MBE Advance Access published July 28, 2012

Evolutionary History of Continental Southeast Asians: "Early Train" Hypothesis Based on Genetic Analysis of Mitochondrial and Autosomal DNA Data

Timothy A. Jinam,[1,2] Lih-Chun Hong,[3] Maude E. Phipps,[4] Mark Stoneking,[5] Mahmood Ameen,[3] Juli Edo,[6] HUGO Pan-Asian SNP Consortium,[7] and Naruya Saitou*[1,2]

[1]Department of Genetics, The Graduate University for Advanced Studies (SOKENDAI), Mishima, Japan

[2]Division of Population Genetics, National Institute of Genetics, Mishima, Japan

[3]Department of Molecular Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia

[4]Jeffrey Cheah School of Medicine and Health Sciences, Monash University (Sunway Campus), Selangor, Malaysia

[5]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[6]Department of Anthropology, Faculty of Arts and Social Sciences, University of Malaya, Kuala Lumpur, Malaysia

[7]Human Genome Organisation, Singapore

***Corresponding author:** E-mail: saitounr@lab.nig.ac.jp.

Associate editor: Anne Stone

The complete mtDNA sequences for all 86 individuals have been submitted to DDBJ/EMBL/Genbank (accession nos. AP012346–AP012431).

Abstract

The population history of the indigenous populations in island Southeast Asia is generally accepted to have been shaped by two major migrations: the ancient "Out of Africa" migration ~50,000 years before present (YBP) and the relatively recent "Out of Taiwan" expansion of Austronesian agriculturalists approximately 5,000 YBP. The Negritos are believed to have originated from the ancient migration, whereas the majority of island Southeast Asians are associated with the Austronesian expansion. We determined 86 mitochondrial DNA (mtDNA) complete genome sequences in four indigenous Malaysian populations, together with a reanalysis of published autosomal single-nucleotide polymorphism (SNP) data of Southeast Asians to test the plausibility and impact of those migration models. The three Austronesian groups (Bidayuh, Selatar, and Temuan) showed high frequencies of mtDNA haplogroups, which originated from the Asian mainland ~30,000–10,000 YBP, but low frequencies of "Out of Taiwan" markers. Principal component analysis and phylogenetic analysis using autosomal SNP data indicate a dichotomy between continental and island Austronesian groups. We argue that both the mtDNA and autosomal data suggest an "Early Train" migration originating from Indochina or South China around the late-Pleistocene to early-Holocene period, which predates, but may not necessarily exclude, the Austronesian expansion.

Key words: Austronesian, Negrito, mitochondrial DNA, Southeast Asia, Orang Asli.

Introduction

The Southeast Asian region is home to a rich variety of human populations, each with their own ethnic cultures and traditions. The history and diversity of the various indigenous groups that still populate the region had been described by using archaeological, linguistic, and most recently, genetic data. Archaeological evidence points to the presence of modern humans in Southeast Asia at least 40,000 years before present (YBP) (Brothwell 1960; Barker et al. 2007). Those early migrants eventually reached the Sahul landmass, which today is split into Papua New Guinea, Australia, and Tasmania (Leavesley and Chappell 2004; O'Connell and Allen 2004). That ancient wave of migration was believed to have brought the ancestors of several "Australoid" populations found in Southeast Asia and Australia. These include the Papuans and Australian Aboriginals, as well as several groups in the Andaman, Philippines, and West Malaysia, which are collectively known as Negritos (Cavalli-Sforza et al. 1994). Up until the Last Glacial Maximum approximately 20,000 YBP, the current islands of Sumatra, Java, and Borneo were joined together to the Asian mainland, forming a landmass known as Sundaland (Glover and Bellwood 2004), which was separated from the Sahul landmass by multiple islands collectively referred to as Wallacea; Wallace's line separates Wallacea and Sundaland (fig. 1).

Another significant epoch with respect to human migrations in island Southeast Asia occurred much later during the mid-Holocene period (5,000–7,000 YBP). With Taiwan as the probable starting point, this migration wave eventually spread southward throughout island Southeast Asia and into the islands of Oceania, bringing with it agriculture, domesticated livestock, and Austronesian languages (Bellwood 2005). This "Out of Taiwan" migration was largely supported by

