Koichi Miyaki · Kazuyuki Omae · Mitsuru Murata
Norio Tanahashi · Ikuo Saito · Kiyoaki Watanabe

High throughput multiple combination extraction from large scale polymorphism data by exact tree method

Received: 15 March 2004 / Accepted: 18 May 2004 / Published online: 11 August 2004
© The Japan Society of Human Genetics and Springer-Verlag 2004

Abstract Single nucleotide polymorphisms (SNPs) are increasingly becoming important in clinical settings as useful genetic markers. For the evaluation of genetic risk factors of multifactorial diseases, it is not sufficient to focus on individual SNPs. It is preferable to evaluate combinations of multiple markers, because it allows us to examine the interactions between multiple factors. If all the combinations possible were evaluated round-robin, the number of calculations would rapidly explode as the number of markers analyzed increased. To overcome this limitation, we devised the exact tree method based on decision tree analysis and applied it to 14 SNP data from 68 Japanese stroke patients and 189 healthy controls. From the obtained tree models, we succeeded in extracting multiple statistically significant combinations that elevate the risk of stroke. From this result, we inferred that this method would work more efficiently in the whole genome study, which handles thousands of genetic markers. This exploratory data mining method will facilitate the extraction of combinations from large-scale genetic data and provide a good foothold for further verifactory research.

Keywords Multifactorial disease · Genetic polymorphism · SNP · Interaction · Multiple factor · Combination · Data mining · Exact tree

K. Miyaki (✉) · K. Omae
Department of Preventive Medicine and Public Health,
School of Medicine, Keio University, 35 Shinanomachi,
Shinjuku-ku, Tokyo 160-8582, Japan
E-mail: miyaki@sc.itc.keio.ac.jp

M. Murata · N. Tanahashi
Department of Internal Medicine, School of Medicine,
Keio University, Tokyo, Japan

I. Saito
Health Center, Keio University, Tokyo, Japan

K. Watanabe
Department of Laboratory Medicine, School of Medicine,
Keio University, Tokyo, Japan

Introduction

Biological researchers face an explosion of data arising from human genome projects and recent high throughput experiments. There are many traditional techniques for analyzing data, including statistics and epidemiological approaches. Data mining methods, however, offer new approaches to data analysis by using techniques based on machine learning that has been developed in research on artificial intelligence. These techniques work by learning patterns in data, and then sometimes discover new information overlooked by conventional analyses.

The amount of single nucleotide polymorphism (SNP) data as useful genetic polymorphism markers in clinical settings has been increasing, because SNPs are widespread and frequent in the human genome and high throughput typing technology has become established. To evaluate the risk of multifactorial diseases in detail, the information of single polymorphism alone is insufficient. It is preferable to evaluate combinations of multiple markers, since they reflect interactions between multiple factors. If all the combinations were evaluated round-robin, the amount of computation would easily explode as the number of markers analyzed increased at an exponential rate. To overcome this limitation, we developed the exact tree method based on decision tree analysis (Kass 1980), which we previously used and evaluated in another epidemiological study (Miyaki et al. 2002). The decision tree analysis is the popular classification method in data mining algorithms. This method evaluates all predictor variables and stratifies the study population with the best variable to contrast the high risk group and the low risk group as much as possible. This process is repeated, and the obtained “only one” tree, which is multiply stratified by best variables, becomes an empirical rule for predicting the class of an object from values of predictor variables.

On the other hand, the “exact tree” we devised aims not for constructing the best classification model but for

extracting hidden combinations that are statistically significant. The main principle of multiple stratifications is the same as the former “decision tree,” but this method constructs plural tree models in contrast to the one-tree model by regular decision tree algorithm.

In this article, we introduce an example of application of this new exploratory data mining algorithm to our SNPs data of stroke patients and healthy controls focusing upon the combination extraction.

