February 2006 · Vol. 8 · No. 2

SNP selection at the *NAT2* locus for an accurate prediction of the acetylation phenotype

Audrey Sabbagh, PhD and Pierre Darlu

Purpose: Genetic polymorphisms in the N-acetyltransferase 2 gene determine the individual acetylator status, which influences both the toxicity and efficacy profile of acetylated drugs. Determination of an individual's acetylation phenotype prior to initiation of therapy, through DNA-based tests, should permit to improve therapy response and reduce adverse events. However, due to extensive linkage disequilibrium between markers within *NAT2*, the genotyping of closely spaced markers yields highly redundant data: testing them all is expensive and often unnecessary. The objective of this study is to establish the optimal strategy to define, in the genetic context of a given ethnic group, the most informative set of single-nucleotide polymorphisms that best enables accurate prediction of acetylation phenotype. **Methods:** Three classification methods have been investigated (classification trees, artificial neural networks and multifactor dimensionality reduction method) in order to find the optimal set of single-nucleotide polymorphisms enabling the most efficient classification of individuals in rapid and slow acetylators. **Results:** Our results show that, in almost all population samples, only one or two single-nucleotide polymorphisms would be enough to obtain a good predictive capacity with no or only a modest reduction in power relative to direct assays of all common markers. In contrast, in Black African populations, where lower levels of linkage disequilibrium are observed at *NAT2*, a larger number of single-nucleotide polymorphisms are required to predict acetylation phenotype. **Conclusion:** The results of this study will be helpful for the design of time- and cost-effective pharmacogenetic tests (adapted to specific populations) that could be used as routine tools in clinical practice. ***Genet Med* 2006:8(2):76–85.**

Key Words: *acetylation polymorphism, SNP selection, classification methods, genotyping tests, clinical pharmacy*

The human acetylation polymorphism is one of the first human hereditary traits affecting drug response to be discovered. It occupies a position of singular importance in the history of pharmacogenetics and in the future impact of the field on the practice of medicine.[1] It refers to inter-individual differences in the acetylation capacity of many clinically important drugs, as well as of known carcinogens present in the diet, cigarette smoke and the environment. Two main metabolic phenotypes have been described in human populations: the fast acetylator phenotype, associated with a normal acetylation capacity, and the slow acetylator phenotype characterized by a decreased enzyme activity. The proportions of rapid and slow acetylators vary remarkably in populations of different ethnic or geographic origin. The gene coding for the arylamine N-acetyltransferase 2 (NAT2) enzyme has been established as the site of the classic human acetylation polymorphism[2–4] and the molecular basis of individual and interethnic variation in acetylation capacity is now well documented.[5,6]

From the Unité de Recherche en Génétique Epidémiologique et Structure des Populations Humaines, INSERM U535, Villejuif, France.

Audrey Sabbagh, Inserm U535, Hôpital Paul Brousse, BP 1000 94817 Villejuif Cedex, France.

Received for publication, August 17 2005.

Accepted for publication, October 17 2005.

DOI: 10.1097/01.gim.0000200951.54346.d6

Individual differences in NAT2 activity have been proved to be important determinants of both the effectiveness of therapeutic response and the development of adverse drug reactions and toxicity during drug treatment.[7,8] Slow acetylators are generally more prone to side effects from drugs that are acetylated, due to the build-up of non-metabolized drugs.[9,10] On the contrary, fast acetylators may exhibit therapeutic failure after standard doses. Therefore, routine screening of individuals for their acetylator status prior to initiation of therapy should permit to improve drug efficacy and reduce adverse events, especially during chronic treatment with drugs known to undergo acetylation as a major metabolic pathway. For instance, the classification of patients as fast or slow acetylators facilitates the establishment of the appropriate dosage regimen of isoniazid used for the rational treatment of tuberculosis.[11] Kinzig-Schippers et al.[12] recently showed that, to achieve similar isoniazid exposure, current standard doses should be decreased or increased by approximately 50% for slow acetylator and fast acetylator patients, respectively.

The caffeine metabolite assay is currently the gold standard for assigning acetylator status through the measure of NAT2 activity in vivo.[13,14] However, the several potential limitations of phenotyping assays have led to the development of genotyping methodologies for the direct typing of the most common genetic polymorphisms in *NAT2*. Genotyping is generally ac-

cepted as an accurate and efficient means to determine acetylator status since a high correlation between phenotype and genotype has been demonstrated in several studies. In particular, the analysis of the seven most common SNPs (Single Nucleotide Polymorphism) in *NAT2* has been shown to be highly predictive of the acetylator phenotype with a prediction rate close to 100%.[11,15–19] The small discrepancies between genotyping and phenotyping studies may result from failures of the phenotyping test (sample handling, data reporting errors, assay failure), from confounding factors influencing phenotyping results (age, disease status, diet, compliance of drug intake, drug interaction, etc.), or may be due to the presence of additional undetected disabling mutations. But the relative agreement between phenotyping and genotyping studies indicate that unknown NAT2 mutant alleles should be present at low frequencies and therefore may not substantially influence the phenotype prediction in population studies.

However, the complete typing of a subject can only be achieved at a high cost, and several days are necessary to complete the analyses. For instance, the analysis of the seven major SNPs at the *NAT2* gene locus in one single subject requires several PCR reactions and seven RFLP assays, and further analyses are required to resolve the gametic phase of mutations and reconstruct haplotypes.[20] This would not be feasible in clinical practice. To become routine clinical tools, genotyping tests have to be cost- and time-effective; this implies a reduction in the number of SNPs to be typed. The issue of selecting the most informative markers for the prediction of acetylation phenotype is therefore of high clinical relevance.

The *NAT2* gene displays a strong haplotype structure with extensive linkage disequilibrium (LD) between markers and a limited haplotype diversity.[21,22] This feature makes unnecessary the genotyping of closely spaced SNP markers which would result in a large amount of redundant information. Indeed, in such a case, only a small fraction of SNPs can be used to distinguish a large fraction of the haplotypes.[23] This offers the possibility to dramatically reduce the number of SNPs required to completely genotype a sample without losing much haplotype information.

The objective of the present study is to identify the most independent and informative SNPs within *NAT2* that could be efficiently genotyped on large samples. And specifically, we aim to determine whether there exists a smaller combination of SNPs that permits to assess acetylator phenotype with a predictive power as high as that reached when all common SNPs are typed. Furthermore, because of large interethnic differences in *NAT2* allele frequencies and of variable pattern of LD across populations, the SNPs to be typed in a genotyping test are likely to be different for every ethnic group. We thus examined to what extent the optimal subset of SNPs differs from one population to another.

