ARTICLE

A new set of markers for human identification based on 32 polymorphic Alu insertions

Ilgar Z Mamedov^{*1}, Irina A Shagina², Marya A Kurnikova², Sergey N Novozhilov², Dmitry A Shagin^{1,2} and Yury B Lebedev¹

A number of genetic systems for human genetic identification based on short tandem repeats or single nucleotide polymorphisms are widely used for crime detection, kinship studies and in analysis of victims of mass disasters. Here, we have developed a new set of 32 molecular genetic markers for human genetic identification based on polymorphic retroelement insertions. Allele frequencies were determined in a group of 90 unrelated individuals from four genetically distant populations of the Russian Federation. The mean match probability and probability of paternal exclusion, calculated based on population data, were 5.53×10^{-14} and 99.784%, respectively. The developed system is cheap and easy to use as compared to all previously published methods. The application of fluorescence-based methods for allele discrimination allows to use the human genetic identification set in automatic and high-throughput formats.

European Journal of Human Genetics (2010) 18, 808–814; doi:10.1038/ejhg.2010.22; published online 24 February 2010

Keywords: human identification; MGM; retroelement insertion polymorphism

INTRODUCTION

Human identification based on genomic DNA analysis and profiling has wide application in many fields including mass disasters, crime detection and paternity identification.^{1,2} The majority of methodological approaches for human identification are based on two types of molecular genetic markers (MGMs): short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs). STRs are multiallelic polymorphic markers, which make them more informative compared to usually biallelic SNPs. STR-based human identification systems have been widely used for many years and became common in forensic studies.³ SNP-based MGMs are mostly biallelic and have therefore less discriminative power compared to STRs. An increasing number of well-characterized SNPs and some unique characteristics of this MGM type made them a useful tool for human identification.^{4–7} Both SNP and STR assays include two steps: (i) PCR amplification of multiple loci containing MGM and (ii) identification of the number of tandem repeats or SNP allele in each amplified locus. The application of PCR at the first stage allows to use small amounts of genomic DNA to be analyzed and permits to analyze highly degraded samples. Many methodological approaches are used at the second stage, most of them involving precise determination of PCR products' length by PAGE or capillary electrophoresis with fluorescent labeling of analyzed DNA fragments, as well as quantitative real-time PCR, MALDI or microarray technology.^{8–10} Although the latter methods are usually automated and have become a routine procedure, they still remain expensive because of the use of special facilities and fluorescent dyes. In the past decade, progress in human genome sequencing and methods of comparative genomics resulted in the identification of a new type of MGMs based on retroelement (RE) insertion polymorphism.^{11–15} Recent studies revealed the existence of several thousands of polymorphic RE insertions in the human genome. This type of

MGMs has several peculiarities making them attractive for human genetic identification purposes: (i) RE insertions are stable, that is there is no specific mechanism for withdrawing or removing an element from its insertion point; (ii) the probability of independent RE insertion in the same genomic locus is negligibly small, and the presence of RE indicates identity by descent; and (iii) simple identification.^{16,17} The only required methods for RE polymorphism typing are locus-specific PCR and agarose gel electrophoresis. There are several families of REs known to be polymorphic in the human genome.¹⁸ Among them, the most convenient type of RE for human identification is Alu Y elements that are highly abundant and only about 300 bp long.

In this study, we developed the first MGM set for human genetic identification based on polymorphic RE insertions. The method is based on simultaneous amplification of 32 individual polymorphic Alu insertions by a set of primers complementary to unique regions flanking each insertion. The set was tested on 90 unrelated individuals drawn from four different Russian populations.

MATERIALS AND METHODS

Genomic DNA samples

A total of 111 human genomic DNA samples were analyzed in this study: 20 samples each were obtained from representatives of the Ukrainian, Mordvinian, Kalmyk and Komi populations. An additional set of 10 samples of genomic DNA taken from individuals of unknown ethnical origin was used for initial screening and testing parameters for PCR amplification. This group of individuals was also included in population studies. All these DNA samples were kindly provided by Dr EK Khusnutdinova (Department of Genomics, Institute of Biochemistry and Genetics, Ufa, Russia). Twenty-one DNA samples from six families from Moscow for validation of 32 Alu set in blind experiment for real case of paternity exclusion were obtained from Centre for Molecular Genetics (Moscow, Russia). Genomic DNA from a buccal scrape sample was

¹Laboratory of Comparative and Functional Genomics, Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry RAS, Miklukho-Maklaya, Moscow, Russia; ²Evrogen JSC, Miklukho-Maklaya, Moscow, Russia

*Correspondence: Dr IZ Mamedov, Laboratory of Comparative and Functional Genomics, Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Miklukho-Maklaya 16/10, Moscow 117997, Russia. Tel: +749 5330 4288; Fax: +749 5330 4288; E-mail: Imamedov@mx.ibch.ru

Received 27 July 2009; revised 15 January 2010; accepted 20 January 2010; published online 24 February 2010

isolated using a DIAtom DNA Prep100 kit (IsoGene, Moscow, Russia). All research on human subjects was conducted in accordance with generally accepted ethical standards.

