

Research article

Open Access

## Demographic changes and marker properties affect detection of human population differentiation

Jennifer B Listman<sup>1</sup>, Robert T Malison<sup>2,3</sup>, Atapol Sughondhabirom<sup>4</sup>, Bao-Zhu Yang<sup>3</sup>, Ryan L Raaum<sup>6</sup>, Nuntika Thavichachart<sup>4</sup>, Kittipong Sanichwankul<sup>7</sup>, Henry R Kranzler<sup>5</sup>, Sookjaroen Tangwonchai<sup>4</sup>, Apiwat Mutirangura<sup>4</sup>, Todd R Disotell<sup>1</sup> and Joel Gelernter<sup>\*2,3,8</sup>

Address: <sup>1</sup>Dept Anthropology, New York Univ, NY, USA, <sup>2</sup>Dept Psychiatry, Yale Univ Sch Medicine, New Haven, CT, USA, <sup>3</sup>VA CT, West Haven, CT, USA, <sup>4</sup>Chulalongkorn Faculty of Med, Bangkok, Thailand, <sup>5</sup>Dept Psychiatry, Univ of CT Sch Medicine, Farmington, CT, USA, <sup>6</sup>Dept Anthropology, Univ of FL, Gainesville, FL, USA, <sup>7</sup>Suan Prung Psychiatric Hospital, Chiang Mai, Thailand and <sup>8</sup>Depts Genetics and Neurobiology, Yale Univ Sch Medicine, New Haven, CT, USA

Email: Jennifer B Listman - jenny.listman@nyu.edu; Robert T Malison - robert.malison@yale.edu; Atapol Sughondhabirom - atapol.s@gmail.com; Bao-Zhu Yang - bao-zhu.yang@yale.edu; Ryan L Raaum - raaum@anthro.ufl.edu; Nuntika Thavichachart - fmednta@md.chula.ac.th; Kittipong Sanichwankul - kittipong@suanprung.go.th; Henry R Kranzler - kranzler@psychiatry.uchc.edu; Sookjaroen Tangwonchai - chstw@hotmail.com; Apiwat Mutirangura - mapiwat@chula.ac.th; Todd R Disotell - todd.disotell@nyu.edu; Joel Gelernter\* - joel.gelernter@yale.edu

\* Corresponding author

Published: 11 May 2007

Received: 16 October 2006

BMC Genetics 2007, 8:21 doi:10.1186/1471-2156-8-21

Accepted: 11 May 2007

This article is available from: <http://www.biomedcentral.com/1471-2156/8/21>

© 2007 Listman et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Differentiating genetically between populations is valuable for admixture and population stratification detection and in understanding population history. This is easy to achieve for major continental populations, but not for closely related populations. It has been claimed that a large marker panel is necessary to reliably distinguish populations within a continent. We investigated whether empirical genetic differentiation could be accomplished efficiently among three Asian populations (Hmong, Thai, and Chinese) using a small set of highly variable markers (15 tetranucleotide and 17 dinucleotide repeats).

**Results:** Hmong could be differentiated from Thai and Chinese based on multi-locus genotypes, but Thai and Chinese were indistinguishable from each other. We found significant evidence for a recent population bottleneck followed by expansion in the Hmong that was not present in the Thai or Chinese. Tetranucleotide repeats were less useful than dinucleotide repeat markers in distinguishing between major continental populations (Asian, European, and African) while both successfully distinguished Hmong from Thai and Chinese.

**Conclusion:** Demographic history contributes significantly to robust detection of intracontinental population structure. Populations having experienced a rapid size reduction may be reliably distinguished as a result of a genetic drift -driven redistribution of population allele frequencies. Tetranucleotide markers, which differ from dinucleotide markers in mutation mechanism and rate, are similar in information content to dinucleotide markers in this situation. These factors should be considered when identifying populations suitable for gene mapping studies and when interpreting interpopulation relationships based on microsatellite markers.

## Background

Genetic characterization and differentiation of populations are often necessary for the conduct of valid case-control association studies [1-5], determining the role of ancestry in phenotypic differences [6,7], assigning population groups for valid linkage analysis [8], examining the distribution of neutral genetic variation among populations, and inferring migration histories [9-11]. Such differentiation has been accomplished with relative ease between major continental populations [10,12-15], but it has been asserted that population differentiation *within* a continent may not be possible; and when it appears to be so, may actually be an artifact of study design [16].

The ubiquity and frequently highly variable nature of short tandem repeat polymorphisms (STRs or microsatellites) have made them desirable markers for measuring population stratification. Commercially available marker sets such as those used for forensic purposes make STR genotyping cost effective, eliminating the time and effort required to develop multiplex marker panels. Panels developed for forensic purposes are designed to identify or exclude an individual as a match for another sample and were compiled, in part, for their high levels of variation in many populations [17]. Such panels have been adopted for non-forensic purposes such as inference of population phylogenies [18-21] and quantification of levels of population differentiation [1,2,5,8].

Homoplasy, as applied to STRs, refers to the situation where alleles of the same length have arisen from different mutation events, such that alleles identical-by-state are not necessarily identical-by-descent. Simulations of STR evolution using the stepwise mutation model (SMM) have indicated that homoplasy in STR genotypes may cause individuals or populations to appear to be more genetically similar than they really are. Point mutations, insertion or deletion events (indels), or complex repeat motifs can generate additional forms of size homoplasy that are sometimes revealed by sequencing but are not detectable through size fractionation (electrophoresis) [22-24]. These forms of homoplasious alleles have been observed in a number of the tetranucleotide repeats that are standard in forensic panels (some of which are included among the markers used in this study; see below) [25].

However, it has also been shown that even in the presence of homoplasy, multi-locus genotypes (the combined genotypes from multiple loci) of highly variable STR markers are effective in assigning individuals to known or unknown populations [26-33]. Again, this has typically been true for large continental populations. Population differentiation *within* a continent has been successful, but only with large numbers of markers when applied to pop-

ulation isolates [10]. Here, we used a small set of markers and, in contrast to the majority of past studies, addressed the properties of the markers used.