We handled these issues by using some recently developed classification methods. Three approaches have been explored: the first one implements a tree-based analysis and makes use of decision trees,[24] the second one is based on artificial neural networks,[25] and the last one is the multifactor dimensionality reduction method.[26] By using these classification methods, we aimed to find the smallest set of SNPs within *NAT2* that enables the most efficient classification of individuals into rapid and slow acetylators. Compared to traditional techniques of analysis such as logistic regression, these nonparametric statistical methods offer the possibility to model complex nonlinear relationship between phenotype and genotype, without the explicit construction of a complicated statistical model. Another practical advantage of these methods is their use of unphased multi-locus genotypes as input data, which alleviates the need of reconstructing haplotypes from genotype data. While the primary goal of these approaches was to highlight an association between candidate gene polymorphisms and a disease phenotype, we show in this study that they can also be useful tools for selecting highly informative markers to predict individual metabolizer status using pharmacogenetic data.

MATERIALS AND METHODS

NAT2 molecular data sets

We analyzed *NAT2* molecular data from eight previously published data sets. They concerned 258 Spanish from Central Spain,[27] 137 Nicaraguans with a Central American Indian-European mixed origin,[28] 1,000 Koreans,[19] 101 Black South Africans (Tswana-speaking people),[29] 564 Germans,[30] 248 Polish from the Wielkopolska region,[15] 303 Turks from south-east Anatolia,[31] and 50 non-caste Dogons from Mali, collected in 6 villages in the district of Sangha.[32] A summary description of the study samples is provided in Table 1.

In each population sample, all individuals were genotyped for the same seven nucleotide changes that are commonly found in human populations at *NAT2* (except in Koreans where the C190T mutation was investigated instead of G191A). Four result in an amino acid substitution that leads to a significant decrease in acetylation capacity (G191A, T341C, G590A, G857A). The other three are either silent mutations (C282T, C481T) or a non-synonymous substitution that does not alter phenotype (A803G).

The individual acetylation phenotypes were predicted from the diplotype configuration at *NAT2*. In the first four samples listed above, the mutation linkage phase was resolved directly through molecular haplotyping (combination of allele-specific PCR and restriction mapping) and this procedure was applied to all multiply heterozygous subjects. In the four others, linkage phase patterns were only partially resolved by molecular haplotyping, making haplotype phase information available for 41%–74% individuals. To infer haplotypes from the unresolved multi-locus genotypes, we employed the PHASE program (PHASE v 2.1),[33] using the default parameter values in the Markov chain Monte Carlo simulations. In this way, individual multi-site *NAT2* genotypes were assigned to a particular combination of two multi-locus haplotypes, each being considered as an allele of the *NAT2* gene.[34]

The *NAT2** alleles were classified on grounds of the current knowledge of the functional impact of the variant alleles. Con-

Table 1
Study samples

Sample	Description and selection criteria
258 Spanish[27]	Healthy, unrelated white Spanish volunteers. All subjects were in good health and with no antecedent of disease. Most subjects were medical students from Extremadura (Badajoz, Spain) and the surrounding area.
137 Nicaraguans[28]	Healthy, unrelated subjects from a mixed Nicaraguan population. Most of them were students and staff of the Universidad Nacional Autónoma de Nicaragua (León, Nicaragua). Only subjects with Central American Indian–white mixed origin were included. All were in good health and had no history of serious disease.
1,000 Koreans[19]	Korean individuals who visited the health promotion center at Samsung Medical Center.
101 Black South Africans[29]	Tswana-speaking people from the North-West Province. They were all healthy volunteers and were part of Transition and Health during Urbanization of South Africans (THUSA) study.
564 Germans[30]	Unrelated subjects of German origin comprising healthy volunteers and hospitalized individuals with various diseases but without known malignancy from the departments of Internal Medicine, Pulmonology, and Urology in Berlin, Germany.
248 Polish[15]	Unrelated children and adolescents of Polish origin. They were randomly selected from patients from the Wielkopolska region who had come to the Third Clinic of Children's Diseases of the University School of Medicine in Poznan because of ordinary respiratory or urinary tract infections or during a routine control visit to the outpatient clinic. Patients with autoimmune disease or malignancy were excluded.
303 Turks[31]	Unrelated Turkish individuals from south-east Anatolia (born and living in Gaziantep and surrounding). Except for 18 healthy volunteers, all were outpatients of the Gaziantep University, Faculty of Medicine, with a broad range of non-malignant diseases.
50 Dogons from Mali[32]	Healthy, unrelated black Africans, namely non-caste Dogons from Mali. They were collected in 6 villages in the district of Sangha, Republic of Mali.

sequently, the *NAT2*4*, *NAT2*12* and *NAT2*13* alleles were considered as functional alleles, and the *NAT2*5*, *NAT2*6*, *NAT2*7, NAT2*14* and *NAT2*19* alleles as slow alleles. Individuals with two low activity alleles were classified as slow acetylators, while those with one or two functional alleles were considered rapid acetylators.

The objective of our study was to determine whether a small subset of SNPs among the seven considered is able to recover the same classification of individuals into rapid and slow acetylators as that reached when all common SNPs are taken into account.

Classification trees

Zhang and Bonney[24] described an innovative use of classification trees for identifying disease genes and susceptibility alleles in association studies. We followed the same approach with the purpose of pointing up the SNP markers within *NAT2* that enable the best discrimination between slow and rapid acetylators.

To perform such a tree-based analysis, we first prepared data in a logistic regression format. In the present application, the response variable is the individual acetylator status, and seven covariates were created which record the number of copies (0, 1, or 2) of the minor allele for each SNP marker at *NAT2*. Then, we used the RTREE program (http://peace.med.yale.edu) for tree construction.

A detailed technical description for constructing classification trees is given elsewhere.[24,35] Briefly, the first step of tree construction is to build an initially large tree using recursive partitioning. During this step, the partition of an internal node into two offspring nodes is carried out by the values of one of the covariates, and it is aimed at improving the distribution homogeneity of the outcome, i.e., the acetylator status. Then, a second step called pruning is applied: it removes from bottom up those splits that may be "superficial" or based on an unreliably small samples. A split is regarded as unnecessary if the chi-square tests from this split as well as its further splits are not significant at a prespecified level. The pruning procedure was applied at various significance levels, from 0.01 to 10^{-6}.

