Legal Medicine 17 (2015) 245–250

Contents lists available at ScienceDirect

Legal Medicine

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

Ancestry informative markers for distinguishing between Thai populations based on genome-wide association datasets

Kornkiat Vongpaisarnsin [a,*], Jennifer Beth Listman [b], Robert T. Malison [b,c], Joel Gelernter [b,c,d,e]

[a] Department of Forensic Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand
[b] Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA
[c] VA Connecticut Healthcare System, West Haven Campus, West Haven, CT, USA
[d] Department of Genetics, Yale University School of Medicine, New Haven, CT, USA
[e] Department of Neurobiology, Yale University School of Medicine, New Haven, CT, USA

ARTICLE INFO

Article history:
Received 16 December 2014
Received in revised form 16 February 2015
Accepted 19 February 2015
Available online 25 February 2015

Keywords:
Ancestry informative markers
Thai
Population
Fst
Forensic
SNP

ABSTRACT

The main purpose of this work was to identify a set of AIMs that stratify the genetic structure and diversity of the Thai population from a high-throughput autosomal genome-wide association study. In this study, more than one million SNPs from the international HapMap database and the Thai depression genome-wide association study have been examined to identify ancestry informative markers (AIMs) that distinguish between Thai populations. An efficient strategy is proposed to identify and characterize such SNPs and to test high-resolution SNP data from international HapMap populations. The best AIMs are identified to stratify the population and to infer genetic ancestry structure. A total of 124 AIMs were clearly clustered geographically across the continent, whereas only 89 AIMs stratified the Thai population from East Asian populations. Finally, a set of 273 AIMs was able to distinguish northern from southern Thai subpopulations. These markers will be of particular value in identifying the ethnic origins in regions where matching by self-reports is unavailable or unreliable, which usually occurs in real forensic cases.

1. Introduction

A new genotype sequencing technology providing large human genome information in a genome-wide association (GWA) study identified many candidate single nucleotide polymorphisms (SNPs). These SNPs could serve as potential genetic markers to assess the vulnerability to a variety of common human diseases [1] and could be useful in forensic human identification [2,3]. An abundance of SNPs demonstrate allelic frequency differences at either the individual or the population level. SNPs are useful markers for population genetic, anthropological and forensic studies that observe the genetic components that are shared in particular populations. SNPs expected to have high heterozygosity are proposed as informative loci to identify individuals, whereas those with low heterozygosity loci that are restricted for each ethnic group are valuable for distinguishing populations or for ancestry studies [4].

Most GWA studies ascertain population stratification with minimal errors by employing universal SNP markers from various population groups and ancestries that reflect global ancestry information; however, there are no local ancestral variations. Regional assessment in the latter group is more considerable, particularly when populations have recently been admixed [5]. Selecting the appropriate SNPs and ancestry informative markers (AIMs) to evaluate the overall genetic admixture proportion could help to identify individual geographic origins and distinguish between diverse population groups. Many highlighted AIM panels have been studied to classify subgroups within European populations [6–11], while Southeast Asian populations have rarely been studied genetically. The purpose of this study is to describe the genetic admixture and identify the best AIMs to classify Thai population. We used a population statistics model to cluster groups with a particular set of markers, and we present a genetic structural pattern that explains the origin of ancestry in this area.

2. Materials and methods

2.1. Population datasets

Genetic sequences of the following 11 human populations (1301 individuals) were obtained from the international HapMap

* Corresponding author at: Department of Forensic Medicine, Faculty of Medicine, Chulalongkorn University, 1873 Rama 4 Road, Pathumwan, Bangkok 10330, Thailand. Tel.: +66 22564269.
E-mail addresses: kornkiat.v@chula.ac.th, v_kornkiat@yahoo.com (K. Vongpaisarnsin).

http://dx.doi.org/10.1016/j.legalmed.2015.02.004
1344-6223/

phase 3 release 2 NCBI build 36 [12]: individuals of African descent in the southwest USA (ASW, 90 individuals), Utah residents of Northern and Western European descent from the CEPH collection (CEU, 180 individuals), Han Chinese in Beijing, China (CHB, 90 individuals), Chinese in Metropolitan Denver, Colorado (CHD, 100 individuals), Yoruba in Ibadan, Nigeria (YRI, 180 individuals), Gujarati Indians in Houston, Texas (GIH, 100 individuals), Tuscans in Italy (TSI, 100 individuals), Japanese in Tokyo, Japan (JPT, 91 individuals), Maasai in Kinyawa, Kenya (MKK 180 individuals), individuals of Mexican descent in Los Angeles, California (MEX, 90 individuals) and Luhya in Webuye, Kenya (LWK, 100 individuals). The Thai population data (THA, 374 individuals) were obtained from a Thai depression GWA study (186 cases and 188 controls) [13] with self-identified ethnic groups in accordance with Thai geographic sub-regions comprising northeast (THA-NOE, 150 individuals), north (THA-NOR, 64 individuals), south (THA-SOU, 67 individuals) and central (THA-CEN, 93 individuals).

2.2. Genotyping and data filtering

The international HapMap dataset comprised 1,440,616 SNPs. Thai population data covering 570,706 SNPs were genotyped on Illumina 650Y platforms. Sex chromosomes and individuals who originated from the same family were excluded from the analysis. Large-scale genotype data were filtered using PLINK software [14]. The SNPs were subsequently excluded according to the following criteria: SNPs with more than 10% missing genotypes, SNPs with a minor allele frequency less than 0.05, SNPs with a genotype frequency that failed a Hardy–Weinberg test at a significant threshold ($p < 10^{-7}$ by the chi-square test) and SNPs with A/T or C/G variants. The SNPs merged with some mismatch strands between each set were flipped.

