Prediction of serotonin transporter promoter polymorphism genotypes from single nucleotide polymorphism arrays using machine learning methods

Ake Tzu-Hui Lu[a], Steven Bakker[c], Esther Janson[d], Sven Cichon[e,f], Rita M. Cantor[a,b] and Roel A. Ophoff[b,d]

Background **The serotonin transporter gene (*SLC6A4*) and its promoter (*5-HTTLPR*) polymorphism have been the focus of a large number of association studies of behavioral traits and psychiatric disorders. However, large-scale genotyping of the polymorphism has been very difficult. We report the development and validation of a *5-HTTLPR* genotype prediction model.**

Methods **The single nucleotide polymorphisms (SNPs) from the 2000 kb region surrounding *5-HTTLPR* were used to construct a prediction model through a newly developed machine learning method, multicategory vertex discriminant analysis with 2147 individuals from the Northern Finnish Birth Cohort genotyped with the Illumina 370K SNP array and manually genotyped for *5-HTTLPR* polymorphism. The prediction model was applied to SNP genotypes in a Dutch/German schizophrenia case–control sample of 3318 individuals to test the association of the polymorphism with schizophrenia.**

Result **The prediction model of eight SNPs achieved a 92.4% accuracy rate and a 0.98±0.01 area under the receiving operating characteristic. Evidence for an association of the polymorphism with schizophrenia was observed ($P=0.05$, odds ratio = 1.105).**

Conclusion **This prediction model provides an effective substitute of manually genotyped *5-HTTLPR* alleles, providing a new approach for large scale association studies of this polymorphism. *Psychiatr Genet* 22:182–188 © 2012 Wolters Kluwer Health | Lippincott Williams & Wilkins.**

Psychiatric Genetics 2012, **22**:182–188

Keywords: genotype prediction, *5-HTTLPR* polymorphism, machine learning methods, schizophrenia, serotonin transporter gene

[a]Department of Human Genetics, David Geffen School of Medicine at UCLA, [b]Center for Neurobehavioral Genetics, UCLA, Los Angeles, California, USA, [c]Department of Psychiatry, [d]Department of Medical Genetics, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, The Netherlands, [e]Institute of Human Genetics and [f]Department of Genomics, Life and Brain Center, University of Bonn, Bonn, Germany

Correspondence to Rita M. Cantor, PhD, Department of Human Genetics, David Geffen School of Medicine at UCLA, 695 Charles E. Young Dr South, Los Angeles, CA 90095 7088, USA
Tel: +1 310 267 2440; fax: +1 310 794-5446;
e-mail: rcantor@mednet.ucla.edu

Received 15 April 2011 **Revised** 9 November 2011
Accepted 28 December 2011

Introduction

The serotonin transporter (*5-HTT* or *SLC6A4*) is probably the most frequently investigated gene in association studies of psychiatric disorders (Caspi *et al.*, 2010). Its gene product mediates the reuptake of monoamine serotonin (5-HT), a key neurotransmitter in the brain. Many effective antidepressant drugs selectively inhibit *5-HTT* function (Pacheco *et al.*, 2009). Under certain physiologic conditions, the expression of *5-HTT* is modulated by genetic variants, and of these, the most frequently studied is a 43-base pair insertion/deletion polymorphism in the promoter region, where *5-HTTLPR* has a long (L) and a short (S) allele. There are other *5-HTTLPR* polymorphisms that are also good candidates for association testing with psychiatric disorders. For example, we now know there is a triallelic variation (S, L_A, L_G) within this gene that is the result of an A to G single nucleotide polymorphism (SNP) that splits the L allele. This was identified by Hu *et al.*, 2006, and it has been shown to alter expression levels, and can make association studies with traits and disorders more precise.

A number of studies have implicated *5-HTTLPR* genotypes in normal behavior traits (Lesch *et al.*, 1996) and psychiatric disorders (Lin and Tsai, 2004; Lopez-Leon *et al.*, 2008; Grabe *et al.*, 2009). This variant is postulated to modulate the effects of stress on the development of psychiatric illnesses (Caspi *et al.*, 2003). However, a recent meta-analysis failed to establish a genetic association of psychiatric illness (Risch *et al.*, 2009) with this polymorphism. Thus, similar to many reported associations in complex disorders, the results of *5-HTTLPR* studies have been inconsistent, warranting further studies in larger samples for resolution.