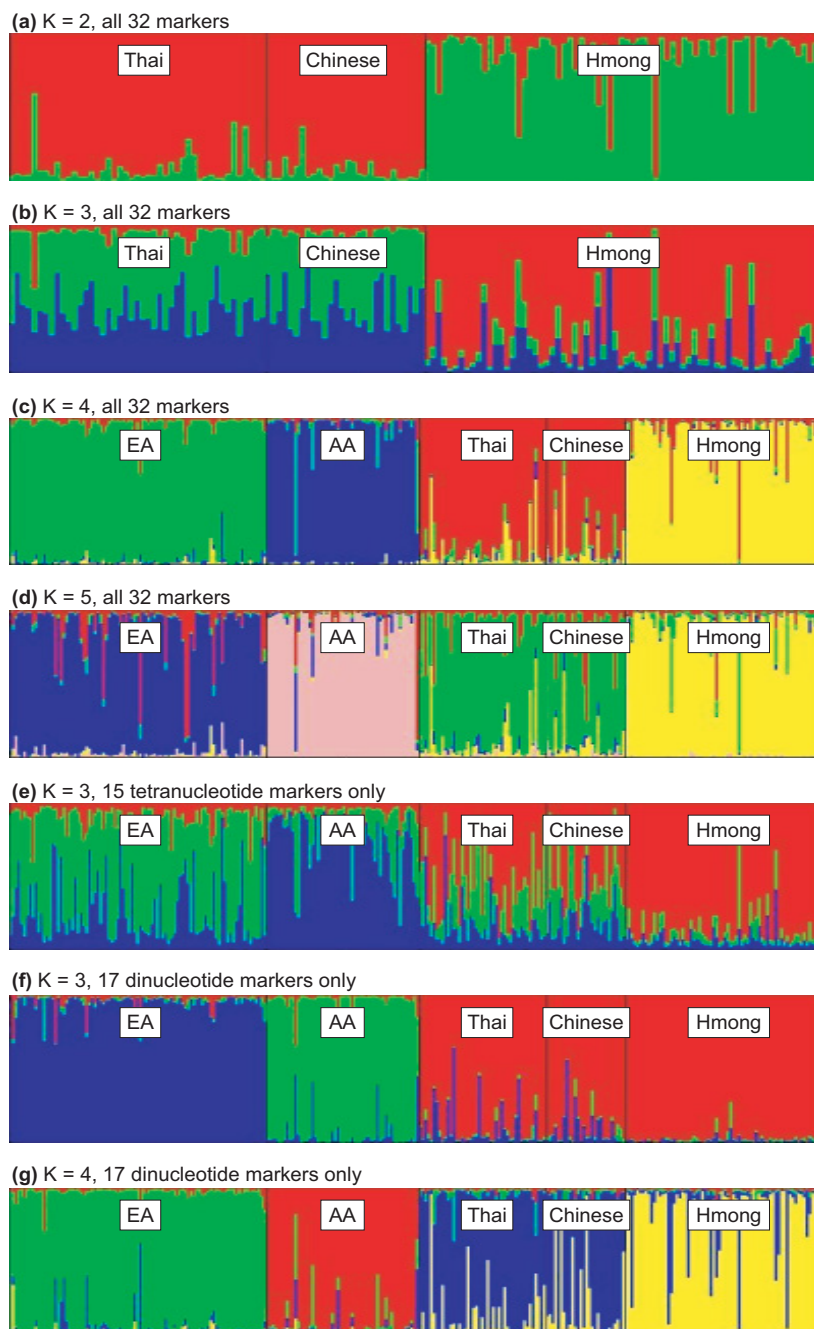
Conditions such as small population size or recent founding of a population may enable statistical differentiation using a small panel of highly variable markers, due to increased effects of genetic drift and decreased incidence of homoplasy. To evaluate this possibility, we investigated whether empirical genetic differentiation could be accomplished efficiently among three closely related Asian populations (Hmong, Thai, and Chinese) using a small set of STRs that includes both tetranucleotide and dinucleotide markers. In addition, we studied the relative information content of tetranucleotide versus dinucleotide markers for discriminating among these three Asian populations, as well as European Americans (EA) and African Americans (AA). We then evaluated the populations for evidence of recent changes in effective population size.

## Results

### Population differentiation

The program STRUCTURE 2.1 [32,33] uses Bayesian clustering of multilocus genotypes to assign individuals to populations, estimate admixture proportions for individuals, and infer the number of parental populations ( $K$ ) for a sample. For STRUCTURE runs which included the three East Asian populations only and all 32 markers, the Hmong were allocated into a cluster distinct from a single Thai/Chinese cluster with 86.0% estimated ancestry for  $K = 2$  with a posterior probability ( $\Pr(K = 2)$ ) of 1, indicating  $K$  with the best fit for the data (Figure 1a). Separate Thai and Chinese clusters were not inferred with  $K = 3$  and  $\Pr(K = 3)$  was effectively zero ( $3.3 \times 10^{-156}$ ) (Figure 1b). When the three East Asian populations were analyzed with EA and AA samples, the Hmong were then allocated to a separate cluster with an average of 90.0% estimated ancestry when  $K = 4$  and  $\Pr(K = 4) = 1$  (Figure 1c). Under these same conditions, the Thai and Chinese were assigned together to a single cluster with 86.5% and 84.2% estimated ancestry, respectively. When  $K$  was increased to 5, the Thai and Chinese populations continued to form a single cluster (Figure 1d) and  $\Pr(K = 5)$  was  $1.7 \times 10^{-48}$ . For  $K = 2$ ,  $K = 3$ , or  $K = 6$ ,  $\Pr(K)$  was similarly effectively zero (barplots not shown).

When the markers were separated by repeat size the extent of successful population assignment differed greatly between the two panels; for STRUCTURE runs using the 15 tetranucleotide markers alone  $\Pr(K = 3)$  was 1, while for the 17 dinucleotide markers alone  $\Pr(K = 4)$  was 1. In addition, although  $K = 3$  had the best fit for the data for tetranucleotide markers, assignment of individuals to major continental populations was not robust (Figure 1e) (EA 63.2%, AA 75.3%, Thai 54.1%, Chinese 48.7%,



### Figure 1

Hmong populations are consistently and reliably distinguished from all others in STRUCTURE analyses using a small number of either dinucleotide or tetranucleotide markers while for other populations successful assignment of individuals varies with marker type. In (a) and (b) Chinese, Hmong, and Thai samples were assigned by STRUCTURE to 2 or 3 populations respectively. In (c) and (d) European-American (EA), African-American (AA), Chinese, Hmong and Thai samples were assigned to 4 or 5 populations respectively. Finally, tetranucleotide markers (e) are less useful for differentiating among EA, AA, Thai and Chinese populations than dinucleotide markers (f & g). These plots were produced using the STRUCTURE software; each individual is represented by a vertical line depicting the estimated percent assignment of the individual into K assumed populations. Each assumed population is represented by one color. Vertical black lines separate individuals by self-reported ancestral population.

