ISSN 0026-8933, Molecular Biology, 2016, Vol. 50, No. 3, pp. 387–397. © Pleiades Publishing, Inc., 2016.
Original Russian Text © V.A. Stepanov, K.V. Vagaitseva, V.N. Kharkov, A.A. Cherednichenko, A.V. Bocharova, 2016, published in Molekulyarnaya Biologiya, 2016, Vol. 50, No. 3, pp. 445–456.

GENOMICS. TRANSCRIPTOMICS

UDC 575.174:577.21

Panel of X-Linked Single-Nucleotide Polymorphic Markers for DNA Identification (XSNPid) Based on Multiplex Genotyping by Multilocus PCR and MALDI-TOF Mass Spectrometry

V. A. Stepanov[a, b, *], K. V. Vagaitseva[a, b], V. N. Kharkov[a, b], A. A. Cherednichenko[a], and A. V. Bocharova[a]

[a] *Research Institute for Medical Genetics, Tomsk, 634050 Russia*
[b] *National Research Tomsk State University, Tomsk, 634050 Russia*
**e-mail: vadim.stepanov@medgenetics.ru*

Received July 28, 2015; in final form, October 26, 2015

Abstract—Human genetic markers linked with the X chromosome (X-linked) are used in the field of population and medical genetics, as well as for DNA identification of individuals in forensic science and forensic medicine. We proposed an XSNPid panel that consists of 66 unlinked single nucleotide X chromosome markers and developed a protocol for their multiplex genotyping using multilocus PCR and MALDI-TOF mass spectrometry. The XSNPid panel is genotyped within two multiplexes (36 and 30 markers). The developed protocol provides an efficient genotype reading; the fraction of determined genotypes is 98.29%. The high level of gene diversity (0.461) for the X-linked SNPs included in the panel is characteristic of the Russian population. A total of 63 out of 66 markers that provide a high efficiency of genotyping and independent inheritance are suitable for DNA identification purposes. The XSNPid panel is characterized by a very high discriminating ability when studying the Russian population. The probability of genotype coincidence in two unrelated individuals is 9×10^{-27} for women and 2×10^{-18} for men. Also, the XSNPid panel has a greater multiplex capacity in addition to a higher discriminating ability compared to the other closest analogues of the X chromosome SNP sets, which makes it more cost effective and less time consuming. The XSNPid panel is a convenient tool, not only for individual DNA identification, but also for population genetic studies.

Keywords: single nucleotide polymorphic markers (SNPs), X chromosome, MALDI-TOF mass spectrometry, multiplex genotyping, DNA identification, population genetics, genetic diversity, human populations

DOI: 10.1134/S0026893316030158

INTRODUCTION

Genetic markers of the X chromosome (X-linked genetic markers) are used in the field of population and medical genetics, as well as for the DNA identification of individuals in forensic science and forensic medicine. The data on genetic variability of the X chromosome complement the picture of the genetic diversity and genetic differentiation of populations in population studies, which can be obtained using uniparental lines (in relation to markers of Y-chromosome and mtDNA), as well as panels of autosomal polymorphic loci [1–4]. In medical genetics X chromosome markers are used for direct and indirect diagnosis of X-linked diseases [5] and search for genetic causes of common diseases. According to the catalog of Genome Wide Association Studies (GWAS), there are about 150 SNPs (single nucleotide polymorphisms), associated with various common chronic diseases, on the X chromosome [6]. In forensic science and forensic medicine, X-chromosome markers are used primarily when testing biological kinship. Features of the X-chromosome transfer in a number of generations (males inherit one of mother's X chromosomes, females inherit one X chromosome from the mother and one chromosome inherited from the paternal grandmother) are the most useful in complex cases of kinship determination. For example, they are applied to determine paternity when the DNA of intended father is not available, to determine paternity between close relatives, and to determine kinship within paternal lineage when only women are available from the genealogic tree [7]. In addition to complex cases of kinship determination, genetic markers of the X chromosome that have sufficient informativeness contribute to addressing standard tasks of individual identification [8].

Microsatellite markers (STRs), including X-linked markers, are commonly used to identify DNA [4, 9–14]. These multiallelic markers are highly heterozygous; therefore, a relatively small number of them (15–20) is sufficient for individual human identifica-

Abbreviations: MALDI-TOF, matrix assisted laser desorption/ionisation-time of flight; SNP, single nucleotide polymorphism.

tion [9]. Single nucleotide markers (SNPs) are less variable, but much more prevalent in the human genome. In order to achieve a discriminatory potential comparable with that of STRs, the number of SNP markers should be three to four times more [15]. The X-linked SNP panels developed based on traditional technologies (real-time PCR, capillary gel electrophoresis) do not reach the level of informativeness of systems based on autosomal STR markers [16–18]. The technologies of massive parallel sequencing and new methods of genotyping makes evident the prospects of STR marker displacement by SNP based systems for DNA identification [19–21]. Among genotyping technologies that are optimal for medium scale tasks (the typing of dozens or hundreds of markers in hundreds or thousands of samples), MALDI-TOF mass spectrometry holds one of the leading places [22, 23].

The aim of this study was to develop an X-linked single nucleotide polymorphic marker panel for population genetic analysis and individual DNA identification in the Russian population, as well as to develop a multiplex genotyping method based on multilocus PCR and MALDI-TOF mass spectrometry. The establishment of an informative panel based on XSNP will allow one to both study genetic variations in populations and carry out DNA identification, while retaining the advantages of X-chromosome markers in complex kinship cases.

EXPERIMENTAL

Materials and experimental procedures. DNA samples from 180 Russian residents of Tomsk city (94 men and 86 women) were studied. SNP markers on the X chromosome were selected using the NCBI database [24] focusing on an even arrangement of the markers along the chromosome with a distance of no less than 1 Mb between two neighboring markers, the lack of linkage between them in the Caucasoid populations from the HapMap project [25], and assuming that the level of genetic diversity (expected heterozygosity, H_e) is no more than 0.40 in these populations.

