Citation: Transl Psychiatry (2012) 2, e154; doi:10.1038/tp.2012.80

www.nature.com/tp

Epistasis network centrality analysis yields pathway replication across two GWAS cohorts for bipolar disorder

A Pandey[1], NA Davis[1], BC White[1], NM Pajewski[2], J Savitz[3,4], WC Drevets[3,5] and BA McKinney[1,3]

Most pathway and gene-set enrichment methods prioritize genes by their main effect and do not account for variation due to interactions in the pathway. A portion of the presumed missing heritability in genome-wide association studies (GWAS) may be accounted for through gene–gene interactions and additive genetic variability. In this study, we prioritize genes for pathway enrichment in GWAS of bipolar disorder (BD) by aggregating gene–gene interaction information with main effect associations through a machine learning (evaporative cooling) feature selection and epistasis network centrality analysis. We validate this approach in a two-stage (discovery/replication) pathway analysis of GWAS of BD. The discovery cohort comes from the Wellcome Trust Case Control Consortium (WTCCC) GWAS of BD, and the replication cohort comes from the National Institute of Mental Health (NIMH) GWAS of BD in European Ancestry individuals. Epistasis network centrality yields replicated enrichment of Cadherin signaling pathway, whose genes have been hypothesized to have an important role in BD pathophysiology but have not demonstrated enrichment in previous analysis. Other enriched pathways include Wnt signaling, circadian rhythm pathway, axon guidance and neuroactive ligand-receptor interaction. In addition to pathway enrichment, the collective network approach elevates the importance of *ANK3, DGKH* and *ODZ4* for BD susceptibility in the WTCCC GWAS, despite their weak single-locus effect in the data. These results provide evidence that numerous small interactions among common alleles may contribute to the diathesis for BD and demonstrate the importance of including information from the network of gene–gene interactions as well as main effects when prioritizing genes for pathway analysis.
Translational Psychiatry (2012) **2,** e154; doi:10.1038/tp.2012.80; published online 14 August 2012

Introduction

Genome-wide association studies (GWAS) of psychiatric disorders (schizophrenia, bipolar disorder (BD), major depressive disorder and others) have suggested a highly polygenic architecture[1] with a high degree of heterogeneity. Given the relative lack of replicated common risk variants[2–4] with a large effect size, interest has turned to other potential explanations (including rare variants and epistasis[5–9]) for the presumed missing heritability.[10,11] Recent analyses have suggested that a substantial proportion of additive genetic variability is in fact well tagged by common variants when considered in aggregate, for example, explaining ~37–40% of the genetic variability for BD.[12,13] These analyses have also suggested that the remaining missing heritability may be a function of imperfect linkage disequilibrium with rare causal risk variants. Although a large degree of additive genetic variance is supported both theoretically and empirically, it is important to note that a large additive contribution to genetic variance does not preclude the contribution of models involving epistasis between single-nucleotide polymorphisms (SNPs).[14,15] The variation encoded in the nodes and edges may be used to estimate the amount of additional variation accounted for by the epistasis network. However, the goal of the current study is to demonstrate the senstivity of epistasis networks to discover new susceptibility genes in GWAS.

The recognition that numerous variants act together to increase disease susceptibility has also led to the development of gene-set or pathway enrichment approaches, which aggregate association evidence at the level of a single gene or biological pathway.[16–18] As applied to SNP data, these approaches typically rely on association evidence calculated marginally for each SNP, thus ignoring potential effects due to interactions.[19,20] Here, we consider a network approach that prioritizes genes and pathways based on the aggregation of effects due to gene–gene interactions as well as marginal (main) effects. This approach consists of four main steps, summarized in Figure 1: (1) filtering to remove noise SNPs from consideration, (2) representing association evidence in terms of an epistasis network, (3) prioritizing SNPs/genes in the network using an eigenvector centrality algorithm and

[1]Tandy School of Computer Science, Department of Mathematics, University of Tulsa, Tulsa, OK, USA; [2]Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA; [3]Laureate Institute for Brain Research, Tulsa, OK, USA; [4]Department of Medicine, Tulsa School of Community Medicine, University of Tulsa, Tulsa, OK, USA and [5]Department of Psychiatry, University of Oklahoma College of Medicine Tulsa, Tulsa, OK, USA
Correspondence: Dr BA McKinney, Tandy School of Computer Science, Department of Mathematics, University of Tulsa, Rayzor Hall, 800 South Tucker Drive, Tulsa, OK 74104, USA.
E-mail: brett.mckinney@gmail.com
Keywords: eigenvector centrality; epistasis network; evaporative cooling machine learning feature selection; pathway enrichment analysis; regression-based genetic association interaction network (reGAIN); SNPrank
Received 5 July 2012; accepted 8 July 2012

