doi: 10.2306/scienceasia1513-1874.2006.32.361

Minimal Y-chromosomal Haplotype Selection for Phylogenetic Study using the Bootstrapped DTI Method

Metawee Srikummool[a], Jeerayut Chaijaruwanich[b], Jatupol Kampuansai[a] and Daoroong Kangwanpong[a*]

[a] Department of Biology, Faculty of Science, Chiang Mai University, Chiang Mai 50202, Thailand.
[b] Department of Computer Science, Faculty of Science, Chiang Mai University, Chiang Mai 50202, Thailand.

* Corresponding author, E-mail: scidkngw@chiangmai.ac.th

Received 6 Jan 2006
Accepted 24 Apr 2006

ABSTRACT: In our previous report, a decision tree induction (DTI) algorithm proved to be an effective tool to measure impurities of categorizing populations and to select the minimum number of Y-microsatellite markers, with the highest discriminating power. By using the DTI algorithm to determine the most informative loci in four studied hilltribe populations, i.e. the Akha, Hmong, Karen, and Lisu, seven powerful loci for phylogenetic reconstruction were selected and validated. However, the accuracy of the DTI algorithm on haplotype selection of other hilltribe populations was still in question. Therefore, a bootstrap method was applied in this study to the same data and contexts to verify the results of the DTI algorithm. We also examined the bootstrap method in a larger sample size and with a larger number of populations. Three informative microsatellite loci were selected to differentiate among the populations. When all loci were ranked, the three selected loci usually appeared at high potential levels. It can be concluded that we were able to regenerate the minimal haplotype— a combination of three selected loci— with a high level of confidence for effective phylogenetic study of human populations.

KEYWORDS: Y chromosome, microsatellite, minimal haplotype, decision tree induction, bootstrap method.

INTRODUCTION

For decades, the Y chromosome has proven to be a very powerful tool to study the history of human populations. The human Y chromosome, with a large non-recombining block, known as a haplotype, is passed down only through the male line. Without recombination during meiosis, this chromosome can be tracked back to the male origin. A variety of polymorphic markers, ranging from a single-base change to many repeated unit changes, are now available on the human Y chromosome[1, 2].

A microsatellite or a short tandem repeat is a polymorphic marker which contains valuable information on population demographic and evolutionary processes. Studies of human population structures, based on a number of Y microsatellite loci, have been reported and the relatedness of populations examined. An important finding was that not all studied loci were informative. In our previous study,[3] a decision tree induction (DTI) algorithm was employed to select the most informative loci for studying human populations in Thailand to reduce laboratory time and lower costs. Seven out of 15 highly-informative microsatellite loci from northern hilltribe populations were selected by the DTI algorithm. Results of Y chromosomal diversity and genetic distances among the studied populations, using either the 7 selected loci or all 15 microsatellite loci, were equivocal. They also showed that the selected loci were appropriate and had high discriminating power for differentiating populations.

However, the accuracy of the DTI algorithm for haplotype selection was remained an issue, especially when considering the sample size. Therefore, we introduced a bootstrap method into this study. This simple but reliable concept has been extensively used for assessing statistical confidence in phylogenetic trees[4]. In our present report, the bootstrap method was employed to the same studied populations and samples as in the previous paper[3], as well as to additional samples and populations in an attempt to increase the accuracy and robustness of the DTI algorithm prediction.

MATERIALS AND METHODS

Volunteer study subjects were unrelated hilltribe males living in northern Thailand. Sample data were classified into three sets (Set I, II and III), as shown in Table 1. The protocols used for blood sample collection, DNA extraction and detection of genetic variation were described in our previous paper[3].

Table 1. Sample sets and the study populations.

Samples	Description	Studied populations (N)	Total
Set I	Original sample[3]	Akha (14), Lisu (11), Karen (19), Hmong (7)	51
Set II	Set I with more samples	Akha (14), Lisu (11), Karen (83), Hmong (97)	205
Set III	Set II with Yao	Akha (14), Lisu (11), Karen (83), Hmong (97), Yao (52)	257