Hmong 86.1%). Dinucleotide markers alone resulted in higher assignment rates than those of the tetranucleotide markers when  $K = 3$  (Figure 1f) (EA 94.0%, AA 91.4%, Thai 88.3%, Chinese 88.4%, Hmong 97.3%) or under the best fit for the data,  $K = 4$  (Figure 1g) (EA 91.4%, AA 90.4%, Thai 81.0%, Chinese 73.6%, Hmong 82.8%).

Out of concern that each Hmong village in which samples were collected could consist of its own apparent cluster due to close relatedness within each village, the villages were analyzed initially as separate populations in STRUCTURE under the same conditions as all STRUCTURE runs reported here. In all cases, the two source villages formed one cluster and the average assignment values for all population samples, including Hmong, were no different than assignment values reported here when the two villages were combined and assumed to be one population (data not shown).

#### Effective population size

The Hmong sample was found to have a heterozygosity deficiency ( $p = 0.004$ ), based on a sign test in BOTTLENECK [34], indicating a possible recent population expansion. Given the number of observed alleles, if the Hmong population was at equilibrium heterozygosity is expected to be higher than that which is observed. All other samples had neither excess nor deficiency for this measure.

#### Relatedness

Based on maximum-likelihood estimates of pair-wise relationships, potential parent-offspring pairs and sibling pairs were discovered in the Chinese and Hmong samples. In each case, one individual was then deleted from the sample and excluded from all other analyses.

#### Hardy Weinberg Equilibrium (HWE)

No population showed significant deviation from HWE over all loci (EA  $p = 0.07$ , AA  $p = 0.82$ , Chinese  $p = 0.57$ , Thai  $p = 0.87$ , Hmong  $p = 0.99$ ) (Table 1). If a Bonferroni correction is applied to correct for multiple testing, (requiring a  $p$  value of  $0.05/32 = 0.00156$  for significance) none of these  $p$ -values for individual loci are significant (Additional File 1).

#### Heterozygosity

The mean observed heterozygosity ( $H_o$ ) (Table 2) for all loci was not statistically different for any of the Asian population pairs, based on paired two-sample  $t$ -test (Chinese/Hmong  $p = 0.07$ , Thai/Chinese  $p = 0.34$ , Thai/Hmong  $p = 0.27$ ). With the exception of EA/Chinese, mean observed heterozygosity was significantly different for all other population pairs (EA/AA  $p = 0.01$ , EA/Thai  $p = 0.03$ , EA/Chinese  $p = 0.37$ , EA/Hmong  $p < 0.01$ , AA/Hmong  $p < 0.01$ , AA/Thai  $p < 0.01$ , AA/Chinese  $p = 0.04$ ).

Mean tetranucleotide  $H_o$  was not significantly different from mean dinucleotide  $H_o$  for any population other than Chinese based on a two-sample  $t$ -test (Table 3) (AA  $p = 0.49$ , EA  $p = 0.17$ , Thai  $p = 0.30$ , Chinese  $p = 0.05$ , Hmong  $p = 0.57$ ).

#### Marker information content

The mean Hmong/Thai and Hmong/Chinese  $\delta$  values are nearly equivalent, and the Hmong were similarly differentiated from these two populations (delta values for each locus and mean delta values for all loci and by repeat size are reported for each population pair in Additional File 2). The low mean Chinese/Thai  $\delta$  appears to explain the inability of this marker panel to assign the Thai and Chinese to separate clusters. Overall, the dinucleotide markers provide more information than the tetranucleotide markers, but this difference is not as great for population pairs that include the Hmong; for the Hmong/Chinese and Hmong/Thai population pairs, the difference in the average dinucleotide  $\delta$  and the average tetranucleotide  $\delta$  is negligible (Figure 2).

#### Discussion

In this study, we successfully differentiated between closely related populations using a marker set much smaller than that previously suggested to be minimally necessary for such studies. We used a set of highly polymorphic microsatellite markers of which some were specifically selected for high  $\delta$  between EA, AA, and Asian populations [15], however, the value of this marker set for differentiating populations within Asia was previously unknown. To explain our results, we investigated the evolutionary histories of the samples, and found evidence for changes in  $N_e$  for the Hmong population, based on an excess of rare alleles. This tribal population has a recent history of repeated fractioning and migration throughout Southeast Asia as well as loss of numbers due to military conflict, which is consistent with our results [35]. Further suggestive evidence of a recent Hmong population bottleneck followed by expansion can be found in the delta values of tetranucleotide markers compared to that of dinucleotide markers. Delta measures absolute values of allele frequency differences which can arise over time via accumulated mutations or through deviations from neutral conditions such as drift caused by a bottleneck.

If time since divergence determines differences in allele frequencies delta should be correlated with time since divergence. The tetranucleotide markers consistently provide as much information for the Hmong as the dinucleotide markers provide, while this is not so for any other population. We propose that this suggests forces other than mutation as measured by divergence time contributing to differences in population allele frequencies between Hmong and other populations examined here.

**Table 1: Results from Fisher's test for deviation from HWE for all 32 loci combined**

	EA	AA	Thai	Chinese	Hmong
chi square	81.3	53.6	51.5	61.3	<b>41.6</b>
df	64	64	64	64	64
p-value	0.07	0.82	0.87	0.57	0.99

The effects of this can be seen in the differences in clustering behavior using STRUCTURE when either tetranucleotide or dinucleotide genotypes are analyzed alone – specifically, the difficulty in using the tetranucleotide panel to assign individuals to major continental groups for all populations, with the exception of the Hmong. Rosenberg et al [30] found dinucleotide markers to be more informative than tetranucleotide markers for population assignment in a larger study based on a different measure of marker informativeness. In their study, populations from the Americas or Oceania were exceptions to this pattern. The authors proposed genetic drift during founding events as one explanation for their results.

