Juan J. Sanchez[1]
Chris Phillips[2]
Claus Børsting[1]
Kinga Balogh[3]
Magdalena Bogus[3]
Manuel Fondevila[2]
Cheryl D. Harrison[4]
Esther Musgrave-Brown[4]
Antonio Salas[2]
Denise Syndercombe-Court[4]
Peter M. Schneider[3]
Angel Carracedo[2]
Niels Morling[1]

[1]Department of Forensic Genetics, Institute of Forensic Medicine, University of Copenhagen, Copenhagen, Denmark
[2]Institute of Legal Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain
[3]Institute of Legal Medicine, Johannes Gutenberg University, Mainz, Germany
[4]Centre for Haematology, ICMS, Barts and The London, Queen Mary's School of Medicine and Dentistry, London, UK

Received September 10, 2005
Revised October 15, 2005
Accepted October 16, 2005

Research Article

A multiplex assay with 52 single nucleotide polymorphisms for human identification

A total of 52 SNPs reported to be polymorphic in European, Asian and African populations were selected. Of these, 42 were from the distal regions of each autosome (except chromosome 19). Nearly all selected SNPs were located at least 100 kb distant from known genes and commonly used STRs. We established a highly sensitive and reproducible SNP-typing method with amplification of all 52 DNA fragments in one PCR reaction followed by detection of the SNPs with two single base extension reactions analysed using CE. The amplicons ranged from 59 to 115 bp in length. Complete SNP profiles were obtained from 500 pg DNA. The 52 loci were efficiently amplified from degraded samples where previously only partial STR profiles had been obtained. A total of 700 individuals from Denmark, Greenland, Somalia, Turkey, China, Germany, Taiwan, Thailand and Japan were typed, and the allele frequencies estimated. All 52 SNPs were polymorphic in the three major population groups. The mean match probability was at least 5.0×10^{-19} in the populations studied. Typical paternity indices ranged from 336 000 in Asians to 549 000 in Europeans. Details of the 52 SNP loci and population data generated in this work are freely available at http://www.snpforid.org.

Keywords: Autosomes / Human identification / Multiplex PCR / Single base extension / Single nucleotide polymorphism DOI 10.1002/elps.200500671

1 Introduction

SNPs have a number of characteristics that make them ideal markers for human identification. First, they have lower mutation rates than the STR and VNTR (variable number tandem repeat) loci typically used for relationship analysis in paternity and immigration testing. Second, SNPs can be analysed after PCR amplification of very short DNA-regions surrounding the substitution site, making SNPs preferable for anthropological and crime case investigations where the DNA is often degraded. Third, SNPs can be genotyped with a growing range of high-throughput technologies; an important factor in the implementation of large criminal DNA databases [1, 2]. Finally, SNPs, as binary polymorphisms, are comparatively easy to validate, because precise allele frequency estimates, required for the accurate interpretation of forensic genotyping data, can be obtained by analysing fewer samples compared to those needed for allele frequencies estimates of STRs and VNTRs. Seeking to match the discriminatory power of the 10–15 multiple allele STRs routinely used in forensic investigations, a set of about 50 polymorphic SNP markers are predicted to be required [3, 4]. Furthermore, it has been suggested that 50 unlinked SNP loci with high overall heterozygosity should be sufficient to adjust for population stratification in population-based associations studies [5]. SNPs that are polymorphic in one population may be almost or completely monomorphic in another population [6, 7], while others are known to be polymorphic in all major population groups. Thus, it should be possible to select SNPs that are useful for human identification purposes in the majority of populations, and to supplement these with SNPs

Correspondence: Dr. Juan J. Sanchez, Department of Forensic Genetics, Institute of Forensic Medicine, University of Copenhagen, 11 Frederik V's Vej, DK-2100 Copenhagen, Denmark
E-mail: juan.sanchez@forensic.ku.dk
Fax: +45-35-32-61-20

Abbreviations: RFU, relative fluorescence unit; **SBE**, single base extension

showing highly contrasting allele frequency distributions in particular populations. These latter SNPs can provide valuable information for population admixture detection, in addition to the estimation of biogeographical ancestry.

The SNPforID group (http://www.snpforid.org) is a consortium supported by the EU GROWTH programme with the following objectives: (i) selection of at least 50 autosomal SNPs suitable for the identification of persons of unknown population origin and determination of allele frequencies in the major population groups; (ii) development of a highly efficient DNA amplification strategy for the simultaneous analysis of up to 50 independent SNPs in a single assay; (iii) assessment of automated, high-throughput DNA-typing platforms for reliable and accurate multiplex SNP typing; (iv) assessment of the forensic application of the high-throughput SNP-typing methods developed.

If SNP typing is to be a realistic alternative to STR typing in forensic analyses and other fields of investigation where the target material is scarce and often of poor quality, the relevant SNP loci must be amplified efficiently in a single multiplexed PCR reaction and preferably analysed with a method that is well established and robust in routine use.

In this work we present five stages in the development of an SNP-based human identification genotyping assay: (i) a set of 52 unlinked autosomal SNPs that are highly polymorphic in European, Asian and African populations; (ii) a multiplex PCR amplification strategy that allows the simultaneous amplification of 52 fragments in a single reaction from as little as 500 pg DNA; (iii) an SNP detection system based on two sets of multiplexed single base extension (SBE) reactions with 23 and 29 SBE primers, respectively, that can be analysed in one CE run and automated computer based allele detection; (iv) validation of assay reproducibility, sensitivity and robustness; (v) determination of allele frequency distributions of each SNP in 700 individuals from 9 European, Asian and African populations as well as 9 animal species, including 46 samples from 6 primate species.

2 Materials and methods

2.1 Samples and DNA purification

A total of 700 samples (numbers in parentheses) from Denmark (156), Greenland (149), Somalia (104), Turkey (96), China (63), Germany (49), Taiwan (43), Thailand (33) and Japan (7) were typed in duplicate. In addition, samples from unrelated chimpanzees (29), gorillas (3), orangutan (1), baboons (2), rhesus macaques (5) and Cynomolgous monkeys (6) were tested together with dog (1), cat (1) and horses (2).

We used blood on FTA® cards (Whatman) and DNA purified by phenol/chloroform extraction or the QIAamp DNA blood mini kit (Qiagen). DNA concentrations were determined by real-time PCR using the Quantifiler™ Human DNA Quantification Kit (Applied Biosystems) with the ABI 7300 real-time PCR system (Applied Biosystems) or by using SYBR Green I (Roche) with the LightCycler system (Roche). The performance of the 52 SNP-plex assay was also tested on DNA purified from paraffin-embedded tissues and DNA from bones and muscle tissue samples taken from seven human cadavers found under various environmental conditions in crime cases.

All protocols were approved by the Danish ethical committee (KF-01-037/03).

2.2 Criteria for selection of SNP loci

The following SNP selection criteria were used when choosing suitable candidate loci: (i) the size of the amplicon generated from optimum primer designs less than 120 bp; (ii) reported minimum 30% heterozygosity (0.28 minor allele frequency) in at least one population, and minimum 20% heterozygosity (0.17 minor allele frequency) in all three populations; (iii) a freely assorting marker set using SNPs from the distal parts of the *p* and *q* arms of each autosome; (iv) a minimum distance of 100 kb between candidate SNPs and neighbouring genes; (v) no likely association with the STR loci most commonly used in forensic analysis; and (vi) flanking DNA sequence reliably reported and free from interfering polymorphisms, such as nucleotide substitutions in potential primer binding sites.