Cross-validation methods were used to estimate the prediction error of the constructed decision tree by leaving out a portion of the data as an evaluation data. With five-fold cross-validation, each data set was divided into five groups with randomizing and alternating the data. Four groups were used to construct the classification tree, and one group was used as evaluation data; this construction and evaluation process was repeated five times, so that each group was assessed once as evaluation data. Then, the prediction accuracy of evaluation data across all five trials was calculated and averaged for the overall prediction accuracy of the decision tree. To ensure that the analysis was not influenced by a chance division of the data (i.e., an order effect), the analysis was repeated 10 times with randomizing the data.

Artificial neural networks

An artificial neural network (ANN) is a powerful data modeling tool that is able to capture and represent complex input/output relationships without having to code an explicit algorithm for deciding on the appropriate output. It is configured for a specific application, such as pattern recognition or data classification, through a learning process. The pattern-recognition properties of neural networks have been shown to be efficient tools to investigate association between a disease phenotype and a multi-locus genotype.[25,36–38]

We performed neural network analysis with the NNPERM package as described in North et al.[25] In the current application the initial inputs to the first layer of the network consist of *NAT2* multi-locus genotypes while the output consists of individual acetylator status. The following procedure was applied. For each subject we presented the SNP genotypes as input with that subject's acetylator status as the target output. This was repeated for each subject in order to train the network to predict acetylator status from SNP genotypes, and this training can only be successful if a significant association exists between the markers and the acetylation phenotype. This training process was repeated for all subjects over a number of training epochs. Once training was completed, a T statistic was computed to compare the outputs for slow and rapid acetylators in the same way as an unpaired t statistic, and the statistical significance of any observed association between genotype and acetylator status was estimated using a permutation test, as described in North et al.[25]

Since the goal of the present study is to select the most informative set of SNPs for the prediction of acetylation phenotype, we constructed an ANN model for each combination of N SNPs where N varies from one to seven. We compared the performance of each constructed ANN model (i.e., its ability to correctly classify subjects into slow and rapid acetylators from the multi-locus genotypes) through the computation of a classification error rate with the NeuroSolutions v4.0 software (NeuroDimension, Inc., Gainesville, FL).

To ensure that the network performs well on data that it has not been trained on, we estimated the prediction accuracy of the trained network using five-fold cross-validation: four-fifths of the data were assigned as learning data and were used to train the network (providing a classification error rate) and the one-fifth piece of the data left out as an independent test piece was assigned as evaluation data and was used to test the model's ability to generalize to independent data (providing a prediction error rate). The procedure was repeated for each of the five pieces of the data and the classification and prediction errors were averaged across all five trials. We ran the analysis 10 times, randomizing the order of data before presenting it to the network.

All data sets were analyzed using a neural network with two hidden layers of three nodes each. We showed indeed that a more complex architecture did not improve the neural network performance for the studied samples, and obtaining a permutation test *P*-value from 1,000 permutations each training over 200 training cycles.

Multi-factor dimensionality reduction

Ritchie et al.[26] developed a nonparametric and genetic model-free approach called multifactor dimensionality reduction (MDR) that reduces the dimensionality of multi-locus information to improve identification of polymorphism combinations associated with the risk for common complex multi-factorial diseases. A theoretical study has proved that MDR is ideally suited for discriminating between binary clinical endpoints using multi-locus genotypes.[39] The kernel of the MDR algorithm is comprised of three general steps: attribute selection, attribute construction, and classification. Model selection and evaluation is carried out using cross-validation and permutation testing. See Ritchie et al.[26,40] for the original descriptions of the MDR method.

We performed MDR analyses on the *NAT2* data sets using the MDR software package (http://www.epistasis.org/mdr.html).[41] We considered a number of N-factor models where N varies from one to seven. All possible combinations of N factors were evaluated sequentially for their ability to classify rapid and slow acetylators and the best N-factor model was selected. An MDR model is developed using 4/5th of the data and a classification error is estimated from this training set. Then, cross-validation methods are used to estimate the prediction error of the selected MDR model using 1/5th of the data as evaluation data. This procedure was repeated for each of the five pieces of the data and the classification and prediction errors were averaged across all five runs.

Single best models were selected from among each of the one-factor, two-factor, three-factor, up to seven-factor combinations. Among this set of best multifactor models, the combination of SNPs that minimizes the prediction error and maximizes the cross-validation consistency was selected. When two or more models had the same prediction error and cross-validation consistency, statistical parsimony was used to select the smaller model as the more likely candidate. An empirical *P*-value for the result was determined using a permutation testing strategy by randomizing the rapid and slow acetylator status in the original data set.[41]

We analyzed the data using five-fold cross-validation and 1,000-fold permutation testing. To ensure that the analysis was not influenced by a chance division of the data or by initial conditions, the analysis was repeated 10 times using 10 different random number seeds.

RESULTS

We will first detail results regarding the German sample. Figure 1 depicts the classification tree provided by the tree-based analysis when applied to the German sample. Only one tree was found after pruning whatever the significance level used, and all χ^2 tests performed at each internal node were highly significant (P-value $< 10^{-6}$). Only two SNPs are used in this tree (T341C and C282T) and they are both employed twice. For instance, in the case of T341C, a first split categorizes individuals with two 341C alleles on one side, and a further split in the tree distinguishes individuals with zero or one 341C allele. This suggests an additive effect of these alleles since individuals with one or two alleles are not in the same terminal nodes: an extra 341C or 282T allele increases the probability of being classified as a slow acetylator.

