Egyptian Journal of
Forensic Sciences

ORIGINAL ARTICLE Open Access

Evaluation and comparison of population genetics software in Rabari Tribe of Gujarat population

Aditi Mishra[1†], Archana Kumari[2†], Sumit Choudhary[1] and Ulhas Gondhali[3*†]

Abstract

Background: Today, when forensic experts talk about quantifiable hereditary traits, they do not just depend on the assessment and examination of DNA profiles but also relate them to the population structures. The use of high-throughput molecular marker technologies and advanced statistical and software tools have improved the accuracy of human genetic diversity analysis in many populations with limited time and resources. The present study aimed to investigate the genomic diversity in Gujarat's Rabari population, using 20 autosomal genetic markers. Numerous bio-statistical software programs are available for the interpretation of population data in forensics. These statistics deal with the measurement of uncertainty and also provides a probability of a random match. The present paper aims to provide a practical guide to the analysis of population genetics data. Three statistical software packages named Cervus, Genepop, and Fstat are compared and contrasted. The comparison is performed on the profiles obtained from fifty unrelated blood samples of healthy male individuals. DNA was extracted using the organic extraction method, 20 autosomal STR loci were amplified using PowerPlex 21 kit (Promega, Madison, WI, USA) and detected on 3100 Genetic Analyser (Life Technologies Corporation, Carlsbad, CA, USA).

Results: A total of 170 alleles were observed in the Rabari Tribe of Gujarat population, and allele frequencies ranged from 0.010 to 0.480. The highest allele frequency detected was 0.480 for allele 9 at locus TH01. Based on heterozygosity and the polymorphism information content, FGA may be considered as the most informative markers. Both the combined power of discrimination (CPD) and the combined power of exclusion (CPE) for the 20 analyzed loci were higher than 0.999999. The combined match probability (CPM) for all 20 loci was 2.5 × 10^{-22}.

Conclusions: With respect to the results, the 20 STR loci are highly polymorphic and discriminating in the Gujarat population and could be used for forensic practice and population genetics studies. However, Fstat demonstrated better genetic software for analysis of the demographic structure of a specific or set of populations.

Keywords: Population study, Forensic genetics, Cervus, Genepop, and Fstat

Background

Short tandem repeats (STRs) markers have gained much popularity in forensic DNA analysis for human identity testing, paternity testing, and population genetics studies (Wyner et al. 2020). The genomic characteristics such as short sequence lengths, high polymorphism, and amplifying minute quantities of template DNA make these STR useful genetic markers in forensic DNA typing (Butler 2011; Nwawuba Stanley et al. 2020). Allelic frequency data obtained from unrelated individual in a population is essential. It is the key to obtain reliable results in an analysis of DNA profiles (Butler 2009). However, to date, few studies have been reported on autosomal STRs in the Gujarat population. Hence, there is a need to report more data in the studied population.

* Correspondence: ugondhali@jgu.edu.in
†Aditi Mishra and Ulhas Gondhali contributed equally to this work.
[3]O.P. Jindal Global University, Sonipat, Haryana, India
Full list of author information is available at the end of the article

Here, we have reported allele frequencies and forensic parameters of 20 autosomal STR loci in a sample of 50 unrelated healthy adults from the Rabari population.

'Rabari', also known as Rewari or Desai, derived from the word Sanskrit, means 'outsiders' (Kohler-Rollefson 1992). They are settled in the western part of India, which includes the states of Gujarat and Rajasthan. The settlements are divided into 133 sub-tribes. This study reports the genetic portrait of the Rabari population using the PowerPlex 21 system (D1S1656, D2S1338, D3S1358, D5S818, D6S1043, D7S820, D8S1179, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, Amelogenin, CSF1PO, FGA, Penta D, Penta E, TH01, TPOX, and vWA). Genotype data was compared and evaluated using three population genetics software. Genetic analysis based on sizeable datasets can provide high statistical confidence that can be useful for forensic cases (Arenas et al. 2017). Powerful new methods have been developed to analyze genetic data, sometimes relying on massive computations. These methods are implemented in various software packages and programs, which have grown in number tremendously in the past few years (Butler 2006; Kumawat et al. 2020). Genetic software functions as per the data that needs to be analyzed. The population's demographic and genetic structure is defined by various parameters such as allelic frequencies, gene diversities, heterozygosity, F-statistics, kinship relation, parentage analysis, deviation from Hardy–Weinberg equilibrium (Mishra et al. 2019). In this study, three genetics software named Cervus, Genepop (Rousset 2017), and Fstat are compared and contrasted using the same dataset. The different software were selected based on (i) ease of downloading, (ii) open access software, (iii) the ability to analyze co-dominant data, and (iv) ease of running using a Microsoft Window interface (Coombs et al. 2008).

This research paper offers a concise and straightforward guide to the principles that form the basis of the most common analyses. It focuses on some of the most widely used computer software in population genetics that runs on the Windows operating system. A detailed comparative study reveals all the software's insides and applications, thus facilitating appropriate selection and use.

Methods

Sample collection

The University Research Ethics Review Board approved the study. Settlements of the Rabari population were identified in the state of Gujarat. Individuals from these settlements were approached in person with the help of a Gramsevak (village co-ordinator) or village head of that area. All the participants were briefed about the purpose of the study. With the aim to investigate the genetic diversity of the Rabari population of Gujarat, 50 randomly selected healthy male individuals were chosen for this study. Peripheral blood from 50 unrelated male individuals was collected and stored into EDTA tubes. The participants were duly informed, and consent was obtained, as per the Helsinki Declaration (Rickham 1964). Participants ranged from 20 to 50 years of age, respectively.

DNA extraction and quantification

Genomic DNA from whole blood samples was extracted using organic extraction method. Isolated DNA was quantified with Real-Time PCR ABI 7500 (Applied Biosystems, Foster City, CA, USA) using the Quantifiler DNA Quantification Kit (Applied Biosystems, Foster City, CA, USA).

Amplification

Extracted DNA was amplified for 20 autosomal STR loci (D1S1656, D2S1338, D3S1358, D5S818, D6S1043, D7S820, D8S1179, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, CSF1PO, FGA, Penta D, Penta E, TH01, TPOX, and vWA) and one sex identification marker using PowerPlex 21 PCR (Promega,Madison, WI, USA) Amplification kit. PCR conditions were set as per the manufacturer's instructions in a total volume of 25 μl and using Gene Amp PCR System 9700 Thermal Cycler (Applied Biosystems, Foster City, CA, USA). Positive and negative controls were also used throughout the reactions.