2.3 Marker selection *in silico*

Suitable regions were chosen for scrutiny using NCBI Map Viewer (http://www.ncbi.nlm.nih.gov/mapview/maps.cgi/) with alignments of the gene and variation maps following the general guidelines described by Phillips [8] (2004). Only loci marked on the gene map as 'confirmed gene models based on mRNA alignments' were used to define gene-free regions. In addition, SNP Browser™ (http://www.allsnps.com/snpbrowser) and HapMap (http://www.hapmap.org/) genome browsers were used as these became publicly available during the search phase. Three searches were performed using dbSNP builds 112 (*p*-arm loci), 115 (*q*-arm loci) and 118 (supplementary loci) corresponding to genome builds 28, 33 and 34, respectively. From these searches, sets of 46, 67 and 25 SNPs were selected giving a median two loci *per* p-arm and four loci *per* q-arm from each autosome. We primarily selected SNPs genotyped by The SNP Consortium, although these

were supplemented by a small number of SNPs validated by the Perlegen and HapMap genotyping initiatives. SNPs from The SNP Consortium had been validated in most cases using genotyped individuals from the three major populations with an average sample size of 38 from each group. Approximately 60% of SNPs scrutinized during the selection process had insufficient variability in one or more population groups using the comparative allele frequency criterion previously described [8] (see also Section 2.2). In the case of the third SNP selection of 25 loci, several markers with limited variability in one of the three major population groups were included to add predictive power for population of origin. The final screening of SNPs before assimilation into candidate pools involved examination of the flanking sequence to ensure that the region available for primer design (approximately 100 bp on each side of the SNP) was free from clustering SNPs and low-complexity sequence. In several cases, a single clustering SNP was permitted and subsequently circumvented during primer design. Simple sequence quality checks were performed using the Fasta sequence report in dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). Candidate SNPs with a proportion of low-complexity flanking sequence greater than 20% were rejected to ensure primer specificity before starting the multiplex design process. Table S1 lists the selected SNPs and the flanking sequences.

2.4 Selection of PCR amplification primers

PCR primers were designed to give amplicon lengths in the range from 59 to 115 bps (Table S1). The aim was to obtain a theoretical melting temperature of 60 $\pm$ 2°C at a salt concentration of 180 mM and a purine:pyrimidine content close to 1:1. All primer candidates were analysed for primer-dimer formation, hairpin structures, homology and complementarity to other primers in the multiplex using Primer 3.2 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). Primer characteristics were chosen to ensure equal PCR amplification efficiency for all DNA fragments as previously described [9]. The primers were HPLC-purified and checked for homogeneity by MALDI-TOF MS (DNA Technology A/S, Denmark). Table S1 shows the sequences and the concentrations of the amplification primers in the final multiplex PCR.

2.5 PCR conditions and purification of PCR products

PCR amplification conditions were: 1–10 ng DNA in a 25 μL reaction volume containing 1 × PCR buffer, 8 mM $MgCl_2$, 700 μM of each dNTP, 0.01–0.17 μM of each primer and 2 U AmpliTaq Gold DNA polymerase (Applied Biosystems). For the PCR inhibition study, haematin (Sigma-Aldrich) was diluted to 15 mM in 0.1 N NaOH and added to the mixture. Cycling was performed in a Gene Amp 9600 (Perkin Elmer) or Eppendorf Mastercycler gradient (Eppendorf) thermal cycler with the following cycle programme: denaturation at 94°C for 5 min followed by 35 cycles of 95°C for 30 s, 60°C for 30 s and 65°C for 30 s, followed by 7 min at 65°C.

Excess primers and dNTPs were removed by using Min Elute PCR purification spin columns (Qiagen). The PCR products were eluted in 20 μL of Milli-Q water. Alternatively, 1 μL ExoSAP-IT kit (Amersham Pharmacia Biotech) or 0.75 μL (1 U/μL) shrimp alkaline phosphatase (Amersham Pharmacia Biotech) and 0.023 μL (10 U/μL) Exonuclease I (Amersham Pharmacia Biotech) were added to 2.5 μL PCR product and incubated at 37°C for 15 min, 80°Cfor 15 min or 37°C for 1 h, and 75°C for 15 min, respectively.

2.6 Design of SBE primers

Two SBE multiplexes with 23 and 29 SBE primers, respectively, were developed according to the guidelines described by Sanchez *et al.* [9] (2004). The lengths of the SBE primers were between 16 and 92 nucleotides. Size intervals of five to six nucleotides were used for primers shorter than 35 nucleotides and size intervals of four nucleotides were used for primers longer than 35 nucleotides (Table S2). The primers were checked for primer-dimer formation using the AutoDimer program (http://www.cstl.nist.gov/biotech/strbase/AutoDimerHomepage/AutoDimerProgramHomepage.htm). The lengths of the SBE primers were increased with tails of nonhuman sequence and/or poly-dNTP stretches (see Table S2). Opposite allele combinations (*e.g.* A/G with C/T) were analysed in the same size interval whenever possible.

2.7 SBE reaction and purification of the SBE products

SBE reactions were performed in 8 μL with 1 μL of purified PCR product, 4 μL of SNaPshot reaction mix (Applied Biosystems), 1 μL of SBE primer mix (0.01–0.27 μM, Table S2) and 2 μL of Milli-Q water. The SBE primer mix was diluted in 160 mM ammonium sulfate (Sigma-Aldrich) to minimize primer-dimer artefacts. The SBE reaction was performed in a GeneAmp 9600 or Eppendorf Mastercycler gradient instruments with 30 cycles of 96°C for 10 s, 50°C for 5 s and 60°C for 30 s. Excess nucleotides were removed by addition of 1 μL (1 U/μL) shrimp alkaline phosphatase to the SBE mix and incubation at 37°C for 45 min followed by 75°C for 15 min.

2.8 Detection and analysis of the SBE products

Two microlitres of SBE product was mixed with 20 μL of Hi-Di formamide (Applied Biosystems) and analyzed by CE using ABI Prism 310, ABI Prism 3100 or ABI Prism 3100-Avant Genetic Analyzers (Applied Biosystems) with 36 cm capillary arrays and POP-4 polymer (Applied Biosystems). The following modifications were made to the SNP36-POP-4 default module of the GeneScan analysis 3.7 software: Three extra washing steps were added (i) after injection of the polymer, (ii) after the prerun and (iii) after the injection of the samples using the spare water reservoir, SIT Water2. The time that the capillary tips were immersed in the wash buffers was increased to 60 s to minimize carry-over. The run time was decreased to 300 s to allow sufficient spacing between the two injections. Another GeneScan method file was created to allow injection of the second SBE multiplex without a capillary polymer fill stage by removing the steps that fill the syringe and the capillary array along with the prerun. The run time of the second injection was 1000 s. The two GeneScan methods were run sequentially by saving the methods with number sequence default module names in the ABI 3100 Data Collection 1.1 software using the Module Editor. We used 22 s injections at 3000 and 2000 V in the first and second injections, respectively. Analysis was made using GeneScan Analysis 3.7 with peak thresholds set to a minimum of 120 relative fluorescence units (RFUs) (blue color), 60 RFUs (green color) and 30 RFUs (yellow, red and orange color). GeneScan-120 Liz internal size standard (Applied Biosystems) was used in both injections but the reference sizes were modified in the second injection to the standard size plus 200 (Fig. 1).

