International Congress Series 1239 (2003) 665–671

Statistical pattern analysis of D1S80 alleles in Northwestern Russians and worldwide database using COLLAPSE software

A.G. Smolyanitsky[a], N.N. Khromov-Borisov[a,*], V.L. Popov[a], G.I. Zaslavsky[a], I.B. Rogozin[b], J.A.P. Henriques[c], T.B.L. Kist[c], H.-G. Scheil[d]

[a]*Forensic Medicine Bureau of Leningrad District, Shkapin str., 36/40, Saint Petersburg, 198092, Russia*
[b]*Institute of Cytology and Genetics, Russian Academy of Sciences, Novosibirsk, Russia*
[c]*Center of Biotechnology and Department of Biophysics, Federal University of Rio Grande do Sul, 91501-970, Porto Alegre, Brazil*
[d]*Institute of Human Genetics and Anthropology, Heinrich-Heine-University, D-40001, Düsseldorf, Germany*

Abstract

Polymorphism at the locus D1S80 was studied in 283 unrelated habitants of the Northwestern Federal Region of Russia that includes the cities Arkhangel'sk, Kaliningrad, Murmansk, Novgorod, Petrozavodsk, Pskov, Saint Petersburg and surrounding districts. The sample distribution obtained was compared with the published data on 128 worldwide samples from the DNA-PCR Databank. For this purpose a new statistical technique—Similarity Pattern Analysis (SPAN) and corresponding COLLAPSE software were used and their validity and applicability were demonstrated. The main features of the method are the following: First, to measure the similarity (homogeneity) between any pair of populations in the study, the relevant sufficient statistics, Kastenbaum–Hirotsu squared distance (*KHi*^{2}), is used. Second, the method is based on the collapsing principle, which permits similar subsets of the sample distributions to be combined (to collapse) into distinct (locally homogeneous) groups (homoclusters). Such a procedure permits the most likely (optimal) version of the similarity pattern to be revealed. Thirdly, the discrimination among homo- and hetero-clusters is based on the so-called χ2Reduction Principle, according to which the corresponding statistics, χ2reduction (CSR), as a measure of intra-cluster homogeneity should be kept non-significant after each step of collapsing. Evaluated similarities and/or dissimilarities appeared to be rather reasonable and interpretable.

Keywords: Forensic population genetics; Databases; Statistical analysis

* Corresponding author. Tel.: +7-812-389-9095.
E-mail address: Nikita@NH8333.spb.edu (N.N. Khromov-Borisov).

0531-5131/03 © 2003 Elsevier Science B.V. All rights reserved.
PII: S0531-5131(02)00477-6

1. Introduction

Several methods of multivariate analysis (cluster, factor, correspondence, etc.) are commonly used in forensic population genetics. However, most of them are descriptive (exploratory) in nature. That means they do not use statistical inference, which should be based on the logic of statistical hypothesis testing. Thus, one of the main objectives of this study was to find and/or to elaborate and evaluate the method, which would be adequate, reliable and efficient for the statistical analysis of forensic population databases.

2. Material and methods

Standardized methods of DNA typing certified and approved for the forensic studies were used. For determination of D1S80 alleles and genotypes the corresponding commercial DNA-PCR kit manufactured by Helix (Saint Petersburg, Russia) was utilized. A related allelic ladder was from Applied Biosystems (USA) and 283 unrelated Northwestern Russians were included in the study. Published data on the distribution of D1S80 alleles in 128 worldwide samples including 25,865 individuals [1,2] were involved in the comparative study. Statistical tests and estimations were conducted with the software Arlequin and GDA designed for the analysis of population genetics data [3,4]. Trees were drawn with TreeExplorer [5].