DNA electrophoresis and analysis

The PCR products were size separated via capillary electrophoresis using ABI 3100 Genetic Analyzer (Life Technologies Corporation, Carlsbad, CA, USA) and sized with GeneScan500-LIZ internal lane size standard (Thermo) as per the manufacturer's recommended protocol. GeneMapper ID-X Software Version 1.4 (Applied Biosystems Foster City, CA, USA) was used to determine amplified fragments' fragment size. All alleles' designations were based on a comparison with allelic ladders provided in the PowerPlex 21 system. All steps were carried out according to the quality assurance standards recommended by the Scientific Working Group on DNA Analysis Methods (SWGDAM 2010).

Statistical analyses

Allelic frequency and parameters of forensic interest such as genetic diversity, polymorphism information content (PIC), Hardy–Weinberg test (HWE), observed heterozygosity (HO), expected heterozygosity (HE), null allele frequency, and F-statistics were calculated using these software programs. The latest versions of the software were studied for functions and features. All three software are Freeware and operate on Windows, Linux, and Mac operating systems. All three programs were tested using a fixed data set of 50 individuals of the targeted population. Twenty genotype markers have been considered for this comparison. The first problem to be addressed was the input data file format which varied between different software packages. A minute error of space or comma could make the data unreadable or missorted. Organizing data into the proper format is time-consuming and often takes longer than the analysis. There are some programs available that facilitate importing or exporting of data as per the requirement rather than reformatting data manually. These different software programs allow experts to prepare the input data file in the required format and make the analysis easier and faster. It is significant where the data set may have to be subjected to more than one application for analysis. An overview of software has been illustrated with data generated from these twenty autosomal loci in the studied population. The alleles generated from a genetic analyzer were separated by Gene mapper software and exported into an excel sheet. It is a universal method to enter the population data.

As such, Cervus reads the text-based file of genotypes for analysis purpose. This software reads the data in (.csv) format. Cervus software, as its 3.0.7 version, can be downloaded from (www.fieldgenetics.com). It provides a template that functions for both co-dominant and diploid data. The data inserted in Cervus can be analyzed and converted in the Genepop format, i.e., (.txt) (if unable to read, try using a double extension like txt.txt). Genepop software, as its 4.7.5 version, can be downloaded from (http://kimura.univ-montp2.fr/~rousset/Genepop.htm). Genepop can convert the input file into different software readable formats such as Fstat and Biosys. For this study, it was converted into Fstat format, i.e., (DAT extension). Fstat is a computer program that calculates F statistics and can be downloaded from (https://www2.unil.ch/popgen/softwares/fstat.htm). All the necessary results were compiled and compared with each other. The significant features and functions of all the three software were noted in the comparative chart (Table 1).

Tools for the population genetic analyses

Cervus (field genetics) version 3.0.7

Cervus software analyses genetic data generated from co-dominant markers, namely microsatellites and SNPs. This software functions on two principles. Firstly, the genetic markers are independently inherited or in linkage equilibrium. Secondly, the nature of species is diploid and genetic markers are autosomal. Cervus software offers the statistical likelihood method. It is mainly employed for parentage analysis and occasionally for genetic analysis. Cervus offers other additional features such as allele frequency analysis, simulation of parentage analysis) (Marshall et al. 1998), parentage and identity analysis, and convert the genotype file into another format such as gene pop, genetix, and kinship. The software can detect those datasets containing thousands of loci. It calculates the following parameters: (1) Hardy–Weinberg equilibrium; (2) polymorphism information content; (3) observed heterozygosity; (4) expected heterozygosity; (5) alleles per locus (k); (6) F-test; (7) non-exclusion probability for the first parent, second parent, pair parent, identity, and sib identity (Kalinowski et al. 2010).

Table 1 Major features of reviewed software programs

Features	Programs		
	Cervus	Genepop	Fstat
Allelic frequency	✓	✓	✓
Gene diversity		✓	✓
Parentage analysis	✓		
Heterozygosity deficit	✓	✓	✓
Rho statistics	✓	✓	✓
F-statistics		✓	✓
Mantel test		✓	✓
Population differentiation		✓	✓
Identity analysis	✓		
Hardy–Weinberg test	✓	✓	✓
Test of significance		✓	✓
Biased dispersal		✓	✓
Haploid data		✓	✓
Global tests		✓	✓
Linkage disequilibrium		✓	✓
Composite disequilibrium			✓

Input data files

Cervus reads input data files in comma-delimited (.csv) and text format (.txt). All input files can be created in spreadsheet packages such a Microsoft excel.

Output data files

Cervus reports for each analysis independently in a text file with (.txt) extension. For example, the results are different for each analysis (.sim) for simulation parentage analysis, (.alf) for allele frequencies (refer to Table 2 and Table 3).

Comments

Floating-point overflow can occur in the case of a large number of loci. Reported bugs in the older version of Cervus have been resolved in Cervus 3.0.7 (Kalinowski et al. 2010). Selected input files can occasionally crash,

Table 2 Cervus output representing the number of alleles per locus (k), number of individuals (N), observed (Hobs) and expected (Hexp) heterozygosity, polymorphic information content (PIC), combined non-exclusion probability for first parent (NE-1P), second parent (NE-2P), parent pair (NE-PP), identity (NE-I) and sib identity (NE-SI), the Hardy–Weinberg equilibrium significance (HW), and the F test (F)

