | 232/1264/1615      | 0.503       |

This table shows the number of samples with the homozygous reference genotype (HomRef), heterozygous genotypes (Het), and homozygous alternative genotype (HomAlt), as well as $P$ values for violation of the Hardy-Weinberg equilibrium (HWE) for both CONVERGE study samples and the replication cohort from northern China at the top SNPs rs12415800 in the SIRT1 locus and rs35936514 in the LHPP locus from the GWAS on MDD. The top two rows show these measures for all samples in both the CONVERGE and replication study, the next two rows show these measures for just cases in CONVERGE and the replication cohort, and the last two rows show these measures for just the controls. The genotype distributions for CONVERGE are obtained from hard-called genotypes from maximum imputed genotype probabilities for each sample at each of the two sites. As a genotype will not be called if the maximum genotype probability at a site is lower than 0.9 for any single sample, the total number of CONVERGE samples showing called HomRef/Het/HomAlt genotypes does not equal 10,640 for either SNP. For rs12415800, 19 samples (9 cases, 10 controls) have no genotype calls owing to a maximum genotype probability smaller than 0.9, giving a total of 10,621 CONVERGE (5,294 cases, 5,327 controls) samples with genotype calls. For rs35936514, 87 (39 cases, 48 controls) samples have no genotype calls owing to a maximum genotype probability smaller than 0.9, giving a total of 10,553 (5,264 cases, 5,289 controls) CONVERGE samples with genotype calls.
Extended Data Table 3 | Single-marker association results of top CONVERGE hits in the PGC study of MDD

| CONVERGE (10,640 samples) | PGC MDD (18,759 samples) |
|---------------------------|--------------------------|
| **CHR** | **POS** | **RSID** | **REF** | **ALT** | **FREQ** | **OR** | **SE** | **P** | **RSID** | **LD r²** | **FREQ** | **INFO** | **OR** | **P** |
| 1 | 11493832 | rs2922240 | C | T | 0.3846 | 1.141 | 0.028 | 2.80E-06 | rs2922240 | 1.00 | 0.495 | 0.999 | 0.947 | 0.011 |
| 1 | 175151950 | rs3766688 | C | T | 0.394 | 0.875 | 0.028 | 1.83E-06 | rs3766688 | 1.00 | 0.546 | 0.926 | 0.970 | 0.163 |
| 1 | 228052027 | rs57047840 | G | A | 0.2843 | 1.138 | 0.031 | 4.64E-05 | rs10916214 | 0.48 | 0.133 | 0.992 | 0.969 | 0.300 |
| 5 | 9161674 | rs55713588 | G | A | 0.0956 | 1.278 | 0.050 | 6.04E-07 | rs13360003 | 0.18 | 0.018 | 0.774 | 0.920 | 0.442 |
| 6 | 4386107 | rs55800092 | T | C | 0.1512 | 0.824 | 0.039 | 1.35E-06 | rs17138114 | 0.89 | 0.073 | 0.975 | 0.972 | 0.440 |
| 10 | 69624180 | rs12415800 | A | G | 0.4519 | 1.164 | 0.028 | 1.92E-08 | rs16924945 | 0.32 | 0.005 | 0.355 | 1.629 | 0.496 |
| 10 | 126244970 | rs35936514 | T | C | 0.2609 | 0.839 | 0.032 | 1.27E-06 | rs35841851 | 0.98 | 0.023 | 0.92 | 0.970 | 0.553 |
| 13 | 107659212 | rs61967003 | T | C | 0.0172 | 1.645 | 0.109 | 6.70E-06 | rs16969540 | 0.98 | 0.055 | 0.982 | 0.941 | 0.211 |
| 14 | 66833851 | rs17827252 | G | C | 0.4624 | 0.887 | 0.028 | 1.44E-05 | rs2319184 | 0.96 | 0.028 | 0.941 | 1.114 | 0.115 |
| 19 | 34493757 | rs11880240 | G | C | 0.0679 | 1.291 | 0.055 | 8.02E-06 | rs7254953 | 0.65 | 0.092 | 0.785 | 1.054 | 0.253 |
| X | 24656658 | rs1921918 | A | G | 0.7206 | 0.883 | 0.031 | 3.22E-05 | rs1921918 | 1.00 | 0.581 | 1.238 | 0.960 | 0.035 |
| X | 25011374 | rs11573525 | T | C | 0.2602 | 1.160 | 0.032 | 5.86E-06 | rs11573525 | 1.00 | 0.158 | 1.081 | 1.031 | 0.275 |

The table compares results from 12 SNPs genotyped in the CONVERGE cohort with either the same SNPs, or best available proxies within a 500 kb window, as reported by the Major Depressive Disorder Working Group of the PGC. The first five columns give the SNP identifier (RSID), chromosome (CHR), genomic position (POS), reference allele (REF) on Human Genome Reference GRCh37.p5, and alternative allele (ALT) called in CONVERGE. The next four columns show the alternative allele frequency (FREQ) and results of association testing with MDD at the 12 SNPs in CONVERGE: odds ratio (OR) of association with MDD with respect to the alternative allele and standard error (SE) in the odds ratio were obtained from a logistic regression model with PCs as covariates; P-values of association (P) were obtained from a linear mixed model with a genetic relatedness matrix containing all samples. The next three columns show the SNP identifier (RSID) of best available proxy of each SNP reported in PGC-MDD, the linkage disequilibrium correlation (LD r²) expressed as the r² value between the SNP in PGC-MDD and SNP in CONVERGE, and the alternative allele frequency (FREQ) at the SNP in PGC-MDD. The last three columns show the information scores (INFO), odds ratios (OR) and P-values of association with MDD in PGC-MDD from a logistic regression model. Bold type indicates the genome-wide significant markers.
### Extended Data Table 4 | Polygenic risk profiling and binomial sign tests

| $p_T$     | Polygenic risk profiling | Binomial sign test |
|-----------|--------------------------|--------------------|
|           | $r^2$ | $P$ | No. SNPs (%) | $P$ |
| 0.000001  | 0.00715 | 0.0174 | 3 (100) | 0.125 |
| 0.00001   | 8.40E-05 | 0.415 | 12 (66.7) | 0.194 |
| 0.001     | 2.57E-05 | 0.652 | 62 (58.1) | 0.126 |
| 0.01      | 5.87E-06 | 0.829 | 481 (53.6) | 0.0605 |
| 0.1       | 8.67E-05 | 0.407 | 3632 (51.1) | 0.101 |
| 0.2       | 0.00142 | 0.000797 | 25106 (50.4) | 0.126 |
| 0.3       | 0.00126 | 0.00156 | 45166 (50.6) | 0.00331 |
| 0.4       | 0.00116 | 0.00246 | 61074 (50.5) | 0.00627 |
| 0.5       | 0.0011 | 0.00317 | 86429 (50.4) | 0.00758 |
| 1         | 0.000924 | 0.00684 | 124361 (50.3) | 0.0116 |

The table shows the predictive value of a PGC-trained polygenic risk score on the CONVERGE results. Predictive values are shown at varying $P$-value thresholds ($p_T$) from $P = 1 \times 10^{-4}$ to 1 (that is, all results). $P$ is the $P$-value of the prediction and $r^2$ is the amount of variance explained (thus the table shows that including all independent SNPs from the PGC study of MDD, irrespective of individual $P$ value, explained 0.09% of MDD risk in CONVERGE.). The number of independent SNPs at each threshold is presented (No. SNPs); the significance of the observed fraction (%) demonstrating a consistent direction of effect was assessed by a one-sided binomial sign test.