Automated allele calls were made using Genotyper 3.7 (Applied Biosystems) macros. The macros used are available on the SNPforID website (www.snpforid.org). All peaks in the size standard had to be detected and the peak height of the largest peak in the electropherogram had to be a minimum of 1000 RFUs before analysis could proceed. Peaks detected in predefined allele windows with peaks heights larger than 10% (blue color), 7% (green color) or 5% (yellow and red colour) of the maximum peak height in the respective color were labelled with allele names, peak heights and sizes. The predefined windows were determined from prior analyses of 96 samples. The widths of the predefined windows varied from 1 to 2 bp depending on the lengths of the extended SBE primers. The ratios of the fluorophore emissions of the dR110 (blue), dR6G (green), TAMRA™ (yellow) and dROX™ (red) dyes were approximately 4:2:1:1. A few windows had slightly different ratios due to locus-specific variations and the peak heights were normalized accordingly. A maximum peak height ratio of 3:1 after normalization was accepted as a heterozygote. A minimum peak height of 400 RFUs (blue), 200 RFUs (green) or 100 RFUs (yellow and red) and a peak height ratio of minimum 5:1 after normalization were accepted as a homozygote.

2.9 Reproducibility and sensitivity

Validation of the 52 SNP-plex assay was, in principle, conducted according to the revised guidelines of DNA analysis of the 'Scientific Working Group on DNA Analysis Methods' (SWGDAM, http://www.fbi.gov/hq/lab/fsc/backissu/july2004/index.htm). Multiplex PCR performance was assessed by analyses of dilution series of genomic DNA (0.07, 0.14, 0.27, 0.55, 1.09, 2.19, 4.37, 8.75, 17.50, 35 and 70 ng) from two individuals. The RFUs of each of the four dyes were collated and normalised by dividing homozygote allele values by two.

2.10 Population studies

SNP characterization including allele frequencies, Hardy–Weinberg equilibrium and linkage disequilibrium tests was carried out using the SNP Assistant Program v. 1.0.9.

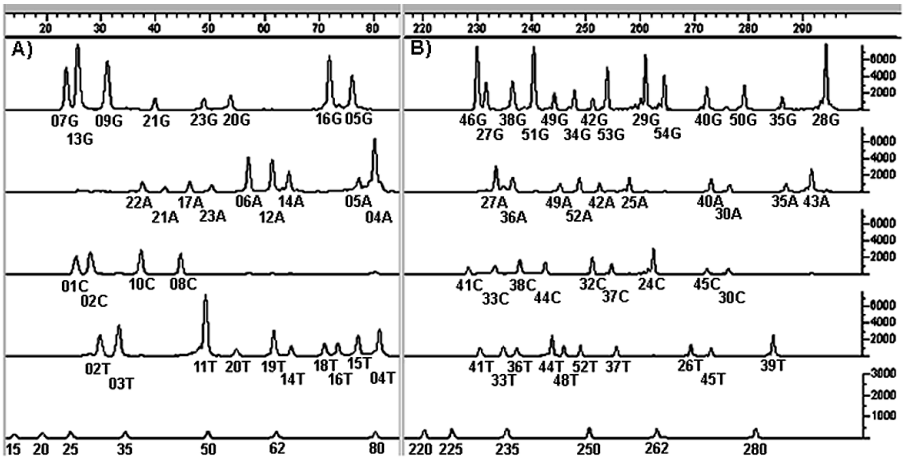

Figure 1. Analysis of the 52 SNP-plex assay. (A) Results of the first injection (23 SNPs). (B) Results of the second injection (29 SNPs).

The Genetic Data Analysis software (http://hybrodictyon.eeb.uconn.edu/people/plewis/software.php) was used to estimate F_{ST} values [10] and exact tests for associations between alleles at different loci. The software package Arlequin v. 2 000 (http://anthro.unige.ch/arlequin) was used to determine the molecular variance. Pearson coefficients (*r*) were calculated using Excel spreadsheets. Cluster analysis was performed by assigning individuals to a predefined number (*K*) of 'inferred populations' according to their genotypes using the Bayesian approach implemented in Structure v. 2.0 [11]. A model of admixture was used that assumes individuals may have mixed ancestry. All Structure v. 2.0 runs used 200 000 Markov Chain Monte Carlo steps after a burn-in of length 200 000. Five independent replicates were performed for each value of *K*, and these gave consistent results. Posterior probabilities of *K* were calculated using the values of ln $P(X|>K)$ where *X* denotes the genotypes of the sampled individuals.

2.11 Autosomal STR typing and sequencing

STR typing was performed using the AmpF/STR Identifiler or the AmpF/STR SGM Plus amplification kit (Applied Biosystems). The PCR was performed as recommended by the manufacturer.

Fourteen PCR amplicons carrying the SNPs rs1028528, rs1029047, rs1979255, rs2016276, rs2046361, rs2056277, rs2107612, rs2831700, rs354439, rs733164, rs735155, rs901398, rs907100 and rs938283 were sequenced in at least three individuals using the BigDye Terminator Kit (Applied Biosystems) as recommended by the manufacturer and analysed on an ABI Prism 377 Genetic Analyzer (Applied Biosystems).

2.12 Forensic statistical analysis

The power of discrimination was calculated as described by Jones [12] (1972) using PowerStats (Promega). The matching probability, mean exclusion probability and the typical paternity index were calculated as described by Brenner and Morris [13] (1989) using DNAVIEW™ 27.19.

3 Results

3.1 Selection of SNP loci and primers for PCR and SBE

Three groups of candidate SNP loci fulfilling the criteria described in Section 2 were collected from the NCBI SNP database, AB SNP Browser and the HapMap genome browser. The first group of 46 SNPs comprised 2–3 SNPs from the distal region of the *p*-arm of each autosome (with the exception of the gene dense chromosome 19), the second group of 67 SNPs comprised 3–4 SNPs from the distal region of the equivalent *q*-arm and the third group contained 25 SNPs selected mostly from the large autosomes. The final selection of SNP loci for the multiplex was determined during the phase of PCR primer design. The SNP from each chromosome end that gave the best chance to be amplified as a short fragment in a large multiplex under the conditions set for primer design was selected from the first two candidate groups. No suitable loci on the distal regions of chromosome 19 were found in this selection, a finding not completely unexpected, since chromosome 19 has a very high gene density [14]. Another 12 loci, including one SNP from chromosome 19, were selected from the third group of loci bringing the total number of loci to 54. All loci were successfully amplified in singleplex PCR and then divided into two large multiplexes with 23 and 31 fragments, respectively. The two PCR multiplexes were optimized and finally combined into a single multiplex with 54 loci. At no point during this process was it necessary to replace any of the originally designed PCR primers, emphasizing the importance of strict adherence to clearly defined guidelines for primer design when developing large multiplexes. Similarly, only 5 of the 54 SBE primer designs were altered during the development of the SBE multiplexes. Alternative primers on the opposite DNA strand were needed in three instances because the SBE reactions for these primers were so efficient that it had a deleterious effect on the balance and stability of the multiplex extension reaction. In two cases, the lengths of the SBE primers were altered to allow more efficient analysis of the electropherograms.

