Forensic Science International: Genetics 3 (2009) e111–e116

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Announcement of Population Data

Developing STR databases on structured populations: The native South Siberian population *versus* the Russian population

Lev A. Zhivotovsky [a,*], Boris A. Malyarchuk [b], Miroslava V. Derenko [b], Marcin Wozniak [c], Tomasz Grzybowski [c]

[a] *Institute of General Genetics, The Russian Academy of Sciences, Gubkin Str. 3, Moscow 119991, Russia*
[b] *Institute of Biological Problems of the North, Russian Academy of Sciences, Portovaya St. 18, 685000 Magadan, Russia*
[c] *Forensic Medicine Institute, The Ludwik Rydygier Medical College, The Nicolaus Copernicus University in Torun, 85-094 Bydgoszcz, Poland*

ARTICLE INFO

Article history:
Received 21 March 2008
Received in revised form 25 July 2008
Accepted 1 August 2008

Keywords:
Siberians
Russian
Ethnic subdivision
STR database
Population differentiation

ABSTRACT

Developing a forensic DNA database on a population that consists of local ethnic groups separated by physical and cultural barriers is questionable as it can be genetically subdivided. On the other side, small sizes of ethnic groups, especially in alpine regions where they are sub-structured further into small villages, prevent collecting a large sample from each ethnic group. For such situations, we suggest to obtain both a total population database on allele frequencies across ethnic groups and a list of θ-values between the groups and the total data. We have genotyped 558 individuals from the native population of South Siberia, consisting of nine ethnic groups, at 17 autosomal STR loci of the kit packages AmpF*l*STR SGM Plus и AmpF*l*STR Profiler Plus. The groups differentiate from each other with average θ-values of around 1.1%, and some reach up to three to four percent at certain loci. There exists between-village differentiation as well. Therefore, a database for the population of South Siberia is composed of data on allele frequencies in the pool of ethnic groups and data on θ-values that indicate variation in allele frequencies across the groups. Comparison to additional data on northeastern Asia (the Chukchi and Koryak) shows that differentiation in allele frequencies among small groups that are separated by large geographic distance can be even greater. In contrast, populations of Russians that live in large cities of the European part of Russia are homogeneous in allele frequencies, despite large geographic distance between them, and thus can be described by a database on allele frequencies alone, without any specific information on θ-values.

Population: The DNA database is comprised of allelic frequencies on 558 individuals from South Siberian ethnic groups that patchily dispersed across southern Siberia (Supplementary Fig. S1): 95 Khamnigans, 80 Tuvinians, 80 Altaians-Kizhi, 42 Mongolians, and additional raw genotypic data on 68 Altaians, 78 Buryats, 29 Sojots, 35 Tofalars, and 51 Khakassians described in [1]. Here we use term 'South Siberia' in a geographic context as an area including the Altai-Sayan Mountains, the Lake Baikal region, and northern Mongolia. For the purposes of comparison, we studied a few other Asian populations: 49 Koreans from South Korea and samples from northeastern Asia: 15 Chukchi, 32 Koryaks, and 14 Evens.

As a contrasting example of a homogeneous pattern of population structure, we studied 371 Russians that live in the European part of the Russian Federation: 66 individuals from Belgorod city, 62 from Pskov city, and 59 from Velikiy Novgorod city; also we used raw genotypic data on 52 individuals from Saratov city, 60 from Mineralnie Vody city, and 72 from Orel city described in [2] (Supplementary Fig. S1). Additionally, we obtained a sample of 13 individuals that have Asian types of mitochondrial DNA (all them belong to macro-haplogroup M) and identified themselves as ethnical Russians. These individuals were from the following cities of the European part of Russia: Kaluga ($n = 2$), Pskov ($n = 1$), Vladimir ($n = 1$), Velikiy Novgorod ($n = 2$), Volot ($n = 5$), Yaroslavl ($n = 1$), and Belgorod ($n = 1$). Each population sample comprises unrelated healthy donors from whom appropriate informed consent was obtained.