Locus	K	N	Hobs	HExp	PIC	NE-1P	NE-2P	NE-PP	NE-I	NE-SI	HW	F(Null)
D3S1358	5	50	0.760	0.755	0.706	0.659	0.481	0.298	0.105	0.403	NS	− 0.0163
D1S1656	10	50	0.860	0.857	0.832	0.463	0.298	0.128	0.039	0.336	ND	− 0.0072
D6S1043	11	50	0.840	0.832	0.801	0.519	0.347	0.169	0.054	0.351	ND	− 0.0064
D13S317	8	50	0.720	0.786	0.745	0.605	0.427	0.241	0.082	0.382	NS	0.0391
PENTA-E	12	50	0.900	0.862	0.837	0.453	0.291	0.124	0.038	0.333	ND	− 0.0267
D16S539	7	50	0.820	0.818	0.781	0.560	0.382	0.205	0.065	0.361	ND	− 0.0063
D18S51	14	50	0.780	0.829	0.798	0.522	0.350	0.170	0.055	0.353	ND	0.0305
D2S1338	10	50	0.880	0.852	0.826	0.474	0.308	0.135	0.042	0.339	ND	− 0.0204
CSF1PO	6	50	0.720	0.735	0.680	0.688	0.514	0.334	0.121	0.417	NS	0.0082
PENTA-D	9	50	0.660	0.827	0.795	0.533	0.358	0.180	0.057	0.355	ND	0.1090
TH01	6	50	0.740	0.705	0.661	0.704	0.523	0.328	0.128	0.433	NS	− 0.0281
VWA	7	50	0.820	0.810	0.773	0.569	0.391	0.211	0.068	0.366	ND	− 0.0092
D21S11	7	50	0.760	0.798	0.758	0.592	0.413	0.231	0.076	0.374	ND	0.0186
D7S820	8	50	0.880	0.786	0.746	0.607	0.428	0.244	0.082	0.381	NS	− 0.0658
D5S818	6	50	0.680	0.704	0.651	0.719	0.544	0.361	0.138	0.436	NS	0.0091
TPOX	4	50	0.700	0.671	0.616	0.753	0.581	0.400	0.161	0.458	NS	− 0.0411
D8S1179	8	50	0.800	0.836	0.805	0.515	0.342	0.166	0.052	0.349	ND	0.0192
D12S391	10	50	0.860	0.856	0.830	0.468	0.302	0.133	0.041	0.336	ND	− 0.0087
D19S433	11	50	0.840	0.773	0.732	0.619	0.440	0.251	0.088	0.389	NS	− 0.0495
FGA	11	50	0.900	0.866	0.841	0.447	0.285	0.120	0.036	0.330	ND	− 0.0237
Mean	8.5			0.797	0.760							

most commonly a genotype file. The new version has a feature of a workaround (turn off “preview input files” on the options menu) to resolve this issue. The reasonable error rate is set to 1% for the starting point. If the kinship relationships are known, then Cervus can estimate the actual proportion of loci mistyped from the frequency of mismatches between parents and offspring (Konuma et al. 2000). It has its application in conservation genetics as it gives an accurate parentage and identity analyses that might help wildlife researchers to carry out the population study of wildlife species.

Genepop Version 4.7.5

It is a software package available on the R platform. This software is developed and maintained by Francois Rousset (Package et al. 2020). This population genetic software is used for both haploid and diploid data. Genepop has two major functions: (1) calculates linkage disequilibrium, allele frequency, gene diversity, Hardy–Weinberg exact tests, population differentiation test, null allele frequency, analyze a single genotypic matrix, basic information such as genotypic matrices, observed and expected homozygotes and heterozygotes, estimates Nm, and F-statistics such as *Fst* and other correlation and isolation by distance; (2) convert file into other formats such as Fstat (data.DAT), Biosys (data.BIO), and Linkdos (data.LKD). The missing data in the datasets can be easily handled by Genepop software. It does not have any restrictions on the number of populations or loci (Raymond and Rousset 1995).

Input data files

It accepts the input file in (.txt) format, which can be converted by using Cervus software. The input file of Genepop software should be in ASCII format file data. Once the program is launched, statistical parameters appear and we can choose any of the options that need to be analyzed.

Output data files

Results are stored automatically with the title data.D, data.E, (data is a preferred name of a file). Different analyzed options save their results in their specific extensions. The Genepop outputs are reported in Table 4.

Table 3 Observed allele frequency distribution in Rabari tribe of Gujarat population based on 20 autosomal STRs (*n*=50)

Allele	D3S1358	D1S1656	D6S1043	D13S317	PENTA-E	D16S539	D18S51	D2S1338	CSF1PO	PENTA-D	TH01	VWA	D21S11	D7S820	D5S818	TPOX	D8S1179	D12S391	D19S433	FGA
6	–	–	–	–	–	–	–	–	–	–	0.12	–	–		–	–	–	–	–	–
7	–	–	–	0.01	–	–	–	–	–	0.03	0.18	–	–		–	–	–	–	–	–
8	–	0.05	–	0.33	–	0.02	–	–	–	0.01	0.15	–	–	0.15	0.01	0.21	–	–	–	–
9	–	–	–	0.1	0.01	0.21	–	–	0.02	0.18	0.48	–	–	0.07	0.01	0.15	–	–	0.01	–
9.1	–	–	–	–	–	–	–	–	–	–	–	–	–	0.01	–	–	–	–	–	–
9.3	–	–	–	–	–	–	–	–	–	–	0.04	–	–	–	–	–	–	–	–	–
10	–	0.01	0.01	0.07	0.02	0.15	0.01	–	0.22	0.15	0.03	–	–	0.22	0.13	0.14	0.15	–	–	–
11	–	0.14	0.22	0.21	0.23	0.2	–	–	0.28	0.28	–	–	–	0.33	0.45	0.5	0.05	–	0.01	–
12	–	0.13	0.18	0.23	0.21	0.24	0.03	–	0.37	0.13	–	–	–	0.19	0.24	–	0.06	–	0.12	–
13	–	0.14	0.26	0.03	0.09	0.16	0.18	–	0.09	0.17	–	–	–	0.02	0.16	–	0.16	–	0.26	–
14	0.08	0.08	0.1	0.02	0.09	0.02	0.29	–	0.02	0.03	–	0.17	–	0.01	–	–	0.16	–	0.36	–
14.2	–	–	–	–	–	–	–	–	–	–	–		–	–	–	–		–	0.02	–
15	0.36	0.09	0.1	–	0.11	–	0.18	–	–	0.02	–	0.06	–	–	–	–	0.27	–	0.14	–
15.2	–	–	–	–	–	–	–	–	–	–	–		–	–	–	–	–	–	0.01	–
16	0.21	0.27	–	–	0.1	–	0.16	–	–	–	–	0.26	–	–	–	–	0.14	–	0.05	–
17	0.25	0.07	0.02	–	0.1	–	0.05	0.03	–	–	–	0.24	–	–	–	–	0.01	0.16	0.01	–
17.2	–	–	–	–	–	–			–	–	–		–	–	–	–	–	–	0.01	–
18	0.1	0.02	0.04	–	0.01	–	0.02	0.08	–	–	–	0.19	–	–	–	–	–	0.25	–	0.01
19	–	–	0.12	–	–	–	0.02	0.27	–	–	–	0.06	–	–	–	–	–	0.17	–	0.05
20	–	–	0.01	–	0.02	–	0.01	0.11	–	–	–	0.02	–	–	–	–	–	0.1	–	0.1
21	–	–	0.03	–	–	–	0.01	0.05	–	–	–	–	–	–	–	–	–	0.09	–	\0.12
22	–	–	–	–	–	–	0.02	0.11	–	–	–	–	–	–	–	–	–	0.11	–	0.14
22.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.01
23	–	–	–	–	0.01	–	0.01	0.19	–	–	–	–	–	–	–	–	–	0.06	–	0.13
24	–	–	–	–	–	–	0.01	0.11	–	–	–	–	–	–	–	–	–	0.02	–	0.24
25	–	–	–	–	–	–	–	0.04	–	–	–	–	–	–	–	–	–	0.03	–	0.14
26	–	–	–	–	–	–	–	0.01	–	–	–	–	–	–	–	–	–	0.01	–	0.04
27	–	–	–	–	–	–	–	–	–	–	–	–	0.01	–	–	–	–	–	–	0.02
28	–	–	–	–	–	–	–	–	–	–	–	–	0.08	–	–	–	–	–	–	–
29	–	–	–	–	–	–	–	–	–	–	–	–	0.2	–	–	–	–	–	–	–
30	–	–	–	–	–	–	–	–	–	–	–	–	0.3	–	–	–	–	–	–	–
31	–	–	–	–	–	–	–	–	–	–	–	–	0.17	–	–	–	–	–	–	–
32	–	–	–	–	–	–	–	–	–	–	–	–	0.21	–	–	–	–	–	–	–
33	–	–	–	–	–	–	–	–	–	–	–	–	0.03	–	–	–	–	–	–	–

