

# Detailed mtDNA Genotypes Permit a Reassessment of the Settlement and Population Structure of the Andaman Islands

S.S. Barik,<sup>1</sup> R. Sahani,<sup>1</sup> B.V.R. Prasad,<sup>1</sup> P. Endicott,<sup>2</sup> M. Metspalu,<sup>3</sup> B.N. Sarkar,<sup>1</sup> S. Bhattacharya,<sup>1</sup> P.C.H. Annapoorna,<sup>1</sup> J. Sreenath,<sup>1</sup> D. Sun,<sup>1</sup> J.J. Sanchez,<sup>4</sup> S.Y.W. Ho,<sup>2</sup> A. Chandrasekar,<sup>1</sup> and V.R. Rao<sup>1\*</sup>

<sup>1</sup>*Anthropological Survey of India, 27 Jawaharlal Nehru Road, Kolkata 700 016, India*

<sup>2</sup>*Henry Wellcome Ancient Biomolecules Centre, Department of Zoology, University of Oxford, Oxford OX1 3PS, UK*

<sup>3</sup>*Institute of Molecular and Cell Biology, University of Tartu and Estonian Biocentre, Tartu University, Tartu, Estonia*

<sup>4</sup>*National Institute of Toxicology and Forensic Science, Canary Islands Delegation, Campus de Ciencias de la Salud, 38320 La Laguna, Tenerife, Spain*

**KEY WORDS** human migration; phylogeography; Jarawa; Munda; Pauri bhuiya; India

**ABSTRACT** The population genetics of the Indian subcontinent is central to understanding early human prehistory due to its strategic location on the proposed corridor of human movement from Africa to Australia during the late Pleistocene. Previous genetic research using mtDNA has emphasized the relative isolation of the late Pleistocene colonizers, and the physically isolated Andaman Island populations of Island South-East Asia remain the source of claims supporting an early split between the populations that formed the patchy settlement pattern along the coast of the Indian Ocean. Using whole-genome sequencing, combined with multiplexed SNP typing, this study investigates the deep structure of mtDNA haplogroups M31 and M32 in India and the Andaman Islands. The identification of a so far unnoticed rare polymorphism

shared between these two lineages suggests that they are actually sister groups within a single haplogroup, M31/32. The enhanced resolution of M31 allows for the inference of a more recent colonization of the Andaman Islands than previously suggested, but cannot reject the very early peopling scenario. We further demonstrate a widespread overlap of mtDNA and cultural markers between the two major language groups of the Andaman archipelago. Given the “completeness” of the genealogy based on whole genome sequences, and the multiple scenarios for the peopling of the Andaman Islands sustained by this inferred genealogy, our study hints that further mtDNA based phylogeographic studies are unlikely to unequivocally support any one of these possibilities. *Am J Phys Anthropol* 136:19–27, 2008. ©2008 Wiley-Liss, Inc.

The genetic origin of the inhabitants of the Andaman Islands (a group of islands in the Bay of Bengal) continues to drive a healthy debate that encompasses the prehistory of anatomically modern humans in general. The interest in the inhabitants of this archipelago stems from their distinctive phenotype and languages, which give them the appearance of enduring isolation (Barnard-Davis, 1867; Portman, 1884; Man, 1885; Cooper, 2002). Central to the discussion is whether the various indigenous people of South and South-East Asia who resemble the Andaman Islanders derive from a common pre-Neolithic regional population, or if their shared features represent convergent evolution (Endicott et al., 2003). These so-called negritos (from the Spanish diminutive for “black”) are found from India to the Philippines and are predominantly tribal groups who pursue a subsistence strategy of mobile resource procurement.

The Andaman Islanders, because of their distinctive languages and history of isolationism (Cooper, 1989), are often touted as a group of “Palaeolithic survivals” who might represent the direct descendants of an early wave of human migrants passing through the region (Thangaraj et al., 2005, 2006a). The archaeological record of the Andaman Islands is limited and does not extend beyond the first millennium BC (Cooper, 2002). This leaves open a range of other possible scenarios including a relatively recent settlement of the islands by migrating

resource exploiters (i.e. during the Holocene), presumably from what is now Myanmar (Morrison, 2007). With such an ephemeral physical record of occupation, the onus is on genetic archaeology to clarify the prehistory of these peoples and to produce a regional narrative, to fit into a broader understanding of the global prehistory of anatomically modern humans.

Although mitochondrial DNA (mtDNA) evolves as a single genetic marker, it has proved a very valuable one in terms of reconstructing human population histories. Its recombination-free, uni-parentally female inheritance

This article contains supplementary material available via the Internet at <http://www.interscience.wiley.com/jpages/0002-9483/suppmat>.

Grant sponsor: Anthropological Survey of India, Ministry of Culture, Government of India.

\*Correspondence to: Dr. V.R. Rao, Anthropological Survey of India, Government of India, 27, Jawaharlal Nehru Road, Kolkata 700 016, India. E-mail: [dr-raovr@yahoo.com](mailto:dr-raovr@yahoo.com)

Received 30 July 2007; accepted 31 October 2007

DOI 10.1002/ajpa.20773

Published online 10 January 2008 in Wiley InterScience ([www.interscience.wiley.com](http://www.interscience.wiley.com)).

pattern enables one to reconstruct the emergence of new variation in the form of a true genealogy. Also, the rate at which new mtDNA variation is generated is suitable for decoding the past of our species, because it is fast enough to provide a substantial amount of information (i.e. segregating sites in DNA), but slow enough to prevent multiple hits (saturation) from obscuring the actual genealogy. Given the normal allelic richness of an mtDNA pool, random genetic drift events (bottlenecks, founder effects) will leave explicit patterns on standing variation. The current resolution of the genealogy of human mtDNA lineages provides evidence that the major regions of the world harbor distinct sets of maternal lineages, and that the coalescent dates of the region-specific mtDNA lineages are consistent with a movement of people out of Africa during the late Pleistocene (Quintana-Murci et al., 1999; Maca-Meyer et al., 2001; Kivisild et al., 2003). Although estimates of these dates vary according to the substitution rates that are assumed (Mishmar et al., 2003; Kivisild et al., 2006), the branching order of the tree remains the same.