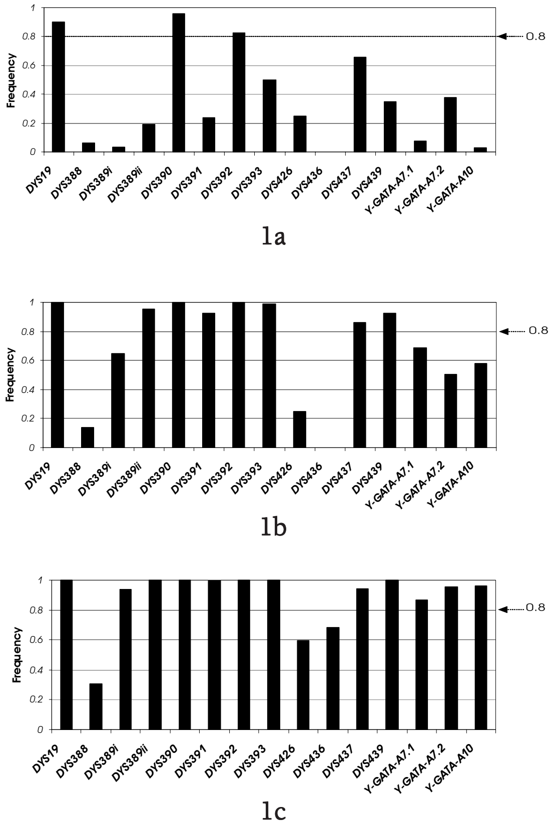
The Bootstrap Method and Loci Ranking

Genetic information of 15 Y-microsatellite loci from each data sample was randomly obtained and re-sampled from the chromosome pool to create a new ideal population of equal population size. The ideal population was used to compute the DTI algorithm, following the published method[3]. One thousand ideal populations for each sample data were randomly generated and their trees of loci were calculated by the DTI algorithm and collected. A locus, located at the root, or level 1, of the tree indicated its importance for discriminating populations. The discriminative power of each locus was lower when it was absent from the tree or located at a level with a higher number. The discriminating power of each locus was evaluated and ranked, based on its frequency of occurrence at each level of the trees. In order to obtain powerful discriminative loci, a highly stringent cut-off level of 0.8 (80% presence) was applied. In the other words, only the loci showing greater frequencies than the cut-off point were accepted.

Results

In order to verify the accuracy of the DTI algorithm and its results, as reported in our 2004 publication[3], the bootstrap method was employed to analyze the previous data set (Set I). Bootstrapped decision trees revealed 3 informative loci (DYS19, DYS390, and DYS392) (Fig 1a) at the 0.8 cut-off level (80% presence). When more samples of the Karen and the Hmong were considered, with 83 and 97 individuals respectively (Set II), 8 loci (DYS19, DYS389ii, DYS390, DYS391, DYS392, DYS393, DYS437 and DYS439) were selected (Fig 1b). These covered all 3 loci found from data Set I. However, when the Yao, which was a newly studied population, was added (as Set III), the results indicated that most loci, except DYS388, DYS426, and DYS436, had frequencies of occurrence greater than 0.8 (Fig 1c).

The loci from each sample set were ranked, in order to estimate the discriminative power. Level 1 was ranked as having the most discriminative power, using the percentage of occurrences of loci in 1,000 bootstrapped decision trees. Table 2 shows the percentage of occurrences of selected loci at each level. Among the loci selected from all three data sets, the DYS19 and DYS392 loci were highly ranked (level 1 and 2) in all data sets, indicating high discriminative power. The DYS390 locus was presented in level 2 of Set I with rather a high percentage (58%). If we considered the percentage of occurrence from levels 1 to 3 in Set II, and levels 1 to 4 in Set III, the candidate loci which could be added to the above 3 loci would be DYS393 (73% presence in Set II) and DYS439 (73% presence in Set III) which were the loci that had the highest summed percentage of presence in both set.



1a

1b

1c

Fig 1. Occurrence frequency of each locus in 1,000 bootstrapped decision trees. a) Set I (51 individuals), b) Set II (205 individuals), c) Set III (257 individuals).

Table 2. Percentages of each locus present at different levels from 1,000 bootstrapped decision trees of 3 sample sets.

Loci	Set I					Set II									Set III														
	Presence (%) level				Absence (%)	Presence (%) level								Absence (%)	Presence (%) Level														Absence (%)
	1	2	3	4		1	2	3	4	5	6	7	8		1	2	3	4	5	6	7	8	9	10	11	12	13	14	
DYS19	69	21	0	0	10	40	43	10	6	1	0	0	0	0	69	10	3	3	0	0	0	0	0	0	0	0	0	15	0
DYS388	1	6	0	0	93	0	3	9	2	0	0	0	0	86	0	4	26	1	0	0	0	0	0	0	0	0	0	0	69
DYS389i	0	3	0	0	97	0	1	12	17	12	16	6	1	35	0	2	11	17	14	9	1	0	0	0	0	0	0	39	7
DYS389ii	0	2	16	1	81	0	3	21	29	27	9	5	1	5	0	3	16	33	27	13	4	3	1	0	0	0	0	0	0
DYS390	9	58	24	5	4	0	9	30	33	21	4	3	0	0	0	2	21	37	26	13	1	0	0	0	0	0	0	0	0
DYS391	0	16	8	0	76	0	4	35	21	14	13	6	1	8	0	0	10	26	32	9	19	3	0	0	0	0	0	0	1
DYS392	10	55	17	0	18	59	19	13	7	2	0	0	0	0	30	7	8	24	7	2	5	17	0	0	0	0	0	0	0
DYS393	0	5	26	19	50	0	41	32	17	7	2	0	0	1	0	10	21	37	22	3	1	2	4	0	0	0	0	0	0
DYS426	0	25	0	0	75	0	25	0	0	0	0	0	0	75	0	0	0	0	0	0	0	16	33	11	0	0	0	0	40
DYS436	0	0	0	0	100	0	0	0	0	0	0	0	0	100	0	0	5	7	1	0	0	0	15	30	10	0	0	0	32
DYS437	0	5	57	3	35	0	1	64	13	8	0	0	0	14	0	1	10	19	12	4	1	0	0	14	28	5	0	0	6
DYS439	0	14	20	0	66	0	26	32	23	10	1	0	0	8	0	8	38	27	22	2	1	0	0	0	1	1	0	0	0
Y-GATA-A7.1	2	3	2	1	92	0	26	32	9	1	1	0	0	31	0	7	29	8	0	0	0	0	0	0	12	24	7	0	13
Y-GATA-A7.2	9	22	6	2	61	0	15	15	18	2	0	0	0	50	1	35	6	6	2	0	0	0	0	0	0	16	30	0	4
Y-GATA-A10	0	2	1	0	97	1	19	22	15	0	0	0	0	43	0	6	16	30	19	5	3	1	0	0	0	0	16	0	4