Mean tetranucleotide  $H_o$  was not statistically different from mean dinucleotide  $H_o$  for any population other than Chinese. Although these markers suggest high intrapopulation variation for all populations in this study, this does not provide information on differences in the sources of that variation either within or between populations for each type of marker.

The decrease in difference between  $\delta$  values between the two subsets of markers for population combinations that include the Hmong indicates that genetic drift (random changes in allele frequencies from one generation to the next that are more likely to affect small populations) rather than mutation has been a major force of evolution contributing to observed allele frequencies in this population. Differences in marker information content between the tetranucleotide and dinucleotide panels for all other populations in this study indicate that mutation rate and mechanism have shaped allele frequency distributions in these populations more than genetic drift, as would be expected for large populations at mutation-drift equilibrium.

The dinucleotide markers were previously selected for differentiation between European and African populations and high variation [15] and the tetranucleotide markers were chosen for forensic purposes for their high rates of variation in multiple populations.

Total sample size, unequal sample size between populations, and number of markers can affect the stability of clustering in STRUCTURE [36,37]. We cannot exclude biases introduced through these study design elements influencing our observations, however, stable clustering patterns were inferred in this case by repeated STRUCTURE runs. Although increasing the number of markers or population sample size can strengthen clustering patterns where clustering exists, the number of individuals in a sample or the minimum number of markers necessary to differentiate between all populations is dependant on the evolutionary histories of the population samples. Sample sizes similar to ours have been demonstrated previously to be generally sufficient for stable and accurate clustering [36].

Some of the tetranucleotide markers in this study have been shown to consist of complex repeats including more than one repeat motif, as well as insertions or deletions of partial repeats [17] (structure of observed alleles and their amplicon sizes can be found for tetranucleotide repeats typically used for forensic purposes and in this study at [25]). These factors, as well as historically large effective population size such as those of the EA and AA populations, increase the likelihood of size homoplasy. We hypothesize that the accumulation of homoplasious alleles of tetranucleotide loci may contribute to their lower information content when compared to that of the dinucleotide markers in populations other than Hmong.

**Table 2: Mean Nei's gene diversity ( $H_z$ ) and mean observed heterozygosity ( $H_o$ ) for all markers for each population**

	N	$H_z$	$H_z$ SD	$H_o$	$H_o$ SD
EA	91	0.77	0.01	0.76	0.01
AA	54	0.81	0.01	0.81	0.01
Thai	45	0.74	0.02	0.72	0.01
Chinese	28	0.75	0.03	0.74	0.01
Hmong	70	0.71	0.02	0.69	0.01

**Table 3: Mean observed heterozygosity ( $H_o$ ) for each marker type for each population**

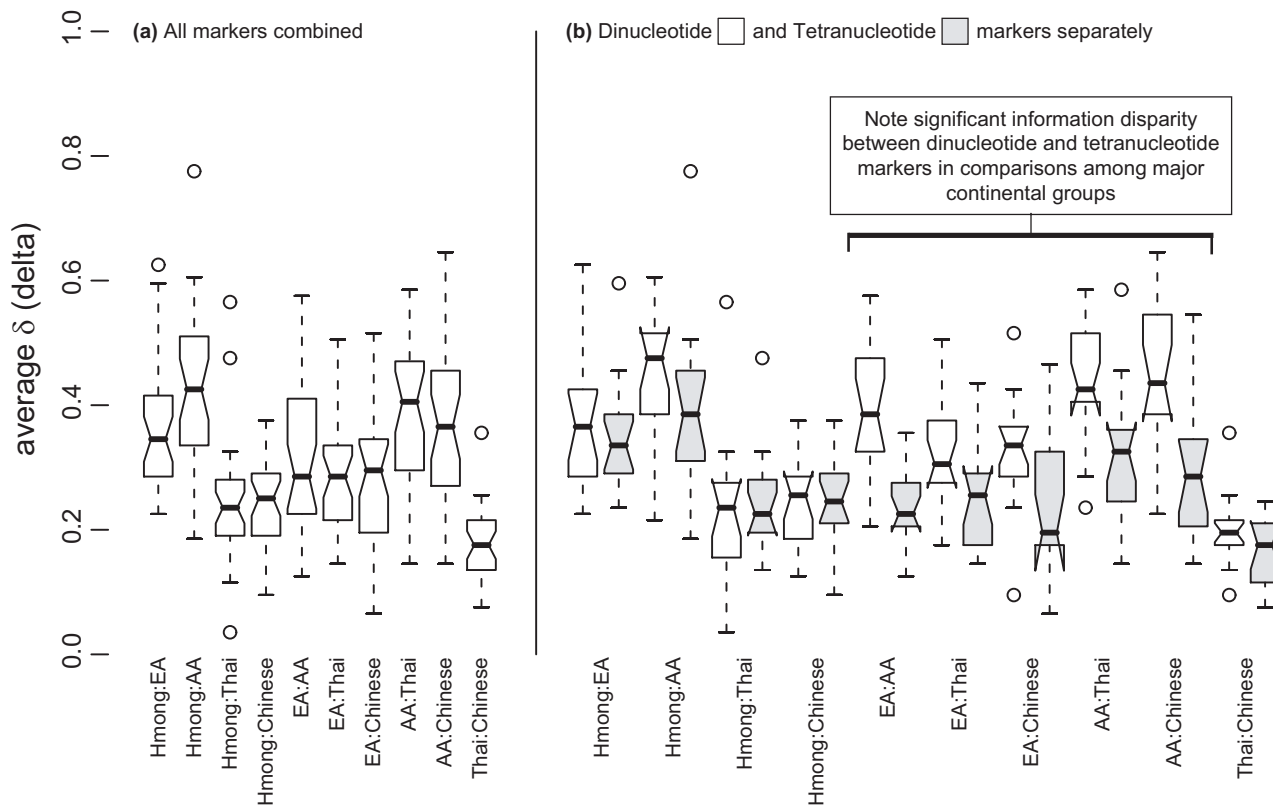
	EA	AA	Thai	Chinese	Hmong
$H_o$ tetranucleotide	0.78	0.80	0.74	0.80	0.71
$H_o$ dinucleotide	0.75	0.82	0.69	0.68	0.68

**Conclusion**

When STR loci are used either to detect association, linkage, or population substructure, population history and marker choice both affect study results. Demographic history and marker properties are often overlooked when determining population or marker suitability for gene mapping studies (i.e. to identify variants that affect traits), but have bearing on the efficiency and feasibility of such studies.