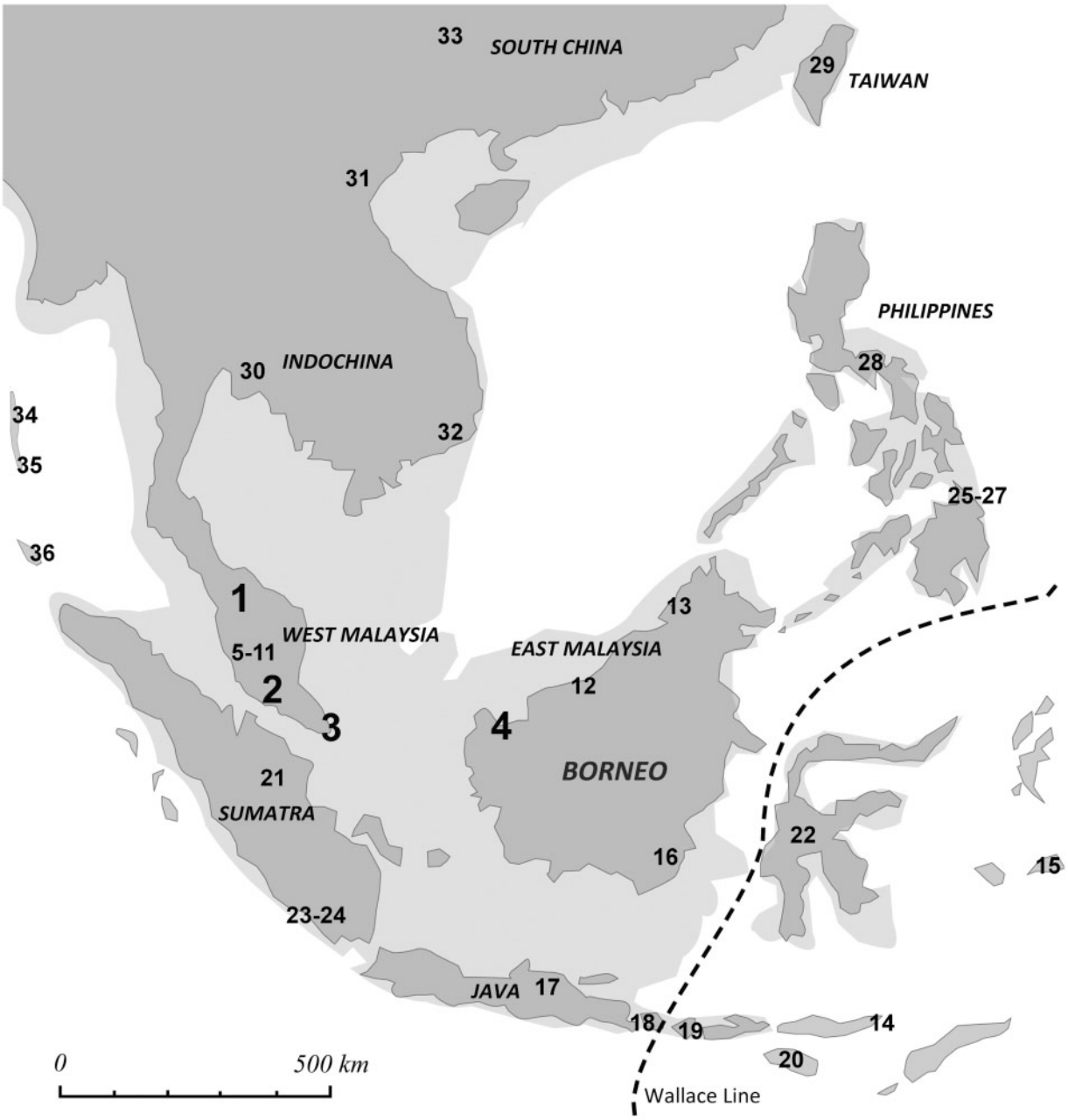

FIG. 1. A map of Southeast Asia. Numbers indicate locations of populations listed in supplementary table S1, Supplementary Material online. Areas shaded light gray indicate the extent of the landmass up to the Last Glacial Maximum.

archaeological data (Bellwood 2005, 2007) and the linguistic phylogeny of Austronesian languages (Diamond 1988; Gray and Jordan 2000). This migration model, termed the "express train," assumes a rapid expansion from Taiwan to Polynesia with little or no admixture between the expanding and extant populations. This model was later expanded to involve a series of pulses and pauses but remains fundamentally similar (Gray et al. 2009). On the other side of the discussion is the possible origin in island Southeast Asia. The "Southeast Asian" model posits complex interactions between Melanesians, Southeast Asians, and Polynesians and ultimately an Austronesian origin probably in East Indonesia instead of Taiwan (Terrell 1988; Oppenheimer and Richards 2001).