Discussion

The advantage of this method is that any estimation error or bias can be corrected. In this study, a subset of data was randomly drawn 1,000 times from each particular data set. The final results were then summarized and displayed graphically as the selected frequency of each locus. The higher the frequency appeared in 1,000 bootstrapped decision trees, the more discriminative power the locus possessed, and the level of confidence increased accordingly. With the bootstrapped DTI method, we were able to select the most powerful discriminative loci for our population genetics study so far.

In our previous study, 7 selected loci, i.e. DYS19, DYS390, DYS392, DYS393, DYS426, DYS439 and Y-GATA-A7.2, were obtained using the DTI method. After introducing the bootstrapped DTI method which gave higher statistical confidence, 3 informative loci (DYS19, DYS390, and DYS392) from the 7 loci mentioned above were selected from the analysis of Set I data. These 3 microsatellite loci were repeatedly presented at high-level ranks in all studied data sets. Thus, they are the most highly informative and discriminative loci.

The potential of the DYS19, DYS390, and DYS392 loci for population distinction were reported in other human population studies. The allele frequency distribution of DYS19 exhibited a remarkable heterogeneity in Caucasian, African, Asian and Oceanic populations[5]. Locus-specific differences of DYS19, DYS390, and DYS392 loci were observed with regard to the pairwise allele-frequency comparisons between populations in the extensive analysis of the Y chromosomes from 20 globally dispersed human populations[6]. The universal usage of these 3 selected loci among many populations world-wide and the reported locus sharing among populations could probably reflect their single geographically restricted origin[1,7].

When data set II (larger sample sizes) and set III (with the additional population) were analyzed with the bootstrapped DTI method, the numbers of selected loci needed to categorize the studied populations increased. The increased locus number needed was most probably due to the heterogeneity of the added samples and populations, which also increased overall genetic variation. Therefore, if high genetic variances are found among populations, more loci should be added to ensure their affinity.

With our validation technique using the bootstrapped DTI method, we strongly suggest that the combination of 3 selected loci (DYS19, DYS390, and DYS392) could be the minimal haplotype required for effective differentiation among Thai hilltribe populations. In case of more samples or populations, the other 2 loci, DYS393 and DYS439, would be good candidates. This minimal haplotype might also be a good starting point for other human populations.

Acknowledgements

The authors thank all volunteers for their collaboration. We thank the staff of the Tribal Research Institute, Chiang Mai, Thailand, for organizing field work, as well as Dr. Sukon Prasitwattanaseree from the Department of Statistics, Faculty of Science, Chiang

Mai University, Chiang Mai, Thailand, for his statistical consulting. This project was supported by grant numbers PHD/0011/2544 and BGJ/26/2544 of the Thailand Research Fund.

References

1. Hammer MF and Zegura SL (1996) The role of the Y chromosome in human evolutionary studies. *Evol Anthropol* **5**, 116-34.
2. Thomas MG, Bradman N and Flinn HM (1999) High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. *Hum Genet* **105**, 577-81.
3. Kangwanpong D, Chaijaruwanich J, Srikummool M and Kampuansai J (2004) Selection of Y-chromosomal microsatellies for phylogenetic study among hilltribes in northern Thailand using the decision tree induction algorithm. *ScienceAsia* **30**, 239-45.
4. Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783-91.
5. Santos FR, Gerelsaikhan T, Munkhtuja B, Oyunsuren T, Epplen JT and Pena SDJ (1996) Geographic differences in the allele frequencies of the human Y-linked tetranucleotide polymorphism DYS19. *Hum Genet* **97**, 309-13.
6. Kayser M, Krawczak M, Excoffier L, Dieltjes P, Corach D, Pascali V and Gehrig C, et al (2001) An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. *Am J Hum Genet* **68**, 990-1018.
7. Relethford JH (2001) Genetic diversity and recent human evolution. In: *Genetics and the search for modern human origins* (Edited by Relethford JH), pp 94-118. Wiley-Liss, New York, NY.