All the mtDNA lineages present between South Asia and Australia are rooted directly to the three pan-Eurasian founding haplogroups (hgs) M, N, and R (Palanichamy et al., 2004; Friedlaender et al., 2005; Macaulay et al., 2005; Merriwether et al., 2005; Sun et al., 2006; Thangaraj et al., 2006b; Hudjashov et al., 2007). Hence, there is no nested structure in the mtDNA genealogy along the proposed track of peopling of Eurasia; for example, the Australian mtDNA types do not derive from East Asian types but from the same founder types that were the inocula for the East Asian mtDNA types. It is parsimonious to see this as evidence of the pioneer settlement by anatomically modern humans across the region. This branching pattern also suggests that the peopling of the world beyond Africa was relatively rapid because of the apparent lack of a nested structure within the region-specific haplogroups, which have evolved *in situ* from the three founder types subsequent to the initial wave of settlement (Macaulay et al., 2005; Metspalu et al., 2006).

The phylogeography of macro-haplogroup M is an important piece of evidence for the recent Out-of-Africa migration of modern humans taking a southern route to Australia. The virtual lack of hg M (with the exception of M1 in Africa (Quintana-Murci et al., 1999) in the regions west of South Asia (Metspalu et al., 2004; Quintana-Murci et al., 2004) is seen as a signature of the route that followed the coast of the Indian Ocean from East Africa to South Asia.

The major Andaman mtDNA lineages were shown to belong to haplogroup M by Endicott et al. (2003) and Thangaraj et al. (2003). Subsequent whole mitochondrial genome data from the Onge and Greater Andamanese populations defined the two Andaman M lineages as M31 and M32 (Thangaraj et al., 2005, 2006b), presuming them to follow the region-specific pattern and to be remnants of this single rapid dispersal of modern humans during the late Pleistocene. The discovery of two examples of M31b (a sister clade to the Andamanese M31a) in West Bengal ran counter to this view and suggested that the Andaman-specific M31a may have originated in India (Palanichamy et al., 2006). The subsequent publication of 13 members of a sister clade (M31a2) to the Andaman variant M31a1, from East India, re-opened the possibility of a South-East Asian origin for M31 but also requires a later back-migration to India to explain the

contemporary distribution of this haplogroup (Endicott et al., 2006).

The fact that these inter-regional links were not discovered earlier is, in part, a result of insufficient information in the mtDNA control region; for example, the Indian and Andamanese variants of M31a harbor completely different motifs, rendering this approach to analyzing genealogical relationships over long time-scales unreliable. One solution to this lacuna is the generation of whole-genome coding-region data to place lineages onto the genealogy regardless of their control region signatures. Unfortunately, as the sheer number of sequenced nucleotides increases, so does the possibility for mistakes to occur, a problem exemplified in the published whole-genome data for the Andamanese (Thangaraj et al., 2005, 2006a) (the respective errata is in press). This has produced an incorrect topology for both M31 and M32, with implications for both inter-regional and intra-regional interpretations (Endicott et al., 2006).

The present study addresses the need for accurate whole genome data for hgs M31 and M32, and includes the additional Andaman population of the Jarawa. A multiplexed Single-Base-Extension (SBE) assay is used to accurately genotype the fine structure of mtDNA from historical samples (extracted from teeth held in the collections of the Natural History Museum London, the Oxford University Museum of Natural History, and the Royal College of Surgeons Edinburgh) providing important additional data on the Greater Andamanese, who have experienced a sustained genetic bottleneck. On the inter-regional scale, it broadens the search for links between the mtDNA of South-East Asia and South Asia by screening Indian tribal populations (see Fig. 1) for markers of M31 and M32 and conducting further whole genome sequencing. The structure of M31 and M32 is estimated by Bayesian phylogenetic analysis using *BEAST* 1.4 (Drummond et al., 2006), to assess their branching within a total data set of 165 mitochondrial genomes. The combination of this model-based approach, using protein-coding and control region data, combined with the use of additional phylogenetic partitions within a hand-drawn tree, yields the most parsimonious genealogy for both M31 and M32 and provides the basis for explaining the distribution of M31 in both regions.

This extant phylogeography of haplogroup M31/32 lineages is then used to explore three different scenarios for the peopling of the Andaman Islands: i) as part of the pioneer settlement (“very early”) of South-East Asia by anatomically modern humans ~45 kya; ii) as a later settlement (“recent”) with an upper limit around the Last Glacial Maximum (~24 Kya); and iii) as a “very recent” settlement during the Holocene from the South-East Asian mainland. The implications for inter-regional prehistoric migrations of human populations are considered in the context of each hypothesis.

## MATERIALS AND METHODS

Samples included in this study were selected from a survey of 3,026 individuals from 30 Indian mainland tribal populations and two from the Andaman archipelago (see Fig. 1). They were screened for membership of M31 and M32 by a combination of control region sequences and coding region SNP data (Kivisild et al., 2003; Metspalu et al., 2004; Sun et al., 2006; Endicott et al., 2006). Five samples from the Pauri Bhuyia and two Munda with HVS1 motifs including 16017–16126–



**Fig. 1.** Locations of the Indian tribal populations studied, listed as ethno-linguistic categories. Red and blue circles indicate tribes where haplogroups M31 and M32, respectively, were found. In Andaman Islands, all the three tribal groups—Great Andamanese, Jarawa, and Onge—possess both M31 and M32 haplogroups, whereas tribes in mainland India have haplogroup M31 only (indicated by red circles) (Palanichamy et al., 2006).