Unfortunately, *5-HTTLPR* genotypes are not present on available SNP arrays. In addition, genotyping of *5-HTTLPR* in large samples is only marginally feasible for technical reasons. The polymorphism is located in a highly repetitive and GC-rich DNA region that negatively affects the efficiency of PCR amplification and possibly results in the preferential amplification of the smaller allele. That is, the relative amplification of the L and S alleles of *5-HTTLPR* has been shown to be dependent on

0955-8829 © 2012 Wolters Kluwer Health | Lippincott Williams & Wilkins DOI: 10.1097/YPG.0b013e328353ae23

the Mg^+ concentration, and several groups have reported genotyping errors biased toward the S alleles (Sen *et al.*, 2004; Yonan *et al.*, 2006; Wray *et al.*, 2009).

If *5-HTTLPR* could be investigated in the large study samples, unresolved questions about its role in behavioral traits and psychiatric disorders could be addressed. SNPs in genome-wide arrays have been selected for their ability to 'tag' haplotypes of multiple other SNPs (De Bakker *et al.*, 2005), and we postulated that it was possible that SNPs surrounding *5-HTTLPR* may tag this tandem repeat polymorphism in a similar manner. Recently, Wray *et al.* (2009) investigated a series of SNPs surrounding *5-HTTLPR*, and found that a 2-marker SNP haplotype predicted *5-HTTLPR* with an R^2 of 0.7. Unfortunately, an SNP that is not present in standard genome-wide association studies (GWAS) arrays or in the comprehensive HapMap dataset is central to its prediction, requiring additional genotyping of study samples in most cases. The current study was designed to identify a set of SNPs present on commonly used Illumina GWAS arrays that predict *5-HTTLPR* genotypes with substantial sensitivity and specificity and without additional genotyping. Using a newly developed machine learning method, we were able to reconstruct and validate *5-HTTLPR* genotypes in European Whites with a model based on eight SNPs. Once validated, we applied the model to a schizophrenia sample of 3318 to assess its association with *5-HTTLPR*.

Materials and methods

Generating the prediction model

Samples and genotypes

Two study samples with manually generated *5-HTTLPR* genotypes based on biallelic variation (S, L) and array-based SNP genotypes were used to develop and test a prediction model for *5-HTTLPR*. The first includes 2147 normal participants from the 1966 Northern Finnish Birth Cohort (Sabatti *et al.*, 2009) genotyped with the Illumina 370K Infinium BeadChip (Illumina, San Diego, California, USA), referred to hereafter as 'Finn'. The second includes 276 Dutch study participants, 126 normal and 150 diagnosed with schizophrenia, genotyped with the Illumina HumanHap550 BeadChip (Illumina), referred to hereafter as 'Dutch1'. The *5-HTTLPR* polymorphism was genotyped manually in both samples. The first sample was used to generate and evaluate the prediction model and the second was used to further validate the model.

Genotyping of Finn is described in Munafo *et al.* (2009); the Dutch participants were genotyped using primers (5′–3′) GGCGTTGCCGCTCTGAATGC and GAGGGACTGAGCTGGACAACCAC and PCR amplification in 20 µl volumes, containing 25 ng of genomic DNA, 0.25 µmol/l of each primer, using AccuPrime GC-Rich DNA Polymerase (Invitrogen, Grand Island, New York, USA). The PCR program was as follows: 95°C (3′); 33 × [95°C (30″); 65–54°C (30″); 72°C (1′)]; and then 72°C (10′), followed by 4°C (∞). A measure of 10 µl PCR product was size-separated on a 2% agarose gel. Scoring was performed by two independent raters (S.B. and E.J.).

Selection of single nucleotide polymorphism predictors

SNPs found between 24 523 266 and 26 462 684 bp on 17q11.2 were used to construct a prediction model for *5-HTTLPR* (Genome build hg18). The criteria for inclusion in the model building panel were a minor allele frequency more than 0.05, Hardy–Weinberg equilibrium (HWE) $P > 0.05$, and missing genotypes less than 0.01. A 0.8 pairwise R^2 threshold was used to remove redundant SNPs. The 77 SNPs fulfilling these criteria were available for the prediction of the *5-HTTLPR* genotypes. Stepwise linear regression implemented under SAS 9.1 (SAS Institute Inc., Cary, North Carolina, USA) PROC REG was used with the significance levels of entry (SLENTRY) and staying (SLSTAY) set at 1.0E-09 to substantially reduce the number of SNPs, to make this model accessible.

