Published in final edited form as:

J Autism Dev Disord. 2012 June ; 42(6): 971–983. doi:10.1007/s10803-011-1327-5.

Single nucleotide polymorphisms predict symptom severity of autism spectrum disorder

Yun Jiao^{a,b,c}, Rong Chen^c, Xiaoyan Ke^{a,d}, Lu Cheng^b, Kangkang Chu^d, Zuhong Lu^{a,b,*}, and Edward H Herskovits^{c,*}

^a Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education

^b State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University

^c Department of Radiology, University of Pennsylvania School of Medicine

^d Child Mental Health Research Center of Nanjing Brain Hospital, Nanjing Medical University

Abstract

Autism is widely believed to be a heterogeneous disorder; diagnosis is currently based solely on clinical criteria, although genetic, as well as environmental, influences are thought to be prominent factors in the etiology of most forms of autism. Our goal is to determine whether a predictive model based on single-nucleotide polymorphisms (SNPs) can predict symptom severity of autism spectrum disorder (ASD). We divided 118 ASD children into a mild/moderate autism group (n = 65) and a severe autism group (n = 53), based on the Childhood Autism Rating Scale (CARS). For each child, we obtained 29 SNPs of 9 ASD-related genes. To generate predictive models, we employed three machine-learning techniques: decision stumps (DSs), alternating decision trees (ADTrees), and FlexTrees. DS and FlexTree generated modestly better classifiers, with accuracy = 67%, sensitivity = 0.88 and specificity = 0.42. The SNP *rs878960* in *GABRB3* was selected by all models, and was related associated with CARS assessment. Our results suggest that SNPs have the potential to offer accurate classification of ASD symptom severity.

* Correspondence concerning this article should be addressed to either: Zuhong Lu zhlu@seu.edu.cn Telephone: +86-25-83793779 Address: Key Laboratory of Child Development and Learning Science, Southeast University, Nanjing, Jiangsu, 210096, China Edward H Herskovits ehh@ieee.org Telephone: +1-215-615-3705, Facsimile: +1-215-615-3681 Address: 3600 Market Street, Science Center, Suite 370, Philadelphia, PA 19104, USA.

Author Note

Yun Jiao: Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education, Nanjing, China; State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China; Department of Radiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA.

Rong Chen: Department of Radiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA.

Xiaoyan Ke: Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education, Nanjing, China; Child Mental Health Research Center of Nanjing Brain Hospital, Nanjing Medical University, Nanjing, China.

Lu Cheng: State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China.

Kangkang Chu: Child Mental Health Research Center of Nanjing Brain Hospital, Nanjing Medical University, Nanjing, China.

Zuhong Lu: Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education, Nanjing, China; State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China.

Edward H Herskovits: Department of Radiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA. No changes in the author affiliation subsequent to the time of the study.

Keywords

autism-spectrum disorder; single-nucleotide polymorphisms; diagnostic model; genotype-phenotype analysis; data mining

Introduction

Autism-spectrum disorder (ASD) is a pervasive neurodevelopmental disorder characterized by abnormal social behavior, impaired communication, and repetitive/stereotyped behavior (Gmitrowicz & Kucharska, 1994). Family-based genetic studies in ASD have demonstrated high heritability of the narrow and broad phenotypes of ASD (Freitag, 2007; Geschwind, 2009), and have shown significant associations between ASD and genetic factors (Belmonte et al., 2004). Researchers have shown that ASD is associated with many genes, such as *GABRA4*, *GABRA2*, *GABRB1*, *GABRB2*, *GABRB3*, *TDO2*, *SLC25A12*, and brain-derived neurotrophic factor (*BDNF*) (Belmonte et al., 2004; Cheng et al., 2009; S.-J. Kim et al., 2008; Nabi, Serajee, Chugani, Zhong, & Huq, 2004).

Single-nucleotide polymorphisms (SNPs) are genetic markers that enable researchers to search for genes associated with complex diseases. Accordingly, many studies have centered on genetic changes in ASD patients based on SNP analysis. Freitag (Freitag, 2007; Freitag, Staal, Klauck, Duketis, & Waltes, 2010) reviewed ASD-related SNP studies, reporting that SNPs in chromosomes 2, 3, 4, 6, 7, 10, 15, 17, X and Y were associated with ASD.

Most of these studies were family studies. For example, chromosome 15q11-q13 (including ASD-related genes, such as *GABRB3*, *GABRA5*) has been considered to be an autism candidate region (Kim et al., 2008), primarily because of reports of maternal interstitial duplication associated with autism and positive linkage and association studies in chromosomally normal autism families (Nurmi, Amin et al., 2003). Other researchers (Ashley-Koch et al., 2006; Buxbaum et al., 2002; Cook et al., 1998; McCauley et al., 2004) also found evidence, based on family studies, that chromosome 15q11-q13 is an autism candidate region. Ma et al. (2005) identified association and gene-gene interactions of GABA receptor subunit genes with ASD based on family studies, and found that *GABRA4*, *GABRR2*, *GABRA2* and *GABRB1* are associated with ASD. Collins et al. (2006) investigated ASD and GABA-receptor subunit genes in several ethnic groups; they found that *GABRA4* and *GABRB1* contribute to ASD susceptibility. Other researchers (Lerer et al., 2008; Wermter et al., 2010) reported that genetic variation in the oxytocin receptor (*OXTR*) gene was relevant to the etiology of ASD. Nabi et al. (2004) investigated 5 SNPs in the *TDO2* gene for association with autism, and found that transmission of a promoter variant in differs between normal and autistic subjects.

There exist many case-control studies of SNPs associated with ASD. For example, Cheng et al. (2009) showed significant differences in allele frequencies between the ASD and control groups in an association study of four *BDNF* polymorphisms. Li et al. (2005) recruited 40 ASD patients and 51 normal controls in their analysis of the *FOXP2* gene—which encodes a putative transcription factor containing a polyglutamine tract and a forkhead DNA-binding domain—for possible causative mutations in autism; their results suggest a relationship between autism and the *FOXP2* gene or a gene located nearby. Ramoz et al. (2004) and Segurado et al. (2005) compared two SNPs (*rs2056202* and *rs2292813*) in the gene *SLC25A12* in family-based and case-control association studies, and found an increased risk for autism associated with the haplotype GG (reverse strand) = CC (sense strand), consisting of the two SNPs.

Little is known about the degree to which genetic variations underlie symptom variability in ASD; this is especially true for children with the more common form of ASD (typical autism), at least in part because the phenotypic patterns of the three ASD core impairments is extremely variable (Parks et al., 2009). To shed greater light on this issue, researchers have sought to study overall symptom severity in young children with ASD, by comparing known or strongly suspected ASD-related genes, or SNPs in such genes, with overall symptom severity. Lerer et al. (2008) found SNPs and haplotypes in the *OXTR* gene that are associated with IQ and total VABS (Vineland adaptive behavioral scales, (Sparrow & Cicchetti, 1985)) scores (as well as the communication, daily living skills and socialization subdomains). However, Wermter et al. (2010) found that three categories (social, communication and behavior) of the ADI-R (Autism Diagnostic Interview-Revised (Lord, Rutter, & Le Couteur, 1994)) scores did not significantly differ between haplotype carriers and noncarriers with respect to the *OXTR* gene. Kim et al. (2008) also attempted to associate the ADI-R subdomain to SNPs that they had determined to be associated with ASD; they found no significant difference between different genotypes of SNPs that they selected.

