Published in final edited form as:

Leg Med (Tokyo). 2015 July ; 17(4): 245–250. doi:10.1016/j.legalmed.2015.02.004.

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

Kornkiat Vongpaisarnsin^{1,*}, Jennifer Beth Listman², Robert T Malison^{2,3}, and Joel Gelernter^{2,3,4,5}

¹Department of Forensic Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand ² Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA ³VA Connecticut Healthcare System, West Haven Campus, West Haven, CT, USA ⁴Department of Genetics, Yale University School of Medicine, New Haven, CT, USA ⁵Department of Neurobiology, Yale University School of Medicine, New Haven, CT, USA

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.

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

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

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of interest

The authors declare that they have no conflicts of interest.

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.

Materials and methods

Population datasets

Genetic sequences of the following 11 human populations (1,301 individuals) were obtained from the international HapMap 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).

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.

Selected AIMs

The measurement of genetic differentiation was conducted using *Fst* 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 *Fst* using GENEPOP (16). The high-*Fst* 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 *Fst* values over loci were calculated using the Arlequin software with bootstrap resampling using 1,000 replications (17).

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 1,000 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/>).

Results

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 1,012 SNPs with *Fst* 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 SNPforID 34-plex (9). An *Fst* 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 3,373 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.

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 conducted with four reference Asian populations (CHB/CHD, JPT and GIH) (Figs. 1B and 1C), the set of 89 AIMs separated the Thai population from other Asian populations, but it was not possible to stratify a Thai subpopulation. An analysis of the Thai subpopulation was performed without reference populations (data not shown), and a homogenous admixture pattern of northern (THA-NOR), northeast (THA-NOE), southern (THASOU) and central (THA-CEN) Thai regions revealed the co-ancestry of the Thai population. After excluding the central and northeast groups, a clear stratification between the southern and northern Thai populations was presented with a set of 273 AIMs, including an Asian reference population, as shown in Fig. 1C. A block plot revealed the individual genetic components within the Asian group, in which CHB/CHD and GIH shared a common genetic component with both northern Thai (THA-NOR) and southern Thai (THA-SOU) populations. However, the northern and southern Thai individuals had different

common genetic proportions shared by CHB/CHD and GIH, as depicted by the different block colors (Fig. 1C).

Population divergence analyzed with principal component analysis

The PCA results are presented in plots of two-dimensional principal components (PCs), with similar AIM sets as analyzed in the genetic structure examination. The first and second PCs (PC1 = 40.06%, PC2 = 9.79%) with 124 AIMs were obviously clustered as Thai (THA), East Asian (CHB/CHD and JPT) and African groups (ASW, MKK, LWK and YRI), whereas the overlapping South Asian (GIH), South America (MEX) and European groups (CEU and TSI) had relatively low population heterogeneity (Fig. 2.1). The eigenvalue of each PC was 26.40% for PC1 and 5.93% for PC2. PCA performed within an Asian reference population using 89 AIMs markers notably demonstrated divergence between THA, GIH and East Asian (CHB, JPT) individuals with PC1 = 25.2% and PC2 = 4.6%, whereas strong homogeneity was observed in CHB and JPT individuals (Fig. 2.2). Analysis within the Thai population (excluding CE_THA and NE_THA) using 273 AIMs (PC1 = 4.9%, PC2 = 3.4%), visibly differentiated between a northern (THA-NOR) and southern (THA-SOU) Thai population (Fig. 2.3).

Discussion

The developing panel of individualized and ancestry informative SNPs was introduced in the forensic community within the past few years (8–11). In earlier studies, a 128-SNP panel was demonstrated to have the capability to cluster continentals corresponding to European, East Asian, Amerindian, African, South Asian, Mexican, and Puerto Rican (21) ancestry. A similar panel studied in large populations (4,871 individuals: included from the HapMap 3 (12), the Human Genome Diversity Project (HGDP) (22) and ALFRED (23) databases) provided the ability to distinguish between African American and Mexican Americans ancestries (24). Another study used the Eurasiaplex panel to stratify within continents and to estimate the admixture proportions of European and South Asian populations (11). Southeast Asian populations with small sample sizes were included in the analysis; nevertheless, genetic structure and diversity studies of Han Chinese, Japanese, Korean, Vietnamese and Filipino individuals identified substructures with a set of 1,500 AIMs (25). Genetic studies in Southeast Asia mainly focused on tracing part of the migration routes based on Y chromosome haplotypes and mitochondrial DNA haplogroups (26–30), but these studies did not provide strong inferences regarding the genetic population structure.

