Available online at www.sciencedirect.com
SCIENCE DIRECT
Forensic Science International 148 (2005) 107–112
Forensic Science International
www.elsevier.com/locate/forsciint

Selection of twenty-four highly informative SNP markers for human identification and paternity analysis in Koreans

Hwan Young Lee[a,b], Myung Jin Park[a], Ji-Eun Yoo[a], Ukhee Chung[a,b], Gil-Ro Han[b,c,d], Kyoung-Jin Shin[a,b,d,*]

[a]*Department of Forensic Medicine, College of Medicine, Yonsei University, Seoul 120-752, Republic of Korea*
[b]*Biometrics Engineering Research Center, Yonsei University, Seoul 120-752, Republic of Korea*
[c]*Department of Forensic Medicine, National Institute of Scientific Investigation, Seoul 158-707, Republic of Korea*
[d]*Human Identification Research Institute, Yonsei University, Seoul 120-752, Republic of Korea*

Received 5 December 2003; received in revised form 26 April 2004; accepted 29 April 2004
Available online 24 June 2004

Abstract

A number of DNA marker types suitable for human identification and parentage testing have been developed, of which single nucleotide polymorphisms (SNPs) merit attention as they are abundant, genetically stable, and amenable to high-throughput automated analysis. In this regard, 24 highly informative SNP markers representing each 22 autosome and both sex chromosomes were selected, and the allele and genotype frequencies of these SNPs were determined in a group composed of 30 unrelated Koreans. Based on frequency data from this group, the estimated probability of identity (P_I) and probability of paternity exclusion (P_E) with 22 autosomal SNP loci were 1.905×10^{-10} and 98.9%, respectively. The SNPs in this study offer a small but highly accurate database that will be an essential reference for SNP-based forensic application in the future.

Keywords: Single nucleotide polymorphism; Human identification; Paternity analysis; Koreans

1. Introduction

Until now, of a number of DNA marker types for human identification and parentage testing, short tandem repeat (STR) markers have been successfully used since they are highly polymorphic within various populations and simply interpreted to digital records of data [1,2].

Meanwhile, recent advances in genomic sequencing technology have enabled the routine acquisition of sequence data and allelic single nucleotide polymorphism (SNP) data from a broad range of sources [3,4]. SNPs are stable, inherited, biallelic, single base pair differences that are present in the human genome at a density of 1–10 per 1000 nucleotides [5,6]. Also, finding SNPs for assessment purposes is not a problem since more than 1 million can be accessed on public databases. The most notable is the SNP Consortium Ltd. (http://www.snp.cshl.org/) [7], which is a non-profit organization founded to provide genomic data. By the end of 2001, 1.4 million SNPs had been released into the public domain without intellectual property restrictions [8]. Whether SNPs will replace STRs is a matter of conjecture at present. However, the throughput potential of SNPs offers significant time and cost advantages. In addition, SNPs are better suited for the analysis of highly degraded DNA, since the distance between the primer binding sites can be designed to be very short. In this regard, we anticipated that SNPs will be used as efficient genetic markers for forensic DNA typing.

In an attempt to use an SNP-based marker system for forensic application, it is prerequisite to score the allele and genotype frequencies of each SNP in populations to develop an efficient marker system with an increased power of discrimination. Here, we report the identification of a set

* Corresponding author. Tel.: +82 2 361 5775; fax: +82 2 362 0860.
E-mail address: kjshin@yumc.yonsei.ac.kr (K.-J. Shin).

0379-0738/$ – see front matter
doi:10.1016/j.forsciint.2004.04.073

of SNPs with significant informativity and estimate the potential utility of these markers for human identification and paternity testing by using statistical methods.

2. Materials and methods

2.1. Genomic DNA samples

Thirty unrelated Korean DNAs comprised of 15 male and 15 female DNAs were analyzed in the present study. DNA was extracted from buccal swabs using a QIAamp DNA mini kit according to manufacturer's instructions (Qiagen, Hilden, Germany). K562 and 9947A standard DNAs were analyzed as well (Promega, Madison, WI).