During the multiplex development phase, the genomic positions or the sequences surrounding the selected SNPs underwent revision in no less than 35 of the 54 loci. This arose from one major and several minor updates of the reference sequence builds on which the dbSNP database is founded, principally correcting numerous segments of inverted sequences. In most cases, this had no impact on the PCR or SBE primer designs. However, the new information reported in dbSNP disqualified 2 of the 54 SNP loci selected for the multiplex. In one case, SNP rs2145294 was reported as duplicated, and in the other case, a revised position was given for SNP rs1360674 only 4.24 kb from another selected locus on chromosome 1. This reduced the final combination to 52 SNPs.

3.2 Analysis of the SBE reaction

We chose to use the ABI Prism SNaPshot™ Multiplex System for generation of the SBE products because SBE primers extended with a fluorescently labelled ddNTP can

be detected with the CE equipment available in most forensic genetic laboratories, including each of the consortium laboratories. However, it was immediately clear that all 52 SNPs could not be detected in one SBE reaction in a five-colour electropherogram as it was impossible to construct suitably spaced size windows for 104 possible alleles when four to six nucleotide spacing between size windows in the same colour was needed and the maximum size of extended SBE primers could not realistically exceed 90 nucleotides [15, 9]. Therefore, we decided to develop two SBE multiplexes with 23 and 29 SBE primers, respectively, and analyse the 52 SNPs in two different electrophoretic runs.

On the ABI Prism 310 Genetic Analyser, it was possible to inject the two SBE multiplexes shortly after each other in the same capillary and thus analyse all 52 SNPs in one data file. This was not possible with the ABI Prism 3100 Genetic Analysers because there is a default capillary polymer fill stage between each injection. To address this problem, we created a new GeneScan method file for the ABI 3100 that allowed sequential injection into the same capillary (see Section 2). During this work it also became clear that the same SBE product analysed on an ABI 310 and on an ABI 3100 gave very different signal strengths. On average, the sensitivity of the ABI 310 was only 20% of the sensitivity of the ABI 3100, and most of the work presented here (see below) was analysed using ABI 3100 or ABI 3100-Avant Genetic Analysers.

In the early phase of the optimisation of the SBE multiplexes, we observed some small, distinct peaks that did not have the appearance of background noise. Addition of ammonium sulfate to the SBE primer mix significantly reduced the number of these nonspecific peaks, indicating that several of the peaks originated from extension of primer-dimers [16]. However, even after extensive optimisation of the sample purification and the SBE reaction, certain peaks were observed occasionally, mostly in the green dye electropherograms. We assume that some of these peaks originate from nontemplate addition of fluorescent ddATP to the 3'-end of the shortest amplicons. Since the electrophoretic mobilities of these amplicons are constant, we designed the GenoTyper 3.7 allele calling macros to label only the peaks in the predefined allele windows and ignore peaks outside the prescribed size windows.

The largest concern during analysis of the electropherograms was the difference in fluorophore emission from the four fluorophores dR110 (blue), dR6G (green), TAMRA (yellow) and dROX (red) used in the SNaPshot Multiplex System. On average, the ratio between the signal strengths of the colours was 4:2:1:1, and the peak thresholds in GeneScan 3.7 and the rules for heterozygous and homozygous allele calls had to be adjusted accordingly (see Section 2). For the large majority of the samples, this was sufficient to ensure robust and reproducible analysis of the electropherogram. However, for some weakly amplified samples, the differences in signal strengths of the four colours gave rise to 'no calls' for certain SNPs. All 700 samples analysed in the population validation work were typed twice and 80% of the samples gave results for all 52 SNPs in the first and second runs. In the other 20% of samples, between one and five of the 52 SNPs were designated no call in one of the typing runs, and a third analysis using increased amount of DNA in the PCR was necessary to make the correct allele call. No call results were equally likely in all SNPs and not confined to particular loci.

3.3 Validation of the 52 SNP-plex assay

PCR and SBE primers were distributed to the four consortium laboratories along with ten blood samples spotted on FTA cards. The ten samples were typed in all laboratories with identical results and this test was used to validate the successful implementation of the 52 SNP-plex assay in each laboratory before more extensive typing of samples began. In general, genomic DNA purified from blood by phenol/chloroform or column based methods were used in the PCR reaction, but Chelex-purified DNA also gave satisfactory amplification of all 52 SNP loci.

In order to analyse the performance of the multiplex PCR and SBE reactions, sequential dilutions of DNA from two different samples purified by the QIAamp DNA blood mini kit were made and different amounts of DNA from 68 pg to 70 ng were used in the PCR reaction (Fig. 2). As expected, allele drop-outs and unusual peak height ratios for heterozygote SNPs due to stochastic phenomena in the PCR reaction [17] were frequently observed when less than 200 pg of DNA was used. However, a complete SNP profile was obtained for both samples from only 500 pg DNA. The electropherograms with the best peak balance were obtained with 1–17 ng DNA, but acceptable results were also obtained with up to 70 ng DNA.

The influence of PCR inhibitors on the multiplex PCR was tested by addition of porcine haematin, a haeme derivative that is known to inhibit *Taq* polymerase [18]. Three samples with 10 ng template DNA were treated with haematin to final concentrations of 5, 10, 15, 20, 25, 30, 35 and 40 μM. Reduced PCR amplification products were observed with 5 μM haematin and complete inhibition of the PCR with 10 μM haematin (data not shown).

The DNA-regions surrounding 14 SNPs were sequenced to verify the allele calling of the SBE reactions. For each SNP, sequences for at least three individuals were obtained. All sequences were in agreement with the SNP-typing results.

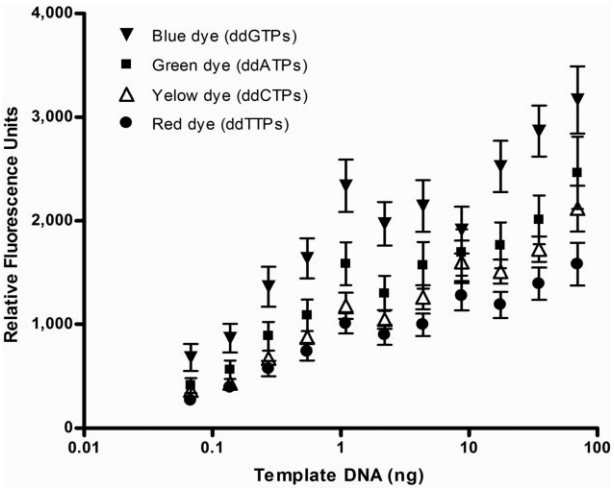

Figure 2. Sensitivity of the 52 SNP-plex assay. Different amounts of genomic DNA (0.07, 0.14, 0.27, 0.55, 1.09, 2.19, 4.37, 8.75, 17.50, 35 and 70 ng) from two individuals were used. RFUs of each of the four dyes were collated and normalized by dividing the RFU values from homozygote alleles by two. Normalised average RFUs are shown as a function of the amounts of template DNA in the multiplex PCR on a $\log_{10}$ scale. Error bars indicate the S.E.Ms.

