Supplementary Methods

Quality control

Quality control procedures were applied in PLINK (1), initially at the level of marker and then at the level of individual. Markers were retained if they had a minor allele frequency (MAF) of 0.01 or more as effects of rare markers would be uninterpretable with the present sample size. Markers were filtered for genotyping completeness of 99% so that all analyses were performed in a comparable set of individuals (this is important for a continuous trait analyses, as individuals with extreme trait values contribute disproportionately to an association). Hardy-Weinberg Equilibrium (HWE) was calculated using the exact test in PLINK, but was not used as a filter as departures from HWE are expected in a case-only sample (2). Probability of departure from HWE is given with all reported markers.

At the individual level, genotypes were first tested for sex mismatch with phenotypic data. Samples with ambiguous genotypic sex and outliers on autosomal heterozygozity were identified for exclusion as these may indicate sample contamination. Related individuals were ascertained through estimation of identity by descent (IBD) applied in PLINK (1) to an LD-pruned dataset (the same as for analysis of population stratification, see below) and one of each pair of first- or second-degree relatives (the one with less complete data) was excluded. Finally, genotyping completeness was assessed for each individual and outliers with genotyping completeness less than 95% were excluded from further analyses.

Population stratification

Although recruitment was restricted to individuals of white European parentage, uniform self-reported ethnicity does not exclude genetic admixture and stratification within European populations has been reported (3). Therefore, we used principal component analysis applied in EIGENSTRAT (4) to detect population structure and control for it in the analyses. To avoid confounding by local linkage disequilibrium (LD), the principal component analysis was performed on an LD-pruned dataset of 39,658 SNPs in low LD and excluding known regions of long-range LD (5). After exclusion of outlier individuals, detected significant principal components (at $p<0.05$) were explored for association with centre of recruitment and principal components that were significant and unevenly distributed across national groups were used as covariates in the main analyses.

In the iterative EIGENSTRAT analysis of population stratification, five individuals were identified as outliers and removed as a joint analysis with HapMap data indicated that they had strong Asian or African admixtures. These individuals were excluded from all further analyses (Figure 1). After these exclusions, the first five principal components were nominally significant ($p<0.05$), with eigenvalues 2.021, 1.833, 1.571, 1.541 and 1.438, and the first four had a clear relationship with subject's geographic origin (Supplementary Figure S1, A, B). Principal components 1 and 2 roughly corresponded to the North-South and East-West geographical distinctions respectively, and jointly provided an almost perfect separation between populations of Norman/Anglo-Saxon (Danish, English, German) and Slavonic (Polish, Slovenian, Croatian) origins. Principal components 3 and 4 jointly distinguished most participants recruited in the UK from the rest of the sample. Principal components 5, 6 and further had no obvious relationship with centre of recruitment and explained progressively less variance (Supplementary Figure S1, C). Therefore, the first four principal components were used as covariates in the main analyses to control for genetic stratification.

Power analysis

As this is one of the first genome-wide pharmacogenetic studies, it is unclear what strengths of associations can be expected. If a single genetic marker is to be used as a clinical test to inform decisions about the care of an individual patient, it would have to explain at least 5% of variation in outcome. On the other hand, commonly reported effects in genetic case-control or qualitative trait studies rarely explain more than 1-2% of variance. To facilitate the interpretation of both positive and negative results, a power analysis was performed using the program QUANTO(6) for pharmacogenetic associations explaining 2 to 5% of variance in outcome. The sample of 706 individuals provides a power of 0.71 to detect a pharmacogenetic effect or gene-drug interaction explaining 5% of variance in antidepressant response at the genome-wide significance threshold of $\alpha=5\text{x}10^{-8}$ and a power of 0.93 to detect such effect at the suggestive significance threshold of $\alpha=5\text{x}10^{-6}$. For a pharmacogenetic effect explaining 2% of variance, the power would be only 0.05 at $\alpha=5\text{x}10^{-8}$ and 0.22 at $\alpha=5\text{x}10^{-6}$. The sample of 394 individuals treated by escitalopram provides a power of 0.17 to detect a pharmacogenetic effect explaining 5% of variance at $\alpha=5\text{x}10^{-8}$ and a power of 0.47 to detect such effect at $\alpha=5\text{x}10^{-6}$. The sample of 312 individuals treated by nortriptyline provides a power of 0.07 to detect a pharmacogenetic effect explaining 5% of variance at $\alpha=5\text{x}10^{-8}$ and a power of 0.29 to detect such effect at $\alpha=5\text{x}10^{-6}$. In summary, the genome-wide pharmacogenetic analysis of the GENDEP sample is powered to detect strong clinically significant pharmacogenetic effects explaining 5% of variance or more but is bound to miss a large proportion of moderate or weak pharmacogenetic effects.

Bioinformatics Analysis

To explore the extent of the associated genomic regions, all HapMap phase II SNPs within 100kb upstream and downstream of markers identified at genome-wide or suggestive significance levels of significance were imputed and tested for association using Markov Chains algorithm (with 50 iterations) implemented in the MACH1 software (7,8). Associated markers were evaluated for LD with other known markers, using the SNAP tool (http://www.broad.mit.edu/mpg/snap/). Common polymorphic CNVs in the associated regions were identified using the DGV Structural Variation track hosted by the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgTracks). The UCSC genome browser was also used to identify genes and expressed sequence tags (EST) mapping to the associated regions and for production of supplementary figures. Identification of Exonic Splice Enhancer sites was performed by analysis of both SNP alleles using ESE finder (http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi (9).

References

1. Purcell S, 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(3):559-575

2. Wittke-Thompson JK, Pluzhnikov A, Cox NJ: Rational inferences about departures from Hardy-Weinberg equilibrium. Am J Hum Genet 2005; 76(6):967-986

3. Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, Silva G, Belmont JW, Klareskog L, Gregersen PK: European population substructure: clustering of northern and southern populations. PLoS Genet 2006; 2(9):e143

4. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006; 38(8):904-909

5. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, Ge D, Rotter JI, Torres E, Taylor KD, Goldstein DB, Reich D: Long-range LD can confound genome scans in admixed populations. Am J Hum Genet 2008; 83(1):132-135

6. Gauderman WJ, Morrison JM: Quanto 1.1: A computer program for power and sample size calculations for genetic-epigemiology studies. http://hydra usc edu/gxe 2006;

7. Li Y, GR A: Rapid Haplotype Reconstruction and Missing Genotype Inference. Am J Hum Genet 2006; S792290

8. Li Y, Willer CJ, Cristen J, Ding J, Scheet P, Abecasis GR: Markov Model for Rapid Haplotyping and Genotype Imputation in Genome Wide Studies. (unpublished manuscript) 2006;

9. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR: ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res 2003; 31(13):3568-3571