The three Asian populations in this study have mean  $H_o$  values which are not significantly different from each other. It is likely that genetic drift, in conjunction with long-standing endogamy, allow the Hmong to be statistically differentiated from the Thai and Chinese using mul-

tilocus genotypes, despite the high level of within-population variation of the Hmong. Potential homoplasy in populations at equilibrium warn against the use of STRs (particularly those with known homoplasious alleles) for phylogenetic analysis or linkage or association mapping purposes in large populations, other than quantifying population stratification. Since the tetranucleotide repeats used in commercially available kits designed for forensic purposes have been well-characterized and have been demonstrated to contain many instances of size homoplasy, these markers should not be relied upon for phylogenetic analyses. Risks of homoplasy interfering with association or linkage analysis, in which identical-by-state is often assumed to mean identical-by-descent,



**Figure 2**

The left side of the figure shows the average delta for all 32 loci for each population pair and the right side of the figure shows the average delta for 15 tetranucleotide (shaded) and 17 dinucleotide (unshaded) markers, separately, for each population pair.

should be less of a concern in populations that have undergone recent bottlenecks.

In addition, a history of migrations or bottlenecks in an isolated population is expected to initially reduce levels of intrapopulation variation, and increase interpopulation differentiation [38]. Linkage disequilibrium (LD) will be higher in such populations [39,40]. A subsequent population expansion will recover allelic variation faster than LD will degrade for a given genomic region. Low intrapopulation variation, the corresponding increase in interpopulation variation, and higher LD, have been identified as desirable characteristics in a population for gene mapping and admixture detection [41,42], making geographically or culturally isolated populations with a history of bottlenecks potentially more valuable for gene mapping than populations whose size has remained stable and large. In addition to reducing genetic heterogeneity for the trait of interest, use of such populations also could reduce the costs of association mapping studies because the number of subjects needed for a specified power level is inversely related to the population's level of LD [43]. It would therefore be useful to identify such populations prior to designing a study.

Existence of a large number of rare alleles within a population can be the result of a bottleneck followed by expansion [44-46]. It is possible to infer these events from a significant heterozygosity deficiency in the Hmong sample based on results from BOTTLENECK. The data from the EA, AA, Chinese and Thai samples do not violate the assumptions of mutation-drift equilibrium. These data suggest that the populations in this study have been large and at equilibrium for a relatively long period, with the exception of the Hmong population (as represented by the individuals we sampled). Unknown migrants or recent admixture can introduce new alleles into a population at initial low frequencies, mimicking the pattern caused by population expansion. Such recent admixture is not a likely explanation of the data in this case in light of successful clustering of the Hmong sample when analyzed with samples from the two populations most likely to contribute to hypothetical admixture: Thai and Chinese.

STRUCTURE, and other clustering algorithms, detect admixture and quantify population differentiation through differences in population allele frequencies. These differences which allow for successful clustering arise through various evolutionary forces and are shaped by ascertainment processes which must also be considered when identifying populations suitable for gene mapping studies or interpreting estimates of inter or intrapopulation genetic distance. The ability to differentiate between East Asian populations that have diverged recently relative to major continental populations indi-

cates that it may also be possible to use more easily-accessible closely related populations, such as European, for admixture mapping if marker choice and population history are taken into account.

## Methods

### Populations and sampling

The Asian populations in this study were collected as part of an ongoing gene mapping study. Samples of self-identified Thai (N = 45) and Chinese (N = 29) were obtained from a blood drive in Bangkok, Thailand. The Thai and Chinese samples used in this study were selected to include only subjects for whom all four grandparents were reported to have the same self-identified ethnicity as the subject. The Hmong, a Miao-Yao-speaking group of the Austro-Thai language family, are an endogamous tribal population with an estimated total population throughout China, Laos, Vietnam, and Burma of eight million, approximately 120,000 of whom reside in Thailand. Chinese written history documents the presence of Hmong in Central China at least 2,300 years ago and their migration to Southern China several hundred years later. Migrations farther south have occurred since the seventeenth century [47]. Hmong refugees fleeing military conflict in Laos have periodically been resettled since 1975 in the U.S., France, and Australia. Hmong samples (N = 103) were obtained in two Hmong villages in northern Thailand. Data on grandparents' reported ethnic affiliation were not available for the Hmong subjects. The dataset also included samples of unrelated African Americans (AA, N = 54) and European Americans (EA, N = 91), a subset of a sample described elsewhere [15]. Both EA and AA samples were self-identified as such, and these identifications were previously confirmed via Bayesian marker clustering [15]. After immediate relatives were discovered and excluded from analysis (see below), sample sizes were reduced as follows: Hmong (N = 70) and Chinese (N = 28). No close relative pairs were found within the remaining three population samples. All subjects provided informed consent as approved by the appropriate institutional review boards.

### Markers and genotyping

For the three East Asian populations, DNA was extracted directly from blood using PaxGene materials and the manufacturer's specified protocol (Qiagen, Valencia CA, USA) (Hmong) or standard phenol/chloroform methods (Thai and Chinese). All samples were genotyped for thirty-two unlinked autosomal STR markers. The panel is comprised of the 15 tetranucleotides in the AmpF/STR Identifiler PCR Amplification kit (PE Applied Biosystems, Foster City, CA, USA) (D8S1179 [GenBank:AX412206], D21S11 [GenBank:AJ550387], D7S820 [GenBank:NC\_000007], CSF1PO [GenBank:AF076965], D3S1358 [UniSTS:148226], TH01 [UniSTS:240639],

D13S317 [GenBank:G09017], D16S539 [GenBank:AF249681], D2S1338 [GenBank:G08202], D19S433 [GenBank:G08036], vWA [UniSTS:240641], TPOX [GenBank:M25706], D18S51 [GenBank:L18333], D5S818 [GenBank:G08446] and FGA [GenBank:G3347] and an additional 17 dinucleotide repeats (D17S799 [GenBank:Z16830], D8S272 [GenBank:Z17250], D7S640 [GenBank:Z23671], D8S1827 [GenBank:Z50970], D22S274 [GenBank:Z16730], D5S407 [GenBank:Z16723], D2S162 [GenBank:Z17035], D10S197 [GenBank:Z16611], D11S935 [GenBank:Z17148], D9S175 [GenBank:Z17021], D5S410 [GenBank:Z16825], D7S2469 [GenBank:Z53000], D16S3017 [GenBank:Z52036], D10S1786 [GenBank:Z51854], D15S1002 [GenBank:Z53242], D6S1610 [GenBank:Z53131], and D1S2628 [GenBank:Z52173]). The amelogenin locus, included in the AmpF/STR Identifiler PCR Amplification kit for sex identification, was not included in any analyses. All STR markers were analyzed on an ABI PRISM 3100 semiautomated capillary fluorescence sequencer. Data were scored using Genemapper (ABI). We have previously used this marker panel to determine and statistically correct for ancestry in case-control studies and genome-wide linkage studies [1,2,5,8].