More than 200 SNPs selected at this stage (when using the Sequenom Assay Design software available on-line on the website www.sequenom.com) formed two multiplexes with a maximum number of the markers compatible in one multiplex. The first multiplex included 36 SNPs (W1, 36-plex), while the second included 30 SNPs (W2, 30-plex). The composition of the multiplexes is given in Table 1. Two PCR primers (forward and reverse) were constructed for each marker, as well as a prolonged primer for the iPLEX reaction. The primers were synthesized by the Evrogen company (Russia). The primer sequences are available upon request to the authors.

Multilocus PCR for producing amplicons that contain the studied SNPs was performed separately for each multiplex in 96-well plates with a volume of 5 μL (Thermocyclers Thermo Scientific, Germany, and Applied Biosystems, United States), in the presence of 5 ng of genomic DNA matrix and PCR primer mixture of corresponding multiplex. PCR Accessory Set reagents supplied with a kit for genotyping by MALDI-TOF mass spectrometry (Agena Bioscience, United States) were used for PCR. The composition of the PCR mixture was 2 mM of $MgCl_2$, 500 μM of each from dNTP, 0.1 μM of primer mixture, and 1 U of DNA polymerase. The PCR scheme was initial denaturation at 94°C (5 min) followed by 42 amplification cycles under the following conditions: denaturation at 94°C (20 s), annealing at 56°C (40 s), and elongation at 72°C (60 seconds), after which the samples were incubated for 5 min at 72°C.

The subsequent stages of the experiment (SAP reaction, iPLEX reaction, transfer of the samples to spectro chip, sample ionization, and analysis of the mass spectra) were performed as previously described [22]. SAP is dephosphorylation of deoxynucleotide triphosphates (dNTPs), which are not included in the amplicons, by an alkaline phosphatase (SAP). During the iPLEX reaction, the multiplex elongation of prolonged primers occurs through the inclusion of this modified dideoxy triphosphate (ddNTP) with altered mass, which is complementary to the nucleotide located in the polymorphic position of each SNP. As a result, a mixture of short PCR products is formed that corresponds to the alleles represented in the multiplex of polymorphic markers. The reaction mixture was purified from salts that accumulated during the multiplex PCR and iPLEX reaction by the treatment with a cationic resin SpectroCLEAN (Agena Bioscience, Germany). The samples were transferred to spectrochip and ionized on a Sequenom MassARRAY 4 device. To analyze real-time mass spectra, as well as for primary processing and the documentation of the experimental results, the MassARRAY TYPER 4.0 software (Agena Bioscience, Germany) was used.

Several samples (two to three of each of three genotypes) were subjected to chain-terminating sequencing according to Sanger with fluorescent labeled terminators of each SNP, followed by capillary gel electrophoresis using a genetic analyzer ABI PRISM 3730 (Life Technologies) for genotype verification. A total of 600 samples were sequenced. Differences in genotype characteristics obtained by MALDI-TOF and direct sequencing were not detected.

Statistical processing of the results. The standard methods of population genetics and mathematical statistics were used. Linkage disequilibrium between pairs and groups of markers was evaluated using coefficient *D*', proposed by Lewontin, and Pearson correlation coefficient (r^2) in the Haploview software package as previously described [26–28], as well as by the exact test using Markov chains implemented in the Arlequin software package [26, 29, 30]. Identification characteristics of X-linked single nucleotide marker

Table 1. Allele frequencies, level of genetic diversity, and discriminatory potential of X-linked SNPs in the XSNPid panel

No.	SNP	Position	MAF	Alleles	Multiplex	He	*PD*f	*PD*m
1	rs2694742	3127322	0.494	G:A	W1	0.4999	0.6250	0.5000
2	rs1405303	4120689	0.411	A:C	W1	0.4842	0.6175	0.4856
3	rs4826682	5119276	0.328	C:A	W1	0.4408	0.5892	0.4394
4	rs5962008	6325520	0.311	A:G	W1	0.4286	0.5777	0.4232
5	rs2130835	7470499	–	–	W1	–	–	–
6	rs2404797	8795378	0.441	A:G	W1	0.4930	0.6213	0.4928
7	rs5934683	9751474	0.378	C:T	W1	0.4702	0.6082	0.4692
8	rs7888207	11916455	0.491	G:T	W2	0.4998	0.6250	0.5000
9	rs952076	13946956	0.357	G:A	W2	0.4591	0.5995	0.4550
10	rs2317327	15407061	0.419	A:G	W2	0.4869	0.6192	0.4888
11	rs4484871	22751065	0.482	A:G	W2	0.4994	0.6244	0.4988
12	rs1351260	26948596	0.427	C:T	W1	0.4893	0.6241	0.4925
13	rs1389433	28128729	0.286	G:T	W2	0.4084	0.5692	0.4118
14	rs225067	29149024	0.314	A:G	W1	0.4308	0.5924	0.4442
15	rs4454452	30985342	0.232	G:A	W2	0.3564	0.5222	0.3564
16	rs3005641	34029930	0.4	T:C	W1	0.4800	0.6130	0.4775
17	rs761913	37859510	0.306	G:T	W1	0.4247	0.5805	0.4270
18	rs5963641	39216082	0.452	T:G	W2	0.4954	0.6230	0.4961
19	rs5917990	40387891	0.375	C:A	W2	0.4688	0.6054	0.4646
20	rs6609159	41573566	0.497	T:C	W1	0.5000	0.6250	0.5000
21	rs205847	42791946	0.444	A:G	W1	0.4937	0.6207	0.4916
22	rs766117	43816206	0.405	G:A	W2	0.4820	0.6159	0.4827
23	rs5953326	49373567	0.483	C:T	W1	0.4994	0.6246	0.4992
24	rs5915291	50379075	0.226	G:T	W2	0.3498	0.5319	0.3669
25	rs4131729	51697194	0.274	T:C	W1	0.3978	0.5583	0.3978
26	rs11799030	53056518	0.453	T:C	W1	0.4956	0.6230	0.4961
27	rs4826609	54765913	0.383	A:G	W1	0.4726	0.6107	0.4736
28	rs6624701	63749154	0.194	G:T	W1	0.3127	0.4889	0.3224
29	rs471205	66238317	0.196	G:A	W1	0.3152	0.4851	0.3188
30	rs5919529	67358208	0.268	C:T	W2	0.3924	0.5625	0.4032
31	rs5937091	70745323	0.206	A:G	W1	0.3271	0.5053	0.3387
32	rs2207739	75644692	0.259	A:G	W2	0.3838	0.5450	0.3819
33	rs2411976	78383858	0.419	A:G	W1	0.4869	0.6195	0.4893
34	rs5969528	81668873	0.394	A:G	W1	0.4775	0.6113	0.4745
35	rs5922869	83049688	0.262	C:T	W2	0.3867	0.5561	0.3951
36	rs5968597	84946832	0.381	A:G	W2	0.4717	0.6094	0.4712
37	rs222108	86910110	0.341	T:C	W1	0.4494	0.6250	0.5000
38	rs1474970	90394689	0.357	T:C	W2	0.4591	0.6034	0.4614
39	rs5941047	91431385	0.425	T:C	W1	0.4888	0.6198	0.4899
40	rs5949581	94756278	0.335	G:A	W2	0.4456	0.5942	0.4469
41	rs5921682	100130437	0.476	G:A	W2	0.4988	0.6248	0.4996
42	rs4898334	101387968	0.447	G:A	W1	0.4944	0.6226	0.4952
43	rs5945770	102594936	0.483	T:C	W1	0.4994	0.6249	0.4998
44	rs1285715	106308416	0.476	G:T	W2	0.4988	0.6248	0.4996

