*J Forensic Sci*, July 2003, Vol. 48, No. 4
Paper ID JFS2002079_484
Available online at: www.astm.org

*Tony Frudakis,[1] Ph.D.; Venkateswarlu K,[1] Ph.D.; Matthew J. Thomas,[1] Ph.D.; Zach Gaskin,[1] B.S.; Siva Ginjupalli,[1] M.S.; Sitarama Gunturi,[1] Ph.D.; Viswanathan Ponnuswamy,[1] M.S.; Sivamiani Natarajan,[1] Ph.D.; and Ponnuswamy Kolathupalayam Nachimuthu,[1] Ph.D.*

# A Classifier for the SNP-Based Inference of Ancestry

**ABSTRACT:** Ancestral inference from DNA could serve as an important adjunct for both standard and future human identity testing procedures. However, current STR methods for the inference of ancestral affiliation have inherent statistical and technical limitations. In an effort to identify bi-allelic markers that can be used to infer ancestral affiliation from DNA, we screened 211 SNPs in the human pigmentation and xenobiotic metabolism genes. Allele frequencies of 56 SNPs (most from pigmentation genes) were dramatically different between groups of unrelated individuals of Asian, African, and European descent, and both observed and simulated log likelihood ratios revealed that the markers were of exceptional value for ancestral inference. Log likelihood ratios of the multilocus estimates of biological ancestry (EAE/EBA) ranged from 7 to 10, which are on par with the best of the STR batteries yet described. A linear classification method was developed for incorporating these SNPs into a classifier model that was 99, 98, and 100% accurate for identifying individuals of European, African, and Asian descent, respectively. The methods and markers we describe are therefore an important first step for the development of a practical multiplex test for the inference of ancestry in a forensics setting.

**KEYWORDS:** forensic science, DNA typing, single nucleotide polymorphism, battery, classification, ancestry, ethnicity, genotype, TYR, TYRP1, OCA2, MCIR, DCT, AP3B1, CYP3A4, CYP2C8, CYP2D6, CYP2C9, CYP1A1, AHR

Human identity testing relies on the segregation of polymorphic alleles into unique combinations in individual human beings. Because a balance of dispersive and systematic forces has shaped the genetic structure of modern-day humanity, most human polymorphisms are characterized by alleles that are unevenly distributed among the world's various populations. In the case of STR markers, interpopulation differences in allele frequencies can impact exclusion calculations (1–6), and a classifier for the inference of ancestry could more objectively delineate the appropriate reference database(s) for these calculations. Moreover, there is a critical need for genetic tests that can function in a predictive or inferential sense before suspects have been identified. For example, ancestral classification markers could be (and actually are) used to assist with the identification of remains and to guide other types of criminal investigations towards individuals that 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, creating a leverage crux for maximizing the efficacy of our criminal justice system.

Various probabilistic methods have been proposed for using interpopulation allele frequency differences to infer the ancestral origin of a DNA specimen (7–13). For example, Bayesian statistical schemes have been employed to use allele frequencies in given populations (class conditional probabilities) to calculate the posterior probability that a DNA sample was derived from an individual of each particular population. However, most STR markers currently in use (i.e., F13A, TH01, FES/FPS, and vWA) offer little power to distinguish between ancestral groups. Log likelihood values for distinguishing individuals of African from European descent average $\log_{10} r = 0.4$ per locus, and, assuming a prior probability of 50% classification in alternative, this means that wrong decisions would be made 20% of the time (12,14). Although a collection of such markers may effectively resolve ancestral origin in most cases, the statistical distributions are such that an unacceptable number (5 to 10%) of classifications are ambiguous (12). Thus, markers are needed that show more dramatic ancestral bias, or a very large collection of modestly biased markers needs to be identified. In fact, screens for STR markers of dramatic ancestral bias have already been conducted and resulted in the discovery of numerous non-CODIS loci capable of resolving individuals of European descent from those of African descent (7). Statistical–inference methods incorporating these STR markers (among other marker types) appear to be fairly robust, but there is considerable debate on their rigor (7,9,12). STR markers typically have a relatively large number of alleles (often 20 or more) with some relatively rare compared to alleles from bi-allelic markers, and population databases of inordinate sample sizes are required for precise allele frequency estimation. In contrast, bi-allelic tests (i.e., SNPs) usually involve the examination of larger numbers of loci with a simpler allelic structure. Because there are only two alleles per loci, more SNPs must be examined to obtain the same statistical power, but the frequency of minor alleles are higher, requiring fewer individuals from each population to obtain reliable allele frequency estimates. Thus, smaller reference databases can be used for SNP-based identity testing and ancestral inference calculations. In addition, the statistical power to unambiguously infer ancestral affiliations using SNP-based methods is potentially greater than with STRs because of the sheer number of SNPs that can be ana-

[1] DNAPrint Genomics, Inc., 900 Cocoanut Ave. Sarasota, FL.

Received 21 Feb. 2002; and in revised form 11 June 2002, 17 Aug. 2002, and 24 Feb. 2003; accepted 24 Feb. 2003; published 22 May 2003.

lyzed simultaneously. If recent advances in high-throughput genotyping technologies can render SNPs technically and economically more attractive for routine use, it is likely that future identity determinations will, at some level, involve SNP typing.

Although SNP-based methods appear to be the wave of the future, relatively few SNP-based human identity testing or ancestral inference products have been developed and/or published. Further, though most forensic assays are focused on the so-called "junk" DNA sequences between genes, polymorphisms with biological and/or functional relevance may represent the best targets for developing tests capable of the inference of physical characteristics, which, unlike STR profiles, are shared among individuals. In this work, we targeted pigmentation and xenobiotic metabolism genes in our search for ancestrally informative SNPs, because it is likely that these genes have been subject to unusual selective pressures over the course of human evolution. For example, higher melanin content protected our African ancestors from UV damage, while unique xenobiotic metabolism sequences were probably beneficial to our ancestors who were exposed to specific alkaloids or tannins in their diet. We show here that the human pigmentation and xenobiotic metabolism genes in fact do exhibit extraordinary ancestral diversity. In particular, alleles for 56 SNP loci within these genes can be used with a linear statistical method to comprise a "classifier" for inferring the ancestral origin of a DNA specimen with exceptional accuracy. Our results comprise what we believe to be the first SNP-based method for inferring the ethnic origin of a DNA specimen. Since size separation steps would be obviated with our method, the battery we describe may constitute a practical and economical substitute for STR testing when ancestral inference is the primary objective or when results are needed in the field. Combined with STR results, our method offers an independent method by which to validate STR-based ancestral inferences that are useful for selecting the appropriate reference database for exclusion calculations.

## Methods

### *Data Collection*

Specimens and basic biographical data were obtained from a convenient sample of individuals of self-reported African, Asian, and European descent within the state of Florida under informed consent guidelines. We offered our subjects ample opportunity to express ambiguity when self-reporting race—there was a set of boxes for reporting racial mixtures. In this study, we used only individuals that reported themselves to be part of a single racial (ancestral) group. We extracted DNA from circulating lymphocytes or buccal swabs using commercial (Promega, Madison, WI) preparation kits.

### *SNP Identification*

Vertical resequencing (sequencing the same region in many individuals) was performed by amplifying promoter, exon, flanking intron, and 3′UTR sequences from a multiethnic panel of 370 unrelated individuals for whom only ancestry was known. PCR amplification was accomplished using *pfu* Turbo, according to the manufacturer's guidelines (Stratagene). We developed a program to design resequencing primers to ensure the region of interest was amplified without co-amplification of pseudo genes or other homologous genes. This is accomplished by analyzing the sequence file of interest in tandem with all other flat files identified through BLAST searches to have homology with this sequence. The program also ensures that the maximum number of relevant regions is included in the fewest possible number of amplicons. Amplification products were subcloned into the pTOPO (Invitrogen) sequencing vector. Ninety-six insert positive colonies were grown, and plasmid DNA was isolated and sequenced using PE Applied Biosystems BDT chemistry and an ABI3700. Sequences were deposited into a commercial relational database system (iFINCH, Geospiza, Seattle, WA). The resulting sequences were aligned and analyzed using another program that we developed to align sequences (using Clustal X) within each amplification region, identify discrepancies between these sequences, and qualify the discrepancies as candidate SNPs using PHRED quality metrics. The collection of candidate SNPs identified via resequencing was augmented with candidates obtained from the NCBI:dbSNP database. We developed a java-based program to download, organize, and format candidate SNP sequences for primer design and assay formatting. Genotyping assays were formatted using the Autoprimer software (Orchid Biosciences, Princeton, NJ).

### *Genotyping*

We used a novel nested PCR approach to front-end a primer extension protocol employing a 25K SNPstream genotyping system (Orchid Biosciences, Princeton, NJ). A first round of PCR was performed on these samples using the high-fidelity DNA polymerase *pfu* turbo. Because the primers for this step were the same primers that were used for resequencing, they were known to not cross-react with other competing sequences in the genome. The resulting PCR products were checked on an agarose gel, diluted, and then used as a template for a second round of PCR incorporating phosphothionated primers. We observed a higher specificity when using this nested genotyping approach than when using a single amplification protocol, presumably because most of the genes we targeted were members of multi-gene families and because of BLAST algorithm deficiencies and public sequence database limitations (incompleteness).