The prediction model

Machine learning methods, such as the support vector machine (SVM) (Vapnik, 1995 and 1998), have recently been successfully applied to solve classification problems such as this one (Capriotti *et al.*, 2006; Kong and Choo, 2007; Zhou and Wang, 2007; Lin and Hwang, 2008; Liu *et al.*, 2008; Zeller *et al.*, 2008). Here, the individuals are classified into their *5-HTTLPR* genotypes. A newly developed multicategory machine learning method of vertex discriminant analysis (VDA) (*http://www.amstat.org/publications/jcgs/*) (Lange and Wu, 2008) was used to predict the three *5-HTTLPR* genotypes S/S, S/L, and L/L, capitalizing on the partial linkage disequilibrium with surrounding SNPs. This analytic method was selected because a study by its developers showed that it performs better in multicategory prediction than a number of other methods, and an additional strength is that it allows for a nonlinear relationship between the genotypes and the predictors, providing greater flexibility in the prediction model.

The VDA approach is described here. A learning model is constructed using a training dataset and evaluated with a test dataset. For k category classifications (here k is 3), VDA constructs k equidistant points in the R^{k-1} space to assign the coordinates of the response variable (the predicted *5-HTTLPR* genotypes, L/L, L/S, and S/S) and is denoted below by y. The learning model is searched numerically to optimize a loss function with two terms:

$$Loss\,(A,\,b)=\frac{1}{n}\sum_{i=1}^{n}\parallel y_i-A^T x_i+b\parallel_{\varepsilon}+\lambda\sum_{j=1}^{k-1}\parallel a_j\parallel^2,$$

where $(A,\,b)$ is a $p\times(k-1)$ matrix of regression coefficient used for genotype prediction with the jth column denoted by a_j and a $(k-1)$ by 1 vector of intercepts, n is the number of observations, where they

are indexed by i, and (x, y) is a vector of p predictors (here there are eight SNPs for x) and their response values (the *5-HTTLPR* genotypes for y). The first term represents the distance between y and the fitted response with a Euclidean distance that is insensitive to ϵ and robust to outliers. The second term represents the ridge penalty, which penalizes irrelevant predictors with a positive tuning parameter λ. This is included to handle a large number of predictors, although they have been limited to eight here. The technical strength of this approach is efficient optimization, where the majorization–minimization algorithm is used to minimize this loss function with fixed (ϵ, λ) with iterative optimization (De Leeuw and Heiser, 1977).

For these analyses, the prediction models were generated using Finn. Here, we present data preparation details for how the models were generated. Before analysis and for each individual, their eight predictor SNPs were coded as an 8×1 vector of 0, 1, or 2, representing the number of minor alleles that person has at that SNP. All predictors were standardized on the basis of the number of minor alleles a person has at a particular SNP in the following way. For each individual in the sample and for a given SNP, the number of minor alleles was recorded and the mean and variance of those data were estimated. The standardized value for 0 minor alleles was obtained by subtracting the mean from 0 and dividing by the SD. For one minor allele, 1 is used in the place of 0, and for two minor alleles, 2 is used in the place of zero. The 8×2 matrix of coefficients A and the 2×1 vector of intercepts b were estimated in the model, which is then used to

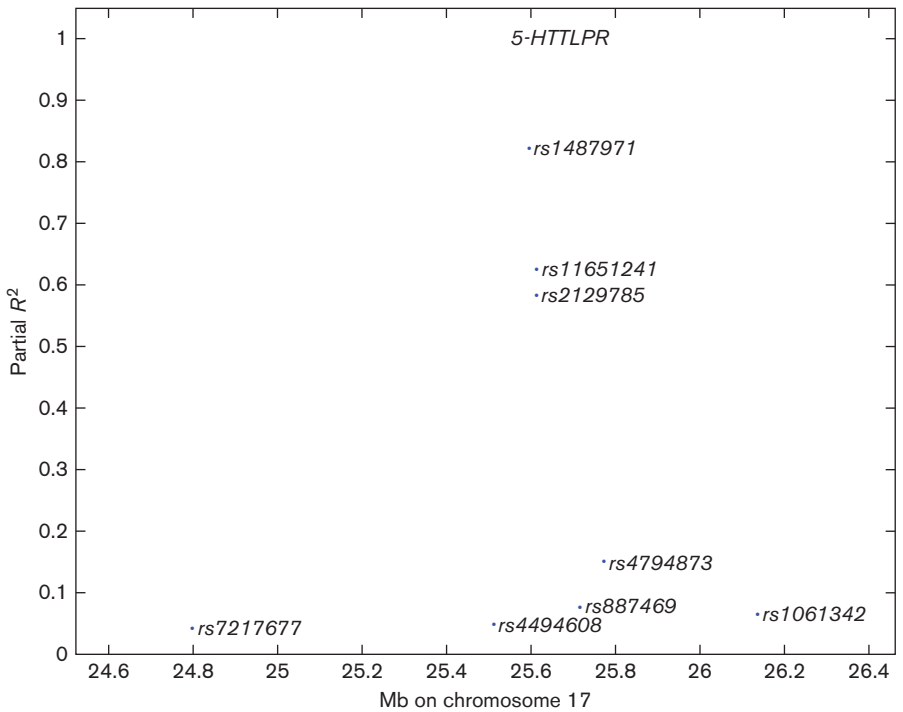
Table 1 Distributions of observed *5-HTTLPR* genotypes (%) in Finn and Dutch1