Materials and methods

The patients and healthy controls analyzed here were from the same population published in our previous studies (Ishii et al. 2004; Ito et al. 2002; Oguchi et al. 2000; Sonoda et al. 2000 etc.). We recruited 235 unrelated Japanese patients under 70 years of age with symptomatic ischemic cerebrovascular disease (CVD) from Keio University Hospital and 189 age- and gender-matched healthy controls. CVD patients with cardio-embolic cerebral infarction and cerebral hemorrhage were excluded. Control subjects were enlisted from people who came to the hospital for regular checkups; those who had a clinical history of CVD or myocardial infarction or peripheral vascular diseases were excluded. Informed consent was obtained from all subjects. Brain CT and/or MRI were performed in all CVD patients. MR angiography and/or extracranial duplex ultrasonography were available for more than 80% of CVD patients. On the basis of classification of CVDs III (ad hoc committee of the National Institute of Neurological Disorders and Stroke 1990), the CVD patients were divided into three clinical categories: an atherothrombotic infarction group (69 patients), a lacunar infarction group (142 patients), and a transient ischemic attack group (24 patients). In this study, we focused on the atherothrombotic infarction group to clarify the differences from healthy controls by limiting the cases to one pathophysiological state. We typed 14 genetic polymorphism markers in 69 atherothrombotic stroke patients and 189 healthy controls. These 14 markers were chosen because they were already known to be related to stroke by effecting blood coagulation, platelet function, and lipid metabolism (target gene approach). These genetic polymorphism markers and their abbreviations used in this analysis are shown in Table 1, and the characteristics of the stroke patients and healthy controls are compared in Table 2.

For conventional statistical analysis, we used SPSS 11.0 and Clementine 7.1 (Chicago, IL, USA). For advanced statistical analysis, we devised the exact tree method based on decision tree analysis. The procedure of the exact tree method is described below. First, we made cross-tabs of genotypes between the groups of stroke patients and healthy controls. Then, we compared the prevalence of these alleles between the groups and calculated the *P* value of Fischer's exact test (See “Appendix”). We defined these *P* values as

Table 1 List of genetic polymorphism markers and abbreviations used for analysis. Genetic polymorphism markers are listed in alphabetical order by abbreviations

Genetic polymorphism markers	Abbreviations
5HT2A receptor (Exon1 102 T/C)	5HT2A
Angiotensin-converting enzyme (Exon 16)	ACE
Human apolipoprotein E	APO-E
Cholesteryl ester transfer protein	CETP
Fibrinogen β	FIB-B
Factor XII gene (Exon1 46 C/T)	F-XII
Glycoprotein Ib (145 Thr/Met)	GPIb (145Thr/Met)
Glycoprotein Ib kozak sequence (5 T/C)	GPIb (kozak)
Mitochondrial DNA mutation (5178 A/C)	Mitochondria
Methylene tetra hydro folate reductase (677 C/T)	MTHFR
Endothelial nitric oxide synthase (Intron 4)	NOS
NADPH oxidase p22phox	p22phox
Plasminogen activator inhibitor-1 (promotor region)	PAI-1
Paraoxonase (192 Arg/Gln)	PON

splitting indices of the data set to make the decision tree models. The smaller the *P* value, the more influential the allele was assumed to be. In regular decision tree methods, we adopt the best splitting point that shows the best index score (e.g., split the first data set by the marker that shows the smallest *P* value). But this method discards the second-best or the third-best onward, even when they are also meaningful. In the exact tree method, we also adopt the second-best or third-best onward as well as the first-best, as the occasion demands. How many best points we decide to adopt is the number of the tree models obtained later. For the effectiveness in data exploration, we must decide the number of the tree analytically. One simple approach is counting the variables that show the *P* value at an arbitrary statistical significance level depending on the number of variables analyzed. Another approach we adopted here was to make the talus plot of *P* values of each variable. The talus plot is a line graph of *P* values and is plotted in ascending order (Fig. 1). Using this plot, we could figure out the point where the slope changed sharply. If we chose the variables plotted on the left side of the large slope change, we adopted the variables that contained relatively larger amounts of information efficiently. With reference to Fig. 1, we chose the first five variables visually and made five tree models. In each tree model, we adopted the genetic polymorphism marker that had the smallest *P* value as the second splitting point and grew the trees with the next smallest *P* values continuously. Such a repetitive process needs stopping rules. As is common for all decision tree methods, the more the root node was split deeply, the smaller the number of subjects becomes in each stratified node, and it tends to be difficult to be interpreted. How many times the tree model should be split depends on the number of study subjects. In this report, we empirically defined the depth of tree as four, which is the level less than