Selection of highly informative polymorphic Alu

The search for polymorphic Alu insertions suitable for the MGM set was performed in a recently published¹⁸ integrated polymorphic repeats database available at <http://labcfg.ibch.ru/Home.html>. To select polymorphic Alu repeats, we used the following criteria: (i) nine pairs located each on human chromosomes from 1 to 9 and 13 individual insertions each located on chromosomes from 10 to 22 (in total, 31 autosomal Alu polymorphisms). In case of two Alu elements per chromosome, the distance between the insertions was about 50 million base pairs to ensure independent inheritance. In addition, a non-polymorphic Alu insertion present on chromosome X and absent from the homologous locus on chromosome Y was selected for sex identification; (ii) the insertions are polymorphic in human populations with the allele frequencies between 0.25 and 0.75 where the population data were available; (iii) the insertions are located in loci containing few or no repetitive elements of other type.

Primer design and PCR amplification

All primers were designed according to the following criteria: (i) PCR product length is 200–300 bp for empty alleles and 500–600 bp for Alu-containing alleles; (ii) PCR primers are complementary to unique genomic regions (tested by NCBI Blast and UCSC *in-silico* PCR); and (iii) the annealing temperature for all primers is 58°C. The Gene Runner program was used to test and avoid possible primer–dimer formation. The amplification profile for the primer set was as follows: 95°C for 20 s, 58°C for 25 s and 72°C for 40 s, 30 cycles (PTC-200 MJ Research, Waltham, MA, USA). The PCR mixture contained 10 ng genomic DNA in 15 µl of 1× PCR buffer for Encyclo (Evrogen, Moscow, Russia) containing 200 µM of each dNTPs, 0.2 µM each of primers and 0.3 µl of 50× Encyclo polymerase mix (Evrogen).

Testing the set for template DNA input quality and quantity

The amount of template genomic DNA required for genotyping using our primer set was estimated in a separate experiment; 1, 2, 5, 10, 20 or 50 ng genomic DNA per reaction were used as a template for PCR amplification. All other reagents were added to the PCR mixture as described above. Four aliquots of genomic DNA were treated with an ultrasonic horn (Cole-Parmer CP750, Cole-Parmer, Vernon Hills, IL, USA) for 5, 10, 20 and 30 cycles for 3 s each with 10 s intervals at the highest power level. In all, 10 ng of the sonicated genomic DNA from each aliquot was then amplified with the primer set using the same conditions and reagents as described above.

Allele frequencies evaluation

Candidate polymorphic Alu markers were tested for Alu-containing and Alu-lacking alleles in all 90 individuals by locus-specific PCR. At this stage, Alu insertions that did not meet the selected allele frequency limit were discarded, and new markers from the same chromosome were taken and tested for allele frequency.

Statistical calculations

For population studies, the gene counting method was used to calculate the allele frequencies for each locus. Fisher's exact test and the χ^2 test were performed using the Genetic Data Analysis software to evaluate the compliance with Hardy–Weinberg equilibrium and independent inheritance of pairs of polymorphic Alu markers located on the same human chromosome. The probability (P_i) for two randomly selected individuals to have identical multi-locus profiles was estimated using genotype frequencies as described in Ref.¹⁹ P_e was defined as the probability for a random male candidate to be excluded from paternity when the maternal genotype is known.¹⁹

Real-time PCR

To evaluate the use of our 32 Alu marker set with fluorescence-based methods for detecting allelic variants SYBR Green I-based RT–PCR amplification was performed in an Mx3005P quantitative PCR system (Stratagene, La Jolla, CA, USA).

The amplification was performed in a 15 µl final volume containing 10 ng genomic DNA, 1× PCR buffer for Encyclo (Evrogen), 200 µM each of dNTPs, 1:30 000 dilution of SYBR Green I (Invitrogen, Carlsbad, CA, USA), 0.2 µM each of primers (Al-4a For/Rev or Al-8b For/Rev) and 0.3 µl of 50× Encyclo polymerase mix (Evrogen). The thermal profile was a 5-min denaturation step at 95°C followed by 40 cycles of 95°C for 10 s, 58°C for 20 s and 72°C for 30 s. The fluorescent product was detected at the last step of each cycle. After amplification, the melting curve was obtained by heating the product to 95°C followed by cooling to 55°C and then slow heating at 0.2°C/s to 95°C with fluorescence measurement at 0.5°C intervals.

