Forensic Science International: Genetics 8 (2014) 187–194

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Developing a novel panel of genome-wide ancestry informative markers for bio-geographical ancestry estimates

Jing Jia [a,b], Yi-Liang Wei [b,c], Cui-Jiao Qin [b], Lan Hu [b], Li-Hua Wan [a,*], Cai-Xia Li [b,**]

[a] Department of Laboratory Medicine, Chongqing Medical University, Chongqing 400016, PR China
[b] Key Laboratory of Forensic Genetics, Institute of Forensic Science, Ministry of Public Security, Beijing 100038, PR China
[c] Key Laboratory of Medical Epigenetics, Tianjin Research Center of Basic Medical Sciences, Tianjin 300070, PR China

ARTICLE INFO

Article history:
Received 22 April 2013
Received in revised form 6 September 2013
Accepted 9 September 2013

Keywords:
SNPs
Ancestry informative markers
Ancestry inference
Population genetics
Multiplex assay

ABSTRACT

Inferring the ancestral origin of DNA samples can be helpful in correcting population stratification in disease association studies or guiding crime investigations. Populations throughout the world vary in appearance features and biological characteristics. Based on this idea, we performed a genome-wide scan for SNPs within genes that are related to physical and biological traits. Using the HapMap database, we screened 52 genes and their flanking regions. Thirty-five SNPs that displayed highly contrasting allele frequencies ($F_{st} > 0.3$, linkage disequilibrium $r^2 < 0.2$, and Hardy–Weinberg equilibrium $P > 0.001$) among Africans, Europeans, and East Asians were selected and validated. A multiplexed assay was developed to genotype these 35 SNPs in 357 individuals from 10 populations worldwide. This panel provided accurate estimates of individual ancestry proportions with balanced discriminatory power among the three continental ancestries: Africans, Europeans, and East Asians. It also proved very effective in evaluating admixed populations living in joint regions of continents (e.g., Uyghurs and Indians) and discriminating some subpopulations within each of the three continents. Structure analysis was performed to establish and evaluate the panel of ancestry-informative markers, and the components of each population were also described to indicate the structural composition. The 21 population structures in our study are consistent with geographic patterns, and individuals were properly assigned to their original ancestral populations with proportion analyses and random match probability calculations. Thus, the panel and its population information will be useful resources to minimize the effects of population stratification in association analyses and to assign the most likely origin of an unknown DNA contributor in forensic investigations.

1. Introduction

Population stratification among groups poses a general concern in genome-wide association studies (GWAS). Even minor stratification can have a substantial impact on population-based studies with large sample sizes [1]. Analyzing the genetic structure of populations and individuals serves as a powerful tool to correct population stratification in association studies, as well as for ancestry inference in modern police investigation [2]. According to genetic profiles, more specific information can be obtained to estimate the relevance among suspects, victims, and databases of individuals [3,4]. Empirical analyses show that continental population groups can be identified by examining differences in allele frequencies [5,6]. Polymorphisms that exhibit large allele frequency differences between populations are named ancestry-informative markers (AIMs) and can be used to infer a population's or individual's ancestral origins. Over the last several years, AIMs have been broadly applied to infer individual bio-geographical ancestry and admixture proportions [2,7–9]. Methods or algorithms to measure and, therefore, assess differences in population structure have also been developed [10–13]. Several studies demonstrate that SNPs distributed throughout the genome have very large differences in allele frequencies between two or more continental populations [14–16]. AIM sets of ≤200 markers have the ability to discern continental structure [2,7,17,18]. Therefore, to accurately and cost-effectively evaluate bio-geographical ancestry, a small but robust set of AIMs is possible and highly desirable for association analyses and forensic practice.

