

# PopAffiliator: online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile

Luísa Pereira · Farida Alshamali · Rune Andreassen · Ruth Ballard · Wasun Chantratita · Nam Soo Cho · Clotilde Coudray · Jean-Michel Dugoujon · Marta Espinoza · Fabricio González-Andrade · Sibte Hadi · Uta-Dorothee Immel · Catalin Marian · Antonio Gonzalez-Martin · Gerhard Mertens · Walther Parson · Carlos Perone · Lourdes Prieto · Haruo Takeshita · Héctor Rangel Villalobos · Zhaoshu Zeng · Lev Zhivotovsky · Rui Camacho · Nuno A. Fonseca

Received: 13 January 2010 / Accepted: 17 May 2010 / Published online: 16 June 2010  
© Springer-Verlag 2010

**Abstract** Because of their sensitivity and high level of discrimination, short tandem repeat (STR) marker systems are currently the method of choice in routine forensic casework and data banking, usually in multiplexes up to 15–17 loci. Constraints related to sample amount and quality, frequently encountered in forensic casework, will

not allow to change this picture in the near future, notwithstanding the technological developments. In this study, we present a free online calculator named *PopAffiliator* (<http://cracs.fc.up.pt/popaffiliator>) for individual population affiliation in the three main population groups, Eurasian, East Asian and sub-Saharan African, based on

Luísa Pereira and Nuno A. Fonseca contributed equally to this work.

L. Pereira (✉)  
Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP),  
R. Dr. Roberto Frias s/n,  
4200-465 Porto, Portugal  
e-mail: lpereira@ipatimup.pt

L. Pereira  
Faculdade de Medicina da Universidade do Porto,  
Porto, Portugal

F. Alshamali  
General Department of Forensic Sciences & Criminology,  
Dubai Police GHQ,  
Dubai, UAE

R. Andreassen  
Faculty of Health Sciences, Oslo University College,  
Oslo, Norway

R. Ballard  
Department of Biological Sciences,  
California State University,  
Sacramento, CA, USA

W. Chantratita  
Department of Pathology, Faculty of Medicine,  
Ramathibodi Hospital, Mahidol University,  
Bangkok, Thailand

N. S. Cho  
Department of Forensic Medicine,  
Central District Office,  
National Institute of Scientific Investigation,  
Daejeon, Republic of Korea

C. Coudray · J.-M. Dugoujon  
Laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse (AMIS), CNRS and University Toulouse III Paul Sabatier,  
Toulouse, France

M. Espinoza  
Departamento de Ciencias Forenses, Organismo de Investigación Judicial, Poder Judicial,  
Unidad de Genética Forense,  
San José, Costa Rica

F. González-Andrade  
Department of Medicine, Metropolitan Hospital,  
Quito, Ecuador

S. Hadi  
School of Forensic & Investigative Sciences,  
University of Central Lancashire,  
Preston, UK

U.-D. Immel  
Institute of Legal Medicine, Martin-Luther-University Halle,  
Halle, Germany

genotype profiles for the common set of STRs used in forensics. This calculator performs affiliation based on a model constructed using machine learning techniques. The model was constructed using a data set of approximately fifteen thousand individuals collected for this work. The accuracy of individual population affiliation is approximately 86%, showing that the common set of STRs routinely used in forensics provide a considerable amount of information for population assignment, in addition to being excellent for individual identification.

**Keywords** Online calculator · Genotype profile · Autosomal STRs · Individual affiliation

### Population affiliation

Because of their high discriminating power, microsatellites or short tandem repeats (STRs) are the preferred genetic markers used in forensic genetics. These markers are characterized by size variation of short (2–8 bp) tandem repetitive motifs, with a mutation rate of  $10^{-3}$  per loci per year. The improvement on high-throughput technologies and the need for high-quality assurance in forensic investigation led to the development of reliable commercial multiplex kits. These kits have high detection sensitivity, allowing results to be obtained from

residual and degraded samples. Additionally, as several STRs are screened in the same reaction, sample amount is conserved and the opportunity for laboratory errors and contamination is reduced. Two commercial kits are very successful in the forensic community: the AmpF/STR® Identifiler® PCR Amplification Kit from AB Applied Biosystems (Foster City, CA, USA) with 15 STR loci (CSF1P0, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D19S433, D21S11, FGA, TH01, TPOX, vWA) and the gender marker Amelogenin; the PowerPlex® 16 System from Promega (Madison, WI, USA) with 15 loci (CSF1P0, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, Penta D, Penta E, TH01, TPOX, vWA) and the gender marker Amelogenin. When used in tandem, the two kits generate information for 17 STRs and provide quality control because 13 of the STRs amplified by the kits are the same [1].