Table 3 Observed allele frequency distribution in Rabari tribe of Gujarat population based on 20 autosomal STRs (n=50) *(Continued)*

Allele	D3S1358	D1S1656	D6S1043	D13S317	PENTA-E	D16S539	D18S51	D2S1338	CSF1PO	PENTA-D	TH01	VWA	D21S11	D7S820	D5S818	TPOX	D8S1179	D12S391	D19S433	FGA
PD	0.884	0.944	0.932	0.914	0.9472	0.923	0.926	0.9472	0.864	0.922	0.862	0.924	0.925	0.89	0.851	0.826	0.939	0.945	0.897	0.947
PE	0.527	0.714	0.675	0.459	0.7954	0.636	0.562	0.755	0.46	0.369	0.493	0.637	0.527	0.755	0.398	0.428	0.599	0.715	0.675	0.795
PI	2.083	3.57	3.125	1.785	5	2.777	2.272	4.17	1.79	1.47	1.92	2.78	2.08	4.17	1.56	1.67	2.5	3.57	3.13	5
Pm	0.116	0.055	0.068	0.085	0.053	0.077	0.073	0.053	0.136	0.078	0.138	0.076	0.075	0.11	0.149	0.174	0.061	0.055	0.103	0.053
GD	0.755	0.857	0.832	0.786	0.862	0.818	0.829	0.851	0.735	0.829	0.704	0.81	0.798	0.786	0.704	0.67	0.836	0.856	0.772	0.772

PD Power of discrimination, *PE* Power of exclusion, *PI* Paternity index, *Pm* Matching Probability, *GD* Genetic Diversity

Table 4 Genepop output reporting the Hardy–Weinberg tests, global Hardy–Weinberg tests [score (*U*) test], mean squared allele size difference, tests and tables for linkage disequilibrium, population differentiation, Fis, and other correlation

Locus	Null allele frequencies	*P* val	S.E	F_{IS} estimates		MSDinter	Generic differentiation (G test)	Geneotypic differentiation (G test)	Fwc(*IS*)
				W&C	R&H				
D3S1358	0.0797	0.1845	0.0034	–0.0068	0.0299	2.6812	–	–	–0.0068
D1S1656	0.0084	0.3019	0.0082	–0.0033	–0.0130	12.02	–	–	–0.0033
D6S1043	0.0000	0.3604	0.0135	–0.0093	–0.0244	17.599	–	–	–0.0093
D13S317	0.0429	0.4764	0.0080	0.0843	0.0399	6.5347	–	–	0.0843
PENTA-E	0.0000	0.7326	0.0105	–0.0443	–0.0278	13.057	–	–	–0.0443
D16S539	0.0039	0.5402	0.0059	–0.0030	–0.0112	4.5114	–	–	–0.0030
D18S51	0.0297	0.1208	0.0108	0.0593	0.0045	10.991	–	–	0.0593
D2S1338	0.0000	0.7253	0.0079	–0.0336	–0.0417	10.397	–	–	–0.0336
CSF1PO	0.0151	0.2302	0.0052	0.0203	0.0130	2.1976	–	–	0.0203
PENTA-D	0.1084	0.0017	0.0004	0.2034	0.1744	5.7478	–	–	0.2034
TH01	0.0171	0.6526	0.0054	–0.0504	–0.0431	3.1653	–	–	–0.0504
VWA	0.0141	0.7957	0.0045	–0.0121	–0.0288	4.7722	–	–	–0.0121
D21S11	0.0360	0.9557	0.0019	0.0476	0.0212	3.6559	–	–	0.0476
D7S820	0.0000	0.5389	0.0090	–0.1203	–0.0582	8.046	–	–	–0.1203
D5S818	0.1652	0.5699	0.0060	0.0348	0.0359	2.0094	–	–	0.0348
TPOX	0.4990	0.2971	0.0003	–0.0445	–0.0155	2.9988	–	–	–0.0445
D8S1179	0.0143	0.6571	0.0057	0.0434	0.0121	8.0188	–	–	0.0434
D12S391	0.0000	0.3666	0.0094	–0.0043	–0.0063	10.048	–	–	–0.0043
D19S433	0.0000	0.8676	0.0081	–0.0875	–0.0304	6.7996	–	–	–0.0875
FGA	0.0049	0.6252	0.0101	–0.0398	–0.0343	11.971	–	–	–0.0398

*Global Hardy–Weinberg tests [score (*U*) test]: *P* value 0.4412, S.E. 0.0263
*Estimation of exact *P* values by the Markov chain method
*Markov chain parameters for all tests: dememorization: 10,000, batches: 20, iterations per batch: 5000, Hardy–Weinberg: probability test
*MSD: mean squared allele size difference
Only 1 population, no differentiation test, no F_{ST} and no F_{IT}

Comments

Up to version 4.3, Genepop performs the Mantel test based on rank correlation, but presently, the rank test can be executed using setting Markov rank test = (no value needed). Haploid loci/3 digits may not convert into a valid input file for other programs (Crawford 2010). Genepop needs a fast-working processor to obtain accurate results within a reasonable length of time (Rousset 2008). It is a software that has no limitations for the number of population or loci. There is also a web-based version of this program (Excoffier and Heckel 2006).