### Statistical analyses

#### Population differentiation

Because variance of STRUCTURE results increases with small sample sizes [15], each run was repeated five times. However, results did not vary notably for each of the five runs given a set of conditions. For analysis of the three East Asian populations alone, the parameters used were  $K = 2$  and  $K = 3$ , 50,000 burn-in and 50,000 Markov chain Monte Carlo (MCMC) iterations. For analysis of all five populations in this study, the parameters used were  $K = 2$ ,  $K = 3$ ,  $K = 4$ , and  $K = 5$ , with 50,000 burn-in and 50,000 MCMC iterations. These STRUCTURE runs were each carried out with all 32 markers and then with the 15 tetranucleotide markers and the 17 dinucleotide markers separated into two marker panels. The posterior probability for each value of "K" was calculated to determine the "K" that best fit the data for each set of populations and markers. The self-reported population of origin was not used as additional data by STRUCTURE and the presence of admixture was assumed.

#### Effective population size

The program BOTTLENECK evaluates populations for evidence of a recent rapid change in effective population size, according to differences between Nei's gene diversity, or unbiased expected heterozygosity ( $H_z$ ) based on observed allele frequencies versus expected equilibrium gene diversity ( $H_{eq}$ ), simulated based upon an assumed mutation model, number of alleles, and number of gene copies (2N for a diploid system) for each locus. Based on simulations

of a coalescent process in which observed alleles at a locus are traced back to a hypothetical common ancestral allele, BOTTLENECK predicts present-day allele frequencies assuming constant population size. This results in a  $H_{eq}$  value for the present-day population. Significant deviations from this predicted value are used to infer drastic changes in effective population size which have occurred in the recent past. A significant heterozygosity excess ( $H_z > H_{eq}$ ) indicates a possible bottleneck while a significant deficit ( $H_z < H_{eq}$ ) indicates a possible expansion. Observed heterozygosity is the percentage of heterozygous individuals in a sample for a locus and is based on observed genotypes while ( $H_z$ ) is the probability that two alleles chosen at random from the population sample will not be identical, correcting for sample size, and is an indirect measure of the extent to which allele frequencies for a locus are evenly distributed.

Significance of deviations from  $H_{eq}$  was tested under the two-phased model of mutation (TPM) which assumes that the majority of mutations are single step mutations, as in the stepwise mutation model but allows for some multi-step mutations, which are more likely to be observed in dinucleotide repeats and may be a more accurate model for microsatellite mutation than the SMM [48]. BOTTLENECK allows the user to specify the percent of multi-step mutations assumed and the variance of allele size for the mutation model. BOTTLENECK authors suggest a percent of multi-step mutations between 5 and 10. It has been shown that incidence of type I error (detecting a bottleneck when a population has been at equilibrium) for the algorithm used in BOTTLENECK increases when assumed parameters are overestimated [49]. Therefore, based on detection of predicted bottlenecks for the AA, Chinese and EA populations when larger values were used, variance was set conservatively at 20 and percent of multi-step mutations was set at 5. The number of iterations of the simulated coalescent process under the TPM was 1000.

#### Relatedness

We used marker genotypes to identify, and then exclude from the analysis sample, closely related subjects who may not have identified themselves as such. The admixture model in STRUCTURE assumes HWE and linkage equilibrium within subpopulations; the use of close relatives within a sample would violate those assumptions and possibly result in false cluster detection [33]. Similarly, BOTTLENECK software assumes no close relatives in a population sample [34]. Although potential subjects may be instructed that multiple family members should not participate, cultural differences in kin definitions, lack of understanding of instructions, or financial compensation of subjects may result in individuals disregarding such instructions. Maximum likelihood estimates of pair-



wise relationships (parent-offspring, full sib, half-sib, or unrelated) were produced using the program ML-Relate [50] for all possible pairs within each population. ML-Relate does not require pedigree information and therefore can be applied to a large anonymous sample.

#### Hardy Weinberg Equilibrium (HWE)

Tests for deviation from Hardy-Weinberg equilibrium were conducted for each locus within each population using the exact test for HWE based on a Markov chain method implemented in the web-based version of GENE-POP [51]. The parameters used were 5000 dememorizations, 500 batches, and 5000 iterations per batch. The parameter values were increased from defaults until the observed standard error for p-values was less than 0.01.

#### Heterozygosity

Allele frequencies, observed heterozygosity ( $H_o$ ) values, and Nei's gene diversity ( $H_z$ ) for each locus were calculated using MStools [52]. For a diploid system,  $H_z$  is calculated as  $H_z = 2N(1 - \sum p_i^2) / 2N - 1$ , where  $N$  is the number of individuals sampled, and  $p_i$  is the frequency of the  $i$ th allele [53].

#### Marker information content

Markers were evaluated for delta ( $\delta$ ) [54], a measure of marker information content, reflecting the ability of a marker to statistically differentiate between populations. To arrive at  $\delta$ , the absolute values of allelewise frequency differences between two populations are added and this

sum is divided in half,  $\delta = \frac{1}{2} \sum_{i=1}^L |p_i^A - p_i^B|$  where  $p_i^A$  and

$p_i^B$  are the allele frequencies for the  $i$ th allele in population A and B. The more effective the marker is at differentiating between populations, the higher the value for  $\delta$  [15]. In comparison to  $F_{ST}$ , the measure  $\delta$  is easily calculated and independent of mutation model assumptions.