The 52 SNP-plex assay was tested with DNA from dog, cat and horse, as well as 46 individual samples representing 6 different primate species. In total, 46 of the 52 SBE primers were extended in the SBE multiplexes in one or more of the primate species. In chimpanzees, 44 loci were typed, 37 loci were typed in orangutans, 36 in gorillas, 23 in baboons, 22 in rhesus monkeys and 21 in Cynomolgous monkeys. No amplification was detected from the dog, cat and horse samples. Table S3 shows the results obtained from the 46 primate samples, each sample being homozygous for all loci with the exception of 2 out of 29 chimpanzees that were heterozygous for SNP rs2056277 (#28). For the loci where SBE products were detected in more than one primate species, the same allele was observed in most cases. However, in five loci both the known human alleles were detected among the primates tested. For the other 41 loci, the frequencies of the alleles observed in primates indicated that the proposed ancestral alleles (alleles found in the last common ancestor of humans and chimpanzees) were marginally more frequent in Somalis (average: 0.53) than in Europeans and Asians (average for both: 0.48).

3.4 SNP allele distributions in nine different populations

A total of 700 samples from 9 different populations were typed with the 52 SNP-plex assay and the allele frequencies in each population determined (Table S4). The populations were divided in three groups: Somali, Asian (Chinese, Taiwanese, Thais, Greenlanders, Japanese) and European (Danes, Germans, Turks), and the combined allele frequencies compared to the allele frequencies in three major population groups; African–American, East Asian and European reported by The SNP Consortium, Celera and the RealSNP databases (Table S4, March 2005 release). The observed allele frequencies deviated by an average of 0.08 (156 comparisons) from the database allele frequencies, with a range from 0.00 to 0.31. Pearson correlation coefficients (r) were calculated to measure correlations between the allele frequency estimates in our grouped populations and the database allele frequencies. The correlation coefficients were r = 0.80 (Danes, Germans, Turks *vs.* European), r = 0.72 (Somalis *vs.* African-American) and r = 0.83 (Chinese, Taiwanese, Thais, Greenlanders and Japanese *vs.* East Asian). Similarly, correlation coefficients between allele frequencies from combinations of our nine populations and those of three of the four listed populations in the HapMap database were calculated. Not all of the 52 SNPs have been genotyped by HapMap, but where allele frequency estimates were listed, appropriate population groupings were compared: r = 0.51 (39 SNPs for Somalis *vs.* Yoruba in Ibadan, Nigeria), r = 0.94 (41 SNPs for combined Danes and Germans *vs.* Utah residents with ancestry from northern and Western Europe) and r = 0.96 (36 SNPs for combined Taiwanese and Chinese *vs.* Han Chinese in Beijing, China).

Cluster analysis using the model-based approach implemented in Structure v. 2.0 was performed with the 700 individuals. The splitting order illustrated in Fig. 3, was as follows: at K = 2, one cluster contained Somalis, Turkish, Danes and Germans, whereas Asians and Greenlanders were grouped together in another cluster. Each increase in K split one of the clusters obtained with the previous value. Thus, at K = 3, Somalis were separated from Turkish, Danes and Germans without altering the Asian and Greenland group. At K = 4, Greenlanders clearly split from the East Asian group. Higher values of K produced further equitable membership coefficients within the European and Middle East individuals without modifying the patterns of the other clusters (the proportion of the sample assigned to each population is roughly symmetric indicating lack of further substructure within these populations). The posterior probability was almost equal to one for K = 4 and virtually zero for other K values. Several individuals had high membership coefficients in more than one cluster. For instance, at K = 4, Somalis included two individuals with clear ancestry coefficients from the main European group, while the opposite pattern was seen in individuals in the European cluster (Fig. 3).

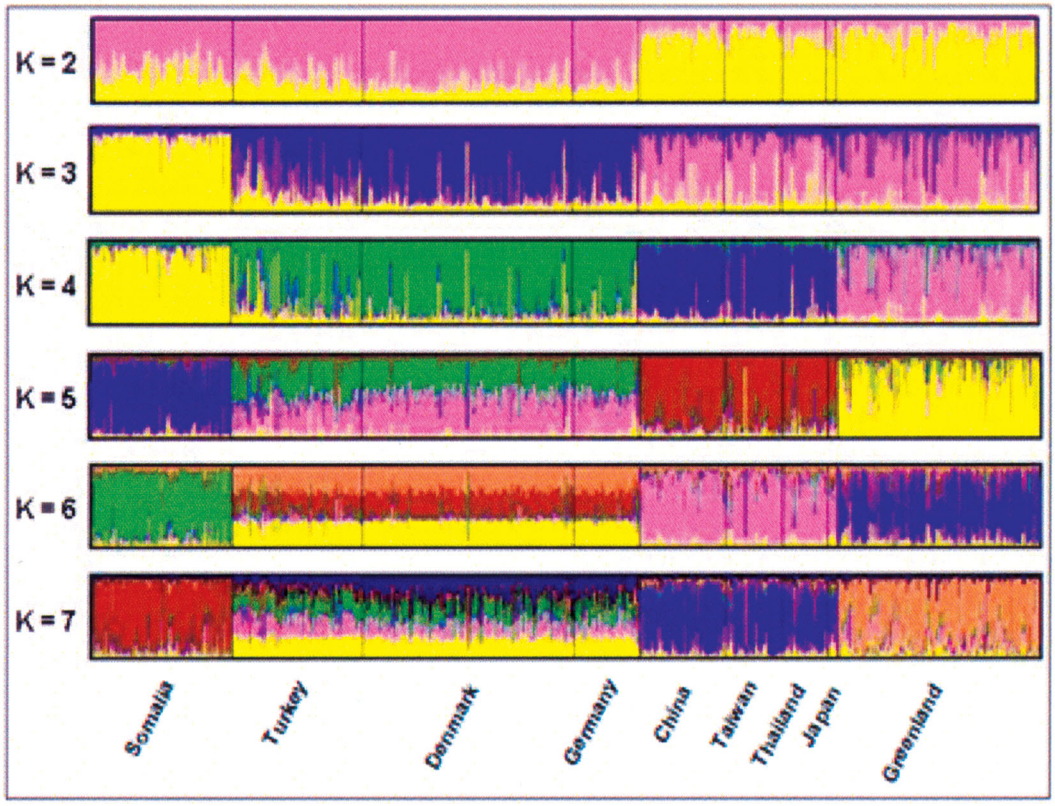

Figure 3. Cluster analysis. Each bar represents a single individual, and colours correspond to the coefficients of ancestry.

The genotype distributions of SNP rs907100 (#38) differed from expectations based on Hardy–Weinberg equilibrium in six of the nine populations due to an excess of individuals typed as homozygotes (χ^2 = 12.04, $p < 0.05$) and the deviations were still significant in three populations after Bonferroni correction [19]. In response to this, we are currently testing a new reverse PCR primer in 100 mother–father–child trios from Denmark and Somalia in order to determine whether the observed disequilibrium is related to a recently described SNP in position 9 from the 3' end of the reverse primer. The remaining SNP genotype distributions all matched expectations. Average heterozygosities were 0.44, 0.41 and 0.38 in Europeans, Somalis and Asians, respectively. The lowest heterozygosity was found in the Greenland and Taiwanese groups (both 0.37).