Genetic data have been used to argue for and against those two models: the Taiwanese or Southeast Asian origin of Austronesians. The "Southeast Asian origin" model was initially based on the age estimates of mitochondrial DNA (mtDNA) haplogroup B4a1a, which was found at high frequencies in the Polynesians, hence the name Polynesian motif (Oppenheimer and Richards 2001). This was later supplemented by a study based on mtDNA haplogroup E distribution, which proposed a dispersal originating from insular Southeast Asia driven by climate change (Soares et al. 2008). Genetic evidence for the "express train" model, which predicts a rapid expansion from Taiwan with little or no admixture with extant populations has been lacking. Instead, most genetic studies advocate a slower movement,

which allows for sex-biased admixture between migrants of Asian ancestry (possibly from Taiwan) with existing Melanesian populations with respect to Polynesian origins (Kayser et al. 2000, 2008; Wollstein et al. 2010; Mirabal et al. 2012). In addition, some studies offered a differing perspective: possible earlier migration(s) from the Asian mainland during the late-Pleistocene to early-Holocene period, which predates the Austronesian expansion, based on mtDNA (Hill et al. 2006, 2007) and Y-chromosomal (Karafet et al. 2010) analyses of island Southeast Asian populations.

Analysis of autosomal DNA markers in Southeast Asian populations tended to be scattered and very limited in contrast to analyses using uniparental mtDNA and Y-chromosomal markers. However, the development of high-throughput single-nucleotide polymorphism (SNP) genotyping platforms allowed for more comprehensive genetic studies such as that by the Pan-Asian SNP Consortium (PASNP) (HUGO PASNP Consortium 2009). Although the article mainly focused on the migration histories of East Eurasian populations as a whole, the data generated should allow for a more focused analysis of Southeast Asian populations.

In addressing questions regarding the origin and genetic diversity of Austronesians, genetic studies using both uniparental markers and autosomal SNP data mostly focused on Polynesian populations. A more comprehensive examination of both mtDNA and autosomal SNP diversity in island Southeast Asian populations may provide more insight into the origins and migration patterns of humans in Southeast Asia. Given this current backdrop, we conducted an in-depth genetic analysis of Southeast Asian populations using available SNP data from PASNP (HUGO PASNP Consortium 2009) and HGDP-CEPH panel database (Li et al. 2008), together with newly generated complete mtDNA sequences in four indigenous Malaysian groups. These include three Austronesian groups (Temuan, Seletar, and Bidayuh) and a Negrito (Jehai). The Negritos from West Malaysia, who are also referred to as Semang, currently speak Austro-Asiatic languages and may be the descendants of ancient migrants to the Southeast Asian region. By exploring the diversity of the maternally inherited mtDNA and genome-wide autosomal SNP in these four groups and comparing them with other populations within the Southeast Asian region, we attempt to shed light on some questions regarding their demographic and migration histories.

Materials and Methods

Population Samples

DNA samples from the Jehai, Temuan, and Bidayuh were previously collected on several occasions as part of studies on general health among indigenous communities (Jinam et al. 2008) and genetic analysis of autosomal markers (HUGO PASNP Consortium 2009; Jinam et al. 2010). We also included samples collected from a Proto Malay subpopulation called Seletar. Based on their family information obtained during field sampling, only samples from unrelated individuals were used in genetic analyses. The Jehai, Temuan, and Seletar represent the *Orang Asli* groups from West Malaysia, whereas the Bidayuh is one of the many indigenous groups from East Malaysia on the island of Borneo. The geographical locations for the Jehai, Temuan, Seletar, and Bidayuh are 1–4 as depicted in figure 1. This study was approved by the respective institutional review boards of the National Institute of Genetics in Japan, University of Malaya in Malaysia, Max Planck Institute for Evolutionary Anthropology in Germany, Ministry of Health Malaysia, Monash University in Malaysia, and the Department of Indigenous Affairs (Jabatan Kemajuan Orang Asli Malaysia, JAKOA). Informed consent was obtained from all participants.

Complete mtDNA Genome Sequencing

Sequencing of mtDNA complete genomes was performed in a total of 68 samples (8 Temuan, 22 Jehai, 17 Bidayuh, and 21 Seletar) using 11 pairs of polymerase chain reaction (PCR) primers and 32 sequencing primers from Torroni et al. (2001). A slight modification to their protocol involved optimizing annealing temperatures for all PCR reactions to 60 °C, instead of 55 °C. PCR products were purified using ExoSAP-IT reagent before being subjected to sequencing reactions using the BigDye Terminator kit (Applied Biosystems). Capillary separation was performed on the ABI3130xl Genetic Analyzer (Applied Biosystems). For each sample, the resulting 32 traces were aligned to the revised Cambridge Reference Sequence (Genbank ID NC_012920) using CodonCode Aligner software (CodonCode Corporation, Dedham, MA, USA) to obtain the consensus sequence. In addition, complete mtDNA sequences from 18 samples (10 Temuan, 2 Jehai, and 6 Bidayuh) were generated using a high-throughput sequencing platform as described previously (Gunnarsdottir et al. 2011a), resulting in a total of 86 complete mtDNA genome sequences from four groups.

