# Common Genetic Variants and Gene-Expression Changes Associated with Bipolar Disorder Are Over-Represented in Brain Signaling Pathway Genes

Inti Pedroso, Anbarasu Lourdusamy, Marcella Rietschel, Markus M. Nöthen, Sven Cichon, Peter McGuffin, Ammar Al-Chalabi, Michael R. Barnes, and Gerome Breen

**Background:** Despite high heritability, the genetic variants influencing bipolar disorder (BD) susceptibility remain largely unknown. Low statistical power to detect the small effect-size alleles believed to underlie much of the genetic risk and possible heterogeneity between cohorts are an increasing concern. Integrative biology approaches might offer advantages over genetic analysis alone by combining different genomic datasets at the higher level of biological processes rather than the level of specific genetic variants or genes. We employed this strategy to identify biological processes involved in BD etiopathology.

**Method:** Three genome-wide association studies and a brain gene-expression study were combined with the Human Protein Reference Database protein–protein interaction network data. We used bioinformatic analysis to search for biological networks with evidence of association on the basis of enrichment among both genetic and differential-expression associations with BD.

**Results:** We identified association with gene networks involved in transmission of nerve impulse, Wnt, and Notch signaling. Three features stand out among these genes: 1) they localized to the human postsynaptic density, which is crucial for neuronal function; 2) their mouse knockouts present altered behavioral phenotypes; and 3) some are known targets of the pharmacological treatments for BD.

**Conclusions:** Genetic and gene-expression associations of BD cluster in discrete regions of the protein–protein interaction network. We found replicated evidence for association for networks involving several interlinked signaling pathways. These genes are promising candidates to generate animal models and pharmacological interventions. Our results demonstrate the potential advantage of integrative biology analyses of BD datasets.

**Key Words:** Bipolar disorder, depression, GWAS, mania, networks, pathways, postsynaptic density, systems biology

Bipolar disorder (BD) is a chronic and episodic psychiatric illness characterized by extremes of mood ranging from mania to severe depression. Despite a convincing and substantial genetic contribution to the etiology of the disorder (1), its genetic and molecular underpinnings remain largely unknown. Its diagnosis is based solely on observed clinical features. Individual genome-wide association studies (GWAS) and linkage studies have highlighted several genomic regions, and recently replicated evidence implicating specific loci have also been reported (2–6). The GWAS of common genetic variation have reinforced the notion that many low-risk genetic variants are involved in the etiology of BD. Therefore, an important challenge of genetic studies is to devise analytical strategies to extract biologically relevant associations from those under the genome-wide significance threshold needed for multiple testing correction, $p = 5 \times 10^{-8}$ (7). Currently, large meta-analyses of GWAS represent the major approach used to increase power to detect BD risk alleles (8). A potential limitation to these studies is allelic and locus heterogeneity (i.e., two or more polymorphisms within a gene being independently associated, and different sets of genes associated in different studies). This has been reported in different diseases (9–11). There is some evidence to suggest this might also be true for BD (12), although the extent to which this occurs in BD or other traits is not yet clear.

Prior knowledge can be used to boost signal-to-noise ratio and tackle heterogeneity in large-scale genomic experiments (13,14). Prior information can be used to filter out data, on the basis that they are unlikely to carry useful information, or to aggregate them into biologically relevant groups to allow their signal to stand out above the noise generated by multiple testing. Such an approach can be used with multiple data sources, and increasing evidence suggests that gene-expression studies can help prioritize GWAS results (15–17). For example, Zhong *et al.* (16) showed that gene-expression changes and disease-susceptibility alleles cluster in common biological pathways conferring risk for type 2 diabetes.

Analysis of the Wellcome Trust Case Control Consortium (WTCCC) BD GWAS and its meta-analysis with another GWAS provided evidence of association within biological processes involved in the modulation of transcription and cellular activity, including that of hormone action and adherens junctions (11,18,19).

