

SCIENTIFIC REPORTS



OPEN

Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm

Li-Chung Chuang^{1,2} & Po-Hsiu Kuo^{2,3}

Received: 03 October 2016

Accepted: 29 November 2016

Published: 03 January 2017

A genetic risk score could be beneficial in assisting clinical diagnosis for complex diseases with high heritability. With large-scale genome-wide association (GWA) data, the current study constructed a genetic risk model with a machine learning approach for bipolar disorder (BPD). The GWA dataset of BPD from the Genetic Association Information Network was used as the training data for model construction, and the Systematic Treatment Enhancement Program (STEP) GWA data were used as the validation dataset. A random forest algorithm was applied for pre-filtered markers, and variable importance indices were assessed. 289 candidate markers were selected by random forest procedures with good discriminability; the area under the receiver operating characteristic curve was 0.944 (0.935–0.953) in the training set and 0.702 (0.681–0.723) in the STEP dataset. Using a score with the cutoff of 184, the sensitivity and specificity for BPD was 0.777 and 0.854, respectively. Pathway analyses revealed important biological pathways for identified genes. In conclusion, the present study identified informative genetic markers to differentiate BPD from healthy controls with acceptable discriminability in the validation dataset. In the future, diagnosis classification can be further improved by assessing more comprehensive clinical risk factors and jointly analysing them with genetic data in large samples.

An accurate risk score model has substantial benefits in assisting the early screening of diseases, patient management, and clinical diagnosis. In addition, a risk score model has been applied to the prognosis prediction for complex traits, such as cardiovascular diseases^{1,2} and cancers³. With the advances in generating and the accumulation of genetic information, especially the increasing popularity in genome-wide association (GWA) studies to provide more comprehensive information about genetic variation for the trait of interest, some studies have incorporated such data to construct risk score models for disease diagnosis, such as schizophrenia and post-traumatic stress disorder^{4–7}.

Although many GWA studies have been conducted for heritable traits, only a few susceptible loci are reproducibly reported, with a small effect size between 1.2 and 2.0^{8,9}. In a recent massive GWA study of schizophrenia, 108 loci were identified. However, each of these loci contributed a tiny fraction of the risk in the population; even the polygenic risk score explained less than 3.4% of the variance¹⁰. The general consensus is that multiple genes may act synergistically and interactively to influence the risk for complex diseases¹¹. Jointly considering the main and interaction effects among multi-loci in whole-genome analyses is advantageous in constructing the genetic risk score model. Traditional statistical approaches in GWA studies usually focus on detecting the main effects without considering the non-linear patterns among markers and their interactions. Machine learning methods have fewer assumptions about analysis models, and have been frequently used in the analysis of high-dimensional data¹². The random forest (RF) method is one commonly used algorithm in machine learning techniques. Because the hierarchical decision tree structure in RF can model non-linear associations, this method is often used to capture interactions among loci, and is likely to perform well in uncovering the genetic causes underlying the etiology of complex disorders.

¹Department of Nursing, Cardinal Tien Junior College of Healthcare & Management, I-Lan, Taiwan. ²Department of Public Health & Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan. ³Research Center for Genes, Environment and Human Health, National Taiwan University, Taipei, Taiwan. Correspondence and requests for materials should be addressed to P.-H.K. (email: phkuo@ntu.edu.tw)

Top	NO. of SNPs	No. of SNPs after excluding markers in complete LD	AUROC	Hosmer-Lemeshow test
10	19	16	0.615	0.535
20	36	29	0.663	0.729
50	85	68	0.763	0.054
100	168	135	0.846	0.732
150	258	211	0.908	0.506
200	348	289	0.944	0.945

Table 1. The performance of discrimination ability for the union marker numbers of the top ranked in the two indices, the Gini Index and the conditional variable importance from the random forest procedure.

Note: AUROC: the area under the receiver characteristic curve; 95% C.I.: 95% confidence interval. *Markers in complete linkage disequilibrium (LD, $D' = 1$) were removed before regression analysis and only one SNP was kept for each case of complete LD situation.

The RF is an ensemble-based machine learning method that uses multiple classification and regression trees as classifiers. After taking the majority vote over all classifiers, the RF method combines information across all trees to reveal variable importance. It can assess associations and interactions simultaneously¹³. The RF has been applied to many biological studies, such as gene expression¹⁴, metabolomics¹⁵, proteomics¹⁶, and GWA¹⁷. These studies showed that the RF method provides good accuracy, less internal examination of error, and high variable importance from mass biological data. Thus, important variables (e.g. biomarkers or SNPs in genetic data) that are selected by the RF procedure form a useful basis to construct risk score models for the risk of developing complex disorders.

In the current study, we focused on a complex psychiatric trait, bipolar disorder (BPD) with a high heritability of around 0.6 to 0.8^{18,19}, as the target for the risk score model construction. BPD is a severe psychiatric disorder, which affects approximately one in a hundred people worldwide. Many studies have suggested that the prevalence of BPD has increased worldwide in the last decade^{20–22}. Without proper treatment, 15% of patients committed suicide²³. Patients with BPD experience intermittent manic and depressive episodes, and usually exhibit chronic courses. Several large-scale GWA studies have been conducted to provide a list of susceptible genetic loci for BPD, without considering the join or interaction effects among loci^{8,24,25}. So far, reliable genetic markers or objective biological indices are unavailable for clinical use in assisting diagnosis or prognosis. The diagnosis of BPD largely depends on a subjective report of patient's syndromes and clinical characteristics.

