Forensic Science International: Genetics 34 (2018) 49–56

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Forensic genetic informativeness of an SNP panel consisting of 19 multi-allelic SNPs

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

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

ARTICLE INFO

Keywords:
Non-binary SNP
Tetra-allelic SNP
AIM-SNP
Forensic genetics

ABSTRACT

Current research focusing on forensic personal identification, phenotype inference and ancestry information on single-nucleotide polymorphisms (SNPs) has been widely reported. In the present study, we focused on tetra-allelic SNPs in the Chinese Han population. A total of 48 tetra-allelic SNPs were screened out from the Chinese Han population of the 1000 Genomes Database, including Chinese Han in Beijing (CHB) and Chinese Han South (CHS). Considering the forensic genetic requirement for the polymorphisms, only 11 tetra-allelic SNPs with a heterozygosity > 0.06 were selected for further multiplex panel construction. In order to meet the demands of personal identification and parentage identification, an additional 8 tri-allelic SNPs were combined into the final multiplex panel. To ensure application in the degraded DNA analysis, all the PCR products were designed to be 87–188 bp. Employing multiple PCR reactions and SNaPshot minisequencing, 511 unrelated Chinese Han individuals from Sichuan were genotyped. The combined match probability (CMP), combined discrimination power (CDP), and cumulative probability of exclusion (CPE) of the panel were 6.07×10^{-11}, 0.9999999999393 and 0.996764, respectively. Based on the population data retrieved from the 1000 Genomes Project, Fst values between Chinese Han in Sichuan (SCH) and all the populations included in the 1000 Genomes Project were calculated. The results indicated that two SNPs in this panel may contain ancestry information and may be used as markers of forensic biogeographical ancestry inference.

1. Introduction

Due to the reducible amplicon sizes, lower mutation rates [1,2], flexible detection methods [3–6], and phenotypic and ancestry information content [7–12], single-nucleotide polymorphisms (SNPs) have attracted the attention of forensic geneticists. Over the last decade, several forensic human identification panels based on bi-allelic and tri-allelic SNPs have been established [13,14], on the capillary electrophoresis system (CE) as well as on the next-generation sequencing system (NGS). Tri-allelic SNPs were also considered as an ideal choice for 2 sources DNA mixture analysis when there is presence of the third allele on one locus [2,15]. It was proposed that at least 50–60 binary SNP loci are required to achieve the same power of routine STR discrimination panels [16,17]. Such a large number of SNPs will increase the difficulty of multi-amplification and detection on either the CE or the NGS platform, particularly when applied to trace biological samples. One method for increasing the information content of SNP panels, and of decreasing the difficulty of constructing a PCR multiplex, is to use more informative SNPs.

During the course of the 1000 Genomes Project, the human genomic variation data released by the project have always been considered as one of the sources of highly informative DNA polymorphic loci for the purpose of forensic individual discrimination. Along with the publication of the Phase III human variant data set, in total, > 84.7 million SNPs, 3.6 million short insertions/deletions (indels) and 60,000 structure variants were characterized and phased onto high-quality haplotypes [18]. From this set of variant data, Phillips et al. [17] collated 961 tetra-allelic SNPs, of which four different nucleotide substitution alleles (A/T/C/G) can be detected in 2504 individuals from 26 populations. Among these tetra-allelic SNPs, only 160 loci meet the forensic genetic requirement for the allele frequency distributions among the four variant nucleotides, and may be used as forensic identification markers. They finally identified 24 tetra-allelic SNPs with good discriminatory power in Europeans or Africans, and few informative loci for East Asians. Their data also suggested that the 24 SNPs produced a cumulative random match probability (RMP) of 2.0E-15 in African

* Corresponding authors at: Ren Min Nan Road 3-17, Chengdu 610041, 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.2018.01.006
Received 15 November 2017; Received in revised form 7 January 2018; Accepted 26 January 2018
Available online 31 January 2018
1872-4973/

Table 1
Summary of the 48 tetra-allelic SNPs retrieved from 1000 Genomes data.