Fstat version 2.9.4

It is a computer program to calculate F-statistics, developed and maintained by Jérôme Goudet (Goudet 1994). Being a user-friendly software, it has an easy access interface. This software performs gene diversities (per sample and locus) and F-statistics from co-dominant genetic markers. It includes global tests like HW within samples, HW overall samples, HW test per locus or samples, and pair-wise differentiation tests. It performs some more functions such as allelic frequency (per sample and overall), allelic richness (per locus, sample and overall), Hardy–Weinberg equilibrium, genotype frequency (per sample and locus), genotypic disequilibrium, and multiple regression/partial mantel test.

Similarly, Fstat also estimates the Wright's fixation indices (*Fis, Fst,* and *Fit* values), which assess population structures' different levels. *Fis* is a measure of within-population heterozygosity deficit; also called a Wahlund effect and *Fit* is a measure of the global heterozygote deficit. However, *Fst* is a measure of between-population heterozygosity deficit. It can have a limit of 3000 individuals, and it can run up to 200 samples. It can also be used for haploid datasets, and missing data can be easily handled. It is a powerful tool for analyzing various aspects of population genetics over other software like Powerstats (which is more time-consuming and labor-intensive).

Input data files

For Fstat, it is necessary to create an input file named data (.DAT). If we have a three-digit number of alleles, we have to code three (001-999) and be separated by any number or space. Genepop software has the feature to convert the input file in .DAT format.

Output data files

There are tap separators that allow the direct reading of the output file in different available spreadsheets. It has the feature of facilitating printing options and graphical presentation of data. The outputs of Fstat are reported in Table 5.

Comments

Version 1.2 has many fewer features that have been updated and modified in the newer version (Goudet 1994). Fstat can process a large number of data set in a shorter time. As it only supports one type of input data format, that may create a problem for a researcher to calculate the data in a single software. Fstat has many performing features that can be helpful to define the demographic structure of the population.

Result

A researcher may face difficulty in creating an input file. The three programs studied here are linked indirectly as cervus can convert the specific file into genepop format and genepop can convert that file into Fstat format. Software programs employed in this study make it convenient by creating a readable file format. These software tools were used to calculate various forensic parameters. To analyze a large data set, it is necessary to have such a time-saving and user-friendly program. The conversion of an input data file in the appropriate format is a must. The software needs to support an input file in all possible extensions. A graphical presentation makes the understanding of parameters easy.

Table 5 FSTAT output reporting FIS Value, Rho (is), and allelic richness

Multilocus estimates for diploid data				
Locus	**Fis value**	**Rho(is)**	**Allelic richness**	***P* value for Fis within samples**
D3S1358	– 0.007	0.1571	5	0.5950
D1S1656	– 0.003	0.1247	10	0.6150
D6S1043	– 0.009	– 0.2273	11	0.6475
D13S317	0.084	0.2195	8	0.1425
PENTA-E	– 0.044	0.2172	12	0.8425
D16S539	– 0.003	0.1443	7	0.5625
D18S51	0.059	0.2047	14	0.2025
D2S1338	– 0.034	0.1304	10	0.7575
CSF1PO	0.02	0.0261	6	0.4475
PENTA-D	0.203	0.1927	9	0.0075
TH01	– 0.05	– 0.0425	6	0.7575
VWA	– 0.012	– 0.3075	7	0.6275
D21S11	0.048	0.2067	7	0.2800
D7S820	– 0.12	– 0.116	8	0.9675
D5S818	0.035	– 0.0948	6	0.3875
TPOX	– 0.044	– 0.047	4	0.7325
D8S1179	0.043	0.0447	8	0.2800
D12S391	– 0.004	0.0147	10	0.5825
D19S433	– 0.04	– 0.053	11	0.9300
FGA	– 0.087	– 0.0792	11	0.8400

In the Rabari population, a total of 170 alleles with corresponding allele frequencies ranging from 0.010 to 0.480 were observed (Table 3). All the loci fall under Hardy–Weinberg equilibrium after applying Bonferroni correction (Bland and Altman 1995) at a 95% confidence level. The locus D18S51 showed the maximum number of observed alleles, i.e., 14, whereas loci TPOX showed the least number of observed alleles, i.e., 4. The mean number of alleles per locus among the studied loci was found to be 8.500. The allele 9 (0.48) of locus TH01 was the most frequent allele in this population. The observed heterozygosity (Hobs) ranged from 0.660 (PENTA-D) to 0.900 (PENTA-E, FGA) and expected heterozygosity (Hexp) ranged from 0.671 (TPOX) to 0.866 (FGA). The most polymorphic locus among the studied population was FGA, with a value of 0.841, and the least polymorphic locus observed was TPOX with a value of 0.616.

The other forensic parameters such as a power of discrimination (PD), power of exclusion (PE), paternity

Table 6 The most common allele (MCA) and least common allele (LCA) in Rabari Tribe of Gujarat population

Allele	MCA	LCA
D8S1179	15	17
D21S11	30	27
D7S820	11	9.1, 14
CSF1PO	12	9,14
D3S1358	15	14
THO1	9	10
D13S317	8	7
D16S539	12	8,14
D2S1338	19	26
D19S433	14	9,11,15.2,17,17.2
vWA	16	20
TPOX	11	10
D18S51	14	10,20,21, 23,24
D5S818	11	8,9
FGA	24	18, 22.2

MCA most common allele, *LCA* least common allele

index (PI), and matching probability (PM) were calculated through PowerStats v1.2 spreadsheet program (Tereba 1999). The power of discrimination among all the studied loci ranged from 0.826 to 0.947 and was considered highly discriminating for forensic and population genetics studies. The combined probability of match (CPM) and combined paternity index (CPI) for the studied loci are 2.5 × 10^{-22} and 2.42 × 10^{8}. The combined probability of exclusion (CPE) and the combined power of discrimination (CPD) are observed as 0.999999996 and 1, respectively. Locus wise distribution of the most common allele (MCA) and least common allele (LCA) in Rabari Tribe is shown in Table 6.

