Rui Pereira[1,2]
Christopher Phillips[2]
Cíntia Alves[1]
António Amorim[1,3]
Ángel Carracedo[2,4]
Leonor Gusmão[1]

[1]Institute of Molecular Pathology and Immunology, University of Porto, Porto, Portugal
[2]Institute of Legal Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain
[3]Faculty of Sciences, University of Porto, Porto, Portugal
[4]Genomic Medicine Group, CIBERER, University of Santiago de Compostela, Santiago de Compostela, Spain

Received April 28, 2009
Revised July 1, 2009
Accepted July 8, 2009

Research Article

A new multiplex for human identification using insertion/deletion polymorphisms

Human identification is usually based on the study of STRs or SNPs depending on the particular characteristics of the investigation. However, other types of genetic variation such as insertion/deletion polymorphisms (indels) have considerable potential in the field of identification, since they can combine the desirable characteristics of both STRs and SNPs. In this study, a set of 38 non-coding bi-allelic autosomal indels reported to be polymorphic in African, European, and Asian populations were selected. We developed a sensitive genotyping assay, which is able to characterize all 38 bi-allelic markers using a single multiplex PCR and detected with standard CE analyzers. Amplicon length was designed to be shorter than 160 bp. Complete profiles were obtained using 0.3 ng of DNA, and full genotyping of degraded samples was possible in cases where standard STR typing had partially failed. A total of 306 individuals from Angola, Mozambique, Portugal, Macau, and Taiwan were studied and population data are presented. All indels were polymorphic in the three population groups studied and the random match probabilities of the set ranged in orders of magnitude from 10^{-14} to 10^{-15}. Therefore, the indel-plex represents a valuable approach in human identification studies, especially in challenging DNA cases, as a more straightforward and efficient alternative to SNP typing.

Keywords:
Forensic genetics / Human identification / Insertion/deletion polymorphism / Multiplex PCR
DOI 10.1002/elps.200900274

1 Introduction

The study of genetic variation, using DNA polymorphisms distributed throughout the genome, has allowed better understanding of the history and diversity of human populations as well as providing a system for the genetic identification of individuals. Insertion/deletion polymorphisms (indels) are length polymorphisms created by insertions or deletions of one or more nucleotides in the genome. Only in the last few years indels have received the attention of major studies. In 2002, Weber *et al.* [1] identified and characterized 2000 human bi-allelic indels which varied greatly in length of alleles observed and highlighted the utility of indels for genetic studies, with reference to their abundance and ease of analysis. Since then, a number of studies have been published using indels for a wide range of purposes including ancestry affiliation [2], addressing the genetic structure of human populations [3, 4], and their use as genetic markers in natural populations [5]. In 2006, Mills *et al.* [6] conducted a milestone study to identify indels, and reported an initial map of indel variation in the human genome containing more than 415 000 unique polymorphisms, with an average density of one indel *per* 7.2 kb. A particular class including insertions/deletions of apparently random DNA sequences represents ~41% of all indels and harbors polymorphisms with a wide range of allele length variation from 2 bp up to ~10 kb, with nearly all of these under 100 bp [6]. These small length indels are amenable to analysis through a simple PCR amplification and electrophoresis, or even using methodologies already developed for SNPs without the need to use direct sequencing methods [1, 2, 6].

At present, STR typing can be considered a standard approach and the method of choice in the forensic field, allowing a high discrimination power adequate for addressing most problems of human identification [7, 8]. As alternatives to STRs, the use of SNPs has proved valuable in specific applications, mainly in the analysis of highly degraded samples. As a consequence, various SNP sets have been selected for individual identification [9–15] or assignment of population of origin [2, 16], although some criticisms have been raised concerning their isolated use in forensics [17].

By combining many of the desirable features of STRs and SNPs, indels can bridge the gap between these

Correspondence: Dr. Rui Pereira, IPATIMUP, Rua Dr. Roberto Frias, s/n, 4200–465 Porto, Portugal
E-mail: rpereira@ipatimup.pt
Fax: +351-225570799

Abbreviations: DP, discrimination power; **indel,** insertion/deletion polymorphism; **RMP,** random match probability

established strategies, as they show the following characteristics: (i) a widely spread distribution throughout the genome [1, 6]; (ii) origination from a single mutation event which occurs at a low frequency and is subsequently stable (unlikely to present recurrent mutations) [18]; (iii) significant differences in allele frequencies among geographically separated population groups, hence they have potential as ancestry informative markers [1, 2]; (iv) small indels can be analyzed in short amplicons, opening perspectives for large-scale multiplexing capability while improving the chances of successful amplification of highly degraded DNA; (v) the genotyping of small indels is relatively easy and inexpensive with a simple PCR and standard dye-linked CE systems; and (vi) small indels are also suitable for automation and analysis with high-throughput technologies [1, 6].

Considering the potential of indels as genetic markers in forensic analysis, we aimed to develop a new multiplex for human identification combining a number of desirable features found in SNPs or STRs. An assay including bi-allelic indels with defined short-interval allele length variation (2–5 bp) allows the benefit of using small amplicon sizes, in the same way as SNPs, thus improving the successful analysis of degraded DNA samples while at the same time using a straightforward system of analysis through a PCR and direct electrophoretic detection of the amplified alleles, in the same way as STRs, taking advantage of methodologies already well established in forensic laboratories with no additional requirements.