Currently, symptom severity in ASD is determined based on behavioral criteria. The promises of computer-generated diagnostic models (classifiers) to predict symptom severity of ASD based on SNPs are twofold. First, the classifiers could identify genetic markers for subsequent investigation. Second, computer-based classifiers may provide prognostic information with respect to symptom severity that is complementary to that obtained by behavioral assessment.

In fact, SNPs have been used to help predict whether an individual has a certain disease, or a subtype of a certain disease (Schwender, Ickstadt, & Rahnenfuhrer, 2008). SNP-based diagnostic models have been constructed for several diseases. For example, Huang et al. (2004) used Flextree to differentiate hypertension from hypotension and obtained sensitivity 0.65 and specificity 0.54. Bureau et al. (2005) used the random forest method to distinguish the asthma group from normal controls, and were able to classify samples with accuracy of 0.55. Nunkesser et al. (2007) applied GPAS (genetic programming for association studies, an advanced version of logic regression (Koopberg, Ruczinski, LeBlanc, & Hsu, 2001)) to discriminate breast cancer cases from normal controls; they obtained a classification accuracy of 0.61. Park and Hastie (2008) used penalized logistic regression to classify bladder-cancer patients from normal controls, and obtained sensitivity of 0.61 and specificity of 0.68.

To our knowledge, SNP-based diagnostic models have not been reported in previous ASD symptom-severity studies. We therefore designed experiments to test the hypothesis that diagnostic models based on SNPs could distinguish children with different degrees of ASD-symptom severity. In particular, we assessed 29 SNPs in 9 ASD-related genes, based on a review of the literature. We then applied three data-mining approaches to generate three diagnostic models based on these SNPs. Finally, we identified important markers for overall symptom severity based on SNP-based diagnostic models.

Methods

Patients

Subjects with ASD were recruited by the Child Mental Health Research Center of Nanjing Brain Hospital. The diagnosis of ASD was based on the criteria of the fourth edition of Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) (Gmitrowicz & Kucharska, 1994), the Autism Diagnostic Inventory-Revised (ADI-R) (Lord et al., 1994),

and the Childhood Autism Rating Scale (CARS) (Schopler, Reichler, DeVellis, & Daly, 1980).

Exclusion criteria for all subjects included history of seizure, head trauma, genetic or neurological disorder, and major medical problem. One hundred eighteen subjects with a diagnosis of ASD met criteria for inclusion in this study. Participations ranged in age from 1.5 to 14 years (mean age = 5.45 ± 2.24) and were from Chinese Han families. Figure 2 and Appendix 6 show that there is no significant difference in genotype or allele distributions between our study and other Asian groups.

Group membership

The CARS is a behavior-based clinical scale derived from interaction and observation; this score has been shown to have a high degree of internal consistency, inter-rater and test-retest reliability, high criterion-related validity, and good discriminant validity (Parks et al., 2009). Behavior is rated on 15 items on a scale of 1 (age appropriate) to 4 (severely autistic). These items include language and communication skills, response to sensory information, and socioemotional and interactional skills (Schopler et al., 1980). A total score of 30 or more is indicative of autism; this score distinguishes between mild/moderate and severe autism (Schopler et al., 1980). In this study, we used the CARS score to verify a diagnosis of ASD, as well as to rate ASD symptoms as mild/moderate or severe. According the CARS manual, the scores of individuals with ASD are bimodally distributed; Appendix 1 displays the histogram of CARS scores for this study. Mild to moderate ASD is defined by scores between 30 and 36.5, and severe ASD is defined by scores between 37 and 60 (Parks et al., 2009). As a result, we set CARS scores of 36.5 as the threshold to divide participants into two groups. This resulted in 65 participants (57 boys, 8 girls) in the mild/moderate group, and 53 (46 boys, 7 girls) participants in the severe group; Table 1 lists details of these two groups.

Genotyping

Genomic DNA was extracted from peripheral-blood leukocytes using an improved potassium iodide salting-out procedure (Cheng et al., 2009). Based on literature review, we selected nine ASD-related genes (*GABRA4*, *GABRB1*, *TDO2*, *GABRB2*, *GABRA2*, *GABRB3*, *GABRA5*, *SLC25A12*, and *BDNF*; details in Table 2) for this study; these genes had been found to be associated with ASD symptoms (Belmonte et al., 2004; Cheng et al., 2009; Freitag, 2007; S.-J. Kim et al., 2008; Nabi et al., 2004). We selected SNPs within each ASD-related candidate region and identified them from NCBI's SNP database (dbSNP) (<http://www.ncbi.nlm.nih.gov/SNP>). In Appendix 4 we describe further details regarding these SNPs and genes.

We performed genotyping via 3-dimensional polyacrylamide gel-based microarray hybridized with dual-color fluorescent probes, which was developed by Dr. Lu (Hou, Ji, Li, & Lu, 2004; Ji, Hou, Li, He, & Lu, 2004). This method is based on immobilizing amino-modified polymerase chain reaction (PCR) products onto poly-L-lysine coated glass slides to fabricate a microarray, which is then interrogated by hybridization with dual-color probes to determine the SNP genotype of each sample (Xiao et al., 2006). Three genotypes can be analyzed using green, red and yellow colors. The entire SNP-detection procedure consists of five steps: PCR, immobilization of PCR products, hybridization, electrophoresis of the microarray, and scanning for genotyping. Appendix 5 describes genotyping details. Finally, we labeled each genotyped SNP according to Table 2.

Diagnostic-model generation

To generate diagnostic models from these data, we employed R (<http://www.r-project.org/>) and WEKA (<http://www.cs.waikato.ac.nz/ml/weka/>) (Witten & Frank, 2005). A classification model (diagnostic model) includes two components: the structural form of the model, S , and model parameters, θ . The input to SNP-based diagnostic-model generation consisted of 118 instances of 29 SNP variables; the group-membership variable was ASD symptom severity (either “M” for the mild/moderate group or “S” for the severe group).

To avoid model-generation bias associated with respect to the model's structural form, we applied three machine-learning methods—decision stumps (DSs), alternating decision trees (ADTrees), and FlexTrees—to generate diagnostic models. Our rationale for choosing tree-based models includes their declarative nature, ease of interpretation of the resulting models, and the ability to handle categorical variables (e.g. SNP data). Tree-based models work well with categorical variables, whereas other methods (e.g., margin-based) are better at prediction based on continuous variables. In addition, tree-based algorithms can identify variables with high predictive power (Huang et al., 2004). As a result, tree-based algorithms have been widely used in SNP-based classification (Bureau et al., 2005; Huang et al., 2004; Nunkesser et al., 2007; Park & Hastie, 2008).

We selected the DS classifier because it is a single-level decision tree, which should indicate the most important predictive variable in the model. We selected the ADTree classifier because it is an advanced version of a standard decision tree, which is a generalization of decision trees based on boosting; the ADTree classifier has been shown to be accurate across a variety of applications (Freund & Mason, 1999). Finally, we selected FlexTrees because they can be seen as an advanced version of standard CART (classification and regression trees), and performed better than CART, QUEST (quick, unbiased and efficient statistical tree (Loh & Shih, 1997)), logic regression, and random forest (Huang et al., 2004). FlexTree has been used to determine genetic predisposition to multifactorial diseases (Huang et al., 2004). We provide additional details regarding these approaches in Appendix 3.