RESULTS

The strategy to select MGMs for human identification in the case of biallelic polymorphism is well known from the SNP-based analysis.^{4,7} Similar to SNPs, polymorphic REs are distributed differently among human populations, some of them being restricted to a particular human group(s), whereas the others are distributed evenly among many populations.^{13,20,21} Evenly distributed polymorphic RE insertions are suitable for human genetic identification. We have used our recently created database PRED¹⁸ to select highly informative Alu insertions for human identification. In this database, we have collected data about polymorphic REs identified in our lab and by other research groups worldwide. The database contains comprehensive information about each polymorphic RE: genome location, population frequencies and references. However, it should be mentioned that the collected data may be partly incomplete. First, some REs suggested to be polymorphic may be not true polymorphic insertions.²² This drawback is usually a consequence of inaccurate primer design resulting in simultaneous amplification of multiple genomic loci both containing and lacking the RE insertion. It is already known that the human genome contains large duplications,²³ and amplification of such loci leads to mimicry of true polymorphisms. Second, many RE insertions were predicted by computational screening, and their polymorphism was not confirmed by subsequent experimental studies.^{11,24} Consequently, the selection of Alu insertions required some additional studies to meet the criteria indicated in 'Materials and methods.' At the first computational stage we selected 31 autosomal loci: 18 on chromosomes from 1 to 9 (2 per chromosome) and 13 each on 1 of the remaining human chromosomes (from 10 to 22). The Alu insertions within these loci met the following criteria (Figure 1): (i) they were located in unique genomic loci having no paralogs in other genomic regions and containing no or a few other repetitive

Figure 1 Strategy of polymorphic Alu selection to create a human genetic identification set.

elements 300 bp upstream and downstream of the insertion point. Those Alu insertions for which the data are available were evenly distributed in human population and their allele frequencies were in the range between 0.25 and 0.75. We also selected an Alu insertion to use for gender determination. For this purpose, we identified two paralogous loci located on human chromosomes X and Y, one of them containing an Alu insertion (on chromosome X) and the other lacking the insertion (on chromosome Y). In this case, amplification of genomic DNA with primers complementary to the sequences flanking the Alu insertion on chromosome Y will give two PCR products if male DNA is used as a template, but only one product when female genomic DNA is used. At the second stage, we designed 32 pairs of primers corresponding to unique genomic sequences flanking the points of selected Alu insertions taking into account that PCR products should be of similar length and primers should have close annealing temperatures (see Materials and methods). At the next stage, each primer pair was tested on a set of genomic DNA samples taken from 10 unrelated individuals of unknown ethnic origin. The final set of selected primers meeting all criteria of PCR amplification and polymorphic Alu frequencies (see below) is listed in Table 1. An example of PCR amplification of all 32 polymorphic Alu loci of a single individual is shown in Figure 2.

By varying the DNA content in a PCR reaction from 50 to 1 ng, we estimated the minimal amount of DNA required for human identification by our method. This test showed that for all concentrations of genomic DNA and for all 32 primers, PCR products were visible on agarose gel after 30 cycles of PCR, and the resulting patterns of polymorphic insertions were the same. We can thus conclude that as little as 32 ng genomic DNA of one individual is sufficient to perform an unambiguous and reproducible human genotyping test. Even less template DNA can be successfully analyzed using a reasonably increased number of PCR cycles. We also performed a human identification test on genomic DNA purified from a buccal scrape. These results suggest a rather high sensitivity of the human genotyping using our set of polymorphic markers and its applicability to various sources of genomic DNA.

In many cases, such as mass disasters, it is necessary to identify victims when their genomic DNA is highly degraded. In genomic PCR, significant DNA degradation will result in preferential amplification of shorter fragments thus leading to misinterpretation of the obtained amplification pattern, especially in the case of heterozygotes. To evaluate the level of DNA degradation that might critically affect the results of genotyping, we performed a special experiment. Genomic DNA was ultrasonically degraded to a different extent as described in 'Materials and methods.' The degradation level was evaluated by agarose electrophoresis. We observed a gradual decrease in the average length of DNA fragments with increasing sonication time, and no difference between samples after 20 and 30 cycles of ultrasonic degradation. The latter means that 20 cycles is sufficient to achieve the maximum degree of genomic DNA fragmentation by means of ultrasonic treatment. We performed amplification with the 32 polymorphic Alu primer set using genomic DNAs degraded for 5, 10, 20 and 30 cycles. The polymorphic Alu pattern was in all cases the same as for non-degraded DNA. These results indicate that even significantly degraded DNA samples could be genotyped correctly using the proposed set of polymorphic markers.