2 Materials and methods

2.1 Marker selection

The initial candidate pool of markers for this study was based on the previous work by Weber *et al.* [1]. We started from a list of ~4000 bi-allelic indels previously confirmed and characterized in major population groups, available through the Marshfield diallelic indels database website (http://www.marshfieldclinic.org/mgs/pages/default.aspx?page = didp). Markers were then selected according to the following criteria: (i) non-coding, autosomal bi-allelic indels; (ii) minimum allele frequency ≥ 0.25 in European, African, and Asian population groups; (iii) average heterozygosity ≥ 0.40; and (iv) allele length variation of 2–5 bp.

After applying the described criteria, we obtained ~220 candidate markers. Flanking sequences of these indels (+/−150 bp), as well as reported sequence variants within this region, were obtained using the University of California Santa Cruz Genome Browser (Human March 2006 Assembly; SNPs 128 track) at http://genome.ucsc.edu/. Subsequently, flanking sequences presenting known polymorphisms or mononucleotide repeats (≥ 7 bp) were identified, and in many cases removed, in order to avoid primer binding problems and amplicon length variations other than those expected.

2.2 Primer and multiplex assay design

Primer design was performed using the Primer3 software (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi), applying the following main criteria for the PCR primers: amplicon size of 60–160 bp (as obtained from the input sequence); optimum Tm = 60°C (minimum of 58°C); optimum CG content = 50% (minimum of 45%). The primer pairs obtained were checked for non-specific hybridizations in other genome regions using the National Center for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (BLAST) at http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Subsequently, aiming to set up a robust multiplex assay, we refined the selection of indels considering the results from a hairpin and primer–dimer secondary structures check, using AutoDimer [19]. In order to achieve an even distribution throughout the genome as well as a sufficient distance to avoid linkage disequilibrium between markers in the same chromosome, we were able to select 38 bi-allelic indels spread across all human autosomes (detailed information in Table 1). All markers were then organized by expected amplicon length and assigned to four different dye-labelling fluorochromes in order to achieve an evenly balanced genotyping assay from a single PCR and electrophoretic separation. When necessary, tails of random nucleotides were added to the designed primers, to adjust the mobility of the final amplicon.

2.3 Indel amplification and detection

All indels were initially analyzed in singleplex, in order to evaluate primer performance and expected allele sizes. After optimization, the amplification of the 38 indels was performed in a single multiplex PCR using the Qiagen Multiplex PCR kit (Qiagen) at 1 × concentration, 0.1 μM of all primers except for rs2308137 (0.2 μM), and rs3047269 (0.3 μM) and 0.5 ng of genomic DNA in a 10 μL final reaction volume. Thermocycling conditions were initial incubation at 95°C for 15 min; 10 cycles at 94°C for 30 s, 60°C for 90 s and 72°C for 60 s; 20 cycles at 94°C for 30 s, 58°C for 90 s and 72°C for 60 s; with a final extension at 72°C for 60 min.

PCR products were subsequently prepared for CE by adding 1 μL of each amplified product to 14 μL of a 24:1 mixture of deionized Hi-Di formamide (Applied Biosystems) and GS-500 LIZ size standard (Applied Biosystems) respectively. Separation and detection were performed with an ABI PRISM 3130 Genetic Analyzer (Applied Biosystems) using filter set G5 and POP7 polymer (Applied Biosystems). Samples were genotyped with GeneMapper v4.0 software (Applied Biosystems).

9947A and 9948 human cell line DNA samples (Promega) were routinely used as amplification positive controls and to test the overall performance of the multiplex genotyping protocol.

Table 1. Indels selected for this study, including the location of each marker in the genome, reported alleles and amplicon expected size for short (S) and long (L) alleles