#### Authors' contributions

JBL designed the study, carried out statistical analyses and drafted the manuscript. BZY participated in the design and execution of statistical analyses. AS, NT, ST, AM, and KS participated in sample collection in Thailand. HRK carried out sample collection in the United States. RLR and TRD assisted in the writing of the manuscript and TRD is JBL's advisor. RTM participated in project coordination and sample collection, and the writing of the manuscript. JG participated in study design and supervision, project coordination, sample collection, and the writing of the manuscript. All authors read and approved the final manuscript.

## Additional material

### Additional file 1

Results from Fisher's test for deviation from HWE for each marker. The data provided represent, for each marker, for each population, the probability of the observed sample given the conditions for HWE are met and the standard error of the probability.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2156-8-21-S1.doc]

### Additional file 2

$\delta$  (delta) for each marker, for each population combination. The data provided show, for each marker, the values for  $\delta$  (delta), a measure of marker informativeness between each pair of populations.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2156-8-21-S2.doc]

## Acknowledgements

Greg Kay and Ann Marie Lacobelle provided excellent technical assistance. Papitchaya Sangoahangkul, Bussakorn Somboon, and Annop Thongkam provided valuable assistance in collection of the Hmong sample. This work was supported in part by NIH grants R01-DA12849, R01-DA12690, K24-DA15105, K24-AA13736, and M01-RR06192 (University of Connecticut General Clinical Research Center) and the NIH/NIDA/Fogarty International Center (Thai-US Drug Dependence Genetics Research Training Grant D43-TWO6166). AM was supported in part by the Thailand Research Fund. JBL is supported in part by an NYU Henry M. MacCracken graduate fellowship, NIH/NIDA Ruth L. Kirschstein National Research Service Award for Individual Predoctoral Fellows: grant FDA019761A, NSF Doctoral Dissertation Improvement Award 0622348, and a Wenner Gren Foundation for Anthropological Research Dissertation Fieldwork Award.

## References

- Stein MB, Schork MJ, Gelernter J: **A polymorphism of the  $\beta$ 1-adrenergic receptor is associated with low extraversion.** *Biological Psychiatry* 2004, **56**:217-224.
- Kaufman J, Yang B-Z, Douglas-Palumberi H, Houshyar S, Lipschitz D, Krystal JH, Gelernter J: **Social supports and serotonin transporter gene moderate depression in maltreated children.** *Proc Natl Acad Sci USA* 2004, **101**:17316-17321.
- Green EK, Raybould R, Macgregor S, Gordon-Smith K, Heron J, Hyde S, Grozeva D, Hamshe M, Williams N, Owen MJ, O'Donovan MC, Jones L, Jones I, Kirov G, Craddock N: **Operation of the schizophrenia susceptibility gene, neuregulin 1, across traditional diagnostic boundaries to increase risk for bipolar disorder.** *Arch Gen Psychiatry* 2005, **62**:642-648.
- Larson GP, Ding Y, Cheng LS, Lundberg C, Gagalang V, Rivas G, Geller L, Weitzel J, MacDonald D, Archambeau J, Slater J, Neuberg D, Daly MB, Angel I, Benson AB 3rd, Smith K, Kirkwood JM, O'Dwyer PJ, Raskay B, Sutphen R, Drew R, Stewart JA, Werndl J, Johnson D, Ruckdeschel JC, Elston RC, Krontiris TG: **Genetic linkage of prostate cancer risk to the chromosome 3 region bearing FHIT.** *Cancer Res* 2005, **65**:805-814.
- Luo X, Kranzler HR, Zuo L, Wang S, Blumberg HP, Gelernter J: **CHRM2 gene predisposes to alcohol dependence, drug dependence and affective disorders: results from an extended case-control structured association study.** *Hum Mol Genet* 2005, **14**:2421-2434.
- Parra FC, Amado RC, Lambertucci JR, Rocha J, Antunes CM, Pena SDJ: **Color and genomic ancestry in Brazilians.** *Proc Natl Acad Sci USA* 2003, **100**:177-182.
- Brutsaert TD, Parra EJ, Shriver MD, Gamboa A, Rivera-Ch M, Leon-Velarde F: **Ancestry explains the blunted ventilatory response to sustained hypoxia and lower exercise ventilation of Que-**