Nucleotide Sequence Data Analysis

Nucleotide sequences of all mtDNAs determined in this study were assigned to mtDNA haplogroups according to nomenclature at http://www.phylotree.org (last accessed April 2011; see van Oven and Kayser 2009). mtDNA haplogroup frequencies from the Kensiu, another Negrito subgroup from West Malaysia (Hong LC, Fong MY, Phipps ME, unpublished data), and other Southeast Asian populations listed in supplementary table S1, Supplementary Material online, were used for principal component analysis (PCA) using R software package (http://www.R-project.org; last accessed April 2009).

Coding region sequences (nucleotide positions 577–16,023) were extracted from the newly determined complete mtDNA sequences and also from available literature (supplementary table S1, Supplementary Material online). A neighbor-joining tree (Saitou and Nei 1987) was generated using MEGA software version 5 (Tamura et al. 2011) with 500 bootstrap replications. Using a subset of sequences that represent the haplogroups found in our samples in addition with a chimpanzee mtDNA sequence (Genbank ID D38113; Horai et al. 1992) as an outgroup, we estimated the mutation rate

and age of haplogroups using a Bayesian Markov chain Monte Carlo (MCMC) method as implemented in the BEAST software (Drummond and Rambaut 2007). A divergence time of 6.5 million years between humans and chimpanzee (Goodman et al. 1998; Mishmar et al. 2003) was used as a calibration point. Using Tamura and Nei (1993) substitution model and assuming a strict molecular clock, the mutation rate was estimated using a normally distributed prior with a mean of 1.71×10^{-8} for coding-region sequences (Soares et al. 2009). The trees were generated on a run of 40,000,000 steps, sampling every 4,000 steps, and the first 4,000,000 steps were regarded as burn-in. Bayesian skyline plots (BSP) were generated for the Jehai, Temuan, Seletar, and Bidayuh groups using the above parameters but with a coalescent-based tree prior with a piecewise linear model. A maximum-likelihood (ML) tree was also constructed using the same set of coding-region sequences as above to estimate the time depth of mtDNA haplogroups. Using the MEGA5 software, the Tamura and Nei (1993) substitution model was implemented, and two mutation rates were used for molecular clock calibration: 1.71×10^{-8} (Soares et al. 2009) and 1.36×10^{-8}, which was estimated using Bayesian MCMC as described earlier.

Genome-Wide SNP Data Analysis

Genotype data from the Jehai, Kensiu, Temuan, and Bidayuh were retrieved from the Pan-Asian SNP database (http://www4a.biotec.or.th/PASNP, last accessed April 2011). The geographical locations of the PASNP populations are depicted in supplementary figure S1, Supplementary Material online. In addition, other Southeast Asian populations from the PASNP (HUGO PASNP Consortium 2009) and HGDP-CEPH (Li et al. 2008) data sets listed in supplementary table S2, Supplementary Material online, were also retrieved. There were 12,150 SNPs that were common to both data sets. To assess the relatedness between individuals, PCA was conducted using the *smartpca* program in the EIGENSOFT software package (Patterson et al. 2006) by using the SNP genotype data.

The populations were further grouped into 17 broader categories listed in supplementary table S3, Supplementary Material online, for phylogenetic tree and network analysis. PHYLIP software (Felsenstein 2005) was used to generate 5,000 bootstrap replicates of SNP allele frequencies, and genetic distance matrices between populations were calculated using Nei's (1972) standard genetic distance. An unrooted Neighbor-Joining tree and a phylogenetic network based on the Neighbor-Net method was constructed from the pairwise genetic distances between populations using the SplitsTree4 software (Huson and Bryant 2006).

Results

mtDNA Analysis

Sequence Diversity and Haplogroup Distribution

Complete mtDNA sequences were newly determined from 86 individuals (24 Jehai, 18 Temuan, 21 Seletar, and 23 Bidayuh). The summary statistics for the sequence variation in those groups are listed in table 1. The highest haplotype diversity was observed in the Temuan, whereas the lowest was in the Seletar, in which only five distinct mtDNA haplotypes were observed. Negative values for Tajima's D test were observed in the Bidayuh, Seletar, and Temuan, suggesting a history of population expansion. However, the *P* values did not indicate statistically significant deviation from expectation assuming a constant population size.