Table 1
Distribution of D1S80 genotypes evaluated in a random sample of 283 Northwestern Russians

Genotype	*n*	Genotype	*n*	Genotype	*n*	Genotype	*n*
16/22	1	18/32	1	23/29	1	25/25	1
16/24	1	18/34	2	24/24	26	25/28	5
16/29	1	18/36	1	24/25	19	25/29	1
17/20	1	18/38	1	24/26	4	25/31	3
18/18	22	19/31	1	24/27	1	26/28	2
18/19	1	20/22	4	24/28	9	26/31	2
18/20	4	20/24	3	24/29	2	26/33	1
18/22	8	20/25	1	24/30	4	28/28	2
18/24	55	21/24	3	24/31	11	28/31	4
18/25	8	22/24	8	24/32	1	29/30	2
18/26	4	22/25	3	24/33	1	29/31	2
18/28	6	22/28	4	24/34	1	30/37	1
18/29	4	22/31	4	24/36	3	31/31	2
18/30	3	22/32	1	24/37	4	31/37	1
18/31	8	22/34	2	24/40	1	Total	283

Observed heterozygosity: $H_o = 0.813$; expected heterozygosity: $H_e = 0.806$. Overall fixation index $f = -0.008$.
Results of the HWE testing: GDA software gave P=0.38 and P=0.39 for the exact tests based on χ^2 and probability, respectively.
P-values produced by Arlequin software were depended on the number of randomization steps: 10^4 steps—P=0.7815±0.0008; 10^5 steps—P=0.2681±0.0005; 10^6 steps—P=0.373±0.012; 10^{-7} steps—P=0.424±0.006; 10^8 steps—P=0.460±0.003.

3. Results

Polymorphism at the D1S80 locus was studied in 283 unrelated habitants of North-western Russia, which is a Federal Region with about 14.5 million Russian residents (more than 10% of the total population). It includes 152 cities among which are Saint Petersburg, Arkhangel'sk, Kaliningrad, Murmansk, Novgorod, Petrozavodsk, Pskov and surrounding districts. The observed genotype frequencies were in good agreement with Hardy–Weinberg equilibrium: estimated overall fixation index was close to zero, $f = -0.008$; observed and expected heterozygosities were $H_o = 0.813$ and $H_e = 0.806$ (Table 1). However, we occasionally found that the two software packages Arlequin and GDA, commonly used for the population genetics data analysis gave substantially