### *Statistical Analysis*

To use the SNP alleles we have identified for ancestral inference, we wrote a software program for using a parametric, multivariate linear classification (14), and quadratic classification technique (15,16) with their modifications for genomics data (17,18). Under the assumption that the samples have been taken from multivariate normal distributions with different mean vectors and common variance covariance matrix, linear classification procedures introduced previously (14,19–21) can be applied. However, if the populations have different variance covariance matrices, a quadratic classification procedure should be used. We used the same scoring method as Smouse and Neel (17) used. We have given a score of 1 if the individual is homozygous for the first allele, score of ½ if the individual is heterozygous, and score 0 if the individual is homozygous for the minor allele (last allele). For the linear classification method, the pooled within-population variance-covariance matrix can be computed from:

$$S = \Sigma_{i=1}^{p} \Sigma_{j=1}^{Ni}(Y_{ij} - \mu_i)(Y_{ij} - \mu_i)'/\Sigma(N_i - 1) \qquad (1)$$

where $Y_{ij}$ is the vector of scores for the $j$th individual in the $i$th population, and $\mu_i$ and $N_i$ are the vector of means and sample size for the $i$th population. By scoring one allele only, we avoid the linear dependence problem that could lead to matrix singularity. The components for these vectors could be surrogate values for SNP alleles, each dimension of the vector representing a different locus. The components may or may not be linked to one another in

gametic disequilibrium (i.e., may or may not be part of a haplotype system). Indeed, this is a strength of the method—it is equally applicable to SNPs on different chromosomes as to those within a particular gene. The generalized distance of the *ij*th individual from the mean of the *k*th population can be computed from:

$$D^2_{ij,k} = (Y_{ij} - \mu_k)' S^{-1} (Y_{ij} - \mu_k) \text{ for } k \neq i \quad (2)$$

The vector $Y_{ij}$ is used to calculate $\mu_k$, the mean of its own population. To avoid circularity caused by this, Smouse and Neel (17) used a correction when comparing an individual with the mean of its own population:

$$D^2_{ij,i} = (N_i/(N_i - 1))^2 (Y_{ij} - \mu_i)' S^{-1} (Y_{ij} - \mu_i) \quad (3)$$

We allocate the *ij*th individual to that population for which Eq 2 or Eq 3 is minimum. The result of applying Eqs 2 and 3 is an inclusion or exclusion probability matrix for the various populations.

We also implemented a quadratic classification procedure for genetic classification, where the quadratic discriminant score for the *i*th population is:

$$D^2_{ij,k} = \ln |S_k| + (Y_{ij} - \mu_k)' S_k^{-1} (Y_{ij} - \mu_k) \text{ for } k = 1,2, \ldots g(\text{populations}) \quad (4)$$

Classification is then simply the allocation of the *ij*th individual to that population for which Eq 4 is minimum. However, in this work, we restricted our attention to the linear classification procedure because of monomorphic loci in some of the groups for some of the loci, which results in an inability to apply the quadratic method due to singularity of the matrix $S_k$ of Eq 4.

Both linear and quadratic methods can be algebraically simplified for dealing with SNP data. Kurczynski (22) provided the analytical solution for the inverse of the variance-covariance matrix, and Chakraborty (23) described the computational equations for using *n* alleles per loci (when we score *n*–1 alleles per loci). Here we derived the analytical solution to the linear discriminate function for bi-allelic loci. The *i*th individual's discriminate function can be calculated in the following way.

*Case 1.* If the individual is homozygous for the major allele:

$$D_{ij} = P^2_{j,2}/(Q_1\, Q_2) \quad (5)$$

*Case 2.* If the individual is heterozygous:

$$D_{ij} = (1/2 - P_{1,j})^2/(Q_1\, Q_2) \quad (6)$$

*Case 3.* If the individual is homozygous for minor allele:

$$D_{ij} = P^2_{1,j}/(Q_1\, Q_2) \quad (7)$$

where, $Q_1$, $Q_2$ are the global allele frequencies (average allele frequencies over all populations for major and minor alleles), and $P_{1j}$ and $P_{2j}$ are the major and minor allele frequencies in the *j*th population. $D_{ij}$ is the discriminant value of the *i*th individual in the *j*th population. For *L* loci, we repeat calculations (Eqs 5 to7), add the sum, and then calculate the discriminate value for all populations. We assign the *i*th individual to the *j*th population for which $D_{ij}$ is smallest.

## Results

To identify SNP markers useful for ancestral classification, we analyzed SNPs in the human pigmentation and xenobiotic metabolism genes *TYR, TYRP1, OCA2, MC1R, DCT, AP3B1, CYP3A4, CYP2C8, CYP2D6, CYP2C9, CYP1A1* and *AHR*. We specifically targeted SNPs in these genes expecting that their sequences had been subject to unusually strong systematic genetic forces over time (they function in dietary tolerance, physical appearance, and/or ultraviolet radiation protection). To identify novel candidate SNPs in these genes, the promoter, exon, and 3′UTR regions for each was amplified and sequenced from an ancestrally diverse pool of 370 individuals. We used these SNPs to enhance a collection obtained by mining a public database (NCBI:dbSNP); the aggregate number of candidate SNPs per gene obtained from both sources was 70. Genotypes for 175 select candidate SNP loci were obtained from 100 unrelated individuals of European descent, 100 unrelated individuals of African descent, and 30 unrelated individuals of Asian descent (different individuals than those used for resequencing). The frequencies of the minor alleles ranged from zero (unvalidated SNPs) to 48%. Approximately one half of the candidate SNPs revealed clear genotype classes with a minor allele frequency greater than 0.005 in at least one ancestral group. Fifty-six of these SNPs had genotype distributions and allele frequencies that were statistically distinct (sometimes dramatically) between the three major ancestral groups tested (individuals of Asian, African, or European descent) (Appendix I). A breakdown of the ancestral bias for the 15 best markers based on nucleotides shows that there is no relationship between the specific nucleotide composition of a genotype and its ancestral affiliation (data not shown). For example, 2/9 markers for which the A allele was informative were useful for inferring inclusion in the AI group, 4/9 in the CA group, and 3/9 in the AA group. All but three of the SNP markers analyzed had allele distributions that were in the Hardy-Wienberg Equilibrium (HWE) (data not shown). Relative to the number of SNPs tested per gene, the pigmentation genes *OCA2, TYR,* and *TYRP1* (in decreasing order) had minor alleles with frequencies that were most often distinct between the ancestral groups. The frequency of ancestrally informative SNPs of the total observed in the pigmentation genes was 85 versus 61% for xenobiotic metabolism genes and 28% for other genes (the *FDPS* and *HMGCR* genes) (Table 1). Sampling bias does not appear to be the source of these ancestrally informative SNPs, since such a mecha-

TABLE 1—*SNPs/gene with alleles differentially distributed among the ancestral groups.*

| Gene* | Validated SNPs | % Ancestrally Informative SNPs† |
|---|---|---|
| **OCA2** | 21 | 95 |
| **TYRP1** | 9 | 89 |
| **TYR** | 15 | 80 |
| CYP2D6 | 30 | 57 |
| CYP2C9 | 16 | 50 |
| CYP3A4 | 16 | 50 |
| **MC1R** | 6 | 50 |
| CYP1A1 | 14 | 50 |
| AHR | 27 | 33 |
| HMGCR | 13 | 31 |
| FDPS | 8 | 25 |
| Avg. | 16 | 56 |

* Each gene is identified by NCBI nomenclature. Pigmentation genes are shown in bold print.

† The number of SNP loci that were informative for ancestry (determined using the δ value), divided by the total number of SNP loci in each gene.

TABLE 2—*Allele frequency differences (δ) and log likelihood estimates (EAE/EBA) of biological ancestry/ethnic affiliation.*