	Finn ($n=2147$)	Dutch1 ($n=276$)
S/S	359 (17)	59 (21)
S/L	1036 (48)	135 (49)
L/L	752 (35)	82 (30)
Frequency of S	0.41	0.46

Table 2 Single nucleotide polymorphisms predicting *5-HTTLPR* genotypes in order of their contributions to the stepwise regression model in the Finn training sample ($n=1852$)

Reference single nucleotide polymorphism ID	Position in base pairs Genome build hg18	Minor allele frequency	Cumulative R^2
rs1487971	25 596 879	0.40	0.341
rs2129785	25 614 656	0.10	0.523
rs11651241	25 613 604	0.10	0.811
rs4794873	25 772 478	0.15	0.821
rs887469	25 716 700	0.08	0.830
rs1061342	26 138 589	0.10	0.837
rs4494608	25 512 917	0.15	0.844
rs7217677	24 799 793	0.17	0.850

Fig. 1



Locations in Mb and importance (partial R^2 values of stepwise regression) of eight single nucleotide polymorphisms predicting *5-HTTLPR* genotypes.

Table 3 *5-HTTLPR* genotype prediction model for eight single nucleotide polymorphisms

		Covariate (x_i)[b] value selected by number of minor alleles for that SNP			Prediction model[c]	
Predictors[a] (SNPs are indexed by i)	i	x_i if 0	x_i if 1	x_i if 0	a_{i1}	a_{i2}
rs7217677	1	−0.6411	1.2467	3.1346	−0.053427738	−0.017944884
rs4494608	2	−0.5930	1.3751	3.3432	0.053488163	0.010513604
rs1487971	3	−1.1437	0.2732	1.6900	−0.388960505	−0.081862677
rs11651241	4	−0.4700	1.8637	4.1974	−0.212066944	−0.053959400
rs2129785	5	−0.4750	1.9214	4.3178	−0.229855845	−0.048378268
rs887469	6	−0.4288	2.1924	4.8136	−0.073285608	−0.024940628
rs4794873	7	−0.6036	1.3645	3.3326	0.111159158	0.034747773
rs1061342	8	−0.4580	1.9055	4.2690	−0.059042871	−0.021787721
–	–	–	–	–	b_1	b_2
–	–	–	–	–	−0.048664235	−0.201663866

Each person will have eight $a_{i1}x$ terms + b_1 for value 1 and eight $a_{i2}x$ terms + b_2 for value 2.
Values 1 and 2 will be used to calculate the Euclidean distance from the vertex.
SNP, single nucleotide polymorphism.
[a]The SNPs are listed by their base-pair positions in increasing order.
[b]x_i is the standard value from the Finn training sample (n=1852).
[c]Coefficients of prediction models in $A^T x_i + b$, where A is (a_1, a_2) and b is (b_1, b_2).

predict the *5-HTTLPR* genotypes for each individual, similar to what is done to predict values by standard regression methods.

Evaluating the prediction model

The learning model was searched using 10-fold cross validation on the training dataset. Predictions were evaluated using an overall misclassification rate (OMR), the number of misclassified genotypes divided by a total number of those predicted. Accuracy was defined as 1–OMR. A binary misclassification rate (BMR) was also estimated for each of the three genotypes. The BMR, the rate for genotype S/S versus others, $BMR_{S/S}$, is defined as the sum of the observations with either S/S genotypes misclassified or non-S/S genotypes misclassified as S/S divided by the total number of observations. $BMR_{S/L}$ and $BMR_{L/L}$ are defined in an analogous manner. Performance of the learning model was evaluated by the area under the receiving operating characteristic curves (AROC) in addition to OMR and BMR. AROC provides a 0–1 diagnostic value to discriminate one class (here, that class is S allele carriers: S/S, S/L) from the other and reflects the relationship of the sensitivity and specificity of prediction. An AROC with a value of one indicates a perfect discrimination between classes (Bamber, 1975).