rs number	Chr.	Position (bp)	Contig and position (bp)	Alleles	Amplicon expected size (S–L)
rs3047269	1	161077452	NT_004487.18 pos. 13301183	-/CTGA	126–130
rs2307579	1	245878706	NT_004836.17 pos. 12569872	-/ATG	104–107
rs16624	2	234681130	NT_005120.15 pos. 949145	-/GT	65–67
rs2308242	3	8591709	NT_022517.17 pos. 8556709	-/CT	106–108
rs2308026	4	119404855	NT_016354.18 pos. 43733552	-/CA	83–85
rs2307526	5	5178112	NT_006576.15 pos. 5115112	-/ACAC	95–99
rs1160956	5	65414216	NT_006713.14 pos. 15972818	-/AGA	128–131
rs1610871	5	171020572	NT_023133.12 pos. 15897552	-/TAGG	61–65
rs2307710	6	47929222	NT_007592.14 pos. 38679494	-/AGGA	92–96
rs2307839	6	117200251	NT_025741.14 pos. 21262987	-/GA	152–154
rs2308137	6	149655891	NT_025741.14 pos. 53718627	-/GA	61–63
rs2307978	7	83121850	NT_007933.14 pos. 8518190	-/GA	156–158
rs35769550	8	76681235	NT_008183.18 pos. 28372034	-/TGAC	89–93
rs5895447	8	138489776	NT_008046.15 pos. 51638773	-/CA	128–130
rs16402	9	38396788	NT_008413.17 pos. 38396788	-/TTAT	150–154
rs2067294	9	70504241	NT_023935.17 pos. 478953	-/CTT	80–83
rs2307580	9	104626014	NT_008470.18 pos. 12907398	-/AATT	120–124
rs140809	10	6027167	NT_077569.2 pos. 350057	-/CAA	115–118
rs1160886	10	54112392	NT_008583.16 pos. 2993541	-/ACT	75–78
rs10688868	11	258180	NT_035113.6 pos. 208180	-/CT	81–83
rs34811743	11	30134266	NT_009237.17 pos. 28964931	-/TG	108–110
rs33972805	11	125794082	NT_033899.7 pos. 29851288	-/CT	135–137
rs1610919	12	14801263	NT_009714.16 pos. 7668970	-/AT	142–144
rs2067238	12	113772931	NT_009775.16 pos. 5858057	-/GCT	71–74
rs2308171	13	43778155	NT_024524.13 pos. 25860155	-/TCTG	135–139
rs2308189	14	28106508	NT_026437.11 pos. 10036508	-/AACTA	119–124
rs2308020	15	51268809	NT_010194.16 pos. 24272074	-/TT	127–129
rs2067208	16	83139788	NT_010498.15 pos. 38196486	-/GCCAG	93–98
rs3051300	17	10076666	NT_010718.15 pos. 9733290	-/GTAT	63–67
rs3080855	18	21507205	NT_010966.13 pos. 4742309	-/AATT	133–137
rs34511541	18	34677042	NT_010966.13 pos. 17912146	-/CTCTT	143–148
rs36040336	19	1353662	NT_011255.14 pos. 1342662	-/AT	65–67
rs2307689	19	48896180	NT_011109.15 pos. 16472558	-/TTC	74–77
rs33917182	20	11643625	NT_011387.8 pos. 11635625	-/CA	142–144
rs34541393	20	30165066	NT_028392.5 pos. 897497	-/AACT	57–61
rs35605984	21	14556736	NT_011512.10 pos. 1296736	-/TAAAG	151–156
rs10629077	21	30294208	NT_011512.10 pos. 17034208	-/AT	74–76
rs2307700	22	25120901	NT_011520.11 pos. 6181470	-/TCAC	101–105

Mapping data according to dbSNP build 129.

2.4 Population samples

To evaluate the genetic variation of the selected indels in three major population groups, we studied samples from Africans (54 Angolans and 50 Mozambicans), Europeans (100 Portuguese), and Asians (52 Macanese and 50 Taiwanese). All DNAs used were long-standing anonymized population samples, labeled by population of origin only.

2.5 Statistical analysis

Allele frequencies, expected heterozygosities, exact tests of Hardy–Weinberg equilibrium, F_{ST} genetic distances, Analysis of molecular variance (AMOVA) and exact tests of linkage disequilibrium were all assessed with Arlequin v3.0 software [20]. Statistical parameters to evaluate the forensic efficiency, such as discrimination power (DP) and random match probabilities (RMPs) for each locus and profile (*i.e.* accumulated values for the whole multiplex) were calculated using in-house-developed applications.

3 Results and discussion

In this study, we developed a simple and sensitive indel multiplex for human identification, which allows the genotyping of 38 bi-allelic markers presenting high heterozygosity in distinct populations groups.

3.1 Multiplex PCR design and optimization

The *in silico* assay design resulted in a good working base to set up the multiplex reaction. In initial multiplex tests some primer pairs revealed weaker performance than others, as would be expected since the optimum annealing temperature was not exactly the same for all primers. To minimize these effects and aiming for a more balanced output, we tested different primer mix concentrations and annealing temperatures. Accordingly, rs3047269 and rs2308137 primer concentrations were raised in the mix, and PCR thermocycling conditions changed to a touchdown strategy.

During the analyses of the observed genotype distributions, although not significant after correction for multiple analyses [21], a low Hardy–Weinberg equilibrium *p*-value was found for rs35605984, along with a heterozygote deficit in both European and East Asian samples, indicating that null alleles could be occurring. Therefore, a new search for SNPs at the flanking region of this marker was undertaken using a more recent dbSNP build (129). A new SNP (rs57670043) was found, located 62 bp upstream of indel rs35605984, at the antepenultimate base of the initial forward primer annealing sequence. Since this new SNP was not yet validated and no allele frequencies were available, we re-typed all homozygous samples and found a frequency of 5.00% in Africans, 7.38% in Europeans, and 10.00% in East Asians for the variant allele T.

To overcome this problem and avoid null alleles at rs35605984, a second forward primer for the alternative base was added to the initial primer mix.

After the multiplex development and optimization, 38 indels were successfully amplified in a single PCR reaction, following the final optimum conditions reported in Section 2 and as shown in Fig. 1.