| Marker | AA/AI –δ | EAE/EBA | CA/AI -δ | EAE/EBA | CA/AA -δ | EAE/EBA |
|---|---|---|---|---|---|---|
| 712064 | 0.54444 | 1.274 | 0.55 | 1.27498 | 0.00556 | 0.00003 |
| 664803 | 0.60556 | 1.17436 | 0.02841 | 0.00337 | 0.57715 | 0.9928 |
| 886994 | 0.51667 | 0.9449 | 0.22765 | 0.2915 | 0.28902 | 0.15849 |
| 886993 | 0.51111 | 0.92979 | 0.22765 | 0.2915 | 0.28346 | 0.1527 |
| 712058 | 0.59444 | 0.70691 | 0.73182 | 1.2717 | 0.13737 | 0.07535 |
| 712057 | 0.57778 | 0.67214 | 0.67424 | 0.95882 | 0.09646 | 0.02496 |
| 712052 | 0.52778 | 0.61509 | 0.3947 | 0.29256 | 0.13308 | 0.05646 |
| 886895 | 0.45556 | 0.39263 | 0.62462 | 0.81055 | 0.16907 | 0.07368 |
| 869772 | 0.27222 | 0.37557 | 0.01098 | 0.00519 | 0.28321 | 0.52445 |
| 217438 | 0.23889 | 0.35171 | 0.17045 | 0.09993 | 0.06843 | 0.06063 |
| 869797 | 0.25 | 0.30938 | 0.22159 | 0.25764 | 0.02841 | 0.00195 |
| 712055 | 0.25 | 0.26484 | 0.02348 | 0.00569 | 0.22652 | 0.18508 |
| 554371 | 0.32778 | 0.21146 | 0.24356 | 0.10895 | 0.08422 | 0.01666 |
| 217452 | 0.19444 | 0.21076 | 0.10795 | 0.0809 | 0.08649 | 0.02593 |
| 217485 | 0.18333 | 0.19233 | 0.625 | 1.23512 | 0.44167 | 0.38454 |
| 712037 | 0.31667 | 0.1854 | 0.19848 | 0.13821 | 0.51515 | 0.66032 |
| 217486 | 0.15556 | 0.1694 | 0.60833 | 1.21223 | 0.45278 | 0.40916 |
| 869785 | 0.15556 | 0.14835 | 0 | 0.00508 | 0.15556 | 0.22926 |
| 615926 | 0.22778 | 0.1366 | 0.1447 | 0.06193 | 0.08308 | 0.01426 |
| 217489 | 0.18889 | 0.11604 | 0.56439 | 0.73705 | 0.37551 | 0.25969 |
| 712047 | 0.21111 | 0.10499 | 0.06515 | 0.0077 | 0.27626 | 0.17004 |
| 217455 | 0.21111 | 0.09643 | 0.34432 | 0.22042 | 0.55543 | 0.60927 |
| 869769 | 0.22778 | 0.09442 | 0.30606 | 0.18041 | 0.07828 | 0.01369 |
| 869813 | 0.11667 | 0.0924 | 0 | 0.00508 | 0.11667 | 0.15376 |
| 869798 | 0.11111 | 0.08502 | 0 | 0.00508 | 0.11111 | 0.14355 |
| 756239 | 0.1 | 0.07077 | 0 | 0.00508 | 0.1 | 0.12361 |
| 886896 | 0.11667 | 0.06764 | 0.56364 | 0.83399 | 0.44697 | 0.40222 |
| 869810 | 0.07222 | 0.06558 | 0.00947 | 5.40E−04 | 0.06275 | 0.05341 |
| 756251 | 0.09444 | 0.06393 | 0.1875 | 0.19919 | 0.09306 | 0.0321 |
| 951526 | 0.09444 | 0.06393 | 0 | 0.00508 | 0.09444 | 0.11389 |
| 664793 | 0.06667 | 0.03295 | 0.02841 | 0.00337 | 0.03826 | 0.01484 |
| 712051 | 0.06667 | 0.03295 | 0 | 0.00508 | 0.06667 | 0.06826 |
| 615921 | 0.06111 | 0.02752 | 0.01136 | 9.80E−04 | 0.04975 | 0.03746 |
| 886933 | 0.09444 | 0.0267 | 0.05947 | 0.01697 | 0.15391 | 0.08743 |
| 217468 | 0.05556 | 0.0224 | 0.36932 | 0.55426 | 0.31376 | 0.31315 |
| 886937 | 0.05556 | 0.0224 | 0.07955 | 0.04659 | 0.02399 | 0.00401 |
| 951497 | 0.08889 | 0.01971 | 0.05833 | 0.01145 | 0.14722 | 0.06154 |
| 869784 | 0.1 | 0.01775 | 0.1053 | 0.01966 | 0.0053 | 0.00005 |
| 664802 | 0.05 | 0.01762 | 0 | 0.00508 | 0.05 | 0.04375 |
| 886894 | 0.06667 | 0.01216 | 0.38598 | 0.28099 | 0.45265 | 0.41206 |
| 869794 | 0.02222 | 0.01082 | 0.1428 | 0.11316 | 0.16503 | 0.21113 |
| 869745 | 0.03889 | 0.00929 | 0 | 0.00508 | 0.03889 | 0.02902 |
| 217480 | 0 | 0.00525 | 0.03409 | 0.0063 | 0.03409 | 0.02315 |
| 554353 | 0 | 0.00525 | 0.02841 | 0.00337 | 0.02841 | 0.01668 |
| 869777 | 0.05 | 0.00464 | 0.04318 | 0.00332 | 0.09318 | 0.01581 |
| 869802 | 0.02778 | 0.00384 | 0.08333 | 0.09412 | 0.11111 | 0.14355 |
| 217459 | 0.02778 | 0.00309 | 0 | 0.00508 | 0.02778 | 0.01599 |
| 712054 | 0.03333 | 0.00201 | 0.02652 | 0.00125 | 0.05985 | 0.00642 |
| 554363 | 0.01111 | 0.00127 | 0.05 | 0.04321 | 0.03889 | 0.02902 |
| 217441 | 0.02222 | 0.00111 | 0.07955 | 0.04659 | 0.05732 | 0.03325 |
| 886892 | 0.02222 | 0.00111 | 0.10227 | 0.07363 | 0.08005 | 0.05604 |
| 554368 | 0.01111 | 0.00107 | 0.02841 | 0.00337 | 0.0173 | 0.00718 |
| 886934 | 0.01111 | 0.00107 | 0.08523 | 0.05302 | 0.07412 | 0.06809 |
| 869809 | 0.01667 | 2.40E−04 | 0.05114 | 0.01857 | 0.03447 | 0.01732 |
| 217456 | 0 | 0 | 0.06856 | 0.05074 | 0.06856 | 0.05074 |
| 712043 | 0 | 0 | 0.0803 | 0.0458 | 0.0803 | 0.0458 |

nism would not explain why 80% of the SNPs found in pigmentation genes were ancestrally informative, while only 20% of those found in nonpigmentation/xenobiotic metabolism genes were informative.

Average log-likelihood ratios, which are called ethnic affiliation estimates or estimates of biological ancestry (EAE/EBAs), and δ values representing allele frequency difference between two populations are presented in Table 2 for all 56 markers. Some of the markers were better at resolving between AA and AI individuals than AA and CA or CA and AI individuals (i.e., marker 886994, Row 3, Table 2), and others were better at resolving between AA and CA individuals (i.e., Marker 217455, Row 22, Table 2), while others still were better at resolving between CA and AI individuals (i.e., Marker 217486, Row 17, Table 2).

We developed an algorithm to construct a linear classifier incorporating alleles of these SNPs. The algorithm used a representation of individual samples (individuals) as *n*-dimensional vectors (where $n$ = number of markers) and average distances between individual vectors and population (ancestry) mean vectors to compute a pooled variance-covariance matrix for each population.

Using this matrix, the algorithm binned the sample (individual) into the population for which its distance is lowest. Using the algorithm with data for all 56 markers in 208 of the 230 genotyped individuals of African (AA, $n = 90$), Asian (AI; $n = 30$), and European (CA, $n = 88$) descent (same individuals genotyped previously, no known ancestral mixtures; 22 individuals with missing data excluded), we observed high corrected probabilities of including an AA individual in the AA group ($pr = 0.98$), an AI individual in the AI group ($pr = 1.0$) and a CA individual in the CA group ($pr = 0.99$) (Table 3). It may be noted that in total, only 2 of 90 AA individuals, 1 of 88 CA individuals, and none of 30 AI individuals were misclassified by the linear classification procedure. A linear classifier incorporating the 30 and 15 strongest SNPs from the battery of 56 was capable of correct classification 96% (30 markers) and 91.1% (15 markers) of the time for AAs, 96.7% of the time for AIs (both 30 and 15 markers), and 99% (30 markers) and 98% of the time for CAs (Table 3*B* and 3*C*). Since uncorrected and corrected ancestral classification probabilities were identical for each pair-wise comparison, using any number of markers, the results indicate that the sample size of each population was reasonable. We also calculated the variance-covariance matrix using 95% of the individuals and blindly classified the remaining 5% based on this matrix 1000 times and obtained similar probabilities of correct classification, suggesting that the classifier will generalize well to other samples of the same populations (data not shown).

TABLE 3—*Linear classification probabilities using all 56, 15, or 30 markers.*

TABLE 3*A*

| 56 Markers | African (AA) Probability* | Asian (AI) Probability* | European CA) Probability* |
|---|---|---|---|
| AA ($n = 90$) | 0.9778 | 0 | 0.0222 |
| AI ($n = 30$) | 0 | 1 | 0 |
| CA ($n = 88$) | 0.0114 | 0 | 0. 9886 |

TABLE 3*B*

| 15 Best Markers | African (AA) Probability* | Asian (AI) Probability* | European (CA) Probability* |
|---|---|---|---|
| AA ($n = 90$) | 0.9111 | 0 | 0.0889 |
| AI ($n = 30$) | 0 | 0.9667 | 0.0333 |
| CA ($n = 88$) | 0.0227 | 0 | 0.9773 |

TABLE 3*C*

| 30 Best Markers | African (AA) Probability* | Asian (AI) Probability* | European (CA) Probability* |
|---|---|---|---|
| AA ($n = 90$) | 0.9556 | 0 | 0.0440 |
| AI ($n = 30$) | 0.0333 | 0.9667 | 0.0000 |
| CA ($n = 88$) | 0.0227 | 0 | 0.9773 |

* The lower of the uncorrected and corrected probabilities are shown for classification into the proper group, and the higher of the uncorrected and corrected probabilities are shown for classification into improper groups (17).

We desired to compare the linear classification method with the log-likelihood ratio approach described by Shriver et al. (7) for ancestral affiliation from STR genotypes. Given the exponential relationship between the number of loci and the number of multilocus genotypes, however, it is not possible to directly determine the distribution of log-likelihood levels when more than a few loci are used. Instead, we used a Monte Carlo simulation approach for using the 56 SNP markers to estimate the log likelihood ratios for correctly discriminating between the three ancestral groups. Specifically, we generated the distribution of ethnic affiliation estimation (EAE/EBA) log-likelihood ratios (7,12,24) and calculated their summary statistics and confidence intervals (CI). The equations used for the calculation of the EAE/EBA log-likelihood ratios are fully described in Ref 7. Using a random number generator, and the observed allele frequencies in the various populations, an individual was created in the first and second populations for a pair-wise population comparison. For this exercise, we assumed that the allele not observed has a frequency of $1/(2n+1)$, where $n$ is the sample size, and that there was linkage equilibrium among all alleles. A sample size of 200 individuals was created in each population, and each time a multi-locus EAE/EBA ancestral log likelihood ratio was calculated. We repeated this procedure 10,000 times to obtain the distribution of multi-locus EAE/EBA log-likelihood ratios for the pair-wise comparison between ancestral groups, and we repeated this experiment for each pair-wise comparison of populations (CA/AA, CA/AI, AA/AI). Simulation data for the most ancestrally informative 7, 10, and all 56 markers (markers with the greatest δ values) are presented in Table 4.