Our analysis shows that the Thai population is notably stratified into the following two groups using 273 AIMs (Fig. 2.3): northern Thai and southern Thai. Indian and Chinese genetic influences exist in mainland Southeast Asia. Upon investigating the ancestry proportion of GIH in both groups (Fig. 1C), the Indian genetic influence seemed to be more dominant in the southern Thai population compared with the northern population, which is particularly apparent in K = 3 (green color) and K = 4 (light blue color). This evidence suggests the migration of the Austroasiatic linguistic population along the coastal South and Southeast Asian territory (31–33). In addition, the diversity of genetic components to reference individuals (Fig. 1C) in K=3 for THA-NOR (blue color) and THA-SOU (red

color) apparently revealed a genetic proportion suggesting an influence of Chinese ancestry in both groups. These data are supported by a record of Chinese military history described that these tribes migrated throughout Asia for up to 5,000 years (34, 35), moving continuously southward into mainland Southeast Asia, including Thailand and its neighbors, over the past one thousand years (36). However, the northern has geographically characterized by complex mountain ranges, there has a gene flow from a northern to southern migration from southern China (Yunnan Province) to northern Thailand that consistent with the earlier study of Y chromosomes and mitochondrial SNPs of hill tribes in northern Thailand (30, 32,37). In addition to the *Fst* of 273 AIMs (Table 1), a southern Thai population has a stronger affinity with Chinese Han (*Fst*=0.0071) and Japan population (*Fst*=0.0093) than northern Thai population (*Fst*=0.0109). It could be inferred that Chinese Han may have a migratory route from eastern China (Fujian and Guangdong provinces) to islands southeast Asia and Malay Peninsula, the southernmost point of the Asian mainland, which comprised of Myanmar, Malaysia, and Southern Thailand (38,39). The close relatedness of Chinese ethnic in this group has revealed with a similar culture, religion as well as linguistic as Austronesian language family.

In conclusion, the analysis of genome-wide SNP genotypes with a few hundred AIMs apparently differentiates the Thai population from other HapMap populations. The ancestry informative SNPs (AIMs) of this study are suitable for successfully stratifying the Thai population. A total of 124 AIMs represent clusters of the geographic continents. It appears that the use of the small number of 89 AIMs could stratify the Thai population and East Asian populations. Moreover, a set of 273 AIMs could also distinguish northern from southern Thai subpopulations. This study also provides a list of AIMs (as shown in the supplemental data) that could be employed as markers of choice to develop multiplex ancestry panels. Using the current next-generation sequencing technology, it could be inferred that our AIMs might be useful in identifying an unidentified body in regions where ethnic origin matching by self-reporting is not available or unreliable for use in anthropology or forensics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study was supported by an NIH/Fogarty Drug Dependence Through the Lifespan: US– Thai training grant (D43 TW009087). The sample collection and genotyping of Thai depression samples were supported by Rachanukul Hospital and BIOTEC Thailand (<http://www4a.biotech.or.th/thaisnp2/>).

References

1. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature [Internet]. Jun 7. 2007; 447(7145):661–78. Available from: <http://dx.doi.org/10.1038/nature05911>.
2. Amorim A, Pereira L. Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. Forensic Sci Int [Internet]. May 28; 2005 150(1):17–21. Available from: <http://www.sciencedirect.com/science/article/pii/S0379073804003937>.