Table 1. (Contd.)

No.	SNP	Position	MAF	Alleles	Multiplex	He	*PD*f	*PD*m
45	rs5973840	107319029	0.461	T:C	W1	0.4970	0.6226	0.4952
46	rs5974348	112218701	0.483	T:C	W1	0.4994	0.6246	0.4991
47	rs7058109	113228559	0.317	A:G	W1	0.4330	0.5853	0.4338
48	rs9329406	115740985	0.411	A:G	W1	0.4842	0.6171	0.4849
49	rs217937	118511843	0.406	T:C	W1	0.4823	0.6153	0.4816
50	rs5909923	121978166	0.5	C:C	W1	0.5000	0.6250	0.5000
51	rs5977571	124496989	0.497	G:A	W2	0.5000	0.6250	0.5000
52	rs5974708	125675579	0.444	T:C	W1	0.4937	0.6211	0.4923
53	rs4830049	126759219	0.363	T:C	W1	0.4625	0.6051	0.4641
54	rs916208	127941251	0.491	G:A	W2	0.4998	0.6246	0.4992
55	rs926640	129462353	0.429	G:T	W2	0.4899	0.6190	0.4884
56	rs2797125	130815979	0.372	G:A	W1	0.4672	0.6085	0.4697
57	rs17391	131910697	0.389	G:A	W2	0.4754	0.6120	0.4758
58	rs5977991	133429960	0.405	A:G	W2	0.4820	0.6178	0.4862
59	rs5975695	135268469	0.408	C:T	W1	0.4831	0.6175	0.4856
60	rs4825220	139214478	0.383	G:T	W2	0.4726	0.6184	0.4872
61	rs4825002	140430391	0.423	A:G	W2	0.4881	0.6110	0.4740
62	rs2869922	141490609	0.437	C:T	W2	0.4921	0.6197	0.4896
63	rs4825213	143435373	–	–	W2	–	–	–
64	rs1781486	144674621	0.274	C:T	W2	0.3978	0.5679	0.4101
65	rs2504169	146358005	0.5	G:G	W2	0.5000	0.6250	0.5000
66	rs614511	149537834	0.47	G:A	W2	0.4982	0.6245	0.4990

Designations: position is the nucleotide position according to the reference human genome sequence, genome assembly 38.2 (GRCh38.p2); MAF is the frequency of the rare allele; alleles are frequent and rare alleles are specified; multiplex is the multiplex number; H_e is the gene diversity; *PD*f is the discriminatory potential for females; *PD*m is the discriminatory potential for males.

system were evaluated using standard indicators for forensic science and forensic medicine [9, 28].

RESULTS

Mass-Spectra Analysis and Genotyping

Figure 1 shows the examples of mass spectra of XSNPid panel W1 and W2 multiplexes for one of the samples. Analyzed molecule are distributed by weight from 4500 to 9000 Da. Here, 108 possible products are detected in the 36-plex W1, i.e., three molecules for each of the SNP, including the original prolonged primer and two products of the iPLEX reaction, which consists of a prolonged primer with an attached nucleotide that corresponds to one of two alternative alleles. In the multiplex W2 composed of 30 SNPs, 90 possible products are detected. The positions of mass spectrum peaks of all possible products are shown by vertical dotted lines. In the present sample (male) the amplification and elongation of all 66 markers was successful, and 36 peaks of specific iPLEX reaction products (36 SNPs in the hemizygous state) for multiplex W1 and 30 peaks (30 SNPs in the hemizygous state) for the second multiplex are observed. Here, the height of the peaks for not prolonged primers do not exceed the background signal intensity, which indicates the high efficiency of iPLEX reaction and the inclusion of most of the primer molecules in specific iPLEX product.

Figure 2 shows genotyping example of certain SNP on the mass spectrum fragment of W1 reaction in the three samples with three different genotypes at rs1405303. The peak of C allele corresponds to the iPLEX reaction product with a mass of 5963 Da, the peak of A allele corresponds to the product with a mass of 5988 Da. Differences of 15 Da between allele-specific molecules enables easily identify the alleles and genotypes even if mass spectrum analyzed manually.