TABLE 4—*Multi-locus EAE/EBA log-likelihood ratio summary statistics for pair-wise comparisons between populations of African (AA), European (CA), and Asian (AI) descent.*

| | 7-BM* AA/CA | 10-BM† AA/CA | All 56‡ AA/CA | 7-BM* AA/AI | 10-BM† AA/AI | All 56‡ AA/AI | 7-BM* AI/CA | 10-BM† AI/CA | All 56‡ AI/CA |
|---|---|---|---|---|---|---|---|---|---|
| Min§ | 3.31 | 4.17 | 7.40 | 5.91 | 6.96 | 11.5 | 7.34 | 9.18 | 12.8 |
| Q1‖ | 3.92 | 4.87 | 8.29 | 6.71 | 7.84 | 12.5 | 8.09 | 10.1 | 14.0 |
| Mean | 4.06 | 5.03 | 8.49 | 6.91 | 8.06 | 12.8 | 8.28 | 10.3 | 14.3 |
| Median | 4.06 | 5.03 | 8.48 | 6.91 | 8.06 | 12.8 | 8.28 | 10.3 | 14.3 |
| Q3¶ | 4.20 | 5.19 | 8.68 | 7.12 | 8.28 | 13.0 | 8.47 | 10.5 | 14.5 |
| Max | 4.83 | 6.05 | 9.74 | 8.33 | 9.30 | 14.2 | 9.42 | 11.6 | 16.2 |
| S.D. | 0.22 | 0.23 | 0.29 | 0.30 | 0.32 | 0.37 | 0.28 | 0.31 | 0.36 |
| 99 CI | 3.54, 4.65 | 4.45, 5.66 | 7.77, 9.27 | 6.21, 7.72 | 7.29, 8.9 | 11.80, 13.71 | 7.59,9.03 | 9.53,11.13 | 13.38,15.25 |
| 95 CI | 3.66, 4.51 | 4.59, 5.5 | 7.93, 9.07 | 6.34, 7.52 | 7.44, 8.72 | 12.04, 13.50 | 7.76,8.85 | 9.64,10.91 | 13.57,15.00 |
| Observed.EAE/EBA | 4.01 | 4.97 | 7.93 | 6.32 | 7.44 | 10.5 | 7.60 | 9.18 | 12.1 |

* Using the seven best markers (BM) for each particular pair-wise comparison of populations.
† Using the ten best markers (BM) for each particular pair-wise comparison of populations.
‡ Using all 56 markers for each particular pair-wise comparison of populations.
§ Minimum value obtained.
‖ Estimate obtained for the first quartile of individuals.
¶ Estimate obtained for the third quartile of individuals.

The simulation results reveal that for all three of the pair-wise population comparisons, virtually all of the multi-locus EAE/EBA summary statistics increased as the number of markers increased (namely minimum, mean, median, and maximum). With respect to the standard deviations for the pair-wise comparisons, the increases in discriminatory power are significant. Mean (or median) discrimination powers for a given number of markers show that the power to resolve between AA and CA individuals is less than the power to resolve between AA and AI individuals, which in turn is less than the power to resolve between AI and CA individuals:

$$\text{AA vs. CA} < \text{AA vs. AI} < \text{CA vs. AI}$$

From the summary statistics presented in Table 4, the minimum multi-locus EAE/EBA log-likelihood ratio for the seven best markers for AA/CA is 3.3, for the ten best markers, 4.17, and for all 56 markers, 7.4. It may be noted that (using all 56 markers), the minimum multi-locus EAE/EBA value derived from the simulation study was greater than the observed multi-locus EAE/EBA for CA/AI and AA/AI pair-wise comparisons. This anomaly did not occur when we considered the seven and ten best markers, and it was likely due to the relatively large number of homozygous loci present in the Asian population.

For the purpose of comparing our linear classifier to previous results described in Ref 7, we re-calculated the linear classification probabilities for specific pair-wise–ancestry comparisons, using the 7, 10, and 20 SNPs with the most dramatic allele frequency differences (best δ and EAE/EBA log likelihood ratios) between each pair of groups (Tables 5, 6, and 7). As with the previous linear classification results, where SNP markers were selected based on their minor allele frequency differences among all three ancestral groups, the Smouse correction (17) was not observed to have a significant effect (i.e., the sample sizes imposed little bias on the probabilities).

TABLE 5—*Probabilities of correct and incorrect ancestral classification using the linear classifier with the most informative markers for the inference of European and Asian ancestry.*

TABLE 5*A*

| 7 Best Markers† | European (CA) Probability* | Asian (AI) Probability* |
|---|---|---|
| CA ($n$ = 88) | 1.0000 | 0.0000 |
| AI ($n$ = 30) | 0.1000 | 0.9 |

TABLE 5*B*

| 10 Best Markers† | European (CA) Probability* | Asian (AI) Probability* |
|---|---|---|
| CA ($n$ = 88) | 0.9333 | 0.0667 |
| AI ($n$ = 30) | 0.0 | 1.0 |

TABLE 5*C*

| 20 Best Markers† | European (CA) Probability* | Asian (AI) Probability* |
|---|---|---|
| CA ($n$ = 88) | 0.9667 | 0.0333 |
| AI ($n$ = 30) | 0.0 | 1.0 |

* The lower of the uncorrected and corrected probabilities are shown for classification into the proper group, and the higher of the uncorrected and corrected probabilities are shown for classification into improper groups (17).

† Using the best set of markers for resolving between CA and AI individuals, as determined using the δ values of Table 2.

TABLE 6—*Probabilities of correct and incorrect ancestral classification using the linear classifier with the most informative markers for the inference of European versus African ancestry.*

TABLE 6*A*

| 7 Best Markers† | European (CA) Probability* | African (AA) Probability* |
|---|---|---|
| CA ($n$ = 88) | 0.9886 | 0.0114 |
| AA ($n$ = 90) | 0.0444 | 0.9556 |

TABLE 6*B*

| 10 Best Markers† | European (CA) Probability* | African (AA) Probability* |
|---|---|---|
| CA ($n$ = 88) | 0.9667 | 0.0333 |
| AA ($n$ = 90) | 0.0114 | 0.9886 |

TABLE 6*C*

| 20 Best Markers† | European (CA) Probability* | African (AA) Probability* |
|---|---|---|
| CA ($n$ = 88) | 0.9444 | 0.0556 |
| AA ($n$ = 90) | 0.0 | 1.0 |

* The lower of the uncorrected and corrected probabilities are shown for classification into the proper group, and the higher of the uncorrected and corrected probabilities are shown for classification into improper groups (17).

† Using the best set of markers for resolving between CA and AA individuals, as determined using the δ values of Table 2.

TABLE 7—*Probabilities of correct and incorrect ancestral classification using the linear classifier with the most informative markers for inference between Africans and Asians.*

TABLE 7*A*

| 7 Best Markers† | African (AA) Probability* | Asian (AI) Probability* |
|---|---|---|
| AA ($n$ = 90) | 0.8 | 0.2 |
| AI ($n$ = 30) | 0.0424 | 0.9576 |

TABLE 7*B*

| 10 Best Markers† | African (AA) Probability* | Asian (AI) Probability* |
|---|---|---|
| AA ($n$ = 90) | 0.9778 | 0.0222 |
| AI ($n$ = 30) | 0.1 | 0.9 |

TABLE 7*C*

| 20 Best Markers† | African (AA) Probability* | Asian (AI) Probability* |
|---|---|---|
| AA ($n$ = 90) | 0.9778 | 0.0222 |
| AI ($n$ = 30) | 0.0667 | 0.9333 |

* The lower of the uncorrected and corrected probabilities are shown for classification into the proper group, and the higher of the uncorrected and corrected probabilities are shown for classification into improper groups (17).

† Using the best set of markers for resolving between CA and AA individuals, as determined using the δ values of Table 2.

Though the three-way linear classification results showed that 56 SNPs were necessary for resolving between the three ancestral groups with at least 98% accuracy, when SNPs were selected based on pair-wise resolving power, only 20 SNPs were necessary to obtain classification probabilities of 97% when attempting to resolve between CA and AI individuals (Table 5*C*), 10 SNPs were required for 97% probability when resolving between CA and AA individuals (Table 6*B*), and 20 SNPs were necessary for a 93% accuracy resolving between AA and AI individuals (Table 7*C*).