The genetic diversity value was observed to be highest (0.862) at a locus PENTA-E and lowest (0.670) at locus TPOX (Fig. 1). Fstat software also analyzed the *F*is (correlation of genes within individuals within the population) values for each locus. *F*is value can help determine the level of inbreeding in one population compared to another one. This *p* value must have 95% confidence levels which make the data more robust and informative. For example, if the *F*is *value* of any population observed to be 0.25 and two individuals from that population were mated, then the resulting offspring would be inbred. Their inbreeding coefficient would be ½*0.5 = 0.25. The highest *F*is value was found at locus PENTA-D, with 0.203 followed by the lowest (− 0.003) at D1S1656 and D16S539.

Discussion

With the aim of estimating the genetic relatedness among the populations included in this study, their intrinsic genetic distance was also calculated.The neighbor joining (NJ) dendrogram was derived based on Nei's genetic distance (DA) through the POPTREE2 software (Takezaki et al. 2010).The robustness of the phylogenetic relationship established by the NJ dendrogram was estimated using bootstrap analysis with 1000 replications.The test was applied to compare the allelic frequencies of the presently studied population (Gujarat) with the previously studied eight populations and their published data set—Balmiki (Punjab) (Ghosh et al. 2011), Konkanastha Brahmin (Maharashtra) (Ghosh et al. 2011), (Iyengar (Tamilnadu) (Ghosh et al. 2011), Gond (Madhya Pradesh) (Ghosh et al. 2011), Riang

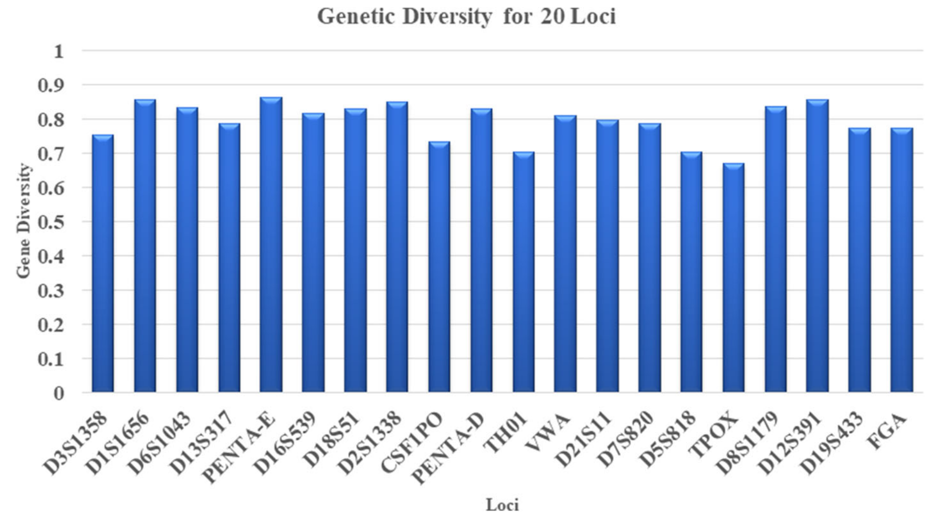

Fig. 1 Genetic diversity of Rabari Tribe of Gujarat population-based on 20 autosomal STRs (*n* = 50)

Table 7 F_{ST} values among nine populations based on the same set of 15 STR markers

	GUJ	BAL	KON	IYEN	GON	TRI	MUN	NEP	SER
GUJ	–	0.017	0.013	0.011	0.001	0.012	0.001	0.018	0.081
BAL	–	–	0.008	0.001	0.014	0.014	0.008	0.024	0.065
KON	–	–	–	0.008	0.015	0.018	0.018	0.033	0.083
IYEN	–	–	–	–	0.006	0.006	0.004	0.018	0.069
GON	–	–	–	–	–	0.017	0.004	0.023	0.072
TRI	–	–	–	–	–	–	0.009	0.016	0.075
MUN	–	–	–	–	–	–	–	0.024	0.072
NEP	–	–	–	–	–	–	–	–	0.084

[a] *GUJ* Rabari (Gujarat, studied population), *BAL* Balmiki (Punjab), *KON* Konkanastha Brahmin (Maharashtra), *IYEN* Iyengar (Tamil Nadu), *GON* Gond (Madhya Pradesh), *TRI* Tripuri (Tripura), *MUN* Munda (Jharkhand), *NEP* Nepal, *SER* Serbia

(Tripura) (Ghosh et al. 2011), Munda (Jharkhand) (Ghosh et al. 2011), Nepal (Kraaijenbrink et al. 2007), and Serbia (Takić Miladinov et al. 2020) (Table 7).

As depicted in Fig. 2, the NJ dendrogram revealed the clustering of the studied populations into three groups. The Iyengar (Tamilnadu) and Konkanastha Brahmin (Maharashtra) formed one group. Populations from the Riang (Tripura) and Nepal formed the second group. Populations from the Rabari Tribe (Gujarat), Munda (Jharkhand), and Gond (Madhya Pradesh) formed the third group. In accordance with the observations recorded through NJ dendrogram, close genetic affinity could be seen between the studied population (Gujarat) and population of Munda (Jharkhand) and Gond (Madhya Pradesh).