Diagnostic-model evaluation

We evaluated the diagnostic models generated by these three machine-learning algorithms based on 10-fold cross-validation. In particular, during each iteration, we randomly divided the subjects into 10 equal-sized groups, with mild/moderate subjects being mixed with severe subjects. We then used 90% of the subjects for classifier generation (i.e., training) and the remaining 10% for testing. After cycling through all 10 partitions for classification, we used every group for testing, and each group appeared in a training set every time except when it was used for testing.

There were 38 missing values (missing value rate = 1.1%) in our study. We estimated all missing values for SNP attributes using the modes from the training data.

We evaluated the performance of each diagnostic model based on three metrics: the true-positive rate (TPR), the false-positive rate (FPR), and accuracy (ACC). Accuracy is the proportion of correctly labeled instances. The true-positive rate (sensitivity) is the proportion of positive instances (severe ASD children) that were correctly reported as being positive. The false-positive rate (1 – specificity) is the proportion of negative instances (mild/moderate ASD children) that were erroneously reported as being positive.

Genotype-Phenotype Analyses

In our study, there were three SNP-genotype groups: homozygous wild type, heterozygote, and homozygous mutant type. We used one-way ANOVA with the Tukey honest significant differences (HSD) test (Miller, 1981) to detect differences between CARS score (phenotype) across SNPs (genotype).

Genetic analysis

We analyzed differences in genotype and allele distributions, and performed haplotype analysis, to identify associations between SNPs included in our classification model and ASD symptom severities. We analyzed genotype and allele distribution differences using methods described in Shi & He (2005). We performed haplotype analysis using the methods described in (Li et al., 2009).

Results

Table 1 lists demographic and behavioral variables for the subjects in the two groups.

For SNP-based diagnostic models, both DS and FlexTree selected only *rs878960* in *GABRB3* as a predictive variable. ADTree built a complex model that included ten predictive variables (please see image in Appendix 2), and selected *rs878960* as a priority root node. All three tree-based methods selected *rs878960* in *GABRB3* as the root node.

The models generated by DS and FlexTree were identical: if a subject carries A/A of *rs878960*, that subject will be classified as having severe autism; carriers of A/G or G/G will be classified as having mild/moderate autism. We found that CARS scores were significantly different between the A/A group and the (A/G + G/G) group (two-sample t-test $p = 0.0004$). Figure 1 show the results of analysis of *rs878960* genotypes and CARS scores. The mean CARS scores for the three genotypes A/A, A/G, and G/G were 37.5 (standard deviation, SD = 3.9), 33.7 (SD = 5.1), and 35.6 (SD = 4.7), respectively. CARS scores differed significantly among these three genotype groups (ANOVA p -value = 0.0013), particularly between the A/A and A/G groups (Tukey HSD p -value < 0.001). However, there was no significant difference between the A/G and G/G groups (Tukey HSD p -value = 0.17), or between the A/A and G/G groups (Tukey HSD p -value = 0.27).

In our ASD sample, the *rs878960* diversity scores (numbers of each genotype divided by the total numbers of subjects) for the three genotypes A/A, A/G, and G/G were 0.25, 0.51, and 0.24, respectively. We compared these diversity scores to those reported for other samples (Gibbs et al., 2003) on this SNP in Figure 2. We found no significant differences (details in Appendix 6) in genotypes and allele distributions with respect to this SNP between our ASD subjects and normal Asian populations in the International HapMap Project (Gibbs et al. 2003).

In contrast, genotypes and the allele distributions differed between the severe and mild/moderate groups (p -value < 0.05), as shown in Table 3. In addition, we performed haplotype analysis for each gene, the details of which are presented in Appendix 7.

Discussion

To our knowledge, there has been no previous attempt to distinguish children with ASD with different symptom severities based on SNPs. We found that SNP-based diagnostic models can differentiate mild/moderate ASD from severe ASD; although classification performance is not very high, this analysis was based on only 29 genes, and we would expect performance to increase as we obtain more genetic information as input to the data-

mining algorithms. The performance metrics of our classification models are comparable to those of SNP-based diagnostic models for other types of diseases (hypertension (Huang et al., 2004), asthma (Bureau et al., 2005), breast cancer (Nunkesser et al., 2007), and bladder cancer (Park & Hastie, 2008)).

We used three machine-learning methods to generate classifiers, in order to avoid bias with respect to the functional form of the classifier. We found that: 1) all methods yielded similar performance. 2) DS and FlexTree performed modestly better than ADTree in terms of accuracy. DS and FlexTree generated a very simple tree model consisting of one predictor variable (*rs878960* in *GABRB3*), which is the most predictive SNP, whereas ADTree generated a complex model that may dilute the contribution of SNP *rs878960* in *GABRB3*. 3) The SNP *rs878960* in *GABRB3* was selected as a predictive variable by all three tree-based methods, indicating that this SNP is potentially central in distinguishing ASD children with different symptom severities.

During fetal life, the 15q GABA_A receptor subunit cluster (including *GABRB3*, *GABRA5*, and *GABRG3*) plays a developmental role in GABAergic signaling for establishing neuronal connectivity, and a critical role in the maintenance of inhibitory tone in the adult brain (Nurmi, Dowd et al., 2003). Previous gene-based studies found that mutations in *GABRB3* are associated with ASD. For example, Kim et al. (S. A. Kim, Kim, Park, Cho, & Yoo, 2006) found an allele at *rs2081648* in *GABRB3* that was preferentially transmitted in a family-based ASD association study. Similarly, Ashley-Koch et al. (2006) and McCauley et al. (2004) used SNP-based methods and found that *GABRB3* is significantly associated with ASD in family-based association studies. Both Buxbaum et al. (2002) and Cook et al. (1998) found that the microsatellite marker 155CA-2 in *GABRB3* was related to ASD in family-based association studies.

The SNP *rs878960* was found to be associated with ASD status in an overall pedigree disequilibrium test (McCauley et al., 2004). The microsatellite marker 155CA-2, which lies ~80 kb centromeric to SNP *rs878960*, has been found to be associated with ASD status (Cook et al., 1998; Buxbaum et al., 2002). Linkage studies (McCauley et al., 2004) in autism subsets have pointed to the region corresponding to the microsatellite marker 155CA-2, with peak linkage occurring at the 5' end of *GABRB3* at D15S511 (Nurmi et al., 2003), ~40 kb from 155CA-2.

To our knowledge, little is understood about the functional effects of SNP *rs878960*. Delahanty et al. (2011) reported that maternal transmission of *rs25409* related to a *GABRB3* signal peptide variant (P11S) is associated with autism. SNP *rs878960* lies ~87 kb from SNP *rs25409*, which may indicate that these two SNPs have a functional relationship. Sutcliffe & Nurmi (2003) suggested that dup(15)-mediated autism is a contiguous gene-duplication effect requiring the GABA_A subunit genes in addition to imprinted, maternally-expressed genes.

The existence of several autism risk alleles for this gene may be the explanation for the observation of an association between *rs878960* (*GABRB3*) and autism symptom severity (or autism) (McCauley et al., 2004).