DNA samples, amplification, genotyping: Blood samples were taken by venipuncture and collected into EDTA vacutainer tubes. DNA was extracted by following the standard phenol-chlorophorm extraction method. The loci were amplified using PowerPlex® 16 System (Promega, Madison, WI) according to manufacturer's instructions. Products of amplification were analyzed using ABI PRISM 3100 Genetic Analyzer (PE Applied Biosystems). All samples were genotyped at loci D18S51, D21S11, TH01, D3S1358, FGA,

* Corresponding author. Tel.: +7 499 135 52 90; fax: +7 499 135 52 90.
E-mail address: levazh@gmail.com (L.A. Zhivotovsky).

1872-4973/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2008.08.001

TPOX, D8S1179, vWA, CSF1PO, D16S539, D7S820, D13S317, and D5S818. Additionally, the samples collected in this study are typed for loci PentaD and PentaE, whereas those from [1,2] have been typed for D2S1338 and D19S433. Allelic ladders and control DNA samples provided with the kit were used for quality control. GeneScan v. 3.7 and Genotyper v. 3.6 software (PE Applied Biosystems) were used for sizing and genotyping. Allele names are concordant with the recommendations of the DNA Commission of the International Society of Forensic Haemogenetics [3]. GEDNAP and CTS quality control exercises passed successfully by the Forensic Medicine Institute, the Ludwik Rydygier Medical College. The raw genotypic data and allele frequencies in each population sample are presented in Supplementary Tables A1 and A2, respectively; they are available online.

Data analysis: Statistical tests for genetic equilibrium (Hardy–Weinberg's ratios and two-locus linkage equilibrium) and estimates of θ-values were performed with the methods described in [4,5] using software GDA [6]; software SPSS 11.5.0 [7] was applied to principal component analysis.

Results:

A database on the population of South Siberia. The total database for the population of South Siberia provides an exclusion power Q of 0.999999924 across the analyzed 17 loci, with average heterozygosity *per* a locus of 0.788. Significant deviations from Hardy–Weinberg proportions and linkage equilibrium were found only for *vWA* (heterozygote excess) and *D8S1179/D7S820*, with p-values 0.007 and 0.01, respectively, after the Bonferroni correction.

Between-ethnic differentiation in South Siberia. The South Siberian ethnic groups significantly differentiate from each other in allele frequencies, with average θ-value of 0.011 and its 95% bootstrap confidence interval across loci of (0.0090, 0.0131). Since the latter figures do not exceed much the lower level of θ-values suggested by [8], 0.01, we pooled the samples together to obtain a combined reference database on the South Siberian population (Table 1). To characterize the extent of differentiation of the ethnic groups in allele frequencies, we additionally computed single-locus θ-values between allele frequencies in each ethnic group and those in the combined database (Table 2).

Table 1
Allele frequencies in the South Siberian population (n is the number of chromosomes examined).

Allele	D3S1358 (n = 1116)	vWA (n = 1030)	FGA (n = 1116)	TH01 (n = 1116)	TPOX (n = 1116)	CSF1PO (n = 1116)	D5S818 (n = 1116)	D13S317 (n = 1116)	D7S820 (n = 1116)
6				0.12545					
7				0.30824	0.00090		0.03495		0.00448
8				0.11559	0.52957	0.00179	0.00090	0.22133	0.29211
9				0.28853	0.09498	0.03763	0.04211	0.14606	0.08961
10				0.00896	0.02240	0.24014	0.08423	0.14875	0.20430
11	0.00090				0.28763	0.32168	0.43369	0.20789	0.21953
12	0.00000	0.00388			0.05914	0.31631	0.25627	0.17832	0.16935
13	0.00000	0.00388			0.00538	0.05735	0.13710	0.07258	0.01613
14	0.03405	0.11165				0.02419	0.01075	0.02151	0.00448
15	0.40502	0.05825				0.00090		0.00358	
16	0.30108	0.24854							
17	0.20520	0.27670	0.00090						
18	0.05287	0.20874	0.00358						
19	0.00090	0.06602	0.02867						
20		0.02233	0.04659						
21			0.08602						
22			0.14427						
23			0.18369						
24			0.27061						
25			0.14964						
26			0.05197						
27			0.00717						
28			0.00269						
29			0.00090						
9.3				0.15323					
10.3				0.00000					
20.2			0.00269						
21.2			0.00000						
22.2			0.00090						
23.2			0.00896						
24.2			0.00179						
25.2			0.00806						
26.2			0.00090						
Q	0.45	0.60	0.68	0.55	0.37	0.49	0.49	0.65	0.58
H	0.70	0.80	0.84	0.77	0.62	0.73	0.72	0.83	0.79