All individuals were assigned to specific haplogroups belonging to M and N macrohaplogroups by following the nomenclature in www.phylotree.org as much as possible. A total of 23 haplogroups were observed, and the specific mutations that define M and N haplogroups are shown in supplementary figures S2 and S3, Supplementary Material online, respectively. In addition to haplogroup-defining mutations, additional mutations that were population specific were observed. For example, additional mutations in haplogroup N9a6a differ between the Bidayuh, Jehai, and Temuan (supplementary fig. S3, Supplementary Material online). In most cases, the haplogroups observed have been reported previously, but there are instances where some haplotypes share only the basal mutations with known haplogroups, and additional mutations did not match any existing ones. Those haplotypes were therefore assigned to the closest basal haplogroup, for example, B4a and F1a'c in supplementary figure S3, Supplementary Material online. The mtDNA haplogroup frequencies of the four populations and the Kensiu (Hong LC, Fong MY, Phipps ME, unpublished data) are presented in table 2. The most frequent haplogroup in the Bidayuh is N9a6a, whereas in the Temuan, the most frequent haplogroups are M21a, N22, and N21, which are lineages that branch off directly from basal M and N haplogroups (Macaulay et al. 2005). The low nucleotide diversity observed in the Seletar was further demonstrated by the limited

Table 1. Summary statistics for complete mtDNA sequences in four Malaysian groups.

Statistics	Bidayuh	Jehai	Seletar	Temuan
No. of sequences	23	24	21	18
No. of haplotypes	13	11	5	14
Haplotype diversity ± SD	0.88 ± 0.05	0.89 ± 0.04	0.54 ± 0.11	0.94 ± 0.05
Mean no. pairwise difference	30.5	32.1	15.9	29.1
No. of polymorphic sites	157	87	65	135
Nucleotide diversity	0.00184	0.00194	0.00096	0.00175
Tajima's D (*P* value)	−1.14495 (NS)	1.4924 (NS)	−0.49193 (NS)	−1.09762 (NS)

NOTE.—NS, not significant; SD, standard deviation.

number of observed haplogroups and the very high frequency of one particular haplogroup, N9a6 at 71%. Haplogroups M21a and R21 are the most frequent in the Jehai and Kensiu, respectively, who are both Negrito groups.

Table 2. mtDNA haplogroup frequencies (%) in five indigenous Malaysian groups.

Haplogroup	Bidayuh	Jehai	Seletar	Temuan	Kensiu[a]
G1c			4.8		
M20	4.3				
M74b	4.3				
M21a		37.5		27.8	43.2
M22a				5.6	
M7b1				5.6	
M7c2				5.6	
M7c3c	8.7				2.7
E1b	4.3		19.0		
B4a	4.3				
B4a1a1a	4.3				
B4b1a2a				5.6	
B4c2			4.8		
B5a					2.7
B5b2	4.3				
B6				5.6	
F1a'c	30.4				
F1a1a		12.5			2.7
F1a1a1		8.3			
R21		25.0			43.2
N21				22.2	
N22				16.7	
N9a6			71.4		
N9a6a	34.8	16.7		5.6	5.4
Total individuals	23	24	21	18	37

[a]Hong LC, Fong MY, Phipps ME, unpublished data.

Relationship with Other Populations

To further elucidate the relationship between the indigenous Malaysian populations and other surrounding populations, PCA was performed using haplogroup frequencies from this study and from selected populations from the literature (supplementary table S1, Supplementary Material online). The resulting PCA plot is shown in figure 2. We added population ID after the name of each population in this paragraph for a better understanding of population identity. A clear division appears along the first principal component (PC) on the *X* axis, which places the Negrito populations from West Malaysia (Jehai_1, Kensiu_5, Batek_6, and Mendriq_7) at one end and most other Austronesian-speaking populations from Southeast Asia at the other end. On the second PC on the *Y* axis, there appears to be a geographical divide between the Austronesians groups. Populations in West Malaysia, Sumatra, and Java (Malay_11, Jakun_10, Java_17, Bali_18, and Lombok_19) tend to cluster with mainland or continental groups (Thai_30, Vietnamese_31, South Chinese_33, and Cham_32). On the other hand, populations from Taiwan, Philippines, and other Indonesian islands to the east (Alor_14, Ambon_15, and Sulawesi_22) tend to group together. It seems that PC1 represents a Negrito-Austronesian divide, whereas PC2 corresponds to a continental-island division of Austronesian groups. However, not all population affinities fall nicely into these two generalized trends.

The Temuan_2 in West Malaysia are Austronesian speakers and are physically distinct from the Malaysian Negritos, but they clustered with them in the PCA plot of figure 2. This may be due to the high frequencies of haplogroup M21a in both the Temuan and Negritos, suggesting either gene flow between these populations or parallel increases of this haplogroup frequency. The populations from Borneo (Bidayuh_4, Iban_12, Banjarmasin_16, and Kota_Kinabalu_13) also displayed somewhat irregular patterns. The Iban_12 and

Fig. 2. PCA plot based on haplogroup frequencies. Numbers in circles are population labels as listed in supplementary table S1, Supplementary Material online. Red circles are populations from current study and black circles are from literature.