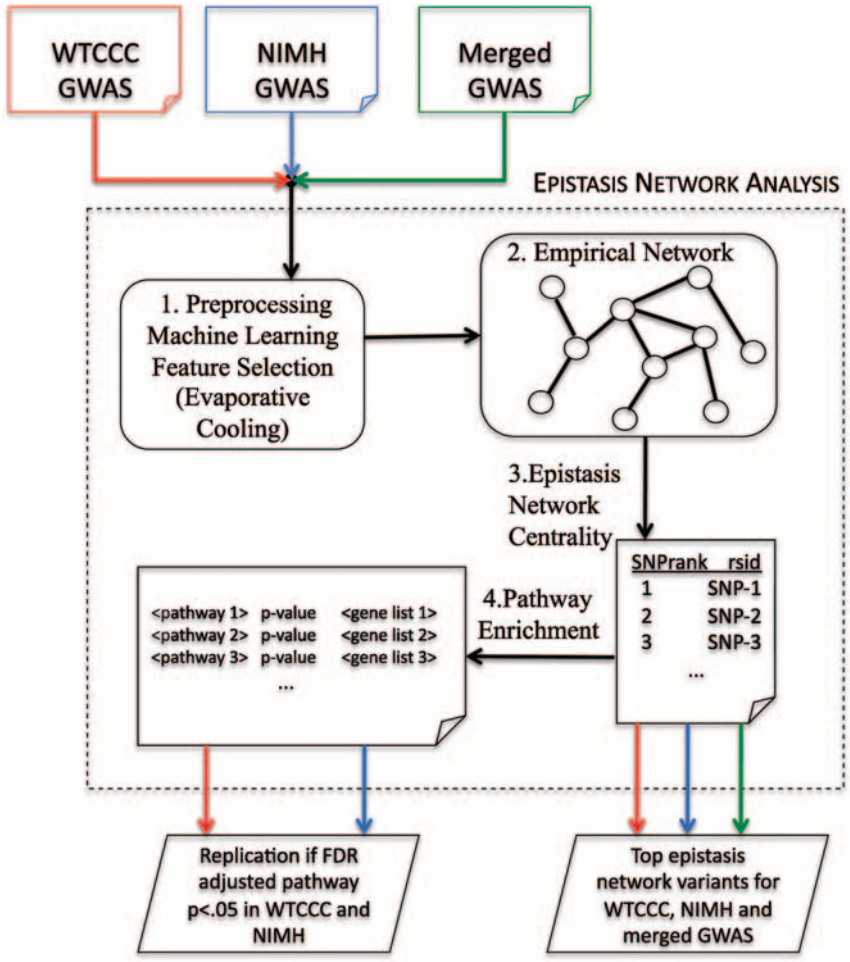

Figure 1 Epistasis network analysis flowchart. Overview of the data analysis workflow used to identify variants due to epistasis network centrality and test for replication of pathways. The analysis steps in the dotted frame are carried out for the three GWAS at the top (WTCCC, NIMH and, as a secondary analysis, the two GWAS combined). On the bottom left, the enriched pathways are compared between the WTCCC and NIMH GWAS, and replication is defined when a pathway has an FDR-adjusted *P*-value less than 0.05 for both. On the bottom right, tables are created for the top genes based on their epistasis network centrality for each of the data combinations.

(4) pathway enrichment based on epistasis network centrality prioritization. We first remove noise SNPs with an optimized version of the evaporative cooling machine learning (ECML) filter. We have shown that the ECML filter, which is based on the combination of Relief-F and Random Forests, has the power to detect both epistatic and main effects, whereas Random Forest alone has very weak power to detect epistatic effects in high dimensional data.[21]

We have previously used information theory to construct epistasis networks (which we label as a genetic association interaction network, itGAIN); however, in the present study, we rely upon regression models, primarily to be able to assign statistical significance to nodes and edges inferred in the network (which we label as a regression-based genetic association interaction network (reGAIN) to differentiate it from an information theory-based approach). Other groups have recently investigated the graph properties of epistasis networks, illustrating, for example, that hub (highly connected) SNPs do not necessarily correspond to SNPs with large main effects.[22] For the final step in our approach, we prioritize edges and nodes in the epistasis network using an eigenvector centrality algorithm we have developed called SNPrank.[21,23] SNPrank can be understood by the analogy of a random SNP surfer circulating through the network, accumulating bits of interaction and main effect information from each SNP regarding association with the phenotype. In a previous application of this approach to a genetic association study of the immune response to smallpox vaccine, we identified an intronic SNP in the retinoid X receptor α (*RXRA*) gene, which is known to be a mediator of vitamin D signaling and has recently been shown to be involved in innate immune response.[23]

Here, we apply the combined approach of ECML + reGAIN + SNPrank to two previous GWAS of BD: the Wellcome Trust Case Control Consortium (WTCCC)[24] and a more recent National Institute of Mental Health (NIMH) GWAS.[25] The original WTCCC study of BD, consisting of 1868 cases and 2938 controls, did not find any single SNP associations surpassing commonly accepted thresholds for genome-wide significance ($P<5\times10^{-8}$). However, a recent collaborative analysis of BD, which combined the WTCCC data with other studies for an overall sample of 4387 cases and 6209 controls, found a strong association for the imputed SNP rs10994336

(*ANK3*) on chromosome 10q21 ($P = 9.1 \times 10^{-9}$).[26] In the recent NIMH GWAS of European ancestry (EA) and African ancestry (AA) individuals, no SNP reached genome-wide significance. However, in the EA samples (1001 cases and 1033 controls), which we analyze in the current study, a sliding-window analysis yielded a high proportion of haplotypes with $P < 0.05$ in the *ANK3* region. In the current work, we observe a highly connected SNP in *ANK3* that is ranked third by SNPrank in our epistasis network analysis of the original WTCCC GWAS, and the network rank of this variant is second when the WTCCC and NIMH-EA GWAS are merged. The network analysis of the merged data yields a top-10 ranking to a SNP in diacylglycerol kinase eta (*DGKH*), which was implicated for BD in a previous study,[27] and top 15 for *ODZ4*, which has been identified in Sklar *et al*.[28] The top genes based on epistasis network analysis for the merged GWAS are given in Table 3.

The epistasis network prioritization also results in enrichment of plausible biological pathways for BD that replicate between the WTCCC[24] and the NIMH BD GWAS.[25] Using the epistasis network centrality for gene prioritization based on the Reactome FI database,[29] we find replication of enrichment of the cadherin signaling pathway and evidence consistent with replication in the Wnt signaling pathway. Genes in the Cadherin pathway have been implicated in BD pathophysiology.[30] In addition, it has been suggested that BD is affected by genes in the Wnt Signaling pathway as well as the circadian rhythm pathway,[31] which are both enriched in the WTCCC GWAS by this approach. Other enriched pathways include axon guidance and neuroactive ligand-receptor interaction. The identification of replicated pathways suggests that network aggregation of gene–gene interactions and main effects can provide statistical power to expose hidden variation associated with complex diseases. These results also indicate the importance of taking into account the information concerning epistasis as well as main effects when prioritizing genes for pathway analysis.