## Discussion

We have described a battery of 56 human pigmentation and xenobiotic metabolism SNPs that can be used to reliably classify an individual DNA specimen into one of three major ancestral groups. Though it appears that the discriminatory power for the 56 SNP battery is inherent to 15 especially powerful SNPs, the entire battery of 56 is necessary for accuracy levels conducive for forensic use. In terms of simulated EAE/EBA log likelihood values, the power of discrimination for this battery of 56 SNP markers (log likelihood of about 2, or 1 in 100 misclassification rate) appears to be similar to that of previously described STR collections (7). Though one might expect that, given the nature of the problem and differences in variance/covariance matrices between the populations we have studied, quadratic discriminate methods would be more appropriate than linear. However, use of the quadratic method led to matrix singularity problems because, given our sample size, some of the most powerful markers had frequencies that were too low to be detected in at least one population. Rather than introduce measurement error by assuming a minimum frequency of $(1/2n+1)$ in these populations, we opted to use the linear technique instead, and the results were generally satisfactory. In addition, we presented simulated log likelihood ratios as calculated (and criticized) by others, but we did so only to facilitate a direct comparison of marker strength with those presented by Shriver et al. (7). The values we obtained using the linear discriminate method suggested that about 1% of the cases would be unresolvable with our battery, but the average simulated EAE/EBA from our work was about 10 (which would correspond to a misclassification rate significantly lower than 1%). However, these EAE/EBAs are likely to be gross overestimates. First, the SNPs we are using come from a small number of genes, which would imply the possibility, indeed the probability, that several are linked to one another in gametic disequilibrium. In fact, LD calculations for several reveal this to be the case, and, as such, the log likelihood values are not strictly additive as we have presented (meaning our log likelihood EAE/EBAs are overestimates). Second, the log likelihood ratios are derived from simulations. Whether one uses a Gibbs sampler or a Monte Carlo approach such as we employed, simulation is probably not the best approach for the estimation of EAE/EBAs from our data because a number of loci were monomorphic in one or more groups. For the simulation, we addressed this problem by assuming a minimum frequency of $(1/2n+1)$ for unobserved alleles, but this adds to estimation error, the impact of which may be most acute for those markers that are the most powerful (those for which the minor allele is rare in some groups but frequent in others). This leads to an overestimation of the log likelihood EAE/EBAs. Thus, though the log likelihood method is useful for ascribing value to particular markers or marker sets for ancestral inference, as others have pointed out previously (12), it is probably not best suited for predictions of classification accuracy. Therefore, we conclude that, though our SNP battery shows a theoretical power for EAE/EBA that is similar to previously reported STR batteries (before criticism by 12), its true accuracy as practically and realistically demonstrated with the linear classifier is closer to 99% (linkage between markers and ancestral mixtures notwithstanding). This gives a log likelihood EAE/EBA of about 2. Though not validated with actual classifications, the best (criticized) estimates obtained for STR markers also give a log likelihood of about 2 for the distinction between individuals of European and African ancestry (7,9,12). Thus, the classification accuracy of our SNP classifier rival the (criticized) projections obtained from previous STR data, though, as one would expect from their simple-allelic structure, more SNPs (56) are required to attain this power of resolution than STRs (6,10). Ultimately, we expect that blind sample classifications, not simulations, will be required to learn the true accuracy of both methods.

Given uncertainties with self-reported (or other) ancestry determinations (i.e., were any unreported or unperceived mixtures present?), the accuracy rate we report herein (99%) seems to be adequate and realistic. In fact, due to ancestral mixture and reporting uncertainty, one might effectively argue that it is unreasonable to expect any classifier for the inference of major ancestral affiliation to test better than a log likelihood of 2 in discrimination power (1 in 100 misclassified) (25). However, as promising as these results appear to be, there remain several other issues of a more practical nature that need to be solved before they will be of practical forensics use. For example, the methods and markers we have described are relevant only for the inference of major ancestral affiliation, but many populations have significant levels of admixture. Thus, a test for the inference of ancestral proportions in individuals may be more useful than the method described here for the inference of majority ancestral affiliation (though using our markers with other statistical methods for this purpose would seem relatively straightforward). In addition, over 70% of the crime scene samples received by the average forensics laboratory contain a mixture of two donors. When one of the donors is known, data for the second can be obtained through the process of subtraction and major ancestral affiliation can be inferred using the methods and markers described herein. However, for the panel to be useful in cases where both are unknown, other statistical methods for inference will be required. Thus, the markers and methods we have described are merely a first step towards the development of an efficient and resilient multiplex-based system for the inference of ancestry in a practical forensics setting. Based on our results, and the observations on unusually high ancestrally informative SNP frequencies in the pigmentation and xenobiotic metabolism genes, it seems that the markers we have described herein are well suited to be part of such an efficient and resilient system. Nonetheless, in the present form, the SNP battery we have identified may be a good replacement or compliment to existing STR methods for the inference of majority ancestry. In particular, our battery could be useful in cases where STR-based inferences are not statistically satisfying or where sample integrity is a problem (in which case, STR or RFLP tests are less useful due to the length of their amplification/digestion targets). Until previous works described how STR markers could be used for ancestral profiling (2,7,8), DNA testing was merely a quantitative tool capable of producing numeric "bar-codes" for matching specimens with individuals. The classifier we describe here is one of a handful of forensics tools for the inference of ancestry, and the very first SNP-based method for this purpose that we know of.

## References

1. Budowle B, Shea B, Niezgoda S, Chakraborty R. CODIS STR loci data from 41 sample populations. J Forensic Sci 2001 May;46(3):453–89.

2. Levadokou EN, Freeman DA, Budzynski MJ, Early BE, McElfresh KC, Schumm JW, et al. Allele frequencies for fourteen STR loci of the PowerPlex 1.1 and 2.1 multiplex systems and Penta D locus in Caucasians, African-Americans, Hispanics, and other populations of the United States of America and Brazil. J Forensic Sci 2001 May;46(3): 736–61.
3. Budowle B, Monson KL. Greater differences in forensic DNA profile frequencies estimated from ancestral groups than from ethnic subgroups. Clin Chim Acta 1994 July;228(1):3–18.
4. Kersting C, Hohoff C, Rolf B, Brinkmann B. Pentanucleotide short tandem repeat locus DXYS156 displays different patterns of variations in human populations. Croat Med J 2001 Jun;42(3):310–4.
5. Meyer E, Wiegand P, Brinkmann B. Phenotype differences of STRs in 7 human populations. Int J Legal Med 1995;107(6):314–22.
6. Gallo JC, Thomas E, Novick GE, Herrera RJ. Effects of subpopulation structure on probability calculations of DNA profiles from forensic PCR analysis. Genetica 1997;101(1):1–12.
7. Shriver MD, Smith MW, Lin J, Marcini A, Akey JM, Deka R, et al. Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 1997 Apr;60(4):957–64.
8. Lowe AL, Urquhart A, Foreman LA, Evett IW. Inferring ethnic origin by means of an STR profile. Forensic Sci Int 2001 Jun;119(1):17–22.
9. Erikson B, Svensmark O. DNA polymorphism in Greenland. Allele and profile frequencies in a Greenland population sample using the VNTR probes MS1, MS31, MS43a and YNH24. Int J Legal Med 1994;106(5): 254–7.
10. Evett IW, Pinchin R, Buffery C. An investigation of the feasibility of inferring ethnic origin from DNA profiles. J Forensic Sci 1992 Oct;32: 301–6.
11. Brenner CH. Probable ancestry of a stain donor. Proceedings from the Seventh International Symposium on Human Identification. Madison, WI: Promega, 1996;4852.
12. Brenner CH. Difficulties in the estimation of ethnic affiliation. Am J Hum Genet 1998 Jun;62(6):1558–60.
13. Smith MW, Lautenberger JA, Shin HD, Chretien JP, Shrestha S, Gilbert DA, et al. Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. Am J Hum Genet 2001 Nov;69(5):1080–94.
14. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eug 1936;7:179–88.
15. Anderson TW. Introduction to multivariate statistical analysis. New York: Wiley, 1958.
16. Srivastava MS, Khatri CG. An introduction to multivariate statistics. Amsterdam; North Holland, 1979.
17. Smouse PE, Neel JV. Multivariate analysis of gametic disequilibrium in the Yanomama. Genetics 1977 Apr;85(4):733–52.
18. Spielman, RS, Smouse, PE. Multivariate classification of human populations. Allocation of Yanomama Indians to villages. Am J Hum Genet 1976; Jul28;(4):317–31.
19. Rao CR. The utilization of multiple measurements in problems of biological classification [with discussion]. JRSS(B) 1948;10:159–203.
20. Rao CR. The problem of classification and distance between two populations. Nature 159: 30–31;1947;160: 835–6.
21. Smith CAB. Some examples of discrimination. Ann of Eugen 1948;13: 272–82.
22. Kurczynski TW. Generalized distance and discrete variables. Biometrics 1970;26:525–34.
23. Chakraborty, R. Multiple alleles and estimation of genetic parameters: computational equations showing involvement of all alleles. Genetics 1992;130:231–43.
24. Shriver MD, Smith MW, Lin J. Reply to Brenner. Am J Hum Genet 1998;Jun;62:1560–1.
25. Goodman AH. Why genes don't count (for ancestral differences in health). Am J Public Health 2000 Nov;90(11):1699–702.

Additional information or reprints requests:
Tony Frudakis, Ph.D.
DNAprint Genomics, Inc.
900 Cocoanut Ave.
Sarasota, FL 34236
E-mail: tfrudakis@dnaprint.com

# APPENDIX I

IUB codes indicate degenerate SNP sequences. Brackets indicate an insertion/deletion polymorphism.