16145–16223, which were not linked to hg M3 by the coding region mutation at np 4,580 (Sun et al., 2006), together with 10 Jarawa (five M31 and five M32) samples were sequenced for the complete mitochondrial genome using 24 pairs of forward and reverse primers published elsewhere (Rieder et al., 1998) in order to obtain double coverage for all the samples. The Ethical Committee of the Anthropological Survey of India approved the protocols. Samples were collected in Vacutainer as per standard protocols and extraction of DNA was performed according to the enzymatic extraction procedure followed by phenol purification (Sambrook et al., 1989), which was standardized at Anthropological Survey of India, C.R.C. laboratory, Nagpur. Sequences were assembled, and edited using SeqScape 2.0. Deviations from the rCRS (Anderson et al., 1981; Andrews et al., 1999) were confirmed by manual checking of their electropherograms. All the sequences have been deposited in the NCBI database (Accession Numbers: DQ149511 to DQ149520, EF060262 to EF060266 and EU075305–EU075306).

Twenty historical samples from the Andamans, together with contemporary individuals from the mainland Indian populations of Chenchu, Lambadi, and Lodha ( $n = 13$ ) were previously genotyped for 20 SNPs associated with the main structure of hgs M31 and M32 (Endicott et al., 2006). Here, a second two-stage multiplexed SBE reaction was optimized to investigate the markers of G1438A and an insertion A at np 2,156 from the main trunk of M31/32 (this study), together with 12 additional markers previously reported for these haplogroups (Thangaraj et al., 2005). The inclusion of four markers (9,581; 9,617; 11,014; and 15,530) defining various parts of M31, which were previously reported for these samples, provides security for the results from this second multiplex. Eight Andamanese samples were not tested due to insufficient template but the existing hierarchical

phylogenetic markers provide sufficient resolution to accurately place them onto the tree. The two-stage reactions used the primers in Supplementary Tables S1 and S2. Details of protocols for the design, operation, and interpretation of this methodology are available elsewhere (Endicott et al., 2006; Sanchez and Endicott, 2006). No evidence of contamination amongst the samples, in the form of secondary peaks in the SBE assays, was found in either the modern or historical samples (Sanchez and Endicott, 2006). All SBE assays were performed twice and the results were consistent with the haplogroup assignments previously given to the samples (Endicott et al., 2006). The genotypes obtained match those of the main branches of the M31/32 genealogy reported here without exception.

Coalescence time estimates were calculated using a substitution rate estimate for protein-coding synonymous changes of  $3.5 \times 10^{-8}$ , which gives 6,764 years per synonymous transition (Kivisild et al., 2006). This has the advantage of excluding those non-synonymous mutations that may be slightly deleterious and therefore subject to purifying selection. We also consider the protein-coding region estimates to be less biased than those that include a) RNA and inter-genic regions due to the difficulty in accounting for variations in their phylogenetic signals, and b) the control region, because of less rate heterogeneity amongst sites and between lineages (Hasegawa and Horai, 1991; Excoffier and Yang, 1999; Meyer et al., 1999; Heyer et al., 2001). Nevertheless, all dates calculated without evidence to sustain the assumption of the molecular clock mean that estimation of the associated error values (Saillard et al., 2000) are only an approximation.

To infer the position of the sequenced mitochondrial genomes in the wider geographic context, Bayesian phylogenetic analysis was performed using *BEAST* 1.4 (Drummond et al., 2006) on an alignment of 165 mito-

chondrial genomes sampled from across the worldwide diversity of humans, with a particular focus on haplogroups M and N. The alignment was partitioned into first and second codon sites of protein-coding genes, third codon sites of protein-coding genes, and entire control region. The substitution model for each partition was selected by comparison of Akaike Information Criterion scores. An uncorrelated lognormal relaxed-clock was assumed to accommodate rate heterogeneity among lineages during genealogical reconstruction (Drummond et al., 2006). Posterior distributions of parameters, including the tree, were obtained by Markov chain Monte Carlo (MCMC) sampling. The MCMC was run for 20,000,000 steps, with samples drawn every 5,000 steps. The analysis was repeated and the two chains were combined. Acceptable mixing and convergence to the stationary distribution were checked. The maximum clade credibility tree was computed from the set of posterior samples.

## RESULTS

The 10 Jarawa, five Pauri Bhuiya, and two Munda fully sequenced mtDNA genomes are arranged into a genealogical tree (see Fig. 2) along with the published M31 and M32 complete (Thangaraj et al., 2005; Palanichamy et al., 2006; Thangaraj et al., 2006b) and partial mtDNA sequences (Metspalu et al., 2004; Endicott et al., 2006; this study, Fig. S1). The branching structure of M31 and M32, and their independence from other hg M lineages included in the analysis, is confirmed by the results from Bayesian phylogenetic analysis (Fig. S2). The major adjustment to the previously established genealogy is the apparent monophyly of hgs M31 and M32 through an insertion of an A at np 2,156 (2156insA). Here we note that, according to this topology, the aforementioned insertion has reverted in hg M31b, so far represented by a single fully sequenced mtDNA (Palanichamy et al., 2006).

Alternative reconstructions of the 2156insA would imply either a double occurrence (both in M31 and M32) of this very rare mutation or four parallel coding region mutations in M31b (see Fig. 2). The 2156insA has been reported in a few African mtDNAs (Ingman et al., 2000; Howell et al., 2004) where it occurs once and is one of the defining mutations for hg L1c2b1 (based on an analysis of 600 African complete mtDNA sequences; Doron Behar, personal communication). Thus, the most parsimonious explanation for the presence of 2156insA in both M31 and M32 amongst the Andaman Islanders is that they are derived from the same common ancestor carrying 2156insA.