Materials and methods

Study samples and initial filtering. For the primary/discovery epistasis network analysis, we used the WTCCC-BD GWAS, which included bipolar I, bipolar II and schizoaffective bipolar in the case diagnosis.[24] Samples (including 1868 cases and 2938 controls after exclusions) were genotyped on the Affymetrix 500K array (Santa Clara, CA, USA). For replication, we used the NIMH-BD GWAS genotyped on the Affymetrix 6.0 platform.[25] The NIMH BD study involved a sample of individuals of EA ($n = 1001$ cases; $n = 1033$ controls), and one involving a sample of individuals of AA ($n = 345$ cases; $n = 670$ controls). We focus on the EA individuals from the NIMH study because the effect of admixture on these machine learning and network techniques has not been fully investigated. The case diagnosis included bipolar I and schizoaffective bipolar. For both studies, we removed SNPs with call rates $< 95\%$, minor allele frequency $< 1\%$, or with evidence of deviation from Hardy–Weinberg equilibrium ($P < 0.001$). As a secondary analysis, we merged the top SNPs from the WTCCC and NIMH-EA cohorts. In the merged data, we only include overlapping SNPs between the Affy 6.0 and 500K chips rather than impute missing SNPs. Imputation may allow for the discovery of additional genes and pathways.

We now detail the methods used in the steps of the analysis pipeline, which is summarized in Figure 1. To limit the number of noise (irrelevant) SNPs used in the network analysis, we filtered SNPs based on ECML, which has power to detect main and interaction effects.[21] We used the 1000 SNPs with the top ECML score to construct a reGAIN, as described below. Any filter increases the risk of excluding pure interaction effects that exhibit negligible marginal effects as well as excluding some weak main effects. However, filtering reduces the number of pairwise interactions that must be calculated, eliminates many irrelevant variants and improves interpretability of the network. The filter used herein retains many more potential interaction effects and is approximately two order of magnitude more SNPs than the threshold used by WTCCC to define moderate associations in their Supplementary data ($P < 0.0001$).

Regression-based epistasis network construction (reGAIN). From the 1000 SNPs remaining after the ECML filter, we construct a GAIN/epistasis network composed of main effects and gene–gene interactions between all pairs. Our previous data-driven GAIN network approach for GWAS used Shannon information theory for epistasis calculations and network construction.[21,23] However, casting the network in the statistical framework of a general linear model has some advantages over information theory. For example, use of a general linear model framework provides the flexibility to handle environmental covariates, longitudinal data, missing data, censoring and cluster structure (for example, family studies) through the inclusion of appropriate random effects. For the BD GWAS, we use a likelihood ratio test of association between disease and a genetic locus, allowing for the possibility that the genetic effect may be modified by another genetic factor.

$$\log \frac{\Pr(D=1 \mid G_1, G_2)}{\Pr(D=0 \mid G_1, G_2)} = b_b + b_1 G_1 + b_2 G_2 + b_{12} G_1 G_2$$

The coefficient b_b gives the baseline risk of disease and coefficients b_1 and b_2 correct for main effects in the interaction regression model. For defining gene–gene edge weights b_{12} in the reGAIN, we are interested in the b_{12} regression coefficients that are statistically different from zero. The statistical framework also allows false discovery rate procedures to be applied to correct for multiple gene–gene hypotheses. The diagonal element b_{ii} of the reGAIN is simply the main effect regression coefficient without interactions. These interaction and main effect regression coefficients for all SNPs in the filter become matrix elements in the SNPrank Markov transition matrix, discussed next.

Eigenvector network centrality (SNPrank) for gene prioritization in pathway enrichment. We use the SNPrank[23] network centrality/importance score to prioritize the 1000 SNP nodes in the reGAIN for pathway enrichment. This score accounts for main effects and gene–gene interactions encoded in the reGAIN matrix. Briefly, SNPrank constructs a stochastic transition matrix from the reGAIN

matrix B (see above). The matrix accounts for single-locus effects through the main effects along the diagonal b_{ii} and accounts for pair-wise interactions through the interaction coefficients b_{ij} on the off-diagonal elements. Higher-order interactions (linear combinations of multiple pair-wise interactions and main effects) are incorporated through a recursive power method to calculate the dominant eigenvector of the transition matrix. The elements of the dominant eigenvector are the SNPrank scores of each genetic node in the reGAIN. The eigenvector is normalized so the elements sum to one, like a probability field. Thus, we use a QQ plot to estimate the number of genes to include in pathway enrichment below; we use the top $n = 200$ genes for both GWAS (WTCCC and NIMH).

Pathway enrichment analysis. To identify enriched pathways from the $n = 200$ top genes, we used the Reactome FI database[29] of expert-curated human biological pathways. Reactome pathways are described as a series of molecular events that transform one or more input entities into one or more output entities catalyzed or regulated by other entities. Entities include small molecules, proteins, complexes, post-translationally modified proteins and nucleic acid sequences. SNPs are assigned to genes based on proximity to the 5′ and 3′ ends of the first and last exons. For SNPs whose proximity is greater than 20 kb, we look for linkage disequilibrium information that may inform gene assignment.[32] If a SNP is not easily assigned, we do not use it in pathway analysis. We use this conservative approach to limit false positive assignments and false positive enriched pathways. Genes are not repeated in the enrichment if more than one SNP from a gene is found in the top list. We calculated the *P*-value for the significance of the overrepresentation of a biological pathway π_i with the hypergeometric distribution

$$P(\pi_i) = 1 - \frac{C^{m(i)}_{M(i)} C^{n-m(i)}_{N-M(i)}}{C^{n}_{N}},$$

where N is the number of background genes (genes annotated to any pathway), n is the number of top genes prioritized by SNPrank, $M(i)$ is the total number of genes in pathway π_i, whereas $m(i)$ is the number of top SNPrank genes that intersect the set of pathway genes π_i.

Two corrective measures were taken to reduce false positive pathway enrichments. The first is correction due to multiple hypothesis testing. All pathways tested for enrichment were sorted in ascending order and the corrected *P*-value was given by

$$\bar{p}(\pi_i) = P \frac{p(\pi_i)}{R(\pi_i)},$$

where P is the total number of pathways and R(x) is the rank order of pathway x. Second, we generated pathway-specific and GWAS-specific enrichment distributions to correct for gene-size bias. Gene length can bias pathway enrichment,[33] which can be particularly significant for large brain-function genes.[8] We select $n = 200$ SNPs randomly from the GWAS, map SNPs to genes and calculate $m_{rand}(i)$ (the number of the randomly selected genes that intersect the set of genes in pathway π_i). We repeat this sampling 1000 times to create a null distribution of m_{rand} for each pathway. If a pathway has a gene-size bias, this should be reflected in the random distribution of m_{rand}. We use the mean and standard deviation of $m_{rand}(i)$, to calculate a *z*-score and *P*-value for each observed $m(i)$ (from the epistasis network centrality ranking of the GWAS). The gene-size corrected *P*-value for Wnt signaling is $P = 0.000337$ for the WTCCC data and $P = 0.06$ for the NIMH data; and for cadherin signaling $P = 0.032$ for both WTCCC and NIMH. Cadherin signaling meets our replication criteria when corrected for multiple tests and gene length. Although Wnt signaling does not technically replicate when corrected for gene length, the consistency of high significance in WTCCC and near significance in NIMH make this pathway very suggestive for involvement in BD.