To test the association between *GABRB3* and ASD symptom severity, we first obtained the genotypes and allele distributions (Shi & He, 2005) for 6 *GABRB3* SNPs (Table 4). We found that only *rs878960* was associated with ASD symptom severity scores. Then, we performed haplotype analysis (Li et al., 2009) of *GABRB3*. We found that none of the haplotypes of all 6 SNPs in *GABRB3* were significantly associated with ASD symptom severity scores; however, when we removed from the analysis *rs1432007*—the SNP with the highest *p*-value in genotype-distribution analysis—and repeated the haplotype analysis, we

obtained $\chi^2 = 19.5$, p -value = 0.007. That is, we found that the haplotypes of the other 5 SNPs in *GABRB3* were significantly associated with symptom severities of ASD when we removed *rs1432007* from the analysis.

We found that CARS scores differed significantly with respect to SNP *rs878960* in *GABRB3*, particularly between the genotypes A/A (homozygous wild type) and A/G (heterozygous wild/mutant). The association between the *GABRB3* gene and CARS may suggest a genetic validation of recent ASD postmortem studies, which indicated that *GABRB3* expression in the ASD group was significantly reduced in parietal cortex (Brodmann's Area 40, BA40) (Fatemi, Reutiman, Folsom, & Thuras, 2009) and in frontal cortex (BA9) (Samaco, Nagarajan, Braunschweig, & LaSalle, 2004), relative to the normal control group.

In addition, we found that CARS scores differed significantly between the A/A and (A/G + G/G) groups (p -value = 0.0004). The A/G group for SNP *rs878960* had the greatest number of subjects in our study; this group had the lowest average CARS score of the three genotype groups (in Figure 1), and was significantly different from the A/A group, which had the highest average CARS score. These were the reasons that we found high sensitivity (0.88) for the DS and FlexTree algorithms. This result indicates that ASD patients carrying A/G copies of *rs878960* are likely to have mild/moderate ASD, whereas ASD patients carrying the A/A genotype of *rs878960* are likely to have severe ASD. We also found that the CARS scores for subjects with the G/G genotype overlapped with those for the A/A and A/G genotype groups (Figure 1). However, only a few subjects carried the G/G genotype of *rs878960*, which therefore contributed less to overall classification accuracy. On the basis of our results, when an ASD individual carries the G/G genotype of *rs878960*, investigation of other ASD-related genes should be included. As a result, in ASD patients, the A/A genotype in *rs878960* may indicate a high risk of severe autism; A/G and G/G indicate less risk. In the future, we will further investigate the results of Delahanty et al. (2011), who found potential associations between *rs878960* genotypes and autism severity; we will also investigate the genotype of *rs25409*, and linkage of these two SNPs, in our samples. In summary, *rs878960* in *GABRB3* is a potential biomarker for predicting symptom severity for ASD patients.

The results reported in this study, and contemporaneous studies, should be interpreted with caution, because the diagnostic criteria for ASD may be changed in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V). In particular, studies using the new diagnostic criteria may not be comparable to studies based on the current diagnostic criteria.

One limitation of this work is that this study includes only 29 SNPs; we did not include all genes potentially related to ASD (Freitag, 2007; Freitag et al., 2010), such as *OXTR* (Wermter et al., 2010), *SLC6A4* (Belmonte et al., 2004), *GABRG3* (Ashley-Koch et al., 2006), *ATP10C* (Nurmi, Amin et al., 2003), and *UBE3A* (Samaco, Hogart, & LaSalle, 2005), due to budgetary constraints. Our SNP selection was based on literature review with attention to ASD-related candidate genes.

Our approach to classifier generation, given sufficient computational resources, would readily scale to a genome-wide study. The primary advantage of a genome-wide analysis is that no prior hypotheses have to be made regarding which genes are associated with ASD. In the future, we plan to analyze additional genes, and will ultimately scale to full genome-wide analysis.

Another limitation of our work is that current machine-learning techniques cannot address overfitting very well. Some genes represent noise in the setting of classification: they reduce the contribution of genes that are useful for classification (Huang et al 2004), and thereby

lead to overfitting. For example, the genotypes and the allele distributions of *rs878960* polymorphisms differed significantly between severe and mild/moderate groups (p -value < 0.05 , Table 3). When we analyzed the remaining SNPs of *GABRB3*, we found that there were no significant differences in genotypes and allele distributions between these two groups (Table 4), possibly obscuring the contribution of *rs878960*. FlexTree can partly solve the overfitting problem (Huang et al 2004); indeed, it did not select the other *GABRB3* SNPs. We will develop a Bayesian algorithm to reduce overfitting.

In conclusion, this study represents the first attempt to classify individuals with mild/moderate or severe ASD symptoms by constructing SNP-based diagnostic models. We found that SNP-based classification is moderately accurate for three different classification approaches, and three classification metrics. The DS and FlexTree approaches generated simple declarative models: one SNP—*rs878960* in *GABRB3*—was included in all three models, and was associated with CARS scores.

Acknowledgments

Yun Jiao was supported by the China Scholarship Council (No. 2008101370), the National Natural Science foundation of China (No. 30570655), and the Scientific Research Foundation of Graduate School of Southeast University (No. YBJJ1011). Drs. Chen and Herskovits are supported by National Institutes of Health grant R01 AG13743, which is funded by the National Institute of Aging, the National Institute of Mental Health, and the National Cancer Institute. They are also supported by NIH R03 EB009310. Drs. Ke and Chu were supported by the Natural Science Foundation of Jiangsu, China (No. BK2008082). Drs. Lu and Cheng were supported by the National Natural Science foundation of China (No. 30570655). The authors also thank the International HapMap Project for the data of normal Asian population diversity on *rs878960* in *GABRB3*, and Richard Olshen and Jing Huang for the code of FlexTree.

Appendix 1: Histogram of CARS scores

Please see Figure 3.

Appendix 2: Tree model generated by ADTree

Please see Figure 4.

Appendix 3: Data-Mining Methods

In this section, we provide an overview of DS, ADTree, and FlexTree.

DS is a single-level decision-tree model with a categorical or numeric class label (Iba & Langley, 1992). It tends to find the main predictor variable in one step. It is widely used when researchers seek the single most significant feature with respect to classification (Iba & Langley, 1992).

An ADTree is a method based on combining weak hypotheses generated during boosting into a single interpretable representation (Freund & Mason, 1999). An ADTree model is more compact than standard boosting-based decision-tree models, which generate more than one tree (Freund & Mason, 1999). As a result, an ADTree model is relatively straightforward to interpret. The application of boosting procedures may improve classification performance for ADTrees. The structure of an ADTree has three characteristics: 1) the root node is a prediction node, and has a numeric score only, which is based on the total weights of the positive and negative instances that satisfy the conditions in the training data (Holmes, Pfahringer, Kirkby, Frank, & Hall, 2002); 2) the nodes in the next layer are decision nodes, and are essentially a collection of decision-tree stumps; 3) the subsequent layers alternate layers of prediction nodes and decision nodes. To classify a new

instance with an ADTree model, all paths for which all decision nodes are true are followed, summing any prediction nodes that are traversed by these paths.