Allele frequencies for each of 31 autosomal polymorphic Alu insertions were analyzed for 80 individuals from genetically and geographically distant Russian populations: the Ukrainians, Mordovians, Kalmyks and Komi. Ten unrelated individuals from Moscow region used for the initial polymorphism evaluation were also

included into the population study. The observed allele frequencies are listed in Table 2. Alu-containing/Alu-lacking allele frequencies varied between 0.29:0.71 and 0.74:0.26 in the set of 90 individuals. However, the allele frequencies for each population analyzed varied more significantly – from 0.075:0.925 to 0.8:0.2 (see Supplementary Table 1) indicating that some populations have their genetic peculiarities and a complex character of ethnic groups formation. We did not observe significant ($P < 0.05$) Hardy–Weinberg disequilibrium for each of 31 autosomal Alu markers (except Alu 8b) using exact Fisher and the χ^2 test for the set of 90 individuals. Pairs of polymorphic Alu insertions located on the same chromosome showed no significant ($P < 0.05$) linkage. The mean match probability was calculated at 5.53×10^{-14} for 32 markers (31 autosomal + 1 X/Y), and the probability of paternal exclusion when the maternal genotype is known for 31 autosomal markers at 99.784%. Thirty-one autosomal polymorphic Alu insertions generally had an even distribution among inhabitants of the Russian Federation. We compared the distribution of allele frequencies for the Alu insertions with previously studied populations.^{13,14,25,26} No significant differences between our and previously published results obtained for genetically and geographically distant populations (see Supplementary Table 1) were observed. These results suggest that the selected 32 Alu set can be successfully used for human genetic identification worldwide.

To validate our 32 Alu marker set in a real paternity identification blind experiment we obtained 21 DNA samples from six families from Centre for Molecular Genetics without any data on paternity and genotyped all individuals using our set. We analyzed obtained Alu profiles in each family marker by marker and as a result in three families fathers were excluded based on mismatch in 4–10 Alu loci (see Supplementary Table 1, families). For the other three families all 32 markers matched perfectly between suspected father, known mother and children. The results of our paternity identification were identical with those obtained by our colleagues from Centre for Molecular Genetics using commercial AmpFLSTR Identifier PCR amplification kit (Applied Biosystems, Carlsbad, CA, USA). The probability of paternal exclusion for real cases varied from 99.9571 to 99.4655% in families with confirmed paternity. To further validate the power of our 32 Alu marker set for individual identification in blinded study we provided all primer pairs along with randomly selected 10 DNA samples from our set of 90 individuals to the other laboratory. After 32 Alu genotyping they were able to correctly identify all the 10 samples tested in our 32 Alu profile database. The 32 Alu genotypes obtained in the other lab were identical to genotypes identified by us in population studies.

Fluorescence-based methods for detecting allelic variants are now widely used in routine laboratory practice.^{3,7,8} The use of fluorescence-based approaches allows performing large-scale human genetic tests and significantly reduces contamination because of keeping laboratory microtubes closed after PCR amplification. Accordingly, we checked the possibility of polymorphic Alu allelic discrimination by means of real-time PCR with Sybr Green dye and dissociation curves analysis using two primer pairs from our set. As a result, we observed two distinct dissociation peaks corresponding to Alu-containing and Alu-lacking PCR fragments (Figure 3). All three possible genotypes could be clearly distinguished by dissociation curve analysis after PCR reaction.

DISCUSSION

Mass human genotyping is of increasing importance. Creation of human genotype databases can resolve many problems related to identification of victims of mass disasters, suspect identification using