Automatic reading of genotypes according to multiplex mass spectrum in the MassARRAY TYPER 4.0 software package is based on the determination of the ratio of the peak height for iPLEX reaction products of two alleles. An example of sample clustering in accordance with genotypes for rs217937 in the reaction is shown in Fig. 3. In this case, genotypes for all of 95 samples were successfully identified. One sample

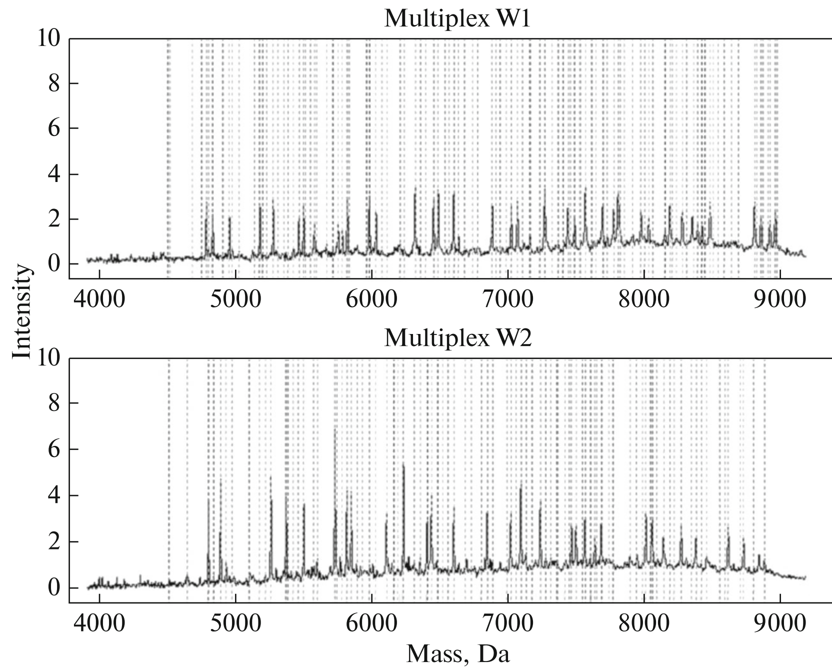

Fig. 1. Mass spectra of XSNPid panel multiplexes W1 and W2 in one of the samples.

without peaks (alleles not found) located at the zero point of coordinate axes corresponds to the control sample without DNA.

Genotyping Efficiency and Evaluation of XSNPid Panel Characteristics

The quality of the developed marker panel (sensitivity, specificity, reproducibility and accuracy of genotyping) were evaluated and genotypes were verified according to the recommendations of the Scientific Working Group on DNA Analysis Methods (SWGDAM) for validation of new DNA identification techniques [31], as well as according to the Promega Company instructions on internal validation of marker systems for forensic laboratories [32].

Panel sensitivity was evaluated by the proportion of defined genotypes to all possible genotypes (call rate) at various concentrations of the DNA matrix (from 20 ng to 500 pg). In the case of two SNPs included in the panel (rs2130835 from W1 (36-plex) and rs4825213 from W2 (30-plex)), a low call rate (less than 60%) was observed and they were excluded from further analysis. Under the standard matrix amount, for our protocol (5 ng of genomic DNA), the call rate for the W1 multiplex is 99.27% and, for the W2 multiplex, it was 97.12% of all possible genotypes. The total call rate for 64 SNPs is 98.29% (11323 of 11520 genotypes were determined).

To evaluate the possibility of our approach using in forensic science, when a small amount of material is available, we tested genotyping efficiency at different DNA concentrations, lower than the standard for our protocol. When reducing the matrix amount in the reaction up to 1 ng the value of call rate falls to 89% in the case of the W1 multiplex and 87% in the case of the W2 multiplex. The proportion of defined genotypes is reduced to 83% at the DNA content of 500 pg in the reaction. Thus, the maximum panel sensitivity is observed under the analysis of 10 ng of DNA matrix (5 ng of each multiplex).

The specificity of the test system was evaluated by determining the intensity ratio of PCR product peaks to nonspecific peaks taking into account the value of the signal to noise ratio (SNR) for ten control samples, which is calculated for each peak in the MassARRAY TYPER 4.0 software. The SNR values ranged between 9.1–35.9; i.e., the proportion of specific product is at least 90% for each of the peaks. Thus, all of the peaks that correspond to alleles included in the SNP panel are uniquely defined, and the specificity of the panel is 100%.

Genotyping accuracy was evaluated by genotype verification using direct sequencing (as described in Experimental section); it is equal to 100%. The reproducibility of the results was evaluated according to the results of a series of repeated genotyping for ten control DNA samples towards all the markers included in the test system. All of the genotypes were reproduced in 100% of experiments.

Analysis of Allele Frequencies and Linkage Disequilibrium

The frequencies of alleles and level of genetic diversity for 64 studied SNPs are given in Table 1. The lowest frequency of the rare allele in the Russian popula-