In this decision tree, each subject is classified either as a rapid or a slow acetylator according to his *NAT2* two-locus genotype with a 100% probability; there are indeed no misclassified individuals in any of the terminal nodes. An identical tree topology was obtained when four-fifths of the data were used to

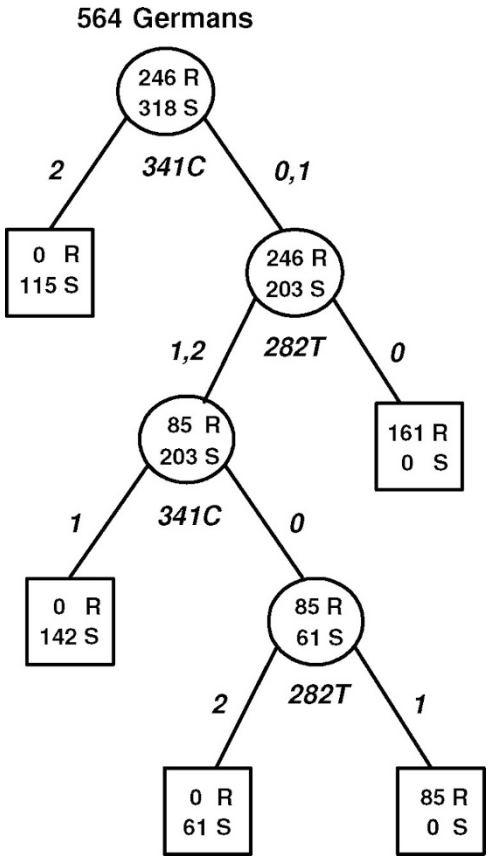

Fig. 1. The pruned tree at significance level 10^{-6} derived from the tree-based analysis of the German sample, when using the whole data set for tree construction.[30] Internal and terminal nodes are respectively represented by circles and boxes. The top node contains the entire study sample, and all other nodes are subsets of the study sample, which are some of the 564 German subjects investigated. Inside each node are the numbers of rapid (R) and slow (S) acetylators. Under each internal node is the split based on the genotype at one SNP marker (in italics). For example, the first internal node is split based on the number (2 vs. 0 or 1) of the minor allele 341C at position 341 of the *NAT2* coding sequence. Among all the single binary splits allowed by the alleles on the seven markers considered, this partition offers the "best possible" performance by attempting to send more slow acetylators in one offspring node and more rapid acetylators in the other one. Individuals with two 341C alleles are classified as slow acetylators.

construct the tree, in all five cross-validation trials and across all ten runs. The overall prediction accuracy of this classification tree was 100%. Therefore, in this German sample only two SNPs are needed to predict the individual acetylator status with a prediction power as high as that reached when all the seven SNPs are considered. One of these SNPs (T341C) is, in fact, a functional polymorphism which entails a decreased acetylation capacity but the other one (C282T) is a silent polymorphism with no impact on phenotype. It is nonetheless informative for the prediction of acetylation phenotype since the 282T allele is almost always associated with two functional polymorphisms in the *NAT2* gene (590A, 857A) in this population sample. This allele can be therefore considered as a predictive marker for the presence of these two inactivating mutations.

The German sample was also analyzed using an artificial neural network. Highly significant *P*-values ($P \leq 0.000999$) were obtained for all the combinations of SNPs tested, except when G191A and G857A were considered either in isolation or in combination (non-significant *P*-values). Figure 2 displays the different classification rates achieved by the network when analyses were performed with all subsets of N SNPs among the seven investigated where N varies from one to seven. Obviously, the best performance of the neural network is observed when all the seven SNPs are considered: the prediction accuracy of the network achieves the maximal value of 100%. However we can note that the same performance is achieved when smaller sets of markers are used as input data in the network. In particular, a combination of two SNPs, C282T and T341C, can predict acetylator status with the same ability as the entire set of SNPs. They are the same as those pointed out by the tree-based method.

The results of the MDR analysis of the German data set for each number of factors considered are presented in Table 2. The model with the lowest prediction error and highest cross-validation consistency was selected for each SNP combination level performed. The reported cross-validation consistency is the number of cross-validation intervals (maximum of 5) that a particular combination of SNPs was selected as the best model by MDR averaged across the 10 runs. The average classification and prediction errors of each selected model are the averages across all cross-validation intervals and all runs. The most parsimonious model that minimized prediction error and maximized the cross-validation consistency was the two-factor model that included again the SNPs C282T and T341C. The permutation testing indicated the cross-validation consistency and the prediction error are statistically significant at the 0.001 level. The prediction rate provided by this two-SNP model is as high as that displayed by the seven-factor model.

There is a sharp contrast with the results provided by the MDR analysis of the African samples. For instance, in the case of the Malian sample (Table 3), the maximal values of the prediction rate and cross-validation consistency are only reached when the seven SNPs are considered. The most parsimonious model consists of the three-factor model composed of G191A, T341C, and G590A; it can predict acetylator status with a prediction rate of only 98% compared with the 100% rate achieved with the seven-factor model.

The results of all analyses performed on the eight studied samples with the three methods investigated are presented in Table 4. In all samples except the African ones, the tree-based analysis provided only one tree topology whatever the significance level used for pruning: a single combination of SNPs was thus found in these samples. In the African samples, we chose to select the subset of SNPs involved in the tree enabling the best discrimination between slow and rapid acetylators. In the MDR analysis, the algorithm selects only one combination of SNPs for each number of factors considered, whereas in the neural network analysis, more than one combination of markers can be selected for each SNP combination level since all combinations of SNPs are evaluated by the user. For instance, in the Spanish sample, the SNP C481T can be used instead of T341C without changing the network's prediction accuracy. All approaches provided concordant results for all studied