During the last decade, a large amount of allele frequency data have accumulated for these STRs at a worldwide level. In an online database reporting published data on these STR markers in the main forensic science journals [2], the last update summed up to a total of 842,826 individuals sampled on average for each of the 17 STRs, from 92 countries (2 in Australasia; 1 in North America; 14 in Central and South America; 27 in Europe;

C. Marian  
Carcinogenesis, Biomarkers and Epidemiology Program,  
Lombardi Comprehensive Cancer Center,  
Georgetown University Medical Center,  
Washington, DC, USA

A. Gonzalez-Martin  
Department Zoology and Physical Anthropology,  
Faculty of Biology,  
University Complutense of Madrid,  
Madrid, Spain

G. Mertens  
Forensic DNA Laboratory,  
Antwerp University Hospital,  
Edegem, Belgium

W. Parson  
Institute of Legal Medicine,  
Innsbruck Medical University,  
Innsbruck, Austria

C. Perone  
Núcleo de Ações e Pesquisa em Apoio Diagnóstico,  
Faculdade de Medicina,  
Universidade Federal de Minas Gerais (NUPAD/FM-UFMG),  
Belo Horizonte, Minas Gerais, Brazil

L. Prieto  
DNA Laboratory, Comisaría general de Policía Científica,  
University Institute of Research Police Sciences (IUICP),  
Madrid, Spain

H. Takeshita  
Department of Legal Medicine,  
Shimane University School of Medicine,  
Izumo, Shimane, Japan

H. Rangel Villalobos  
Instituto de Investigación en Genética Molecular,  
Centro Universitario de la Ciénega (CUCI-UdeG),  
Universidad de Guadalajara,  
Ocotlán, Jalisco, México

Z. Zeng  
Department of Legal Medicine, School of Basic Medical Sciences,  
Zhengzhou University,  
Zhengzhou, Henan, China

L. Zhivotovsky  
Institute of General Genetics, The Russian Academy of Sciences,  
Moscow, Russia

R. Camacho  
Laboratory of Artificial Intelligence and Decision Support  
(LIAAD-INESC),  
Porto, Portugal

R. Camacho  
DEI, Faculdade de Engenharia da Universidade do Porto,  
Porto, Portugal

N. A. Fonseca  
CRACS-INESC Porto LA,  
Porto, Portugal

11 in Near East; 6 in North Africa; 11 in sub-Saharan Africa; 7 in South Asia; 5 in East Asia; 8 in Southeast Asia). Unfortunately, most of these publications only report allele frequencies, which are not as informative as the genotype profiles. For instance, many classifiers in machine learning methods, as the ones applied in this work, can take into account the information of which alleles are present in the individual for each biallelic marker. Recently, authors have been advised to publish the genotype profiles along with the allele frequencies, but many forensic laboratories have ethical concerns in publishing them due to the high capacity of individual identification attained by the typing of these markers (for instance, for the AmpF/STR® Identifiler® PCR Amplification Kit, the probability that two individuals selected at random will have an identical profile is  $5.01 \times 10^{-18}$  for US Caucasians; company's information). Moreover, these publications usually do not provide information concerning ethnic group affiliation and the strategy for sample collection, which would be very useful for application in population genetic studies. Nonetheless, this is not a major concern for clearly European, African and East Asian populations.

