Communications for Statistical Applications and Methods
2019, Vol. 26, No. 1, 1–9

https://doi.org/10.29220/CSAM.2019.26.1.001
Print ISSN 2287-7843 / Online ISSN 2383-4757

Evaluation of the classification method using ancestry SNP markers for ethnic group

Hyo Jung Lee[a], Sun Pyo Hong[b], Soong Deok Lee[c], Hwan seok Rhee[d], Ji Hyun Lee[b], Su Jin Jeong[e], Jae Won Lee[1, e]

[a]Product Development HQ, Korea; [b]Research & Development Center, GeneMatrix Inc., Korea;
[c]Department of Forensic Medicine, Seoul National University, College of Medicine, Korea;
[d]Bioinformatics Research Center, Macrogen Inc, Korea;
[e]Department of Statistics, Korea University, Korea

Abstract

Various probabilistic methods have been proposed for using interpopulation allele frequency differences to infer the ethnic group of a DNA specimen. The selection of the statistical method is critical because the accuracy of the statistical classification results vary. For the ancestry classification, we proposed a new ancestry evaluation method that estimate the combined ethnicity index as well as compared its performance with various classical classification methods using two real data sets. We selected 13 SNPs that are useful for the inference of ethnic origin. These single nucleotide polymorphisms (SNPs) were analyzed by restriction fragment mass polymorphism assay and followed by classification among ethnic groups. We genotyped 400 individuals from four ethnic groups (100 African-American, 100 Caucasian, 100 Korean, and 100 Mexican-American) for 13 SNPs and allele frequencies that differed among the four ethnic groups. Additionally, we applied our new method to HapMap SNP genotypes for 1,011 samples from 4 populations (African, European, East Asian, and Central-South Asian). Our proposed method yielded the highest accuracy among statistical classification methods. Our ethnic group classification system based on the analysis of ancestry informative SNP markers can provide a useful statistical tool to identify ethnic groups.

Keywords: single nucleotide polymorphisms (SNP), allele, ethnic group, classification, Korean population

1. Introduction

Ancestral classification markers could be used to assist with the identification of remains as well as guide criminal investigations towards individuals who cannot be excluded on the basis of ancestry. In some cases, an ancestral classification result could provide probable cause for the legal request of DNA from suspects that can create a leverage crux to maximize the efficacy of the criminal justice system.

Single nucleotide polymorphisms (SNPs) are biallelic and also have several attractive features as genetic markers. Due to the low mutation rate of SNP, their information refers to longer periods of time, compared to those obtained with Short Tandem Repeat (STR), and Variable Number Tandem

First and second authors contributed equally to this work.

[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea.
E-mail: jael@korea.ac.kr

Published 31 January 2019 / journal homepage: http://csam.or.kr

Repeat (VNTR) (Mountain *et al.*, 2002). Both STR and SNP have been used for ancestral classification markers, but most STR markers currently in use (i.e., F13A, TH01, FES/FPS, and vWA) offer low power indistinguishing between ancestral groups (Taillon-Miller *et al.*, 1999). In addition, in forensic and anthropological investigations, where biological samples are often poor or degraded, the particular advantage of SNPs is that the studied DNA sequences are shorter than those used for classical DNA analysis (Schafer and Hawkins, 1998).

A wide variety of methods for genotyping SNPs have recently been developed that include restriction fragment length polymorphism (RFLP), melting-curve analysis with fluorescence resonance energy transfer (FRET) probe, and single base extension (SBE) (Bray *et al.*, 2001). In this study, we use a restriction fragment mass polymorphism (RFMP) method for genotyping. The RFMP method is based on the matrix assisted laser desorption-ionization-time-of-flight (MALDI-TOF) mass spectrometry that has already been shown to be an efficient method to identify SNP sequences (Hong *et al.*, 2008; Hwang *et al.*, 2007).

Various probabilistic methods have been proposed for using interpopulation allele frequency differences to infer the ethnic group of a DNA specimen (Brenner, 1998; Evett *et al.*, 1992; Frudakis *et al.*, 2003; Lowe *et al.*, 2001). Furthermore, the selection of the statistical method is critical because the accuracy of the statistical classification results vary depending on the statistical methods used. Therefore, this study analyzes SNP frequencies obtained using the multiplex RFMP analysis method with the various statistical classification methods as well as suggests and optimal classification method based on the analysis result.