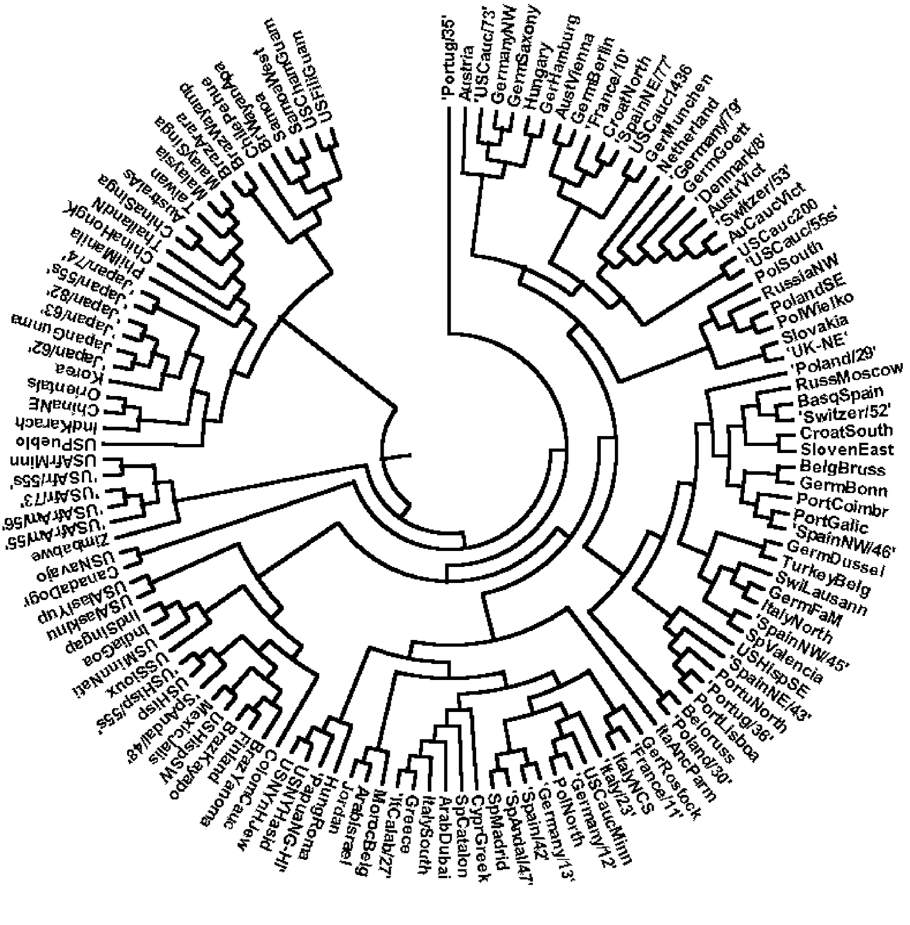

Fig. 1. Circle tree for the original data on the distribution of D1S80 alleles in 129 worldwide population samples. Drawn with TreeExplorer [5]. P-values corresponding to the observed value of KHi^2 statistic calculated with COLLAPSE software were used as a measure of pair-wise similarity between rows and/or columns in the analyzed contingency table [6].

different estimations of the exact P-values. GDA provides more stable estimations whereas estimations produced by Arlequin appeared to be depended on the number of randomization steps applied (see legend in Table 1).

The data set obtained was compared with published data on the distribution of D1S80 alleles in 128 worldwide samples including 25,865 individuals [1,2]. For this purpose, COLLAPSE software [6] was exploited. Results are presented in Fig. 1 as a circle tree drawn with the program Tree Explorer [5]. Of 129 samples, 113 were collapsed into 29 statistically non-distinguishable (internally homogeneous) blocks. Most of them seem rather reasonable as follows from the identity of the names for the sources of samples

Table 2
The 29 internally homogeneous blocks (homoclusters) revealed with COLLAPSE software in the distributions of D1S80 alleles among 129 worldwide samples

I	(Austria, (((*Germany*Northwest, *Germany*Saxony), USCaucasoids/73), (Hungary, *Germany*Hamburg)));
II	((((AustriaVienna, *Germany*Berlin), ((CroatiaNorth, SpainNortheast/77), France/10)), (USCaucasoids1436, *Germany*Munchen)), (((((Denmark/8, ((Switzerland/53, *Auatralia*CaucasoidsVictoria), *Australia*Victoria)), *Germany*Göttingen), *Germany*/79), Netherlands), (*USCaucasoids*200, *USCaucasoids*/55s)));
III	((((Basques*Spain*, Switzerland/52), RussiaMoscow), (CroatiaSouth, SloveniaEast)), ((BelgiumBrussels, GermanyBonn), (*Portugal*Coimbra, (*Portugal*Galicia, *Spain*Northwest/46))));
IV	(Byelorussia, Poland/30);
V	((Finland, *Brazil*Kayapo), *Brazil*Yanomama);
VI	((France/11, ItalyNorthCenterSouth), GermanyRostock);
VII	((((*Germany*/12, PolandNorth), USCaucasoidsMinnesotta), Italy/23), ((*Germany*/13, *Spain*/42), (*Spain*Andalusia/47, *Spain*Madrid)));
VIII	((*Germany*Düsseldorf, TurkeyBelgium), (((*Germany*Frankfurt-am-Main, ItalyNorth), SwitzerlandLausanna), (*Spain*Northwest/45, *Spain*Valencia)));
IX	(((((*Greece*, *Italy*Calabria/27), *Italy*South), ArabsDubai), SpainCatalonia), Cyprus*Greek*);
X	(*Poland*South, ((*Poland*Southeast, *Poland*Wielkopolska), RussiaNorthwest);
XI	((((*Portugal*/36, *Portugal*Lisboa), *Portugal*North), SpainNortheast/43), USHispanicsSoutheast);
XII	(Slovakia, UK-Northeast);
XIII	(SpainAndalusia/48, (*USHispanics*, *USHisanicsp*/55s));
XIV	(MoroccoBelgium, (ArabsIsrael, Jordan));
XV	(*USAfroamericans*/55, (*USAfroamericans*/56, *USAfroamericans*/73));
XVI	(USHispanicsSouthwest, MexicanJalisco);
XVII	(*USAlaska*Inupiaq, *USAlaska*Yupik);
XVIII	(*Brazil*Arara, *Brazil*Wayampi);
XIX	(BrazilWayanaApalai, (*Samoa*, *Samoa*West));
XX	(ChinaNortheast, Orientals);
XXI	(ChinaSingapore, AustraliaAsiatics);
XXII	(*Japan*/62, *Japan*Gunma);
XXIII	(*Japan*/63, *Japan*/82);
XXIV	((*Malaysia*, *Malays*Singapore), Taiwan);
XXV	((*India*Goa, *India*Singapore), USMinnesottaNatives);
XXVI	(PapuaNewGuinea, (USNYHasid*Jewish*, USNYnon-Hasid*Jewish*));
XXVII	(*USAfroamericans*/55s, *USAfroamericans*Minnesotta);
XXVIII	(*Japan*/55s, *Japan*/74);
XXIX	(USCham*Guam*, USFili*Guam*);

Identical names in a given block are emphasized in italics. Homonymous samples are differentiated with the reference numbers from [1,2].