Here we present an integrative biology analysis aimed to identify biological processes associated with BD susceptibility. Three GWAS of BD susceptibility were integrated via a gene-wide analysis followed by protein–protein interaction network (PPIN) analysis

From the Medical Research Council Social, Genetic and Developmental Psychiatry Centre (IP, AL, PM, MRB, GB); National Institute for Health Research Biomedical Research Centre for Mental Health at South London and Maudsley National Health Service Foundation Trust (IP, AL, PM, GB); Medical Research Council Centre for Neurodegeneration Research (AA-C), Department of Clinical Neuroscience, Institute of Psychiatry, King's College London, London; Computational Biology (MRB), GlaxoSmithKline Pharmaceuticals, Stevenage, Hertfordshire, United Kingdom; Central Institute of Mental Health (MMN, SC), Division of Genetic Epidemiology in Psychiatry, Mannheim; Institute of Human Genetics (MMN, SC), University of Bonn; Department of Genomics (MMN, SC), Life & Brain Center, University of Bonn; German Center for Neurodegenerative Diseases (DZNE) (MMN), Bonn; Institute of Neuroscience and Medicine (INM-1) (SC), Research Center Juelich, Juelich, Germany.

Address correspondence to Gerome Breen, Ph.D., King's College London, Institute of Psychiatry, Medical Research Council Social, Genetic and Developmental Psychiatry Centre, P080 SGDP, De Crespigny Park, London SE5 8AF, UK; E-mail: gerome.breen@kcl.ac.uk.

Received Apr 28, 2011; revised Dec 13, 2011; accepted Dec 15, 2011

0006-3223/$36.00
doi:10.1016/j.biopsych.2011.12.031

and comparison with a brain gene-expression study of BD patients and matched control subjects. Our integrative approach revealed convergent evidence for association of genes and biological processes with BD susceptibility.

## Methods and Materials

### Samples and Genotype Data

We reanalyzed GWAS of BD from the Wellcome Trust Case Control Consortium (20), Cichon *et al.* (4), and Sklar *et al.* (21) studies, which we refer to as WTCCC, Bonn, and Sklar, respectively. We used individual level genotype and phenotype data from the Bonn study and summary statistics from each of the other studies. Genotype data from the Bonn study were quality controlled by the sample and single nucleotide polymorphism (SNP) missing rate and Hardy-Weinberg. All BD samples met DSM-IV criteria to establish BD diagnoses. We also analyzed summary statistics from six GWAS in common nonpsychiatric disorders reported by the WTCCC (20). See Supplement 1 for additional method descriptions.

### Gene-Based Association

We calculated gene-wide *p* values with the FORGE software suite (see Supplement 1 for a detailed description of the software). We included in our analyses approximately 21,000 protein-coding, long noncoding RNA and micro RNA genes annotated in Ensembl v59 (www.ensembl.org) and mapped them to SNPs if the SNP was within 20 kb of the annotated coordinates. The FORGE software combined the *m* association *p* values within genes with the fixed-effects *Z* score method

$$Z_{fix} = \left( \frac{\sum_{i=1}^{m} W_i Z_i}{\sum_{i=1}^{m} W_i} \right) \cdot \sqrt{V_{fix}}$$

where $z_i$ are the *p* values transformed to *Z* scores with the standard normal distribution inverse cumulative distribution function (c.d.f.) and $V_{fix}$ is the variance of $Z_{fix}$. With the approximation of the multivariate-normal distribution

$$V_{fix} = \sum_{i=1}^{m} \sum_{j=1}^{m} w_i w_j \rho_{ij}$$

where *w* are weights that we set to $1/m$ and $\rho_{ij}$ is the correlation between the $z_i$ and $z_j$ that we approximate as the correlation between the $SNP_i$ and $SNP_j$, because we only use summary statistics. We used the simulation-based strategy of Liu *et al.* (22) to estimate the significance of the $Z_{fix}$ statistics, because it was shown to provide very good correlation with empirical estimates. Briefly, N gene-wide statistics for each gene were calculated with sets of *m* random Gaussian variables (*Z* scores) with correlations defined by the correlation matrix between the SNPs (see Liu *et al.* [22] for details). We set N to a maximum of $10^6$ simulations or until the value of $Z_{fix}$ was observed 10 times. The gene-wide association test significance is equal to $(R+1)/(N+1)$, where R is the number of times a statistic $\geq Z_{fix}$ was observed.