Network pruning with edge significance for visualization of network. For SNPrank gene ranking, we used the full network of ECML-filtered SNPs because we suspect multiple small interactions with potentially weaker significance will contribute to the overall expression of the phenotype. False connections have the potential to bias the network, but we expect the false edges to be randomly distributed. We did not observe a gene length bias that might artificially inflate the network importance of longer genes. For improved interpretation of the network, we pruned the network based on edge strength. We used an edge strength threshold of $b_{ij} = 0.575$ to highlight the gene nodes and edges that have the strongest effects and to reduce the obscuring effect (network hairball) of many weak connections. The maximum threshold was chosen (edges below this threshold were pruned) subject to the constraint of minimizing the number of network islands. Gene symbols are used to label nodes. If more than one SNP from a gene is found in the network, then the SNP with the highest SNPrank score represents the gene and its interactions.

Results

The Materials and methods section contains details of the regression-based epistasis-network pathway-enrichment analysis as well as descriptions of the WTCCC-BD[24] and NIMH-BD[25] GWAS data sets. In brief, the WTCCC-BD GWAS was used for discovery and NIMH-BD for replication. We retained the top 1000 SNPs based on ECML feature selection, which has demonstrated power to detect both main effects and gene–gene interactions in GWAS.[21] From these top 1000 SNPs, we constructed an epistasis network of main effects and gene–gene interactions between all pairs using the reGAIN method discussed below and in McKinney *et al.*[20] We then applied SNPrank[23] to the epistasis network to further remove noise SNPs and enrich the top list of SNPs for main effects and interactions. We retained the top genes for pathway enrichment analysis based on the QQ plot of the SNPrank eigenvector scores, which resulted in a cutoff of approximately 200 genes. This cutoff removes network nodes whose SNPrank scores are consistent with a uniform distribution in the range (0,1). We used the same cutoff for both the discovery and replication data sets to define the number of top genes for use in the hypergeometric distribution for pathway enrichment. We used pathway annotations from the Reactome FI pathway database.[29]

Table 1 WTCCC pathway enrichment

Pathway	P-*value*	*Genes in network*
Wnt signaling pathway(P)[a]	0.0008	*CTNNA2, DACT1, FBXW11, CDH16, CDH18, CDH10, CDH11, GNA14, SMARCA2, CDH2, CHD1L, FHL2, PRICKLE1, FAT3, HOXA6*
Neuroactive ligand-receptor interaction(K)	0.0008	*GRIN2B, GRIK2, GABRB1, NTSR1, CYSLTR2, ADRA2A, GABRG3, ADRB2, HRH2, LEP*
Cadherin signaling pathway(P)[a]	0.004	*CTNNA2, CDH16, CDH18, CDH10, CDH11, CDH2, FAT3*
Shigellosis(K)	0.0054	*ELMO1, FBXW11, DOCK1, ABL1*
Bacterial invasion of epithelial cells(K)	0.0085	*CTNNA2, ELMO1, CAV3, DOCK1*
Calcium signaling pathway(K)	0.0141	*ATP2B1, GNA14, NTSR1, CYSLTR2, ADRB2, HRH2*
CFTR and beta 2 adrenergic receptor (b2ar) pathway(B)	0.019	*AGT, ADRB2*
Circadian rhythm—mammal(K)	0.027	*FBXW11, BHLHE40*
Signaling events mediated by HDAC class III(N)	0.0292	*PPARGC1A, FHL2*
Receptor-ligand complexes bind G proteins(R)	0.0302	*AGT, GNA14, ADRA2A, ADRB2, HRH2*
ID(C)	0.0315	*ADD1, ID2*
Corticosteroids and cardioprotection(B)	0.0315	*AGT, ADRB2*
β-Arrestins in gpcr desensitization(B)	0.0338	*AGT, ADRB2*
Activation of camp-dependent protein kinase pka(B)	0.0338	*AGT, ADRB2*
Role of β-arrestins in the activation and targeting of map kinases(B)	0.0387	*AGT, ADRB2*
O-glycan biosynthesis(K)	0.0438	*GCNT1, GALNTL4*
Roles of β arrestin-dependent recruitment of src kinases in gpcr signaling(B)	0.0491	*AGT, ADRB2*

Genes were prioritized by epistasis network analysis as described in the Materials and methods. Pathways are shown with adjusted hypergeometric enrichment P-value < 0.05.
[a]These pathways suggest replication in the NIMH-BD GWAS for European ancestry (see Table 2).

Table 2 NIMH-EA pathway enrichment

Pathway	P-*value*	*Genes in network*
M phase(R)	0.009	*RPS27, NUF2, PPP2CA, SGOL1, KIF2A*
Cadherin signaling pathway(P)[a]	0.0094	*CDH10, PCDH7, CDH6, CDH8, CDH7, CDH9, FYN*
Glycosphingolipid biosynthesis—globo series(K)	0.0149	*B3GALT5, ST3GAL1*
Glycosaminoglycan biosynthesis—keratan sulfate(K)	0.017	*ST3GAL1, B4GALT1*
Syndecan-3-mediated signaling events(N)	0.0214	*FYN, MC4R*
Protein processing in endoplasmic reticulum(K)	0.0233	*STT3B, UGGT1, SEC61A1, SEL1L, SEC23B, PARK2*
Map kinase inactivation of smrt corepressor(B)	0.0238	*RXRA, THRB*
Axon guidance(K)	0.0282	*ARHGEF12, FYN, LRRC4C, CXCL12, ROBO2*
PDGFR-alpha signaling pathway(N)	0.0289	*RAPGEF1, CAV3*
LPA receptor-mediated events(N)	0.0397	*LPAR3, GNAL, TIAM1, TNFAIP3*
Ephrin B reverse signaling(N)	0.0403	*FYN, TIAM1*
RXR and RAR heterodimerization with other nuclear receptor(N)	0.0465	*RXRA, THRB*
Glycosphingolipid biosynthesis—lacto and neolacto series(K)	0.0497	*B3GALT5, B4GALT1*
Pyruvate metabolism and TCA cycle(R)	0.053	*PDHX, SUCLA2*
Reelin signaling pathway(N)	0.053	*RAPGEF1, FYN*
NR transcription pathway(R)	0.0599	*PGR, NR3C2*
Alpha-synuclein signaling(N)	0.0599	*FYN, PARK2*
Wnt signaling pathway(P)[a]	0.0606	*PPP2CA, CDH10, MYH13, PCDH7, CDH6, CDH8, CDH7, CDH9, SMARCAD1*