- chua altitude natives. *Am J Physiol Regulatory Integrative Comp Physiol* 2005, **289**:R225-R234.
8. Gelernter J, Panhuysen C, Weiss R, Brady K, Hesselbrock V, Rounsaville B, Poling J, Wilcox M, Farrer L, Kranzler HR: **Genomewide linkage scan for cocaine dependence and related traits: significant linkages for a cocaine-related trait and cocaine-induced paranoia.** *Am J Med Genet B Neuropsychiatr Genet* 2005, **136**:45-52.
  9. Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G: **Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms.** *Genome Res* 2002, **12**:602-612.
  10. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW: **Genetic structure of human populations.** *Science* 2002, **298**:2381-2385.
  11. Fejerman L, Carnese FR, Goicoechea AS, Avena SA, Dejean CB, Ward RH: **African ancestry of the population of Buenos Aires.** *Am J Physical Anthropology* 2005, **128**:164-170.
  12. Shriver MD, Smith MV, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE: **Ethnic-affiliation estimation by use of population-specific DNA markers.** *Am J Hum Genet* 1997, **60**:957-964.
  13. Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB: **Human population genetic structure and inference of group membership.** *Am J Hum Genet* 2003, **72**:578-589.
  14. Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N, Baron A, Jackson T, Argyropoulos G, Jin L, Hoggart CJ, McKeigue PM, Kittles RA: **Skin pigmentation, biogeographical ancestry and admixture mapping.** *Hum Genet* 2003, **112**:387-399.
  15. Yang BZ, Zhao H, Kranzler HR, Gelernter J: **Practical population group assignment with selected informative markers: Characteristics and properties of Bayesian clustering via STRUCTURE.** *Genet Epi* 2005, **28**:302-312.
  16. Serre D, Paabo S: **Evidence for gradients of human genetic diversity within and among continents.** *Genome Res* 2004, **14**:1679-1685.
  17. Butler JM: **Genetics and genomics of core STR loci used in human identity testing.** *J Forensic Sci* 2006, **51**:253-265.
  18. Agrawal S, Khan F: **Reconstructing recent human phylogenies with forensic STR loci: A statistical approach.** *BMC Genet* 2005, **6**:47.
  19. Rowold DJ, Herrera RJ: **Inferring recent human phylogenies using forensic STR technology.** *Forensic Sci Int* 2003, **133**:260-265.
  20. Sun G, McGarvey ST, Bayoumi R, Mulligan CJ, Barrantes R, Raskin S, Zhong Y, Akey J, Chakraborty R, Deka R: **Global genetic variation at nine short tandem repeat loci and implications on forensic genetics.** *Eur J Hum Genet* 2003, **11**:39-49.
  21. Budowle B, Chakraborty R: **Population variation at the CODIS core short tandem repeat loci in Europeans.** *Legal Medicine* 2001, **3**:29-33.
  22. Adams RI, Brown KM, Hamilton MB: **The impact of microsatellite electromorph size homoplasy on multilocus population structure estimates in a tropical tree (*Corythophora alta*) and an anadromous fish (*Morone saxatilis*).** *Mol Ecol* 2004, **13**:2579-2588.
  23. Angers B, Estoup A, Jarne P: **Microsatellite size homoplasy, SSCP, and population structure: a case study in the freshwater snail *Bulinus truncatus*.** *Mol Biol Evol* 2000, **17**:1926-1932.
  24. Van Oppen JH, Rico C, Turner GF, Hewitt GM: **Extensive homoplasy, nonstepwise mutations and ancestral polymorphism at a complex microsatellite locus in the lake Malawi Cichlids.** *Mol Biol Evol* 2000, **17**:489-498.
  25. **National Institute of Standards and Technology Short Tandem Repeat DNA internet database** [<http://www.cstl.nist.gov/biotech/strbase/>]
  26. Queller DC, Goodnight KF: **Estimating relatedness using genetic markers.** *Evolution* 1989, **43**:258-275.
  27. Rannala B, Mountain JL: **Detecting immigration by using multilocus genotypes.** *Proc Natl Acad Sci USA* 1997, **94**:9197-9201.
  28. Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M: **New methods employing multilocus genotypes to select or exclude populations as origins of individuals.** *Genetics* 1999, **153**:1989-2000.
  29. Dawson KJ, Belkhir K: **A Bayesian approach to the identification of panmictic populations and the assignment of individuals.** *Genet Res* 2001, **78**:59-77.
  30. Rosenberg NA, Li LM, Ward R, Pritchard JK: **Informativeness of genetic markers for inference of ancestry.** *Am J Hum Genet* 2003, **73**:1402-1422.
  31. Mountain JL, Ramakrishnan U: **Impact of human population history on distributions of individual-level genetic distance.** *Hum Genomics* 2005, **2**:4-19.
  32. Pritchard JK, Stephens M, Donnelly PJ: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155**:945-959.
  33. Falush D, Stephens M, Pritchard JK: **Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies.** *Genetics* 2003, **164**:1567-1587.
  34. Piry S, Luikart G, Cornuet J-M: **BOTTLENECK: a computer program for detecting recent reductions in the effective size using allele frequency data.** *J Hered* 1999, **90**:502-503.
  35. Quincy, Keith: *Hmong: History of a People* Cheney, Wash: Eastern Washington University Press; 1988.
  36. Yang BZ, Zhao H, Kranzler HR, Gelernter J: **Characterization of a likelihood based method and effects of markers informativeness in evaluation of admixture and population group assignment.** *BMC Genet* 2005, **6**:50.
  37. Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW: **Clines, clusters, and the effect of study design on the inference of human population structure.** *PLoS Genet* 2005, **1**:e70.
  38. Chakraborty R, Nei M: **Bottleneck Effects on Average Heterozygosity and Genetic Distance with the Stepwise Mutation Model.** *Evolution* 1977, **31**:347-356.
  39. Aulchenko YS, Heutink P, Mackay I, Bertoli-Avella AM, Pullen J, Vaessen N, Rademaker TA, Sandkuijl LA, Cardon L, Oostra B, van Duijn CM: **Linkage disequilibrium in young genetically isolated Dutch population.** *Eur J Hum Genet* 2004, **12**:527-534.
  40. Varilo T, Laan M, Hovatta I, Vibe V, Terwilliger JD, Peltonen L: **Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo.** *Eur J Hum Genet* 2000, **8**:604-612.
  41. Vitart V, Carothers AD, Hayward C, Teague P, Hastie ND, Campbell H, Wright AF: **Increased level of linkage disequilibrium in rural compared with urban communities: A factor to consider in association-study design.** *Am J Hum Genet* 2005, **76**:763-772.
  42. Escamilla MA: **Population isolates: their special value for locating genes for bipolar disorder.** *Bipolar Disord* 2001, **3**:299-317.
  43. Wang WY, Barratt BJ, Clayton DG, Todd JA: **Genome-wide association studies: theoretical and practical concerns.** *Nat Rev Genet* 2005, **6**:109-118.
  44. Luikart G, Allendorf FW, Cornuet JM, Sherwin WB: **Distortion of allele frequency distributions provides a test for recent population bottlenecks.** *J Hered* 1998, **89**:238-247.
  45. Cornuet JM, Luikart G: **Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data.** *Genetics* 1996, **144**:2001-2014.
  46. Maruyama T, Fuerst PA: **Population bottlenecks and nonequilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability.** *Genetics* 1984, **108**:745-763.
  47. Schliesinger J: *Ethnic Groups of Thailand: Non-Tai-Speaking Peoples* Bangkok: White Lotus Press; 2000.
  48. Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB: **Mutational processes of simple-sequence repeat loci in human populations.** *Proc Natl Acad Sci USA* 1994, **91**:3166-3170.
  49. Williamson-Natesan EG: **Comparison of methods for detecting bottlenecks from microsatellite loci.** *Conservation Genetics* 2005, **6**:551-562.
  50. Kalinowski ST, Wagner AP, Taper ML: **ML-Relate: a computer program for maximum likelihood estimation of relatedness and relationship.** *Molecular Ecology Notes* 2006, **6**:576-579.
  51. Raymond M, Rousset F: **GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism.** *J Heredity* 1995, **86**:248-249.
  52. Park SDE: **Trypanotolerance in West African Cattle and the Population Genetic Effects of Selection.** In PhD thesis University of Dublin; 2001.
  53. Nei M: *Molecular Evolutionary Genetics* New York: Columbia University Press; 1987.
  54. Smith MW, Lautenberger JA, Shin HD, Chretien JP, Shrestha S, Gilbert DA, O'Brien SJ: **Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations.** *Am J Hum Genet* 2001, **69**:1080-1094.