Forensic Science International: Genetics 34 (2018) 71–80

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Construction and forensic genetic characterization of 11 autosomal haplotypes consisting of 22 tri-allelic indels

Xiaohong Zhao[1], Xiaogang Chen[1], Yuancun Zhao, Shu Zhang, Zehua Gao, Yiwen Yang, Yufang Wang[**], Ji Zhang[*]

West China School of Basic Medical Sciences and Forensic Medicine, Sichuan University, Chengdu, PR China

ARTICLE INFO

Keywords:
Tri-allelic
Indels
Haplotype
Forensic genetics

ABSTRACT

Insertion/deletion polymorphisms (indels), which combine the advantages of both short tandem repeats and single-nucleotide polymorphisms, are suitable for parentage testing. To overcome the limitations of the low polymorphism of di-allelic indels, we constructed a set of haplotypes with physically linked, multi-allelic indels. Candidate haplotypes were selected from the 1000 Genomes Project database, and were subject to the following criteria for inclusion: (i) each marker must have a minimum allele frequency (MAF) of ≥0.1 in the Han population of China; (ii) markers must exist in a non-coding region; (iii) the physical distance between a pair of candidate indels must be < 500 bp; (iv) the allele length variation of each indel from 1 to 20 bp; (v) different haplotypes must be located on different chromosomes or chromosomal arms, or be more than 10 Mb apart if on the same chromosomal arm; and (vi) they must not be located across a recombination hotspot. A multiplex system with 11 haplotype markers, comprising 22 tri-allelic indel loci distributed over 10 chromosomes was developed. To validate the multiplex panel, we investigated the haplotype distribution in sets of two and three-generation pedigrees. The results demonstrated that the haplotypes consisting of multi-allelic indel markers exhibited higher polymorphism than a single indel locus, and thus provide Supplementary information for forensic kinship identification.

1. Introduction

Insertion/deletion polymorphisms (indels) are a class of DNA polymorphism involving from 1 up to 1000 s of nucleotides and have been used in the field of forensic genetics for several years. Indels are classified into five different groups, namely single-nucleotide insertions/deletions, mononucleotide tandem repeats, 2–15-nucleotide tandem repeats, transposon insertions, and insertions/deletions of random DNA sequences. Among them, single-nucleotide insertions/deletions account for approximately 29.1% of all indels; mononucleotide tandem repeats and 2–15-nucleotide tandem repeats account for 29.5% of indels; transposons represent a very small fraction of indels (0.59%); and the remaining indels (40.8%) comprise a wide spectrum of random DNA sequences, ranging from 2 to 9989 bp in length, over 99% of which are less than 100 bp in length [1]. Indels are distributed throughout the human genome, with approximately one indel per 7.2 kb of DNA [1,2].

DNA re-sequencing traces were first used in 2006 to compute a comprehensive map of human indel variation with 415,436 unique indel polymorphisms identified [1]. With the completion of the 1000 Genomes Project in 2015 around 3.6 million indels were identified in 2504 individuals from 26 populations, with ~5.4% comprising multi-allelic indels [3]. Indels are considered to be potential genetic markers, but have been underutilized in forensic applications. Indels merge the advantages of both short tandem repeats (STRs) and single-nucleotide polymorphisms (SNPs); they can be analyzed on the same PCR and capillary electrophoresis (CE)-based platforms utilized for routine STRs genotyping, and they also show a low mutation rate similar to that of SNPs. The mutation rate of small indels has been reported to be approximately 0.20×10^{-9} per site per generation for insertions, and 0.53×10^{-9}–0.58×10^{-9} per site per generation for deletions [4]. However, as the majority of indels are di-allelic, their discrimination power is relatively lower than that of STRs, and more indel loci are required for forensic identification at the same level of discrimination

* Corresponding author at: No. 16, 3rd Section, South Renmin Road, Department of Forensic Genetics, West China School of Basic Medical Sciences & Forensic Medicine, Sichuan University, Chengdu, Sichuan, PR China.
** Corresponding author at: No. 17, 3rd Section, South Renmin Road, Department of Pathophysiology, West China School of Basic Medical Sciences & Forensic Medicine, Sichuan University, Chengdu, Sichuan, PR China.
E-mail addresses: wangyufang@scu.edu.cn (Y. Wang), zhangj@scu.edu.cn (J. Zhang).
[1] These authors contributed equally to this study.

https://doi.org/10.1016/j.fsigen.2018.02.001
Received 3 July 2017; Received in revised form 20 January 2018; Accepted 1 February 2018
Available online 06 February 2018
1872-4973/

as STRs [2,5–7]. To overcome the limitation of low polymorphism in di-allelic markers, multi-allelic indels have been advocated in recent years [4,8–11]. Recently, the potential usefulness of haplotypes for kinship analysis has been described [12–15]. The use of multiple SNP haplotypes, or even the integration of SNP-STRs [16] can increase the effectiveness of individual identification.

To improve the informative power of indel markers, we constructed 11 haplotypes, each comprising two tri-allelic indel loci that were tightly linked in terms of their physical map position, which may be used as a new type of forensic genetic marker. The results suggest that these haplotypes exhibited higher levels of polymorphism and could provide efficient Supplementary information for kinship testing.

2. Materials and methods

2.1. Family samples and DNA extraction

EDTA blood samples and cotton-swab oral samples were obtained from each individual of two- and three-generation families, which included 99 trios (father, mother and child), 6 two-generation families with two or more children (4 families comprising parents with two children, 1 family comprising parents with three children, and 1 family comprising parents with four children), and 5 three-generation pedigrees (grandparent-parent-child) from the Southern Han Chinese (CHS) population. All the subjects provided written informed consent for participation. The genomic DNA from each bio-sample was extracted using Chelex 100, as described previously [17].