In the present study, we used the RF-based method to construct a genetic risk prediction model for BPD, using pre-screened potentially associated markers in large-scale GWA datasets. The procedures are as follows. First, the RF method was used to select important variables as candidate risk markers in a BPD GWA dataset from the Genetic Association Information Network (GAIN). Second, the multivariable logistic regression with variable-selection methods was used to further select a smaller optimal subset of the variables from all candidate risk markers. We could then build risk score models using the identified optimal markers for the GAIN-BPD dataset. Third, to estimate the performance of the risk score models, leave-one-out cross-validation was performed for internal validation. We also used the other BPD-GWA dataset, Systematic Treatment Enhancement Program (STEP), as an external validation data. Finally, we performed a gene sets analysis to uncover the underlying biological mechanisms for those loci that were identified as the candidate risk markers.

Result

The accuracy of the RF procedure was evaluated in the GAIN training set created during the forest growing process. The accuracy of the RF classification in the GAIN data was 0.939 in controls and 0.852 in BPD patients. We ranked markers based on values of the two indices, the Gini Index (GI) and the conditional variable importance (VI) from the RF procedure. Because the two indices did not completely agree with each other, we used the union markers of the top ranked 200 in each index. As Table 1 shows, we included 348 single nucleotide polymorphisms (SNPs) to calculate the performance of discrimination ability. After excluding SNPs with complete linkage disequilibrium (LD), 289 SNPs were retained as the candidate risk markers in the regression model. As shown in Fig. 1A, the 289 candidate markers had a good discrimination ability with an area under the receiver operating characteristic (AUROC) of 0.944 (95% confidence interval (CI), 0.935–0.953), and calibration ability measured by the Hosmer-Lemeshow test (p -value = 0.933). The multivariable logistic regression with stepwise selection retained 121 SNPs as the final optimal marker set. A good discrimination ability was observed with an AUROC of 0.924 (95% CI, 0.913–0.935) based on the 121 markers (Fig. 1B).

The genetic risk score based on this final prediction model was calculated for each individual from accumulating numbers of the risk alleles and weighted by the beta regression coefficient. The risk scores among all participants were ranged from 143.8 to 228.4 in the GAIN dataset, with a mean of 175.4 in the controls and 191.3 in the BPD patients. To find an optimal cutoff point, we used the Youden Index to obtain the risk score cutoff as 184, and the corresponding sensitivity and specificity for BPD were 0.777 and 0.854, respectively (Table 2). The likelihood ratios of the risk score with the optimal cutoff point were 5.322 for a positive result and 0.261 for a negative result, indicating moderate evidence for the differentiation between BPD patients and healthy controls.

We then used leave-one-out cross-validation to conduct internal validation for classification accuracy (Table 3). The predicted error rates were around 0.2 using either the 289 candidate risk markers, the 121 optimal markers and the risk score. The STEP dataset was used as an external validation dataset to evaluate the

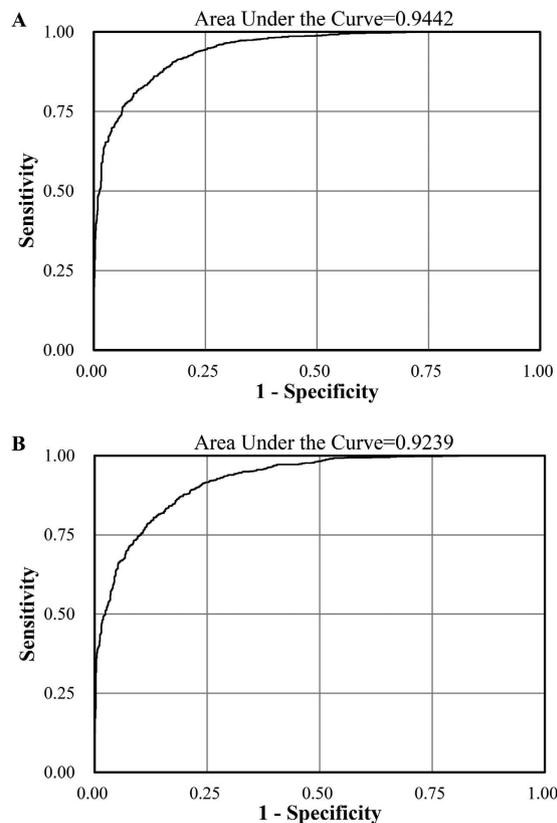


Figure 1. The receiver characteristic curves for (A) the 289 candidate markers and (B) the 121 optimal markers in the GAIN bipolar disorder dataset.

Score	Sensitivity	Specificity	Youden Index	PPV	NPV	LR+	LR–
180	0.905	0.685	0.590	0.735	0.882	2.871	0.139
181	0.880	0.723	0.604	0.755	0.862	3.182	0.166
182	0.854	0.776	0.630	0.787	0.846	3.807	0.188
183	0.813	0.817	0.630	0.812	0.819	4.449	0.229
184	0.777	0.854	0.631	0.837	0.798	5.322	0.261
185	0.747	0.879	0.626	0.857	0.782	6.181	0.288

Table 2. Sensitivity, specificity, Youden Index, and the likelihood ratios of the risk score model in the GAIN bipolar disorder dataset as a training set. Note: PPV: positive predictive value; NPV: negative predictive value; LR+: the likelihood ratio for a positive test result; LR–: the likelihood ratio for a negative test result.