217485
AACCTTTTCAAATTAATGTTCCAGTTTGAAGAC M
CAATCAAATATATTATTTAGTCAACATTTTGTC
TTTTTATTTTTATCTTCCTTTCCAAATAGGTCGG
GAGTTTAGTGTACCTGAGAT

217486
TATCAGCCTTTTATGTATTTTCCAAGTAAAATA W
T T A A C A T A T T A Y T T C A T T G G T C T T C T T
ACTGATCAGTATCAATGCTATGCTG[AAGA]AT
ATGAAAAACTCCAGAATCCTAATCAGT

217487
TTTATCTGGTTCTATATGAATGCTATTTTTTCCC [AATT]
TTCTCTTCTAACATGAAATATATTTTTTGAATA
TAATAGATTGAGTTATTAACTGTATTTTCTTTC
ACTTTATTACCTTCTTTCTA

217489
AGAAACAAGTTTAAGTTATGTATCCCTGATTGG Y
TTACTGGGTTTTCCTATATTCAAAAATATTAAT
TAAAGAA[AATT]AATTAATTATGTGTAGTTATA
AACCAATGAAATTTTGATTA

217468
AGGTCAGCACCCCACAAATCCTAACTTACTCA M
GCCCAGCATCATTCTTCTCCTCTTGGCAATCAC
TGTAGTAGTAGCTGGAAAGAGAAATCTGTGAC
TCCAATTAGCCAGTTCCTGCAGA

217473
GAACACTTTAAATCCTGAAAGTGCATTATAATC R
CTTAATTTATTACCAGTTTATTATCACTATTTTT
GAAGTATAAAGAATATATTCAACATCTTTCCAT
GTCTCCAGATTTTAATATAT

217480
CACATTTTTATTCTCTTCAGAAAGGATGATATT R
CCCCTTTATTTTACATTTCTGCTCCAATCCCATT
TTTCTGATGAAGAAACTGAGGCTTTGGAGTATT
AGGTGTAACTTTCCCAAGCT

217438
ACCTCCCTGGTCCCCGTTTGTCAAAGAGGATGG Y
ACTAAATGATCTCTGAAAGTGTTGAAGGGAGA
GGGTGTGAGGGCAGATCTGGGGGTGCCCAGAT
GGAAGGAGGCAGGCATGGGGGAC

217439
ACCTCCCTGGTCCCCGTTTGTCAAAGAGGATGG Y
ACTAAATGATCTCTGAAAGTGTTGAAGGGAGA
GGGTGTGAGGGCAGATCTGGGGGTGCCCAGAT
GGAAGGAGGCAGGCATGGGGGAC

217441
ACCTCCCTGGTCCCCGTTTGTCAAAGAGGATGG Y
ACTAAATGATCTCTGAAAGTGTTGAAGGGAGA
GGGTGTGAGGGCAGATCTGGGGGTGCCCAGAT
GGAAGGAGGCAGGCATGGGGGAC

217452
TAGCGTGTCCCTCTCTCTAGGTAGAAAGGGAA Y
CCATACAGGAATATTTGCTGAATCTTGGCCTAT
GTCTCACGCCTGCTGCCTGTGCTCACTGCTCTT
CCAGCTGTGATATTGGGCGTTG

217455
TTGCCCAAGAACCATGCTAGAGGTATGAACTA R
ACAAGCTACAGCATTGAAGAGTACTTTTCAAG
CAGCTTCCCTTAGATGGCACGTTGGTGGTAGCT
GTATGTGTCTGTGGGGTGTCCAG

217456
CATTCCAGTCCAGCTCGTGTCTGCTTTGTGTGA R
CTGCAGTACATGCTACAAGCAGTGGGGCCAGA
ATACCGATGGCATTACGGGACTGAGGGTCATC
ACCTTGTGACAAATTAACCATCA

217459
GGTGGGCAGCCTGCCCTGGGAAGAAGGGCGCC K
TTTCCTTTTGGTTTCCTGGGCAGGAGGGGGTTT
CCTTGTAACACAGTACTTTGCCATTTTCTTTCA
AGTTCGAGAGGTTACATTTTTC

217460
TTAACCAGCTTTACCTTAGCCACTGAGAGATTT M
CTGACAGCACTGCGTATTTGTTTTTTTAAAATT
AAGCCAATCTATAGTGAAAGAAAAGAGATGAA
TGGTTTACTGGGAGTGTGGGGG

554363
AAAGCAATGTGGTAGTTCCAACTCGGGTCCCC Y
TGCTCACGCCCTCGTTGGGATCATCCTCGACAT
CTCAGACATGGTCGTGGGAGAGGTGTGCCCGG
GTCAGGGGGCACCAGGAGAGGCC

554368
AAAGCAATGTGGTAGTTCCAACTCGGGTCCCC M
TGCTCACGCCCTCGTTGGGATCATCCTCGACAT
CTCAGACATGGTCGTGGGAGAGGTGTGCCCGG
GTCAGGGGGCACCAGGAGAGGCC

554370
GCCTGCAGCTGGCCTGGACGCCGGTGGTCGTG R
CTCAATGGGCTGGCGGCCGTGCGCGAGGGGGA
GGCAGGGGGTCCACTTGATGTCGAGACTGCAG
TGAGCCATGATCCTGCCACTGCAC

554371
CTGGGCAGAGAGGGCGCGGGGTCGTGGACATG Y
AAACAGGCCAGCGAGTGGGGACAGCGGGAAC
GTTCCCACCAGATTTCTAATCAGAAACATGGA
GGCCAGAAAGCAGTGGAGGAGGACG

554353
TCACAGGTGTGTGCACAGACATAAACACATGG R
AAAAGTTTCACAAAACACTTACCATTATGTATC
ATATATAATTGTATGTGCTATACTTTTTATATG
ACTGGCAACACAGGTTTGCTTC

664784
CCTAGCTCAGGAGGGACTGAAGGAGGAGTCGG R
GCTTTCTGCGCGAGGTGCGGAGCGAGAGCAGC
AGAGGGCAAAGGCCATCATCAGCTCCCTTTAT
AAGGGAAGGGTCACGCGCTCGGTG

664785
CCTAGCTCAGGAGGGACTGAAGGAGGAGTCGG Y
GCTTTCTGCGCGAGGTGCGGAGCGAGAGCAGC
AGAGGGCAAAGGCCATCATCAGCTCCCTTTAT
AAGGGAAGGGTCACGCGCTCGGTG

664793
TCCCTCTATCCAATTTATAGCAGCAGGTTTCTT M
GTCAGTACAATAGTTACCACTAACGGCAGCCA
ATCCAGACAAACATTTATATTTAAACATTTATA
TTTAAACAAAAGGCCTCTCTGA

664802
CCCAATTCTTGAAGTATTAAATATCTGTGTGTT Y
TCCAAGAGAAGTTACAAATTTTTTAAGCTGGG
ACTAGAGTCTGCACATTTAACTATGGGTGGTGT
TGTGTTTTGTGCTTAGATGGTC

664803
TTAGTTTTTCATAATTTTTTAGATAATATACAT K
ATGATCAGTGCAGTTACCTGTATGTTTTCTCCC
AAGATGGGGCAGCTCCGATGAGGAGGTGGGGC
AGCTGGAGGAAAAGGATCTTCT

712047
TAACAAAAAGTCTTCATCCCATCCCTGTCTACC Y
ATCTGCCCAGTTCTTTGCTACCTACACGAGTCT
CACTCTGTGGCCCAGGCTGGAGTGCAGTGGCT
TGATCTTGGCTCACTGCAACAT

712043
AGAGAAACCCAGAGAGTCAGAACTAGGCTTGT Y
GGACTCTATGCCTGATACATCATACCTGAGCCA
ATCCAGACAAACATTTATATTTAAACATTTATA
TTTAAACAAAAGGCCTCTCTGA

712064
AAATTTGAGGTGGTGTCACAGTCTTTTCTTTTA R
CCAAAGCTTTACCCATAGTTTTCCTTCATGAAA
ATAAAAATAAAAATAAATAAATAAATAAATGA
AAGAAAGAAAGAAAGAGAAAGG

712037
TGAGGTGAACACAAAGGGATGTTCTTCAGAGA R
TTACAGTCCAGCCCTGAAGCAACAACTAAGAT
TTTGAATCAGTAGTTCAAGGGTGGGGTTTGAG
ATTTTGCATTTCTAAATGAGCTCT

712051
GAAACAGTTAAATTATTGTCTAAAGACTTAGA W
ATCAATAGAAAGGAATGTCTGGGTCAAGGTGC
TTAGGGATGGAGGACCAGACAAGGTTAGAGGG
ACTTTGGTTCTGAGGCAGCTTCTA

712052
TGTGTTTGTGCCATTTGTATTTGATCAGCTGCT R
GGGGCACTTCTCCCTCTGACTGTGTGTTCTACC
CGCCCGGCCAAAACAGCCCCTACTGCCCCCTG
GCGGCAAGCCTGTGTACGAGGT

712058
ATGGCCAGGGTTAGAAAAGAAAGGTATAGCTG R
TGATACTCTTGCAGGCCCCAAGTTCATAATCAT
TCAGGTCATTATATGTATTTTTTTGGGAAAATA
GAGAGTGAGCACCTTTTCCAGC

712054
TTTTTCTCTTGTTCATTTAATGCCGTTGGGCTTG R
TTTGTGTTTTGTAGGATTCCTGGCGCCATTGAC
TTATTTTTAAAAATATTGCTCCATTGTCGTTTTG
TTTATATCTTGATTTTGGA

712055
CCCCAGGCTGGGCTGCCCAGATGTCTCTTCCTG R
TGGAGAGGAGTTTCAGGTCTGCAGAAGTCCAA
TTCTACATTAATTCCTCCACTATGAGCTTCCAC
AGTAACCTAATCTTACCCTGAG

712057
AGATTTCAAAGGAACCGGGCAGGGTGGGCCAG K
GTCTCCCCTGGTCCCCAAGAGCTGACCTAGATC
GTGGATAGCCCAGAGTGTCTCAGCACCCCTTTG
AGATTGTGCCCTGGGCCTCTGC

756251
CAGCTGGATGAGCTGCTAACTGAGCACAGGAT R
GACCTGGGACCCAGCCCAGCCCCCCCGATGAG
TGCAAAGGCGGTCAGGGTGGGCAGAGACGAG
GTGGGGCAAAGCCTGCCCCAGCCAA