In addition, a family-based sample of 27 trios from HapMap CEU referred to as 'HapMap Trios' was used to further evaluate prediction by assessing the number of detectable Mendelian errors among the predicted genotypes in the trios.

Testing association with schizophrenia

5-HTTLPR genotypes were predicted in a combined ethnically homogeneous White schizophrenia case–control sample from the Netherlands and Germany to test for the association of this disorder, assuming an additive genetic effect. This sample is composed of 3318 individuals (2030 cases and 1288 controls). The first, referred to as 'Dutch', is 803 cases and 685 controls and includes Dutch1, used in developing the prediction model. The second, referred to as 'German', includes 485 cases and 1345 controls. Both samples were genotyped with the same Illumina HumanHap 550 K BeadChip. The Dutch and German Schizophrenia samples have been described previously (Stefansson *et al.*, 2009). In Dutch, one of the SNP predictors in the model was not completely genotyped, and Option 23 of the Mendel software package (Department of Human Genetics, David Geffen School of

Table 4 Distributions of predicted *5-HTTLPR* genotypes for manual genotypes in Finn and Dutch1 samples

Predicted manual	S/S	S/L	L/L	Total
Finn				
S/S	36	6	0	42
S/L	7	139	4	150
L/L	0	4	79	83
Dutch1				
S/S	53	5	1	59
S/L	10	117	7	134
L/L	1	5	73	79

Table 5 Misclassification analyses for predicted *5-HTTLPR* genotypes in training and test data sets

	Finn	Dutch1
Training data sample size (S/S, S/L, L/L)	1852 (314, 878, 660)	–
Test data sample size (S/S, S/L, L/L)	275 (42,150,83)	272 (59, 134,79)
OMR (%)	7.6	10.7
$BMR_{S/S}$ (%)	4.7	6.2
$BMR_{S/L}$ (%)	8.0	9.9
$BMR_{L/L}$ (%)	3.3	5.2
AROC (standard error) S/S, S/L versus L/L	0.98 (0.01)	0.94 (0.02)

AROC, area under the receiving operating characteristic curves; BMR, binary misclassification rate; OMR, overall misclassification rate.

Table 6 Predicted *5-HTTLPR* Genotypes in Dutch, German, and HapMap samples

	Genotype				
	S/S	S/L	L/L	HWE *P*-value	S Frequency
Dutch[a]	Number (%)				
Controls ($n=682$)	132 (19)	322 (47)	228 (33)	0.4	0.43
Schizophrenia ($n=802$)	161 (20)	395 (49)	246 (31)	0.9	0.45
German sample					
Controls ($n=1345$)	213 (16)	670 (50)	462 (34)	0.6	0.41
Schizophrenia ($n=485$)	93 (19)	231 (48)	161 (33)	0.3	0.43
HapMap CEU sample					
Founders ($n=54$)	14 (26)	23 (43)	17 (31)	0.4	0.47

HWE, Hardy–Weinberg Equilibrium.
[a]Four of 1488 missing predictors.

Medicine at UCLA, Los Angeles, California, USA) was used to impute the missing SNP genotypes from those that were genotyped (Ayers and Lange, 2008). The prediction model was applied to the combined Dutch and German to predict the S/S, S/L, and L/L genotypes. The association of the predicted genotypes with schizophrenia was tested using the Cochran Armitage trend analysis test programmed in the PLINK software package (*http://pngu.mgh.harvard.edu/~purcell/plink/dataman.shtml#extract*) (Purcell *et al.*, 2007).

Results

Table 1 presents the distributions of the observed *5-HTTLPR* genotypes and their allele frequencies in the Finn and Dutch1 samples. HWE was not rejected in either sample ($P > 0.05$), indicating that this hallmark of genotyping error was not violated by the predicted genotypes. Stepwise linear regression was applied to select SNPs that were used as input for the learning models predicting *5-HTTLPR* genotypes in Finn. Table 2 presents the reference sequence number of the eight selected SNPs, their base-pair locations, minor allele frequencies, and the cumulative R^2 for these eight predictors. They explain 85% of the variance of *5-HTTLPR* genotypes. The first three SNPs in the table contribute to the majority of the variance. Figure 1 represents the locations of the eight SNPs relative to *5-HTTLPR* with the base-pair locations along the horizontal axis.