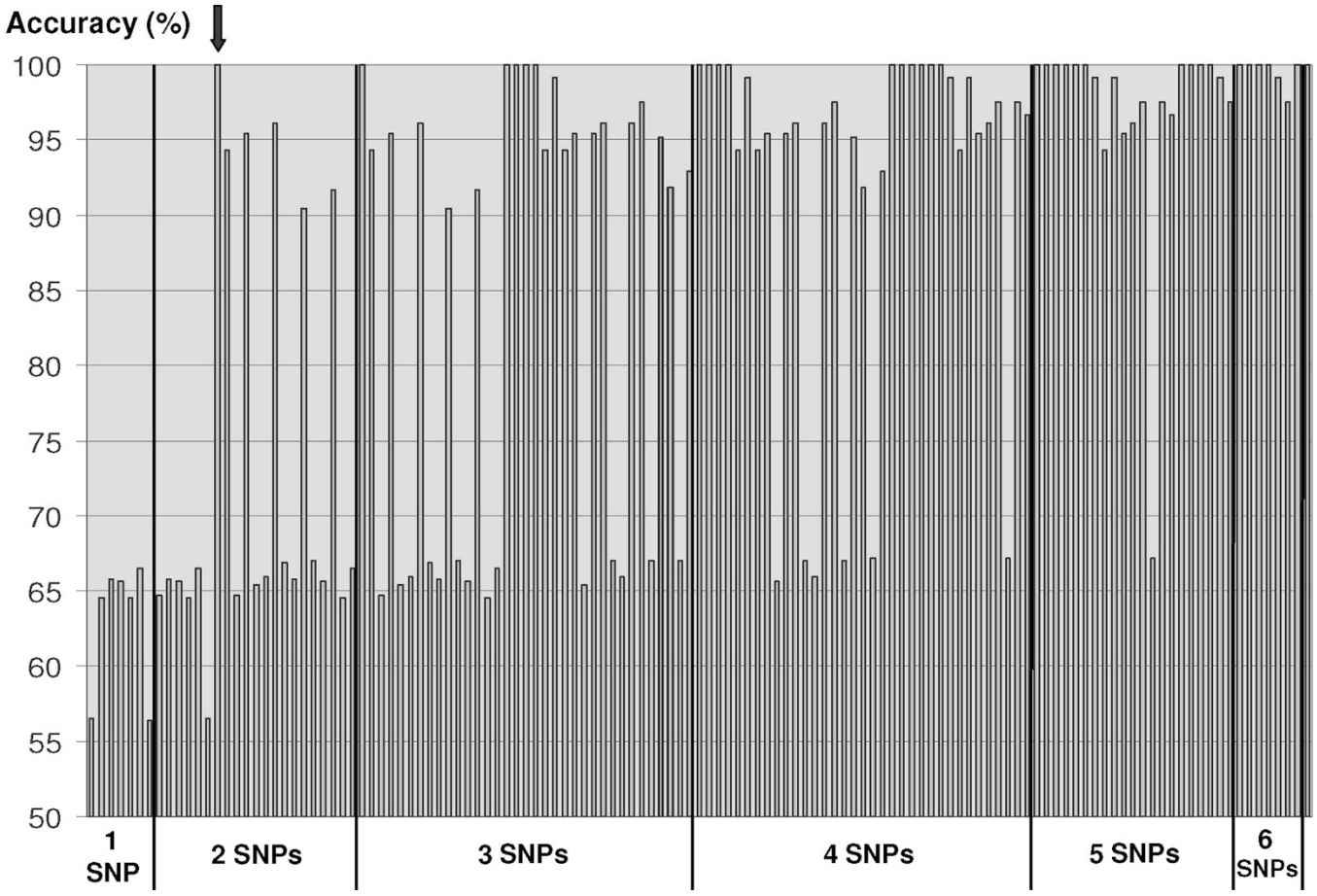

Fig. 2. Results of the neural network analysis of the German sample.[30] The graph shows the different values of the network's prediction accuracy when all possible combinations of SNPs, from one to six SNPs, are considered as input data. They are compared to the value obtained when all the seven SNPs are considered (last bar right in the chart). The best neural network performance (accuracy of 100%) is achieved with several subsets of SNPs, among which a two-SNPs model involving C282T and T341C (pointed by the black arrow).

Table 2
Results of the MDR analysis of the German sample[30]

No. of factors considered	Best candidate model	Average cross-validation consistency	Average classification error (%)	Average prediction error (%)
1	A803G	4.4	33.65	35.67
2	C282T, T341C	5.0[a]	0.00[a]	0.00[a]
3	G191A, C282T, T341C	5.0	0.00	0.00
4	G191A, C282T, T341C, C481T	5.0	0.00	0.00
5	G191A, C282T, T341C, C481T, G590A	5.0	0.00	0.00
6	G191A, C282T, T341C, C481T, G590A, A803G	5.0	0.00	0.00
7	G191A, C282T, T341C, C481T, G590A, A803G, G857A	5.0	0.00	0.00

[a] $P < 0.001$.

samples: the same subsets of SNPs were selected whatever the method used.

As shown above, only 2 SNPs (C282T and T341C) in the German sample can predict acetylator status with the same ability than the seven common SNPs. The same finding was also observed in all other European samples investigated (Polish, Turks and Spanish). In Nicaraguans, another combination of two SNPs was selected (T341C and G590A), but again, it was sufficient to reach a prediction rate of 100%. In Koreans, the best model is composed of the two same SNPs as those found in Europeans; however, it is interesting to note that only one of them, C282T, can predict acetylator status with a very high probability (99%). In contrast, in the two Black African populations, three SNPs (G191A, T341C, and

Table 3
Results of the MDR analysis of the Malian sample[32]

No. of factors considered	Best candidate model	Average cross-validation consistency	Average classification error (%)	Average prediction error (%)
1	C282T	3.7	29.31	36.80
2	T341C, G590A	4.6	12.00	15.20
3	G191A, T341C, G590A	5.0[a]	2.00[a]	2.04[a]
4	G191A, C282T, T341C, G590A, G857A	4.5	0.00	2.05
5	G191A, C282T, T341C, G590A, G857A	4.5	0.00	2.25
6	G191A, C282T, T341C, C481T, G590A, G857A	4.5	0.00	2.77
7	G191A, C282T, T341C, C481T, G590A, A803G, G857A	5.0	0.00	0.00

[a]$P < 0.001$.

Table 4
Best combinations of SNPs selected in each studied sample with the three classification methods

Sample	Tree-based analysis[a]		Neural network analysis[b]		Multifactor Dimensionality Reduction analysis[b]	
564 Germans[30]	C282T, T341C	(100%)	C282T, T341C	(100%)	C282T, T341C	(100%)
248 Polish[15]	C282T, T341C	(100%)	C282T, T341C	(100%)	C282T, T341C	(100%)
303 Turks[31]	C282T, T341C	(100%)	C282T, T341C	(100%)	C282T, T341C	(100%)
258 Spanish[27]	C282T, T341C	(99.6%)	C282T, T341C	(99.61%)	C282T, T341C	(99.61%)
			C282T, C481T	(99.61%)	G191A, C282T, T341C	(100%)
			G191A, C282T, T341C	(100%)		
			G191A, C282T, C481T	(100%)		
137 Nicaraguans[28]	T341C, G590A	(100%)	T341C, G590A	(100%)	T341C, G590A	(100%)
1000 Koreans[19]	C282T, T341C	(99.0%)	C282T	(99.00%)	C282T	(99.00%)
			C282T, T341C	(99.70%)	C282T, T341C	(99.70%)
			T341C, G590A, G857A	(100%)	T341C, G590A, G857A	(100%)
101 Black South Africans[29]	G191A, T341C, G590A	(95.0%)	G191A, T341C, G590A	(100%)	G191A, T341C, G590A	(100%)
50 Dogons from Mali[32]	G191A, T341C, G590A	(98.0%)	G191A, T341C, G590A	(98.00%)	G191A, T341C, G590A	(98.00%)
			G191A, T341C, G590A, G857A	(100%)	G191A, T341C, G590A, G857A	(100%)