The second striking feature of the revised genealogy is the divergence of M31a into Andaman-specific M31a1 and mainland India-specific M31a2. The Andaman populations are characterized by M31 and M32 lineages of mtDNA, but there are no traces of M32 lineage in mainland India (see Fig. 1). Andaman-specific M31 lineages are now nested within a largely mainland-specific genealogy. Relative to previous reconstructions (Thangaraj et al., 2006b), we now move the transition at np 16,126 from M31b to the trunk of M31 because all mtDNAs of the novel M31a2 also harbor this substitution. This implies a reversion of 16,126 in M31a1. We note, however, that the alternative reconstruction would be a dual mutation of np 16,126 in M31b and M31a2.

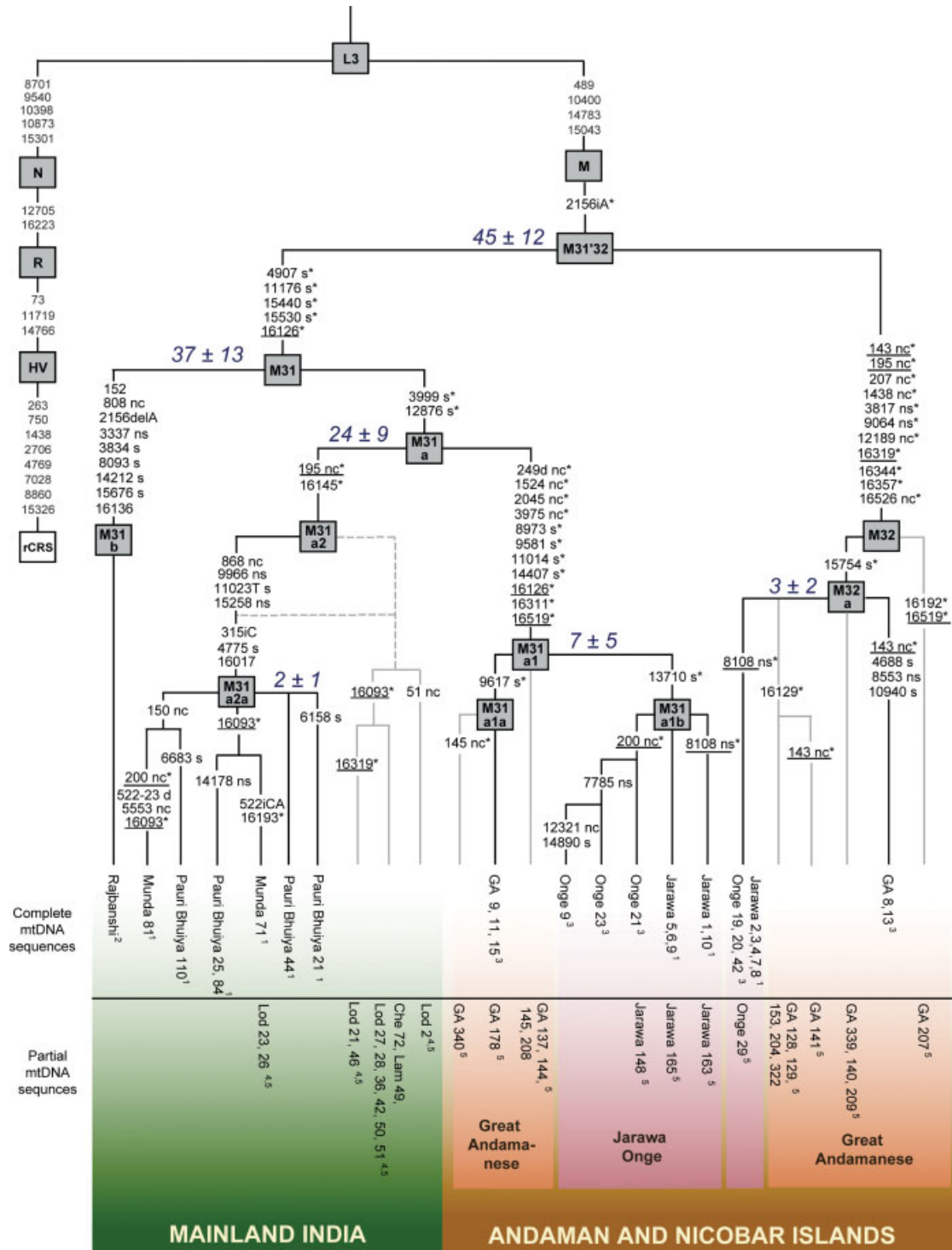
The coalescence estimate for hg M31'32 is essentially a coalescence estimate for hg M using only M31'32 data and is therefore of little interest. Those for the splitting of hg M31a1 and M31a2 yielded dates well into the late Pleistocene at 24 ( $\pm 9$ ) thousand years ago (kya), whereas the coalescence estimate for the Andaman-specific branches ( $<12$  kya) clearly postdates the Last Glacial Maximum (LGM). Although these dates carry the normal levels of uncertainty, our Bayesian analysis indicates no substantial deviation from the molecular clock in the M31'32 genealogy, providing additional security to these estimates.

Another important feature of the reconstruction of the M31'32 genealogy is the general separation of the culturally differentiated Greater Andaman tribes from the Onge and Jarawa populations. This is apparent in both hgs M31a1 and M32. Hg M31a1b is exclusively present among the Jarawa and the Onge whereas the Great Andamanese sport M31a1a. In hg M32 the situation is similar and one finds the Jarawa and Onge samples only on one sub-lineage of M32. It is difficult to assess the time depth of the split within both lineages, but the appearance of only a single synonymous substitution in both M31a1a and M31a1b is consistent with a time frame within the period since the terminal Pleistocene.