3. Butler JM, Coble MD, Vallone PM. STRs vs. SNPs: Thoughts on the future of forensic DNA testing. *Forensic Sci Med Pathol.* 2007; 3:200–5. [PubMed: 25869164]
4. Budowle B, Van Daal A. Forensically relevant SNP classes. *BioTechniques.* 2008:603–10. [PubMed: 18474034]
5. Baye TM. Inter-chromosomal variation in the pattern of human population genetic structure. *Hum Genomics.* 2011; 5:220–40. [PubMed: 21712187]
6. Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, Silva G, et al. European population substructure: Clustering of northern and southern populations. *PLoS Genet.* 2006; 2:1339–51.
7. Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesian K, et al. Measuring European population stratification with microarray genotype data. *Am J Hum Genet.* 2007; 80:948–56. [PubMed: 17436249]
8. Prestes PR, Mitchell RJ, Santos C, van Oorschot RAH. The SNPforID 34-plex-Its ability to infer level of admixture in individuals. *Forensic Sci Int Genet Suppl Ser.* 2013:4.
9. Amigo J, Phillips C, Lareu M, Carracedo Á. The SNPforID browser: An online tool for query and display of frequency data from the SNPforID project. *Int J Legal Med.* 2008; 122:435–40. [PubMed: 18491122]
10. Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Álvarez-Dios J, et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet.* 2007; 1(3-4):273–80. [PubMed: 19083773]
11. Phillips C, Aradas AF, Kriegel AK, Fondevila M, Bulbul O, Santos C, et al. Eurasiaplex: A forensic SNP assay for differentiating European and South Asian ancestries. *Forensic Sci Int Genet.* 2013; 7:359–66. [PubMed: 23537756]
12. Thorisson GA, Smith A V, Krishnan L, Stein LD. The International HapMap Project Web site. *Genome Res.* 2005; 15:1592–3. [PubMed: 16251469]
13. Ngamphiw C, Assawamakin A, Xu S, Shaw PJ, Yang JO, Ghang H, et al. PanSNPdb: The Pan-Asian SNP Genotyping Database. *PLoS One.* 2011:6.
14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81:559–75. [PubMed: 17701901]
15. Cockerham CC, Weir BS. Estimation of Gene Flow from F-Statistics. *Evolution (N Y)* [Internet]. 1993; 47:855–63. Available from: <http://www.jstor.org/stable/2410189><http://www.jstor.org/stable/pdfplus/2410189.pdf?acceptT C=true>.
16. Rousset F. GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. *Mol Ecol Resour.* 2008; 8:103–6. [PubMed: 21585727]
17. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour.* 2010; 10:564–7. [PubMed: 21565059]
18. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick N a, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006; 38:904–9. [PubMed: 16862161]
19. R Development Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing; Vienna Austria: 2013. p. {ISBN} 3-900051-07-0. Available from: <http://www.r-project.org/>
20. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 2009; 19:1655–64. [PubMed: 19648217]
21. Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat.* 2009; 30:69–78. [PubMed: 18683858]
22. Pemberton TJ, Sandefur CI, Jakobsson M, Rosenberg NA. Sequence determinants of human microsatellite variability. *BMC Genomics.* 2009; 10:612. [PubMed: 20015383]
23. Osier M V, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, Kidd KK. ALFRED: an allele frequency database for diverse populations and DNA polymorphisms--an update. *Nucleic Acids Res.* 2001; 29:317–9. [PubMed: 11125124]

24. Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, De La Vega FM, Kidd KK. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investig Genet.* 2011; 2(1):1.
25. Tian C, Kosoy R, Lee A, Ransom M, Belmont JW, Gregersen PK, et al. Analysis of east asia genetic substructure using genome-wide SNP arrays. *PLoS One.* 2008:3.
26. Ballinger SW, Schurr TG, Torroni A, Gan YY, Hodge JA, Hassan K, et al. Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. *Genetics.* 1992; 130:139–52. [PubMed: 1346259]
27. Jin L, Su B. Natives or immigrants: modern human origin in east Asia. *Nat Rev Genet.* 2000; 1:126–33. [PubMed: 11253652]
28. Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. *Nature* [Internet]. 1987; 325:31–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3025745>.
29. Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, et al. Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. *Am J Hum Genet.* 1999; 65:1718–24. [PubMed: 10577926]
30. Harihara S, Saitou N, Hirai M, Gojobori T, Park KS, Misawa S, et al. Mitochondrial DNA polymorphism among five Asian populations. *Am J Hum Genet.* 1988; 43:134–43. [PubMed: 2840820]
31. Blench R. Was there an Austroasiatic Presence in Island Southeast Asia prior to the Austronesian Expansion? *Bulletin of the Indo-Pacific Prehistory Association.* 2011
32. Lertrit P, Poolsuwan S, Thosarat R, Sanpachudayan T, Boonyarit H, Chinpaisal C, et al. Genetic history of Southeast Asian populations as revealed by ancient and modern human mitochondrial DNA analysis. *Am J Phys Anthropol.* 2008; 137:425–40. [PubMed: 18615504]
33. Anderson GDS. Austroasiatic Languages. *Encyclopedia of Language & Linguistics.* 2006:598–600.
34. Young, G. The hill tribes of northern Thailand[uni2009]: a socio-ethnological report. Siam Society; Bangkok: 1962.
35. Duncan, CR. Civilizing the margins[uni2009]: Southeast Asian government policies for the development of minorities. Cornell University Press; Ithaca: 2004.
36. American University. Minority groups in Thailand. Schrock, Joann L., editor. Washington Cultural Information Analysis Center; Headquarters, Dept. of the Army; DC: Washington: 1970.
37. Besaggio D, Fuselli S, Srikumool M, Kampuansai J, Castri L, Tyler-Smith C, et al. Genetic variation in Northern Thailand Hill Tribes: origins and relationships with social structure and linguistic differences. *BMC Evol Biol.* 2007; 7(Suppl 2):S12. [PubMed: 17767728]
38. Gupta A. Landforms of Southeast Asia. *The Physical Geography of Southeast Asia.* 2005:38–64.
39. Soils P, Asia S. Southeast Asia. *Science* (80–) [Internet]. 2004; 36:553–85. Available from: <http://books.google.com/books?id=uaFaDUyeCOcC>.