Chr.	rs-number	allele frequency [a]				Het [c]	DP [d]	Chr.	rs-number	allele frequency [a]				Het [c]	DP [d]
		A	T	C	G					A	T	C	G		
6	rs623327 [b]	0.334	0.289	0.188	0.190	0.7339	0.954	6	rs1071652 [b]	0.010	0.010	0.820	0.161	0.3019	0.530
8	rs4872345	0.284	0.161	0.447	0.108	0.6820	0.926	2	**rs170764**	0.017	0.132	0.010	0.841	0.2743	0.486
6	rs9272608 [b]	0.276	0.435	0.231	0.058	0.6777	0.926	6	rs35406650 [b]	0.861	0.050	0.002	0.087	0.2494	0.444
6	rs1071816 [b]	0.195	0.252	0.075	0.478	0.6640	0.915	8	**rs348146**	0.002	0.002	0.892	0.103	0.1939	0.359
6	rs1131115 [b]	0.103	0.502	0.180	0.214	0.6586	0.908	9	**rs10867536**	0.002	0.070	0.012	0.916	0.1562	0.292
6	rs1059546 [b]	0.188	0.040	0.435	0.341	0.6577	0.916	1	rs112146896	0.921	0.005	0.043	0.031	0.1495	0.279
7	rs59807595	0.305	0.010	0.250	0.435	0.6549	0.916	3	**rs74951201**	0.055	0.002	0.012	0.930	0.1313	0.248
6	rs2894257 [b]	0.034	0.163	0.344	0.459	0.6432	0.907	22	rs71315381	0.007	0.002	0.947	0.043	0.1010	0.194
14	rs200088812	0.495	0.327	0.161	0.017	0.6217	0.892	19	rs114930005	0.002	0.031	0.964	0.002	0.0698	0.136
6	rs9265769 [b]	0.063	0.075	0.423	0.440	0.6180	0.892	15	**rs74019021**	0.029	0.002	0.002	0.966	0.0653	0.127
6	rs1289632	0.005	0.310	0.168	0.517	0.6084	0.883	5	rs140053180	0.007	0.007	0.012	0.974	0.0519	0.101
6	**rs7767914** [b]	0.038	0.067	0.529	0.365	0.5808	0.864	8	rs55711507	0.005	0.002	0.017	0.976	0.0472	0.092
6	rs9256983 [b]	0.094	0.575	0.303	0.029	0.5686	0.8482	6	rs143117724	0.005	0.014	0.978	0.002	0.0426	0.084
7	**rs11767982**	0.002	0.171	0.630	0.197	0.5354	0.812	10	rs77338932	0.002	0.002	0.978	0.017	0.0425	0.083
19	**rs36622**	0.536	0.002	0.036	0.425	0.5303	0.832	9	rs186440475	0.012	0.002	0.981	0.005	0.0379	0.075
18	rs7241559	0.007	0.567	0.005	0.421	0.5011	0.808	2	rs111315781	0.005	0.988	0.002	0.005	0.0238	0.047
4	rs6811747	0.002	0.425	0.002	0.570	0.4944	0.803	19	rs12976854	0.002	0.005	0.005	0.988	0.0238	0.047
3	**rs76507337**	0.132	0.010	0.680	0.178	0.4880	0.761	6	rs150281870	0.988	0.002	0.007	0.002	0.0238	0.047
10	rs11594891	0.002	0.019	0.637	0.341	0.4773	0.774	9	rs183781474	0.002	0.002	0.988	0.007	0.0238	0.047
6	rs9268487 [b]	0.700	0.108	0.002	0.190	0.4629	0.735	3	rs114067668	0.005	0.002	0.002	0.990	0.0191	0.038
1	**rs596734**	0.002	0.668	0.005	0.325	0.4481	0.742	2	rs571214782	0.002	0.993	0.002	0.002	0.0144	0.029
9	rs10122930	0.786	0.070	0.058	0.087	0.3664	0.608	2	rs113171341	0.002	0.993	0.002	0.002	0.0144	0.029
21	rs11910088	0.026	0.010	0.149	0.815	0.3129	0.543	8	rs728155	0.002	0.002	0.002	0.993	0.0144	0.029
18	rs112321814	0.002	0.005	0.813	0.180	0.3073	0.542	12	rs58573421	0.002	0.002	0.002	0.993	0.0144	0.029