2.2. Tri-allelic indel haplotype marker selection

The indel loci were derived from the 1000 Genomes Project database (http://www.ncbi.nlm.nih.gov/variation/tools/). Candidate markers were selected to meet the following criteria: (i) each marker must have a minimum allele frequency (MAF) of ≥ 0.1 in the Han population of China; (ii) markers must exist in a non-coding region; (iii) the physical distance between a pair of candidate indels must be < 500 bp; (iv) the allele length variation of each indel from 1 to 20 bp; (v) different haplotypes must be located on different chromosomes or chromosomal arms, or be more than 10 Mb apart if on the same chromosomal arm; and (vi) they must not be located across a recombination hotspot. After selecting potential haplotypes based on these criteria, each locus was tested for conformity to the Hardy-Weinberg equilibrium (HWE) model and for potential linkage disequilibrium (LD). The threshold *p*-values for the HWE test was set at < 0.002 after Bonferroni correction, and for LD tests was set at < 0.05. Finally, 11 candidate haplotype markers comprising 22 tri-allelic indel loci distributed over 10 chromosomes were acquired. Each haplotype was positioned on a different chromosome, except for rs199741060-rs545009081 and rs368149163-rs144664785, which were both located on the long arm of chromosome 4, with an inter-region distance of ~121.54 Mb. To assess the applicability of these markers in other populations, allele frequency and phased haplotype of 22 tri-allelic indels in all 2504 unrelated individuals from 26 populations were retrieved from 1000 Genomes Project Phase III.

2.3. PCR amplification and capillary electrophoresis (CE) genotyping

Table 1 shows the 22 tri- allelic indels selected for this study, including their primer sequences and amplicon lengths. Primers were designed using Primer 3 software (http://primer3.wi.mit.edu/). PCR amplification was performed in a 10 μL final reaction volume, including 1 μL of DNA template (2 ng/μL), 5 μL of Multiplex PCR Mix (Qiagen, Germany), 0.03–0.4 μM of PCR primers, and ddH_2O. Additionally, Q-Solution (Qiagen, Germany) was used to facilitate the amplification. The working concentrations of Q-Solution were 1× for 14 indel multiplexes and 0.5× for 8 indel multiplexes. The thermal cycling conditions consisted of an initial step at 95 °C for 15 min; 30 cycles of 94 °C for 30 s, 58 °C for 90 s, and 72 °C for 20 s; and a final extension at 72 °C for 90 min.

The amplified products were detected separately by a 3130 Genetic Analyzer (Applied Biosystems, USA) with POP-7™ polymer (Life Technologies, USA) as a separation matrix and a 36-cm Capillary Array. The data were analyzed by GeneMapper ID v3.2 software (Applied Biosystems). Negative control (ddH_2O) reactions were subject to the same conditions.

2.4. Allele nomenclature

In this study, alleles were named according to the structure of each indel shown in 1000 Genome Project. New alleles observed in this study were also named based on the allelic structure. All alleles at each locus were confirmed by bi-directional Sanger sequencing.

2.5. Haplotype construction

Although indels were typed separately, the genotype at a haplotype could only be identified unambiguously when both indels were homozygous, or when only one indel was heterozygous. However, the haplotype phase could not be inferred when both indels were heterozygous, as two haplotype combinations, i.e. cis and trans, are possible in this situation. PHASE version 2.1 [18] was used to estimate the unknown haplotypes in each region. The -p and -q options were used to set the thresholds at which the inferred-phase would be the most likely haplotype; the default settings for -p and -q were 90%.

2.6. Forensic parameters

The frequencies of haplotypes were estimated using PHASE version 2.1, and were compared to those retrieved from 1000 Genome Project Phase III using chi-square test, the threshold *p*-value was set at < 0.05. Statistical computations, including the allele frequencies of the indel markers and the HWE and LD tests, were performed to compare haplotype markers using PowerMarker V3.25. The expected heterozygosity (eHET), power of exclusion (PE) and probability of matching (PM) were calculated using an application developed in-house, and mean exclusion chance (MEC) of each haplotype region was calculated with the online application ChrX-STR.org 2.0 (http://www.chrx-str.org/).

2.7. Inference of haplotype phase

Even if indels are typed separately, the phase can be known unambiguously if both indels are homozygous or if only one is heterozygous. When the haplotype phase is ambiguous and the alleged father (AF) cannot be excluded, it is possible to compute the relative likelihood of the alternative possibilities based on Mendelian transition probabilities and genotype-to-haplotype probabilities. In terms of duo cases [one AF, one child (C) and no mother (M)], there are three genotype situations from which the explicit haplotype phase cannot be established, and, therefore, for which paternity cannot be excluded: First, the genotypes of the two indels in a haplotype are the same between the C and AF, and both indels are heterozygous; second, there is only one different allele between the C and AF, and both indels are heterozygous; or third, there are two distinct alleles between the C and AF, and both indels are heterozygous. The theoretical phases and shared haplotypes are illustrated in Fig. 1, and the calculations of the corresponding PIs are presented in Table 2. For trios (one AF, one C and one M), haplotype transmission between the M and C must be considered. Therefore, there are four genotype situations that must be considered to assess the possible transmission patterns. In some cases, the haplotype transmission pattern between AF and C is clarified by the M's genotype results. All possible genotyping results are illustrated in Fig. 2, and the calculations of the corresponding PIs are listed in