2. Materials and methods

2.1. Population DNA samples and selection of SNP genotypes

SNPs analysis was conducted on a set of 400 unique anonymized individuals of diverse geographic origin with self-described ethnic group affiliation information, chosen to represent four ethnic groups. These included 100 African-American, 100 Caucasian, 100 Mexican-American, and 100 Korean individuals. The 300 anonymized samples were obtained from the Coriell Cell Repositories (Coriell Institute for Medical Research), and 100 Korean anonymized individuals were obtained from all regions of Korea.

The 400 individuals from four different population groups were genotyped for the 13 SNP markers. These markers are rs492602, rs485186, rs140864, rs2814778, rs1042602, rs7495174, rs11855019, rs6497268, 1176–1174 dupAAT, rs6867641, rs13289, rs1801133, and rs1994798 (Table 1).

SNP is the marker to determine various human characteristics, and studies on the SNP marker have been reported in the fields of pharmacogenomics and forensics. We conducted the investigation on SNPs, which are known to be differently distributed among diverse human races, and determined final markers. In this study, we targeted ABO-Secretor genes, ancestry informative markers (AIMs), and a pigmentation metabolism related gene in our search for ancestrally informative SNPs because these genes are likely to have been subject to unusual selective pressures over the course of human evolution. The human alpha 1, 2 fucosyltransferase 2 (FUT2) gene plays a key role for tissue expression of the H antigen, and recent studies have indicated that rs492602 and rs485186 polymorphism showed an ethnic group-specific pattern (Koda *et al.*, 2001). AIM is a set of polymorphisms for a locus that exhibits substantially different frequencies between populations from different geographical regions (Pastinen and Hudson, 2004; Shriver *et al.*, 1997). Oculocutaneous albinism 2 (OCA2) gene plays important roles in control of pigmentation, and rs7495174, rs11855019 and rs6497268 poly-

Table 1: Allelic frequencies of 13 SNP markers in African-American, Caucasian, Mexican-American, and Korean populations

Gene	Marker (allele)	African-American			Caucasian			Mexican-American			Korean		
		MA	MAF	HWE	MA	MAF	HWE	MA	MAF	HWE	MA	MAF	HWE
FUT2	rs492602 (A/G)	A	0.480	0.106	G	0.470	0.401	G	0.270	0.245	G	0.005	0.960
	rs485186 (A/G)	A	0.460	0.038	A	0.455	0.601	G	0.310	0.031	G	0.000	-
AIMs	rs140864 (TTC+/TTC-)	-	0.150	0.556	-	0.005	0.960	-	0.405	0.319	+	0.320	0.136
	rs2814778 (A/G)	A	0.200	1.000	G	0.020	0.838	G	0.040	0.677	G	0.000	-
	rs1042602 (C/A)	A	0.040	0.677	A	0.370	0.248	A	0.225	0.092	A	0.000	-
OCA2	rs7495174 (A/G)	G	0.245	0.031	G	0.175	0.516	G	0.105	0.913	A	0.375	0.031
	rs11855019 (A/G)	A	0.260	0.244	G	0.200	0.532	G	0.220	0.098	A	0.390	0.000
	rs6497268 (A/C)	C	0.430	0.842	A	0.295	0.075	A	0.440	0.795	C	0.085	0.397
MATP	1176-1174 (dupAAT dup+/dup-)	-	0.365	0.000	+	0.400	0.000	+	0.465	0.292	+	0.325	0.798
	rs6867641 (C/T)	T	0.125	0.026	T	0.300	0.000	T	0.365	0.469	T	0.085	0.003
	rs13289 (C/G)	C	0.335	0.727	G	0.265	0.616	G	0.475	0.000	G	0.275	0.000
MTHFR	rs1801133 (C/T)	T	0.100	0.267	T	0.235	0.002	C	0.465	0.031	T	0.470	0.000
	rs1994798 (C/T)	T	0.360	0.652	C	0.495	0.000	C	0.270	0.000	C	0.375	0.000

MA = minor allele; MAF = minor allele frequency; HWE = Hardy-Weinberg equilibrium.