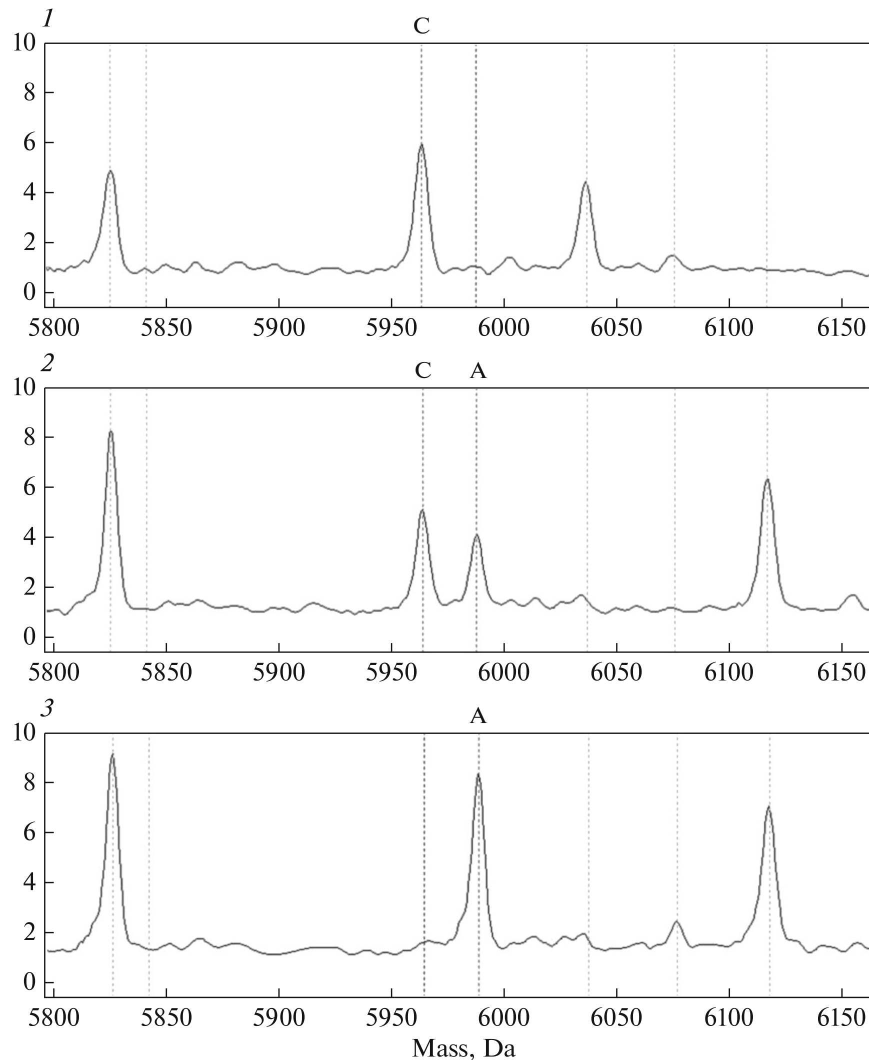

Fig. 2. Mass spectrum fragment of W1 reaction (36-plex) for three samples with three different genotypes at rs1405303. The peak for allele C corresponds to the product of an iPLEX reaction with a mass of 5963 Da, the peak of the allele A is the product with a mass of 5988 Da. (*1*) C; (*2*) CA; (*3*) A.

tion is typical of rs6624701 (0.194), the highest is typical of rs5909923 and rs2504169 (0.5). Gene diversity (H_e) varies from 0.31 to 0.5. The average gene diversity level for 64 SNPs is 0.461. The level of genetic variation for the Russian population towards studied markers corresponds to the level of diversity in the Caucasoid populations (from the HapMap project, 0.447–0.462) and greater than that of the Mongoloid and Negroid populations (0.310–0.419). No deviations from Hardy–Weinberg equilibrium in the genotype distribution for females are observed towards any of the markers.

X-chromosome markers for the XSNPid panel were selected based on their even distribution over the chromosome and lack of linkage. The physical distance between neighboring markers varies in the range of 0.993–8.9 Mb (the average distance between two neighboring SNPs is 2.252 Mb). The analysis of linkage disequilibrium using the Haploview software package showed that the linkage is possible for nine pairs of markers (LOD score > 2; $D' = 0.237$–0.507). However, SNPs neither as part of pairs nor in more extensive configurations do not form blocks of linkage disequilibrium. All of the pairs showed possible linkages, consist of markers located at a distance of several tens of millions of base pairs of each other, and are characterized by low values of the Pearson correlation coefficient ($r^2 = 0.016$–0.073). The exact test for linkage disequilibrium taking into account the Bonferroni correction for multiple comparisons, shows that only one marker pair (rs1351260–rs2411976) is characterized by a significant value of linkage disequilibrium ($p = 0.00000$). The distance on the chromosome between these markers is 43.79 Mb and physical linkage between them in the absence of linkage with inter-

mediate markers can be excluded. Most likely, the gametic disequilibrium between these two remote SNPs is explained by the sample characteristics. Nevertheless, we decided to exclude one of the markers of the pair, rs2411976, which is characterized by a lower frequency of the minor allele, from further analysis of differentiating potential of X-linked SNP panel.

Identification Potential of X-Linked SNP Panel

When evaluating the possibility of using the studied marker system for DNA identification, the standard population statistical indexes that characterize the identification potential of marker system were determined. The most important of these indexes is the power of unrelated individual discrimination (*PD*, power of discrimination) that represent the probability of genotype mismatch in unrelated individuals. Due to the features of X chromosome transfer in a number of generations, the identification potential of X chromosome markers is significantly different in both sexes, so the power of discrimination is calculated separately for females (*PD*f) and males (*PD*m). The *PD*m and *PD*f values are shown in Table 1.

In general, 63 X-linked SNP markers of the XSNPid panels have a very high discriminating ability towards the Russian population. The power of discrimination (*PD*f) in unrelated females for the whole markers is 0.999999999999999999999999991, and the mismatch probability of multilocus genotypes in unrelated males is 0.999999999999999998. The genotype coincidence probability for two unrelated individuals (*PI*, probability of identity) is 9×10^{-27} and 2×10^{-18}, respectively. It has been shown in our previous work that the set of 15 STR loci, including markers of the Combined DNA Index System (CODIS) and European Standard Set (ESS) standards, also provides a very high power of discrimination for unrelated individuals ($PI = 2.81 \times 10^{-17}$ for the Russian population of Tomsk and $PI = 3.04 \times 10^{-15}$–$1.56 \times 10^{-17}$ for other Russian populations) [9]. A comparison of the discriminatory potential of these two systems in one population shows that the XSNPid panel has a much higher identification ability even for men than a standard set of 15 autosomal STR loci.

A comparison of the informativeness of the used panel and its identification potential in the world's populations of the HapMap project (Table 2) shows that the XSNPid system is characterized by a high average heterozygosity in the Caucasoid and Mongoloid populations and a relatively low diversity in the Negroid population. The power of discrimination for unrelated individuals in the European and Asian populations is comparable to that of the Russian population investigated in the present study. The probability of the genotype coincidence in the case of Europeans and Asians ranges from 2×10^{-24} to 5.7×10^{-27} for men and 4×10^{-16} to 1.3×10^{-18} for women. At the same time, the probability of genotype coincidence for unrelated individuals is much higher in African populations.

Fig. 3. Sample clustering according to genotypes in the MassARRAY TYPER 4.0 program using rs217937 as an example.