[a] Allele frequency calculated based on the CHB and CHS data of 1000 Genomes Project.
[b] SNPs sited in the MHC.
[c] Het: heterozygosity.
[d] DP: discrimination power.
SNP, single-nucleotide polymorphism; MHC, major histocompatibility complex. Markers marked in bold was used to construct the final multiplexed amplification system.

populations, but only 5.2E-10 in East Asian populations [17]; this means that it may be more difficult to identify suitable tetra-allelic SNPs for forensic genetic practice in Chinese individuals.

To construct a highly informative SNP panel for forensic application in the Chinese population, we screened the Phase III variant data of two Chinese Han populations, namely Chinese Han in Beijing (CHB) and Chinese Han South (CHS), employing the VCFtools package. A multi-allelic SNP panel consists of 11 tetra-allelic SNPs, and 8 tri-allelic SNPs were constructed based on SNaPshot and the CE system. The combined discrimination power (CDP) of this panel reached 0.9999999999393 in the Chinese Han population. The Fst values suggested that rs596734 in this panel may provide ancestry information between two major continental groups, namely African (AFR) and East Asian (EAS), whereas rs7767914 may have population discrimination power within East Asian biogeographic regions.

2. Materials and methods

2.1. Tetra-allelic and tri-allelic SNP selection

Target SNPs were filtered from the published 1000 Genomes Phase III using VCFtools (https://vcftools.github.io/man_latest.html). The original aim of this study was to construct a tetra-allelic SNP panel for forensic genetic applications. Based on a literature review and our preliminary study, we concluded that there may not be enough tetra-allelic SNPs suitable for constructing a highly informative panel. Thus, in the initial step, we retrieved all the tetra- and tri- allelic SNPs from 1000 Genomes Phase III. The only filter we set was that these SNPs should contain at least three alleles in two Chinese Han populations, including CHB and CHS (208 individuals in total). The candidate SNPs should match the following criteria: (1) the heterozygosity of tetra-allelic SNP should be > 0.06 in Chinese Han individuals; (2) the heterozygosity of candidate tri-allelic SNPs should be > 0.65; (3) the candidate SNPs should be located on the non-coding sequence; (4) and the SNP sites should be checked of the uniqueness and validity.

2.2. DNA samples

A total of 511 EDTA blood samples were collected from non-related and healthy Chinese Han individuals who resided in Chengdu, in the Southwestern part of China. DNA was extracted using the TIANamp Genomic DNA kit (TIANGEN, China), according to the manufacturer's protocol. The collected DNA was quantified with NanoDrop ND-1000 (Thermo Scientific, USA).

2.3. Primer design for multi-PCR and single-base extension (SBE)

PCR primers were designed with the Oligo 7 software (http://www.oligo.net/). PCR amplicons were kept to < 200 bp. SBE primers were designed with the SBEprimer program [19]. To distinguish between different loci on the CE system, SBE primers were tailed at the 5′-end with a poly-C sequence or a poly-GCCTCC $(TCCCC)_n$ sequence to produce serial SBE products of different sizes. To avoid the effect of primer-dimer and hairpin structures, the homology and complementarity of primers was analyzed using the AutoDimer program [20]. All the designed primers were checked for the absence of any other forms of polymorphism at the 3′-end to avoid potential null alleles, apart from the SBE primer of rs623327. As shown in Supplementary Table 1, a degenerate SBE primer was designed for this tetra-allelic SNP, since this SNP has a closely connected di-allelic SNP, which is only three nucleotides away from the polymorphic site. Primers were synthesized by Thermo Fisher Scientific, Inc. (Shanghai, China), and SBE primers were purified with polyacrylamide gel electrophoresis (PAGE). All primer information is listed in Supplementary Table 1.

2.4. PCR and SNaPshot™

A multiplex PCR system was constructed with the 19 validated