morphisms in intron 1 are highly associated with eye color (Duffy, 2007). The membrane-associated transporter protein (MATP) plays an important role as it is involved in intracellular processing and trafficking of melanosomal proteins, and duplication (1176–1174 dupAAT) of promoter and polymorphisms of rs6867641 and rs13289 are known to have significantly different frequencies between population groups (Graf *et al.*, 2005, 2007). Methylenetetrahydrofolate reductase (MTHFR) is a key enzyme for intracellular folate homeostasis and metabolism, and rs1801133 and rs1994798 polymorphisms varies with geographical origin (Botto and Yang, 2000; Rosenberg *et al.*, 2002) Additionally, we also used HapMap SNP genotypes for 1,011 samples from 4 populations (African, European, East Asian, and Central-South Asian), which were connected to SPSmart [http://spsmart.cesga.es]. These included 246 African (African-American 61, Kenya 97, Nigeria 88), 380 European (European 87, Finnish 93, British 88, Iberian 14, Toscani 98), 286 East Asian (Han Chinese 97, Southern Han Chinese 100, Japanese 89) and 99 Central South Asian (Gujarati Indian). Following the reference (Phillips *et al.*, 2013), we selected the 17 SNP markers known as optimal loci in Eurasian and East Asian populations. There are rs1785864, rs10131666, rs2472304, rs17625895, rs2156208, rs1363345, rs2227203, rs7354930, rs6026972, rs2835133, rs1519654, rs984038, rs39897, rs756913, rs2196051, rs11779571, rs10962599.

Half of the total individuals were selected by random sampling as the training set, and the rest were selected as the test set. A predictor or classifier for the 4 ethnic classes was built from the training set and applied to the test set to predict the class. The predicted and true classes were then compared to estimate the misclassification error rate of the predictor. We repeated the above process 1000 times. Finally, the average misclassification error rate was calculated by 1,000 iterations with a different training set in order to compare the performances of the classification methods.

2.2. Classical classification methods

In this paper, we applied statistical classification methods such as linear discriminant analysis (LDA) (Fisher, 1936), diagonal linear discriminant analysis (DLDA) (Bickel and Levina, 2004), diagonal quadratic discriminant analysis (DQDA) (Dudoit *et al.*, 2002), K-nearest neighbor (KNN) (Altman, 1992) as well as more modern ones, like classification and regression trees (CART) (Breiman, 1984).

Recent machine learning approaches, like support vector machines (SVM) (Vapnik, 2000), random forest (RF) (Breiman, 2001), nearest shrunken centroids (NSC) (Tusher *et al.*, 2001), and partial least squares discriminant analysis (PLSDA) (Nguyen and Rocke, 2004) were also considered. All data were analyzed using R, version 3.1.0. (https://www.r-project.org/).

We also applied two widely used the algorithms for forensic analysis: Snipper and STRUCTURE. Both Snipper and STRUCTURE are open programs frequently used for the classification of ethnic groups. The Snipper is the online Bayesian classification system (http://mathgene.usc.es/ snipper/). This system contains the training set and profile to classify single SNP profiles. STRUCTURE is the software with a model-based clustering method that is based on the systematic Bayesian clustering approach that applies Markov chain Monte Carlo (MCMC) (Pritchard *et al.*, 2000; Porras-Hurtado *et al.*, 2013). We estimated ancestral population using STRUCTURE 2.3.4. Applied parameters are Burnin Period: 5,000, MCMC steps: 500,000. So many iterations in the burnin process results in a progressive convergence toward reliable allele frequency estimates in each population and membership probabilities of individuals to a population (Porras-Hurtado *et al.*, 2013). Our study did not consider the mixing groups; however, admixture between populations is a common characteristic that can have ancestors from multiple populations. Therefore, Admixture/POPFLAG model can be used in conjunction with an alternative approach to USEPOPINFO and the POPFLAG model. POPFLAG considers specified information about the population of origin for a portion of the individuals to help infer the ancestry of other samples with unknown origin (Porras-Hurtado *et al.*, 2013). The other classification methods are statistical models commonly used for general classification analysis as well as the identification of Ethnic group. The optimal model may be different depending on the characteristics of the data; therefore, we applied various models because it is effective to find and apply the model with the highest accuracy through validation.