Table 2 Comparative table of characteristics

	All participants (<i>n</i> = 257)	Stroke patients (<i>n</i> = 68)	Healthy controls (<i>n</i> = 189)
Demographic			
Age (years) ^a **	60.3 (5.2)	63.4 (7.7)	59.5 (3.7)
Male/female	201/56	56/12	145/44
Clinical			
Body-mass index (kg/m ²) ^a *	22.8 (2.7)	23.7 (2.6)	22.6 (2.7)
Smoking (%) never/ex or current **	65.4/34.6	42.1/57.9	72.5/27.5
Hypertension, yes/no *	82/173	30/38	52/135
Diabetes, yes/no **	35/222	22/46	13/176
Hyperlipemia, yes/no **	93/138	25/17	68/121
Family history of cerebellar vein disease, yes/no	37/141	13/39	24/102
Family history of cardiac artery disease, yes/no	17/163	4/51	13/112
Biochemical			
Fast blood sugar (mg/dl) ^a *	101.1 (18.8)	109.1 (31.7)	99.2 (13.7)
Total cholesterol (mg/dl) ^a	203.4 (32.7)	201.1 (37.1)	204.1 (31.1)
Triglyceride (mg/dl) ^a **	117.3 (77.4)	160.0 (90.8)	108.5 (71.5)
Serum uric acid (mg/dl) ^a	5.7 (1.3)	5.6 (1.4)	5.7 (1.3)
Platelet count (μl) ^a	235.5 (70.0)	229.4 (101.4)	237.5 (55.4)
High density lipoprotein (mg/dl) ^a **	56.2 (15.3)	44.0 (11.6)	57.5 (15.1)
CETP density (μg/ml) ^a	2.5 (0.6)	2.4 (0.6)	2.5 (0.6)

^amean (SD)**P* < 0.05, *t*-test; ***P* < 0.01, chi-square test **Fig. 1** Talus plot of *P* values for Fischer's exact test

half of the terminal nodes that consist of 30 subjects. Before reaching the defined depth, the extension was terminated when the number of cases or controls in the node became zero.

In addition, we calculated the odds ratios (OR) of each node with 95% confidence intervals (Figs. 2, 3, 4, 5, 6). Each OR stands for odds ratio of each node to "root node," which is the whole population we analyzed shown as Node #0 in each figure. It thus makes it possible to remove the influence of the sampling ratio of cases to controls on the estimated OR. We added the *P* value of each OR in addition to the 95% confidence interval of each OR. The OR shows the estimate of the risk ratio of the disease, so the odds ratio in this study indicates the risk of stroke. We circled the node with statistically significant odds ratio in each tree. These processes (which we call the "exact tree" method) enabled us to appraise risks quantitatively and extract meaningful combinations from the tree models constructed.

Results

The alleles used as variables in this analysis are listed in Table 3. The *P* values according to Fischer's exact test in the case-control tables are listed in Table 4, and the talus plot of these *P* values, which was plotted in ascending order, is shown in Fig. 1. From the view point of *P* value, p22phox (*P* value = 0.012) and APO-E (*P* value = 0.031) had statistically significant differences in allele frequencies between stroke patients and healthy controls. Based on the talus plots, we adopted F-XII (*P* value = 0.056), ACE (*P* value = 0.112), and GPIIb (kozak) (*P* value = 0.220) as useful genetic polymorphism markers for the first node splitting in the tree models, since it is more likely to be statistically significant in some combinations than markers with lower *P* values. By exact tree method described above, we constructed five exact tree models (Figs. 2, 3, 4, 5, 6). As to the second or later splitting point in each exact tree, we adopted the variables that had the smallest *P* value continuously. Subsequently, we calculated the OR of each node of the trees and drew circles around nodes whose 95% confidence interval of the odds ratio was not across 1 (i.e., significant at 5% alpha-error level). In the first exact tree (Fig. 2), which is the usual decision tree in itself, we could not extract any significant combination, since the only circled Node #2 indicated the elevated risk [Odds ratio: 2.34 (1.14–4.81)] in p22phox t allele-positive people (single polymorphism). Significant combination extraction needs more than twice-split circled node. In the second exact tree (Fig. 3), we observed three significant combinations in three nodes (Node #5, #8 and #11). In Node #5, for example, when with a combination of APO-E E4 allele-positive and PAI-1 5G allele-negative, the odds ratio became 3.24 (1.05–9.99). In Node #8,