DISCUSSION

Until recently, only the panels of microsatellite markers have been used for DNA identification of unknown individuals in forensic science and forensic medicine. Current official marker standards for DNA identification accepted in United States (CODIS) and Europe (ESS) are composed of 13 and 16 STR markers, respectively. These standards and extended sets of STR markers for their genotyping, containing up to 23 microsatellites, provide a high and sufficient for practice probability of identification ($PI = (3–5) \times 10^{-16}$ for CODIS and ESS according to Promega) [33]. However, the complexity of genotyping and multiplexing of STR loci coupled with the development of new technologies and progress in the description of the genetic variability in the human genome, obviously, will lead to the replacement of STR markers by

Table 2. Average expected heterozygosity and discriminatory potential of 66 XSNPid panel markers in populations of the world

Population	H_e	Probability of genotype coincidence (*PI*) for women	Probability of genotype coincidence (*PI*) for men
Europeans (CEU)	0.458	6.5×10^{-27}	2×10^{-18}
Italians (TSI)	0.461	5.7×10^{-27}	1.3×10^{-18}
Mexicans (MEX)	0.43	2×10^{-25}	5.9×10^{-17}
Japanese (JPT)	0.406	5×10^{-24}	4×10^{-16}
Chinese (CHB)	0.414	2×10^{-24}	2×10^{-16}
Chinese (CHD)	0.421	8×10^{-25}	8×10^{-17}
Indians (GIH)	0.445	3×10^{-26}	9×10^{-18}
African Americans (ASW)	0.371	3×10^{-22}	2×10^{-14}
Yoruba (YRI)	0.309	1.2×10^{-18}	4×10^{-12}

easy-to-handle systems. SNP marker panels for forensic science and forensic medicine, although are not included in the official standards, but developed actively in the world. Comparative characteristics of some panels for DNA identification based on different technologies are shown in Table 3.

The first of the developed SNP panels included 40–50 SNPs, which also provide the probability of identifying comparable with the standard STR panels, but the processibility of their use was imperfect. For example, a panel of 49 SNPs, proposed previously [34, 35], uses the Genplex technology based on capillary gel electrophoresis as STR fragment analysis does. A set of 45 unlinked autosomal SNPs designed previously provides the probability of identity (*PI*) ranging from 10^{-16} to 10^{-19}; however, it is based on monoplex amplification by TaqMan probes [36, 37].

The SNPforID panel proposed by the European consortium EU GROWTH, includes 52 markers and genotyped in a single multiplex PCR and single base extension (SBE) reactions, as our panel does. The final detection of the products is performed using capillary gel electrophoresis [38]. The Sequenom Company developed the iPLEX Sample ID panel based on the SNPforID panel and included 47 of 52 SNPs into it [39]. This panel, which is genotyped in one multilocus reaction, is the first set for identification using tMALDI-TOF mass spectrometry for DNA. However, the iPLEX Sample ID panel is designed for identifying samples in biobanks, not identifying DNA in forensic science. It can potentially be used for sample identification in criminal cases, but as shown later, the value of panel call rate is sharply reduced in the analysis of degraded or contaminated samples and when the matrix amount less than 10 ng [40]. In contrast to the aforementioned study results, our data indicate the high efficiency of our XSNPid panel with a standard DNA amount of 5 ng, as well as the fairly significant efficiency in the case of the much smaller amount of DNA (0.5–1 ng).

The emergence of relatively inexpensive instruments that use the technology of massive parallel sequencing (MPS) resulted in attempts to develop a SNP marker panel for DNA identification based on MPS. There are two sets available on the market now, i.e., HID-Ion AmpliSeq Identity Panel, which is designed for personal MPS sequencers of Life Technologies [21, 41], and ForenSeq DNA Signature Prep Kit for Illumina devices [42]. The first set consists of 124 SNP markers, including 48 SNPs from the SNP for ID panel, and 43 markers from the K. Kidd panel and 34 Y chromosome SNPs, which enables to determine the sample belonging to the Y chromosome clade of upper layer. The cumulative probability of genotype coincidence for this panel varies in the range from 1×10^{-31} to 1×10^{-33} for different populations [39] (Table 2).

The second panel, ForenSeq DNA Signature Prep Kit, consists of 230 various markers (autosomal, Y- and X-linked STRs, 94 SNPs for identification, SNPs for determining bio-geographical origin and prediction of certain phenotypes).

Despite the significant progress in the development of approaches to DNA identification and the emergence of SNP panels for this purpose, a niche for X chromosome systems remains free both due to the range of tasks that can be solved with their help and the status of practical developments. The XSNP panels designed based on traditional technologies (real-time PCR, capillary gel electrophoresis) include not more than 25 markers and do not reach the level of informativeness for systems based on STR markers [16–18]. The closest analogue of our development is the panel of 52 SNP markers proposed by Chinese authors [43] (Table 2). This panel is also based on MALDI-TOF mass spectrometry; however, it has a small capacity of multiplexes (17–18 markers) and assumes carrying out

Table 3. Comparative characteristics of some panels for DNA identification

Panel	Marker type	Number of marker*	Technology	Probability of genotype coincidence (PI)	References
CODIS	aSTR	14/13	Fragment analysis/Capillary gel electrophoresis	5.02×10^{-16}	[44]
ES	aSTR	17/16	Fragment analysis/Capillary gel electrophoresis	3.04×10^{-16}	[44]
PowerPlex 16	aSTR	16/15	Fragment analysis/Capillary gel electrophoresis	2.81×10^{-17}	[9]
PowerPlex Fusion	aSTR	24/23	Fragment analysis/Capillary gel electrophoresis	6.58×10^{-29}	[44]
SNPforID	aSNP	52	SBE/Capillary gel electrophoresis	$10^{-17} \times 10^{-20}$	[38]
Genplex 49-plex	aSNP	49	Genplex/Capillary gel electrophoresis	$10^{-17} \times 10^{-19}$	[35]
K. Kidd	aSNP	45	TaqMan	$10^{-16} \times 10^{-19}$	[36, 37]
iPLEX Pro Sample ID Panel	aSNP	52/44	MALDI-TOF MS	–**	[39]
HID-Ion AmpliSeq Identity Panel	aSNP + YSNP	124/90	MPS	$1 \times 10^{-31} - 1 \times 10^{-33}$	[41]
ForenSeq DNA Signature Prep Kit	aSTR +YSTR + XSTR + SNP	230/94	MPS	ND	[42]
53 XSNP	XSNP	67/52	MALDI-TOF MS	6.9×10^{-16} to 1×10^{-19}	[43]
XSNPid	XSNP	66/63	MALDI-TOF MS	2×10^{-18} to 9×10^{-27}	Present work

* The total number of markers and the number of markers for DNA identification is specified. Some panels also include additional markers for sex determination; evaluation of DNA amount and other purposes.