Allele	D16S539 (n = 1116)	D2S1338 (n = 522)	D8S1179 (n = 1116)	D21S11 (n = 1116)	D18S51 (n = 1112)	D19S433 (n = 520)	PentaD (n = 592)	PentaE (n = 592)
4								
5								0.02365
6							0.01689	0.00000
7							0.00845	0.03209
8	0.01254		0.00179				0.01351	0.01520
9	0.25806		0.00090				0.34797	0.00000
10	0.12724		0.08423		0.00180		0.14358	0.06081
11	0.21326		0.05287		0.00360	0.00769	0.17399	0.07939

Table 1 (*Continued*)

Allele	D16S539 (n = 1116)	D2S1338 (n = 522)	D8S1179 (n = 1116)	D21S11 (n = 1116)	D18S51 (n = 1112)	D19S433 (n = 520)	PentaD (n = 592)	PentaE (n = 592)
12	0.24194		0.07975		0.02878	0.04808	0.12331	0.10811
13	0.11380		0.37276		0.14119	0.32308	0.12331	0.04730
14	0.03136		0.21057		0.28867	0.22692	0.03209	0.07939
15	0.00179		0.13710		0.13489	0.08846	0.01689	0.09122
16		0.00575	0.05018		0.09712	0.05000		0.10980
17		0.08429	0.00717		0.08004	0.00192		0.08277
18		0.09195	0.00269		0.05486			0.06419
19		0.18966			0.05036			0.07264
20		0.13027			0.04766			0.04054
21		0.02107			0.02338			0.03547
22		0.04789			0.03147			0.04054
23		0.18199			0.01169			0.01351
24		0.13602			0.00180			0.00338
25		0.08238			0.00270			
26		0.02299		0.00090				
27		0.00192		0.00627				
28		0.00383		0.04928				
29				0.21147				
30				0.39068				
31				0.09229				
32				0.00986				
33				0.00090				
12.2								
13.2						0.04615		
14.2						0.07885		
15.2						0.10000		
16.2						0.02885		
17.2								
18.2								
28.2				0.00896				
29.2				0.00000				
30.2				0.01703				
31.2				0.07168				
32.2				0.09767				
33.2				0.03315				
34.2				0.00806				
35.2				0.00179				
Q	0.60	0.74	0.59	0.59	0.72	0.64	0.61	0.85
H	0.80	0.87	0.78	0.78	0.85	0.81	0.80	0.93

Between-village variation in South Siberia. For four ethnic groups, we extracted data on two major villages in each and found average between-village θ-values to be greatly significant within the Altai-Kizhi and the Tuvinian, 0.021 and 0.018, with 95% confidence intervals of (0.0086, 0.0341) and (0.0073, 0.0287), respectively. It is worth noting that the between-village variation exceed 0.01 even for those markers that were shown to be least variable between ethnic groups, PentaE,D, D18S51 and D19S433 (Table 2) For other two ethnics, the Khakas and the Altai, the average between-village θ-value was insignificant, but single-locus θ-values exceeded 0.03

Table 2
θ-Values between the allele frequencies in a single ethnic group and those in the combined data on the South Siberian population (from Table 1).