Model construction: the GAIN dataset				Validation: the STEP dataset				
Models	AUROC	(95% C.I.)	Hosmer-Lemeshow test	Error rate*		AUROC	(95% C.I.)	Hosmer-Lemeshow test
				Controls	BPD			
289 candidate markers	0.944	(0.935–0.953)	0.933	0.208	0.220	0.702	(0.681–0.723)	0.681
121 optimal markers	0.924	(0.913–0.935)	0.458	0.193	0.209	0.639	(0.617–0.662)	0.002
The risk score variable	0.905	(0.893–0.918)	0.264	0.179	0.193	0.506	(0.482–0.530)	0.954

Table 3. The performance of discrimination ability for the genetic risk score model in the training set and the validation dataset. Note: AUROC: the area under the receiver characteristic curve; 95% C.I.: 95% confidence interval. *Error rate was examined by the leave-one-out cross-validation procedure.

performance of discrimination ability. In general, the discrimination performance was acceptable using 289 candidate risk predictors in the STEP data, which had an AUROC of 0.702 (95% CI, 0.681–0.723) and a good calibration ability (Hosmer-Lemeshow test, p -value = 0.681). With only the 121 optimal predictors, the AUROC dropped to 0.639 (95% CI, 0.617–0.662), with a low calibration ability (Hosmer-Lemeshow test, p -value = 0.0002). The risk score method showed the poorest discrimination ability (AUROC = 0.506; 95% CI, 0.482–0.529).

We performed the same set of analyses using STEP data as the training dataset, and the GAIN data as the validation. The results are displayed in Supplementary Table 1. In total, 354 candidate risk markers were identified for the STEP dataset. After excluding SNPs with complete LD, 312 SNPs were retained in the regression model and had the discrimination ability with AUROC of 0.934 (95% CI, 0.925–0.944). In the external validation GAIN dataset, decreased but acceptable discrimination performance was again observed, with an AUROC of 0.732 (95% CI, 0.711–0.754) (Supplementary Table 1). It is worth noting that there were no overlapping candidate risk markers between the two datasets. If we mapped all candidate risk markers from the two datasets to genes, there were in total 233 gene regions, including 98 genes in the GAIN dataset and 144 genes in the STEP dataset. Only 9 genes (3.8%) overlapped between the two datasets, including genes *ALK*, *TACR1*, *LRP1B*, *GALNT17*, *NAV2*, *ODZ4*, *RAD51L1*, *KTNI1*, and *CACNG2*.

Significantly enriched gene-sets were identified for these mapped genes in the two GWA datasets of BPD. Table 4 showed that 43 pathways were identified in the GAIN dataset with a *q*-value of less than 0.01 after correction for multiple comparisons. In the STEP dataset, 28 significant pathways were identified. Important biological pathways were reported, including cation ion channel activity (such as voltage-gated calcium channel activity and complex, regulation of action potential and cation transport), membrane structure (such as plasma membrane, transmembrane receptor activity and establishment of location), neuron function (such as brain development, axon guidance and GABA receptor activity) and cytoskeleton (such as cytoskeletal protein binding and actin filament).

Discussion

It is a common interest to explore the usage of genetic findings for heritable complex traits. There is an absence in the literature of a risk score model based on genetic information for the diagnosis of BPD. Non-replication across datasets is often observed, especially when focusing on specifically significant markers. Hence, using extremely significant markers in one sample to construct a genetic risk model to apply to other samples might not produce good prediction accuracy. On the other hand, informative genetic markers, which are selected by methods of machine learning, have been used for the classification of outcomes or for predicting the risk of developing diseases, such as early detection of prostate cancer²⁶, treatment response in attention deficit hyperactivity disorder²⁷, and identification of idiopathic autism spectrum disorder (ASD) patients²⁸. Among the many machine learning methods, such as support vector machine, linear discriminant analysis, and *k*-nearest neighbour classification, RF is often applied in biomedical research with different data sources, such as gene expression²⁹. Similar applications are reported using GWA datasets for complex traits with low prediction errors, such as severe asthma^{30,31}. To our best knowledge, the present study reports the first prediction results for BPD using an RF approach to select informative markers which jointly consider the main and interaction effects among genetic variants. Our results revealed that these informative markers possess fair to good discriminability for BPD patients in the training and validation datasets.

Diagnoses of psychiatric disorders often largely depend on clinical interview rather than biomarkers. With the RF procedures, the 289 candidate risk predictors in the GAIN dataset perform well with an AUROC of 0.944 and a 0.702 AUROC in the validation dataset. Moreover, high sensitivity and specificity are also observed using the more parsimonious 121 optimal markers and the risk score, with acceptable predicted error rates less than 0.2 in leave-one-out cross-validation. Without the RF procedures, if we selected the same 289 candidate predictors based on *p*-values significance in the GAIN data to estimate the discriminability in the STEP dataset, the AUROC slightly dropped to 0.686 (data not shown). In the literature, some risk score models were built using genetic information in aid of improving disease classification for complex traits, without satisfactory prediction power. For instance, Golan and colleagues (2014) used the random effect approach and reported the discrimination ability with an AUROC of 0.62 for BPD patients from the Wellcome Trust Case Control Consortium dataset³². Using SNPs within 13 significant pathways in a study of ASD, one recent study included 237 SNPs to generate a genetic diagnostic classifier and reported an 85.6% prediction accuracy in the Central European cohort, but the accuracy dropped to 50.6% in the Han Chinese cohort³³. Moreover, based on differentially expressed 762 unique genes, a previous study reported an 82.5% prediction accuracy for ASD³⁴. In our study, the RF approach demonstrated a fine performance in selecting the informative genetic markers from massive GWA data. The classification accuracy for BPD in the current study is at the higher end with low error rates. In particular, we still obtained fair results with an AUROC of 0.702 in the STEP validation dataset.