Table 3 presents the results of generating the prediction model. The eight SNPs are given in the first column, with their locations in base pairs in the second. Columns 3–5 show the values for the prediction model for each SNP. The value chosen for the model will be based on whether the individual has 0, 1, or 2 minor alleles at that SNP. The coefficients of the equations are given in the last two columns. Each individual will have two predicted values. A Euclidean distance between those two values and the two values at each vertex is calculated. The vertex with the shortest distance is the predicted *5-HTTLPR* genotype. For example, if the predicted values are 0.1040 and −0.1807, and vertex S/S has values 0.7071 and 0.7071, then the Euclidean distance is given by:

$$d_1=\sqrt{(0.1040-0.7071)^2+(-0.1807-0.7071)^2}=1.0733,$$

The coordinates of the three vertices are S/S (0.7071, 0.7071), S/L (0.2588, −0.9659), and L/L (−0.9659, 0.2588).

Table 4 presents information regarding the performance of the model in predicting *5-HTTLPR* in the Finn and Dutch1 samples. Here, in Finn, the S/S genotype is predicted correctly 36 times and incorrectly six times, S/L is predicted correctly 139 of 150 times, and L/L 79 of 83 times. An analogous result is seen in the independent Dutch1 sample. Table 5 indicates that predicting the *5-HTTLPR* genotypes in Finn results in a 7.6% OMR, indicating an accuracy rate of 92.4% in the test dataset and a 0.98 AROC to distinguish the S allele carriers. As indicated by the $BMR_{S/L}$, 8.0% of the misclassified genotypes are heterozygous. The same pattern is observed in Dutch1. Predicted genotypes in both samples are consistent with HWE.

Table 6 presents the distributions of predicted *5-HTTLPR* genotypes for the Dutch and German samples taken separately, and the HapMap CEU family-based sample. The model has good predictive ability, as assessed by HWE of the *5-HTTLPR* genotypes ($P > 0.3$), and no detectable Mendelian errors in the HapMap CEU sample, although the number of trios is small.

In the assessment of association with schizophrenia, a small, but significant association ($P = 0.05$, odds ratio = 1.105) was found in the case–control sample of 3318 individuals. This effect size is consistent with those observed for individual SNPs in psychiatric disorders.

Discussion

The promoter polymorphism of the serotonin transporter gene (*5-HTTLPR*) has been reported to be associated with various personality traits and psychiatric disorders. For example, numerous studies (Beevers *et al.*, 2007; Osinsky *et al.*, 2008; Alexander *et al.*, 2009; Beevers *et al.*, 2009;

Crisan *et al.*, 2009) have found that carriers of the S allele are more sensitive to threats and stress; however, the literature lacks consistency (Caspi *et al.*, 2010). One can infer that studies with larger samples may detect significant but smaller effect sizes. Unfortunately, this variant is difficult to genotype. To address the problem, the current study demonstrates the feasibility of using SNP genotype data from standard GWAS arrays to predict *5-HTTLPR* genotypes in White Europeans.

SNPs present in widely used Illumina GWAS arrays were studied, and eight were selected from those in the region of the promoter polymorphism using stepwise regression. VDA was used to develop a model to assign *5-HTTLPR* genotypes. The R^2 of 0.85 between *5-HTTLPR* and the 8-SNP proxy set compares favorably with the R^2 of 0.7 that was recently reported for a two-SNP proxy of *5-HTTLPR* (Wray *et al.*, 2009). SNP array platforms have been evolving at a rapid rate, generating some concern regarding the availability of the eight predicators across platforms. Currently, all of them are assayed in Illumina Human1M-Duo DNA Analysis BeadChip and are in the HapMap and 1000 genomes projects, but three SNPs (*rs1487971*, *rs887469*, and *rs7217677*) are not assayed on the Human-Omni1-Quad BeadChip (1M-Quad). An alternative model on the basis of the markers assayed on 1M-Quad has been developed. Out of 77 tagging SNPs, 45 are assayed on the 1M-Quad across the region surrounding *5-HTTLPR*. An alternative prediction model for this platform is available from the authors, upon request.