Table 1 32 Alu amplification primers and chromosomal location

<i>Alu</i>	<i>Amplification primers</i>	<i>Accession number/position</i>	<i>Product size^a</i>	<i>Locus</i>
Al-1a	GCAGATTCAGGTCATTATTG GGTCACTAGAGACCGTATCTGTA	AK023131.1 1453_1773del	547/224	1q25.3
Al-1b	TTTTGTGTCCATTAGGGTTAC GACCTCACTGTTAGACAATAATATC	AL136124.10 33110_33420del	540/229	1q42.2
Al-2a	TGTTTAAATGATCAAAATTAGG AGATCTCTGGCTAAGTTCTTTAC	AC073577.32 48284_48612del	541/212	2p13.1
Al-2b	ATACTTTAATTTAGGCAATAGACAT TAGAGAAAAGGAGTTTGAATG	AC012492.9 137270_137590del	531/210	2q23.3
Al-3a	ACCAAAATACCCCTCAGAAGA CCTGACCTAGTGATCCCATAG	AY736289 157_483del	534/207	3p22.1
Al-3b	TTTATACAAATACTACAGACAAAAGC CTGATCAAAGCTATAGACCTCTCT	AC107614.3 108968_109298del	549/219	3q11.2
Al-4a	GAGGTTAGACAACATACATTATATCA CCCTATTGTTTACATCTAATTCTA	AC023886.7 89620_89927del	594/287	4q25
Al-4b	TTCTGAACTGTGTACATATGACACT GCTTTTAGTAAAGGTCACACCA	AP002824.1 14206_14536del	596/194	4q13.2
Al-5a	ACGCAAGGACTTAATTGTAAC TAATGTTTGGATTATTCCTGGA	AC114316.2 147755_148109del	551/197	5q14.3
Al-5b	AGGTGCAGTCTTAACCCAGA AAAGCTATTGGAAGATGAAGA	AC034233.5 1030_1355del	555/229	5q12.1
Al-6a	AGCTGATATATCTCAATAAGCAAC GATTAAATCACTGTCTGAGACT	AL031904.1 65549_65886del	555/218	6p24.3
Al-6b	TCACAAGTTCTTATAACTCTTTCC GGACACAATCATGATGATATTAC	AL356454.15 33195_33507del	539/225	6q12
Al-7a	CAGAGTAAAGCTACAGTAAGACAGTA ATAACATAGGGAAAATCACTTCTA	AC004027.1 997_1332del	566/230	7q21.11
Al-7b	GAATTCCAAATGTCTAGAGATTAGT TTGGAGAACCTTGACTTTTCA	AC008039.1 18381_18698del	562/244	7q32.1
Al-8a	TGCTGTTCAGCTCTTCACTA ATACTATTGGTTGACCCTACTG	AP005354.2 118375_118706del	584/252	8q23.1
Al-8b	AGAATTTTGAGTGTGGGTATACT TGCACATTGAAGGACACAATGAGAATAC	AC022555.8 167168_167484	586/269	8q12.1
Al-9a	TTGGAACCTTTGGTAAGTAACT AAAGCCACAGAAATACATCTAAT	AL354726.17 58282_58611del	523/193	9q22.33
Al-9b	ATGCCAAATATTGGTCAGAGA TATACTGGGACAGTGAGTTTAC	AF334830.1 11296_11599del	547/244	9p13.3
Al-10	TGCCAAACAGTGATTATAGAAC CTGTCTAGAAGACAACCTAGGAATAG	AY736292.1 183_495del	588/275	10q23.1
Al-11	CTATGATGTGCTATGGAAGGTAGGT CCTAGTCCTCATAAATTAGTAACCTG	AC080183.7 84023_84345del	631/309	11p14.3
Al-12	ATTTGGGAGTTTCCATTATGA TATTAACCATATTATGTGCCAAGA	AC005868.1 15288_15596del	570/256	12q24.32
Al-13	AGTTACTGGATGCATTATATTGAG CCATGATAGCTAACAAAAGAAG	AY736302.1 97_413del	550/233	13q34
Al-14	TTAAAAATAAGTTCCTAATTGATTC CAAATTAGATAAGTAAGATAGGTAAGA	AL117356.5 8355_8661del	558/251	14q23.1
Al-15	TGGACTACTTTTAATAGTAATTTGTG ATTGCTTACACTGCTACCAAAG	AC091046.2 55784_56101del	564/246	15q22.1
Al-16	GATATCACATACTCATTAGATATAATCA CATATTCAATACTTTTCAGTCATTCT	AC040161.5 145635_145967del	564/231	16q21
Al-17	CTCAGATCAGTTATGCAGACAG AAAGTAAGAGACTTTCCTTTGAG	AC018665.9 136400_136717del	561/244	17q25.1
Al-18	CAAATACAGGAATTATTATAATAATG ATCAATAATGCACAGTAAATCT	AC090241.9 146484_146820del	548/211	18q21.1
Al-19	GTCATACAAAATGATGATGATTG AAAAATATCTAACTTTGAAAATGG	AC092295.2 100287_100602del	578/262	19q13.12
Al-20	CTTTTGCTTCCTACTAAATACCT TTGAGGCATTGCTACTGAGAT	AL109837.22 48313_48636del	551/228	20p12.2
Al-21	AGTTTGGATGTGATGTTATATGTC TGTATACTTGTAGGATACTTGTGT	AP000459.4 38886_39199del	560/246	21q21.2

Table 1 (Continued)

Alu	Amplification primers	Accession number/position	Product size ^a	Locus
AI-22	TCGAGTGAGAAGAGGTAGTGAG TTGAATTTACGCAGCATAATG	AC016026.15 45671_45992del	560/238	22q11.21
AI-XY	CGGTAGAAATATTAAGTAGGGAC GTGTTTCTAACATTCTTACGGT	AL137015.9 46641_46955del	554/239	Xq21.31/Yp11.2

^aPCR product size of Alu-containing/Alu-lacking allele. **Figure 2** Example of PCR amplification of 32 polymorphic Alu loci on genomic DNA of one individual. Gray and white arrows indicate the Alu-containing and Alu-lacking PCR products, respectively.