It is commonly observed that the accuracy of the genetic prediction model is reduced in external validation samples^{33,34}. Schulze and colleagues (2014) constructed a polygenic model for BPD, however, the performance of this model is poor in two external validation datasets, with AUROC ranged between 0.55 to 0.57³⁵. The heterogeneity inherited in different studies and samples is often noted, which might reflect differences in sample ascertainment, population stratification, or experimental variations. An example of this is demonstrated in a large-scale study of Psychiatric Genomics Consortium (PGC) for population stratification. Using the multivariate linear mixed model approach, Maier and colleagues (2015) created genomic risk scores for severe psychiatric disorders, including schizophrenia, BPD, and major depressive disorder, using GWA data in PGC as the training set. In the validation data, the correlation coefficients between the observed status of the psychiatric disorders and their predicted genomic risk scores were low, ranged from 0.076–0.224⁷. To evaluate the population stratification, they calculated ancestry principle components of PGC data and then divided GWA data into four groups of the first ancestry principal component that reflect the population difference between individuals. Their results indicated significant heterogeneity for BPD in PGC GWA datasets (*p*-value = 0.0017), which is likely attributed to the ancestral population differences⁷. Therefore, heterogeneity derived from many sources might result in lowered prediction accuracy in an external dataset and hinders clinical usage and further application to assist diagnosis.

We ran both GWA datasets as the training and the other as the validation dataset for model construction. In either scenario, a very similar classification performance is observed, suggesting the stability of current

Gene set name	No. of gene in gene set	No. of gene ^c	Dataset ^d	q-value
Acetyl-glucosaminyl transferase activity	16	2	STEP	3.4×10^{-03}
Actin binding	76	3	GAIN	2.8×10^{-03}
Actin cytoskeleton organization and biogenesis	104	3	GAIN	6.7×10^{-03}
Actin filament	18	2	STEP	4.3×10^{-03}
Actin filament based process	114	3	GAIN	8.6×10^{-03}
Amine metabolic process	137	4	STEP	6.8×10^{-03}
Anatomical structure morphogenesis	374	6	GAIN	2.6×10^{-03}
Auxiliary transport protein activity	25	2	STEP	8.1×10^{-03}
Axon guidance	22	2	GAIN	2.9×10^{-03}
Axonogenesis	43	3	GAIN	5.3×10^{-04}
Brain development	51	3	GAIN	8.8×10^{-04}
Calcium channel activity	33	2	GAIN	6.5×10^{-03}
Calcium ion transport	27	2	GAIN	4.4×10^{-03}
Calmodulin binding	25	2	STEP	8.1×10^{-03}
Carbohydrate binding	72	3	STEP	7.2×10^{-03}
Cation transport	146	4	STEP	8.4×10^{-03}
Cell migration	93	3	GAIN	4.9×10^{-03}
Cell surface	76	3	STEP	8.3×10^{-03}
Cellular morphogenesis during differentiation	49	3	GAIN	7.8×10^{-04}
Channel regulator activity	23	2	STEP	6.9×10^{-03}
Chr12q23	78	3	STEP	8.9×10^{-03}
Chr2p23	75	4	STEP	7.6×10^{-04}
Chr2q23	25	2	STEP	8.1×10^{-03}
Chr3p14	56	3	STEP	3.6×10^{-03}
Chr4q34	19	2	STEP	4.7×10^{-03}
Chr6q13	23	2	STEP	6.9×10^{-03}
Chr6q26	16	2	STEP	3.4×10^{-03}
Cytoplasmic membrane bound vesicle	112	3	GAIN	8.2×10^{-03}
Cytoplasmic vesicle	116	3	GAIN	9.0×10^{-03}
Cytoskeletal protein binding	158	5	GAIN	3.0×10^{-04}
Cytoskeleton	361	6	GAIN	2.2×10^{-03}
Di-, tri-valent inorganic cation transport	32	2	GAIN	6.2×10^{-03}
Endocytic vesicle	14	2	GAIN	1.2×10^{-03}
Enzyme regulator activity	314	6	STEP	7.6×10^{-03}
Establishment of localization	852	12	STEP	2.5×10^{-03}
G protein signaling coupled to IP3 second messenger phospholipase C activating	41	2	GAIN	1.0×10^{-02}
GABA receptor activity	11	2	STEP	1.6×10^{-03}
Generation of neurons	83	3	GAIN	3.6×10^{-03}
Integrin binding	30	2	GAIN	5.4×10^{-03}
Ion transport	184	5	STEP	3.4×10^{-03}
KEGG-Arrhythmic right ventricular cardiomyopathy	76	5	GAIN	9.0×10^{-06}
KEGG-Calcium signaling pathway	178	4	GAIN	4.2×10^{-03}
KEGG-Cardiac muscle contraction	80	4	GAIN	2.2×10^{-04}
KEGG-Dilated cardiomyopathy	92	4	GAIN	3.7×10^{-04}
KEGG-Hypertrophic cardiomyopathy HCM	85	4	GAIN	2.7×10^{-04}
Membrane	1942	20	GAIN	2.1×10^{-05}
Membrane bound vesicle	114	3	GAIN	8.6×10^{-03}
Membrane organization and biogenesis	133	4	STEP	6.1×10^{-03}
Membrane part	1633	13	GAIN	6.6×10^{-03}
Neurite development	53	3	GAIN	9.8×10^{-04}
Neurogenesis	93	3	GAIN	4.9×10^{-03}
Neuron development	61	3	GAIN	1.5×10^{-03}
Neuron differentiation	76	3	GAIN	2.8×10^{-03}
Neuropeptide binding	23	2	GAIN	3.2×10^{-03}
Continued				