## DISCUSSION

The new reconstruction of the M31'32 genealogy permits various interpretations regarding the demographic history of the Andaman Islands for the initial peopling of the archipelago. The traditional scenario is for a very "early" settlement of the Andaman Islands, perhaps during the initial out-of-Africa population movements. This hypothesis has M31'32 following the established pattern of deep-rooting M haplogroups to be region-specific (Macauley et al., 2005; Sun et al., 2006; Thangaraj et al., 2006b) and to have evolved *in situ* from parts of the early migrations that arrived in Australia and New Guinea at least 45 kya (Hudjashov et al., 2007). The current evidence does not overthrow this interpretation. However, within this scenario, the presence of M31b and M31a2 in mainland India can only be explained by back-migration(s) from the Andaman Islands. Though not impossible, this claim still falls within the category of extraordinary claims that need extraordinary proof, and thus the very early settlement scenario will remain a possibility until we wait for further evidence from other genetic markers (Y-chromosome and autosomal). However, other scenarios can also be maintained with no unequivocal support for any one in particular (Table 1). Here, we advance an alternative hypothesis of a more recent colonization, and consider the evidence for India or Myanmar being the region where M31a1 and M31a2 differentiated.

The first model considered involves a very recent colonization of the Andaman Islands. This accords better with the very shallow archaeological record of the islands, which does not extend beyond 3,000 years ago (Cooper, 1989). Although, the very limited excavation that has taken place there, together with the reduction of land area since the LGM, suggests this could represent an absence of evidence as much as evidence of absence, the coalescence times of the Andaman-specific M31a1 and M32a are well within the Holocene and could be used to support the recent peopling scenario. However, if past genetic diversity has been obscured by fluc-



**Fig. 2.** Genealogical reconstruction of mtDNA haplogroups M31 and M32 lineages. The trunk of the tree is determined by complete sequence data in solid black lines whereas partial sequences are drawn in gray. Numbers along the lines designate substitutions in reference to the rCRS (Anderson et al., 1981; Andrews et al., 1999) whereas letters following the numbers code the following: s, synonymous substitution; ns, non-synonymous substitution; nc, substitution in non-coding or RNA-coding region. HVSI mutations are not coded. Underlined substitutions indicate multiple occurrences. Except for np 4,775 in Chenchu and Lambadi samples, nps 868; 4,775; 9,966; 11,023; and 15,258 are not determined in the Lodha, Chenchu, and Lambadi samples and therefore their genealogical relationship within M31a2 cannot be fully resolved. The considered samples originate as follows:<sup>1</sup> this study;<sup>2,3</sup> (Palanichamy et al., 2006; Thangaraj et al., 2005; Thangaraj et al., 2006a; Thangaraj et al., 2006b) all these sequences have been rechecked and, when necessary, partly re-sequenced (K. Thangaraj and G. Chaubey personal communication, the respective errata has been submitted to Science)<sup>4,5</sup> (Metspalu et al., 2004) (Endicott et al., 2006). Population abbreviations: Che, Chenchu; GA, Great Andamanese; Lam, Lambadi; Lod, Lodha; others are typed in full. The coalescent estimates are calculated following Kivisild et al (2006), using the synonymous substitution rate and given in thousands of years before present. Note that np 16,126 can alternatively be reconstructed as double occurrence within hgs M31b and M31a2. \*Denotes nps genotyped in the historic Andaman samples. [Color figure can be viewed in the online issue, which is available at [www.interscience.wiley.com](http://www.interscience.wiley.com).]

TABLE 1. Alternative scenarios for the settlement of the Andaman Islands

	Very recent	Recent	Very early
Definition	Holocene (<10 kya), from extant mainland South-East Asian population	Terminal Pleistocene, since the Last Glacial Maximum (~20 kya), from mainland South-East Asia	Late Pleistocene, during the initial peopling of coastal Eurasia (~45 kya)
Molecular date estimates based on synonymous protein-coding substitutions	7 ± 5 to 3 ± 2 kya; coalescence of the Andaman-specific mtDNA clades M31a1 and M32a	24 ± 9 to 7 ± 5 kya; coalescence of South-East and South Asian clades of M31a, and Andaman-specific M31a1	45 ± 12 kya; coalescence of M31 and M32, essentially the chronology for the dispersal of macro-haplogroup M within Eurasia
Concordance with archaeological record	Good	Poor	No
Concordance with molecular date estimates for the Andaman-specific clades	Good	Poor	No <sup>a</sup>
Explaining the mtDNA gene-pool of the Andaman Islanders sampled from two sister lineages of a single haplogroup, M31'32	An enigmatic isolated mainland population has to be evoked where M31'32 would have survived, differentiated into its various daughter lineages; this population would then have had to colonize the Andaman Islands, possibly in response to the advancing Neolithic populations		No problem: only one random lineage—which happened to be the root of M31'32—had to be sampled
Explanation for the presence of M31b and M31a2 in India	No back-migration required; the enigmatic mainland M31'32 population should though have left traces of its mtDNA pool on the mainland before settling in the Andaman Islands. Without detailed sampling of Myanmar this it is not possible to fully test this hypothesis		Can only be explained by back migration(s) from the Andaman Islands

<sup>a</sup> We note, however, that in populations especially prone to random genetic drift (e.g. due to constantly relatively small population size like on the Andaman Islands), the coalescence times are expected to underestimate the actual age of the lineages.

tuations in effective population size, such as bottleneck events, then the time depth of the surviving lineages could be seriously underestimated.

If the Andaman Islands were settled recently then we might expect to find M31'32 in adjoining regions, but neither haplogroup has been detected so far in mainland or Island South-East Asia (Hill et al., 2006, 2007), whereas only M31b and M31a2 are present in India at very low frequencies (Endicott et al., 2006; Palanichamy et al., 2006; this study). Unfortunately, the most geographically proximal source region, Myanmar, has not been extensively characterized for mtDNA sequence variation but a survey of data from the neighboring states of Northeast India, Malaysia, Thailand, and the Yunnan province of China for mtDNA control and partial coding region data reveals no clear candidate members of M31'32 (Fucharoen et al., 2001; Oota et al., 2001; Yao and Zhang, 2002; Yao et al., 2002; Cordaux et al., 2003; Metspalu et al., 2004; Wen et al., 2004; Oota et al., 2005; Wen et al., 2005; Hill et al., 2006; Kumar et al., 2006).