Association of alleles across loci (linkage disequilibrium – LD) was estimated using the χ^2-test, the pL10 coefficient [20] and the Exact test [21]. LD tests for pairs of SNPs on the same chromosome demonstrated no significant deviation from the expectations ($p > 0.05$ after sequential Bonferroni correction).

Table S5 shows analyses of variances for all 52 markers in European (F_{ST}), Somali (F_{is}) and Asian populations (F_{ST} including Greenlanders). The F_{ST} values ranged from 0.003 to 0.217 with a mean of 0.07. The distribution of the F_{ST} values is shown in Fig. 4. The tendency to bimodality was mainly due to the contribution of the European populations.

The allele distribution pattern showed uniformity in the studied populations with approximately 93% genetic variability within populations. To determine the distribution of residual genetic variance, we grouped the populations into Europeans, Somalis and Asians (including Greenlanders) and calculated the Φ statistics. Low values of both inter- and intragroup variability (Φ_{CT} = 0.054; Φ_{SC} = 0.035) were observed in each case.

3.5 Forensic statistics

The combined mean match probability using the 52 SNPs was between 5.0×10^{-19} (Asian) and 5.0×10^{-21} (European), corresponding to a combined power of

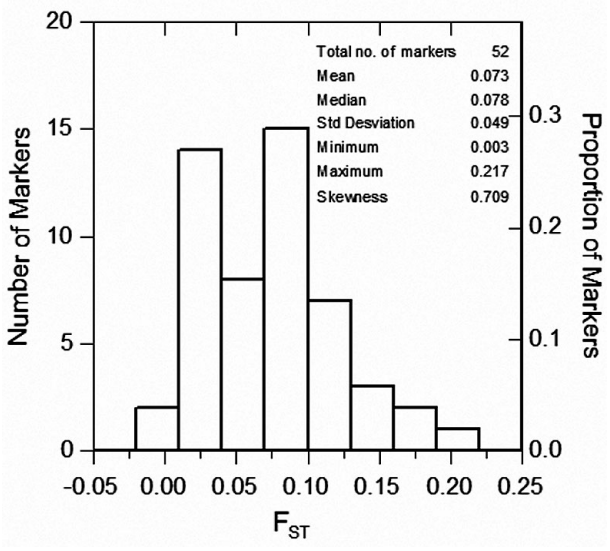

Figure 4. F_{ST} distribution of the 52 autosomal SNPs. Values were calculated for a global population including 700 individuals from Denmark (156), Greenland (149), Somalia (104), Turkey (96), China (63), Germany (49), Taiwan (43), Thailand (33) and Japan (7).

discrimination of 99.9999 and 99.99999%, respectively (Table 1). We performed pair-wise comparisons of the 700 samples typed in the validation study. The average number of identical SNP loci between two unrelated individuals was 21 with a minimum of 7 and a maximum of 41 matching loci (Fig. 6).

The typical paternity indices [13] obtained with the 52 SNPs ranged from 336 000 in Asians to more than 549 000 in Europeans, corresponding to a mean exclusion probability of 99.91 and 99.98%, respectively. In motherless cases, the typical paternity indices ranged from 2880 (Asian) to 4640 (European) (Table 1).

3.6 Typing of partially degraded DNA from crime case samples

Purified DNA from bone, muscle or other tissues collected from seven cadavers found under different environmental conditions and in various stages of decomposition, obtained previously as part of crime case investigations,

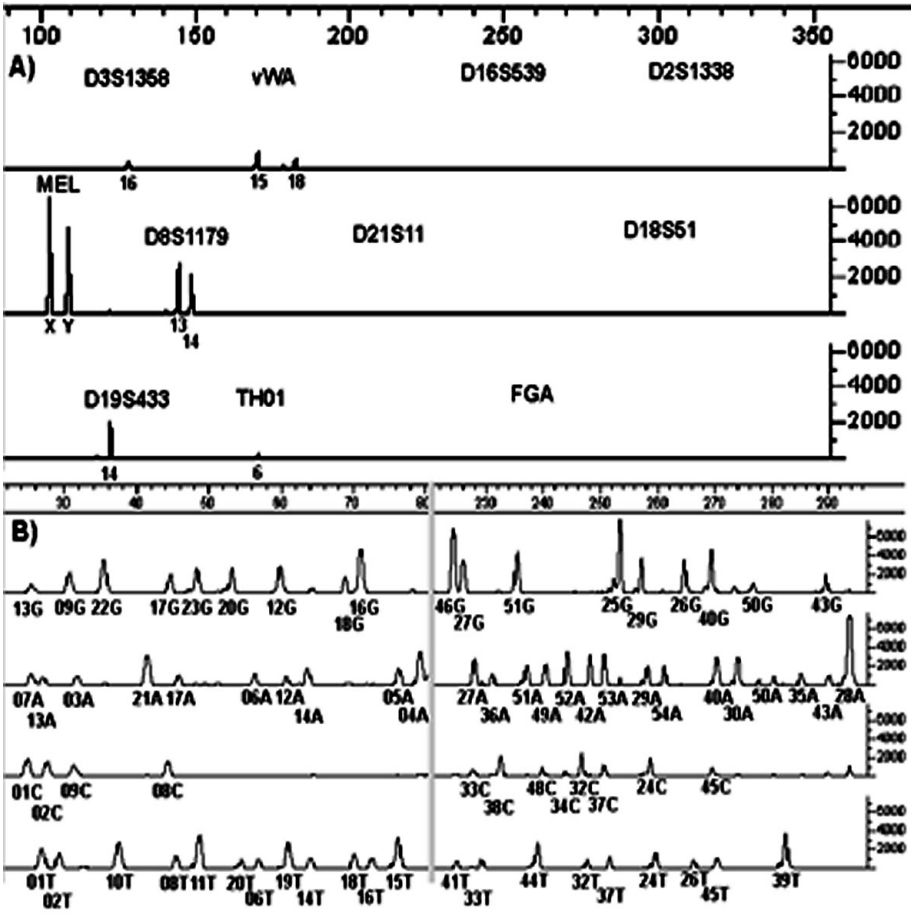

Figure 5. Representative electropherograms of (A) AmpF/STR SGM Plus and (B) 52 SNP-plex assay typing of a sample with partly degraded DNA. A total of 2.6 μL of 0.4 ng/μL DNA was used in the PCR amplifications.