Fig. 2 Exact tree model #1 (p22phox). Circle statistically significant combination, OR odds ratio of each node to root node (Node #0) and its 95% confidence interval. * $P < 0.05$, chi-square test

Fig. 3 Exact tree model #2 (APO-E). Circle statistically significant combination, OR odds ratio of each node to root node (Node #0) and its 95% confidence interval. * $P < 0.05$ and ** $P < 0.01$, chi-square test

with a combination of APO-E E4 allele-positive, PAI-1 5G allele-negative and Mitochondria A allele-negative, the odds ratio became 6.49 (1.63–25.79). And in Node #11, with a combination of APO-E E4 allele-negative, CETP B2 allele-negative, and ACE I allele-positive, the odds ratio became 0.41 (0.19–0.91), which can be a protective combination against stroke. In the same way, we can extract another one, three, and three statistically significant combinations from the tree models in Figs. 4, 5, and 6, respectively.

Fig. 4 Exact tree model #3 (F-XII). Circle statistically significant combination, OR odds ratio of each node to root node (Node #0) and its 95% confidence interval. * $P < 0.05$, chi-square test

Fig. 5 Exact tree model #4 (ACE). Circle statistically significant combination, OR odds ratio of each node to root node (Node #0) and its 95% confidence interval. * $P < 0.05$ and ** $P < 0.01$, chi-square test

Discussion

Although not shown in the Figures, we extended Nodes #7, #8, #9, and #10 in Fig. 2 (which are the same structure as Nodes #3, #4, #5, and #6 in Fig. 3) one more step down and found the same split in Nodes #7, #9, and #10. Node #8 was split by APO-E E2 allele instead of 5HT2A C allele. Because the left-side branch in Fig. 2 started from 222 subgroup [Node #1, p22phox t allele(-)], it is not surprising that the later splits are not

Fig. 6 Exact tree model #5 [GPIb (kozak)]. Circle statistically significant combination, OR odds ratio of each node to root node (Node #0) and its 95% confidence interval. * $P < 0.05$, chi-square test

Table 3 List of alleles used as variates for this analysis. Genetic polymorphism markers are listed in alphabetical order

Genetic polymorphism markers	Variates	
5HT2A	C allele	T allele
ACE	D allele	I allele
APO-E	E2 allele	E3 allele
CETP	B1 allele	B2 allele
FIB-B	H1 allele	H2 allele
F-XII	C allele	T allele
GPIb (145Thr/Met)	M allele	T allele
GPIb (kozak)	C allele	T allele
Mitochondria	A allele	T allele
MTHFR	A allele	V allele
NOS	a allele	b allele
P22phox	c allele	t allele
PAI-1	4G allele	5G allele
PON	Arg allele	Gln allele

completely consistent with the result from 257 whole population cases. When a similar structure was observed such as between Fig. 3 and the left-side branch in Fig. 2, we could speculate the root split (p22phox t allele in this case) influence little on the later splits. However, it does not mean the root split is unnecessary for making the tree, since the other side branch can be meaningful (unfortunately meaningless in this case). Such a consideration can be made because the number of tree models is plural in the exact tree method. Regular decision tree algorithm makes only one tree, and

Table 4 List of acquired P values by Fischer's exact test in case-control tables. Genetic polymorphism markers are listed in ascending order by P value

Genetic polymorphism markers	Variates	P values
P22phox	t allele*	0.012
APO-E	E4 allele*	0.031
F-XII	T allele	0.056
ACE	I allele	0.112
GPIb (kozak)	T allele	0.220
PAI-1	5G allele	0.506
Mitochondria	A allele	0.541
FIB-B	H1 allele	0.611
PON	gln allele	0.654
GPIb (145Thr/Met)	M allele	0.707
CETP	B2 allele	0.756
NOS	a allele	0.840
MTHFR	A allele	0.849
5HT2A	C allele	0.863

* $P < 0.05$, Fischer's exact test between stroke patients and healthy controls

structural similarity cannot be evaluated. This is one of the strong points the exact tree algorithm has over the regular decision tree algorithm.

As shown in Fig. 1, 5HT2A did not show a high score (i.e., low P value) by itself, but was useful for making trees (Figs. 2, 3, 5, 6). It is speculated that this polymorphism has a low risk by itself but has a high risk when combined with the other polymorphisms. Such a combination is the main target the exact tree method aims for.