Loci	Ethnic groups									Maximal θ-values
	Altai	Altai-Kizhi	Buryat	Khakas	Khamnigan	Mongol	Sojot	Tofalar	Tuva	
vWA	0	0	0	0	0.0013	0.0110	0.0066	0.0176	0.0007	0.0176
D8S1179	0.0067	0.0080	0.0121	0.0036	0.0072	0	0.0129	0.0070	0.0062	0.0129
TPOX	0.0089	0	0	0	0	0.0100	0	0.0463	0.0007	0.0463
D3S	0.0039	0	0.0037	0.0091	0.0013	0.0003	0.0140	0.0371	0.0156	0.0371
FGA	0	0	0.0014	0.0072	0.0007	0.0049	0.0123	0.0116	0.0122	0.0123
TH01	0.0016	0	0.0004	0.0112	0.0010	0	0.0024	0.0303	0.0078	0.0303
D21S11	0.0047	0.0071	0.0022	0.0106	0.0003	0	0.0033	0.0021	0.0016	0.0106
D5S818	0.0135	0	0	0.0073	0.0028	0.0199	0.0052	0.0284	0	0.0284
D13S317	0.0024	0.0021	0.0009	0.0005	0.0003	0.0065	0.0095	0.0105	0.0071	0.0105
D7S820	0	0.0023	0	0	0.0042	0.0008	0.0251	0.0244	0.0078	0.0251
D16S539	0	0.0054	0.0098	0.0170	0.0033	0.0010	0.0089	0.0377	0.0003	0.0377
CSF1PO	0	0.0018	0.0050	0.0130	0	0	0.0085	0.0124	0	0.0130
D18S51	0.0000	0.0056	0.0012	0.0014	0	0	0.0019	0.0068	0.0030	0.0068
D2S1338	0	–	0	0	–	–	0.0034	0.0286	–	0.0286
D19S433	0	–	0.0005	0.0003	–	–	0	0.0088	–	0.0088
PentaD	–	0.0006	–	–	0	0	–	–	0	0.0006
PentaE	–	0.0037	–	–	0.0004	0	–	–	0.0039	0.0039

at loci TPOX, D7S820, D16S539, D2S1338, were higher than 0.01 for D19S433 in the Khakas, and exceeded 0.01 at loci D3S1358, D18S51, and D19S433 in the Altai.

A database on the Russian population. The database provides an exclusion power $Q = 0.999999946$ across the 17 loci, with average heterozygosity *per* locus $H = 0.795$. Differentiation between Russian samples is insignificant, with average $\theta = 0.0013$ and its 95% confidence interval (-0.0005, 0.0033); the maximal θ-value, 0.010, was found for locus D16S539 only. Therefore, a common database on allele frequencies in European Russians can be developed by pooling data from the six urban population samples together (Table 3), with no additional information on θ-values.

Between-population differentiation. Differentiation between the South Siberian ethnic groups is high, it increased further when northeastern Asian the Koryak and the Chukchi were added (Supplementary Fig. S2). In contrast, the Russian urban samples are genetically very close to each other, although they are quite distant geographically (Supplementary Fig. S2). Generally, the eastern Slavs seem to be very homogeneous because another Slavic ethnic

Table 3
Allele frequencies in the Russian population (n is the number of chromosomes examined).