The present study on Rabari Population is the first report on data pertaining to polymorphism on FGA autosomal STR locus in this population. A detailed analysis of the polymorphism of the 20 autosomal markers as observed in this study clearly establishes the efficacy of

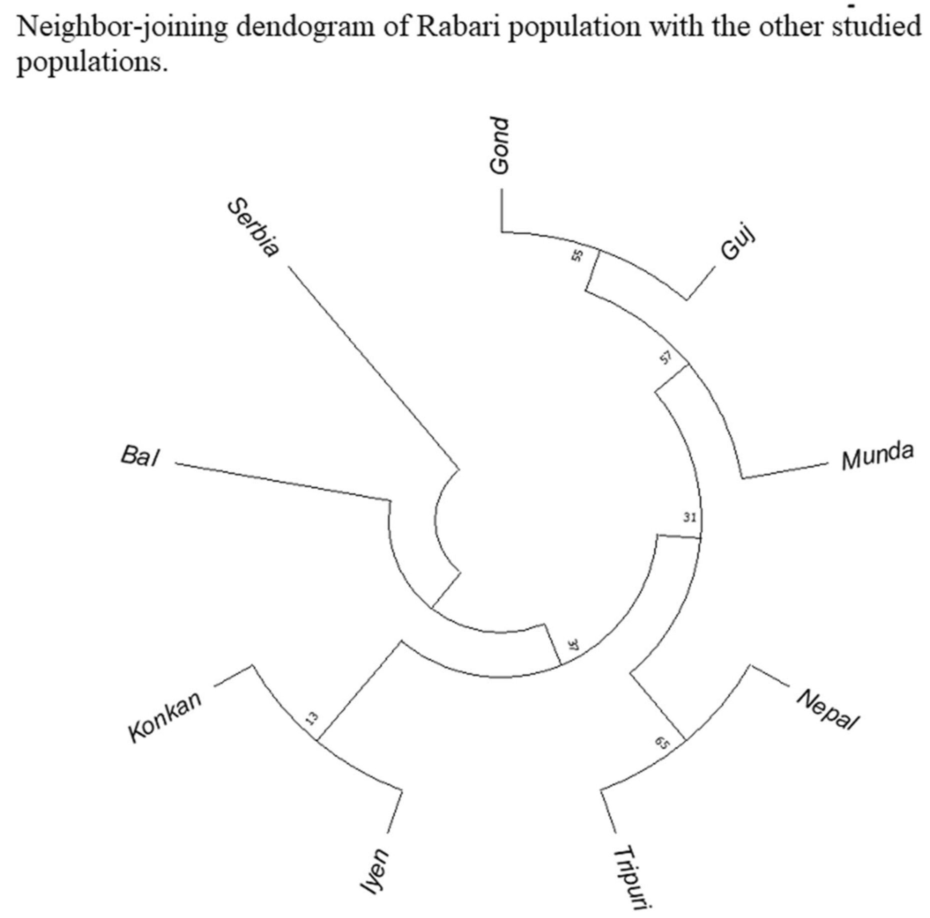
Neighbor-joining dendogram of Rabari population with the other studied populations.
Gond
Serbia
Guj
55
57
Bal
Munda
31
37
13
65
Konkan
Nepal
Iyen
Tripuri

Fig. 2 Neighbor-joining dendogram showing the relationship of Rabari population with the previously reported populations. The tree was constructed based on allele frequencies for 15 autosomal STR loci shared among all populations (D2S1338, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D19S433, D21S11, CSF1PO, FGA, TH01, TPOX, and vWA)

FGA marker for forensic casework, paternity testing, population genetics studies, and familial DNA searching in the Rabari population.This finding has been consistent with the findings in other similar studies on various Indian populations (Dubey et al. 2009; Ghosh et al. 2011; Chaudhari and Dahiya 2014; Ekka et al. 2020; Kakkar et al. 2020) reconfirming that FGA marker exhibits the highest polymorphism and is thus the most useful and unique marker for studying Indian populations.

After testing and evaluating the software (latest versions), the new features and modifications of the software were identified. Genepop has some parallel features like Fstat, but it cannot compete in the new modifications like biased dispersal, simultaneous testing of *Fis*, *Fst*, *Fit* values, among others. Cervus can be vital for a study of the wildlife population, but due to the limitations of performing functions, it may fail to perform some statistical functions. Genepop and Fstat can estimate f-statistics, whereas Fstat can analyze both Nei and Weir and Cockerham families of estimators of gene diversities and F-statistics. All tests were carried out by using randomization methods which effectively displayed the dominance and utility of Fstat program over the remaining two. It overcomes the limitations of the remaining two software as it has the features related to F-statistics and drastically reduces the analysis time by displaying the least inconsistencies between analyses.

In light of the facts discovered in this study, the authors found Genepop and Fstat software to be best suited for forensic applications and strongly advocate using these two over others in the context of similar researches on population genetics. The Cervus software was found to have limited applications in population genetics from forensic perspectives. Its merits and shortfalls have been cataloged for clear understanding of its features. It was also felt that comparisons between some of these software are not appropriate owing to their fundamental differences in purposes for which they have been devised. For example, though the Cervus software helps in conversion of genotype files into Genepop formats, it is predominantly meant for parentage analysis in plant and animal populations and thus it should not be weighed against the other genetic analytics software which has other or additional functionalities. The present research has demonstrated and provides the template guide to the analysis of co-dominant data and selection of appropriate software besides arguing in favor of using more than one software program for getting a comparative evaluation of outputs on any parameter included in the study.

Conclusion

The present study established the valuable genetic information on 20 autosomal STR loci using PowerPlex 21 kit (Promega, Madison, WI, USA) in Rabari Tribe of Gujarat population. The calculated forensic parameters showed that the studied 20 STR markers are highly polymorphic and can be applied in forensic testing as well as in demographic and anthropological studies. According to the geographic or demographic location, differences in population are observed which can be concluded based on genetic distance values. The studied populations (Gujarat) are closely related to Munda (Jharkhand) and Gond (Madhya Pradesh) but distant from geographically distant countries such as Serbia. However, further research is recommended on this population with large sample size to confirm these results**.**

Limitations

The study provided genotype and frequencies data of the autosomal STR genetic markers of the Indian Rabari Tribe for forensic practice albeit all the analyzed samples were male individuals.

Abbreviations

STR: Short tandem repeat; PIC: Polymorphism information content; HWE: Hardy–Weinberg test; HO: Observed heterozygosity; HE: Expected heterozygosity; PD: Power of discrimination; PE: Power of exclusion; PI: Paternity index; PM: Matching probability; CPM: Combined probability of match; CPI: Combined paternity Index; CPE: Combined probability of exclusion; CPD: Combined power of discrimination; LCA: Least common allele; MCA: Most common allele

Acknowledgements

The authors want to acknowledge all the participants for this study.

Authors' contributions

AM and UG planned the study. AK, AM, and UG wrote the manuscript. SKC reviewed the manuscript. All authors have read and approved the manuscript.

Funding

None.