Table 1. Forensic statistical parameters

	European	Somali	Asian
Mean match probability	5.0×10^{-21}	1.1×10^{-19}	5.0×10^{-19}
Combined power of discrimination	>99.99999%	>99.9999%	>99.9999%
Mean exclusion probability	99.98%	99.95%	99.91%
'Typical' paternity index (trios)	549 000	337 000	336 000
Typical paternity index (motherless)	4 640	3 160	2 880

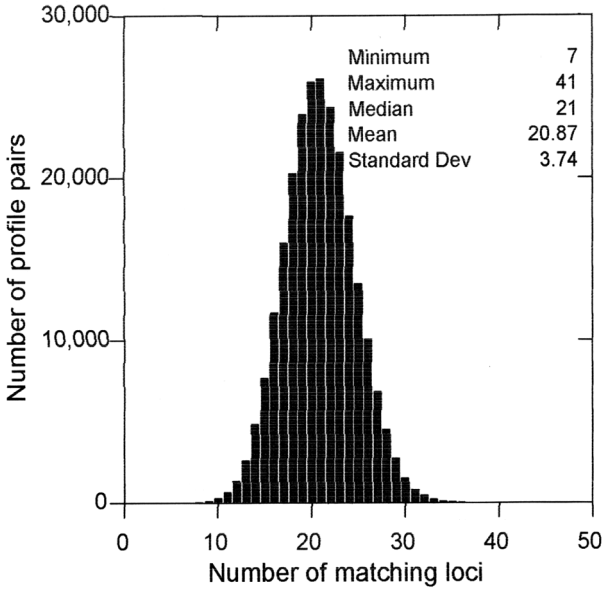

Figure 6. Distribution of the number of identical SNP loci found in unrelated individuals. Values were calculated for 700 individuals belonging to nine different populations. Total number of comparisons was 244 650.

were analyzed for ten STR loci using the AmpFLSTR SGM Plus PCR Amplification kit. The short-PCR fragments in the STR kit were amplified efficiently whereas the longer fragments (more than 200 bp) were very poorly amplified or not amplified at all (example shown in Fig. 5A), suggesting that the DNA was partially degraded. A series of dilutions of DNA from the seven samples were made and different amounts of DNA were used in the 52-plex PCR. When using concentrations identical to those used in the AmpF/STR SGM plus PCR (approximately 0.5 and 1 ng), all 52 SNPs were successfully typed (example shown in Fig. 5B). At lower concentrations (approximately 0.225 ng), the 52 SNP-plex assay also gave full SNP profiles for six of the seven samples (data not shown). Five of the SNPs (rs717302, rs729172, rs1015250, rs1528460 and rs1886510) consistently gave weak signals suggesting that they perform poorly under these conditions, but this did not correlate with the sizes of the amplicons; the sizes of the amplicons for these loci varied from 60 to 115 bp. Similarly, no relationship was observed between the amplicon sizes and successful SNP allele calls when genotyping the 700 blood samples (see above).

4 Discussion

This study demonstrates that it is possible to construct large, sensitive multiplex PCR assays for the detection of more than 50 SNPs for forensic applications. As a result, a set of SNP markers carefully tailored for use in human identification and readily genotyped with established technology is now available to the forensic community. STR typing is the most commonly used typing method for forensic genetic investigations in crime casework, but this technique may not always give conclusive results in cases with heavily degraded DNA [22]. In such instances, the size of the amplicon to be investigated is critical. With the widely used STR markers, the lengths of the amplicons range from approximately 100 to 400 bp. Efforts are being made to reduce the amplicon lengths of STR systems currently in routine use [23, 24]. In addition, new shorter STR loci are being identified and tested for practical forensic use [25]. SNP typing has the advantage that amplicons less than 60 bp in length may be investigated because the length of the amplicon is restricted only by the compositions of the flanking sequences and the lengths of the PCR primers. Therefore, in the 52 SNP-plex assay outlined here, amplicons have a maximum size of only 115 bp.

The most important factors allowing the construction of the present 52 SNP-plex assay were (i) a careful selection of primers to avoid intra- and interactions between the primers; (ii) high-quality primers that are pure and homogeneous; and (iii) careful balancing of the PCR multiplex and the SBE multiplex reactions [9]. Until now, only a few large multiplexes have been reported [26], but larger multiplexes that are constructed based on the same principles as the present 52-plex are emerging, *e.g.* packages with Y chromosome SNPs [15, 27] or autosomal SNPs with contrasting allele frequency distributions in different populations useful for the estimation of the population of origin (Phillips *et al.*, in preparation). In order to speed up the typing process and keep costs low, we developed a double injection method

that has the advantage that both the 23-plex and the 29-plex SBE products can be investigated in the same electrophoresis run and computer analysis. In addition, automatic allele calling macros have been developed and are available for use with this SNP set. These macros reduce both the analysis time involved and the risk of allele calling errors. Overall, the 52 SNP-plex assay performed well in four different laboratories when reaction mixtures were sent to the participating laboratories on dry ice. Preliminary results demonstrated that the 52-plex also performed well when applied to other SNP-typing platforms. Currently, 48 of 52 SNPs have been successfully incorporated into the Sequenom MassArray™ MALDI-TOF-based genotyping assay (Phillips *et al.*, unpublished), and 39 of the 52 SNPs have been implemented successfully onto the Nanochip™ Molecular Biology Workstation (Balogh *et al.*, unpublished). Furthermore, we are also implementing the 52-plex into alternative MALDI-TOF-based assays [28], conventional microarrays and oligonucleotide ligation-based assays.

The lack of polymorphism in various primates allowed the recognition of the ancestral allele in 41 of the 52 SNPs analysed. For five of the SNPs, both human alleles were detected in the primates studied. It is not possible to conclude whether this was caused by a higher mutation rate at these loci compared to the average SNP mutation rate, or if the observation reflects that the substitutions appeared before the genetic divergence of the primates studied and present-day humans.

The observed allele frequencies deviated on average 0.08 (maximum 0.31) from those reported by The SNP Consortium, Perlegen and HapMap databases when our data were pooled into three major groups. In the few cases where large frequency estimate differences were observed, it is possible that these resulted from divergence between the listed populations and those we analysed or that the samples sizes used in the SNP database validation were small or unrepresentative.

Although we had selected SNPs with allele distributions as close to 0.5/0.5 as possible in Europeans, Asian and African populations, using Cluster analysis (Fig. 3), we observed a broadly comparable population-grouping pattern to that found in a previous study in which a considerably higher number of polymorphic markers were used [29]. Although the interpretation of the patterns in our case requires caution due to our limited range of sampled populations and the very limited number of SNPs as well as the low power of the majority of the SNPs for prediction of the origin of population of individuals, grouping individuals into appropriate principal regions gave consistent results that provided a high correlation between predefined populations, geography and the inferred clusters. However, if one wants to use SNPs for prediction of the origin of population, the SNPs should be selected for this specific purpose. The observation that several individuals exhibited high membership coefficients in multiple clusters indicates a degree of coancestry in some of the populations sampled, suggesting some population admixture. This is consistent with the hypothesis that human populations are not discrete groups, since admixture commonly occurs between neighbouring populations [30]. However, the use of a limited number of SNPs in our analysis could also have contributed to the extent to which multiple group membership was observed.

For the European and East Asian population groups, we observed high-correlation (r) values when comparing the allele frequencies of the present study with those of HapMap. In contrast, the coefficient of correlation between the Hapmap Yoruba Nigerian population and our Somali populations was much lower, and this mirrors the high genetic diversity of the African continent [31, 32]. Clearly, further studies and a broader range of population samples from this continent will be needed to refine our knowledge of the allele frequency patterns for the 52 SNPs amongst Africans.

The genotypes of all loci studied except SNP rs907100 (#38) were distributed as expected based on the assumption of Hardy–Weinberg equilibrium. Recently, a new SNP (rs11689319) was reported in position 9 from the 3' end of the reverse PCR primer used to amplify rs907100. A new reverse PCR primer with a degenerate base in position 9 from the 3' end is currently being tested in order to determine whether the unexpectedly high number of homozygotes is caused by this substitution.