Numerous global studies describe correlations between population geographical distribution and variations in the allele frequencies that are linked to several human phenotypes [19,20], including skin, hair, and iris pigmentation [21–23],

* Corresponding author. Tel.: +86 23 68485137; fax: +86 23 68485137.
** Corresponding author. Tel.: +86 01 66269503; fax: +86 01 66269503.
E-mail addresses: jia.jing628@gmail.com (J. Jia), lihuawan@yahoo.com (L.-H. Wan), licaixia@tsinghua.org.cn (C.-X. Li).

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.09.004

biological metabolism [24], biological modification variants [25], disease susceptibility [26], and morphology [27], because these variations are expected to display great population diversity. Thus, to develop a panel of combined SNPs that enable the inference of continental ancestral affiliation with proper discriminating power, we performed a genome-wide scan to identify the physical and biological trait-related genes involved in forming various group-specific features. In our study, a three-stage approach was used to develop a panel of 35 AIMs optimized to characterize individuals from three main continents (Africa, Europe, and East Asia), as well as to estimate relative ancestral proportions, especially in admixed individuals such as Uyghurs and Indians. First, informative AIMs were selected from the genes related to phenotypes using data from HapMap phase 1, phase 2, and phase 3. Second, evaluated population structure analyses were performed based on the dataset of the 11 populations from the HapMap database. Finally, we developed a multiplexed genotyping assay utilizing the SNPstream® 12-plex microarray platform (Beckman Coulter, Brea, CA, USA) based on single base extension technology and genotyped 357 samples collected from 10 populations originating throughout the three continents to estimate ancestral differences and determine population structure. This autosomal AIMs panel can provide reliable ancestry estimates and was validated with respect to both theoretical and actual performance for practical analysis. For individual inference, we described the components of the three continents and calculated the random match probabilities in each of the 21 populations. Our study also provides genotypes of continental populations as a research data resource.

2. Materials and methods

2.1. Population samples and data filtration

Our study involved 21 populations and a total of 1453 individuals from two sources, among which 357 samples from 10 populations were collected in our laboratory, and data from 1,096 individuals of 11 populations were downloaded from the HapMap database (Table S1) [28,29]. All collected samples were obtained with written inform content and self-declared ancestry information from the last three generations. DNA was isolated from circulating lymphocytes using a QIAamp® DNA blood midi kit (QIAGEN, Hilden, Germany) or from swabs using a QIAamp® DNA Mini M48 Kit (QIAGEN, Hilden, Germany). DNA quantification was performed on a 1.5-μL DNA sample in solution using a NanoDrop 2000c Spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA) or with real-time PCR. One DNA sample was quantified with real-time PCR using a Quantifiler™ Human DNA Quantification kit (Life Technologies, Carlsbad, CA, USA) on an AB 7500 (Life Technologies) and serially diluted (0.016 ng/μL, 0.031 ng/μL, 0.062 ng/μL, 0.125 ng/μL, 0.25 ng/μL, 0.5 ng/μL and 1.0 ng/μL) for sensitivity tests.

All samples underwent quality control procedures. For populations including parent/child trios or duos (ASW, CEU, MEX, MKK, and YRI), only genotypes from the parents were used. In addition, samples for known cryptically related individuals were removed [30]. Samples with more than three missing genotypes of the 35 loci were removed.

2.2. SNP typing

SNP typing was performed using the SNPstream® 12-plex microarray platform (Beckman Coulter, Brea, CA, USA). Primer design was performed using the Autoprimer.com online service and modified with PRIMER PREMIER 5.0 (PREMIER Biosoft, Palo Alto, CA, USA). All 35 SNPs were arranged into three PCR pools. Details of markers, PCR primers, extension primers, and product sizes are outlined in Table S2. The 45-bp extension primer was composed of a 5′ tag array tail (20 nt) and a 3′ SNP region (25 nt). Amplification reactions used a total PCR volume of 5 μL comprising: 1× Qiagen HotStar *Taq* PCR buffer (containing 15 mM $MgCl_2$), 25 mM $MgCl_2$ to a final concentration of 5 mM, 10 mM dNTP mix to a final concentration of 0.075 mM, 0.5 U Qiagen HotStar *Taq* polymerase, 0.025 μL of 24 premixed PCR primers with a final concentration of 0.05 μM each, 2 μL of DNA (DNA concentration in the range of 0.1–10 ng/μL), and 1.6375 μL of ddH_2O. PCR cycling comprised: 15 min at 95 °C, 40 cycles of 94 °C for 30 s, 55 °C for 30 s, and 72 °C for 1 min, followed by a final extension at 72 °C for 10 min.