2.3. Development of a new ancestry evaluation method

We propose a modified method applied to classification as an alternative to known statistical classification methods for an ethnic group. The SNP allele associated with ancestry inference, which appears mainly in a certain ethnic group, is a major allele, while the other alleles are minor alleles. The basic concept of method, Ethnicity Index (EI), is to compare the likelihood that a person could receive the genotype from the parents of specific ethnic population, with the likelihood that the genotype could be transmitted from the parents of another ethnic background (Butler, 2009). The EI is analogous to the definition of paternity index (PI). The combined ethnicity index (CEI) was determined by multiplying individual EIs for each locus tested. The CEI is a combined index that indicates how many times more likely it is that the person belongs to the alleged population than to another ethnic background.

The CEI of our modified method can be calculated by the X/Y formula. In this case, X is the chance that the genotype could be observed in the same ethnic group, P(genotype | H_p), and Y is the chance that the genotype could be observed in other populations, P(genotype | H_d). For example, K-score is a CEI of the sample when its alleged population is Korean (KO). In this case, X is the chance that the genotype could be observed in the Korean group, and Y is the chance that the genotype could be observed in the rest of the population (African-American (AA), Caucasian (CA), Mexican-American (MA)) except for Korean. The larger the value of K-score compared to other A-score, C-score, and M-score, the more likely it is that the gene is inherited from the Korean population. Our ancestry evaluation method characterized individuals into the highest score group after calculating all ethnic scores. We proposed an ancestry evaluation method that estimate the CEI and compared its performance with classical methods for the classification of ancestries.

Table 2: Maximum and average score of the ancestry evaluation method in mixed data

Ethnic group		A-score	C-score	M-score	K-score
African-American (AA)	Maximum	1884304.2	2.4	16.8	0.1
	Average	82605.9	0.1	0.3	0.0
	Median	2876.5	0.0	0.0	0.0
Caucasian (CA)	Maximum	38.5	3694.1	55.9	8.7
	Average	0.7	305.2	2.8	0.1
	Median	0.0	56.7	0.2	0.0
Mexican-American (MA)	Maximum	31.0	494.5	30812.0	1304.7
	Average	0.4	11.8	2225.7	23.2
	Median	0.0	0.0	22.5	0.0
Korean (KO)	Maximum	0.0	3.7	11.9	173828.4
	Average	0.0	0.0	0.4	12901.4
	Median	0.0	0.0	0.0	3657.7

3. Results

3.1. Classification of ethnic groups from the mixed data including Korean samples

We compared the performances of the classical statistical classification methods, Snipper, STRUCTURE and our new ancestry evaluation method. Statistical classification methods included LDL, DLDL, DQDA, KNN, CART, SVM, RF, PMA, and PLSPA. Using our new ancestry evaluation method, we calculated four scores (A-score, C-score, M-score, and K-score) for each sample (Table 2).

For African samples, the average A-score was 82605.9 (median 2876.5), which was larger than the other scores. Note that the C-score was 0.1, M-score was 0.3, and K-score was 0.0. When we assigned each sample to the ethnic population with the maximum score, we obtained 96% accuracy in the African samples. For Caucasian samples, the average C-score was 305.2 (median 56.7), which was larger than the other scores. When we assigned each sample to the ethnic population with the maximum score, we obtained 86% accuracy in the Caucasian samples. For Mexican samples, the average M-score was 2225.7 (median 22.5), which was larger than the other scores. When we assigned each sample to the ethnic population with the maximum score, we obtained 77% accuracy in the Mexican samples. Discrimination analyses excepting MA were also performed because of a high misclassification error rate due to the misclassification of MA individuals. The reason for elevating the error rate was MA, and thus more MA specific SNPs were needed to discriminate MA. If the study were to be conducted by adding SNP capable of discriminating MA, the total error rate would also be expected to decline. For Korean samples, the average K-score was 12901.4 (median 3657.7), which was larger than the other scores. When we assigned each sample to the ethnic population with the maximum score, we obtained 98% accuracy in the Korean samples (Tables 2, 3).

As shown in Table 3, the most accurate statistical method that had the lowest error rate among the classical classification methods was the SVM method (87.50%). This method has a classification capability of 95.58% for AA, 92.28% for KO, 91.26% for CA and 70.88% for MA. This method also had higher accuracy than Snipper (86.49%) and STRUCTURE (68.55%).