(Table 2). Some geographical and ethnic relatedness are also obvious. This means that internal structure in such “homoclusters” can be regarded as apparent and more realistic pattern can be defined as in Fig. 2. Data on Northwestern Russians clustered with three samples from Poland population: Southern, Southeastern and Wielkopolska (Fig. 1; Table 2, block X). Historically, Poland was, for a long time, a part of the Russian Empire and the similarity revealed can reflect real genetic relatedness between this populations.

When another measure of similarity (modified Rogers distance) was used to construct the tree the relatively strong similarity between Croatians and Indians (Goa) was observed which was inconsistent with the well-known correlations. It was interpreted as an artefact resulting from the sample errors [1]. COLLAPSE software did not reveal such similarity. Moreover, significant difference was detected between two (Northern and Southern) Croatian samples (marked with arrows in Fig. 1). That means that they could not be combined into single sample.

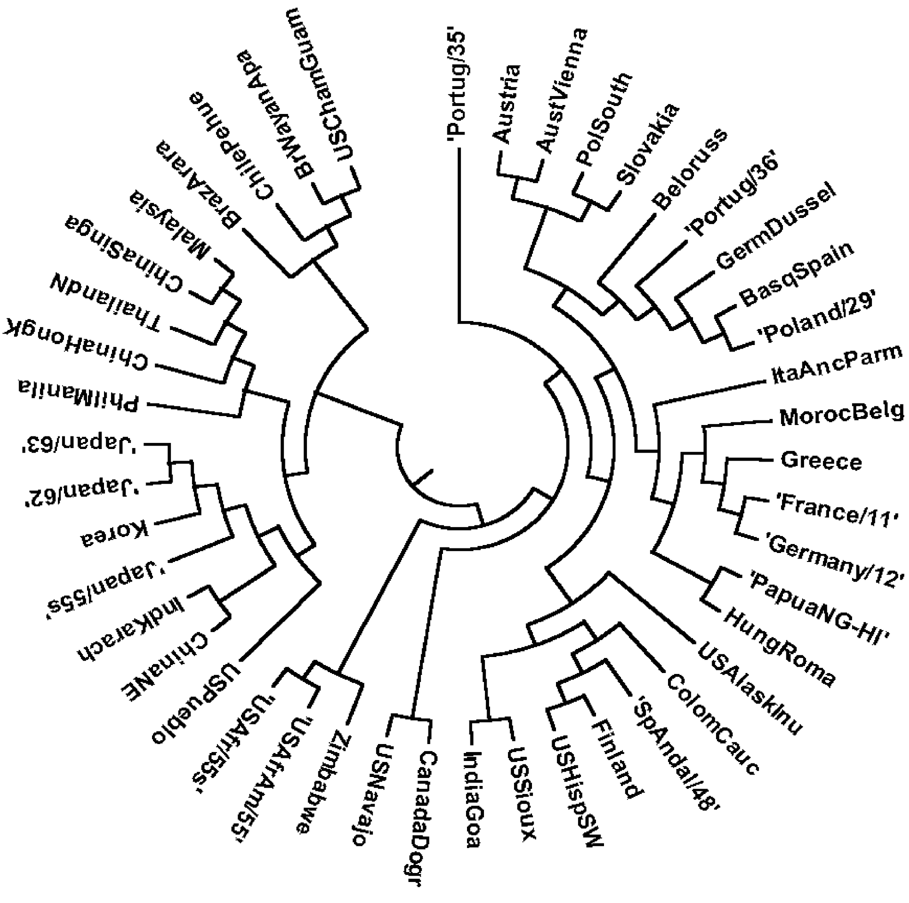

Fig. 2. Circle tree for the worldwide population samples after collapsing the homogeneous blocks. Drawn with TreeExplorer [5]. Content of homogeneous blocks is listed in Table 2.