The individual affiliation in a population group has obvious advantages in forensic genetics, namely, in the identification of a missing person or as an investigative tool. Non-recombining and uniparental transmitted markers, such as those in mitochondrial DNA and on the Y-chromosome, can be informative to ascertain the affiliation of maternal and paternal lineages, respectively, due to a high level of population structure for these markers [3, 4]. They do not allow, however, individual affiliation. Some authors have investigated the use of biallelic markers that have extreme differences in allelic frequencies between population groups, for the purpose of population affiliation (the so-called ancestry-informative-marker single nucleotide polymorphism (SNP) [5–7]). These ancestry-informative SNPs were recently shown to be evenly distributed across the genome [8]. However, these markers have very low informative power for individual identification, being almost fixed in a population, so that they may only be used as in conjunction with the common forensic systems. On the other hand, biallelic markers selected as highly polymorphic to be informative for individual identification (although always less informative than STRs [9]) are not so informative for population affiliation.

As most of the human genetic variation is observed within populations (93–95% as estimated from autosomal STRs [10]), a large set of markers, both STRs and SNPs, is traditionally considered necessary to be informative for population affiliation. For instance, a data set of 377 autosomal STRs in 1056 individuals from 52 populations [10] was used to ascertain the identification of six main genetic clusters, five of which correspond to major

geographical regions (Africa, Europe, the part of Asia south and west of Himalayas, East Asia, Oceania and the Americas). A general trend for clusteredness was noticeably smaller for 10 and 20 loci and for database sample sizes of 100 [11], but comparatively larger for 50 or more loci and database sample sizes of 250 and 500. In another work [12], an average accuracy of at least 90% required a minimum of ~60 markers, and the assignment for a historically admixed southern India sample was of only 87% even using 160 markers, while an average accuracy of 95% was attained to predict ancestral continent of origin from 50 SNPs picked up from the HapMap large data set of SNPs and informative for population affiliation [8].

A few tests on inferring ethnicity were conducted for the common forensic STR package [13], from 6 [14, 15], to 13 [16], to 15 markers [17] and up to 19 [18] STRs. These studies applied very different methods, from empirical evaluations (recalculating allelic frequencies by removing one individual at a time and using this to estimate the percentage of correct affiliation [18]) to application of Bayesian classifiers to a simulated genotype profile database (constructed from allelic frequencies [17]), and concluded in general for correct classifications rates of around 90% for 16–18 STRs (or slightly higher when comparing pairs of very distinct populations [17]). None of these works, however, provided researchers with a tool for evaluating population assignment of an individual in their daily casework investigations.

### STR database

The genotype STR database presented in this work encompasses data gathered from more than 40 different studies and contains a total of 61,212 genotype profiles, distributed by seven major geographical locations (Fig. 1): Eurasia, East Asia, Near East, North Africa, sub-Saharan Africa, North America and Central-South America. Some of these STR profiles are publicly available [19–40]. Since some publications only present allelic frequencies, we have contacted the corresponding authors. A total of 99 corresponding authors were contacted and a few of them provided the data for the STR profiles. Studies referring mixed populations (i.e., studies containing, with high probability, individuals having recent ancestors from several regions) and studies with a number of markers less than ten were excluded from analyses. The complete data set, together with the online calculator, are provided in the site <http://cracs.fc.up.pt/popaffiliator>.

It should be noted that the database is still very unbalanced: 17.00% Eurasian; 1.42% sub-Saharan African; 11.38% East Asian; 2.00% Near Eastern; 1.43% North African; 65.75% Central-South American; 1.02% North



**Fig. 1** Geographical distribution of the samples and regions considered in this work

American. To deal with this problem, some precautions were taken when performing the machine learning analysis. From the initial STR collection database, three different groupings of regions were considered, resulting in the following three data sets:

- Data set *3R*: encompassing three regions (Asia, Eurasia and sub-Saharan Africa) and including data from 14,714 individuals;
- Data set *5R*: encompassing five regions (Asia, Eurasia, sub-Saharan Africa, North Africa and Near East) with 16,090 individuals;
- Data set *7R*: encompassing all seven regions and including data from 54,267 individuals.