Gene set name	No. of gene in gene set	No. of gene ^c	Dataset ^d	q-value
Neuropeptide receptor activity	22	2	GAIN	2.9×10^{-03}
Nitrogen compound metabolic process	150	4	STEP	9.3×10^{-03}
Plasma membrane	1393	14	GAIN	5.4×10^{-04}
Plasma membrane part	1141	10	GAIN	9.0×10^{-03}
RAS guanyl nucleotide exchange factor activity	18	2	STEP	4.3×10^{-03}
Reactome-Depolarization of the presynaptic terminal triggers the opening of calcium channels	12	2	GAIN	8.6×10^{-04}
Reactome-Neurotransmitter release cycle	28	2	GAIN	4.7×10^{-03}
Reactome-Transmission across chemical synapses	130	4	GAIN	1.4×10^{-03}
Receptor mediated endocytosis	33	2	GAIN	6.5×10^{-03}
Regulation of action potential	17	2	STEP	3.8×10^{-03}
Response to external stimulus	306	6	STEP	6.7×10^{-03}
ST interleukin 4 pathway	26	2	STEP	8.8×10^{-03}
System process	558	8	GAIN	1.0×10^{-03}
Transmembrane receptor activity	411	8	STEP	1.9×10^{-03}
Transport	778	11	STEP	3.6×10^{-03}
Voltage-gated calcium channel activity	18	2	GAIN	2.0×10^{-03}
Voltage-gated calcium channel complex	15	2	GAIN	1.4×10^{-03}

Table 4. The significant gene sets for bipolar disorder based on the candidate markers from the GAIN and the STEP datasets. q-value: the value of false discovery rate. ^aNumber of gene in overlap. ^bThe dataset of significant gene set.

procedures. However, we also noticed that there were no overlapping markers selected by RF in the two datasets as candidate risk markers. The agreement increased in gene levels across both datasets, where the same 9 genes are mapped in the two datasets. It may be intuitive, as genetic markers identified are often not causal variants, but rather the proxy for real causal variants. Therefore, the agreement may be the least in marker level, and the increase in gene and pathway levels, especially when heterogeneity exists among datasets. The 9 genes contained both sets of the candidate risk markers, and many studies have indicated that some of these genes are associated with BPD or brain function, such as *ODZ4*³⁶, *TACR1*³⁷, *KTN1*³⁸ and *CACNG2*³⁹. A previous GWA study using 11,974 BPD cases and 51,792 controls, identified a new intronic variant in *ODZ4*³⁶. A recent study indicated that the genetic variants showed specific volumetric effects on the putamen and altered the expression of the *KTN1* gene in both brain and blood tissue³⁸. Similarly, we found a number of significant pathways for the identified genes. These pathway results are quite consistent with pathways findings in previous GWA studies for bipolar disorder^{40,41}.

There were some limitations in the present study. Although risk score models have been used to capture genetic effects from large-scale GWA studies, the power of discrimination was, however, inadequate in previous studies. Dudbridge and colleagues (2013) indicated that the power of polygenic score might be sufficient to use about 2,000 cases and controls, respectively⁴². Even with a good discrimination ability, a smaller sample size of the present study might cause the low power of classification model for BPD patients. In addition, we only used the genetic information to create the risk score model for BPD. The complex psychiatric disorders were caused by the interactions of genetic and environmental risk factors, such as substance dependence and childhood maltreatment. Wong and colleagues (2012) used a risk-classification tree analysis to create a reliable framework based on interactions of genetic variants and environmental factors⁴³. Lacking the information from environmental factors might hinder the application of a prediction model for clinical diagnosis.

In conclusion, we successfully used a machine learning approach to extract informative genetic markers for the construction of a risk score model. Our results indicated a fair discrimination ability for BPD patients with AUROCs of around 0.70 in the external validation datasets. Integration of more comprehensive risk factors from family and environmental data in larger samples is necessary to construct a more precise and applicable risk score model for BPD, to assist with clinical diagnosis in the future.

Materials and Methods

The study design and analysis flow chart are displayed in Fig. 2. Details of the datasets and analytic procedures for the selection of candidate markers in model construction are described below.

Imputation and quality control in the GWA datasets. We used two individual GWA datasets of BPD in the Caucasian populations, the GAIN (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login>)²⁴ and the STEP data (https://www.nimhgenetics.org/available_data/bipolar_disorder/)²⁵. The details of participant enrolment and genotyping of the two GWA studies were provided in their original articles^{24,25}. In brief, in the GAIN dataset, individuals were Americans with European ancestry, including 1,001 BPD cases and 1,034 controls. In the STEP dataset, there were 955 BPD cases and 1,498 healthy subjects from the National Institute of