Fig. 1. (A, Band C) Population structure results using an admixture population model. Each vertical line represents an individual who is proportionally assigned to one of the K clusters, with the proportions represented by the relative lengths of the K in different colors. The populations are identified below the figures. Three sets using 124, 89 and 273 AIMs were analyzed for continental, sub-continental and Thai subpopulation clustering, respectively.

Fig. 2.1.

Principal component analyses of the studied population (1,183 individuals) with 124 AIMs. The different symbols indicate the samples in each population. There are four corresponding clusters: African (right upper, ASW, LWK, MKK and YRI), East Asian (left lower, CHB, CHD and JPT), Southeast Asian (left upper, THA) and combined European, South Asian and South America (right lower, CEU, TSI, GIH and MEX).

Fig. 2.2.

Two-dimensional plot using 89 AIMs of the East Asian and Thai populations. The samples are clusterings of the Thai (red, triangle), GIH (green, plus), JPT (blue, cross) and CHB (black, circle) populations.

Fig. 2.3.

PC1 and PC2 plot of 273 AIMs for the stratification of northern and southern Thai population. The northern (blue, cross) and southern (light blue, diamond) Thai samples are clustered as two separate groups, with GIH (red, triangle), JPT (green, plus) and CHB (black, circle) used as reference populations.

Table 1Pairwise *F_{st}* values across populations with different numbers of AIM sets

POPULATIONS	124 AIMs										
	CHB/CHD	JPT	GIH	THAI	ASW	LWK	YRI	MEX	MKK	CEU	TSI
CHB/CHD		+	+	+	+	+	+	+	+	+	+
JPT	0.0117		+	+	+	+	+	+	+	+	+
GIH	0.1426	0.1174		+	+	+	+	+	+	+	+
THAI	0.3977	0.4049	0.4643		+	+	+	+	+	+	+
ASW	0.4487	0.4156	0.2301	0.6388		+	+	+	+	+	+
LWK	0.4446	0.4148	0.2498	0.6343	0.0065		+	+	+	+	+
YRI	0.5389	0.5114	0.3450	0.6864	0.0309	0.0256		+	+	+	+
MEX	0.2176	0.1800	0.0419	0.5067	0.1885	0.2133	0.3084		+	+	+
MKK	0.3701	0.3378	0.1718	0.5901	0.0218	0.0221	0.0844	0.1404		+	+
CEU	0.2989	0.2576	0.0743	0.5447	0.1716	0.2026	0.2931	0.0471	0.1318		+
TSI	0.2703	0.2329	0.0583	0.5294	0.1639	0.1915	0.2831	0.0363	0.1183	0.0055	

POPULATIONS	89 AIMs			
	CHB/CHD	JPT	GIH	THAI
CHB/CHD		+	+	+
JPT	0.0140		+	+
GIH	0.1817	0.1429		+
THAI	0.1868	0.2223	0.3669	

POPULATIONS	273 AIMs				
	CHB/CHD	JPT	GIH	THA-SOU	THA-NOR
CHB/CHD		+	+	+	+
JPT	0.0001		+	+	+
GIH	0.0250	0.0256		+	+
THA-SOU	0.0071	0.0093	0.0247		+
THA-NOR	0.0109	0.0140	0.0298	0.0283	

*,** = *F_{st}* P values with a significance level < 0.05.