biological material taken at the scene of a crime, paternity identification, and so on. In several countries, projects of human genotype databases for all population or for some population categories are being discussed. Human genetic passports will include many MGMs of different types including STRs and SNPs. Although these ambitious projects appear to be time, money and labor consuming, they might give serious advantages in many aspects of life. Significant progress in human genome sequencing technologies might result in individual genome sequences in the next 10–20 years. The complete genome information for every individual will stimulate medical genetic studies and provide a lot of personal DNA markers. Nevertheless, a compact universal set of neutral markers to create ‘genomic fingerprints’ would certainly be of use. Therefore, the introduction of a genetic marker set that allows rapid, easy and cheap human identification is a very important and urgent task. Many human identification genetic systems based on STRs and SNPs have been developed and used since human genome structure was published and information on human genome diversity accumulated.^{3,5–7} Initially, the most commonly used human genotypic systems were based on STRs, each of them having multiple alleles with different population frequencies. These markers possess a very strong discrimination power but have a higher mutation rate as compared with SNP or RE^{3,27} and laborious and expensive allele identification. The acquisition of data on human

Table 2 32 Alu allele and genotype frequencies and observed heterozygosity

Alu	fAlu	Genotype frequency			H_o^a	P_i^b
		Alu+/Alu+	Alu+/Alu0	Alu0/Alu0		
1a	0.46	0.22	0.48	0.30	0.478	0.369
1b	0.51	0.28	0.45	0.27	0.446	0.352
2a	0.68	0.46	0.43	0.11	0.429	0.409
2b	0.38	0.16	0.44	0.40	0.438	0.380
3a	0.45	0.24	0.42	0.34	0.424	0.350
3b	0.61	0.40	0.42	0.18	0.416	0.369
4a	0.40	0.16	0.46	0.38	0.462	0.380
4b	0.49	0.28	0.42	0.30	0.416	0.344
5a	0.41	0.24	0.42	0.34	0.411	0.347
5b	0.33	0.08	0.49	0.43	0.506	0.435
6a	0.29	0.10	0.38	0.52	0.378	0.425
6b	0.65	0.43	0.45	0.12	0.449	0.400
7a	0.42	0.19	0.45	0.36	0.463	0.372
7b	0.38	0.19	0.38	0.43	0.382	0.365
8a	0.41	0.15	0.52	0.33	0.505	0.395
8b	0.68	0.52	0.32	0.16	0.326	0.398
9a	0.53	0.33	0.40	0.27	0.402	0.342
9b	0.74	0.57	0.34	0.09	0.337	0.450
10	0.29	0.10	0.38	0.52	0.380	0.427
11	0.61	0.38	0.46	0.16	0.457	0.380
12	0.44	0.21	0.46	0.33	0.462	0.365
13	0.64	0.42	0.43	0.14	0.435	0.389
14	0.41	0.20	0.41	0.39	0.411	0.360
15	0.52	0.28	0.47	0.25	0.472	0.363
16	0.34	0.10	0.48	0.42	0.478	0.417
17	0.47	0.21	0.52	0.27	0.516	0.385
18	0.37	0.14	0.46	0.40	0.456	0.388
19	0.46	0.23	0.46	0.31	0.461	0.362
20	0.49	0.23	0.52	0.25	0.522	0.387
21	0.62	0.37	0.49	0.13	0.495	0.402
22	0.66	0.47	0.38	0.15	0.378	0.385

^a H_o , observed heterozygosity.^b P_i , probability of identity.

SNPs has resulted in systems based on this type of MGMs. Each individual SNP is in most cases biallelic and has less discriminative power than STR. However, the use of a greater number of polymorphic loci provides a sufficiently high level of discrimination comparable with that for STR-based systems.⁷ In addition, SNP markers have a reduced mutation rate as compared with STRs. Multiple methodological approaches have been developed for SNP allele detection based on allele-specific PCR, mass-spectrometry, and so on. In most cases, STR and SNP typing requires locus-specific amplification with a set of primers followed by allele detection by capillary electrophoresis or allele-specific real-time PCR.

Figure 3 Example of dissociation curves analysis after PCR amplification.

In the past decade, a new type of MGMs based on RE insertion polymorphism has become popular in population studies.^{28–32} These markers belong to a biallelic indel (insertion/deletion) polymorphism type and are characterized by several features making them good candidates for human genetic identification. Multiple studies intended to identify new polymorphic RE insertions resulted in an increasing number of candidate MGMs scattered across the whole human genome. The data on genomic location and population distribution of many polymorphic RE insertions are collected in two public databases.^{18,26} Recently, we showed the applicability of RE-based genetic markers to cell line fingerprinting.³³ Thirty-eight dimorphic RE insertions (17 L1 and 21 Alu) were used for genotyping of different cell lines, and unique multilocus profiles were obtained for each cell line. Here, we used 32 highly polymorphic and evenly distributed Alu insertions to create a simple and cheap human genetic identification system. Comparison of the results of our studies and earlier published data showed that each of the selected polymorphic Alu insertions was present in genetically distant populations with compared allele frequencies. These data are encouraging; however, validation of the 32 Alu set on larger human sample panel before wide use is needed. The calculated mean match probability 5.53×10^{-14} indicates that the discrimination power of the proposed set is comparable with that of existing genetic identification systems,^{3,4,6,7} and is sufficient for kinship determination or genetic ‘passportization.’ The probability of paternity exclusion, when the maternal genotype is known, was calculated at 99.784% that is comparable with existing SNP-based genetic tests (98.9% for 22 SNPs and 99.91–99.98% for 52 SNPs). The power of proposed Alu-based set was successfully validated in two blinded experiments on paternity exclusion and individual identification.