It is expected that as the number of regions increases from 3R to 5R to 7R, the difficulty of predicting the geographical origin of an individual also increases. This is due to the fact that 5R and 7R data sets include some regions long known as being on the path for many past human migrations, such as North Africa and Near East, and the affiliation of individuals to regions like North and Central-South America is artificial since their ancestor's recent origins is from elsewhere, namely Eurasia, East Asia and sub-Saharan Africa.

Furthermore, for each data set, the machine learning analyses were conducted in two subsets: a 'balanced data set' composed of an even distribution of individuals per

population classes considered and an 'unbalanced test data set' composed of the remaining data.

Not all of the 17 STR markers were typed in all populations. The higher percentages of missing values were observed for markers only present in one of the kits (Penta D and Penta E present in PowerPlex<sup>®</sup> 16 kit, with 90% of missing values; and D2S1338 and D19S433 from the AmpFSTR<sup>®</sup> Identifier<sup>®</sup> PCR Amplification Kit were missing in 15% of the profiles). The other 13 common markers included between 2% and 8% missing values.

### Machine learning methods

Machine learning methods aim at extracting information (knowledge) from data, by applying algorithms that allow computers to automatically construct models for data. The Weka (Waikato Environment for Knowledge Analysis) software package [41] was used in our study to discover relationships between the alleles for each marker and the geographical region. Weka contains a wide collection of data pre-processing and modeling techniques, being, therefore, a good choice to explore different modeling techniques on the data. The method for the construction of the model will be published elsewhere [42], but basically it consists in exploiting the features of several learning and meta-learning methods available in Weka (0R; 1R; DTNB;

## Enter Profile

Allele/Loci	CSF1PO	D2S1338	D3S1358	D5S818	D7S820	D8S1179	D13S317	D16S539	D18S51	D19S433	D21S11	FGA	PENTA_D	PENTA_E	TH01	TPOX	VWA
1	11	17	16	11	8	11	9	12	14	12	30	22	9	11	9.3	8	16
2	11	24	17	12	9	17	12	12	15	14	30	24	9	11	9.3	12	17

34 values expected, 34 provided, and 34 are ok (100%)

## Population Assignments

Population Group	Probability
Eurasia	97.1%
Asia	2.8%
Sub-Saharan Africa	0.1%

**Fig. 2** Output of the online calculator for a south Portuguese genotype profile based on the 17 STRs

SMO; NaiveBayes; J48; PART; DecisionStump; Multi-layerPerceptron; NBTree; RandomForest; BayesNet). These algorithms were applied to each of the data sets 3R, 5R and 7R. To handle the missing values existing in the data, we used each machine learning algorithm capability to handle such missing values.

A direct approach to analyze the data is to use the STRs markers as features. Since humans are diploid, the values of the two alleles for a given STR were ordered and designated as the first (lowest) and second (highest) values of a feature. For instance, the marker CSF1PO is associated with two features: CSF1PO-1 and CSF1PO-2. A total of 34 features were considered for each individual.

The best model to infer population affiliation was evaluated by calculating the predictive accuracy, also known as generalization accuracy. The (predictive) accuracy is the proportion of correct predictions over the whole set of instances. To estimate the accuracy of the classifiers, a tenfold cross-validation procedure was used on the balanced data set and the unbalanced test data set was used as a test set. The evaluation procedure was applied to each of the three data sets: 3R, 5R and 7R. Additionally, sensitivity testes on two variables were also undertaken: (i) the size of the training data set (subsets of 50, 100, 150, 200, 250, 300, 350, 400, 450 and 500 individuals of each class were considered) and (ii) the number of markers (6, 9, 13, 15 and 17 markers were considered).

The best model was obtained by WEKA's DTNB method with boosting. DTNB combines decision tables with naive Bayes and was applied to the data set 3R, a balanced data set with a size of 1200 individuals and 17 STRs. This model achieved an accuracy of 86.77%. This is the model implemented in PopAffiliator, the online calculator. The effect of the data set size and of the number of STRs showed that a big increment in accuracy is observed when increasing data from 150 to 450 genotype profiles per

population group, but then the increment stabilizes. We conjecture that better models can still be obtained with a reduction in the percentage of missing values for the two Penta markers (currently with 90% of missing values).