FlexTree, a general supervised-learning method, extends the binary tree-structured approach (Classification and Regression Trees, CART (Breiman, 1984)), although it differs greatly in its selection and combination of predictors (Huang et al., 2004). It is particularly applicable for assessing gene-gene and gene-environment interactions as they bear on complex diseases. FlexTree creates a simple rooted binary tree with each split defined by a linear combination of selected variables. The linear combination is determined by regression with optimal scoring; the variables are selected by a backward pruning procedure. Using a selected variable subset to define each split increases interpretability, improves predictive robustness, and prevents overfitting. FlexTree deals with additive and interactive effects simultaneously. Sampling units can be families or individuals, depending on the application. Generally, FlexTree demonstrated performance that is better than many alternatives to which it was compared, particularly when a small fraction of candidate genes are useful for classification (Huang et al., 2004).

Appendix 4: Rationale for SNP selection

GABA is the major inhibitory neurotransmitter in the adult brain, although it mediates excitatory transmission during development. As a result, many GABA_A receptors encoding genes were involved in our study. Previous autism pathophysiology studies reported that: 1) The numbers of GABA_A receptors were significantly decreased in brains of children with autism (Blatt et al., 2001); 2) Plasma GABA, and its essential precursor glutamate, were elevated in children with autism (Moreno-Fuenmayor et al., 1996; Dhossche et al., 2002; Aldred et al., 2003); 3) Benzodiazepines, which are effective in treating the seizures, anxiety, and social phobia that occur in the setting of autism, bind to, and act on, GABA_A receptors (Olsen & Macdonald, 2002); 4) GABA-ergic transmission has important trophic actions during development. Based on these data, the GABA_A receptor subunit genes, particularly those in 15q11-q13 (Cook et al., 1998; Buxbaum et al., 2002; McCauley et al., 2004; Ashley-Koch et al., 2006), represent excellent candidates, allelic variants of which could confer genetic susceptibility for development of autism (McCauley, 2005).

TDO2 (Nabi et al., 2004), *SLC25A12* (Ramos et al., 2004; Segurado et al., 2005), and *BDNF* (Cheng et al., 2009) were also found to be associated with ASD, so we included these genes in our studies.

Appendix 5: Genotyping

The first step in genotyping was PCR; primers were designed using Primer Premier 5.0 software, based on published DNA sequences. The primers were synthesized and HPLC purified by the TaKaRa Company (P.R. China). All reverse primers were modified with an acrylamide group at the 5'-terminal, in order to covalently bond to the polyacrylamide gel. After several cycles of PCR amplification, we used ethanol to precipitate PCR products.

Step 2 was immobilization of PCR products. We dissolved acrylamide-modified PCR products, and spotted them on 3-methacryloxypropyltrimethoxy silane-modified glass slides, using a microarrayer (Captial Biochip Corporation, P.R. China). Each slide was placed into a humid, 1,000 Pascal (Pa) pressure-sealed chamber full of tetramethylethylenediamine, to induce copolymerization between acrylamide groups and acryl groups. We then used electrophoresis to obtain single stranded DNA (ssDNA) for hybridization.

Step 3 was hybridization. We designed a pair of probes for every SNP locus, such that the probes could be matched with the polymorphic portion of the targets, and labeled with Cy3

or Cy5. For every SNP genotyped, we mixed the labeled probes in equimolar amounts, and suspended them in unihybridization solution (3:1 dilution) to obtain a final concentration of 2 μ M. We achieved hybridization in a humid chamber at 37 °C for 2-4 hours.

The fourth and fifth steps were post-hybridization and scanning, respectively. We rinsed the slide in water and air dried it, after which we completed electrophoresis at 2 V/cm for 8 min in 1X TBE buffer at 4° C. We scanned the hybridization slides at 70% laser power and 65% photomultiplier tubes gain with a confocal scanner (Luxscan-10K/A, CapitalBio Company, P.R. China) that had been fitted with filters for Cy3 and Cy5. We used QuantArray software (Packard BioChip Technologies, Billerica, MA) to analyze these images.

Appendix 6: rs878960 genotype and allele distributions between our study and other Asian cohorts

The genotypes and the allele distributions for *rs878960* polymorphisms in our study (n = 118) and for the HapMap Han Chinese group (n = 43, total 45 subjects but 2 missing values) are presented in Table 5. There is no significant difference in genotype or allele distribution between these two groups. Similarly, there is no significant difference in genotype or allele distributions between our Chinese subjects and Japanese subjects (HapMap, n = 86).

Appendix 7: Haplotype analysis for each gene

Table 6 shows the results of haplotype analysis using methods described in (Li et al., 2009). We found that haplotypes of *GABRA4*, but not of *GABRB3*, are significantly associated with ASD symptom severity. The reason for the lack of association with *GABRB3* is that we tested 6 SNPs in *GABRB3*, and some of these SNPs may contribute noise to the analysis. In particular, when we remove one SNP—*rs1432007*—and repeat the analysis (see Table 7), we obtain $\chi^2 = 19.5$, *p*-value = 0.007. That is, haplotypes of *GABRB3* are significantly associated with symptom severities of ASD when we remove SNP *rs1432007* from the analysis.

References

- Aldred S, Moore KM, Fitzgerald M, Waring RH. Plasma amino acid levels in children with autism and their families. *J Autism Dev Disord.* 2003; 33:93–7. [PubMed: 12708584]
- Ashley-Koch AE, Mei H, Jaworski J, Ma DQ, Ritchie MD, Menold MM, et al. An analysis paradigm for investigating multi-locus effects in complex disease: examination of three GABA receptor subunit genes on 15q11-q13 as risk factors for autistic disorder. *Ann Hum Genet.* 2006; 70(Pt 3): 281–292. [PubMed: 16674551]
- Belmonte MK, Cook EH, Anderson GM, Rubenstein JLR, Greenough WT, Beckel-Mitchener A, et al. Autism as a disorder of neural information processing: directions for research and targets for therapy. *Molecular Psychiatry.* 2004; 9(7):646–663. [PubMed: 15037868]
- Blatt GJ, Fitzgerald CM, Gupitill JT, Booker AB, Kemper TL, Bauman ML. Density and distribution of hippocampal neurotransmitter receptors in autism: an autoradiographic study. *J. Autism Dev. Disord.* 2001; 31:537–43. [PubMed: 11814263]
- Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and regression trees. Chapman and Hall/CRC; Monterey, CA: 1984.
- Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, et al. Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology.* 2005; 28(2):171–182. [PubMed: 15593090]
- Buxbaum JD, Silverman JM, Smith CJ, Greenberg DA, Kilifarski M, Reichert J, et al. Association between a *GABRB3* polymorphism and autism. *Molecular Psychiatry.* 2002; 7(3):311–316. [PubMed: 11920158]