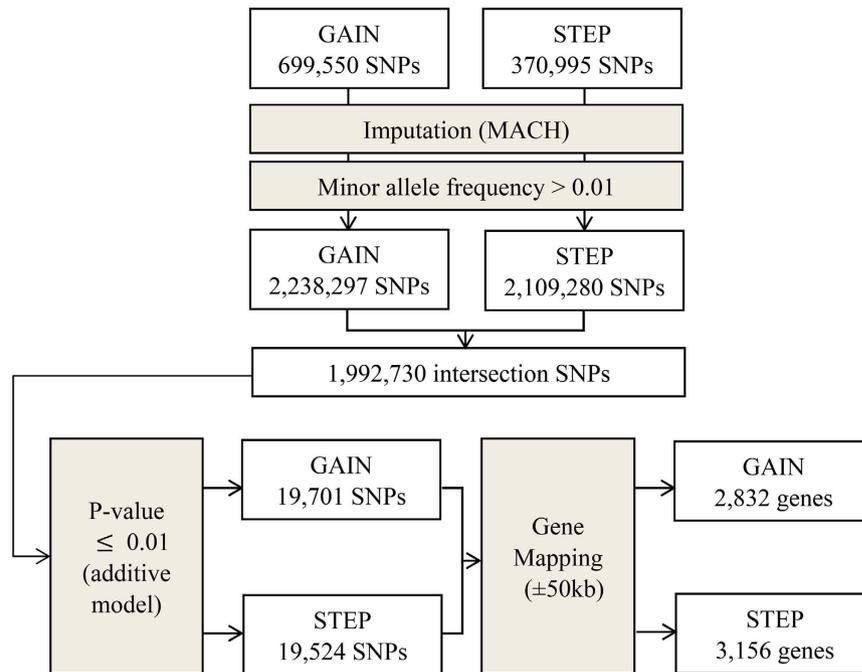


Figure 2. The summary description of the selection of candidate markers for the construction of the genetic risk score model. **GAIN:** the Genetic Association Information Network; **STEP:** the Systematic Treatment Enhancement Program; **MACH:** The program Markov Chain Haplotyping.

Mental Health Genetics Initiative. After quality control for genotypic data, a total of 699,550 and 370,995 autosomal SNPs were retained in the GAIN and STEP datasets, respectively.

Because the two GWA studies used different genotyping platforms, we imputed all the autosomal SNPs for the two GWA datasets based on those genotyped SNPs that passed quality control, to obtain the maximal number of common SNPs for the construction of prediction models for the two datasets. We applied Markov Chain Haplotyping (MACH) 1.0⁴⁴ to perform genotype imputation, and the HapMap II CEU (release 22) samples were used as the reference panel (www.hapmap.org). Well-imputed SNPs had squared correlations ≥ 0.30 between imputed and true genotypes, which were suggested by MACH. After removing markers with minor allele frequency (MAF) less than 1%, the number of well-imputed SNPs was 2,238,297 and 2,109,280 for the GAIN and the STEP datasets, respectively. In total, there were 1,992,730 well-imputed SNPs overlapped in the two datasets.

Criteria for candidate markers in the construction of prediction models. To construct the genetic prediction model for BPD, we first performed association analyses with the additive model for 1,992,730 well-imputed SNPs using PLINK versions 1.07⁴⁵. As the study design flow chart shows in Fig. 1, SNPs with p -value ≤ 0.01 were selected as candidate SNPs in RF analysis, for which 19,701 SNPs were in the GAIN dataset and 19,524 SNPs were in the STEP dataset. Each candidate SNP was then mapped to a gene region (using NCBI build 36) if the SNP was located within the 50 kb of upstream or downstream of a gene. These SNPs were mapped to 2,832 genes in the GAIN dataset, and 3,156 genes in the STEP dataset. In total, there were 802 genes overlapped in both GWA datasets.

Random forest procedures. To select informative risk predictors for model construction, we used the RF method for classification and building regression trees. All candidate SNPs were used to build a training model (*i.e.* 19,701 candidate SNPs in the GAIN dataset). The results from the growing of ensemble trees as a forest could provide a list of important variables for disease outcome. The RF procedures were performed using the Random Jungle package⁴⁶, which facilitates the rapid analysis of large-scale GWA data⁴⁷. Detailed procedures are described in the following steps:

- (1) Two-thirds of the subjects in the GWA dataset were taken as the training set using the bootstrap procedure, and the remaining subjects were treated as the test set.
- (2) Second, is the splitting step. A random subset of markers was chosen from all candidate markers without replacement. The size of each marker subset was equal to the square root of numbers of all candidate markers. The decrease in impurity for all markers was then calculated. The definition of the decrease in impurity was detailed elsewhere⁴⁶. The marker with the best classification by the decrease in impurity was used as a node to split subjects of a training set into two distinct subsets, that is, one node split into two nodes.
- (3) The 2nd splitting step was repeated until the tree is grown with its largest extent in the tree growing step.
- (4) Steps 1 to 3 were repeated to grow 5,000 classification trees to build a forest.
- (5) The prediction error of particular markers was estimated by permutation procedure from the test set.
- (6) Two indices of the RF procedures were used for the selection of relevant risk predictors, the GI and the VI.

The GI represents the total decrease in impurity of the whole dataset by summing the probability of each risk predictor being chosen multiplied by the probability of a mistake in categorizing a subset. The VI means the decrease in accuracy for every predictor. To avoid the bias induced by including highly correlated candidate markers such as SNPs in LD, we used the conditional permutation scheme in the tree building procedure⁴⁸. The conditional importance permutation groups were created, which involve all variables of a Pearson's correlation coefficient of $r \geq 0.2$, that is, the dependency structure between SNPs in linkage disequilibrium was preserved in the calculation of VI. To obtain an appropriate amount of predictors, the corresponding top ranked 200 SNPs of the GI and the conditional VI were considered as the candidate risk predictors for model construction in the next step.

Model construction and performance evaluation. Among the candidate risk predictors identified in the GAIN dataset using the RF analysis, we applied multivariable logistic regression with variable-selection methods (i.e. stepwise selection) to select the optimal predictors and to obtain p-values, odds ratios (OR), and 95% CI. The genetic risk score for each individual was calculated by summing across all predictors in the model using the numbers of risk allele multiplied by the beta regression coefficient of each marker. The highest Youden Index⁴⁹ was used to define the optimal cutoff point, which equals to (sensitivity + specificity) - 1.