From the same population, the allele frequencies observed can differ in each sampling time. However,

these differences can be controlled within reasonable statistical error rates. Of course even statistically significant combinations may contain some false positives, but this problem might be out of scope of exploratory methods by nature. Other verificatory approaches are needed for explored hidden combinations, and the exact tree method will work efficiently in the former steps.

In this paper, we adopted the P value of Fischer's exact test as the splitting criterion for the decision tree. This splitting criterion can be substituted for the chi-square of CHAID (Kass 1980), the improvement score of CART (Breiman et al. 1980), the gain ratio of C4.5 (Quinlan 1993), and so on, by using the exact tree method we proposed. We have already tested the effectiveness of the exact tree method with the odds ratio as the splitting criterion (data not shown). It is impossible to state which splitting criterion is the best in decision tree modeling, because the each criterion is the best in each mathematical modeling theory. So we adopted the P value of Fischer's exact test because it is easily interpreted as the probability and is familiar to biologists and medical researchers. Moreover, we can use the commonly used alpha error rate (0.05 or 0.01) as the standard for judging the magnitude of the value.

Here we need to clarify the advantage of the exact tree method over the regular decision tree method. If we adopt the regular decision tree method, we can get the best (i.e., only one) classification tree model (Fig. 2). On the contrary, the exact tree method makes plural tree models, including the best classification tree model (Figs. 2, 3, 4, 5, 6). In the real data analysis, while the regular decision tree shows us no significant combination (Fig. 2), five tree models constructed by the exact tree method showed us ten significant combinations (Figs. 2, 3, 4, 5, 6). Because the output of the exact tree method always includes the best tree model, it is never inferior to the regular decision tree method as to the extraction of hidden significant combinations. In addition, the amount of computation increases not exponentially but only several times from that of regular decision tree analysis. That is why we report this new method as an efficient and feasible exploratory technique for handling such large-scale genomic data as round-robin approach, which otherwise is impossible.

For reference, we executed the logistic regression analysis in which 14 polymorphisms are incorporated as independent variables. APO-E E4 allele and Factor XII gene t allele are statistically significant polymorphisms by both compulsory and stepwise methods. In the exact tree method, these two polymorphisms are selected as the first splitting variables in Figs. 3 and 4, and four statistically significant combinations are extracted. Regular decision tree method failed to find any combinations of the two polymorphisms above. This finding supports the advantage of the exact tree method over regular decision tree methods.

Since this method is an exploratory screening method, we do not insist that the individual results are solid evidences by themselves. We want to emphasize our

success in extracting combinations of three polymorphisms using real genetic data. From the obtained tree models, we succeeded in extracting ten statistically significant combinations that may elevate or decrease the risk of stroke.

Here we handle a very small data set (only 14 genetic markers) in this pilot analysis, since round-robin explorations of multiple combinations are executable in ordinary computer workstations. So the advantage of the exact tree method is obscure, but the situation will change in handling extremely large-scale data set. For example, when we handle 100,000 genetic markers, which can be real in some whole genome approach researches, we need 10,000,000,000 operations to check all the combinations of two markers. In this situation, we need 1,000,000,000,000,000 operations to check all the combinations of three markers. Such an acute increase of the amount of computation at an exponential rate makes it impossible to extract combinations of large numbers of markers. With this new data mining method, however, the increase of the amount of computation is not exponential, and it can extract multiple combinations efficiently without computational explosion.

To assess the degree of the reduction in calculation task quantitatively, we did the reduction rate estimation trial. Suppose the number of genetic polymorphisms is N and we want to know the statistically significant disease-associated combinations of three polymorphisms. In the round-robin approach, we need to execute chi-square test $N \times (N-1) \times (N-2)$ times. On the other hand, the 3-step splitting exact tree method (i.e., the depth of the tree is 4) demands $3 \times n \times N$ times, where n is the number of tree models we make, which is less than N and usually less than 10. So the degree of the reduction in calculation task is evaluated as $\frac{3 \times n \times N}{N \times (N-1) \times (N-2)}$ and this reduction ratio simply decreases as N increases. And the number of tree models (n) is usually rather smaller than the total number of polymorphisms in consideration (N), this reduction ratio decreases exponentially. In spite of the rather small number of polymorphisms we demonstrated in this paper ($N=14$, $n=5$), this reduction rate is estimated to be 0.096. This estimation suggests the exact tree method works effectively enough in handling rather small genetic data, and this efficiency will increase when handling larger data. This new data mining method will facilitate the extraction of combinations from the huge genetic data that we are now confronted with and may provide a good foothold for further verificatory research.