[a]The classification rate achieved in the decision tree constructed from each sample data is shown in parenthesis. It is averaged across all cross-validation intervals and all runs. The tree-based analysis of the two African samples provided more than one tree when we changed the significance level of the pruning procedure from 0.01 to 10^{-6}. We chose to select the tree enabling the best discrimination between slow and rapid acetylators (i.e., the tree which minimized the misclassification rate).
[b]The average classification rate of each selected model is shown in parenthesis. It is averaged across all cross-validation intervals and all runs, and indicates how well the model performs on the whole data set. We selected for each sample the most parsimonious multifactor models that display an average classification rate equal or very close to the 100% value. All selected models are statistically significant at the 0.001 level.

G590A) are required to assess acetylation phenotype and the corresponding prediction rate is lower than that observed in the other populations.

DISCUSSION

Knowledge of the genetic basis of acetylation polymorphism should led to the development of genotyping tests of high efficiency and accuracy that will become routine tools with which clinicians will select medications and drug doses for individual patients. The procedure requires genotype information about a small number of individuals for an initial set of SNPs and selection of an optimum subset of SNPs that could be efficiently genotyped on larger numbers of samples while retaining most of the genetic variation in samples.

A practical and ethical concern is the transferability of diagnostic tests across ethnic groups. Our results show that the most informative subset of SNPs for the prediction of acetylation phenotype in one population may not necessarily perform well in another if the populations are sufficiently differentiated. Two distinct reasons may explain why the selected SNPs differ across ethnic groups. First, the underlying genetic causes of acetylation phenotype show significant differences in allele frequency across populations. The second reason is the variable pattern of LD across populations with different demographic histories. This is of particular relevance when markers,

instead of causal variants, are used diagnostically. A marker that has been associated with a phenotype in a given population, but that is not itself causal, is likely to have less or even no diagnostic value in other ethnic groups.[42] All this justifies why marker selection strategies should be applied separately, at least within different geographic areas. However, the fact that the subset of SNPs ascertained in German samples was also selected in the other European samples, and the same combination of SNPs was found in the two Black African populations, provides some reassurance that within the major human ancestral geographic groups, the SNPs to be targeted in genotyping tests are portable among populations.

From our findings in the European population samples studied, we can deduce that, for the purpose of phenotype prediction, the analysis of mutations at 282T and 341C would be enough to obtain a good predictive capacity in these populations, with no reduction in power relative to direct assays of all seven common SNPs. The analysis of these two polymorphisms would offer at a low cost a typing methodology that can be carried out in few hours and that avoids the use of probe drugs. This finding should encourage the routine typing of acetylator status in clinical practice in order to ensure adequated drug therapy with minimal or no toxic effects. Since all the major *NAT2* haplotypes are expected to be shared between the general European-derived populations, it is reasonable to expect that this combination of SNPs will also perform well in other populations of European origin. In Asian populations, two main 'slow' alleles, *NAT2*6A* and *NAT2*7B*, have been shown to predominate at *NAT2*,[43] and they can be both characterized by the C282T polymorphism. This explains why this marker alone is able to predict the slow acetylator phenotype with such a high probability in the Korean sample. Since the Chinese, Japanese, Korean, and Thai populations show comparable *NAT2* allele frequencies, this finding is also likely to hold in other populations along the Pacific Asian littoral. In contrast, in Black African populations, more SNPs are required to predict the individual acetylator status. These populations are indeed haplotypically more diverse at *NAT2* and, since they display lower levels of LD at this locus, SNPs are poor markers of each other in these populations.[22] These features have been observed for many other loci than *NAT2*. The geographic patterns reported in most studies of nucleotide variability in humans generally reveal more variation in sub-Saharan African populations than in other continental regions, and this observation is often interpreted as evidence for the out-of-Africa model.[44–47] Furthermore, reviews of published data based on analyses of multiple loci show strong variation of LD patterns among major continental groups, Africans displaying lower levels of LD compared to samples from other parts of the world.[48–51] Because linkage disequilibrium decreases through time, as a result of recombination, levels of disequilibrium can be correlated with the relative "age" of a population, with older populations having less disequilibrium. The linkage disequilibrium results are thus also consistent with an African origin of modern humans. In the West African and South African samples investigated in our study, the same optimal set of SNPs was selected. But since African populations are significantly differentiated, further surveys on the genetic variation of *NAT2* throughout sub-Saharan Africa are needed to determine to what extent this same subset of markers will work adequately in other African populations.

In order to simplify the typing of the *NAT2* gene, several authors[20,28,52] advocated the analysis of only the most prevalent mutations producing a defective NAT2 function to predict acetylation phenotype in clinical settings. However this criterion for marker selection is not necessary the most efficient one: looking at the patterns of LD between the different sites may be also useful. Indeed we showed in this study that a silent polymorphism, C282T, could be predictive of the presence of two enzyme-inactivating mutations due to extensive LD between these markers. This feature offers the possibility to reduce the number of SNPs to be targeted in a genotyping test.

While the three classification methods investigated in this study produced comparable results for all studied samples, they provide different kinds of information that can be used in a complementary manner. Indeed, the tree-based method generates a decision-tree model that provides simple rules to classify subjects into slow and rapid acetylators according to their unphased multi-locus genotypes. For instance, in the case of the German sample, the acetylator status of an individual can be predicted from his genotype at only two SNPs, and there is no need to resolve haplotype phase. Furthermore, one may decide to type only the T341C SNP in a first step, and if the subject is homozygous for the 341C allele, genotype data at the second SNP would be no further needed. Of course each decision-tree is population-specific and can only be used to predict the acetylation phenotype of individuals of the same ethnic background. However, the tree-based approach often yields a unique solution and, since it does not evaluate the performance of all possible combinations of SNPs, it does not provide any information on the additional markers to type if one wants to improve the discrimination power of the classification tree. This information is available when using the two other approaches, neural network and MDR analyses, which determine the best model for each SNP combination level tested. The drawback of the MDR method is that, although it gives useful guidelines to select the most informative markers for phenotype prediction, it is not able to predict acetylator status from individual multi-locus genotypes. In contrast, the neural network approach offers the possibility, once trained with the *NAT2* genotype data of a given ethnic group, to predict the acetylator phenotype of a new subject of the same ethnic background. Furthermore, the ability of this method to identify alternative minimal subsets of SNPs, when available, can be valuable in practice when individual SNPs prove difficult to genotype.