3.2 Multiplex performance

In general, during the genotyping of population samples, the multiplex reaction revealed to be robust and able to successfully amplify all markers in samples extracted by different methods, containing variable DNA quantity and/or quality. Nevertheless, when using highly concentrated DNA samples ($\geq$10 ng), we often observed strong signals and over-scaled peaks; in these cases, the presence of pull-ups and abnormally shaped peaks due to ineffective fluorescence correction can make interpretation of the genotyping results difficult. Furthermore, the use of higher amounts of DNA in the reaction occasionally caused the inhibition of amplification of longer size amplicons. Several DNA concentrations ranging from 0.10 to 10 ng were tested with the optimized multiplex, using dilution series of reference samples 9947A and 9948 (four replicates). The best results, revealing an improved peak balance among markers, were observed

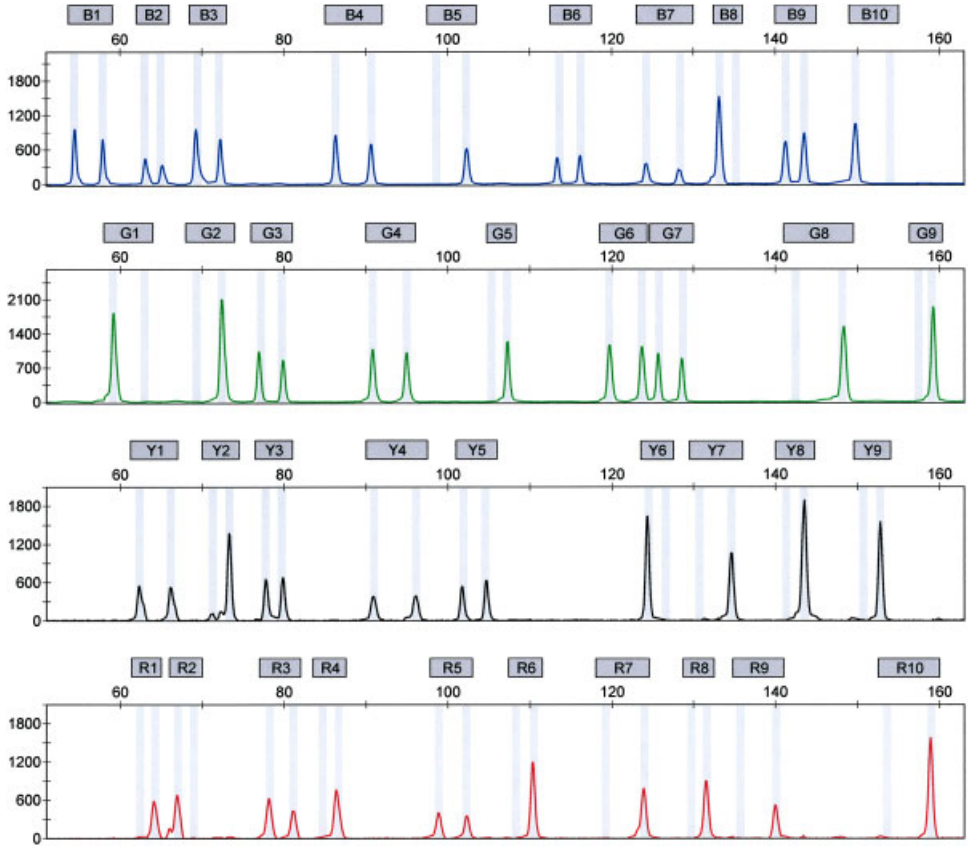

Figure 1. Electropherogram of the 38 indel-plex amplified using 1 ng DNA sample from an European individual.

when using 0.5 ng of DNA in a 10 μL PCR final volume, although good results for all markers were obtained for concentrations from 0.3 to 5 ng. In samples with DNA amounts as low as 0.2 or 0.1 ng, it was still possible to amplify all markers; nonetheless, the results were not always consistent, sometimes presenting unusual allele imbalances and/or allele dropouts observed from the known consensus genotypes of the control DNAs used. The marked heterozygote imbalance or allele dropouts with small quantities of target DNA have been previously reported and are a consequence of stochastic variation in PCR amplification when the number of template molecules is very low [22]. We highlight the fact that, in low copy number applications, results must be interpreted with caution, and duplicate analyses are required to minimize errors.

As a practical application, the assay was employed with a high percentage of success in biological material recovered from paraffin blocks of normal and tumor tissues, aiming to confirm the identity of the different samples [23]. In contrast, genotyping of the same samples using well-established commercial STR kits provided little information, as most STR markers failed to amplify. In the final stages of the indel-plex optimization, four forensic analyses were performed on highly degraded DNA extracted from skeletal material (three bones and one dental pulp extract). These were analyzed with STRs and the indel-plex in parallel and each case showed the same improvement in genotyping success with indel-plex described for the paraffin-embedded samples. A more thorough analysis of forensic and paternity applications of the indel-plex markers is now being conducted on the range of challenging material that can be expected in normal forensic casework. An example of profiles obtained from a DNA sample from a femur is shown in Fig. 2 for PowerPlex® 16 STR kit and indel-plex. The profiles shown highlight the difference in genotyping success between each marker set with indel-plex showing a complete profile, albeit with some imbalance, and STR profile comprising just 5 of 15 loci reliably genotyped, locus dropout in four of the longest fragment length systems and most likely allele dropout from a higher than expected range of homozygote peaks in half of the STRs. Such results from challenging DNA emphasize the importance of methodologies using reduced amplicon sizes to analyze severely degraded DNA, in line with success rates observed for other short-amplicon genotyping approaches based on SNPs or miniSTRs (*e.g.* [11, 13, 24, 25]).

3.3 Genetic variation in human populations

The genetic variation of the 38 indels was studied in three major population groups: Africans, Europeans, and East Asians. In Table 2 we present allele frequencies derived from these populations and the expected heterozygosity values for each marker. Data obtained from each pair of populations from Africa and from East Asia were pooled in the analysis after verifying an absence of significant differentiation for this set of markers between Angolans and Mozambicans (overall $F_{ST} = 0.00246$; $p = 0.15246$) or between Macanese and Taiwanese (overall $F_{ST} = 0.00277$; $p = 0.10959$).