- Cheng L, Ge Q, Xiao P, Sun B, Ke X, Bai Y, et al. Association Study between BDNF Gene Polymorphisms and Autism by Three-Dimensional Gel-Based Microarray. *Int J Mol Sci*. 2009; 10(6):2487–2500. [PubMed: 19582215]
- Collins AL, Ma DQ, Whitehead PL, Martin ER, Wright HH, Abramson RK, et al. Investigation of autism and GABA receptor subunit genes in multiple ethnic groups. *Neurogenetics*. 2006; 7(3):167–174. [PubMed: 16770606]
- Cook EH Jr. Courchesne RY, Cox NJ, Lord C, Gonen D, Guter SJ, et al. Linkage-disequilibrium mapping of autistic disorder, with 15q11-13 markers. *Am J Hum Genet*. 1998; 62(5):1077–1083. [PubMed: 9545402]
- Delahanty RJ, Kang JQ, Brune CW, Kistner EO, Courchesne E, Cox NJ, et al. Maternal transmission of a rare GABRB3 signal peptide variant is associated with autism. *Molecular Psychiatry*. 2011; 16(1):86–96. [PubMed: 19935738]
- Fatemi SH, Reutiman TJ, Folsom TD, Thuras PD. GABA(A) Receptor Downregulation in Brains of Subjects with Autism. *Journal of Autism and Developmental Disorders*. 2009; 39(2):223–230. [PubMed: 18821008]
- Freitag CM. The genetics of autistic disorders and its clinical relevance: a review of the literature. *Mol Psychiatry*. 2007; 12(1):2–22. [PubMed: 17033636]
- Freitag CM, Staal W, Klauck SM, Duketis E, Waltes R. Genetics of autistic disorders: review and clinical implications. *European Child & Adolescent Psychiatry*. 2010; 19(3):169–178. [PubMed: 19941018]
- Freund, Y.; Mason, L. Proceedings of the Sixteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc.; 1999. The Alternating Decision Tree Learning Algorithm; p. 124-133.
- Geschwind DH. Advances in Autism. *Annu Rev Med*. 2009; 60:367–380. [PubMed: 19630577]
- Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu FL, Yang HM, et al. The International HapMap Project. *Nature*. 2003; 426(6968):789–796. [PubMed: 14685227]
- Gmitrowicz A, Kucharska A. Developmental disorders in the fourth edition of the American classification: diagnostic and statistical manual of mental disorders (DSM IV -- optional book). *Psychiatr Pol*. 1994; 28(5):509–521. [PubMed: 7527563]
- Holmes, G.; Pfahringer, B.; Kirkby, R.; Frank, E.; Hall, M. Proceedings of the 13th European Conference on Machine Learning. Springer-Verlag; 2002. Multiclass Alternating Decision Trees; p. 161-172.
- Hou P, Ji M, Li S, Lu Z. Microarray-based approach for high-throughput genotyping of single-nucleotide polymorphisms with layer-by-layer dual-color fluorescence hybridization. *Clin Chem*. 2004; 50(10):1955–1957. [PubMed: 15484337]
- Huang J, Lin A, Narasimhan B, Quertermous T, Hsiung CA, Ho LT, et al. Tree-structured supervised learning and the genetics of hypertension. *Proc Natl Acad Sci U S A*. 2004; 101(29):10529–10534. [PubMed: 15249660]
- Iba, W.; Langley, P. Proceedings of the ninth international workshop on Machine learning. United Kingdom Morgan Kaufmann Publishers Inc.; Aberdeen, Scotland: 1992. Induction of one-level decision trees; p. 233-240.
- Ji M, Hou P, Li S, He N, Lu Z. Microarray-based method for genotyping of functional single nucleotide polymorphisms using dual-color fluorescence hybridization. *Mutat Res*. 2004; 548(1-2):97–105. [PubMed: 15063140]
- Kim SA, Kim JH, Park M, Cho IH, Yoo HJ. Association of GABRB3 polymorphisms with autism spectrum disorders in Korean trios. *Neuropsychobiology*. 2006; 54(3):160–165. [PubMed: 17230033]
- Kim SJ, Brune CW, Kistner EO, Christian SL, Courchesne EH, Cox NJ, et al. Transmission disequilibrium testing of the chromosome 15q11-q13 region in autism. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. 2008; 147B(7):1116–1125.
- Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L. Sequence analysis using logic regression. *Genetic Epidemiology*. 2001; 21:S626–S631. [PubMed: 11793751]

- Lerer E, Levi S, Salomon S, Darvasi A, Yirmiya N, Ebstein RP. Association between the oxytocin receptor (OXTR) gene and autism: relationship to Vineland Adaptive Behavior Scales and cognition. *Molecular Psychiatry*. 2008; 13(10):980–988. [PubMed: 17893705]
- Li H, Yamagata T, Mori M, Momoi MY. Absence of causative mutations and presence of autism-related allele in FOXP2 in Japanese autistic patients. *Brain & Development*. 2005; 27(3):207–210. [PubMed: 15737702]
- Li Z, Zhang Z, He Z, Tang W, Li T, Zeng Z, et al. A partition-ligation-combination-subdivision EM algorithm for haplotype inference with multiallelic markers: update of the SHESis (<http://analysis.bio-x.cn>). *Cell Res*. 2009; 19(4):519–523. <http://analysis.bio-x.cn> [PubMed: 19290020]
- Loh WY, Shih YS. Split selection methods for classification trees. *Statistica Sinica*. 1997; 7(4):815–840.
- Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord*. 1994; 24(5):659–685. [PubMed: 7814313]
- Ma DQ, Whitehead PL, Menold MM, Martin ER, Ashley-Koch AE, Mei H, et al. Identification of significant association and gene-gene interaction of GABA receptor subunit genes in autism. *American Journal of Human Genetics*. 2005; 77(3):377–388. [PubMed: 16080114]
- McCauley, JL. Genetic and phenotypic dissection of autism susceptibility. 2005.
- McCauley JL, Olson LM, Delahanty R, Amin T, Nurmi EL, Organ EL, et al. A linkage disequilibrium map of the 1-Mb 15q12 GABA(A) receptor subunit cluster and association to autism. *Am J Med Genet B Neuropsychiatr Genet*. 2004; 131B(1):51–59. [PubMed: 15389768]
- Miller, RG. Simultaneous Statistical Inference. Second Edition ed.. SPRINGER-VERLAG INC.; NY: 1981.
- Moreno-Fuenmayor H, Borjas L, Arrieta A, Valera V, Socorro- Candanoza L. Plasma excitatory amino acids in autism. *Investigacion Clinica*. 1996; 37:113–28. 1996. [PubMed: 8718922]
- Nabi R, Serajee FJ, Chugani DC, Zhong H, Huq AHMM. Association of tryptophan 2,3 dioxygenase gene polymorphism with autism. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. 2004; 125B(1):63–68.
- Nunkesser R, Bernholt T, Schwender H, Ickstadt K, Wegener I. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics*. 2007; 23(24):3280–3288. [PubMed: 18006552]
- Nurmi EL, Amin T, Olson LM, Jacobs MM, McCauley JL, Lam AY, et al. Dense linkage disequilibrium mapping in the 15q11-q13 maternal expression domain yields evidence for association in autism. *Molecular Psychiatry*. 2003; 8(6):624–634. [PubMed: 12851639]
- Nurmi EL, Dowd M, Tadevosyan-Leyfer O, Haines JL, Folstein SE, Sutcliffe JS. Exploratory subsetting of autism families based on savant skills improves evidence of genetic linkage to 15q11-q13. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2003; 42(7):856–863. [PubMed: 12819446]
- Olsen, RW.; Macdonald, RL. GABAA receptor complex: Structure and function, in Glutamate and GABA receptors and Transporters. Egebjerg, J.; Schousboe, A.; Krosgaard-Larsen, P., editors. Taylor & Francis; London: 2002. p. 202-235.
- Ramos N, Reichert JG, Smith CJ, Silverman JM, Beshalova IN, Davis KL, Buxbaum J. Linkage and association of the mitochondrial aspartate/glutamate carrier SLC25A12 gene with autism. *Am J Psychiatry*. 2004; 161:662–669. [PubMed: 15056512]
- Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. *Biostatistics*. 2008; 9(1):30–50. [PubMed: 17429103]
- Parks LK, Hill DE, Thoma RJ, Euler MJ, Lewine JD, Yeo RA. Neural correlates of communication skill and symptom severity in autism: A voxel-based morphometry study. *Research in Autism Spectrum Disorders*. 2009; 3(2):444–454.
- Rapin I. Autism. *New England Journal of Medicine*. 1997; 337(2):97–104. [PubMed: 9211680]
- Samaco RC, Hogart A, LaSalle JM. Epigenetic overlap in autism-spectrum neurodevelopmental disorders: MECP2 deficiency causes reduced expression of UBE3A and GABRB3. *Human Molecular Genetics*. 2005; 14(4):483–492. [PubMed: 15615769]