Our proposed ancestry evaluation method (89.25%) had higher accuracy than SVM in the discrimination of 98.00% for KO, 96.00% for AA, 86.00% for CA and 77.00% for MA. In particular, the capability to classify the Korea population showed a high accuracy; therefore our proposed ancestry evaluation method is more useful for discriminating the four ethnic groups. In addition, it will be useful to discriminate between native people and foreigners in Japan, China, and Vietnam that have similar SNP distributions to Korea since the discrimination of Koreans has a high accuracy. Note that

Table 3: Accuracy of the various classification methods and the ancestry evaluation method in mixed data

Methods	African-American (AA)	Caucasian (CA)	Mexican-American (MA)	Korean (KO)	Total
LDL	93.48%	85.28%	64.52%	92.16%	83.86%
DLDL	93.60%	82.38%	64.84%	90.60%	82.86%
DQDA	94.60%	77.38%	77.24%	73.48%	80.68%
KNN	89.58%	84.70%	61.18%	96.06%	82.88%
CART	90.62%	84.48%	63.22%	85.24%	80.89%
SVM	95.58%	91.26%	70.88%	92.28%	87.50%
RF	94.74%	89.28%	66.18%	92.50%	85.68%
PMA	91.78%	85.66%	59.16%	92.80%	82.35%
PLSPA	94.30%	83.94%	58.76%	91.62%	82.16%
Snipper	93.80%	83.05%	75.62%	93.49%	86.49%
STRUCTURE	82.10%	50.40%	50.50%	91.20%	68.55%
Ancestry evaluation method	96.00%	86.00%	77.00%	98.00%	89.25%

LDA = linear discriminant analysis; DLDA = diagonal linear discriminant analysis; DQDA = diagonal quadratic discriminant analysis; KNN = K-nearest neighbor; CART = classification and regression trees; SVM = support vector machines; RF = random forest; NSC = nearest shrunken centroids; PLSDA = partial least squares discriminant analysis.

although not tabulated here, the SNP distribution of East Asia, such as of the Japanese, Chinese, and Vietnamese, was shown to be almost the same as the Korean SNP distribution.

3.2. Classification of ethnic groups in published HapMap data

Our proposed ancestry evaluation method was also applied to ethnic groups in published HapMap data to check if it also shows good performance in the other data sets. We compared the performances of the classical statistical classification methods, Snipper, STRUCTURE and our new ancestry evaluation method. Using our new ancestry evaluation method, we calculated four scores (AF-score, CSA-score, EA-score, and EU-score) for each sample (Table 4).

For African samples, the average AF-score was 2826.4, which was larger than other scores. Note that the CSA-score was 1.3, EA-score was 119.4, and EU-score was 0.3. When we assigned each sample to the ethnic population with the maximum score, we obtained 87.80% accuracy in the African samples (Table 5). For Central South Asian samples, the average CSA-score was 19.9. Unlike previous results, the average EA-score and EU-score is higher than the CSA-score in this case. This result is because some people have a very high EA-score (max 6267.6) or EU-score (max 3383.8) despite being actually Central-South Asian. The CSA score of the Central South Asia population was the highest when compared with the median score (8.9). So, even though the Central South Asian group did not have best CSA-score, when we assigned each sample to the ethnic population with the maximum score, we obtained 70.71% accuracy in the Central South Asian samples. For East Asian samples, the average EA-score was 4611.0, which was larger than the other scores. When we assigned each sample to the ethnic population with the maximum score, we obtained 94.76% accuracy in the East Asian samples. For European samples, the average EU-score was 1474109701036.7, which was larger than the other scores. When we assigned each sample to the ethnic population with the maximum score, we obtained 99.74% accuracy in the European samples.

Table 5 shows that the most accurate statistical method with the lowest error rate among the statistical classification methods was Snipper that also had a higher accuracy than STRUCTURE. The proposed ancestry evaluation method had almost the same accuracy as the Snipper and higher accuracy than SVM. The accuracy of Snipper was 92.66% in the discrimination of the four ethnic groups, and the accuracy of the ancestry evaluation method was 92.58%. Therefore, we can suggest a new ancestry evaluation method based on the analysis of ancestry informative SNP markers as a

Table 4: Maximum and average score of the ancestry evaluation method in HapMap data