** Not suitable for DNA identification in forensic science and forensic medicine. The manufacturer positions the panel as a system for identification of DNA samples in biobanks. Abbreviations: aSTR are autosomal microsatellite markers; MALDI-TOF MS is matrix-assisted laser desorption/ionization–time of flight mass spectrometry; MPS is massive parallel sequencing; SBE is single base extension; XSTR is X-linked STR; YSTR is Y-linked STR; ND indicates no data.

four separate multiplex PCR and four iPLEX reactions. Our panel has a greater number of markers, which are combined only in two multiplexes. Remarkably that our panel intersects with the mentioned one [43] only in one marker (rs471205). When using our panel the discrimination probability for unrelated individuals is several orders higher than when using Chinese one. For the Chinese panel the probability of random genotype coincidence (*PI*) for women is 1×10^{-19} (for our panel, it was 9×10^{-27}), and for men, it was 6.9×10^{-16} (for our panel, it was 2×10^{-18}).

Thus, X-linked single nucleotide marker panel, represented in this study, is superior to all existing analogues (except HID-Ion AmpliSeq Identity Panel based on MPS) with respect to the discriminatory potential under individual DNA identification in women, and exceeds the existing standard for STR systems in men. However, it should be noted that, in practice, the threshold for discriminating potential that can be achieved by using markers of the CODIS and ESS standards (10^{-16}) is sufficient for the vast majority of DNA identification tasks. When comparing with the closest analogues of X chromosome SNP sets, the XSNPid panel also has a higher capacity of multiplexes in addition to a higher discriminatory potential, which makes it cheaper and less time-consuming. The XSNPid panel can become universally applicable, since it provides a high discriminating potential in the analysis of at least major world populations of European and Asian origin and can be used in more than just Russia.

Considering the panel sensitivity indexes and the need for relatively large amounts of DNA for analysis, the panel cannot be recommended for solving forensic

problems (e.g., for the DNA analysis of extremely small amounts of biological material from the crime scene). However, it could be used in cases when the biological material is sufficient, e.g., mass disasters, terrorist attacks and military conflicts).

The XSNPid panel can be used not only for individual identification, but also for complex cases of kinship determining. A high level of genetic diversity of markers, included in the panel, and the panel processability also makes it a useful tool for genetic studies of population. An example of such a study will be represented in one of our next papers.

ACKNOWLEDGMENTS

The work was financially supported by the State Contract no. 14.604.21.0019, dated June 17, 2014 in the framework of the Federal Targeted Program for Research and Development in Priority Areas of Advancement of the Russian Scientific and Technological Complex for 2014–2020.

REFERENCES

1. Stepanov V.A. 2002. *Etnogenomika naseleniya Sibiri i Srednei Azii* (Ethnogenomics of the Population of Siberia and Central Asia). Tomsk: Pechatnaya Manufaktura.
2. Stepanov V.A., Khar'kov V.N., Puzyrev V.P. 2006. The evolution and phylogeography of human Y-chromosome lineages. *Inf. Vestn. Vavilovskogo O-va. Genet. Sel.* **10**, 57–73.
3. Khitrinskaya I.Yu., Khar'kov V.N., Stepanov V.A. 2010. Genetic diversity of the chromosome X in aboriginal Siberian populations: The structure of linkage disequilibrium and haplotype phylogeography of the *ZFX* locus. *Mol. Biol.* (Moscow). **44** (5), 709–719.
4. Vagaitseva K.V., Kharkov V.N., Cherpinskaya K.V., Khitrinskaya I.Yu., Stepanov V.A. 2015. Genetic variability of X-linked STR markers in Siberian populations. *Mol. Biol.* (Moscow). **49** (2), 267–274.
5. Dobyns W.B., Filauro A., Tomson B., Chan A.S., Ho A.W., Ting N.T., Oosterwijk J.C., Ober C. 2004. Inheritance of most X-linked traits is not dominant or recessive, just X-linked. *Am. J. Med. Genet.* **129**, 136–143.
6. Burdett T., Hall P.N., Hasting E., Hindorff L.A., Junkins H.A., Klemm A.K., MacArthur J., Manolio T.A., Morales J., Parkinson H., Welter D. 2015. A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies. Accessed July 18, 2015.
7. Szibor R. 2007. X-chromosomal markers: Past, present and future. *Forensic Sci. Int. Genet.* **1**, 93–99.
8. Lancia M., Severini S., Coletti A., Margiotta G., Dobosz M., Carnevali E. 2011. Using X-chromosomal markers in rape investigation. *Forensic Sci. Int. Genet.* **3**, e55–e56.
9. Stepanov V.A., Balanovskii O.P., Melnikov A.V., Kharkov V.N., Tyazhelova T.V., Akhmetova V.L., Zhukova O.V., Shneider Yu.V., Shil'nikova N.N., Borinskaya S.A., Marusin A.V., Spiridonova M.G., Simonova K.V., Khitrinskaya I.Yu., Radzhabov M.O., et al. 2011. Characteristics of populations of the Russian Federation over the panel of fifteen loci used for DNA identification and in forensic medical examination. *Acta Naturae.* **3**, 59–71.
10. Stepanov V.A., Melnikov A.V., Lash-Zavada A.Y., Kharkov V.N., Borinskaya S.A, Tyazhelova T.V., Zhukova O.V., Schneider Y.V., Shil'nikova I.N., Puzyrev V.P., Rybakova A.A., Yankovsky N.K. 2010. Genetic variability of 15 autosomal STR loci in Russian populations. *Legal Medicine*. **12**, 256–258.
11. Shorokhova D.A., Stepanov V.A., Udovenko Yu.D., Hovoselov V.P., Puzyrev V.P. 2005. Genetic variation and discriminating power of four DNA microsatellites in the Russian population. *Mol. Biol.* (Moscow). **39** (6), 845–850.
12. Costa A., Morais P., Vieira da Silva C., Matos S., Marques dos Santos R., Espinheira R, Costa Santos J., Amorim A. 2014. X-chromosome STR markers data in a Cabo Verde immigrant population of Lisboa. *Mol. Biol. Rep*. **41**, 2559–2569.
13. Gill P., Ivanov P.L., Kimpton C., Piercy R., Benson N., Tully G., Evett I., Hagelberg E., Sullivan K. 1994. Identification of the remains of the Romanov family by the DNA analysis. *Nat. Genet*. **6**, 130–135.
14. Gill P., Werrett D.J., Budowle B., Guerrieri, R. 2004. An assessment of whether SNPs will replace STRs in national DNA databases: Joint considerations of the DNA working group of the European Network of Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDAM). *Sci. Justice*. **44**, 51–53.
15. Krawczak M. 1999. Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis*. **20**, 1676–1681.
16. Li C., Zhang S., Zhao S., Liu Y., Lin Y. 2010. Analysis of 14 highly informative SNP markers on X chromosome by TaqMan SNP genotypoing assay. *Forensic Sci. Int.: Genetics*. **4**, e145–e148.
17. Tomas C., Sanchez J.J., Castro J.A., Børsting C., Morling N. 2010. Forensic usefulness of a 25 X-chromosome single-nuleotide polymorphism set. *Transfusion*. **50**, 2258–2265.
18. Pereira V., Tomas C., Amorim A., Morling N., Gusmao L., Prata M.J. 2011. Study of 25 X-chromosome SNPs in Portugese. *Forensic Sci. Int.: Genetics*. **5**, 336–338.
19. Berglund E.C., Kiialainen A., Syvänen A.-C. 2011. Next-generation sequencing technologies and applications for human genetic history and forensics. *Invest. Genet*. **2**, 23.
20. Yaran Yang, Bingbing Xie, Jiangwei Yan. 2014. Application of next-generation sequencing technology in forensic science. *Genom. Proteom. Bioinform*. **12**, 190–197.
21. Churchill J.D., Chang J., Ge J., Rajagopalan S.C. Wootton C.W. 2015. Blind study evaluation illustrates utility of the Ion PGM system for use in human identity DNA typing. *Croat. Med. J.* **56**, 218–229.
22. Stepanov V.A., Trifonova E.A. 2013. Multiplex SNP genotyping by MALDI–TOF mass spectrometry: Frequencies of 56 immune response gene SNPs in human populations. *Mol. Biol.* (Moscow). **47**, 852–862.