All studied markers followed Hardy–Weinberg expectations.

The mean heterozygosity of the indel set was 0.45, and all markers showed heterozygosity values higher than 0.30. Our results confirm that the selected indels are highly polymorphic and very informative in the three major population groups.

Differences in allele frequencies were noticeable for some polymorphisms when compared with previously available data [1] (see also NCBI dbSNP at http://www.ncbi.nlm.nih.gov/SNP/). Our studies showed some loci with a minor allele frequency <0.25 in particular population groups, despite our selection criterion. These differences could be explained by differences in populations used in our studies and others to represent each population group, and variations related to sampling size effects could also have occurred. Further studies involving more populations are required to better understand some of the differences found and in order to establish appropriate databases, which are essential prior to application in forensic casework.

3.4 Population comparisons

Population comparisons performed by genetic distance analysis (see Table 3) revealed levels of genetic variation in the range of those reported by other studies involving intercontinental samples [*e.g.* [4, 26]]. Overall, AMOVA analysis of the 38 indel set indicates that differences between groups represent only 10.50% of the total genetic diversity, meaning that the individual component of genetic variation accounts for most of human genetic diversity for these markers, in line with all other studies to date.

Single locus analysis also showed significant differentiation between groups for the majority of indels. In order to reduce such differentiation shown by certain loci, we would need to be more restrictive in the selection criteria in order to achieve a panel for identification presenting lower F_{ST} values between all population groups [14]. However, it is important to stress that a panel fulfilling such desirable characteristics in terms of population variability usually implies the use of specific platforms allowing individual genotyping of polymorphisms without easily developing multiplexed assays of sufficient size. Such an approach requires high-quality/quantity DNA, which is frequently a limiting factor in forensic casework. Conversely, our aim was to develop a multiplex assay allowing the simultaneous genotyping of an informative set of bi-allelic markers in a single reaction so that individual differences in allele frequency distributions were compensated sufficiently by the marker set as a whole. Therefore, markers such as rs2308026 (heterozygosity in Africans 0.101, Europeans 0.471, and East Asians 0.387) represent the exception rather than the rule and do not markedly reduce the informativeness of the multiplex as a whole for any one population group.

