Supplemental Data for Schulze et al, "Molecular genetic overlap in bipolar disorder, schizophrenia, and major depressive disorder"

Samples

GAIN BD

In the data cleaning process, we identified and excluded 3,654 SNPs with ≥ 2 errors between duplicate subjects, 54 SNPs with a homozygote pp to qq error between duplicates, and three X-linked SNPs that were consistently homozygous in females. An additional 1,675 SNPs were excluded for a minor allele frequency (MAF) below 5%, missing genotypes over 2%, or highly significant (p<10-4) deviation from Hardy-Weinberg-Equilibrium (HWE). From the case sample, we excluded 21 (known) parents, 23 (known) duplicates, 20 subjects who upon final review had not been assigned a high-confidence diagnosis of BD1 or SABP, and one unaffected subject. The total genotyping rate in the remaining 1001 cases and 1033 controls was 99.55 % across 723,918 SNPs. The genomic inflation factor was 1.03.

WTCCC BD

Using PLINK (vers. 1.04), we filtered out a total of 158,075 SNP markers for "poor clustering" (1), different rates of missing data in cases vs. controls excessive heterozygosity, MAF that were more than three-fold different from those reported in HapMap CEU (www.hapmap.org), deviation from HWE at p <0.0001, more than 2% missing data, or MAF < 5%. We used GRR software (2) and 500 unlinked SNPs to identify 112 apparently related individuals (mean identity by state scores of >1.74) who were subsequently dropped. Also using PLINK, we then filtered out 181 individuals for > 3% missing data, genome-wide heterozygosity that was greater or less than four standard deviations from the mean for this sample, gender mismatch, or an outlier position in a multidimensional scaling analysis based on four principal dimensions of identity-by-state (IBS). The cleaned file had 1856 cases and 2945 controls and 342,493 SNPs. This data was then imputed using the methods described below. The imputed files contained 2,094,427 SNPs and 4801 subjects. On final analysis, 61,398 SNPs that deviated from HWE at p<0.00001, and an additional 23,528 SNPs with MAF < 5% in only the cases or only in the controls were also dropped. The total genotyping rate in the remaining 1856 cases and 2945 controls was 99.4411% across 2,009,502 SNPs. After analysis, the genomic inflation factor was 1.09

German BD

The following quality control steps were taken using PLINK (vers. 1.4): 770 SNPs were removed for potential genotyping error based on checking in 20 duplicate samples. Four male cases were removed for potential gender mismatch based on the analysis of heterozygous haploid genotypes. Of the remaining 2062 individuals, 59 with more than 5% missing genotypes were removed. A total of 1188 markers that deviated from HWE at p<0.0001, 15,777 SNPs with more than 2% missing data, and 29,168 with MAF <2% were removed. The total genotyping rate in the remaining 652 cases and 1351 controls was 99.55% across 516,024 SNPs. Finally, we removed six more individuals with a potential gender mismatch identified by the 'sexcheck' procedure in PLINK; 14 individuals identified by GRR as potentially related or duplicated; and 28 individuals identified by the multidimensional scaling (MDS) procedure in PLINK as population outliers. These procedures yielded a final sample size of 645 cases and 1310 controls and 516,024 SNPs. This data was then imputed using the methods described below. The imputed files contained 2,135,560 SNPs and 1955 subjects. On final analysis, 4801 SNPs were dropped that deviated from HWE at p <0.00001, an additional 40,184 SNPs with MAF < 5% in only the cases or only in the controls were also dropped. The total genotyping rate in the remaining 645 cases and 1,310 controls was 99.0681% across 2,090,575 SNPs. After analysis, the genomic inflation factor was 1.05141.