23. *Electrospray and MALDI Mass Spectrometry: Fundamentals, Instrumentation, Practicalities, and Biological Applications,* 2nd ed. Ed. Cole R.B. Wiley, 2012.
24. http://www.ncbi.nlm.nih.gov/projects/SNP/
25. http://hapmap.ncbi.nlm.nih.gov/
26. Guo S., Thompson E. 1992. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics*. **48**, 361–372.
27. Trifonova E.A., Eremina E.R., Urnov F.D., Stepanov V.A. 2012. Genetic diversity and structure of the MTHFR gene linkage disequilibrium in populations of northern Eurasia. *Acta Naturae*. **4**, 80–96.
28. Stepanov V.A., Kharkov V.N., Trifonova E.A., Marusin A.V. 2014. *Metody statisticheskogo analiza v populyatsionnoi i evolyutsionnoi genetike cheloveka. Uchebno-metodicheskoe posobie* (Methods of Statistical Analysis in Human Population and Evolutionary Genetics: A Manual), Ser. Nasledstvennost' i zdorov'e (Heredity and Health Book Series), vol. 12, Tomsk: Pechatnaya Manufaktura.
29. Slatkin M. 1994. Linkage disequilibrium in growing and stable populations. *Genetics* **137**, 331–336.
30. Excoffier L., Lischer H. 2011. *Arlequin v. 3.5: An Integrated Software Package for Population Genetics Data Analysis*. Swiss Institute of Bioinformatics.
31. *Validation Guidelines for DNA Analysis Methods,* Scientific Working Group on DNA Analysis Methods, 2012.
32. Promega. 2013. *Internal Validation Guide of Autosomal STR Systems for Forensic Laboratories: Reference Manual*.
33. https://worldwide.promega.com/products/genetic-identity/str-analysis-for-forensic-and-paternity-testing/powerplex-fusion-str-kits/
34. Sanchez J.J., Phillips C., Borsting C., Balogh K., Bogus M., Fondevila M., Harrison C.D., Musgrave-Brown E., Salas A., Syndercombe-Court D., Schneider P.M., Carracedo A., Morling N. 2006. A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis*. **27**, 1713–1724.
35. Philips C., Fang R., Ballard D. Fondevila M., Harrison C., Hyland F., Musgrave-Brown E., Proff C., Ramos-Luis E., Sobrino B., Carracedo A., Furtado M.R., Syndercombe-Court D., Schneider P.M.; SNPforID Consortium. 2007. Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel. *Forensic Sci. Int. Genetics*. **1**, 180–185.
36. Pakstis A.J., Speed W.C., Fang R. Hyland F.C.L, Furtado M.R, Kidd J.R, Kidd K.K. 2010. SNPs for a universal individual identification panel. *Hum. Genet*. **127**, 315–324.
37. Kidd K.K. 2011. *Population Genetics of SNPs for Forensic Purposes (Updated)*. NIJ Final Report.
38. Musgrave-Brown E., Ballard D., Balogh K., Bender K., Berger B., Bogus M., Børsting C., Brion M., Fondevila M., Harrison C., Oguzturun C., Parson W., Phillips C., Proff C., Ramos-Luis E., et al. 2007. Forensic validation of the SNPforID 52plex assay. *Forensic Sci. Int. Genet.* **1**, 186–90.
39. Agena Bioscience. 2014. *iPLEX Pro Sample ID Brochure*.
40. Johansen P., Andersen J.D., Borsting C., Morling N. 2013. Evaluation of the iPLEX Sample ID Plus Panel designed for the Sequenom MassARRAY system. A SNP typing assay developed for human identification and sample tracking based on the SNPforID panel. *Forensic Sci. Int. Genet.* **7**, 482–487.
41. Life Technologies. 2014. HID-Ion AmpliSeq Identity Panel Data Sheet. 2 P.
42. Illumina. 2014. *ForenSeq DNA Signature Prep Kit Data Sheet*.
43. Li Y., Liu Y., Lin Y. 2015. Typing of 67 SNP loci on X chromosome by PCR and MALDI-TOF MS. *Res. Genet.* **2015**, ID 374688.
44. Promega. 2013. *PowerPlex® Fusion Systems Data Sheet*.

Translated by D. Novikova