Ethnic group		AF-score	CSA-score	EA-score	EU-score
Africa (AF)	Maximum	37354.3	55.0	4206.4	46.2
	Average	2826.4	1.3	119.4	0.3
	Median	470.1	0.0	0.5	0.0
Central South Asia (CSA)	Maximum	330.8	186.3	6267.6	3383.8
	Average	6.8	19.9	95.2	50.5
	Median	0.0	8.9	0.0	0.0
East Asia (EA)	Maximum	6580.6	70.6	34896.4	0.0
	Average	107.6	4.2	4611.0	0.0
	Median	3.3	1.9	1903.6	0
Europe (EU)	Maximum	0.9	4.9	0.0	219643909574865.0
	Average	0.0	0.1	0.0	1474109701036.7
	Median	0.0	0.0	0.0	74973798.8

Table 5: Accuracy of the various classification methods and the ancestry evaluation method in HapMap data

Methods	Africa (AF, 246)	Central South Asia (CSA, 99)	East Asia (EA, 286)	Europe (EU, 380)	Total
LDL	89.39%	77.27%	91.07%	96.48%	91.36%
DLDL	88.63%	78.80%	91.04%	96.27%	91.23%
DQDA	87.11%	75.69%	53.01%	98.95%	80.80%
KNN	85.46%	33.53%	95.05%	97.96%	87.84%
CART	79.19%	34.76%	87.11%	93.41%	82.47%
SVM	90.78%	61.43%	92.87%	99.37%	91.76%
RF	87.54%	49.02%	92.40%	98.49%	89.30%
PMA	87.50%	23.67%	95.92%	99.40%	88.17%
PLSPA	90.52%	15.67%	94.24%	99.11%	87.54%
Snipper	88.13%	83.96%	91.15%	98.98%	92.66%
STRUCTURE	72.00%	70.60%	75.90%	92.70%	80.75%
Ancestry evaluation method	87.80%	70.71%	94.76%	99.74%	92.58%

LDA = linear discriminant analysis; DLDA = diagonal linear discriminant analysis; DQDA = diagonal quadratic discriminant analysis; KNN = K-nearest neighbor; CART = classification and regression trees; SVM = support vector machines; RF = random forest; NSC = nearest shrunken centroids; PLSDA = partial least squares discriminant analysis.

useful statistical tool for identifying the ethnic group.

4. Discussion

We used the multiplex RFMP method to obtain the frequencies of 13 SNPs in mixed data that included Korean samples. We described human SNP markers that can be used to reliably classify individual DNA specimens into one of the four ancestral groups. We proposed an ancestry evaluation method that estimates the combined ethnicity index in order to compare its performance with statistical methods that included Snipper and STRUCTURE for the classification of ancestries based on the 13 SNPs.

Various statistical classification methods (including our proposed method) were also applied to ethnic groups in published HapMap data based on the 17 SNPs to examine if our proposed ancestry evaluation method also shows good performance in the other data set.

We make no claim to insist that the ancestry evaluation method is superior to other classification methods in terms of the classification of all races; however, our method provides the highest accuracy when it comes to identifying the Korean race (98.00%) or European races (99.74%), which might be because the main markers of our method are specific for Koreans or Europeans. In addition, our method is a probabilistic analysis method considering the likelihood ratio that is different from

existing classification methods and more suitable for the classification of races. Therefore, this method is expected to be useful in the identification of a specific race or population rather than a classification of many population groups.

Acknowledgement

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF), funded by the Ministry of Science, ICT & Future Planning (2012-0009833) and supported by a research project of the Supreme Prosecutors Office, Republic of Korea (1333-304-260, 2014) for the practical use and advancement of forensic DNA analysis. It was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2017R1D1A1B03028279).

References

Altman NS (1992). An introduction to kernel and nearest-neighbor nonparametric regression, *The American Statistician*, **46**, 175–185.

Bickel PJ and Levina E (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations, *Bernoulli*, **10**, 989–1010.

Botto LD and Yang Q (2000). 5,10-Methylenetetrahydrofolate reductase gene variants and congenital anomalies: a HuGE review, *American Journal of Epidemiology*, **151**, 862–877.

Bray MS, Boerwinkle E, and Doris PA (2001). High-throughput multiplex SNP genotyping with MALDI-TOF mass spectrometry: practice, problems and promise, *Human Mutation*, **17**, 296–304.