Availability of data and materials

Available with the corresponding author on request.

Declarations

Ethics approval and consent to participate

The study was approved by the ethical committee of the Raksha Shakti University, Gujarat, India (RSU/IRD/RSUIEC/5-2019/72/2019). The written consent was obtained from all the participants.

Consent for publication

All the authors have given written consent for the publication of this manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Forensic Science and Risk Management, Rashtriya Raksha University, Gandhinagar, India. [2]Department of Forensic Science, Mody University of Science and Technology University, Sikar, Rajasthan 332311, India. [3]O.P. Jindal Global University, Sonipat, Haryana, India.

Received: 30 September 2020 Accepted: 31 August 2021
Published online: 16 September 2021

References

Arenas M, Pereira F, Oliveira M, Pinto N, Lopes AM, Gomes V, Carracedo A, Amorim A (2017) Forensic genetics and genomics: much more than just a human affair. PLoS Genet 13(9):e1006960. https://doi.org/10.1371/journal.pgen.1006960

Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. Bmj 310(6973):170. https://doi.org/10.1136/bmj.310.6973.170

Butler JM (2006) Advanced topics in forensic dna analysis - statistics and population genetics, pp 1–18

Butler JM (2009) Fundamentals of forensic DNA typing. Academic press, Cambridge

Butler JM (2011) Advanced topics in forensic DNA typing: methodology. Academic press, Cambridge

Chaudhari RR, Dahiya M (2014) Genetic diversity of 15 autosomal short tandem repeats loci using the AmpFLSTR® Identifiler™ kit in a Bhil Tribe Population from Gujarat state, India. Indian J Hum Genet 20(2):148

Coombs JA, Letcher BH, Nislow KH (2008) Create: A software to create input files from diploid genotypic data for 52 genetic software programs. Mol Ecol Resour 8:578–580. https://doi.org/10.1111/j.1471-8286.2007.02036.x

Crawford NG (2010) Smogd: Software for the measurement of genetic diversity. Mol Ecol Resour 10:556–557. https://doi.org/10.1111/j.1755-0998.2009.02801.x

Dubey B, Meganathan P, Eaaswarkhanth M, Vasulu T, Haque I (2009) Forensic STR profile of two endogamous populations of Madhya Pradesh, India. Legal Med 11(1):41–44

Ekka A, Kujur K, Sirmour R, Guha D, Padhye SA, Verma A, Dixit S, Kumawat RK, Chaubey G, Shrivastava P (2020) Genetic Polymorphism of 15 Autosomal STR Loci in Population of Chhattisgarh, India. Gene Rep 21:100883

Excoffier L, Heckel G (2006) Focus on statistical analysis genetics data analysis : a survival guide. https://doi.org/10.1038/nrg1904

Ghosh T, Kalpana D, Mukerjee S, Mukherjee M, Sharma AK, Nath S, Rathod VR, Thakar MK, Jha GN (2011) Genetic diversity of autosomal STRs in eleven populations of India. Forensic Sci Int : Genet 5(3):259–261

Goudet J (1994) Computer note computer program to calculate F-statistics. J Hered:1994–1995

Kakkar S, Shrivastava P, Mandal SP, Preet K, Kumawat R, Chaubey G (2020) The Genomic Ancestry of Jat Sikh Population from Northwest India Inferred from 15 Autosomal STR Markers Using Capillary Electrophoresis. Ann Hum Biol 47(5):483–489

Kalinowski ST, Taper ML, Marshall TC (2010) Erratum: Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment (Molecular Ecology (2007) 16 (1099-1106)). Mol Ecol 19: 1512. https://doi.org/10.1111/j.1365-294X.2010.04544.x

Kohler-Rollefson I (1992) The Raika dromedary breeders of Rajasthan: a pastoral system in crisis. Nomad People 30:74–83

Konuma A, Tsumura Y, Lee CT et al (2000) Estimation of gene flow in the tropical-rainforest tree Neobalanocarpus heimii (Dipterocarpaceae), inferred from paternity analysis. Mol Ecol 9:1843–1852. https://doi.org/10.1046/j.1365-294X.2000.01081.x

Kraaijenbrink T, van Driem GL, of Gaselô KT, de Knijff P (2007) Allele frequency distribution for 21 autosomal STR loci in Bhutan. Forensic Sci Int 170(1):68–72

Kumawat RK, Mishra A, Shrivastava P (2020) Statistical softwares used in evaluation of forensic DNA typing. In: Forensic DNA typing: principles, applications and advancements. Springer, Singapore, pp 105–134

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. Mol Ecol 7:639–655. https://doi.org/10.1046/j.1365-294x.1998.00374.x

Mishra A, Dixit S, Choudhary SK, Sharma H, Shrivastava P (2019) Forensic genetic analysis of the population of Gujarat with PowerPlex 21 Multiplex System. Forensic Sci Int: Genet Suppl Ser 7(1):167–168

Nwawuba Stanley U, Mohammed Khadija A, Bukola AT, Omusi Precious I, Ayevbuomwan Davidson E (2020) Forensic DNA profiling: autosomal short tandem repeat as a prominent marker in crime investigation. Malays J Med Sci: MJMS 27(4):22–35

Package T, Population T, Data G, et al (2020) Package ' genepop '

Raymond M, Rousset F (1995) Genpop 1.2 Population genetics software for exact test and ecumenicism. Comput Notes 86:248–249

Rickham PP (1964) Human experimentation. Code of ethics of the world medical association. Declaration of Helsinki. Br Med J 2(5402):177–177

Rousset F (2008) GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. Mol Ecol Resour 8:103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Rousset F (2017) Genepop Version 4.7. 0

Takezaki N, Nei M, Tamura K (2010) POPTREE2: Software for constructing population trees from allele frequency data and computing other population statistics with Windows interface. Mol Biol Evol 27(4):747–752

Takić Miladinov D, Vasiljević P, Šorgić D, Podovšovnik Axelsson E, Stefanović A (2020) Allele frequencies and forensic parameters of 22 autosomal STR loci in population of 983 individuals from Serbia and comparison with 24 other populations. Ann Hum Biol 170(1):1–23

Tereba A (1999) Tools for analysis of population statistics. Profiles DNA 2:14–16

Wyner N, Barash M, McNevin D (2020) Forensic autosomal short tandem repeats and their potential association with phenotype. Front Genet 11:884

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.