Appendix

The exact probability of observing a table with cells (a , b , c , d) is when

$$P(a, b, c, d) = \frac{m_1! C_a \cdot m_2! C_b}{N! C_{n_1}} = \frac{n_1! n_2! m_1! m_2!}{N!} \frac{1}{a! b! c! d!}$$

	Disease (+)	Disease (-)	
Polymorphism (+)	a	b	n_1
Polymorphism (-)	c	d	n_2
	m_1	m_2	N (total)

This probability distribution is known as the hypergeometric distribution (Bernard 2000).

Suppose the probability that a man had polymorphism (+) given that he had disease (+) = p_1 , and the probability that a man had polymorphism (+) given that he had disease (-) = p_2 . Here we wish to test the hypothesis $H_0: p_1 = p_2$ versus $H_1: p_1 \neq p_2$.

General procedure and computation of P value using Fischer's exact test are as follows:

1. Enumerate all possible tables with the same row and column margins (i.e., n_1, n_2, m_1, m_2) as the observed table
2. Compute the exact probability of each table enumerated in previous step
3. Suppose the observed table is the x table and the last table enumerated is the k table. To test the hypothesis $H_0: p_1 = p_2$ versus $H_1: p_1 < p_2$, the P value is computed as $P(0) + P(1) + \dots + P(x)$ (left-hand tail area). To test the hypothesis $H_0: p_1 = p_2$ versus $H_1: p_1 > p_2$, the P value is computed as $P(x) + P(x+1) + \dots + P(k)$ (right-hand tail area). To test the hypothesis $H_0: p_1 = p_2$ versus $H_1: p_1 \neq p_2$, the P value is computed as $2 \times \min[P(0) + P(1) + \dots + P(x), P(x) + P(x+1) + \dots + P(k), 0.5]$.

This P value can be interpreted as the probability of obtaining a table as extreme as the observed table.

References

- Ad Hoc Committee of National Institute of Neurological Disorders and Stroke (1990) Classification of cerebrovascular disease III. *Stroke* 21:637–676
- Bernard R (2000) Fundamentals of biostatistics, 5th edn. Duxbury, Pacific Grove
- Breiman L, Friedman JH, Olshen RA, Stone CJ (1980) Classification and regression trees. Wadsworth, Belmont
- Ishii K, Murata M, Oguchi S, Takeshita E, Ito D, Tanahashi N, Fukuuchi Y, Saitou I, Ikeda Y, Watanabe K (2004) Genetic risk factors for ischemic cerebrovascular disease of candidate prothrombotic gene polymorphisms in the Japanese population. *Rinsho Byori* 52:22–27
- Ito D, Tanahashi N, Murata M, Sato H, Saito I, Watanabe K, Fukuuchi Y (2002) Notch3 gene polymorphism and ischemic cerebrovascular disease. *J Neurol Neurosurg Psychiatr* 72:382–384
- Kass G (1980) An exploratory technique for investigating large quantities of categorical data. *Appl Stat* 29:119–127
- Miyaki K, Takei I, Watanabe K, Nakashima H, Watanabe K, Omae K (2002) Novel statistical classification model of type 2 diabetes mellitus patients for tailor-made prevention using data mining algorithm. *J Epidemiol* 12:243–248
- Oguchi S, Ito D, Murata M, Yoshida T, Tanahashi N, Fukuuchi Y, Ikeda Y, Watanabe K (2000) Genotype distribution of the 46C/T polymorphism of coagulation factor XII in the Japanese population: absence of its association with ischemic cerebrovascular disease. *Thromb Haemost* 83:178–179
- Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann, Los Altos
- Sonoda A, Murata M, Ito D, Tanahashi N, Oota A, Tada-Yatabe Y, Takeshita E, Yoshida T, Saito I, Yamamoto M, Ikeda Y, Fukuuchi Y, Watanabe K (2000) Association between platelet glycoprotein Ib α genotype and ischemic cerebrovascular disease. *Stroke* 31:493–497