- Samaco RC, Nagarajan RP, Braunschweig D, LaSalle JM. Multiple pathways regulate MeCP2 expression in normal brain development and exhibit defects in autism-spectrum disorders. *Human Molecular Genetics*. 2004; 13(6):629–639. [PubMed: 14734626]
- Schopler E, Reichler RJ, DeVellis RF, Daly K. Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). *J Autism Dev Disord*. 1980; 10(1):91–103. [PubMed: 6927682]
- Schwender H, Ickstadt K, Rahnenfuhrer J. Classification with High-Dimensional Genetic Data: Assigning Patients and Genetic Features to Known Classes. *Biometrical Journal*. 2008; 50(6):911–926. [PubMed: 19067340]
- Segurado R, Conroy J, Meally E, Fitzgerald M, Gill M, Gallagher L. Confirmation of association between autism and the mitochondrial aspartate/glutamate carrier SLC25A12 gene on chromosome 2q31. *Am J Psychiatry*. 2005; 162:2182–2184. [PubMed: 16263864]
- Shi YY, He L. SHEsis, a powerful software platform for analyses of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci. *Cell Res*. 2005; 15(2):97–98. [PubMed: 15740637]
- Sparrow SS, Cicchetti DV. Diagnostic Uses of the Vineland Adaptive-Behavior Scales. *Journal of Pediatric Psychology*. 1985; 10(2):215–225. [PubMed: 4020603]
- Sutcliffe JS, Nurmi EL. Genetics of childhood disorders: XLVII. Autism, Part 6: Duplication and inherited susceptibility of chromosome 15q11-q13 genes in autism. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2003; 42(2):253–256. [PubMed: 12544187]
- Wermter AK, Kamp-Becker I, Hesse P, Schulte-Körne G, Strauch K, Remschmidt H. Evidence for the Involvement of Genetic Variation in the Oxytocin Receptor Gene (OXTR) in the Etiology of Autistic Disorders on High-Functioning Level. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*. 2010; 153B(2):629–639.
- Witten, IH.; Frank, E. *Data Mining: Practical machine learning tools and techniques*. 2nd ed.. Morgan Kaufmann; San Francisco: 2005. 2005
- Xiao PF, Cheng L, Wan Y, Sun BL, Chen ZZ, Zhang SY, et al. An improved gel-based DNA microarray method for detecting single nucleotide mismatch. *Electrophoresis*. 2006; 27(19):3904–3915. [PubMed: 16960844]

top

Figure 1.
Three genotypes of rs878960 and CARS
Higher CARS score reflects greater symptom severity.

top

Figure 2.

Comparison of population diversity with respect to rs878960 in our study, with that in other Asian populations.

*: data were acquired from the International HapMap Project (Gibbs et al., 2003).

top

Figure 3.
Histogram of CARS scores

top

Figure 4.

Tree model generated by ADTree

A negative number indicates that a subject was classified as belonging to the mild/moderate group, whereas a positive number indicates that a subject was classified as belonging to the severe group. The *rs878960* SNP was selected as the priority root node.

Table 1

Subjects' demographic and behavior variables

Variable	Mild/moderate	Severe
Number of subjects	65	53
Gender (male:female)	57:8	46:7
Age $\pm$ SD (yrs)	5.9 $\pm$ 2.1	5.0 $\pm$ 2.3
ADI-R Social $\pm$ SD	17.7 $\pm$ 6.1	25.2 $\pm$ 5.6
ADI-R communication $\pm$ SD	14.4 $\pm$ 5.1	17.9 $\pm$ 4.5
ADI-R behavior $\pm$ SD	6.3 $\pm$ 3.5	7.5 $\pm$ 3.2
CARS $\pm$ SD	31.4 $\pm$ 3.2	39.6 $\pm$ 2.1

Table 2

Details of 29 SNPs in 9 ASD-related genes

Gene	SNP ID (Reference)	Allele (wild/mutant)	Chromosome	Region
<i>GABRA4</i>	<i>rs1912960</i> (Ma et al., 2004)	G/C	4	intron8
	<i>rs2280073</i> (Ma et al., 2004)	G/C	4	intron7
	<i>rs17599165</i> (Collins et al., 2006)	T/A	4	intron8
	<i>rs17599416</i> (Collins et al., 2006)	G/A	4	intron6
	<i>rs7660336</i> (Collins et al., 2006)	G/C	4	exon9
<i>GABRB1</i>	<i>rs2351299</i> (Ma et al., 2004)	G/T	4	intron3
	<i>rs3832300</i> (Collins et al., 2006)	-/T	4	exon9
<i>TDO2</i>	<i>rs3755908</i> (Nabi et al., 2004)	T/C	4	promoter
	<i>rs3775085</i> (Nabi et al., 2004)	A/C	4	promoter
	<i>rs3755910</i> (Nabi et al., 2004)	C/A	4	promoter
	<i>rs2292537</i> (Nabi et al., 2004)	T/C	4	intron11
<i>GABRB2</i>	<i>rs2617503</i> (Ma et al., 2004)	A/G	5	intron6
	<i>rs12187676</i> (Ma et al., 2004)	C/G	5	intron4
<i>GABRA2</i>	<i>hcv8262334</i> (Ma et al., 2004)	T/A	4	Intron
<i>GABRB3</i>	<i>rs1432007</i> (McCauley et al. 2004)	A/G	15	intron7
	<i>rs2873027</i> (McCauley et al. 2004)	G/A	15	intron3
	<i>rs4542636</i> (McCauley et al. 2004)	C/T	15	intron3
	<i>rs878960</i> (McCauley et al. 2004)	A/G	15	intron3
	<i>rs2081648</i> (McCauley et al. 2004)	A/G	15	intron8
	<i>hcv2911914</i> (McCauley et al. 2004)	C/G	15	intergenic
<i>GABRA5</i>	<i>hcv11298361</i> (McCauley et al. 2004)	A/G	15	intron
	<i>hcv252720</i> (McCauley et al. 2004)	C/T	15	intron6
	<i>hcv27725</i> (McCauley et al. 2004)	A/C	15	intron6
<i>SLC25A12</i>	<i>rs2056202</i> (Freitag 2007)	T/C	2	intron3
	<i>rs2292813</i> (Freitag 2007)	T/C	2	intron16
<i>BDNF</i>	<i>rs6265</i> (Cheng et al., 2008)	A/G	11	exon2
	<i>rs988748</i> (Cheng et al., 2008)	C/G	11	intron1
	<i>rs2049046</i> (Cheng et al., 2008)	A/T	11	intron1
	<i>C270T</i> (Cheng et al., 2008)	C/T	11	5'UTR

Table 3

Genotypes and allele distributions for rs878960 polymorphisms in the severe and mild/moderate groups.