The discrepancies between 'real' and predicted genotypes show that prediction is not perfect. However, some differences may be because of errors in genotyping the *5-HTTLPR* variant in these samples, and our method may be more accurate than what is shown. There is no 'gold standard' for *5-HTTLPR* genotyping, and given the known technical difficulties, both training sets and validation sets may have included incorrect genotypes. The misclassification observed here is not likely to result in a systematic bias in assigning genotypes to cases when compared with controls; however, it is likely to lead to a loss of power when testing for an association. To further evaluate the prediction method, models based on (a) the same eight predictors using an alternative approach, SVM, and (b) all of 77 tagging SNPs using VDA were generated. For those models, the OMR was 8.4% by SVM, just slightly higher than the 7.6% found with VDA, and 7.2% by the VDA with all tag SNPs, just a small improvement over the one based on eight predictors.

Hu *et al.*, 2006 show that a triple allele at this locus (S, L_A, L_G) may be a more specific predictor of disease. Although we have not had the data to conduct analyses to predict this polymorphism with SNPs, the same methods as those reported here can be used to do so in a sample that has been genotyped for the polymorphism and SNPs in the region. The method of VDA can easily be used to predict six *5-HTTLPR* genotypes in an analogous manner. The strength of VDA is that it is suitable for high-dimensional data.

The SNPs in the prediction model are commonly used and present in the HapMap database. Although this model was effective in predicting *5-HTTLPR* in multiple northern European populations, we cannot exclude the possibility that distinct patterns of linkage disequilibrium in specific, non-White populations will render the 8-SNP model less effective in predicting *5-HTTLPR*. In this respect, this model does not differ from other indirect genotyping approaches. Studies in a variety of populations will be needed to demonstrate its general applicability. When the model is inaccurate in a given population, the VDA approach can be applied if *5-HTTLPR* is genotyped in a GWAS subsample, as was done here in Dutch1. In addition, in samples without available GWAS data, the eight SNPs could be genotyped using established SNP genotyping techniques.

This prediction model was used to test for association in a combined schizophrenia case–control samples from the Netherlands and Germany, two populations that are considered relatively homogenous. The sample provided substantial statistical power, and a small but significant association of schizophrenia with *5-HTTLPR* was observed with an odds ratio consistent with those found for individual SNPs. Investigators with case–control schizophrenia study samples should examine this association using the prediction model. Until reliable, high-throughput genotyping or resequencing methods for *5-HTTLPR* become available, this provides an effective substitute that can provide the field of *5-HTTLPR* research with a new approach. In an analogous manner, other polymorphisms that prove difficult to genotype can also be predicted.

The model to predict the *5-HTTLPR* polymorphism genotypes from the genotypes of the eight SNPs is available in R code and can be obtained from the authors.

Acknowledgements

Conflicts of interest

There are no conflicts of interest.

References

Alexander N, Kuepper Y, Schmitz A, Osinsky R, Kozyra E, Hennig J (2009). Gene-environment interactions predict cortisol responses after acute stress: implications for the etiology of depression. *Psychoneuroendocrinology* **34**:1294–1303.

Ayers KL, Lange K (2008). Penalized estimation of haplotype frequencies. *Bioinformatics* **24**:1596–1602.

Bamber D (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol* **12**: 387–415.

Beevers CG, Gibb BE, Mcgeary JE, Miller IW (2007). Serotonin transporter genetic variation and biased attention for emotional word stimuli among psychiatric inpatients. *J Abnorm Psychol* **116**:208–212.

Beevers CG, Wells TT, Ellis AJ, Mcgeary JE (2009). Association of the serotonin transporter gene promoter region (5-HTTLPR) polymorphism with biased attention for emotional stimuli. *J Abnorm Psychol* **118**:670–681.

Capriotti E, Calabrese R, Casadio R (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**:2729–2734.

Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H, *et al.* (2003). Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science* **301**:386–389.

Caspi A, Hariri AR, Holmes A, Uher R, Moffitt TE (2010). Genetic sensitivity to the environment: the case of the serotonin transporter gene and its implications for studying complex diseases and traits. *Am J Psychiatry* **167**:509–527.

Crisan LG, Pana S, Vulturar R, Heilman RM, Szekely R, Druga B, *et al.* (2009). Genetic contributions of the serotonin transporter to social learning of fear and economic decision making. *Soc Cogn Affect Neurosci* **4**:399–408.

De Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D (2005). Efficiency and power in genetic association studies. *Nat Genet* **37**: 1217–1223.

De Leeuw J, Heiser W (1977). Convergence of correction matrix algorithms for multidimensional scalings. In: Roskam JC, Borg EI, editors. *Genometric representations of relational data*. Ann Arbor MI: Mathesis Press.