2.2. Selection of highly informative SNPs

SNP data in the SNP Consortium (TSC) were employed in the present study. The genotype frequencies and genomic locations were used as first-level screen. To ensure that closely linked loci were not chosen, SNPs at separate chromosomal locations were selected as candidate markers. As markers with even allelic distributions have high observed heterozygosities, SNPs with a common 45:55–50:50 allelic distributions in three major populations (i.e., Caucasians, Asians and African–Americans) were screened preferentially from the allele frequency and genotyping database of TSC. Also, since the SNP screen was especially directed towards selecting those markers that would be useful in increasing the theoretical power of discrimination in Koreans as well as in three major populations, the additional selection criteria for candidate SNPs included the characteristics, such as allelic distributions in a range of 30:70–50:50 in Koreans. SNPs that met the above criteria were defined as highly informative, and the selected markers represent 22 autosomes and both sex chromosomes, respectively.

2.3. PCR amplification and SNP scoring

Primers used for DNA amplification are listed in Table 1. Some primers used were those previously described in the SNP Consortium, and others were designed using the Primer3 program (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). PCR was performed in a 25 μl reaction volume in a thermocycler PTC-200 DNA engine (MJ Research, Waltham, MA). For each PCR, 1–2 ng of genomic DNA was mixed with 2.5 μl of 10× GeneAmp PCR buffer (Applied Biosystems, Foster City, CA), 200 μM of each dNTP, 10 pmol each of forward and reverse primer and 1 U of AmpliTaq Gold DNA polymerase (Applied Biosystems). DNA was first denatured at 95 °C for 10 min followed by 35 cycles of 94 °C for 30 s, the appropriate annealing temperature for 30 s (Table 1), and at 72 °C for 30 s. The final extension was carried out at 72 °C for 7 min. To remove remaining single-strand primers and dNTPs, 7.5 μl of the PCR products were treated with 4 U of Exonuclease I (USB, Cleveland, OH) and 0.8 U of shrimp alkaline phosphatase (USB) and then incubated at 37 °C for 45 min. After enzyme inactivation at 75 °C for 15 min, SNP scoring was carried out by the single-base extension method [9] using SNaPshot™ Kit (Applied Biosystems) according to the manufacturer's instructions. Extension reactions were carried out in a final volume of 10 μl, containing 1 μl of the purified PCR product, 1 μl of SNaPshot™ Ready Reaction Premix, 2 μl of buffer (200 mM Tris–HCl, 5 mM $MgCl_2$, pH 9), and 1 pmol of each extension primer (Table 2). The reaction mix was subjected to 25 cycles of 96 °C for 10 s, 50 °C for 5 s and 60 °C for 30 s. To remove unincorporated fluorescent ddNTPs, the reactions were treated with 0.5 U of shrimp alkaline phosphatase (USB) at 37 °C for 45 min and then incubated at 75 °C for 15 min. Samples were analyzed by capillary electrophoresis using an ABI PRISM 310 Genetic Analyzer (Applied Biosystems) and GeneScan software 3.1 (Applied Biosystems). SNP scoring at each locus was confirmed by sequencing two homozygotes for each of the observed alleles, as well as two heterozygotes using a BigDye Terminator Cycle Sequencing v2.0 Ready Reaction kit (Applied Biosystems).

2.4. Statistical analyses

For population data, the gene count method was used to calculate observed allele frequencies and the observed genotypes at each of the 24 SNP loci. P_I was defined as the estimated probability that two unrelated individuals selected at random would possess identical multi-locus genotypes and was calculated for each SNP from genotype frequencies as previously described [10,11]. For parentage testing, P_E, defined as the probability that a random candidate male would be excluded from paternity with a known maternal genotype, was calculated for each SNP from allele frequencies, as previously described [12,13]. The calculation of P_I at X chromosomal SNP was carried out as described by Desmarais et al. [14]. Fisher's exact tests using the GDA (Genetic Data Analysis) program (http://lewis.eeb.uconn.edu/lewishome/gda.html) were performed to evaluate compliance with the Hardy–Weinberg equilibrium.

3. Results

3.1. Identification of highly informative SNPs in TSC database

As the theoretical power of discrimination of a marker set is correlated with allelic distribution and the number of unlinked markers, candidate SNP markers with a common 45:55–50:50 allelic distributions in three major populations were screened to represent each 22 autosome