Application of these methods to GWAS has been reported elsewhere (23). Before calculating gene *p* values we applied genomic control to the SNP *p* values, if the study $\lambda$ median was $>1$ (24).

### Network Analyses

To identify subnetworks of interacting genes enriched with genetic associations, we used the greedy search introduced by Ideker *et al.* (25), which we implemented in a Perl script distributed together with FORGE. The algorithm starts subnetwork searches from each node (seed node) in the PPIN. A subnetwork is defined by sequentially adding the direct neighbors of the nodes of the subnetworks (initially only the seed node). We allowed searches to go to a maximum of five interactions from the seed node and generate subnetworks of 2 to 500 nodes in size. For each dataset, we calculate the aggregate *Z* score of the subnetwork ($S_{Net}$) with

$$S_{Net} = \left( \frac{\sum_{d=1}^{k} W_i g_i}{\sum_{d=1}^{k} W_i} \right) \cdot \sqrt{V_{SNet}}$$

where $g_i$ is the *Z* score of the *i*th gene in the subnetwork, *k* is the number of genes in the subnetwork, and $V_{genes}$ is the variance-covariance matrix of the statistic of the gene that we calculated with the method described by Luo *et al.* (23),

$$V_{SNet} \approx corr(g_i, g_j) = \frac{\sum_{u=1}^{ki} \sum_{v=1}^{kj} corr(z_{iu}, z_{jv})}{\sqrt{\left(\sum_{u=1}^{ki} \sum_{u=1}^{ki} \rho_{ij}\right) \cdot \left(\sum_{v=1}^{kj} \sum_{v=1}^{kj} \rho_{ij}\right)}}$$

To identify groups of highly overlapping networks, we constructed a gene-to-network membership matrix filled with values 0 or 1, depending on whether a gene was part of the network or not. This matrix was used to calculate correlations between networks and to perform hierarchical clustering with the Heatplus R package (http://www.bioconductor.org/packages/2.3/bioc/html/Heatplus.html).

### DNA Microarray Analysis

The gene-expression data of dorsolateral prefrontal cortex (DLPFC) tissue from 61 subjects and orbitofrontal cortex (OFC) tissue from 21 subjects reported by Ryan *et al.* (26) were downloaded from the ArrayExpress database (27) under accession number E-GEOD-5392. Raw intensity values of Affymetrix Human Genome U133A arrays were normalized with the Robust Multi-Average algorithm (28). Pre-filtering removed transcripts not detected (marked as "absent" with MAS5 detection call algorithm) in any sample and were not considered further. The generalized linear model with covariates was used to assess differential expression for each probe in each brain region. Covariates were used as in the original report: for the DLPFC samples we used the generalized linear model with disease status (control/BD) as the main effect while controlling for brain pH and fluphenazine equivalents; and for the OFC samples, we used fluphenazine equivalents as a covariate.

### In Silico Characterization of Significant Networks

We interpreted the biology of significant subnetworks with MetaCore (GeneGo; http://www.genego.com). GeneGO provides gene ontologies as GeneGO Pathways Maps and Network Processes manually constructed from literature review. There are defined as: 1) GeneGO Pathways summaries of established, noncontradictory state-of-the-art knowledge on the major functional categories of human metabolism and cell signaling; and 2) Networks Processes descriptive of a biological function but containing more information than a Pathway Map and possibly having newer published results on them. Enrichment of subnetwork genes in these biological categories is calculated with hypergeometric distribution statistics as has been described elsewhere (29). For all analyses in GeneGO, we used the intersection between the PPIN and the three GWAS dataset genes as a background to account for biases in PPIN gene annotation. This reference list had 7924 genes, of which 3375 could be mapped to GeneGO Pathways and 7917 could be mapped to GeneGO Processes. In addition we compiled genes associated with mouse phenotypes by parsing the files provided by the Mammalian Phenotype Ontology database (30). Genes localized in the human postsynaptic density (hPSD) were obtained from the supplementary material of Bayes *et al.* (31). Enrichment of genes for membership of the Mammalian Phenotype Ontology, and the hPSD categories were calculated with