Genes were prioritized by epistasis network analysis as described in the Materials and methods and pathway enrichment adjusted P-values calculated by the hypergeometric distribution.
[a]These pathways were statistically significant in the WTCCC-BD GWAS (see Table 1).

We list the most significant epistasis network pathway enrichment results in Tables 1–2 for the WTCCC and NIMH GWAS of BD. We find replication evidence of enrichment of the cadherin signaling pathway ($P = 0.004$ in WTCCC and $P = 0.0094$ in NIMH-EA) and evidence of replication in the Wnt signaling pathway ($P = 0.0008$ in WTCCC and $P = 0.06$ in NIMH-EA). Genes in the cadherin pathway as well as protein partners in the Wnt pathway have been implicated as possible components of a molecular pathway in susceptibility to BD pathophysiology.[30] It has also been suggested separately that BD is affected by genes in the Wnt signaling pathway as well as the circadian rhythm pathway,[31] both enriched in the WTCCC GWAS by the epistasis network approach. These pathways are not significantly enriched when SNPs are prioritized by single-locus statistics as observed for example in the WTCCC-BD in Torkamani *et al.*[19] Other enriched pathways of note based on epistasis networks include axon guidance (NIMH-EA ($P = 0.028$)) and neuroactive ligand-receptor interaction (WTCCC ($P = 0.0008$)), which is also the most significantly enriched when the WTCCC and NIMH-EA GWAS are merged. Genes and edges for the WTCCC reGAIN network in Figure 2 are annotated by pathway membership for the replicated pathways.

Table 3 Top genes from epistasis network centrality of combined WTCCC + NIMH GWAS

Chromosome	*SNP rs-id*	*Gene symbol*	*SNPrank score*	*Univaraite odds ratio*	*Univariate* P-*value*
5	rs393291	DAP	7.61E-03	1.05	0.6388
10	rs10509126	**ANK3**	6.64E-03	1.192	0.01619
2	rs10190186	FHL2	6.63E-03	1.195	0.01106
4	rs7679912	ARAP2	6.41E-03	1.209	0.009473
3	rs6773049	ZIC1	6.30E-03	1.143	0.07756
12	rs983421	SUDS3	6.29E-03	1.154	0.05072
13	rs606568	**DGKH**	6.28E-03	0.8816	0.1125
13	rs17088579	OR7E156P	6.27E-03	1.123	0.1374
12	rs4135067	TDG	6.17E-03	1.091	0.2667
10	rs2094179	KLF6	6.05E-03	1.122	0.1266
1	rs640718	KMO	6.00E-03	1.192	0.009732
1	rs17484306	RRAGC	5.97E-03	1.231	0.00339
6	rs3736712	WDR27	5.93E-03	1.137	0.06991
11	rs12275977	GALNTL4	5.92E-03	1.127	0.09964
11	rs6591941	**ODZ4**	5.84E-03	1.04	0.6031
3	rs614566	LAMP3	5.80E-03	1.204	0.005761
14	rs6574988	GPR65	5.80E-03	1.234	0.0003089
1	rs495489	POGK	5.79E-03	0.9191	0.2722
1	rs11161999	LMO4	5.70E-03	1.193	0.007684
18	rs17082921	SOCS6	5.69E-03	1.144	0.07807
9	rs17063814	GNA14	5.62E-03	1.21	0.002639
14	rs12588812	RNASE1	5.55E-03	1.137	0.07456
3	rs16852539	GOLIM4	5.53E-03	1.073	0.2998
4	rs7680321	GABRB1	5.51E-03	1.25	0.0001764
8	rs448578	MSR1	5.50E-03	1.111	0.1176
8	rs17069985	CSMD1	5.49E-03	1.105	0.1615
1	rs1890038	CHD1L	5.48E-03	1.137	0.05786
10	rs10443995	DOCK1	5.48E-03	1.047	0.5138
9	rs13290547	DAB2IP	5.47E-03	1.192	0.01176
3	rs9824570	CLSTN2	5.45E-03	0.92	0.1817
16	rs4843366	LOC732275	5.44E-03	1.162	0.013
10	rs1338007	ADRA2A	5.44E-03	1.075	0.3076
9	rs615928	GCNT1	5.44E-03	1.099	0.2024
14	rs10137389	C14orf106	5.43E-03	1.084	0.2648
7	rs56183050	POT1	5.43E-03	1.095	0.1748
12	rs2468244	CEP290	5.42E-03	1.096	0.1677
9	rs3780621	COL15A1	5.41E-03	1.157	0.01337
1	rs6684324	INADL	5.41E-03	1.204	0.003692
13	rs9514132	SLC10A2	5.38E-03	0.9617	0.6074
1	rs1318222	C1orf94	5.37E-03	1.123	0.08309
18	rs1560398	MC4R	5.36E-03	1.035	0.6496
5	rs17653341	ADRB2	5.32E-03	1.055	0.4495
1	rs12046987	MIR101-1	5.30E-03	1.158	0.02022
6	rs7739908	OGFRL1	5.29E-03	1.165	0.02408
18	rs17739703	C18orf34	5.28E-03	0.9017	0.1245
12	rs1861674	LOH12CR1	5.28E-03	1.204	0.001111
7	rs7785575	ELMO1	5.28E-03	1.117	0.08243

Top genes found by the epistasis network analysis workflow described in the Materials and methods for the merged WTCCC + NIMH-EA data sets. Rows are sorted by SNPrank epistasis network centrality score. Columns are chromosome, SNP rsid, gene symbol, SNPrank score and univariate odds ratio and *P*-value.
Bold gene symbols are genes that have strong evidence from univariate analysis of other larger-scale GWAS of BD. Ranking for unmerged data may be found in Supplementary Table 1.