PCR products were cleaned up prior to extension using Exonuclease I and Shrimp alkaline phosphatase (SAP; USB Products, Affymetrix, Santa Clara, CA, USA). A total clean-up reagent volume of 3 μL was added to the PCR product, comprising 2 U Exonuclease I, 1 U SAP, 1× SAP buffer, and 1.5 μL ddH_2O, then incubated at 37 °C for 30 min and 96 °C for 10 min. Single base extension (SBE) reactions used a total volume of 15 μL, comprising: 8 μL of clean PCR product, 3.97 μL SNPstream extension mix, 0.03 μL of 12 premixed SBE primers with a final concentration of 0.02 μM each, and 3 μL ddH_2O. SBE cycling involved 3 min at 96 °C, followed by 46 cycles of 94 °C for 20 s and 40 °C for 11 s.

For hybridization, SBE products were added to hybridization mix (7.56 μL hybridization solution and 0.44 μL hybridization additive) provided by Beckman Coulter. After washing with wash buffer I, hybridization plates were incubated for 2 h ± 15 min, and then, the plate was washed with wash buffer II. The dried plate was imaged in the SNPstream genotyping system (Beckman Coulter, Brea, CA, USA). Genotyping data were obtained with the GetGenos software package (Beckman Coulter, Brea, CA, USA).

2.3. Stage one: AIMs selection

We defined each population group by combining two HapMap population samples comprising: Africans (77 LWK and 98 YRI), Europeans (104 CEU and 70 TSI), and East Asians (76 CHB and 74 JPT). The δ value representing the allele frequency difference between any of two pair-wise populations was used to filter the SNPs [31,32]. A total of 52 candidate genes were screened: FDPS, ASH1L, DCST1, SLC45A2, HMGCR, COL4A3BP, POLK, AP3B1, EGFR, CYP3A4, TYRP1, PDLIM1, TYR, DRD2, VDR, HMGA2, KITLG, DCT, TGDS, GPR180, SLC24A4, RIN3, OCA2, HERC2, HERC2P9, APBA2, SLC24A5, SLC12A1, CYP19A1, GLDN, BCL2L10, MYO5C, CYP1A2, LMAN1L, SCAMP2, MPI, COX5A, SCAMP5, FANCA, SPIRE2, MC1R, DBNDD1, GAS8, PRDM7, CSH1, ASIP, AHCY, TNFRSF13C, CENPM, SEPT3, WBP2NL, and NAGA. Finally, SNPs were chosen as our panel of AIMs on the basis of the following criteria: 1) any loci with a Hardy–Weinberg equilibrium (HWE) $p < 0.001$ was cut off; 2) population-specific markers, comprising loci with a polymorphism detected in one or two population groups but absent in the other(s) or SNPs with a common allele in one population that is rare in others, using $\delta > 0.5$ to qualify [7]; 3) the F_{st} value of the three ancestral group was calculated using Genepop 4.2 [33,34] and only loci with $F_{st} > 0.35$ were selected; 4) Jensen and Shannon's divergence [35] was also calculated using the online software Snipper (at: http://mathgene.usc.es/snipper/); and 5) Haploview 4.2 [36] was used to ensure this set of AIMs had no linkage disequilibrium (LD) or different haplotypes among the population groups.

2.4. Stage two: validation of the 35 AIMs

To validate the panel of AIMs, estimates of ancestry proportions using the panel were compared to the self-identified ancestry within the HapMap data. Individual ancestry components were