Banjarmasin_16 appeared closer to the continental clusters, whereas Kota_Kinabalu_13 appear closer to the island cluster. The Bidayuh_4 clustered with two other Proto-Malay groups (Seletar_3 and Semelai_9) from West Malaysia. The kinship structure may also have a significant bearing on mtDNA diversity, as shown by the Besemah_23 and Semende_24 of Sumatra. The Semende tribe is matrilocal and is closer to the continental populations, whereas the Besemah tribe is patrilocal and is closer to island Southeast Asians. This may suggest that the mtDNA diversity in the Besemah is shaped by female migrations from island Southeast Asia.

Phylogenetic Analysis and Coalescence Time Estimation

The Neighbor-Joining tree for haplogroups M and N are shown separately in supplementary figure S4A and S4B, Supplementary Material online, respectively. Estimation of the mutation rate for the mtDNA coding region using Bayesian MCMC analysis resulted in a mean value of 1.36×10^{-8}, with a 95% highest posterior density range of 1.07×10^{-8}–1.65×10^{-8}, substitutions per site per year. The mean value of the mutation rate was used to calibrate the molecular clock of the ML tree. In addition, we used the mutation rate of 1.71×10^{-8} as reported by Soares et al. (2009). Age estimates using the ML tree and Bayesian methods were based on the coalescence time of all mtDNA sequences that belong to the same haplogroup. In general, the higher mutation rate of Soares et al. (2009) resulted in younger age estimates of haplogroups compared with the ages obtained using the mutation rate estimated from our own data. The haplogroup ages obtained from the Bayesian MCMC analysis tended to give slightly older time frames compared with the ML estimates and this may be attributed to the differences between Bayesian and ML methods. The resulting ML tree using the mutation rate of 1.36×10^{-8} is shown in figure 3, whereas the other age estimates (ML and Bayesian) are listed in table 3.

Diversity of M Haplogroup Lineages

The M haplogroups observed in this study included those which were considered indigenous to the *Orang Asli*,