### The online calculator

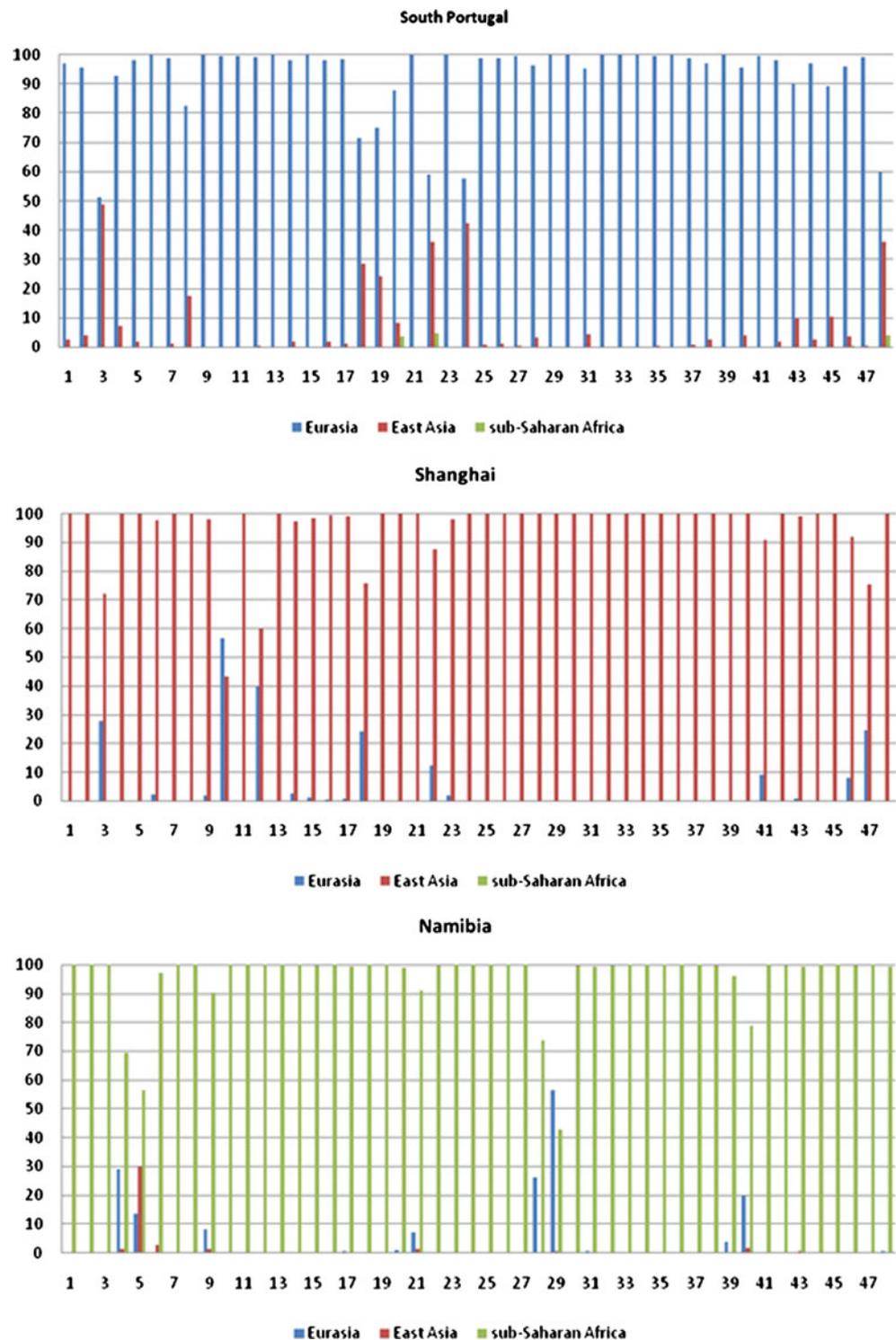
The PopAffiliator online calculator is a very simple and intuitive tool and can be freely accessed from <http://cracs.fc.up.pt/popaffiliator>. Users should insert their study profile, and the output will indicate the probability of assignment to the major population groups. The range for the allele size was restricted to the ones published in the database [http://www.cstl.nist.gov/div831/strbase/str\\_fact.htm](http://www.cstl.nist.gov/div831/strbase/str_fact.htm). Figure 2 shows an example of calculation of population assignment for a South Portuguese individual.