The heterozygosity values of the 52 SNPs in the tested populations were close to those expected with the highest average of 0.44 in Europeans. This is not surprising considering one of the primary SNP selection criteria was a priority for maximum heterozygosity in European populations with less emphasis on the same variability in other populations for most SNPs chosen. The low average F_{ST} value of 0.07 (maximum of 0.22 for the total group) underlines the fact that none of these markers has any known or likely functional relevance, another key factor in locus selection being sufficient distance between SNPs and neighbouring genes.

Due to the limited amount of DNA in many crime cases, it is important that typing of a sufficient number of SNPs can be performed reliably on small amounts of DNA from traces of evidential material. STR typing can often work on as little as 200 pg of target DNA. In our study partially degraded, purified DNA from seven different cadavers were typed for the full set of 52 SNP markers using 200–500 pg DNA (Fig. 5). In contrast, only the short-PCR fragments from the

AmpF*l*STR SGM Plus STR amplification kit were amplified efficiently in these samples. These results are very promising for practical crime casework in which only very small amounts of partially degraded DNA are available.

The mean power of discrimination (5.0×10^{-19} to 5.0×10^{-21}) would be satisfactory in crime cases with a full profile from only one contributor. However, if a mixture of DNA from two or more individuals is found, additional SNPs, including SNPs with unequal allelic distribution (*i.e.* much lower average minor allele frequencies), will be necessary in order to obtain a weight of evidence equivalent to that obtained by STR typing [3]. For paternity testing, the 52 SNP-plex assay will be most valuable in European populations (typical paternity index: 549 000) while the unequal distribution of the alleles in Asians and Somalis resulted in lower paternity indices (typical paternity indices: 336 000 and 337 000, respectively). This demonstrates the importance of selecting SNPs with allele frequencies close to 0.5 for relationship testing applications.

In conclusion, we have identified a set of 52 polymorphic SNPs that can be typed by standard methods and we have developed an assay that allows multiplex PCR amplification of very limited amounts of DNA. The SNP-typing procedure uses standard CE equipment available to most modern forensic genetic laboratories. Furthermore, the 52-plex SNP set can be readily adapted to a range of other genotyping methods offering the possibility of high-throughput solutions in the future. The SNPs presented are freely available and can be part of a future core set of SNPs for forensic genetic investigations.

We thank Ms. Annemette Holbo Birk for technical assistance, Jock Nielsen, PhD, for help in programming Access macros and Bo Simonsen, PhD for helpful discussions. The work was supported by grants to J. J. Sanchez from Ellen and Aage Andersen's Foundation, the European Commission (GROWTH programme, SNPforID project, contract G6RD-CT-2002-00844) and a grant from the 'Ministerio de Ciencia y Tecnología' (BMC2003-09822) to C. Phillips, M. Fondevila, A. Salas and A. Carracedo.

5 References

[1] Schneider, P. M., Martin, P. D., *Forensic Sci. Int.* 2001, *119*, 232–238.

[2] Martin, P. D., Schmitter, H., Schneider, P. M., *Forensic Sci. Int.* 2001, *119*, 225–231.

[3] Gill, P., *Int. J. Legal Med.* 2001, *114*, 204–210.

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

[5] Hao, K., Li, C., Rosenow, C., Wong, W. H., *Eur. J. Hum. Genet.* 2004, *12*, 1001–1006.

[6] Pfaff, C. L., Barnholtz-Sloan, J., Wagner, J. K., Long, J. C., *Genet. Epidemiol.* 2004, *26*, 305–315.

[7] Shriver, M., Kennedy, G., Parra, E., Lawson, H. *et al.*, *Hum. Genomics.* 2004, *1*, 274–286.

[8] Phillips, C., in: Carracedo, A. (Ed.), *Forensic DNA Typing Protocols. Series: Methods in Molecular Biology*, Humana Press, Totowa, NJ 2004, pp. 83–106.

[9] Sanchez, J. J., Børsting, C., Morling, N., in: Carracedo, A. (Ed.), *Forensic DNA Typing Protocols. Series: Methods in Molecular Biology*, Humana Press, Totowa, NJ 2004, pp. 209–228.

[10] Weir, B. S., Hill, W. G., *Annu. Rev. Genet.* 2002, *36*, 721–750.

[11] Pritchard, J. K., Stephens., M., Donnelly, P., *Genetics* 2000, *155*, 945–959.

[12] Jones, D. A., *J. Forensic Sci.* 1972, *12*, 355–359.

[13] Brenner, C. H., Morris, J. W., *Proceedings for The International Symposium on Human Identification* 1989, Promega Corporation, Madison, WI pp. 21–53.

[14] Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M. *et al.*, *Nature* 2001, *409*, 928–933.

[15] Sanchez, J. J., Børsting, C., Hallenberg, C., Buchard, A. *et al.*, *Forensic Sci. Int.* 2003, *137*, 74–84.

[16] Doi, Y., Yamamoto, Y., Inagaki, S., Shigeta, Y. *et al.*, *Leg. Med. (Tokyo)* 2004, *6*, 213–223.

[17] Krenke, B. E., Tereba, A., Anderson, S. J., Buel, F. S. *et al.*, *J. Forensic Sci.* 2002, *47*, 773–785.

[18] Wallin, J. M., Buoncristiani, M. R., Lazaruk, K. D., Fildes, N. *et al.*, *J. Forensic Sci.* 1998, *43*, 854–870.

[19] Rice, W. R., *Evolution* 1989, *43*, 223–225.

[20] Weir, B. S., *Biometrics* 1979, *35*, 235–254.

[21] Zaykin, D., Zhivotovsky, L., Weir, B. S., *Genetica* 1995, *96*, 169–178.

[22] Schneider, P. M., Bender, K., Mayr, W. R., Parson, W. *et al.*, *Forensic Sci. Int.* 2004, *139*, 123–134.

[23] Hellmann, A., Rohleder, U., Schmitter, H., Wittig, M., *Int. J. Legal Med.* 2001, *114*, 269–273.

[24] Butler, J. M., Shen, Y., McCord, B. R., *J. Forensic Sci.* 2003, *48*, 1054–1064.

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

[26] Dixon, L. A., Murray, C. M., Archer, E. J. Dobbins, A. E. *et al.*, *Forensic Sci Int.* 2005, *154*, 62–77

[27] Brión, M., Sanchez, J.J., Balogh, K., Thacker, C. *et al.*, *Electrophoresis* 2005, *26*, 4411–4420

[28] Mengel-Jørgensen, J., Sanchez, J. J., Børsting, C., Kirpekar, F., Morling, N., *Anal. Chem.* 2005, *77*, 5229–5235.

[29] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M. *et al.*, *Science* 2002, *298*, 2381–2385.

[30] Serre, D., Paabo, S., *Genome Res.* 2004, *14*, 1679–1685.

[31] Salas, A., Richards, M., De la Fe, T., Lareu, M. V. *et al.*, *Am. J. Hum. Genet.* 2002, *71*, 1082–1111.

[32] Sanchez, J. J., Hallenberg, C., Børsting, C., Hernandez, A., Morling, N., *Eur. J. Hum. Genet.* 2005, *13*, 856–866.