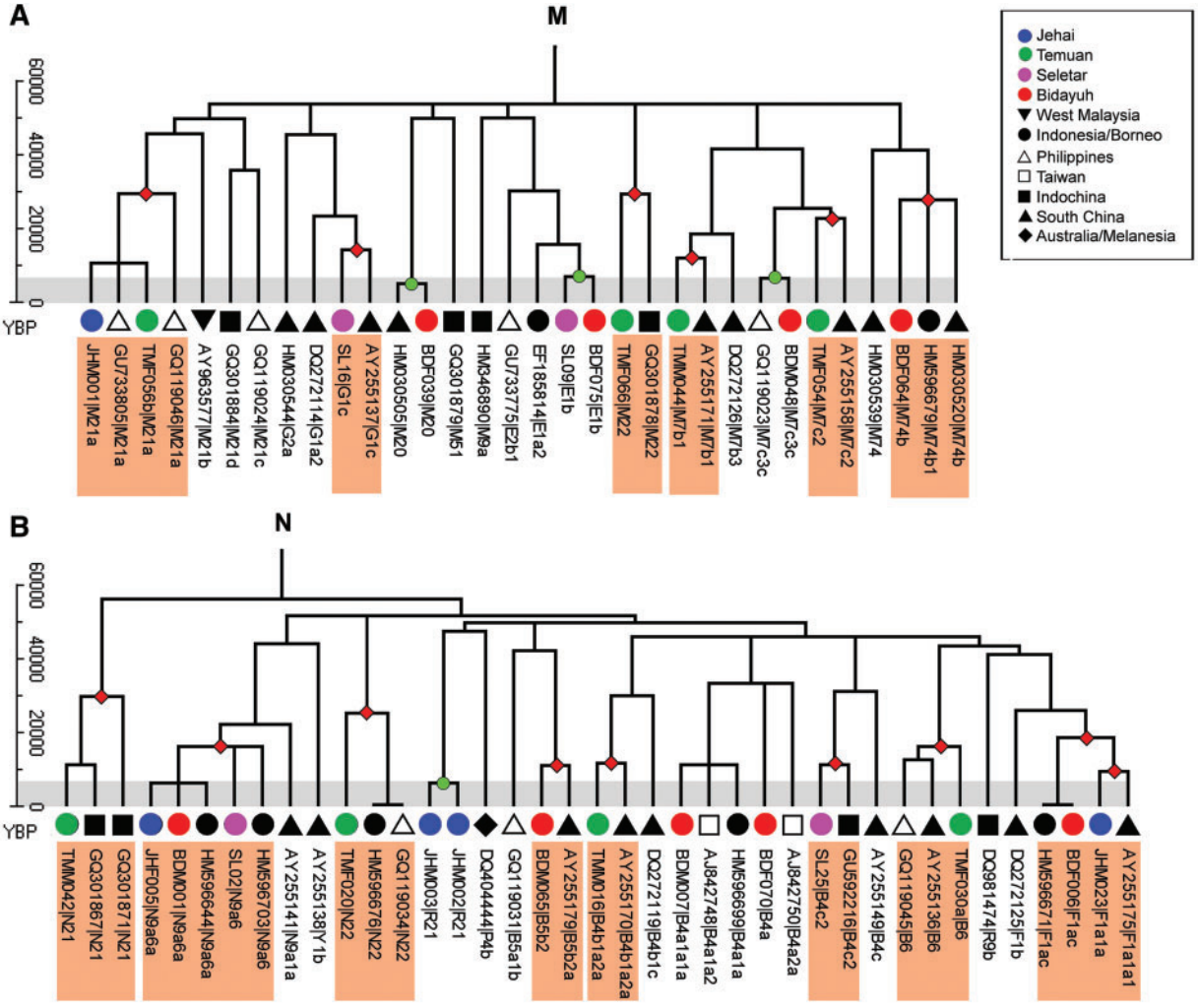

Fig. 3. ML tree constructed using mtDNA coding-region sequences. The molecular clock was calibrated with a mutation rate of 1.36×10^{-8} substitutions per site per year. Gray horizontal bar represents time frame of the Austronesian expansion from 7,000 years ago, and haplogroups with coalescent times within that period are indicated with a green dot. Haplogroups that support the "early train" hypothesis are indicated with red diamonds and orange boxes. (*A*) Subtree of haplogroup M lineages. (*B*) Subtree of hapologroup N lineages.

Table 3. Age estimates of selected haplogroups based on mtDNA coding-region sequences using ML and Bayesian MCMC methods.

Haplogroup	ML	MCMC (95% highest posterior density interval)
M	41,300–51,900	79,000 (59,000–98,700)
M20	4,400–5,500	8,500 (1,600–16,900)
M21a	22,600–28,400	31,100 (16,900–46,600)
M22	22,500–28,500	34,400 (17,500–51,500)
M74	31,300–39,300	52,300 (34,100–70,300)
M74b	20,400–25,600	35,400 (20,800–49,600)
M7c2	17,800–22,400	27,600 (14,300–41,500)
M7c3c	4,900–6,200	9,300 (900–19,200)
M7b1	12,900–16,300	18,600 (6,800–30,600)
E	23,200–29,000	36,300 (20,500–52,100)
E1b	5,100–6,400	8,200 (1,500–15,700)
G1c	10,700–13,500	15,100 (5,400–25,500)
N	38,800–48,000	78,500 (58,400–99,200)
N21	22,900–28,700	35,400 (16,800–55,600)
N9a6	11,500–14,400	17,200 (8,500–26,900)
N9a6a	4,900–6,100	8,200 (2,600–14,900)
N22	19,300–24,100	27,300 (13,700–42,200)
R	30,500–38,200	64,600 (46,100–86,300)
R21	4,100–5,100	8,600 (1,500–17,100)
B5	28,600–35,800	54,700 (37,100–73,500)
B5b2	8,900–11,100	16,300 (5,800–27,100)
B4a	24,100–30,100	37,100 (23,700–52,500)
B4a1a1a	8,100–10,000	12,300 (3,200–22,200)
B4b	21,700–27,200	34,600 (20,600–50,500)
B4b1a2	8,100–10,200	13,200 (4,800–23,000)
B4c	22,900–28,600	36,000 (20,400–53,000)
B4c2	8,400–10,500	13,300 (3,400–24,100)
B6	11,700–14,700	20,900 (9,600–33,500)
F1	19,100–23,800	29,900 (17,700–42,100)
F1a'c	13,800–17,300	21,100 (11,100–31,500)
F1a1a	6,900–8,600	10,300 (3,500–18,200)

The range of dates obtained by ML method was obtained by using the mutation rate reported by Soares et al. (2009) (lower limit) and the rate obtained by Bayesian MCMC estimation (upper limit).

namely M21a and M22 (Macaulay et al. 2005; Hill et al. 2006). M21a was most frequent in the Temuan and Jehai, as well as other Negrito subgroups in West Malaysia (Hill et al. 2006). Outside of West Malaysia, M21a was also present in appreciable frequencies in the Sakai (also a Negrito group) and in the Chiang Mai population from Thailand (Fucharoen et al. 2001) and very rarely in some Philippine populations (Tabbada et al. 2009; Gunnarsdottir et al. 2011a). The other M21 subtypes, M21b and M21c, which were reported at low frequencies in the *Orang Asli* (Hill et al. 2006) but frequent in the Moken of Myanmar (Dancause et al. 2009), were not observed in any of our current samples. M22 was earlier reported in the Proto Malays (Macaulay et al. 2005), and recent reports showed that it was present in the Vietnamese (Peng et al. 2010) and Southern Chinese (Kong et al. 2011) but has so far not been reported in any island Southeast Asians (Hill et al. 2007).