Allele	D3S1358 (n = 742)	vWA (n = 680)	FGA (n = 738)	TH01 (n = 742)	TPOX (n = 742)	CSF1PO (n = 742)	D5S818 (n = 742)	D13S317 (n = 742)	D7S820 (n = 742)
6				0.22237	0.00135				
7				0.13342	0.00000		0.00270		0.01213
8				0.09434	0.57951		0.00135	0.15499	0.17116
9				0.23046	0.08760	0.04582	0.04043	0.08760	0.15094
10				0.00404	0.05930	0.28167	0.08895	0.05795	0.28302
11					0.24394	0.29380	0.33423	0.35175	0.21159
12	0.00270	0.00875			0.02830	0.30593	0.36658	0.20081	0.14420
13	0.00000	0.00588				0.05795	0.15229	0.10108	0.02561
14	0.11590	0.09412				0.00809	0.00943	0.04313	0.00135
15	0.25337	0.11471				0.00539	0.00404	0.00270	
16	0.29380	0.21471				0.00135			
17	0.21833	0.23529							
18	0.11051	0.21618	0.01897						
19	0.00539	0.09853	0.08943						
20		0.01765	0.15854						
21		0.00294	0.14634						
22			0.23171						
23			0.11247						
24			0.12060						
25			0.08266						
26			0.02439						
27			0.00407						
28									
29									
9.3				0.31402					
10.3				0.00135					
20.2									
21.2			0.00407						
22.2			0.00407						
23.2			0.00271						
24.2									
25.2									
26.2									
Q	0.56	0.64	0.71	0.55	0.35	0.49	0.48	0.60	0.61
H	0.78	0.82	0.86	0.77	0.59	0.74	0.72	0.79	0.80

Allele	D16S539 (n = 742)	D2S1338 (n = 368)	D8S1179 (n = 742)	D21S11 (n = 742)	D18S51 (n = 742)	D19S433 (n = 368)	PentaD (n = 372)	PentaE (n = 374)
4								
5								0.07487
6								0.00000
7	0.00135						0.00269	0.15241
8	0.01482		0.00539				0.00269	0.00535
9	0.08356		0.00404				0.22849	0.01604
10	0.05256		0.05256		0.00674		0.12097	0.12299
11	0.28302		0.06199		0.01213	0.00272	0.19355	0.08289
12	0.34906		0.16173		0.09569	0.07609	0.19355	0.13636
13	0.19137		0.34097		0.11995	0.21196	0.17204	0.10695
14	0.02291		0.22372		0.15229	0.37500	0.06720	0.05615
15	0.00135		0.11860		0.17251	0.15489	0.01344	0.05615
16		0.04620	0.02291		0.17116	0.04348	0.00269	0.08021
17		0.22554	0.00674		0.11590	0.00543	0.00269	0.05615
18		0.10326	0.00135		0.07143			0.02941
19		0.14402			0.04582			0.01070
20		0.13315			0.02426			0.01070
21		0.02717			0.00539			0.00267
22		0.01359			0.00539			

Table 3 (*Continued*)

Allele	D16S539 (n = 742)	D2S1338 (n = 368)	D8S1179 (n = 742)	D21S11 (n = 742)	D18S51 (n = 742)	D19S433 (n = 368)	PentaD (n = 372)	PentaE (n = 374)
23		0.10326			0.00135			
24		0.08967						
25		0.09511		0.00135				
26		0.01902		0.00404				
27				0.02022				
28				0.15768				
29				0.18868				
30				0.23315				
31				0.06739				
32				0.01213				
33				0.00270				
12.2						0.00272		
13.2						0.01087		
14.2						0.03804		
15.2						0.04076		
16.2						0.02174		
17.2						0.00815		
18.2						0.00815		
28.2				0.00270				
29.2				0.00270				
30.2				0.07412				
31.2				0.08625				
32.2				0.09704				
33.2				0.04447				
34.2				0.00539				
35.2								
Q	0.53	0.74	0.59	0.71	0.74	0.59	0.65	0.80
H	0.75	0.87	0.79	0.86	0.87	0.78	0.82	0.90

group, the Belarusian, is spatially nearly homogeneous across the Republic of Belarus, and differentiation between Russians, Belarusians, and Ukrainians from large cities is very small [9]. It is interesting that a sample of Russians that have Asian types of mitochondrial DNA is close to other Russian samples (Supplementary Fig. S2); this is explained by rapid substitution of autosomal alleles during generations of absorption by the home population.