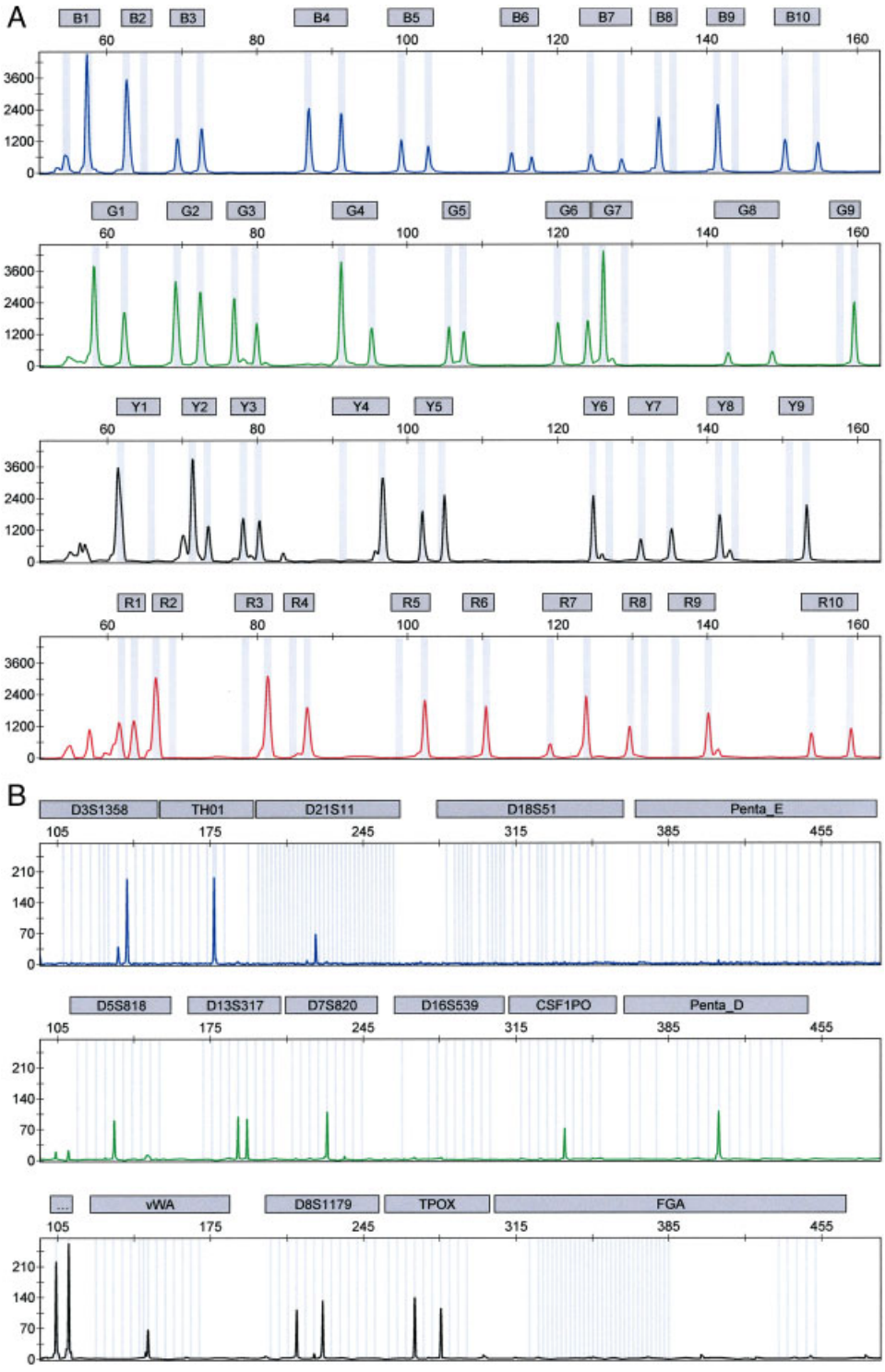

Figure 2. Electropherogram of the 38 indel-plex (A) and PowerPlex® 16 STRs kit (B) using a low quality DNA sample extracted from a femur.

3.5 Linkage disequilibrium analysis

Our assay design aimed to combine independent markers, in order to enable statistical assessment of parameters of forensic interest free from association between the markers incorporated. We selected indels in different chromosomes or with a minimum separation of 30 Mb for markers located in the same chromosome; only for three smaller chromosomes (18, 20, and 21) were markers admitted with less distant separation (~13.3–18.5 Mb), thus allowing the inclusion of three extra indels in the assay. However, these distances are several times larger than the reported maximum extent of linkage disequilibrium blocks [27]. In the end, we achieved the multiplexing of 38 indels

Table 2. Population data and statistical parameters of forensic efficiency for the 38 indel-plex in Africans, Europeans, and East Asians

Internal code	rs number	Allele frequencies (short allele)			Expected heretozygosity				Discrimination power		
		African	European	E. Asian	African	European	E. Asian	Total	African	European	E. Asian
B1	rs34541393	0.388	0.371	0.789	0.477	0.470	0.334	0.500	0.612	0.607	0.499
B2	rs16624	0.278	0.740	0.373	0.403	0.387	0.470	0.498	0.561	0.547	0.607
B3	rs2307689	0.466	0.310	0.162	0.500	0.430	0.273	0.431	0.624	0.581	0.432
B4	rs35769550	0.115	0.410	0.593	0.205	0.486	0.485	0.467	0.346	0.617	0.616
B5	rs2307700	0.209	0.500	0.275	0.332	0.503	0.400	0.440	0.497	0.625	0.559
B6	rs140809	0.525	0.402	0.299	0.503	0.485	0.422	0.483	0.624	0.615	0.575
B7	rs3047269	0.688	0.436	0.613	0.432	0.494	0.477	0.488	0.582	0.621	0.611
B8	rs33972805	0.337	0.465	0.098	0.449	0.500	0.178	0.420	0.594	0.624	0.307
B9	rs33917182	0.534	0.538	0.417	0.500	0.500	0.489	0.501	0.624	0.624	0.618
B10	rs16402	0.284	0.310	0.196	0.408	0.430	0.317	0.388	0.565	0.581	0.481
G1	rs1610871	0.505	0.570	0.353	0.502	0.493	0.459	0.500	0.625	0.620	0.601
G2	rs2067238	0.216	0.650	0.275	0.341	0.457	0.400	0.471	0.506	0.599	0.559
G3	rs2067294	0.106	0.365	0.240	0.190	0.466	0.367	0.360	0.325	0.605	0.530
G4	rs2307710	0.438	0.365	0.245	0.495	0.466	0.372	0.456	0.621	0.605	0.535
G5	rs2308242	0.367	0.185	0.252	0.467	0.303	0.379	0.394	0.605	0.467	0.541
G6	rs2307580	0.250	0.535	0.466	0.377	0.500	0.500	0.486	0.539	0.624	0.624
G7	rs1160956	0.471	0.835	0.529	0.501	0.277	0.501	0.477	0.624	0.437	0.624
G8	rs34511541	0.313	0.430	0.485	0.432	0.493	0.502	0.484	0.582	0.620	0.625
G9	rs2307978	0.413	0.200	0.373	0.487	0.322	0.470	0.443	0.617	0.486	0.607
Y1	rs3051300	0.223	0.415	0.304	0.349	0.488	0.425	0.431	0.513	0.617	0.578
Y2	rs10629077	0.249	0.215	0.260	0.376	0.339	0.387	0.367	0.538	0.504	0.547
Y3	rs10688868	0.188	0.246	0.534	0.306	0.373	0.500	0.438	0.470	0.536	0.624
Y4	rs2067208	0.116	0.266	0.186	0.206	0.393	0.305	0.306	0.347	0.553	0.468
Y5	rs2307579	0.635	0.525	0.167	0.466	0.501	0.279	0.494	0.605	0.624	0.440
Y6	rs2308020	0.726	0.663	0.407	0.400	0.449	0.485	0.481	0.558	0.594	0.616
Y7	rs3080855	0.332	0.260	0.451	0.446	0.387	0.498	0.455	0.592	0.547	0.623
Y8	rs1610919	0.534	0.625	0.299	0.500	0.471	0.421	0.500	0.624	0.608	0.575
Y9	rs2307839	0.182	0.307	0.461	0.300	0.427	0.499	0.434	0.463	0.579	0.623
R1	rs2308137	0.710	0.348	0.441	0.414	0.456	0.495	0.501	0.569	0.599	0.621
R2	rs36040336	0.296	0.800	0.691	0.419	0.322	0.429	0.483	0.573	0.486	0.580
R3	rs1160886	0.258	0.377	0.552	0.386	0.474	0.498	0.482	0.546	0.609	0.622
R4	rs2308026	0.053	0.375	0.260	0.101	0.471	0.387	0.352	0.185	0.608	0.547
R5	rs2307526	0.309	0.325	0.598	0.429	0.441	0.483	0.485	0.581	0.589	0.615
R6	rs34811743	0.577	0.670	0.696	0.491	0.444	0.425	0.457	0.619	0.591	0.578
R7	rs2308189	0.567	0.352	0.485	0.495	0.460	0.503	0.499	0.620	0.600	0.625
R8	rs5895447	0.284	0.325	0.284	0.408	0.441	0.409	0.419	0.565	0.589	0.565
R9	rs2308171	0.577	0.215	0.078	0.491	0.339	0.145	0.415	0.619	0.504	0.258
R10	rs35605984	0.635	0.450	0.368	0.466	0.497	0.467	0.500	0.605	0.622	0.606
				mean:	0.407	0.438	0.417	0.452			
								ac. PD	99.999999999994	99.9999999999995	99.999999999997
								ac. RMP	3.4×10^{-14}	3.6×10^{-15}	1.7×10^{-14}
									1 in $2.9 \times 10^{+13}$	1 in $2.8 \times 10^{+14}$	1 in $6.0 \times 10^{+13}$