Owing to limited amounts of DNA in many crime cases, it is important to perform genetic tests on samples with only traces of genomic DNA. As mentioned above, we have shown that as little as 32 ng of input template is sufficient for typing of all 32 polymorphic Alu insertions, and that limited amounts of DNA do not prevent the correct allele determination. We have also concluded that DNA from

just one buccal scrape is sufficient for correct 32 Alu typing and there is no need to take blood for obtaining human genetic profiles. Ultrasonic DNA degradation was shown to have little effect on the accuracy of the human genotyping using the 32 polymorphic Alu set. However, a deeper DNA degradation can prevent amplification of Alu-containing alleles, and this should be taken into account when analyzing some particular types of human samples (eg fire-affected samples).

We have shown that fluorescence-based dissociation curve analysis after PCR amplification can be successfully used for allele discrimination. Clear separation of two peaks corresponding to Alu-containing and Alu-lacking PCR products is ensured by significant difference in length (about 300 bp) and is additionally enhanced by the high GC content of Alu elements. Thus, the proposed human genotyping system can be easily adapted to high-throughput automatic format and studies where the problem of contamination, such as contamination of PCR products or sample cross-contamination, is critical. Another possible development of the proposed Alu markers system is creation of multiplex reactions where several primer pairs corresponding to different Alu loci can be amplified and analyzed in single PCR reaction. This approach was successfully tested earlier on polymorphic Alu repeats³⁴ and can significantly increase the cost effectiveness of human identification system, however, making data analysis more complex.

In conclusion, in this paper we propose RE-based set for the human identification purposes for the first time. We have shown the use of the proposed set for different forensic applications, however, the package needs to be further developed and carefully validated before routine laboratory use. The proposed approach can be applied as an additional, cheaper and compact variant of existing commercial packages rather than substitute them in all forensic applications.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This research was supported by Russian Foundation for Basic Research Grant RFBR 08-04-12056 ofi and Program of RAS grant. We thank Prof. Elsa K Khusnutdinova (Department of Genomics, Institute of Biochemistry and Genetics, Ufa, Russia) for providing human DNA samples for population studies. We thank Prof. Alexander V Polyakov and Dr Olga A Schagina (Centre for Molecular Genetics Ltd, Moscow, Russia) for providing us DNA samples and help with blinded experiments. We also express our gratitude to Dr B Glotov for helpful and constructive discussion and help in manuscript preparation.

- 1 Budowle B, Bieber FR, Eisenberg AJ: Forensic aspects of mass disasters: strategic considerations for DNA-based human identification. *Leg Med (Tokyo)* 2005; **7**: 230–243.
- 2 Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B: The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. *Electrophoresis* 1999; **20**: 1682–1696.
- 3 Butler JM: Genetics and genomics of core short tandem repeat loci used in human identity testing. *J Forensic Sci* 2006; **51**: 253–265.
- 4 Giardina E, Pietrangeli I, Martone C *et al*: In silico and *in vitro* comparative analysis to select, validate and test SNPs for human identification. *BMC Genomics* 2007; **8**: 457.
- 5 Giardina E, Predazzi I, Pietrangeli I *et al*: Frequency assessment of SNPs for forensic identification in different populations. *Forensic Sci Int Genet* 2007; **1**: e1–e3.
- 6 Lee HY, Park MJ, Yoo JE, Chung U, Han GR, Shin KJ: Selection of twenty-four highly informative SNP markers for human identification and paternity analysis in Koreans. *Forensic Sci Int* 2005; **148**: 107–112.
- 7 Sanchez JJ, Phillips C, Borsting C *et al*: A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis* 2006; **27**: 1713–1724.