We examined the performance of the classification models by several indices. First, the discrimination capability of the established prediction model was assessed with the receiver operating characteristic (ROC) curve, and an AUROC was calculated. The ROC curve was plotted by false-positive rate versus sensitivity measure. The goodness of fit for each prediction model was assessed by the Hosmer-Lemeshow test, which calculates the difference between the predicted and the observed risk. Leave-one-out cross-validation was performed for internal validation to obtain a bias-corrected estimation of error rate in prediction. The STEP GWA dataset was used as external validation for the prediction models. Statistical analyses in this stage were performed with SAS version 9.2 (SAS Institute, Cary, NC).

Identified significantly enriched pathways. We used the Molecular Signatures Database (MSigDB, <http://www.broadinstitute.org/gsea/msigdb/annotate.jsp>) to examine the common processes or the underlying biological gene sets of the selected candidate genes⁵⁰. In the present study, we used databases including GO terms (domains in biological process, cellular component and molecular function), chromosome positional gene sets, and the curated gene sets (e.g. canonical pathways, *KEGG*, *Biocarta* and *Reactome*). In total, 3,100 collections of gene sets were available, which includes 45,956 unique genes. Enriched gene sets were identified using the hypergeometric method, with the false discovery rate less than 0.01⁵⁰.

References

1. D'Agostino, R. B. Sr. *et al.* General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**, 743–753, doi: 10.1161/CIRCULATIONAHA.107.699579 (2008).
2. Wilson, P. W. *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837–1847 (1998).
3. Yang, H. I. *et al.* Risk estimation for hepatocellular carcinoma in chronic hepatitis B (REACH-B): development and validation of a predictive score. *Lancet Oncol* **12**, 568–574, doi: 10.1016/S1470-2045(11)70077-8 (2011).
4. Ayalew, M. *et al.* Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. *Mol Psychiatry* **17**, 887–905, doi: 10.1038/mp.2012.37mp201237 (2012).
5. Glatt, S. J. *et al.* Blood-based gene-expression predictors of PTSD risk and resilience among deployed marines: a pilot study. *Am J Med Genet B Neuropsychiatr Genet* **162B**, 313–326, doi: 10.1002/ajmg.b.32167 (2013).
6. van Hoek, M. *et al.* Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. *Diabetes* **57**, 3122–3128, doi: 10.2337/db08-0425db08-0425 (2008).
7. Maier, R. *et al.* Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am J Hum Genet* **96**, 283–294, doi: 10.1016/j.ajhg.2014.12.006 (2015).
8. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678, doi: 10.1038/nature05911 (2007).
9. Goldstein, D. B. Common genetic variation and human traits. *N Engl J Med* **360**, 1696–1698, doi: 10.1056/NEJMp0806284 (2009).
10. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427, doi: 10.1038/nature13595 (2014).
11. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753, doi: 10.1038/nature08494 (2009).
12. Kruppa, J., Ziegler, A. & König, I. R. Risk estimation and risk prediction using machine-learning methods. *Hum Genet* **131**, 1639–1654, doi: 10.1007/s00439-012-1194-y (2012).
13. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32, doi: 10.1023/a:1010933404324 (2001).
14. Diaz-Uriarte, R. & Alvarez de Andres, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3, doi: 10.1186/1471-2105-7-3 (2006).
15. Kalhan, S. C. *et al.* Plasma metabolomic profile in nonalcoholic fatty liver disease. *Metabolism* **60**, 404–413, doi: 10.1016/j.metabol.2010.03.006 (2011).
16. Gonzales, D. A. *et al.* Protein expression profiles distinguish between experimental invasive pulmonary aspergillosis and *Pseudomonas pneumonia*. *Proteomics* **10**, 4270–4280, doi: 10.1002/pmic.200900768 (2010).
17. Goldstein, B. A., Hubbard, A. E., Cutler, A. & Barcellos, L. F. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genet* **11**, 49, doi: 10.1186/1471-2156-11-49 (2010).
18. Edvardsen, J. *et al.* Heritability of bipolar spectrum disorders. Unity or heterogeneity? *J Affect Disord* **106**, 229–240, doi: 10.1016/j.jad.2007.07.001 (2008).
19. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234–239, doi: 10.1016/S0140-6736(09)60072-6 (2009).
20. Rolim-Neto, M. L. *et al.* Bipolar disorder incidence between children and adolescents: A brief communication. *J Affect Disord* **172**, 171–174, doi: 10.1016/j.jad.2014.09.045 (2015).
21. Medici, C. R., Videbech, P., Gustafsson, L. N. & Munk-Jorgensen, P. Mortality and secular trend in the incidence of bipolar disorder. *J Affect Disord* **183**, 39–44, doi: 10.1016/j.jad.2015.04.032 (2015).
22. Carlborg, A., Ferntoft, L., Thuresson, M. & Bodegard, J. Population study of disease burden, management, and treatment of bipolar disorder in Sweden: a retrospective observational registry study. *Bipolar Disord* **17**, 76–85, doi: 10.1111/bdi.12234 (2015).

