Forensic Science International: Genetics 22 (2016) 161–168

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Exploring the efficacy of paternity and kinship testing based on single nucleotide polymorphisms

Shao-Kang Mo[a], Ya-Cheng Liu[b], Sheng-qi Wang[a], Xiao-Chen Bo[a], Zhen Li[a], Ying Chen[c], Ming Ni[a,*]

[a] Department of Biotechnology, Beijing Institute of Radiation Medicine, 27 Taiping Road, Beijing 100850, People's Republic of China
[b] Beijing Tongda Shoucheng Institute of Forensic Science, 23 Yongtai Middle Road, Beijing 100192, People's Republic of China
[c] Department of Radiation Toxicology & Oncology, Beijing Institute of Radiation Medicine, Beijing 100850, People's Republic of China

ARTICLE INFO

Article history:
Received 27 August 2015
Received in revised form 22 February 2016
Accepted 24 February 2016
Available online 27 February 2016

Keywords:
Paternity and kinship testing
Single nucleotide polymorphisms
Likelihood ratio
Next-generation sequencing

ABSTRACT

Short tandem repeats (STRs) are conventional genetic markers typically used for paternity and kinship testing. As supplementary markers of STRs, single nucleotide polymorphisms (SNPs) have less discrimination power but broader applicability to degraded samples. The rapid improvement of next-generation sequencing (NGS) and multiplex amplification technologies also make it possible now to simultaneously identify dozens or even hundreds of SNP loci in a single pool. However, few studies have been endeavored to kinship testing based on SNP loci. In this study, we genotyped 90 autosomal human identity SNP loci with NGS, and investigated their testing efficacies based on the likelihood ratio model in eight pedigree scenarios involving paternity, half/full-sibling, uncle/nephew, and first-cousin relationships. We found that these SNPs might be sufficient to discriminate paternity and full-sibling, but impractical for more distant relatives such as uncle and cousin. Furthermore, we conducted an *in silico* study to obtain the theoretical tendency of how testing efficacy varied with increasing number of SNP loci. For each testing battery in a given pedigree scenario, we obtained distributions of logarithmic likelihood ratio for both simulated relatives and unrelated controls. The proportion of the overlapping area between the two distributions was defined as a false testing level (FTL) to evaluate the testing efficacy. We estimated that 85, 127, 491, and 1,858 putative SNP loci were required to discriminate paternity, full-sibling, half-sibling/uncle-nephew, and first-cousin (FTL, 0.1%), respectively. To test a half-sibling or nephew, an additional uncle relative could be included to decrease the required number of putative SNP loci to ~320 (FTL, 0.1%). As a systematic computation of paternity and kinship testing based only on SNPs, our results could be informative for further studies and applications on paternity and kinship testing using SNP loci.

1. Introduction

Paternity and kinship testing has useful applications for resolution of inheritance disputes, missing person searches, and even disaster victim identification. The genetic markers used for paternity and kinship testing include short tandem repeats (STRs), single nucleotide polymorphisms (SNPs), and small insertion/deletion polymorphisms (InDels) [1,2]. Core STR Loci systems, such as the U.S Combined DNA Index System (CODIS; 13 STR loci) and Extended European Standard Set (17 STR loci), have been widely used for decades in paternity and kinship testing and other human identity (HID) applications. Compared to STRs, HID SNPs are relatively new and have been incorporated into two main systems: the SNP*for*ID multiplex assay including 52 SNP loci proposed by Sanchez et al. [3], and the SNP panel for individual identification (IISNP) including 92 SNP loci developed by Kidd et al. [4–6].

Due to the binary polymorphisms of almost all SNP loci, the per-locus discrimination powers of SNPs are weaker than those of STRs, which usually have 7 to 16 polymorphisms [3,7,8]. In current research or applications of paternity and kinship testing, SNPs are more frequently regarded as complements to the STR-based approach. For instance, Lindner et al. [9] reported a paternity case in which the STR typing produced negative results due to mutations at three STR loci whereas the SNP*for*ID assay confirmed the relationship. In another study, Phillips et al. [2] used the

* Corresponding author.
E-mail addresses: 505704391@qq.com (S.-K. Mo), yachengliu@163.com (Y.-C. Liu), sqwang@bmi.ac.cn (S.-q. Wang), boxc@bmi.ac.cn (X.-C. Bo), lizhen0131@163.com (Z. Li), yingchen_29@163.com (Y. Chen), ni.ming@163.com (M. Ni).

http://dx.doi.org/10.1016/j.fsigen.2016.02.012
1872-4973/

SNP*for*ID assay alongside 42 STR loci for kinship testing and concluded that the SNP loci could remarkably increase the statistical power for relatively close kinship analysis.

As supplementary markers of STRs, SNPs have specific properties for paternity and kinship testing including a significantly lower mutation rate ($\sim 10^{-8}$ compared to 10^{-3}–10^{-4} for STRs) [7,8,10–13], broader applicability to challenging samples [1,14], and greater availability of loci [14–16]. SNP-based testing may also be preferred in cases dealing with highly degraded DNA (e.g., ancient DNA or samples from disaster victims), as SNP genotyping requires shorter amplicons than STR genotyping [1,14]. The low discriminative power for each SNP locus might also be compensated, to a certain extent, by typing additional independent loci. Compared to the conventional capillary electrophoresis approach, next-generation sequencing (NGS) enables simultaneous typing of a large number of SNP loci [15–27]. Recently, Eduardoff et al. [16] evaluated a NGS panel for 169HID SNPs (51 SNP*for*ID loci, 85 IISNP loci, and 33 Y-chromosomal SNP loci), and concluded that NGS provides sensitive and accurate genotyping applicable to forensic analysis.