Epistasis network centrality (SNPrank) results of the top individual SNPs for the WTCCC, NIMH and merged data sets may be found in Supplementary Table 1. There is consistent evidence in the GWAS literature for the role of *ANK3* for BD susceptibility, yet no *ANK3* SNPs are ranked higher than 600 in a single-locus analysis of the WTCCC data unless the data is merged with other studies to create a larger sample size.[26] Without pooling additional samples, the epistasis network centrality analysis of the WTCCC data yields a variant in *ANK3* (rs10509126) that is ranked third by SNPrank. The network centrality rank (SNPrank) of this variant moves higher in the rankings when the WTCCC and NIMH-EA GWAS are merged (rank second). As shown in Figure 2, this *ANK3* SNP has the largest number of gene–gene interaction connections in the WTCCC GWAS data. The merged network analysis yields a top-10 SNPrank (rank seventh) to a SNP in *DGKH*, which was implicated for BD in a previous study[27] but not in the WTCCC and NIMH data sets. The merged analysis yields a rank of 15 for a variant in *ODZ4*, which was identified in Sklar *et al.*[28]

Discussion

Motivated by the complex, interconnected nature of biological pathways involved in biological processes such as mood regulation, we infer epistasis network signatures of BD from two published GWAS. An underlying assumption of pathway and gene-set approaches is that genes influence phenotypic expression as part of a biological network; however, most gene-set and pathway studies use statistical gene

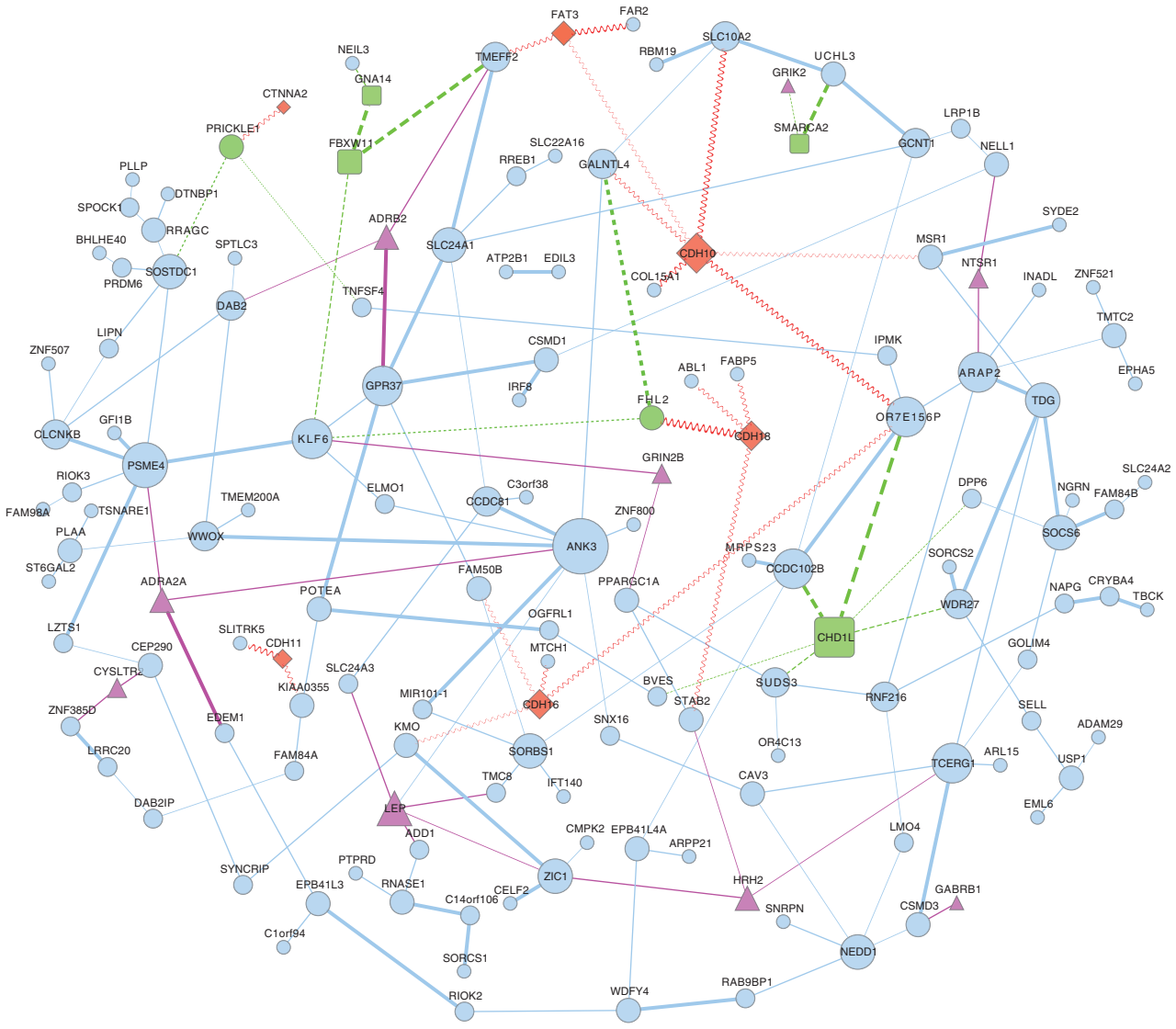

Figure 2 Epistasis network for WTCCC GWAS of bipolar disorder. Network inferred following ECML feature selection and regression-based genetic association interaction network (reGAIN) for the WTCCC GWAS of bipolar disorder, annotated by top enriched pathways. An edge threshold (0.575) was chosen as described in Materials and methods; interactions below this threshold are hidden. The 146 nodes are colored based on membership of the genes in the pathways with evidence of enrichment replication (Tables 1 and 2): red diamond (membership in both Wnt signaling pathway and cadherin signaling pathway), green square (Wnt signaling pathway only) and magenta triangle (Neuroactive ligand-receptor interaction pathway). The weight of an edge is proportional to the gene–gene interaction strength. The 183 edges are colored based on connection of a gene node to a gene in the given pathway using the scheme above (red squiggle, green dashed, magenta solid). The size of a node is proportional to its degree (number of edges). Note, *ANK3* in the middle is the most connected.

prioritization limited to the individual effect of each gene or variant. The goal of the current work was to use pathway replication evidence for the hypothesis that epistasis network signatures contain information about the underlying biological pathways that regulate phenotypic expression of BD. Our approach used ECML filtering and reGAIN to create a data-driven BD-specific network consisting of statistical gene–gene interactions and single-locus associations. We then used SNPrank to integrate these effects and prioritize genes for pathway enrichment analysis.