We further confirmed the applicability of our online calculator to 48 genotype profiles collected from three data sets included in our database: from South Portugal based on 17 STRs (this work); from Namibia for 15 STRs (except Penta markers) [43]; and from Shanghai for 17 STRs [37]. As can be seen in Fig. 3, most of the individuals belonging to each data set were affiliated in the correct population group, with a high probability. There is still the possibility that the few dubious affiliations belong to individuals resulting from mixing crossings, which cannot be confirmed.

### Conclusions

Lowe et al. [15] call the attention to the fact that “[...] as long as it is made clear that the information provided from the DNA profile is probabilistic—not a simple categorical classification—then we believe that it can provide useful

**Fig. 3** Probabilities of affiliation to the three main population groups for 48 genotype profiles collected from three data sets included in the database: South Portugal, Namibia and Shanghai



strategic guidance when set into the context of the other information available to the investigator. An indication that the offender was of Caucasian origin may be of little use in an area where the majority of the inhabitants are Caucasians but may be far more valuable in a locality where they form a minority of the population.” We agree with these authors.

Our confirmation of an 86% accuracy of individual population affiliation for the common 17 STR genotype profiles shows that this well-known forensic set of STRs has also a considerable amount of information for population assignment, besides being excellent for individual identification. We believe that our online calculator will be a valuable tool in helping forensic

researchers to predict population affiliation in a specific forensic casework. However, researchers should always be aware that this information is just a first indication, which should be confirmed by other genetic and nongenetic evidence if the population affiliation is really essential to resolve a case. This is especially true for populations that result from a high miscegenation between population groups, such as populations from the Near East or America, for which, in any case, most individuals will have a real mixed ancestry.

**Acknowledgments** IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education and is partially supported by FCT, the Portuguese Foundation for Science and Technology. CRACS-INESC Porto is supported by Programa Operacional Ciência, Tecnologia e Inovação (POCTI) e Quadro Comunitário de Apoio III. NJ and DH were supported by grant 196-1962766-2751. LZh received grants from the Russian Academy of Science for Mol & Cell Biol and FSM.