Grabe HJ, Spitzer C, Schwahn C, Marcinek A, Frahnow A, Barnow S, *et al.* (2009). Serotonin transporter gene (SLC6A4) promoter polymorphisms and the susceptibility to posttraumatic stress disorder in the general population. *Am J Psychiatry* **166**:926–933.

Hu XZ, Lipsky RH, Zhu G, Akhtar LA, Taubman J, Greenberg BD, Xu K, *et al.* (2006). Serotonin transporter promoter gain-of-function genotypes are linked to obsessive-compulsive disorder. *Am J Hum Genet* **78**:815–826.

Kong W, Choo KW (2007). Predicting single nucleotide polymorphisms (SNP) from DNA sequence by support vector machine. *Front Biosci* **12**: 1610–1614.

Lange K, Wu TT (2008). An MM algorithm for multicategory vertex discriminant analysis. *J Comput Graphical Stat* **17**:527–544.

Lesch KP, Bengel D, Heils A, Sabol SZ, Greenberg BD, Petri S, *et al.* (1996). Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science* **274**:1527–1531.

Lin E, Hwang Y (2008). A support vector machine approach to assess drug efficacy of interferon-alpha and ribavirin combination therapy. *Mol Diagn Ther* **12**:219–223.

Lin PY, Tsai G (2004). Association between serotonin transporter gene promoter polymorphism and suicide: results of a meta-analysis. *Biol Psychiatry* **55**:1023–1030.

Liu Q, Yang J, Chen Z, Yang MQ, Sung AH, Huang X (2008). Supervised learning-based tagSNP selection for genome-wide disease classifications. *BMC Genomics* **9 (Suppl 1)**:S6.

Lopez-Leon S, Janssens AC, Gonzalez-Zuloeta Ladd AM, Del-Favero J, Claes SJ, Oostra BA, Van Duijn CM (2008). Meta-analyses of genetic studies on major depressive disorder. *Mol Psychiatry* **13**:772–785.

Munafo MR, Freimer NB, Ng W, Ophoff R, Veijola J, Miettunen J, *et al.* (2009). 5-HTTLPR genotype and anxiety-related personality traits: a meta-analysis and new data. *Am J Med Genet B Neuropsychiatr Genet* **150B**:271–281.

Osinsky R, Reuter M, Kupper Y, Schmitz A, Kozyra E, Alexander N, Hennig J (2008). Variation in the serotonin transporter gene modulates selective attention to threat. *Emotion* **8**:584–588.

Pacheco J, Beevers CG, Benavides C, Mcgeary J, Stice E, Schnyer DM (2009). Frontal-limbic white matter pathway associations with the serotonin transporter gene promoter region (5-HTTLPR) polymorphism. *J Neurosci* **29**:6229–6233.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**:559–575.

Risch N, Herrell R, Lehner T, Liang KY, Eaves L, Hoh J, *et al.* (2009). Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. *JAMA* **301**:2462–2471.

Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, *et al.* (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* **41**:35–46.

Sen S, Burmeister M, Ghosh D (2004). Meta-analysis of the association between a serotonin transporter promoter polymorphism (5-HTTLPR) and anxiety-related personality traits. *Am J Med Genet B Neuropsychiatr Genet* **127B**:85–89.

Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, *et al.* (2009). Common variants conferring risk of schizophrenia. *Nature* **460**:744–747.

Vapnik VN (1995). *The nature of statistical learning theory*. Berlin: Springer-Verlag.

Vapnik VN (1998). *Statistical learning theory*. New York: Wiley.

Wray NR, James MR, Gordon SD, Dumenil T, Ryan L, Coventry WL, *et al.* (2009). Accurate, large-scale genotyping of 5HTTLPR and flanking single nucleotide polymorphisms in an association study of depression, anxiety, and personality measures. *Biol Psychiatry* **66**:468–476.

Yonan AL, Palmer AA, Gilliam TC (2006). Hardy–Weinberg disequilibrium identified genotyping error of the serotonin transporter (SLC6A4) promoter polymorphism. *Psychiatr Genet* **16**:31–34.

Zeller G, Clark RM, Schneeberger K, Bohlen A, Weigel D, Ratsch G (2008). Detecting polymorphic regions in Arabidopsis thaliana with resequencing microarrays. *Genome Res* **18**:918–929.

Zhou N, Wang L (2007). Effective selection of informative SNPs and classification on the HapMap genotype data. *BMC Bioinformatics* **8**:484.