In the particular case of the *NAT2* gene, the three classification methods appear to perform similarly but it is still possible that, under different conditions (longer gene, larger haplotype diversity, different patterns and/or levels of LD), one method stands out from the others.

More traditional statistical tests exist that permit dealing with the same issue raised in this paper; that is, selection of the most informative SNPs for the prediction of a discret phenotype. To compare the performance of classical statistical approaches with that of the three classification methods investigated in this study, we performed additional analyses using logistic regression and discriminant analysis. Identical results were obtained: the same optimal subsets of SNPs were pointed out in each study sample by both methods (data not shown). This is not surprising since the phenotype-genotype relationship underlying the acetylation polymorphism is quite simple and of the linear type. However, in more complex cases where multiple predictive features interact and correlate with outcomes in complex ways, the use of systems able to afford non linear tasks, like artificial neural networks or MDR method, should allow a better discriminating capacity in comparison with classical statistics. In fact, there have been many successful applications of these classification methods in genetic and epidemiological studies. For instance, the recent study of Di Luca et al.[53] demonstrated that neural networks were more efficient than conventional statistical analyses to predict the presence of Alzheimer disease in early stages. Similarly, Tomita et al.[38] showed that neural networks discriminated cases from controls more precisely than logistic regression for diagnostic prediction of childhood allergic asthma. As these classification methods are easy-to-use, not time-consuming, and all implemented in freely-available and user-friendly softwares, they constitute interesting alternatives to classic parametric statistics. They should be of great use when applied to a large number of polymorphisms within a group of several interacting genes involved in a common pathway of drug response (e.g., genes that encode enzymes that act at different points in the metabolism of a drug or genes that encode a receptor complex).

CONCLUSION

This paper reports the first attempt to use classification approaches in pharmacogenetic analyses to predict one individual's drug response. It presents an innovative use of three classification methods which appear to be efficient and reliable techniques for selecting the most informative set of markers within a gene or a group of genes to predict one individual's metabolizer status. The results of this study will be helpful for the design of cost-and time-effective genotyping strategies, adapted to specific populations, to predict acetylation phenotype. This should facilitate the introduction of pharmacogenetic tests into widespread clinical practice.

REFERENCES

1. Hein DW, McQueen CA, Grant DM, Goodfellow GH, et al. Pharmacogenetics of the arylamine N-acetyltransferases: a symposium in honor of Wendell W. Weber. *Drug Metab Dispos* 2000;28:1425–1432.
2. Blum M, Grant DM, McBride W, Heim M, et al. Human arylamine N-acetyltransferase genes: isolation, chromosomal localization, and functional expression. *DNA Cell Biol* 1990;9:193–203.
3. Ohsako S, Deguchi T. Cloning and expression of cDNAs for polymorphic and monomorphic arylamine N-acetyltransferases from human liver. *J Biol Chem* 1990; 265:4630–4634.
4. Blum M, Demierre A, Grant DM, Heim M, et al. Molecular mechanism of slow acetylation of drugs and carcinogens in humans. *Proc Natl Acad Sci U S A* 1991;88: 5237–5241.
5. Grant DM, Hughes NC, Janezic SA, Goodfellow GH, et al. Human acetyltransferase polymorphisms. *Mutat Res* 1997;376:61–70.
6. Upton A, Johnson N, Sandy J, Sim E. Arylamine N-acetyltransferases - of mice, men and microorganisms. *Trends Pharmacol Sci* 2001;22:140–146.
7. Butcher NJ, Boukouvala S, Sim E, Minchin RF. Pharmacogenetics of the arylamine N-acetyltransferases. *Pharmacogenomics J* 2002;2:30–42.
8. Meisel P. Arylamine N-acetyltransferases and drug response. *Pharmacogenomics* 2002;3:349–366.
9. Hiratsuka M, Kishikawa Y, Takekuma Y, Matsuura N, et al. Genotyping of the N-acetyltransferase 2 polymorphism in the prediction of adverse drug reactions to isoniazid in Japanese patients. *Drug Metab Pharmacokinet* 2002;17:357–362.
10. Tanaka E, Taniguchi A, Urano W, Nakajima H, et al. Adverse effects of sulfasalazine in patients with rheumatoid arthritis are associated with diplotype configuration at the N-acetyltransferase 2 gene. *J Rheumatol* 2002;29:2492–2499.
11. Parkin DP, Vandenplas S, Botha FJ, Vandenplas ML, et al. Trimodality of isoniazid elimination: phenotype and genotype in patients with tuberculosis. *Am J Respir Crit Care Med* 1997;155:1717–1722.
12. Kinzig-Schippers M, Tomalik-Scharte D, Jetter A, Scheidel B, et al. Should we use N-acetyltransferase type 2 genotyping to personalize isoniazid doses? *Antimicrob Agents Chemother* 2005;49:1733–1738.
13. Tang BK, Kadar D, Qian L, Iriah J, et al. Caffeine as a metabolic probe: validation of its use for acetylator phenotyping. *Clin Pharmacol Ther* 1991;49:648–657.
14. Vincent-Viry M, Pontes ZB, Gueguen R, Galteau MM, et al. Segregation analyses of four urinary caffeine metabolite ratios implicated in the determination of human acetylation phenotypes. *Genet Epidemiol* 1994;11:115–129.
15. Mrozikiewicz PM, Cascorbi I, Brockmoller J, Roots I. Determination and allelic allocation of seven nucleotide transitions within the arylamine N-acetyltransferase gene in the Polish population. *Clin Pharmacol Ther* 1996;59:376–382.
16. Meisel P, Schroeder C, Wulff K, Siegmund W. Relationship between human genotype and phenotype of N-acetyltransferase (NAT2) as estimated by discriminant analysis and multiple linear regression:1. Genotype and N-acetylation *in vivo*. *Pharmacogenetics* 1997;7:241–246.
17. Gross M, Kruisselbrink T, Anderson K, Land N, et al. Distribution and concordance of N-acetyltransferase genotype and phenotype in an American population. *Cancer Epidemiol Biomarkers Prev* 1999;8:683–692.
18. Jorge-Nebert LF, Eichelbaum M, Griese EU, Inaba T, et al. Analysis of six SNPs of NAT2 in Ngawbe and Embera Amerindians of Panama and determination of the Embera acetylation phenotype using caffeine. *Pharmacogenetics* 2002;12:39–48.
19. Lee SY, Lee KA, Ki CS, Kwon OJ, et al. Complete sequencing of a genetic polymorphism in NAT2 in the Korean population. *Clin Chem* 2002;48:775–777.
20. Agundez JA, Olivera M, Martinez C, Ladero JM, et al. Identification and prevalence study of 17 allelic variants of the human NAT2 gene in a white population. *Pharmacogenetics* 1996;6:423–428.
21. Xu CF, Lewis K, Cantone KL, Khan P, et al. Effectiveness of computational methods in haplotype prediction. *Hum Genet* 2002;110:148–156.
22. Sabbagh A, Darlu P. Inferring haplotypes at the NAT2 locus: the computational approach. *BMC Genet* 2005;6:30.
23. Zhang K, Calabrese P, Nordborg M, Sun F, et al. Haplotype block structure and its applications to association studies: power and study designs. *Am J Hum Genet* 2002; 71:1386–1394.
24. Zhang H, Bonney G. Use of classification trees for association studies. *Genet Epidemiol* 2000;19:323–332.
25. North BV, Curtis D, Cassell PG, Hitman GA, et al. Assessing optimal neural network architecture for identifying disease-associated multi-marker genotypes using a permutation test, and application to calpain 10 polymorphisms associated with diabetes. *Ann Hum Genet* 2003;67:348–356.
26. Ritchie MD, Hahn LW, Roodi N, Bailey LR, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;69:138–147.
27. Agundez JA, Olivera M, Ladero JM, Rodriguez-Lescure A, et al. Increased risk for hepatocellular carcinoma in NAT2-slow acetylators and CYP2D6-rapid metabolizers. *Pharmacogenetics* 1996;6:501–512.
28. Martinez C, Agundez JA, Olivera M, Llerena A, et al. Influence of genetic admixture on polymorphisms of drug-metabolizing enzymes: analyses of mutations on NAT2 and C gamma P2E1 genes in a mixed Hispanic population. *Clin Pharmacol Ther* 1998;63:623–628.
29. Loktionov A, Moore W, Spencer SP, Vorster H, et al. Differences in N-acetylation genotypes between Caucasians and Black South Africans: implications for cancer prevention. *Cancer Detect Prev* 2002;26:15–22.