## References

- Alves C, Amorim A, Gusmão L, Pereira L (2001) VWA STR genotyping: further inconsistencies between Perkin-Elmer and Promega kits. *Int J Leg Med* 115:97–99
- Pamplona JP, Freitas F, Pereira L (2008) A worldwide database of autosomal markers used by the forensic community. *Forensic Sci Int: Genetics Supplement Series* 1:656–657
- Salas A, Bandelt HJ, Macaulay V, Richards MB (2007) Phylogeographic investigations: the role of trees in forensic genetics. *Forensic Sci Int* 168:1–13
- Jobling MA (2001) Y-chromosomal SNP haplotype diversity in forensic analysis. *Forensic Sci Int* 118:158–162
- Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A, SNPforID Consortium (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet* 1:273–280
- Phillips C, Prieto L, Fondevila M, Salas A, Gómez-Tato A, Alvarez-Dios J, Alonso A, Blanco-Verea A, Brión M, Montesino M, Carracedo A, Lareu MV (2009) Ancestry analysis in the 11-M Madrid bomb attack investigation. *PLoS ONE* 4:e6583
- Sanchez JJ, Børsting C, Balogh K, Berger B, Bogus M, Butler JM, Carracedo A, Court DS, Dixon LA, Filipović B, Fondevila M, Gill P, Harrison CD, Hohoff C, Huel R, Ludes B, Parson W, Parsons TJ, Petkovski E, Phillips C, Schmitter H, Schneider PM, Vallone PM, Morling N (2008) Forensic typing of autosomal SNPs with a 29 SNP-multiplex-results of a collaborative EDNAP exercise. *Forensic Sci Int Genet* 2:176–183
- Allocco DJ, Song Q, Gibbons GH, Ramoni MF, Kohane IS (2007) Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms. *BMC Genomics* 8:68
- Amorim A, Pereira L (2005) Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Sci Int* 150:17–21
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. *Science* 298:2381–2385
- Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW (2005) Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet* 1:e70
- Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB (2003) Human population genetic structure and inference of group membership. *Am J Hum Genet* 72:578–589
- Evett IW, Pinchin R, Buffery C (1992) An investigation of the feasibility of inferring ethnic origin from DNA profiles. *JFSS* 32:301–306
- Meyer E, Wiegand P, Brinkmann B (1995) Phenotype differences of STRs in 7 human populations. *Int J Leg Med* 107:314–322
- Lowe AL, Urquhart A, Foreman LA, Evett IW (2001) Inferring ethnic origin by means of an STR profile. *Forensic Sci Int* 119:17–22
- Fosella X, Marroni F, Manzoni S, Verzeletti A, De Ferrari F, Cerri N, Presciuttini S (2004) Assigning individuals to ethnic groups based on 13 STR loci. *Int Congr Ser* 1261:59–61
- Graydon M, Cholette F, Ng LK (2009) Inferring ethnicity using 15 autosomal STR loci -comparisons among populations of similar and distinctly different physical traits. *Forensic Sci Int Genet* 3:251–254
- Klitschkar M, Füredi S, Egyed B, Reichenpfader B, Kleiber M (2003) Estimating the ethnic origin (EEO) of individuals using short tandem repeat loci of forensic relevance. *Int Congr Ser* 1239:53–56
- Fridman C, dos Santos PC, Kohler P, Garcia CF, Lopez LF, Massad E, Gattás GJ (2008) Brazilian population profile of 15 STR markers. *Forensic Sci Int Genet* 2:e1–e4
- Brisighelli F, Capelli C, Boschi I, Garagnani P, Lareu MV, Pascali VL, Carracedo A (2009) Allele frequencies of fifteen STRs in a representative sample of the Italian population. *Forensic Sci Int Genet* 3:e29–e30
- Herrera-Paz EF, García LF, Aragon-Nieto I, Paredes M (2008) Allele frequencies distributions for 13 autosomal STR loci in 3 Black Carib (Garifuna) populations of the Honduran Caribbean coasts. *Forensic Sci Int Genet* 3:e5–e10
- Jacewicz R, Jedrzejczyk M, Ludwikowska M, Berent J (2008) Population database on 15 autosomal STR loci in 1000 unrelated individuals from the Lodz region of Poland. *Forensic Sci Int Genet* 2:e41–e43
- Juárez-Cedillo T, Zuñiga J, Acuña-Alonzo V, Pérez-Hernández N, Rodríguez-Pérez JM, Barquera R, Gallardo GJ, Sánchez-Arenas R, García-Peña Mdel C, Granados J, Vargas-Alarcón G (2008) Genetic admixture and diversity estimations in the Mexican Mestizo population from Mexico City using 15 STR polymorphic markers. *Forensic Sci Int Genet* 2:e37–e39
- Kraaijenbrink T, Zuniga S, Su B, Shi H, Xiao CJ, Tang WR, de Knijff P (2008) Allele frequency distribution of 21 forensic autosomal STRs in 7 populations from Yunnan, China. *Forensic Sci Int Genet* 3:e11–e12
- Omran GA, Ruttly GN, Jobling MA (2009) Genetic variation of 15 autosomal STR loci in Upper (Southern) Egyptians. *Forensic Sci Int Genet* 3:e39–e44
- Piatek J, Jacewicz R, Ossowski A, Parafiniuk M, Berent J (2008) Population genetics of 15 autosomal STR loci in the population of Pomorze Zachodnie (NW Poland). *Forensic Sci Int Genet* 2:e41–e43
- Sánchez-Diz P, Menounos PG, Carracedo A, Skitsa I (2008) 16 STR data of a Greek population. *Forensic Sci Int Genet* 2:e71–e72
- Sánchez-Diz P, Acosta MA, Fonseca D, Fernández M, Gómez Y, Jay M, Alape J, Lareu MV, Carracedo A, Restrepo CM (2009) Population data on 15 autosomal STRs in a sample from Colombia. *Forensic Sci Int Genet* 3:e81–e82
- Nie S, Yao J, Yan H, Yang Y, Gu T, Tang W, Li W, Wang B, Xiao C (2008) Genetic data of 15 STR loci in Chinese Yunnan Han population. *Forensic Sci Int Genet* 3:e1–e3
- Soták M, Petrejčková E, Bernasovská J, Bernasovský I, Sovicová A, Boronová I, Svicková P, Bôžiková A, Gabriková D (2008) Genetic variation analysis of 15 autosomal STR loci in Eastern