<i>rs878960</i>		Severe	Mild/moderate
Genotype (frequency)	A/A	22 (0.41)	8 (0.12)
	A/G	19 (0.36)	41 (0.63)
	G/G	28 (0.23)	16 (0.25)
p-value	0.0009		
Allele (frequency)	A	63 (0.59)	57 (0.44)
	G	43 (0.41)	73 (0.56)
p-value	0.017		
Odds Ratio	1.88		
95%CI	1.12 - 3.16		

Table 4

Genotypes and allele distributions for GABRB3 polymorphisms in the severe and mild/moderate groups

<i>GABRB3</i> SNP	<i>p</i> -value for allele	<i>p</i> -value for genotype
<i>rs1432007</i>	0.54	0.82
<i>rs2873027</i>	0.75	0.70
<i>rs4542636</i>	0.94	0.38
<i>rs878960</i>	0.017	0.0009
<i>rs2081648</i>	0.77	0.55
<i>hcv2911914</i>	0.25	0.09

Table 5

Genotypes and allele distributions for rs878960 polymorphisms in our study and for other Asian groups

		Our Study	HapMap Han
Genotype (frequency)	A/A	30 (0.25)	6 (0.14)
	A/G	60 (0.51)	25 (0.58)
	G/G	28 (0.24)	12 (0.28)
<i>p</i> -value	0.30		
Allele (frequency)	A	120 (0.51)	37 (0.43)
	G	116 (0.49)	49 (0.57)
<i>p</i> -value	0.21		
Odds Ratio	1.37		
95%CI	0.83~.25		

		Our Study	HapMap Han
Genotype (frequency)	A/A	30 (0.25)	18 (0.21)
	A/G	60 (0.51)	36 (0.42)
	G/G	28 (0.24)	32 (0.37)
<i>p</i> -value	0.11		
Allele (frequency)	A	120 (0.51)	72 (0.42)
	G	116 (0.49)	100 (0.58)
<i>p</i> -value	0.07		
Odds Ratio	1.44		
95%CI	0.98~2.14		

Table 6

Haplotype analysis for each gene. Haplotypes with frequency < 0.05 in both groups have been dropped.

Gene	$\chi^2 (p)$	Haplotype	S (freq)	M (freq)	χ^2	p-value	Odds Ratio [95%CI]
<i>GABRA4</i>	12.4 (0.015)	G G T A G	0.032	0.067	0.97	0.33	0.51 [0.13~2.01]
		G G T A C	0.341	0.191	7.94	0.005	2.55 [1.32~4.93]
		G C T A G	0.275	0.259	0.39	0.53	1.23 [0.65~2.33]
		C G A G C	0.163	0.357	7.80	0.005	0.38 [0.19~0.76]
		C G A A C	0.053	0.067	0.07	0.80	0.86 [0.26~2.79]
<i>GABRB1</i>	0.58 (0.45)	G –(not T)	0.712	0.632	0.58	0.44	1.26 [0.70~2.26]
		T –(not T)	0.260	0.290	0.58	0.44	0.80 [0.44~1.43]
<i>GABRB2</i>	3.5 (0.33)	A C	0.267	0.279	0.039	0.84	0.941 [0.51~1.72]
		A G	0.293	0.212	1.851	0.17	1.541 [0.83~2.88]
		G C	0.373	0.471	2.070	0.15	0.669 [0.39~1.16]
		G G	0.067	0.038	0.903	0.34	1.817 [0.52~6.33]
<i>GABRB3</i>	7.6 (0.10)	AGCAGC	0.043	0.073	0.911	0.34	0.572 [0.18~1.82]
		AGCAGG	0.399	0.260	6.230	0.012	2.180 [1.18~4.04]
		AGCGAG	0.020	0.051	1.589	0.21	0.374 [0.08~1.82]
		AGCGGG	0.225	0.264	0.543	0.46	0.786 [0.41~1.49]
		GATGAC	0.042	0.082	1.574	0.21	0.485 [0.15~1.53]
<i>GABRA5</i>	9.7 (0.097)	A T A	0.178	0.152	0.376	0.54	1.243 [0.62~2.50]
		A T C	0.026	0.085	3.500	0.061	0.294 [0.08~1.14]
		G C A	0.279	0.212	1.714	0.19	1.496 [0.82~2.74]
		G C C	0.107	0.187	2.648	0.10	0.534 [0.25~1.15]
		G T A	0.107	0.159	1.180	0.28	0.650 [0.30~1.42]
<i>TDO2</i>	1.9 (0.17)	G T C	0.256	0.181	2.231	0.14	1.614 [0.86~3.03]
		T A C T	0.826	0.929	1.879	0.17	0.481 [0.17~1.40]
		C C C T	0.087	0.047	1.879	0.17	2.078 [0.72~6.03]
<i>SLC25A12</i>	0.33 (0.85)	T T	0.065	0.059	0.022	0.88	1.084 [0.38~3.13]
		T C	0.048	0.064	0.321	0.57	0.721 [0.23~2.25]
		C C	0.867	0.836	0.087	0.77	1.127 [0.51~2.51]

Gene	χ^2 (p)	Haplotype	S (freq)	M (freq)	χ^2	p-value	Odds Ratio [95%CI]
BDNF	2.4 (0.49)	A G A C	0.409	0.366	0.690	0.41	1.264 [0.73~2.20]
		G C A C	0.019	0.054	1.932	0.16	0.337 [0.07~1.68]
		G C T C	0.355	0.389	0.198	0.66	0.882 [0.51~1.54]
		G C T T	0.087	0.077	0.099	0.75	1.162 [0.45~2.98]

S: Severe, M: Mild/moderate; Haplotypes are organized as follows: GABRA4: rs1912960 - rs2280073 - rs17599165 - rs17599416 - rs7660336; GABRB1: rs2351299 - rs3832300; GABRB2: rs2617503 - rs12187676; GABRB3: rs1432007 - rs2873027 - rs4542636 - rs78960 - rs2081648 - rs2911914; GABRA5: HCV11298361 - HCV252720 - HCV27725; TDO2: rs3755908 - rs3775085 - rs3755910 - rs2292537; SLC25A12: rs2056202 - rs2292813; BDNF: rs6265 - rs988748 - rs2049046 - C270T.

Table 7

Haplotype analysis of GABRB3 after removing rs1432007. Haplotypes with frequency < 0.05 in both groups have been dropped.

Haplotype	S (freq)	M (freq)	χ^2	p-value	Odds Ratio [95%CI]
G C A A C	0.031	0.053	0.746	0.39	0.556 [0.144~.143]
G C A G C	0.044	0.071	0.806	0.37	0.593 [0.187~1.877]
G C A G G	0.399	0.263	4.791	0.029	1.899 [1.066~3.381]
G C G A G	0.017	0.099	6.893	0.009	0.155 [0.032~0.749]
G C G G G	0.225	0.253	0.370	0.54	0.825 [0.443~1.535]
A T G A C	0.041	0.077	1.408	0.24	0.501 [0.157~1.599]
A T G A G	0.059	0.035	0.706	0.40	1.699 [0.488~5.918]
G C A A G	0.052	0.000	6.677	0.010	-

S: Severe, Mild/moderate; Haplotypes were organized: r2873027 - r4542636 - r878960 - rs2081648 - HCV2911914.