GAIN MDD

We obtained a list of cases and controls that passed stringent quality control in the original study (3), and used only those in this analysis. At the SNP level, 335 markers were dropped due to missing/ambiguous chromosomal assignment. This data was then imputed using the methods described below. The imputed files contained 2,119,526 SNPs and 3,573 subjects. On final analysis, 6,295 SNPs were dropped that deviated from HWE at p <0.00001, an additional 29,071 SNPs with MAF < 5% in only the cases or only in the controls were also dropped. The total genotyping rate in the remaining 1722 cases and 1774 controls was 99.312% across 2,083,536 SNPs. After analysis, the genomic inflation factor was 1.04.

GAIN SZ

When data were cleaned, 24 subjects from eight trios and 60 SNPs were removed due to Mendelian errors, 30 intentional duplicates, and one case with missing phenotypes were removed. Before imputing, 44 parents were removed. The pre-imputed file had 729,394 and 1343 cases and 1378 controls. This data was then imputed using the methods described below. The imputed files contained 1,966,598 SNPs and 2,721 subjects. On final analysis, 77 markers were dropped that deviated from HWE at p <0.00001, along with an additional 27,220 SNPs with MAF < 5 in only the cases or only in the controls. The total genotyping rate in the remaining 1,343 cases and 1,378 controls was 99.307% across 1,939,306 SNPs. After analysis, the genomic inflation factor was 1.066.

NIA/NINDS PD

The data provided by A. Singleton and M. Nalls were fully imputed, using the methods described below. The imputed files contained 2,143,749 SNPs and 1793 subjects. On final analysis, 3,346 markers were dropped that deviated from HWE at p <0.0001, along with an additional 41,021 SNPs with MAF < 5% in only the cases or only in the controls. The total genotyping rate in the remaining 984 cases and 809 controls was 99.0545% across 2,099,397 SNPs. After analysis, the genomic inflation factor was 1.05.

Whole-genome imputation

Imputation was performed as follows: genotype data from the respective test samples were used to impute data on 2.1 million HapMap Phase 2 (www.hapmap.org) SNPs using the MArkov Chain Haplotyping (MACH) program, version 1.0 (4). MACH uses Markov chain haplotyping to resolve haplotypes—and therefore missing genotypes—from observed genotypes in unrelated individuals. We used the “greedy” algorithm, as recommended by the authors. SNPs flagged as having different alleles than in HapMap CEU or as monomorphic were reviewed and then subsequently either recoded for the reverse strand (flipped) or dropped. SNPs that were flagged for allele frequencies that were markedly different than HapMap CEU were also reviewed. Palindromic SNPs whose allele frequencies were consistent with reversed coding were flipped. Other SNPs with unexpected allele frequencies were dropped. PLINK was used to flip and drop SNPs as necessary. After all allele coding, monomorphic, and palindromic issues were resolved, imputation was run again. SNPs in the results files were dropped if the MAF in cases or controls was below 5% or if the error rate (as reported in the .erate output file) was >0.01. Finally, the imputed data were formatted into PLINK binaries for analysis.

References

1. Purcell SM, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-75
2. Abecasis GR, Cherny SS, Cookson WO, Cardon LR: GRR: graphical representation of relationship errors. Bioinformatics 2001;17:742-3
3. Sullivan PF, de Geus EJ, Willemsen G, James MR, Smit JH, Zandbelt T, Arolt V, Baune BT, Blackwood D, Cichon S, Coventry WL, Domschke K, Farmer A, Fava M, Gordon SD, He Q, Heath AC, Heutink P, Holsboer F, Hoogendijk WJ, Hottenga JJ, Hu Y, Kohli M, Lin D, Lucae S, Macintyre DJ, Maier W, McGhee KA, McGuffin P, Montgomery GW, Muir WJ, Nolen WA, Nothen MM, Perlis RH, Pirlo K, Posthuma D, Rietschel M, Rizzu P, Schosser A, Smit AB, Smoller JW, Tzeng JY, van Dyck R, Verhage M, Zitman FG, Martin NG, Wray NR, Boomsma DI, Penninx BW: Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. Mol Psychiatry 2009;14(4):359-75
4. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 2010;34:816-34