- 8 Butler JM, Buel E, Crivellente F, McCord BR: Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *Electrophoresis* 2004; **25**: 1397–1412.
- 9 Ross PL, Belgrader P: Analysis of short tandem repeat polymorphisms in human DNA by matrix-assisted laser desorption/ionization mass spectrometry. *Anal Chem* 1997; **69**: 3966–3972.
- 10 Taylor CY, Guillen IA, Nazabal M, Fernandez JR, Silva JA: Electrophoretic techniques applied to the detection and analysis of the human microsatellite DG10s478. *J Biomol Tech* 2007; **18**: 298–305.
- 11 Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE: Natural genetic variation caused by transposable elements in humans. *Genetics* 2004; **168**: 933–951.
- 12 Mamedov I, Lebedev Y, Hunsman G, Khusnutdinova E, Sverdlov E: A rare event of insertion polymorphism of a HERV-K LTR in the human genome. *Genomics* 2004; **84**: 596–599.
- 13 Mamedov IZ, Arzumanyan ES, Amosova AL, Lebedev YB, Sverdlov ED: Whole-genome experimental identification of insertion/deletion polymorphisms of interspersed repeats by a new general approach. *Nucleic Acids Res* 2005; **33**: e16.
- 14 Otieno AC, Carter AB, Hedges DJ *et al*: Analysis of the human Alu Ya-lineage. *J Mol Biol* 2004; **342**: 109–118.
- 15 Mamedov I, Batrak A, Buzdin A, Arzumanyan E, Lebedev Y, Sverdlov ED: Genome-wide comparison of differences in the integration sites of interspersed repeats between closely related genomes. *Nucleic Acids Res* 2002; **30**: e71.
- 16 Batzer MA, Deininger PL: Alu repeats and human genomic diversity. *Nat Rev Genet* 2002; **3**: 370–379.
- 17 Witherspoon DJ, Marchani EE, Watkins WS *et al*: Human population genetic structure and diversity inferred from polymorphic L1(LINE-1) and Alu insertions. *Hum Hered* 2006; **62**: 30–46.
- 18 Mamedov IZ, Amosov AL, Fisunov G, Lebedev Iu B: A new database on polymorphic retroelements in human genome (PRED). *Mol Biol (Mosk)* 2008; **42**: 721–727.
- 19 Brenner CH, Morris JW: Paternity index calculations in single locus hypervariable DNA probes: validation and other studies; The International Symposium on Human Identification Publisher: Promega, Madison, 1989, pp 21–53.
- 20 Carroll ML, Roy-Engel AM, Nguyen SV *et al*: Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J Mol Biol* 2001; **311**: 17–40.
- 21 Watkins WS, Rogers AR, Ostler CT *et al*: Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. *Genome Res* 2003; **13**: 1607–1618.
- 22 Luis JR, Terreros MC, Martinez L, Rojas D, Herrera RJ: Two problematic human polymorphic Alu insertions. *Electrophoresis* 2003; **24**: 2290–2294.
- 23 Redon R, Ishikawa S, Fitch KR *et al*: Global variation in copy number in the human genome. *Nature* 2006; **444**: 444–454.
- 24 Wang J, Song L, Gonder MK *et al*: Whole genome computational comparative genomics: a fruitful approach for ascertaining Alu insertion polymorphisms. *Gene* 2006; **365**: 11–20.
- 25 Carter AB, Salem AH, Hedges DJ *et al*: Genome-wide analysis of the human Alu Yb-lineage. *Hum Genomics* 2004; **1**: 167–178.
- 26 Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P: dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum Mutat* 2006; **27**: 323–329.
- 27 Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD: The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res* 2008; **18**: 30–38.
- 28 Antunez-de-Mayolo G, Antunez-de-Mayolo A, Antunez-de-Mayolo P *et al*: Phylogenetics of worldwide human populations as determined by polymorphic Alu insertions. *Electrophoresis* 2002; **23**: 3346–3356.
- 29 Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB: Human population genetic structure and inference of group membership. *Am J Hum Genet* 2003; **72**: 578–589.
- 30 Dornelles CL, Battilana J, Fagundes NJ, Freitas LB, Bonatto SL, Salzano FM: Mitochondrial DNA and Alu insertions in a genetically peculiar population: the Ayoreo Indians of Bolivia and Paraguay. *Am J Hum Biol* 2004; **16**: 479–488.
- 31 Nasidze I, Risch GM, Robichaux M, Sherry ST, Batzer MA, Stoneking M: Alu insertion polymorphisms and the genetic structure of human populations from the Caucasus. *Eur J Hum Genet* 2001; **9**: 267–272.
- 32 Ray DA, Walker JA, Hall A *et al*: Inference of human geographic origins using Alu insertion polymorphisms. *Forensic Sci Int* 2005; **153**: 117–124.
- 33 Ustyugova SV, Amosova AL, Lebedev YB, Sverdlov ED: Cell line fingerprinting using retroelement insertion polymorphism. *Biotechniques* 2005; **38**: 561–565.
- 34 Kass DH: Generation of human DNA profiles by Alu-based multiplex polymerase chain reaction. *Anal Biochem* 2003; **321**: 146–149.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)