E. Asian – East Asian; ac. DP – accumulated discrimination power; ac. RMP – accumulated random match probability.

distributed across all human autosomes, with 12 chromosomes having more than one marker.

There are no plausible reasons to expect strong associations between loci in different chromosomes and in addition, the exact test of linkage disequilibrium between pairs of indels on the same chromosome did not reveal any significant *p*-value after correction for multiple analyses [21]. For the reasons outlined above, it is possible to conclude that indels included in the assay do not present signs of association and therefore can be treated as independent markers in forensic statistical analysis.

Table 3. Pairwise population F_{ST} estimates between African, European, and East Asian from the 38 indels studied (F_{ST} values below diagonal) and corresponding *p*-value (for 10 100 permutations) (above diagonal)

	African	European	E. Asian
African	–	$\leq 10^{-5}$	$\leq 10^{-5}$
European	0.10758	–	$\leq 10^{-5}$
E. Asian	0.12176	0.09066	–

to combine in the same assay several desirable characteristics in forensic analysis: (i) adequate informativeness level for human identification studies, by multiplexing 38 highly polymorphic indels that when combined provide forensic efficiency levels suitable for standard casework applications; (ii) use of reduced amplicon sizes comparable to those of forensic SNP assays that improves the PCR success when amplifying degraded samples; and (iii) simplicity of analysis through PCR and CE, in the same way as STRs. Being a true single tube reaction for the whole procedure, it minimizes laboratory procedures and sample manipulation, and consequently reduces the risk of contaminations, pipetting errors or other possible causes of failure, when compared with other bi-allelic marker typing approaches.

Table 4. Accumulated random match probabilities for the 38 indel-plex assay and others commonly used in human identification

	African	European	E. Asian
AmpFlSTR® Identifiler®	1 in $5.5 \times 10^{+16}$	1 in $1.9 \times 10^{+16}$	1 in $6.1 \times 10^{+15}$
PowerPlex® 16	1 in $1.4 \times 10^{+18}$	1 in $1.8 \times 10^{+17}$	1 in $3.7 \times 10^{+17}$
SNPforID 52 plex [13]	1 in $9.7 \times 10^{+17}$	1 in $2.7 \times 10^{+20}$	1 in $4.5 \times 10^{+17}$
AmpFlSTR® MiniFiler™	1 in $1.4 \times 10^{+09}$	1 in $6.2 \times 10^{+08}$	1 in $4.8 \times 10^{+08}$
38 indel-plex (this study)	1 in $2.9 \times 10^{+13}$	1 in $2.8 \times 10^{+14}$	1 in $6.0 \times 10^{+13}$

Values obtained from CEPH genotype data (CP unpublished results) except PowerPlex® 16 (from company product notes).

3.6 Forensic efficiency

The forensic efficiency of the 38-plex was evaluated by calculating the DP and RMP in Europeans, Africans, and East Asians, and values obtained from the allele frequencies in each population group are summarized in Table 2. In Table 4 we present comparative values of accumulated RMPs for the indel-plex assay and others commonly used in human identification.

As shown, the assay is highly efficient in all three population groups studied. The cumulative RMP ranges in orders of magnitude from 10^{-14} to 10^{-15}, whereas DP reaches 99.9999999999995%, thus providing satisfactory levels of informativeness for forensic demands. Moreover, selecting a subset of the 25 most informative indels in each population group would be sufficient to obtain globally unique genetic profiles (considering a world population of ~6.6 billion), achieving RMP of 1 in 10.0 billion in Africans, 1 in 12.3 billion in Asians, and 1 in 17.1 billion in Europeans. These values highlight the potential of this set of markers in identification studies, even when incomplete profiles are obtained.