2.3. Selected AIMs

The measurement of genetic differentiation was conducted using F*st* estimation, which evaluates the allele frequency differences between major populations and subpopulations and was originally defined by Weir and Cockerham [15]. Using two data sets (HapMap-Thai and Asian-Thai), SNPs on each chromosome were independently extracted to calculate the F*st* using GENEPOP [16]. The high-F*st* SNPs in each set were used to determine AIMs. SNPs that had strong linkage disequilibrium (LD) were removed, with a SNP window of 50 bases and sliding window of 5 SNPs. One SNP was also removed from each pair of SNPs using a variance inflation factor (VIF) value of 2, as described by Purcell [14]. To determine population differences, the average pairwise F*st* values over loci were calculated using the Arlequin software with bootstrap resampling using 1000 replications [17].

2.4. Population model studies

For the population simulation used in this study, the two most popular approaches were structured association mappings, which use model-based clustering, and principal component analysis (PCA), which uses top principal components (PCs) as covariates for stratification. PCA was analyzed by EIGENSTRAT to calculate the eigenvalues and PCs [18]. The results of the top PCs are presented in a two-dimensional scatter plot to identify clustering. The graphical results are shown using R software [19]. Admixture mapping was conducted using ADMIXTURE to estimate ancestry in unrelated individuals with 10,000 burn-in and 10,000 Markov chain Monte Carlo (MCMC) iterations [20]. The values of predefined cluster numbers (K) were analyzed to select a sensible modeling choice, and parameter standard errors were estimated using bootstrapping with 1000 replications. The suggested K was determined using the lowest cross-validation (CV) values of the number of assumptions. The results are presented in boxplot matrices that show the genetic admixture pattern across populations. Computerized operations were carried out using Information Technology Services by Yale University (http://its.yale.edu/).

3. Results

3.1. Genetic differences among geographic regions

A total of 1,389,511 SNPs were apparent from the HapMap population (809 individuals) and 560,311 SNPs from the Thai population (374 individuals). The data were merged, and 421,925 overlapping SNPs were selected to be included in this analysis. A total of 1012 SNPs with F*st* values over 0.20 were selected to generate HapMap-Thai AIMs (Supplement 1.1). A total of 124 unique HapMap-Thai AIMs that did not overlap with any previous study [9,20] were identified, except for one, rs12913832, which was presented in the SNP*for*ID 34-plex [9]. An F*st* value of 1 was observed in 11 markers (rs2224545, rs34019675, rs2810204, rs2233971, rs11744792, rs12166946, rs10282720, rs12750376, rs35726748, rs13194134 and rs8854) (Supplement 1.4). The level of genetic differentiation between populations existing with pairwise F*st* values revealed a close group within each continent (Table 1). The highest genetic distance between the Thai and the African, European and South American clusters were observed in YRI (F*st* = 0.6864, $p < 0.05$), MEX (F*st* = 0.5067, $p < 0.05$) and CEU (F*st* = 0.5447, $p < 0.05$), respectively.

Within an Asian population study, CHB, GIH, JPT and THA were grouped as an Asian-Thai cluster (655 individuals) that covered 1,446,473 SNPs. After data filtering, 463,265 overlapping SNPs were identified in a total of 632 individuals. A total of 1,506 SNPs had an F*st* value over 0.10 (Supplement 1.2). After LD removal, 89 SNPs were selected as Asian-Thai AIMs. Seven markers (rs186154, rs1447826, rs1036819, rs1975920, rs9572312, rs12884681 and rs8063779) were identified as a subset of 124 HapMap-Thai AIMs. The highest F*st* value was 0.88, observed for rs1455311 (Supplement 1.5). The pairwise F*st* values across the Asian population are presented in Table 1. A relatively high genetic distance was observed between THA and GIH (F*st* = 0.3669, $p < 0.05$), JPT (F*st* = 0.2223, $p < 0.05$) and CHB/CHD (F*st* = 0.1868, $p < 0.05$).

Using a similar reference Asian population, the Thai population was classified into northern and southern groups. There were formerly 554,292 SNPs from the southern and northern Thai population. A total of 3373 SNPs had an F*st* over 0.02 (Supplement 1.3). The highest F*st* value was 0.11, for rs12094795. The pairwise F*st* values were calculated and are presented in Table 1. The southern and northern Thai samples showed slightly different F*st* values (0.0283). The allele frequencies and heterozygosities of 273 AIMs are presented in Supplement 1.6.

3.2. Analysis of genetic structure

In evaluating the genetic structure across the studied populations from $K = 2$ to $K = 7$ using different sets of AIMs, the admixture model results of minimal K values revealed a homogenous pattern within geographic continents and closely related groups. The set of 124 AIMs had clearly differentiated clustering for the Thai population (THA) out of the East Asian and other populations at $K = 4$ (cross validation = 0.422), as shown in Fig. 1A. Separation of the South Asian population (GIH) also resulted in an admixed genetic proportion between European (CEU and TSI) and East Asian (CHB/CHD and JPT). When analysis of the Thai population was