Many previous works have evaluated the performance of SNPs for paternity and kinship testing by simulation investigation and/or testing in real cases as supplementary markers [2,3,7–11,13]. However, few studies have systematically assessed paternity and kinship testing based on SNPs alone, and the performance of SNPs in kinship testing of real cases remains unclear. In this study, we employed a multiplex HID panel and typed subjects from a Chinese Han population involving eight pedigree scenarios ranging from paternity to first-cousin. Moreover, we performed a computational simulation and derived the variations of testing efficacies with numbers of SNP loci in each of the eight pedigree scenarios.

2. Materials and methods

2.1. Sample collection and DNA extraction

Two donor families, referred to as family A ($n = 12$) and family B ($n = 4$), and eleven control donors were recruited. All subjects were from a Chinese Han population. Whole blood or buccal swab samples were collected from individuals. For samples from family A and control donors, total DNA was extracted by using PureLink® Genomic DNA kits (Invitrogen) following the manufacturer's instructions. For samples from family B, DNA was extracted by using DNA IQ™ System (Promega Biotech).

2.2. Target amplification and NGS

DNA samples were first quantified by both quantitative real-time PCR (qPCR) with Quantifiler® Trio DNA Quantification Kit (Applied Biosystems) and Qubit® 2.0 Fluorometer (Invitrogen). HID target amplification was performed by using the Ion Ampliseq™ HID SNP Panel (the community version, Life Technologies), and NGS libraries were prepared by using Ion AmpliSeq Library Kit 2.0 (Life Technologies). DNA samples from family A were amplified for 18 cycles with 1.5 ng (nanogram) DNA input, and samples from family B as well as control donors were amplified for 21 cycles with 1 ng input. Other amplification conditions were performed according to the manufacturer's instructions.

The NGS libraries were barcoded with the Ion Xpress Barcode Adapters (Life Technologies) as the protocol, and were quantified by using qPCR with the Ion Library Quantitation Kit (Life Technologies) on the 7500 Real-Time PCR System (Applied Biosystems). According to the quantification, libraries from family B were concentrated at room temperature for 30 min on the Concentrator plus (Eppendorf). The NGS platform of Ion Personal Genome Machine (PGM, Life Technologies) was employed to generate ~200 bp single-end reads with Ion PGM Template OT2 200 Kit, Ion PGM Sequencing 200 Kit v2, and Ion 316Chip V2. These NGS experiments were performed in three batches, respectively, for libraries from family A, family B, and controls.

2.3. SNP genotyping

The raw sequencing data were deposited in NCBI BioProject under the accession code PRJNA283235. All HID SNP genotypes, as well as numbers of reads for alleles were called by the HID SNP Genotyper Plugin v4.2 developed in Torrent Suite v4.2. The allelic balance for each heterozygous locus was calculated by the quotient of reads of alleles in the order A, C, G, and T.

2.4. Linkage disequilibrium between SNP loci

To examine linkage disequilibrium (LD) between SNPs in this study, the phase3 variant data were downloaded from the 1000 Genomes Project [28]. Based on genotypes of 134 autosomal SNP loci from SNP*for*ID and IISNP [15,16], a total of 853 unrelated individuals were selected by KING v1.4 [29], and the population frequencies of the 88 autosomal SNPs in this study were obtained (rs2269355 and rs938283 were not found). Then, by using PLINK v1.07 [30], the r^2 measures of LD were calculated between any pairs of the 88 autosomal SNPs.

2.5. Likelihood ratio model using autosomal SNPs for real samples

The widely used likelihood ratio (LR) model based on the Elston–Stewart algorithm was used for paternity and kinship testing on the autosomal markers [31–35]. Briefly, for a test person within a given pedigree, the model compared the likelihood value based on the genotypes of autosomal markers, denoted as L, for each of the two alternative hypotheses: H_0, the test person was the specific member in the pedigree; H_1, the test person was an unrelated person. Then the LR value was calculated as the ratio of $L(G|H_0)$ and $L(G|H_1)$ for each specific testing case, where G denoted genotypes. As parameters in calculation, the allele frequencies of the 90 autosomal SNPs were collected and revised from the NCBI dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/), and the mutation rate was set to a constant of 1.2×10^{-8} per bp per generation for all SNP loci according to Scally et al. [35]. For the 15 common STRs (13CODIS loci, D2S1338 and D19S433), the allele frequencies and locus-specific mutation rates were according to Butler et al. [36] and the STRBase (http://www.cstl.nist.gov/strbase/), respectively.

2.6. Putative pedigree simulations

To obtain the theoretical $\log_{10}$(LR) distributions based on the 90 autosomal HID SNP loci, in each pedigree scenario, 100,000 putative pedigrees were simulated for related cases and unrelated controls. Specifically, for a putative test person who is related, the putative pedigree was generated from all ancestors, whose genotypes were randomly generated based on allele frequencies of the population. Then, following the rules of genetic inheritance and mutations, genotypes of other pedigree members were derived, including the test person. For an unrelated control case, the test person was within a special pedigree in which he or she was treated as the ancestor for him/herself, and the genotypes were randomly generated. Based on genotypes of the putative pedigrees, the simulated $\log_{10}$(LR) distributions for both related cases and unrelated controls were calculated.

Moreover, to obtain the variation of testing efficacies with numbers of SNP loci and the limitations of efficacy elaboration, a