615921
CATAGGAGGCAAGAAGGAGTGTCAGGGCCGG M
ACCCCCTGGGTGCTGACCCATTGTGGGGATTTG
CATAGATGGGTTTGGGAAAGGACATTCCAGGA
GACCCCACTGTAAGAAGGGCCTGG

615926
CTCACCCCAGCTCAGCACCAGCACCTGGTGAT Y
AGCCCCAGCATGGCTACTGCCAGGTGGGGGGG
CCTGAGACTTGTCCAGGTGAACGCAGAGCACA
GGAGGGATTGAGACCCCGTTCTGT

756239
CATCTCTAATGAGCCCTAGATTATTCCTGGTGT R
CAGGGAGATTAGGAAACACCTTCATATAACAG
AAAAACAAGCAATCAATCTCTAGTCTCGGTTC
ATACTAAGAGCCATCACCCCAACAC

809125
AAACATAGTAGTTGCTCAAAATATTTGTTAAA Y
AATATTTTTAATGTTAAAATGTAAGTATATCAC
TTGAGGTCAGAAATTAAAGACCAGTCTGGCCA
ACATGGCAAAACTCCGTCTCTACTG

869787
GTGTTAGGTATTATGACTAGTCAATTCAGTAAC K
TCCTTCAGGTAAACATGTTAATTGTCATCTGTG
TCTGGGCCTGGGACAGACCGCTGTGGCTCATC
ATCAGGGAGGGGCAGATGTGAGGC

869777
CCCTTACCCGCATCTCCCACCCCCARGACGCCC S
CTTTCGCCCCAACGGTCTCTTGGACAAATGAGT
GCAAAGGCGGTCAGGGTGGGCAGAGACGAGG
TGGGGCAAAGCCTGCCCCAGCCAAG

869784
GAGGGACTTGGTGAGGTCAGTGGTAAGGACAG R
GCAGGCCCTGGGTCTACCTGGAGATGGCTGCC
GTGAGCAACGTGATCGCCTCCCTCACCTGCGG
GCGCCGCTTCGAGTACGACGACCCTC

869785
GGGAGGGACTTGGTGAGGTCAGTGGTAAGGAC Y
AGGCAGGCCCTGGGTCTACCTGGAGATGGCTG
GCTGCTGGACCTAGCTCAGGAGGGACTGAAGR
AGGAGTCGGGCTTTCTGYGCGAGGTGYGGA

869772
TTGTATTAATAATTCTTTTAACTGAGTGGTCTG Y
TATTTTTTAAAAAGAATATGCTTGTTTAAT
CACTATTTATCTCATCTCAACAAGACTGAAAGC
TCCTATAGTGTCAGGAGAGTAGAAAGGATC

869745
TTTGTGTGAAATGTCATTTTACATATGGGTTCC Y
ATTTTGAAAGTGGTTTGGGAAGGGGGCATAAT
TAATTATCAGGCAGCAATCCACATGCACTTAA
CAGTTCTGACGTGAGAGGACAAGAAACAC

869769
CCTTAAGTCATCCTATTTTACACAAGCCAAACT M
GAGGCTCTAGGAGGTAGGAAGATAGTAGAGAC
CGAGGTTCCTCTGTCCACGCTTGGCACCAGCAG
CRGGCACTGTGCCAGGCCAGGACTGGGT

869794
TCTTGGAGAGGAGTTTTCTGGAAGAGGCATTTT S
CCCACTGGCTGAAAGAGCTAACAGAGGATTTG
AACATCACAGGCCATCTGAGTGGCAAGTATAA
TCATCATCATGTTTCTATTTAAAATTCAG

869797
TTTCTCCCTCATGACGCTGCGGAATTTTRGGAT Y
GGGGAAGAGGAGCATTGAGGACMGTGTTCAA
TGATTGATCTTGGAGAGGAGTTTTCTGGAAGA
GGCATTTTCCCACTGGCTGAAAGAGCTAACAG

869798
GCCCGCTGCCTTGTGGAGGAGTTGAGAAAAAC W
CAAGGGTGGGTGACCMTACTCCATATCACTGA
TGGTAGGTGTGCAWGTGCCTGTTTCAGCATCT
GTCTTGGGGATGGGGAGGATGGAAAACAGAGA

869802
TCCATTATTTTCCAKAAACGTTTTGATTATAAA S
GATCAGCAATTTCTTAACTTAATGGAAAAGT
TGGGAATGTAAATTTAGCATTTGAACAACCATT
ATTTAACCAGCTAGGTTGTAATGGTCAACTC

869809
CMTTGACCTTCTCCCCACCAGCCTGCCCCATGC Y
AGTGACCTGTGACATTAAATTCAGAAACTATTC
CTTTATTGAAGAGAATTTTCTCCACTTATATGT
GTACAGATTTTTCTTAATATCTGGTTTAT

869810
TTTAAACCTCTACCATCACCGGGTGAGAGAAG M
TGCATAACTCATATGTATGGCAGTTTAACTGG
TATAATGATGTTTGGATACCTTCATGATTCATA
TACCCCTGAATTGCTACAACAAATGTGCCAT

869813
TAGGTTGGTTGAATTCTGCCTCTAGGTACACCA Y
GTGAGGTACCCAAGAACTCCTCCTGGAAGATT
CTGGATGAAGGTGGCAATTTTAAGAAAAGTAA
ATACTTCATGCCTTTCTCAGCAGGTAATATA

886933
CCTTACTGGAATTTTGCAACGGGGAAAAATGT Y
CTGTGATATCTGCAYGGATGACTTGATGGGAT
AATCATTTTCAGAAATGTCTGCATAATGAGTTG
AGTTTCATTCCCTCTAATGCCTAAATGACAC

886937
TAGAAGTCATGTGTCTTGTGTTGGAATTTCACA K
GAAAATGTTTCCTAAGAAAATGTGAAAAATAC
TCCTTGGAAGATTATGATACCCTGGGAACACTT
TGTAACAGTAAGTTCCAAATGATAGCTTGG

886895
TGCTTTGTGTGACTGCAGTACATGCTACAAGCA R
GTGGGGCCTCAGAAGCTGGTGGCAGAAATGCG
TCACTAATGAAAGGCTGCCTCTGTTCTACGAGC
CTGCTCACTCTGGCTTGTACTCTCTCTGTG

886896
GAGGTGGAAGACATAGGCCTTGCTTTCCTGGA R
GATTGTGGTCTCATGGGGAGACATGTGGACAA
TGGCCAGGCATACCGGCTCTCCCGGGGACGGG
TGTGGGCCATGATCATCATGCTCTGTCTCATC

886894
TCGGAAGGAGTGGCACTGGGGATGGGGCTCTC Y
ACTGTCAACCGCTGGGCTGTCCCATCTCTCTAT
GCGTCGCCCGGAGGCTGCACACCTTCCACAGG
TACCGGGCGGGGTCCTGCTCAGACTGTGCTT

886892
GGAGGGAGAGAGAGATGCATCTCTGGCCCCTT S
AGACTCTGTGCCATGGGTCCTCAGCCCCTCCAG
CTGCAGGAGTCAGAAGGTTGTGCAGAGTAAAT
GAGCTGTGGTTTCTCTCTTACAGCATAGGAT

886934
TGTAAACAGAAGCAGAGAGTATTAATGTGGTT Y
TCTGTGATCTAGGAAATGTTGCAAGAGCCTTC
TTCTCCCTTCCTTACTGGAATTTTGCAACGGGG
AAAAATGTCTGTGATATCTGCAYGGATGACT

886993
GGGCAGGGTATACTTGCTATGTTAAGTTGTATG R
GCTCTGAGCAGCACTTTCAGCTGCTCAGTAA
AAATCCCTGGACACACATATAGGCACAAAACT
GCTAGCAAGAGGCTCCATTCAAGGAGTGAGTG

886994
GCAAGAGGAGAGCTCAACATGTACCATGCCCT M
GCTAATGCAGTCTAGTGCTGTGCTTGAATATA
TCTGCAGGGCAGGGTATACTTGCTATGTTAAGT
TGTATGGCTCTGAGCAGCACTTTCAGCTGCT

951497
TTATAAAGATAAATTAAAAGAAGGTGGATTAG R
GCAGGATACAAAAGAAAGAAAAGTAAAATAA
GTTTCATTTTTTTTTAATGAACAGGATTTGCTA
GTCCACTTACTGGGATAGCGGATGCCTCTCAA

951526
AGTTCAAGCAGTGAGACTACCTCTGTGCCAGT R
ATCCTGGGCTGTCTCTTCCCTTCACTCTTGGCA
CATTAAAAATAGACATTTTATTACAAGAGTGT
AGAGAAGGGAGACCAATAGAAGGTAATTGAA

*J Forensic Sci,* Sept. 2004, Vol. 49, No. 5
Paper ID JFS12477
Available online at: www.astm.org

# Erratum

**Erratum/Correction of** Frudakis et al. A Classifier for the SNP-based Inference of Ancestry. J Forensic Sci 2003 July;48(4):771–782.

It has come to the attention of the Journal that there is incorrect 5′ and 3′ flanking sequence appearing around the IUB variant for each SNP in Appendix 1. Below is the new/correct Appendix 1.

The Journal regrets this error. Note: Any and all future citations of the above-referenced paper should read Frudakis et al. A Classifier for the SNP-based Inference of Ancestry. [published erratum appears in J Forensic Sci 2004 Sept;49(5)] J Forensic Sci 2003 July;48(4):771–782.