A second hypothesis is for a “recent” origin for the Andaman specific components of M31'32 in India. However, if M31 and M32 are rooted by 2156insA, it seems considerably less parsimonious to invoke an origin in India because it should be possible to find some examples of both clades in the source area. This is because the splitting of M31a1 and M31a2 would give both M31 and M32 tens of thousands of years to develop. Despite our comprehensive survey of populations in India no examples of M32 could be located, whereas M31a2 was located in several locations. The probability of this scenario is reduced further by the need for both lineages to have been sampled by the migrants that eventually colonized the Andaman Islands. However, it must be acknowledged that random genetic drift could have erased M32 from the Indian gene pool, or it may not have been sampled yet.

A recent migration event from Myanmar to the Andaman Islands, sampling both deep-rooting portions of M31'32, would suggest a mainland source population containing substantial amounts of these lineages. The lack of M31'32 in neighboring states except for India suggests a limited axis of regional movement for people with these lineages, and perhaps the same source population. Whether a hypothetical back-migration to India from a source in Myanmar can be linked to any cultural or physical markers in the present is a rather different matter. The current distribution of M31 in India is predominantly tribal, except for the Rajbanshi who are the most numerous scheduled castes in West Bengal (ca. 2.8 million people). However, the Rajbanshi are most likely a composite of former tribal populations who spoke languages from the Munda branch of the Austro-Asiatic languages (Kumar and Reddy, 2003; Thanseem et al., 2006). As most of the Austro-Asiatic speakers of India reside in East India, there is considerable overlap between the distribution of M31 and this linguistic phylum. However, it would be speculative to forward any specific inferences regarding the much disputed origins of the Indian Austro-Asiatic speakers based on a single haploid marker e.g. Basu et al (2003).

Importantly, whether the source of M31a2 was in India or Myanmar, the migration of people to the Andaman Islands with M31a1 cannot have occurred prior to the coalescence between these two clades, which we tentatively date at around 24 kya (±9). A recent colonization scenario implicitly assumes that there was a discrete continental (tribal) source population, within which the autochthonous M31 and M32 elements had survived long-term. This hypothetical population might have moved in the face of the advancing Neolithic farming populations and reached the Archipelago leaving negligible traces of their maternal legacy on the mainland. The wide standard errors on the dates for the linguistic-based divisions in M31a1 and M32

amongst the Andaman Islanders allows for a time of settlement compatible with the archaeological record (Morrison, 2007), but also for one in the Early Holocene when the sea levels were much lower than today.

This “very recent” peopling scenario, which evokes the presence of a small regional ancestral population on the mainland, also provides an explanation for the distinct phenotype of the Andaman Islanders. In this way, the cultural and phenotypic distinctiveness of a source population could have been maintained by relative isolation from the emerging Neolithic populations, thereby explaining the absence of M31/32 amongst contemporary mainland populations from adjoining regions. The more recent history of the Andaman populations appears to be much clearer, with genetic divergence that is broadly along linguistic lines during the Holocene. The fact that two mtDNA lineages dominate the genetic pool of the Andamanese is likely a product of i) random genetic drift fuelled by constant relatively small population sizes throughout (pre) history and ii) physical isolation of the islands from the mainland, which allowed only very low levels of migration into the Andaman maternal genetic pool.

## CONCLUSIONS

The improved genealogy for mtDNA hg M31/32 obtained by accurate whole genome data coupled with multiplexed SNP genotyping has permitted the evaluation of competing hypotheses for the peopling of the Andaman Islands. Of these, the previously prevalent one, that the inhabitants are direct descendants of the pioneer settlement of South-East Asia ~45 kya, is undermined by its requirement for a subsequent back-migration from the Andaman Islands to India.

Moving further from the “early” settlement hypothesis, the coalescence date for M31a1 and M31a2 provides an upper limit of ~24 kya for a migration taking the Andaman-specific variant to its current island home. Although this could have occurred at any time since the LGM, the apparent ages of the Andaman specific clades of M31/32 favors a chronology constrained within the last 10,000 years.

Under either of the “recent” hypotheses, whether M31/32 ultimately originated in South or South-East Asia is still unresolved, but both possibilities are potentially of great importance because of the need to evoke inter regional movements of people during a period of human prehistory about which little is known.

To further evaluate the possibility that the South-East Asian mainland was home to a regional population with phenotypic similarities to the present day Andaman Islanders, and other so-called negrito populations, detailed comparative data are required from Myanmar and the rest of Island South-East Asia.

Future studies would be greatly enhanced by multi-locus data, especially autosomal markers, to provide insights into population histories that are beyond the scope of a single haploid marker. The present study reinforces the need for careful quality control of novel sequence data, and demonstrates the power of using the resulting SNP data to conduct wider population surveys with multiplexed SBE assays. The corrected genealogy for M31/32 obtained in this way affords important insights into the prehistory of the Andaman Islands and advances our understanding of the history of human migrations and settlement of regions from Africa to Australia during the late Pleistocene.