Other remarks: Some populations are subdivided by natural boundaries into small subpopulations, exemplified by ethnic groups in alpine regions, as well as by social, religious or other kinds of barriers, which makes difficult to establish a forensic database on a single local group as sampling a sufficiently large number of unrelated individuals can be unrealistic. Therefore, pooling data from several neighboring small groups together seems to be the only way to establish a reference database on a subdivided population. On the other hand, such a database may be internally heterogeneous and thus needs to be accompanied by information on variation of allelic frequencies in terms of θ-statistic as it has been suggested in [8]. The value $\theta = 0.03$ has been recommended for use in forensic cases that involve small isolated tribes, such as American Indians, and $\theta = 0.01$ in other cases of subdivision ([8], p. 122). This is appropriate for cases with unknown level of population differentiation. However, when the extent of differentiation can be evaluated from observed data, then using actual estimates of θ-values seems to be more practical especially because actual θ-values in a particular population can greatly vary across loci, thus being for some loci larger than the 0.01 and even may exceed 0.03.

The ethnic groups of the population of South Siberia significantly differentiate from each other in allele frequencies [1]. All these suggest using the allele frequencies from Table 1 together with the maximal θ-values from the very right column of Table 2 in those forensic implications that involve the natives from South Siberia. Moreover, taking in account a possibility of between-village variation, the θ-values in the right column of Table 2 that are less than 0.01 must be substituted by 0.01 at least.

In contrast, in forensic cases that involve Russians from large cities of European part of Russia (Table 3), θ-values can be taken as 0.01 at most.

The paper is in agreement with the requirements for publication of population data [10].

Acknowledgements

We are grateful to all the voluntary donors of DNA samples used in this study, to Irina Dambueva and Ilia Zakharov for some Siberian samples, and to Sergey Rychkov for some Russian samples. Research supported by grants of The Russian Academy of Sciences "Fundamental Science to Medicine" and "Biodiversity", and by The Russian Foundation for Basic Research (grants 06-04-48136 and 07-04-00171).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2008.08.001.

References

[1] M.V. Derenko, J. Czarny, B.A. Malyarchuk, M. Wozniak, I.K. Dambueva, T. Grzybowski, I.A. Zakharov, The variation of 15 autosomal microsatellite DNA loci in five indigenous populations of South Siberia, Molecular Biology (Moscow) 41 (2007) 531–538.

[2] B.A. Malyarchuk, M. Wozniak, J. Czarny, M.V. Derenko, T. Grzybowski, D. Miscicka-Sliwka, Variation of 15 autosomal microsatellite DNA loci in the Russian population, Molecular Biology (Moscow) 41 (2007) 3–7.

[3] B. Olaisen, W. Bär, B. Brinkmann, B. Budowle, A. Carracedo, P. Gill, P. Lincoln, W.R. Mayr, S. Rand, DNA recommendations 1997 of the international society for forensic genetics, Vox Sanguinis 74 (1998) 61–63.

[4] B.S. Weir, Genetic Data Analysis II: Methods for discrete population genetic data, Sinauer Associates, Sunderland, MA, 1996.
[5] D. Zaykin, L. Zhivotovsky, B.S. Weir, Exact tests for association between alleles at arbitrary numbers of loci, in: B. Weir (Ed.), Human Identification: The Use of DNA Markers, Kluwer Acad. Publ., London, 1995, pp. 169–178.
[6] P.O. Lewis, D. Zaykin, Genetic data analysis: computer program for the analysis of allelic data. Version 1.0, 2001 (http://lewis.eeb.uconn.lewishome/software.html).
[7] SPSS for Windows, Release 11.5.0, Standard version, SPSS Inc., 2002.
[8] NRC (National Research Council), The Evaluation of Forensic DNA Evidence, National Academy Press, Washington, DC, 1996.
[9] L.A. Zhivotovsky, V.M. Veremeichyk, A.I. Mikulich, I.G. Udina, L.A. Atramentova, S.A. Kotova, N.A. Kartel, I.S. Tsybovsky, A comprehensive population survey on the distribution of STR frequencies in Belarus, Forensic Science International 172 (2007) 156–160.
[10] P. Lincoln, A. Carracedo, Publication of population data of human polymorphisms, Forensic Science International 110 (2000) 3–5.