217485
TACCTTCTTTCTAATACAAGCATATGTTAG M
ATTAAAGTTCTAGGCATACTTTTCAAAGCT

217486
TGGGCATTTCTAAAATGTTAAAACATAAAC W
CATTTCCATTCATGGATATTTGTCAACAGA

217487
TAAAGAAAACCACAGTTATTAATTAAAGAA [AATT]
AATTAATTATGTGTAGTTATAAACCAATGA

217489
GTATTTTCCAAGTAAAATATTAACATATTA Y
TTCATTGGTCTTCTTTTTTATCTGGTTCTA

217468
ATGTGTCAATGGATGCACTGCTTGGGGGAT M
TGAAATCTGGAGAGACATTGATTTT[TCT]GC

217473
TCCTCTGCAGTATTTTTGAGCAGTGGCTCC R
AAGGCACCGTCCTCTTCAAGAAGTTTATCC

217480
CATTTGCAAAATTGTAACCTAATACAAAGT R
TAGCCTTCTTCCAACTCAGGTAGAACACAC

217438
TTGTCGGASCTGCTGGTGAGCGGGASSAAC Y
TGCTGGAGACGGCCGTCATCCTCCTGCTGG

217439
GACCGCTACATCTCCATCTTCTACGCACTG Y
GCTACCACAGCAYCGTGACCCTGCCGYGGG

217441
CTGYGCTACCACAGCAYCGTGACCCTGCCG Y
GGGCGCSGCRASSCGTTGCGGCCATCTGGG

217452
TTCTGCAGAGAGACGGTGTCCATCAGCATC Y
GGGCCTCCYTGCAGCAGACCCAGGCTGTCC

217455
TTCTTTCCAGATCGTGCACAGAACTCTGGC R
GCCATGCTGGGTTCCCTTGCAGCACTGGCA

217456
GTGTGTGTGGCCAGGCATACCGGCTCTCCC R
GGGACGGGTGTGGGCCATGATCATCATGCT

217459
CGGGATTCTGCTCGCCAAATGCCTGACAGT K
TTGGGATTTGTTATCTTCATGTTTTTCCTC

217460
AGTGGAATGGGCAACCCTTCTGTTTTTTGC M
GCGCTCTTTGTTCTGATGGAGGTAAGATTT

554363
TCRCATGCCCTRCAYCACTGCCGTGATTCA Y
GAGGTGCAGCGCTTTGGGGACATCGTCCCC

554368
GGCYACTGCCAGGTGGGCCCASTCTAGGAA M
CCTGGCCACCYAGTCCTCAATGCCACCACA

554370
ACCCGCCCTGGCCTGACTCTGCCACTGGCA R
CACAGTCAACACAGCAGGTTCACTCACAGC

554371
CACCGGCGCCAACGCTGGGCTGCACGCTAC Y
CACCAGGCCCCCTGCCACTGCCCGGGCTGG

554353
ATTTTGCTGTAAAATTAAGCACTTGAATAG R
TAAAATTTGYATTAGTTTCTAAACTGGAGT

664784
GCCGACCGCCCGCCTGYGCCCATCACCCAG R
TCCTGGGYTTCGGGCCGCGYTCCCAAGGCA

664785
CCCGCCTGYGCCCATCACCCAGRTCCTGGG Y
TTCGGGCCGCGYTCCCAAGGCAAGCRGCGG

664793
TTGGACACAATGGATTAGGCTGATATGAC M
AAAGAGTTTGGAAAAGACCAATTAAAATA

664802
GATCACTAGCACATCATTTGGAGTGAACAT Y
GACTCTCTCAACAATCCACAAGACCCCTTT

664803
TGTGGACTACTATTTCCTTTAATTTATCTT K
CTCTCTTAAAAATAACTGCTTTATTGAGAT

712047
TGTCCTTAAAACTCTTCTCATTGCCTTAC Y
TATGATGTATTTTTAAACTGGCAAATATA

712043
CCCAATGCCCATGTTCCAGTTCAGAACTGT Y
GGGCTATTCAGGCTGTCTTCTTGGTGCAAG

712064
GTAAATGAGCTGTGGTTTCTCTCTTACAGC R
TAGGATATCTGACGGGATTCTGCTCGCCAA

712037
TGCAGTTTTACCCAATAAGGTGAGTGGATG R
TACATGGAGAAGGAGGGAGGAGGTGAAACC

712051
CAGCCTAATTACATTCCATAAATGTTGGTA W
CCTTTCCTTTGCTATCTCCGCAAATGTAAG

712052
AGAGAGAATAGACCAGACACCTAGACTTTA R
CAACATTCTCAAAGAACAAGCCATGGAAAG

712058
GTCCCTGCACGTTGCAGGGCCCGCCCTCTG R
CGGGTGTCCCTGACCACCAGCCCGCTCCTG

712054
TAATTAGGATATCCATAACCTCAAACATTT R
TTGTTACTTGTTTTGGGAACATTCCAAGTC

712055
ATCTATCATTTCCATTTGGTTCTTTTTTCT R
TATCTTTTGTTTCTTCCCATAGTTTTTCAT

712057
TGTGTGCACCAGTGTGAACTGTGTAGGTT K
TGTGTGGTCCCCTGTGCCTGCTGTATGTC

756251
GCGACCCCTTACCCGCATCTCCCACCCCCA R
GACGCCCCTTTCGCCCCAACGGTCTCTTGG

615921
GCACAGGATGACCTGGGACCCAGCCCAGCM M
CCCCGAGACCTGACTGAGGCCTTCCTGGCA

615926
ACCTCGACCACGCTGGCCTGGGGCCTCCTG Y
TCATGATCCTACATCCGGATGTGCAGCGTG

756239
AGTCTCACTATTATCCAGTGCAGTGGTGCA R
TCACAGCTCACTGTAGCCTCTAACTCCCAG

809125
GCAATTCTTAAATCTTGTGCTATGAAGAAA Y
GCTATTAATCCTTCCTATTAATGTAAACTG

869787
GAGACTCAAGGCTCTTAATAAAATCTTAAA K
CATTTGAGCTGGAGGAATACCTGGAAATCA

869777
CGCGTGGCGCGAGCAGAGGCGCTTCTCCRT S
TCCACCTTGCGCAACTTGGGCCTGGGCAAG

869784
GACTGGGGCCTCGGAAGAGCAGGATTTGC R
TAGATGGGTTTGGGAAAGGACATTCYAGG

869785
GCRTAGATGGGTTTGGGAAAGGACATTC Y
AGGAGACCCCACTGTAAGAAGGGCCTGG

869772
TGTCATTAACTTTTTAAAAATCTACCAA Y
GTGGAACCAGATTCRGCAAGAAGAACAA

869745
AATCACAGAAATTAACTTGCTGGAAATC Y
GTTCCCAATTCTTCCTTCAGCTCCAAGG

869769
TGGCAGCTCCCCAGATAACTCCCACCCC M
GCCTTAGCCCAGAGTGCCCCTCCCTCTT

869794
AATAGTAACTTCGTTTGCTGTTATCTCT S
TCTACTTTCCTAGCTCTCAAAGGTCTAT

869797
GAGGATGGAAAACAGAGACTTACAGAGC Y
CCTCGGGCAGAGCTTGGCCCATCCACAT

869798
TGGGATCTCCCTCCTAGTTTCGTTTCTC W
TCCTGTTAGGAATTGTTTTCAGCAATGG

869802
GTAAGATAATTTCTAAACTACTATTATCT S
TTAACAAATACAGTGTTTTATATCTAAAG

869809
TTGCTACAACAAATGTGCCATTTTTCTCC Y
TTTCCATCAGTTTTTACTTGTGTCTTATC

869810
ATGCTGTGGTGCACGAGGTCCAGAGATAC M
TTGACCTTCTCCCCACCAGCCTGCCCCAT

869813
TGCCTCCTCTGCAGTGGTACAATTACTCT Y
TGTACATGATCAAGAGCACTGTTCTGAAT

886933
CCAGTACCTTATTGTCTGAAGAGAGCTAA Y
AGAAATAGACTGTCAGAGAGTAGACCAAA

886937
TAAACACCTAGAATGTTCAAGGTACTCTA K
AAGTTGCTCCAGGGGAAACAGAAAGTGCC

886895
TAAGGTACGCAAAGCACCTCTGCCGTGGG R
GTTGCGGCCAGGTTCTGGCAGGCAGGGGC

886896
CCATTCTCAGGTGCATGAAAAGGTGGGGGC R
GTTGAGCCCACAGCTCACTGCATTCCAGTC

886894
TAGTACATTTTATCTAACCCTCACTGAGCT Y
TGCAGGGGGTACACAGCCGAGTTTAAGGAC

886892
GCTTTGGTTCATAGGCTTTGTCACATTCTG S
ATGGGAAGGTTTCAGAGCCTGTTCCCAGAC

886934
TGTGATCTAGGAAATGTTGCAAGAGCCTTC Y
TTCTCCCTTCCTTACTGGAATTTTGCAACG

886993
AAGATTATCCTTGTCTTCTTCTTTTCCCC R
TAGATGATCTTAGTAGCCATATTTTCAGA

886994
TTGCTTTTTGGTGAAATAATTTCCATGATT M
CTTCCTAAATATTGAATATATACACATTTA

951497
AAATAACGTGCTCATTGGATTTAAATAGA R
GGTGCCTATCAAATGTGATTTAAGTTATT

951526
CCCCGCAGACACAAGTCCCCAGCCCCTCC R
GGACAGCAATAAGGGTCTTACAAGGCCAG