Moreover, in paternity investigations, the occurrence of Mendelian incompatibilities between alleged father and child often becomes a problem. Given the high STR mutation rates, this is not an uncommon situation that can be largely overcome by the use of a high number of markers with low mutation rates, which is the case of both SNP and indel multiplex strategies.

4 Concluding remarks

In this study, we report a new strategy for human identification using indels. With our approach we were able

We believe that this assay can be easily implemented in current forensic laboratories without any extra requirements, thus taking advantage of the already well-established methodologies and technologies in place in all facilities. The efficiency and simplicity of the indel-plex make the assay a valuable routine tool in identification studies as well as a robust complementary tool to standard STR typing, especially in cases involving highly degraded DNA. Furthermore, the ease of analysis and time–cost effectiveness in relation to current SNP identification procedures are further arguments in favor of the indel-plex approach we have detailed.

R.P. has a Ph.D. grant (SFRH/BD/30039/2006) from Fundação para a Ciência e a Tecnologia. This work was partially supported by "Programa Operacional Ciência e Inovação 2010" (POCI 2010).

The authors have declared no conflict of interest.

5 References

[1] Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., Marth, G., *Am. J. Hum. Genet.* 2002, *71*, 854–862.

[2] Yang, N., Li, H., Criswell, L. A., Gregersen, P. K., Alarcon-Riquelme, M. E., Kittles, R., Shigeta, R. *et al.*, *Hum. Genet.* 2005, *118*, 382–392.

[3] Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., Feldman, M. W., *PLoS Genet.* 2005, *1*, 660–671.

[4] Bastos-Rodrigues, L., Pimenta, J. R., Pena, S. D., *Ann. Hum. Genet.* 2006, *70*, 658–665.

[5] Väli, U., Brandstrom, M., Johansson, M., Ellegren, H., *Biomed. Chromatogr. Genet.* 2008, *9*, 8.

[6] Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., Devine, S. E., *Genome Res.* 2006, *16*, 1182–1190.

[7] Chakraborty, R., Stivers, D. N., Su, B., Zhong, Y., Budowle, B., *Electrophoresis* 1999, *20*, 1682–1696.

[8] Butler, J. M., *Forensic DNA Typing*, Elsevier Academic Press, Burlington, London 2005.

[9] Petkovski, E., Keyser-Tracqui, C., Niemeyer, D., Hienne, R., Ludes, B., *Int. Congr. Ser.* 2004, *1261*, 21–23.

[10] Inagaki, S., Yamamoto, Y., Doi, Y., Takata, T., Ishikawa, T., Imabayashi, K., Yoshitome, K. *et al.*, *Forensic Sci. Int.* 2004, *144*, 45–57.

[11] Dixon, L. A., Murray, C. M., Archer, E. J., Dobbins, A. E., Koumi, P., Gill, P., *Forensic Sci. Int.* 2005, *154*, 62–77.

[12] Kidd, K. K., Pakstis, A. J., Speed, W. C., Grigorenko, E. L., Kajuna, S. L., Karoma, N. J., Kungulilo, S. *et al.*, *Forensic Sci. Int.* 2006, *164*, 20–32.

[13] Sanchez, J. J., Phillips, C., Børsting, C., Balogh, K., Bogus, M., Fondevila, M., Harrison, C. D. *et al.*, *Electrophoresis* 2006, *27*, 1713–1724.

[14] Pakstis, A. J., Speed, W. C., Kidd, J. R., Kidd, K. K., *Hum. Genet.* 2007, *121*, 305–317.

[15] Phillips, C., Fang, R., Ballard, D., Fondevila, M., Harrison, C., Hyland, F., Musgrave-Brown, E. *et al.*, *Forensic Sci. Int. Genet.* 2007, *1*, 180–185.

[16] Phillips, C., Salas, A., Sánchez, J. J., Fondevila, M., Gómez-Tato, A., Álvarez-Dios, J., Calaza, M. *et al.*, *Forensic Sci. Int. Genet.* 2007, *1*, 273–280.

[17] Amorim, A., Pereira, L., *Forensic Sci. Int.* 2005, *150*, 17–21.

[18] Nachman, M. W., Crowell, S. L., *Genetics* 2000, *156*, 297–304.

[19] Vallone, P. M., Butler, J., *Biotechniques* 2004, *37*, 226–231.

[20] Excoffier, L., Laval, G., Schneider, S., *Evol. Bioinform. Online* 2005, *1*, 47–50.

[21] Bonferroni, C. E., *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 1936, *8*, 3–62.

[22] Jobling, M. A., Gill, P., *Nat. Rev. Genet.* 2004, *5*, 739–751.

[23] Oliveira, C., Sousa, S., Pinheiro, H., Karam, R., Bordeira-Carriço, R., Senz, J., Kaurah, P. *et al.*, *Gastroenterology* 2009, *136*, 2137–2148.

[24] Coble, M. D., Butler, J. M., *J. Forensic Sci.* 2005, *50*, 43–53.

[25] Dixon, L. A., Dobbins, A. E., Pulker, H. K., Butler, J. M., Vallone, P. M., Coble, M. D., Parson, W. *et al.*, *Forensic Sci. Int.* 2006, *164*, 33–44.

[26] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., Feldman, M. W., *Science* 2002, *298*, 2381–2385.

[27] Wall, J. D., Pritchard, J. K., *Nat. Rev. Genet.* 2003, *4*, 587–597.