After collapsing of homogenous parts, the similarity/dissimilarity pattern becomes more vivid and compact (Fig. 2).

4. Discussion

The procedure proposed here can be called Statistical Collapsing. The main idea of the method consists in a stepwise *collapsing of the most homogeneous parts* of analyzed data. To test homogeneity within collapsed blocks the Χ2reduction (CSR) was used as virtually the most relevant statistic [7]. Χ2Reduction (CSR) is a difference between common χ^2 calculated for the original and for the collapsed contingency tables:

$$\chi^2[\text{reduction}] = \chi^2[\text{original}] - \chi^2[\text{collapsed}]$$

$$df[\text{reduction}] = df[\text{original}] - df[\text{collapsed}]$$

By analogy with ANOVA, χ^2[collapsed] and χ^2[reduction] can be interpreted as the (integral) measures of heterogeneity *between* and *within* collapsed blocks of data, respectively. When only one pair of populations is compared CSR is reduced to the known Kastenbaum–Hirotsu squared distance (KHi^2) [6,8–10].

Similarity Pattern Analysis (SPAN) (or Statistical Collapse Analysis (SCAN)) together with supplementary COLLAPSE software seem to be adequate, relevant and reliable tools for defining the pattern of similarity in databases on forensic population genetics. Use of the relevant statistics, χ^2 reduction, adequately solves the problem of multiple comparisons (multiple hypothesis testing) in large databases. Evaluated similarities between different population samples appeared to be rather reasonable and interpretable: very often, the samples of related origin appeared to be indistinguishable statistically.

http://www.ufrgs.br/bioinf/COLLAPSE.zip and ftp://bionet.nsc.ru/pub/biology/dbms/COLLAPSE.zip or under request from the authors: Nikita@NH8333.spb.edu and SAG@cards.lanck.net.

Acknowledgements

The authors are deeply indebted to Drs. Bernd Brinkmann and Carsten Hohoff for their help and to the DFG for sponsoring our participation in 19th International Congress ISFG 2001.

References

[1] W. Huckenbeck, K. Kuntze, H.-G. Scheil, The Distribution of the Human DNA-PCR Polymorphisms, Verlag Dr. Köster, Berlin, 1997, p. 301.

[2] W. Huckenbeck, H.-G. Scheil, The Distribution of the Human DNA-PCR Polymorphisms—a Worldwide Database—Supplement Volume I (1998/99/00). Düsseldorf. Institute of Forensic Medicine and Institute of

Human Genetics and Anthropology, Heinrich-Heine-University, 1998–2001. http://www.uni-duesseldorf.de/WWW/MedFak/Serology/dna.html.

[3] S. Schneider, D. Roessli, L. Excoffier, Arlequin: A software for population genetic data analysis. Ver 2.000, Geneva: Genetics and Biometry Laboratory, Dept. of Anthropology and Ecology, University of Geneva, 1995–2000. http://lgb.unige.ch/arlequin/.

[4] P.O. Lewis, D. Zaykin, Genetic Data Analysis: Computer Program for the Analysis of Allelic Data. Version 1.0 (d16c) 2001, http://lewis.eeb.uconn.edu/lewishome/software.html.

[5] K. Tamura, TreeExplorer, 1997–1999. http://evolgen.biol.metro-u.ac.jp/pub/MolEvol/TE212.zip.

[6] N.N. Khromov-Borisov, I.B. Rogozin, J.A.P. Henriques, F.J. de Serres, Similarity pattern analysis in mutational distributions, Mutat. Res. 163 (1) (1999) 55–74.

[7] Z. Gilula, A.M. Krieger, Collapsed two-way contingency tables and the chi-square reduction principle, J. R. Stat. Soc. B51 (3) (1989) 425–433.

[8] M.A. Kastenbaum, A note on the additive partitioning of chi-square in contingency tables, Biometrics 16 (3) (1960) 416–422.

[9] C. Hirotsu, Multiple comparisons and grouping rows in a two-way contingency table, Rep. Stat. Appl. Res. UJSE 25 (1) (1978) 1–12.

[10] C. Hirotsu, Defining the pattern of association in two-way contingency tables, Biometrika 70 (4) (1983) 579–589.