23. Guze, S. B. & Robins, E. Suicide and primary affective disorders. *Br J Psychiatry* **117**, 437–438 (1970).
24. Smith, E. N. *et al.* Genome-wide association study of bipolar disorder in European American and African American individuals. *Mol Psychiatry* **14**, 755–763, doi: 10.1038/mp.2009.43 (2009).
25. Sklar, P. *et al.* Whole-genome association study of bipolar disorder. *Mol Psychiatry* **13**, 558–569, doi: 10.1038/sj.mp.4002151 (2008).
26. Yucebas, S. C. & Aydin Son, Y. A prostate cancer model build by a novel SVM-ID3 hybrid feature selection method using both genotyping and phenotype data from dbGaP. *PLoS One* **9**, e91404, doi: 10.1371/journal.pone.0091404 (2014).
27. Kim, J. W., Sharma, V. & Ryan, N. D. Predicting Methylphenidate Response in ADHD Using Machine Learning Approaches. *Int J Neuropsychopharmacol*, doi: 10.1093/ijnp/pyv052 (2015).
28. Bruining, H. *et al.* Behavioral signatures related to genetic disorders in autism. *Mol Autism* **5**, 11, doi: 10.1186/2040-2392-5-11 (2014).
29. Chung, R. H. & Chen, Y. E. A two-stage random forest-based pathway analysis method. *PLoS One* **7**, e36662, doi: 10.1371/journal.pone.0036662 (2012).
30. Xu, M. *et al.* Genome Wide Association Study to predict severe asthma exacerbations in children using random forests classifiers. *BMC Med Genet* **12**, 90, doi: 10.1186/1471-2350-12-90 (2011).
31. Botta, V., Louppe, G., Geurts, P. & Wehenkel, L. Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS One* **9**, e93379, doi: 10.1371/journal.pone.0093379 (2014).
32. Golan, D. & Rosset, S. Effective genetic-risk prediction using mixed models. *Am J Hum Genet* **95**, 383–393, doi: 10.1016/j.ajhg.2014.09.007 (2014).
33. Skafidas, E. *et al.* Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. *Mol Psychiatry* **19**, 504–510, doi: 10.1038/mp.2012.126 (2014).
34. Pramparo, T. *et al.* Prediction of autism by translation and immune/inflammation coexpressed genes in toddlers from pediatric community practices. *JAMA Psychiatry* **72**, 386–394, doi: 10.1001/jamapsychiatry.2014.3008 (2015).
35. Schulze, T. G. *et al.* Molecular genetic overlap in bipolar disorder, schizophrenia, and major depressive disorder. *World J Biol Psychiatry* **15**, 200–208, doi: 10.3109/15622975.2012.662282 (2014).
36. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* **43**, 977–983, doi: 10.1038/ng.943ng.943 (2011).
37. Sharp, S. I. *et al.* Genetic association of the tachykinin receptor 1 TACR1 gene in bipolar disorder, attention deficit hyperactivity disorder, and the alcohol dependence syndrome. *Am J Med Genet B Neuropsychiatr Genet* **165B**, 373–380, doi: 10.1002/ajmg.b.32241 (2014).
38. Hibar, D. P. *et al.* Common genetic variants influence human subcortical brain structures. *Nature* **520**, 224–229, doi: 10.1038/nature14101 (2015).
39. Nissen, S. *et al.* Evidence for association of bipolar disorder to haplotypes in the 22q12.3 region near the genes stargazin, IFT27 and parvalbumin. *Am J Med Genet B Neuropsychiatr Genet* **159B**, 941–950, doi: 10.1002/ajmg.b.32099 (2012).
40. Chuang, L. C., Kao, C. F., Shih, W. L. & Kuo, P. H. Pathway analysis using information from allele-specific gene methylation in genome-wide association studies for bipolar disorder. *PLoS One* **8**, e53092, doi: 10.1371/journal.pone.0053092 (2013).
41. Nurnberger, J. I. Jr. *et al.* Identification of pathways for bipolar disorder: a meta-analysis. *JAMA Psychiatry* **71**, 657–664, doi: 10.1001/jamapsychiatry.2014.1761859133 (2014).
42. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* **9**, e1003348, doi: 10.1371/journal.pgen.1003348 (2013).
43. Wong, M. L., Dong, C., Andreev, V., Arcos-Burgos, M. & Licinio, J. Prediction of susceptibility to major depression by a model of interactions of multiple functional genetic variants and environmental factors. *Mol Psychiatry* **17**, 624–633, doi: 10.1038/mp.2012.13 (2012).
44. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816–834, doi: 10.1002/gepi.20533 (2010).
45. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575, doi: 10.1086/519795 (2007).
46. Schwarz, D. F., Konig, I. R. & Ziegler, A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics* **26**, 1752–1758, doi: 10.1093/bioinformatics/btq257 (2010).
47. Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* **10**, 392–404, doi: 10.1038/nrg2579 (2009).
48. Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T. & Zeileis, A. Conditional variable importance for random forests. *BMC Bioinformatics* **9**, 307, doi: 10.1186/1471-2105-9-307 (2008).
49. Biggerstaff, B. J. Comparing diagnostic tests: a simple graphic using likelihood ratios. *Stat Med* **19**, 649–663, doi: 10.1002/(SICI)1097-0258(20000315)19:5<649::AID-SIM371>3.0.CO;2-H (2000).
50. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**, 15545–15550, doi: 10.1073/pnas.0506580102 (2005).

Acknowledgements

This work was supported by the Ministry of Science and Technology (MST 102-2314-B-002-117-MY3) and by the National Taiwan University (Career Development Project: 104R7883) to Dr. Po-Hsiu Kuo). We thank the Genetic Association Information Network (GAIN) and the Systematic Treatment Enhancement Program (STEP), which makes genotyping data publically available. We also thank Po-Chang Hsiao for his IT support.

Author Contributions

Acquisition of data: L.C.C.; Revised manuscript critically for important intellectual content: P.H.K. and L.C.C.; Final approval of the version to be published: P.H.K.; Conceived and designed the experiments: P.H.K. and L.C.C.; Analyzed the data: L.C.C.; Wrote the paper: L.C.C.

Additional Information

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Chuang, L.-C. and Kuo, P.-H. Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm. *Sci. Rep.* **7**, 39943; doi: 10.1038/srep39943 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017