- Slovak Caucasian and Romany (Gypsy) population. *Forensic Sci Int Genet* 3:e21–e25
31. Rubi-Castellanos R, Anaya-Palafox M, Mena-Rojas E, Bautista-España D, Muñoz-Valle JF, Rangel-Villalobos H (2009) Genetic data of 15 autosomal STRs (Identifiler kit) of three Mexican Mestizo population samples from the States of Jalisco (West), Puebla (Center), and Yucatan (Southeast). *Forensic Sci Int Genet* 3:e71–e76
  32. Simms TM, Garcia C, Mirabal S, McCartney Q, Herrera RJ (2008) The genetic legacy of the transatlantic slave trade in the island of New Providence. *Forensic Sci Int Genet* 2:310–317
  33. Budowle B, Moretti TR (1999) Genotype profiles for six population groups at the 13 CODIS Short Tandem Repeat core loci and other PCR Based loci. *Forensic Sci Commun* 1
  34. Zhivotovsky LA, Veremeichyk VM, Kuzub NN, Atramentova LA, Udina IG, Kartel NA, Tsybovsky IS (2009) A reference data base on STR allele frequencies in the Belarus population developed from paternity cases. *Forensic Sci Int Genet* 3:e107–e109
  35. Zhivotovsky LA, Malyarchuk BA, Derenko MV, Wozniak M, Grzybowski T (2009) Developing STR databases on structured populations: the native South Siberian population versus the Russian population. *Forensic Sci Int Genet* 3:e111–e116
  36. Zhivotovsky LA, Akhmetova VL, Fedorova SA, Zhirkova VV, Khusnutdinova EK (2009) An STR database on the Volga-Ural population. *Forensic Sci Int Genet* 3:e133–e136
  37. Li C, Li L, Zhao Z, Lin Y, Que T, Liu Y, Xue J (2009) Genetic polymorphism of 17 STR loci for forensic use in Chinese population from Shanghai in East China. *Forensic Sci Int Genet* 3:e117–e118
  38. Andreassen R, Pereira L, Dupuy BM, Mevaag B (2009) Icelandic population data for the STR loci in the AMPFISTR®SGM Plus™ system and the PowerPlex® Y-system. *Forensic Sci Int Genet* (in press)
  39. Tillmar AO, Bäckström G, Montelius K (2009) Genetic variation of 15 autosomal STR loci in a Somali population. *Forensic Sci Int Genet* 4:e19–e20
  40. Lopes V, Serra A, Gamero J, Sampaio L, Balsa F, Oliveira C, Batista L, Corte-Real F, Vieira DN, Vide MC, Anjos MJ, Carvalho M (2009) Allelic frequency distribution of 17 STRs from Identifiler and PowerPlex-16 in Central Portugal area and the Azores archipelago. *Forensic Sci Int Genet* 4:e1–e7
  41. Witten IH, Frank E (2005) *Data Mining: practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann
  42. Fonseca NA, Camacho R, Pereira L (submitted) On the prediction of an individual affiliation to a major population group based on information from a small set of autosomal STRs—a machine learning approach
  43. Muro T, Fujihara J, Imamura S, Nakamura H, Yasuda T, Takeshita H (2008) Allele frequencies for 15 STR loci in Ovambo population using AmpFISTR Identifiler Kit. *Leg Med (Tokyo)* 10:157–159

LP delineated the project and collected the database. RC and NAF performed the machine learning analyses, interpreted results and constructed the online calculator. The remaining authors contributed genotype profiles for the database and collaborated in the improvement of the manuscript and of the online tool's output.