Breiman L (1984). *Classification and Regression Trees*, Wadsworth International Group, California.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Brenner CH (1998). Difficulties in the estimation of ethnic affiliation, *American Journal of Human Genetics*, **62**, 1558–1560.

Butler JM (2009). *Fundamentals of Forensic DNA Typing*, Elsevier Science, Burlington.

Dudoit S, Fridlyand J, and Speed TP (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, **97**, 77–87.

Duffy DL, Montgomery GW, Chen W, *et al.* (2007). A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation, *American Journal of Human Genetics*, **80**, 241–252.

Evett IW, Pinchin R, and Buffery C (1992). An investigation of the feasibility of inferring ethnic origin from DNA profiles, *Journal of the Forensic Science Society*, **32**, 301–306.

Fisher RA (1936). The use of multiple measurements in taxonomic problems, *Annals of Human Genetics*, **7**, 179–188.

Frudakis T, Venkateswarlu K, Thomas MJ, *et al.* (2003). A classifier for the SNP-based inference of ancestry, *Journal of Forensic Science*, **48**, 771–782.

Graf J, Hodgson R, and van Daal A (2005). Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation, *Human Mutation*, **25**, 278–284.

Graf J, Voisey J, Hughes I, and van Daal A (2007). Promoter polymorphisms in the MATP (SLC45A2) gene are associated with normal human skin color variation, *Human Mutation*, **28**, 710–717.

Hong SP, Ji SI, Rhee H, *et al.* (2008). A simple and accurate SNP scoring strategy based on typeIIS

restriction endonuclease cleavage and matrix-assisted laser desorption/ionization mass spectrometry, *BMC Genomics*, **9**, 276.

Hwang SH, Oh HB, Choi SE, Hong SP, and Yoo W (2007). Effective screening of informative single nucleotide polymorphisms using the novel method of restriction fragment mass polymorphism, *The Journal of International Medical Research*, **35**, 827–835.

Koda Y, Tachida H, Pang M, Liu Y, Soejima M, Ghaderi AA, Takenaka O, and Kimura H (2001). Contrasting patterns of polymorphisms at the ABO-secretor gene (FUT2) and plasma $\alpha(1,3)$ fucosyltransferase gene (FUT6) in human populations, *Genetics*, **158**, 747–756.

Lowe AL, Urquhart A, Foreman LA, and Evett IW (2001). Inferring ethnic origin by means of an STR profile, *Forensic Science International*, **119**, 17–22.

Mountain JL, Knight A, Jobin M, Gignoux C, Miller A, Lin AA, and Underhill PA (2002). SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes, *Genome Research*, **12**, 1766–1772.

Nguyen DV and Rocke DM (2004). On partial least squares dimension reduction for microarray-based classification: a simulation study, *Computational Statistics & Data Analysis*, **46**, 407–425.

Pastinen T and Hudson TJ (2004). Cis-acting regulatory variation in the human genome, *Science*, **306**, 647–650.

Phillips C, Freire AA, Kriegel AK, *et al.* (2013). Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries, *Forensic Science International Genetics*, **7**, 359–366.

Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo A, and Lareu MV (2013). An overview of STRUCTURE: applications, parameter settings, and supporting software, *Frontiers in Genetics*, **29**, 1–13.

Pritchard JK, Stephens M, and Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Rosenberg N, Murata M, Ikeda Y, Opare-Sem O, Zivelin A, Geffen E, and Seligsohn U (2002). The frequent 5,10-methylenetetrahydrofolate reductase C677T polymorphism is associated with a common haplotype in whites, Japanese, and Africans, *American Journal of Human Genetics*, **70**, 758–762.

Schafer AJ and Hawkins JR (1998). DNA variation and the future of human genetics, *Nature Biotechnology*, **16**, 33–39.

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, and Ferrell RE (1997). Ethnic-affiliation estimation by use of population-specific DNA markers, *American Journal of Human Genetics*, **60**, 957–964.

Taillon-Miller P, Piernot EE, and Kwok PY (1999). Efficient approach to unique single-nucleotide polymorphism discovery, *Genome Research*, **9**, 499–505.

Tusher VG, Tibshirani R, and Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. In *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116–5121.

Vapnik VN (2000). *The Nature of Statistical Learning Theory* (2nd ed), Springer, New York.

Received January 10, 2018; Revised May 10, 2018; Accepted December 13, 2018