## ACKNOWLEDGMENTS

The authors acknowledge the ministry of Culture, Government of India for funding the national project, “DNA Polymorphism in Contemporary Indian Populations”. They thank two anonymous reviewers and Thomas Kivisild and Richard Villems for helpful suggestions to improve and clarify the manuscript. They are also indebted to the anonymous blood donors, without which this study would not have been possible. Thanks are due to their colleagues involved at various levels in the larger project and Mr. Gopichand, Mr. J.S.J. Rao, Dr. Bandopadhaya, Mr. P. Dhar, and Dr. Satish Kumar, for their extended cooperation. They also thank Dr. K. Thangaraj and his colleagues, especially G. Chaubey, for open discussion regarding potential sequencing errors and their sharing of corrected sequence information prior to their errata publication. They are also grateful to the staff of the Natural History Museum London, the Oxford Museum of Natural History, and the Royal College of Surgeons Edinburgh for research access to recent skeletal material in their collections.

VRR and SSB conceived the project on Jarawas. SSB, RS, BVRP, BNS, and SB collected samples. VRR, AC, JS, PE, and JJS designed the experiments. JS, CHA, DS, and PE performed the experiments. AC, JS, MM, PE, JJS, and SYWH analyzed the data. PE, MM, and AC drafted the manuscript. VRR, RS, BNS commented on the draft. SYWH and JJS improved the draft. MM, PE, and SYWH provided the figures. All the new contemporary DNA samples were analyzed in AnSI labs.

## LITERATURE CITED

- Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, and Young IG (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290:457–65.
- Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 23:147.
- Barnard-Davis J. 1867. *Thesaurus craniorum: catalogue of the skulls of the various races of man, in the collection of Joseph Barnard-Davis*. London: Printed for the subscribers.
- Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, Dey B, Roy M, Roy B, Bhattacharyya N, Roychoudhury S, Majumder P. 2003. Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res* 13:2277–2290.
- Cooper Z. 1989. Analysis of the nature of contracts with the Andaman Islands during the last two millennia. *South Asian Studies* 5:133–147.
- Cooper Z. 2002. *Archaeology and history: early settlements in the Andaman Islands*. New Delhi and Oxford: Oxford University press.
- Cordaux R, Saha N, Bentley G, Aunger R, Sirajuddin S, Stoneking M. 2003. Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. *Eur J Hum Genet* 3:253–264.
- Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88.
- Endicott P, Gilbert M, Stringer C, Lalueza-Fox C, Willerslev E, Hansen A, Cooper A. 2003. The genetic origins of the Andaman Islanders. *Am J Hum Genet* 72:178–184.
- Endicott P, Metspalu M, Stringer C, Macaulay V, Cooper A, Sanchez JJ. 2006. Multiplexed SNP typing of ancient DNA clarifies the origin of Andaman mtDNA haplogroups amongst South Asian tribal populations. *PLoS ONE* 1:e81.