30. Cascorbi I, Drakoulis N, Brockmoller J, Maurer A, et al. Arylamine N-acetyltransferase (NAT2) mutations and their allelic linkage in unrelated Caucasian individuals: correlation with phenotypic activity. *Am J Hum Genet* 1995;57:581–592.
31. Aynacioglu AS, Cascorbi I, Mrozikiewicz PM, Roots I. Arylamine N-acetyltransferase (NAT2) genotypes in a Turkish population. *Pharmacogenetics* 1997;7:327–331.
32. Deloménie C, Sica L, Grant DM, Krishnamoorthy R, et al. Genotyping of the polymorphic N-acetyltransferase (*NAT2**) gene locus in two native African populations. *Pharmacogenetics* 1996;6:177–185.
33. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003;73:1162–1169.
34. Hein DW, Grant DM, Sim E. Update on consensus arylamine N-acetyltransferase gene nomenclature. *Pharmacogenetics* 2000;10:291–292.
35. Zhang H, Singer B. Recursive partitioning in the health sciences. New York: Springer, 1999.
36. Curtis D, North BV, Sham PC. Use of an artificial neural network to detect association between a disease and multiple marker genotypes. *Ann Hum Genet* 2001;65:95–107.
37. Serretti A, Smeraldi E. Neural network analysis in pharmacogenetics of mood disorders. *BMC Med Genet* 2004;5:27.
38. Tomita Y, Tomida S, Hasegawa Y, Suzuki Y, et al. Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. *BMC Bioinformatics* 2004;5:120.
39. Hahn LW, Moore JH. Ideal discrimination of discrete clinical endpoints using multi-locus genotypes. *In Silico Biol* 2004;4:183–194.
40. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* 2003;24:150–157.
41. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 2003;19:376–382.
42. Goldstein DB, Tate SK, Sisodiya SM. Pharmacogenetics goes genomic. *Nat Rev Genet* 2003;4:937–947.
43. Hamdy SI, Hiratsuka M, Narahara K, Endo N, et al. Genotype and allele frequencies of TPMT, NAT2, GST, SULT1A1 and MDR-1 in the Egyptian population. *Br J Clin Pharmacol* 2003;55:560–569.
44. Harding RM, Fullerton SM, Griffiths RC, Bond J, et al. Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet* 1997;60:772–789.
45. Zietkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, et al. Genetic structure of the ancestral population of modern humans. *J Mol Evol* 1998;47:146–155.
46. Harris EE, Hey J. X chromosome evidence for ancient human histories. *Proc Natl Acad Sci U S A* 1999;96:3320–3324.
47. Kaessmann H, Heissig F, von Haeseler A, Paabo S. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat Genet* 1999;22:78–81.
48. Kidd JR, Pakstis AJ, Zhao H, Lu RB, et al. Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am J Hum Genet* 2000;66:1882–1899.
49. Tishkoff SA, Pakstis AJ, Stoneking M, Kidd JR, et al. Short tandem-repeat polymorphism/alu haplotype variation at the PLAT locus: implications for modern human origins. *Am J Hum Genet* 2000;67:901–925.
50. Reich DE, Cargill M, Bolk S, Ireland J, et al. Linkage disequilibrium in the human genome. *Nature* 2001;411:199–204.
51. Osier MV, Pakstis AJ, Soodyall H, Comas D, et al. A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. *Am J Hum Genet* 2002;71:84–99.
52. Bouchardy C, Mitrunen K, Wikman H, Husgafvel-Pursiainen K, et al. N-acetyltransferase NAT1 and NAT2 genotypes and lung cancer risk. *Pharmacogenetics* 1998;8:291–298.
53. Di Luca M, Grossi E, Borroni B, Zimmermann M, et al. Artificial neural networks allow the use of simultaneous measurements of Alzheimer Disease markers for early detection of the disease. *J Transl Med* 2005;3:30.