Direct replication of a network signature poses a statistical challenge due to the complexity of the models that are to be tested.[19,20] We chose a level of replication that uses pathway enrichment statistics as evidence for network effects in independent GWAS. We constructed filtered epistasis networks and use SNPrank network centrality scores to prioritize genes for subsequent pathway enrichment analysis. In the current study, we replicated the enrichment of the cadherin signaling pathway based on the prioritization of genes through an epistasis network analysis of the WTCCC and NIMH GWA studies of BD. Other enriched pathways of interest were identified including WNT signaling, axon guidance and neuroactive ligand-receptor interaction (see Tables 1 and 2).

The enrichment of genes in the cadherin, Wnt and axon guidance signaling pathways is suggestive of a developmental origin for BD. The Wnt/B-catenin pathway is the canonical pathway controlling cell proliferation and differentiation during embryonic development.[34] Cadherins guide neuronal

migration during development and are involved in neuronal differentiation and synaptogenesis. Interestingly, the schizophrenia susceptibility gene, *DISC1*, appears to have a role in the regulation of cell–cell adhesion and neurite outgrowth via the expression of N-cadherins.[35] Wnt pathway genes may also have a role in synpatic plasticty and adult neurogenesis, possibly explaining why lithium[36] and perhaps valproate,[37] increase gray matter volumes in patients with BD—lithium inhibits GSK3B thereby upregulating WNT signaling.[38] Although the cadherin/WNT pathway has not generally been the focus of genetic studies, a number of genes within this pathway, including FAT[30,39] and PPARD,[40] have been implicated in the development of BD.

In addition to pathways, we find evidence for increased sensitivity to detect SNPs relevant to BD susceptibility by aggregating network effects, including the main effect of nodes. A notable example of this boost in sensitivity is *ANK3* (rs10509126). When ranked by univariate statistical significance in the WTCCC GWAS, *ANK3* SNPs are outside the top 600 SNPs. However, the epistasis network procedure ranks this *ANK3* SNP third in the WTCCC data, and the rank is second when the WTCCC data is merged with the NIMH-EA data (see Table 3 and Supplementary Table). The ability to identify this SNP in the WTCCC data is significant because of the growing body of support for *ANK3* for BD susceptibility since the WTCCC study. The top SNPrank SNP in the WTCCC data is *ARAP2* gene, which contains ankyrin repeats. Both *ANK3* and *ARAP2* are highly connected in the reGAIN in Figure 2 and interact with genes in the neuroactive ligand-receptor interaction pathway. The *DGKH* region, implicated in a previous study,[27] lacks a strong signal in the WTCCC data by itself, but when merged with the NIMH data, the epistasis network approach ranks one of the *DGKH* SNPs seventh.

Baum *et al.*[27] reported the first association between a SNP in DGKH and BD in the context of a GWAS. The association with DGKH was recently replicated in a Han-Chinese population.[41] Moreover, a DGKH haplotype consisting of the SNPs, rs994856, rs9525580 and rs9525584, was recently associated with BD, unipolar depression and attention deficit hyperactivity disorder (ADHD),[42] which comprise psychiatric disorders that share substantial overlap with respect to clinical symptomatology. Interestingly, DGKH is a key protein in the phosphatidyl-inositol pathway that is also regulated by lithium.[43] A recent large-scale analysis (11 974 BD cases and 51 792 controls) identified a new variant in *ODZ4*.[28] The epistasis network analysis of the present study also yielded variants in the *ODZ4* gene for the smaller WTCCC and NIMH GWAS data sets, and the merged analysis yielded a rank of 15 for a variant in *ODZ4*. With the growing number of large-scale GWAS studies, it may be possible to identify novel variants of biological importance through an epistasis network approach.

The general linear model used in reGAIN provides a statistical framework to assign confidence to edges and nodes in the network. In addition, the SNPrank eigenvector centrality scores computed from the reGAIN are well suited to prioritizing genes for pathway enrichment calculations. The SNPrank scores are more difficult to interpret than an odds ratio or a *P*-value; however, the scores have an interpretation as probabilities because the scores come from the elements of a normalized eigenvector so that the scores sum to unity. Thus, we can identify a significance threshold for pathway enrichment by comparing the observed SNPrank score distribution with a uniform probability as a theoretical null.

These results suggest that some of the missing heritability may be due to the neglect of the context of disease-specific networks of epistatic and main effects. A future challenge is to quantify the amount of heritability that may be accounted for in these networks. A strategy toward this end may be to use the variation in the edge and node regression coefficients of the network to estimate the heritability. These data-driven network techniques offer an additional tool to identify new biological pathways, network signatures and markers relevant to phenotypes due to network interactions.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements. This work was supported by NIH Grants no. K25 AI-64625 (PI: BA McKinney), R56 AI-80932 (PI: BA McKinney). Dr McKinney is also supported by the William K Warren foundation.

1. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460**: 748–752.
2. Burmeister M, McInnis MG, Zollner S. Psychiatric genetics: progress amid controversy. *Nat Rev Genet* 2008; **9**: 527–540.
3. Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J *et al.* Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatry* 2009; **166**: 540–556.
4. Gershon ES, Alliey-Rodriguez N, Liu C. After GWAS: searching for genetic risk for schizophrenia and bipolar disorder. *Am J Psychiatry* 2011; **168**: 253–256.
5. Grozeva D, Kirov G, Ivanov D, Jones IR, Jones L, Green EK *et al.* Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Arch Gen Psychiatry* 2010; **67**: 318–327.
6. Moskvina V, Craddock N, Muller-Myhsok B, Kam-Thong T, Green E, Holmans P *et al.* An examination of single nucleotide polymorphism selection prioritization strategies for tests of gene-gene interaction. *Biol Psychiatry* 2011; **70**: 198–203.
7. Patel SD, Le-Niculescu H, Koller DL, Green SD, Lahiri DK, McMahon FJ *et al.* Coming to grips with complex disorders: genetic risk prediction in bipolar disorder using panels of genes identified through convergent functional genomics. *Am J Med Genet B Neuropsychiatr Genet* 2010; **153B**: 850–877.
8. Raychaudhuri S, Korn JM, McCarroll SA, Altshuler D, Sklar P, Purcell S *et al.* Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet* 2010; **6**: e1001097.
9. Zhang D, Cheng L, Qian Y, Alliey-Rodriguez N, Kelsoe JR, Greenwood T *et al.* Singleton deletions throughout the genome increase risk of bipolar disorder. *Mol Psychiatry* 2009; **14**: 376–380.
10. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010; **11**: 446–450.
11. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ *et al.* Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
12. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* 2011; **88**: 294–305.
13. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010; **42**: 565–569.
14. Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* 2008; **4**: e1000008.
15. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* 2012; **109**: 1193–1198.
16. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* 2012; **8**: e1002375.
17. Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P *et al.* Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet* 2009; **85**: 13–24.

18. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM *et al.* A versatile gene-based test for genome-wide association studies. *Am J Hum Genet* 2010; **87**: 139–145.
19. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* 2008; **92**: 265–272.
20. McKinney BA, Pajewski NM. Six degrees of epistasis: statistical network models for GWAS. *Front Genet* 2011; **2**: 109.
21. McKinney BA, Crowe JE, Guo J, Tian D. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS Genet* 2009; **5**: e1000432.
22. Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, Moore JH. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics* 2011; **12**: 364.
23. Davis NA, Crowe JE Jr., Pajewski NM, McKinney BA. Surfing a genetic association interaction network to identify modulators of antibody response to smallpox vaccine. *Genes Immun* 2010; **11**: 630–636.
24. WTCCC Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661–678.
25. Smith EN, Bloss CS, Badner JA, Barrett T, Belmonte PL, Berrettini W *et al.* Genome-wide association study of bipolar disorder in European American and African American individuals. *Mol Psychiatry* 2009; **14**: 755–763.
26. Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L *et al.* Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet* 2008; **40**: 1056–1058.
27. Baum AE, Akula N, Cabanero M, Cardona I, Corona W, Klemens B *et al.* A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry* 2008; **13**: 197–207.
28. Sklar P, Ripke S, Scott LJ, Andreassen OA, Cichon S, Craddock N *et al.* Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* 2011; **43**: 977–983.
29. Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. *Genome Bio* 2010; **11**: R53.
30. Blair IP, Chetcuti AF, Badenhop RF, Scimone A, Moses MJ, Adams LJ *et al.* Positional cloning, association analysis and expression studies provide convergent evidence that the cadherin gene FAT contains a bipolar disorder susceptibility allele. *Mol Psychiatry* 2006; **11**: 372–383.
31. Gould TD, Manji HK. The Wnt signaling pathway in bipolar disorder. *Neuroscientist* 2002; **8**: 497–511.
32. Bush WS, Chen G, Torstenson ES, Ritchie MD. LD-spline: mapping SNPs on genotyping platforms to genomic regions using patterns of linkage disequilibrium. *BioData Min* 2009; **2**: 7.
33. Taher L, Ovcharenko I. Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. *Bioinformatics* 2009; **25**: 578–584.
34. Yang Y. Wnt signaling in development and disease. *Cell Biosci* 2012; **2**: 14.
35. Hattori T, Shimizu S, Koyama Y, Yamada K, Kuwahara R, Kumamoto N *et al.* DISC1 regulates cell-cell adhesion, cell-matrix adhesion and neurite outgrowth. *Mol Psychiatry* 2010; **15**: 778 98-809.
36. Moore GJ, Bebchuk JM, Wilds IB, Chen G, Manji HK. Lithium-induced increase in human brain grey matter. *Lancet* 2000; **356**: 1241–1242.
37. Savitz J, Nugent AC, Bogers W, Liu A, Sills R, Luckenbaugh DA *et al.* Amygdala volume in depressed patients with bipolar disorder assessed using high resolution 3T MRI: the impact of medication. *Neuroimage* 2010; **49**: 2966–2976.
38. Klein PS, Melton DA. A molecular mechanism for the effect of lithium on development. *Proc Natl Acad Sci USA* 1996; **93**: 8455–8459.
39. Abou Jamra R, Becker T, Georgi A, Feulner T, Schumacher J, Stromaier J *et al.* Genetic variation of the FAT gene at 4q35 is associated with bipolar affective disorder. *Mol Psychiatry* 2008; **13**: 277–284.
40. Zandi PP, Belmonte PL, Willour VL, Goes FS, Badner JA, Simpson SG *et al.* Association study of Wnt signaling pathway genes in bipolar disorder. *Arch Gen Psychiatry* 2008; **65**: 785–793.
41. Zeng Z, Wang T, Li T, Li Y, Chen P, Zhao Q *et al.* Common SNPs and haplotypes in DGKH are associated with bipolar disorder and schizophrenia in the Chinese Han population. *Mol Psychiatry* 2011; **16**: 473–475.
42. Weber H, Kittel-Schneider S, Gessner A, Domschke K, Neuner M, Jacob CP *et al.* Cross-disorder analysis of bipolar risk genes: further evidence of DGKH as a risk gene for bipolar disorder, but also unipolar depression and adult ADHD. *Neuropsychopharmacology* 2011; **36**: 2076–2085.
43. Berridge MJ. The Albert Lasker Medical Awards. Inositol trisphosphate, calcium, lithium, and cell signaling. *JAMA* 1989; **262**: 1834–1841.

Supplementary Information accompanies the paper on the Translational Psychiatry website (http://www.nature.com/tp)