- Excoffier L, Yang Z. 1999. Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Mol Biol Evol* 16:1357–1368.
- Friedlaender J, Schurr T, Gentz F, Koki G, Friedlaender F, Horvat G, Babb P, Cerchio S, Kaestle F, Schanfield M, Deka R, Yanagihara R, Merriwether DA. 2005. Expanding southwest Pacific mitochondrial haplogroups P and Q. *Mol Biol Evol* 22:1506–1517.
- Fucharoen G, Fucharoen S, Horai S. 2001. Mitochondrial DNA polymorphisms in Thailand. *J Hum Genet* 46:115–125.
- Hasegawa M, Horai S. 1991. Time of the deepest root for polymorphism in human mitochondrial DNA. *J Mol Evol* 32:37–42.
- Heyer E, Zietkiewicz E, Rochowski A, Yotova V, Puymirat J, Labuda D. 2001. Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *Am J Hum Genet* 69:1113–1126.
- Hill C, Soares P, Mormina M, Macaulay V, Clarke D, Blumbach PB, Vizuete-Forster M, Forster P, Bulbeck D, Oppenheimer S, Richards M. 2007. A mitochondrial stratigraphy for island southeast Asia. *Am J Hum Genet* 80:29–43.
- Hill C, Soares P, Mormina M, Macaulay V, Meehan W, Blackburn J, Clarke D, Raja JM, Ismail P, Bulbeck D, Oppenheimer S, Richards M. 2006. Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Mol Biol Evol* 23:2480–2491.
- Howell N, Elson JL, Turnbull DM, Herrnstadt C. 2004. African haplogroup L mtDNA sequences show violations of clock-like evolution. *Mol Biol Evol* 21:1843–1854.
- Hudjashov G, Kivisild T, Underhill PA, Endicott P, Sanchez JJ, Lin AA, Shen P, Oefner P, Renfrew C, Villems R, Forster P. 2007. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc Natl Acad Sci USA* 104:8726–8730.
- Ingman M, Kaessmann H, Pääbo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* 408:708–713.
- Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, Parik J, Metspalu E, Adojaan M, Tolk H-V, Stepanov V, Gölge M, Usanga E, Papiha SS, Cinnioglu C, King R, Cavalli-Sforza L, Underhill PA, Villems R. 2003. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet* 72:313–332.
- Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, Coppa A, de Knijff P, Feldman M, Cavalli-Sforza LL, Oefner PJ. 2006. The role of selection in the evolution of human mitochondrial genomes. *Genetics* 172:373–387.
- Kumar V, Langsith BT, Biswas S, Babu JP, Rao TN, Thangaraj K, Reddy AG, Singh L, Reddy BM. 2006. Asian and non-Asian origins of Mon-Khmer- and Mundari-speaking Austro-Asiatic populations of India. *Am J Hum Biol* 18:461–469.
- Kumar V, Reddy M. 2003. Status of Austro-Asiatic groups in the peopling of India: an exploratory study based on the available prehistoric, linguistic and biological evidences. *J Biosci* 28: 507–522.
- Maca-Meyer N, González AM, Larruga JM, Flores C, Cabrera VM. 2001. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet* 2:13.
- Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt HJ, Oppenheimer S, Torroni A, Richards M. 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* 308: 1034–1036.
- Man E. 1885. On the Andaman Islands and their inhabitants. *J Anthropol Inst Great Britain Ireland* 14:235–272.
- Merriwether DA, Hodgson JA, Friedlaender FR, Allaby R, Cerchio S, Koki G, Friedlaender JS. 2005. Ancient mitochondrial M haplogroups identified in the southwest Pacific. *Proc Natl Acad Sci USA* 102:13034–13039.
- Metspalu M, Kivisild T, Bandelt H-J, Richards M, and Villems R. 2006. The pioneer settlement of modern humans in Asia. In: Bandelt H-J, Macaulay V, Richards M, editors. *Human mitochondrial DNA and the evolution of homo sapiens*. Heidelberg: Springer. p 179–197.
- Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, Serk P, Karmin M, Behar DM, Gilbert MTP, Endicott P, Mastana S, Papiha SS, Skorecki K, Torroni A, Villems R. 2004. Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet* 5:26.
- Meyer S, Weiss G, von Haeseler A. 1999. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics* 152:1103–1110.
- Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hoeseni S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace DC. 2003. Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci USA* 100:171–176.
- Morrison K. 2007. Foragers and forager-traders in South Asian worlds: some thoughts from the last 10,000 years. In: Petraglia M, Allchin B, editors. *The evolution and history of human populations in South Asia: inter-disciplinary studies in archaeology, biological anthropology, linguistics and genetics*. New York: Springer. pp 321–340.
- Oota H, Pakendorf B, Weiss G, von Haeseler A, Pookajorn S, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M. 2005. Recent origin and cultural reversion of a hunter-gatherer group. *PLoS Biol* 3:e71.
- Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilineal versus patrilineal residence. *Nat Genet* 29:20–21.
- Palanichamy M, Sun C, Agrawal S, Bandelt H-J, Kong Q-P, Khan F, Wang C-Y, Chaudhuri T, Palla V, Zhang Y-P. 2004. Phylogeny of mtDNA macrohaplogroup N in India based on complete sequencing: implications for the peopling of South Asia. *Am J Hum Genet* 75:966–978.
- Palanichamy MG, Agrawal S, Yao YG, Kong QP, Sun C, Khan F, Chaudhuri TK, Zhang YP. 2006. Comment on “Reconstructing the origin of Andaman islanders.” *Science* 311:470. Author reply.
- Portman M. 1884. *The Andaman Islanders*. Calcutta, India: Office of the Superintendent of Government Printing.
- Quintana-Murci L, Chaix R, Wells S, Behar D, Sayar H, Scozzari R, Rengo C, Al-Zahery N, Semino O, Santachiara-Benerecetti A, Coppa A, Ayub Q, Mohyuddin A, Tyler-Smith C, Mehdi Q, Torroni A, McElreavey K. 2004. Where West meets East: the complex mtDNA landscape of the Southwest and Central Asian corridor. *Am J Hum Genet* 74:827–845.
- Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, McElreavey K, Santachiara-Benerecetti AS. 1999. Genetic evidence of an early exit of homo sapiens from Africa through eastern Africa. *Nat Genet* 23:437–441.
- Rieder MJ, Taylor SL, Tobe VO, Nickerson DA. 1998. Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucl Acids Res* 26:967–973.
- Saillard J, Forster P, Lynnerup N, Bandelt H-J, Nørby S. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet* 67:718–726.
- Sanchez JJ, Endicott P. 2006. Developing multiplexed SNP assays with special reference to degraded DNA templates. *Nat Protocol* 1:1370–1378.
- Sambrook J, Fritsch EF, Maniatis T. 1989. *Molecular cloning: a laboratory manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
- Sun C, Kong QP, Palanichamy MG, Agrawal S, Bandelt HJ, Yao YG, Khan F, Zhu CL, Chaudhuri TK, Zhang YP. 2006. The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. *Mol Biol Evol* 23:683–690.
- Thangaraj K, Chaubey G, Kivisild T, Reddy AG, Singh VK, Rasalkar AA, Singh L. 2005. Reconstructing the origin of Andaman Islanders. *Science* 308:996.

- Thangaraj K, Chaubey G, Kivisild T, Reddy AG, Singh VK, Rasalkar AA, Singh L. 2006a. Response to comment on "Reconstructing the origin of Andaman islanders." *Science* 311:470. Author reply.
- Thangaraj K, Chaubey G, Singh V, Vanniarajan A, Thanseem I, Reddy A, Singh L. 2006b. *In situ* origin of deep rooting lineages of mitochondrial Macrohaplogroup 'M' in India. *BMC Genomics* 7:151.
- Thangaraj K, Singh L, Reddy A, Rao V, Sehgal S, Underhill P, Pierson M, Frame I, Hagelberg E. 2003. Genetic affinities of the Andaman islanders, a vanishing human population. *Curr Biol* 13:86–93.
- Thanseem I, Thangaraj K, Chaubey G, Singh V, Bhaskar L, Reddy M, Reddy A, Singh L. 2006. Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. *BMC Genet* 7:42.
- Wen B, Li H, Gao S, Mao X, Gao Y, Li F, Zhang F, He Y, Dong Y, Zhang Y, Huang W, Jin J, Xiao C, Lu D, Chakraborty R, Su B, Deka R, Jin L. 2005. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. *Mol Biol Evol* 22:725–734.
- Wen B, Xie X, Gao S, Li H, Shi H, Song X, Qian T, Xiao C, Jin J, Su B, Lu D, Chakraborty R, Jin L. 2004. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. *Am J Hum Genet* 74:856–865.
- Yao Y-G, Nie L, Harpending H, Fu Y, Yuan Z-G, Zhang Y-P. 2002. Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. *Am J Phys Anthropol* 118:63–76.
- Yao Y-G, Zhang Y-P. 2002. Phylogeographic analysis of mtDNA variation in four ethnic populations from Yunnan Province: new data and a reappraisal. *J Hum Genet* 47:311–318.