Haplogroup E, which was proposed to be a marker for postglacial expansion centering in Island Southeast Asia (Soares et al. 2008), was found in the form of E1b in the Seletar and Bidayuh. Haplogroup M7 lineages that are present in the Malaysian samples included M7c3c in the Bidayuh. This haplogroup seems to be restricted to Southeast Asia and was suggested to be a marker for the Austronesian expansion during mid-Holocene (Hill et al. 2007), consistent with our age estimates. Other M7 lineages found in the Temuan include M7b1 and M7c2, and they coalesce with lineages from the mainland (Kong et al. 2003) (fig. 3). We also observed several haplogroups that have not been reported in any Southeast Asian population to date. These included G1c in the Seletar and M74b and M20 in the Bidayuh. G1c was earlier reported in Koreans (Derenko et al. 2007) and Han Chinese (Kong et al. 2003). The ancestral M74a haplotype was reported in southern Chinese populations (Kong et al. 2011), whereas a derived type M74b was found in the Bidayuh in Borneo and Hani of south China (Kong et al. 2011). The M74b1 subtype has been found in Surigaonon and Mamanwa in the Philippines (Gunnarsdottir et al. 2011a; reported as M*) and also in the Besemah in Sumatra (Gunnarsdottir et al. 2011b, reported as M4). Figure 3

shows that the deepest branch of M74 is found in a Southern Chinese, whereas the subgroups M74b were found in the Bidayuh and Besemah, suggesting a dispersal originating from southern China and into island Southeast Asia. Haplogroup M20 found in one Bidayuh individual coalesces with the sequence found in a southern Chinese group (Kong et al. 2011), and these two M20 lineages clustered with haplogroup M51 found in the Cham of Vietnam (Peng et al. 2010) and the Besemah in Sumatra (Gunnarsdottir et al. 2011b), as shown in supplementary figure S4A, Supplementary Material online.

Diversity of N Haplogroup Lineages

As with haplogroup M, we found rare N lineages, which were previously only reported in the *Orang Asli*, namely N21, N22, and R21 (Hill et al. 2006, 2007). N21 lineages in the Temuan appeared to be derived from an ancestral type found in the Cham of Vietnam (supplementary figure S4, Supplementary Material online), implying an origin in Indochina during late Pleistocene based on our age estimates. N22 appears to be limited to the Temuan, as observed in this study and by Hill et al. (2006), although it also appears in very low frequencies in the Philippines (Tabbada et al. 2009), Sumatra (Gunnarsdottir et al. 2011b), and Sumba islands (Hill et al. 2007). Haplogroup R21 appears to be limited to Negrito populations in West Malaysia, although it was also found at appreciable frequencies in the Senoi (Hill et al. 2006), who are thought to have arrived from Indochina (Glover and Bellwood 2004). Haplogroup N9a is widespread in East Asia, but the subclade N9a6 appears to be restricted to island Southeast Asian populations where it is found at low frequencies in Sumatra and Java, Indonesia, but not in the Philippines or Taiwan (Hill et al. 2007). However, we found N9a6 and its daughter clade N9a6a to be quite frequent in the Malaysian groups, particularly in the Bidayuh and Seletar.

Haplogroup B, which is characterized by a 9-bp deletion at position 8272, is fairly common in island Southeast Asia and particularly in Polynesia. The distribution of this haplogroup is varied among the Malaysian populations, with B4a and B5b found in the Bidayuh, B4b and B6 in the Temuan, and B4c in the Seletar. The two B4a lineages in the Bidayuh included B4a1a1a, also known as the Polynesian motif, and it might reflect recent gene flow from the Pacific during the mid-Holocene period (Soares et al. 2011). The other is an undefined B4a haplogroup, which shares the same basal mutations as B4a but could not be further designated to any of its daughter clades. The branching patterns of the NJ tree (supplementary fig. S4, Supplementary Material online) show that the ancestral types of haplogroups B4b, B4c, and B5b were found in South Chinese populations, suggesting an origin in the mainland and dispersal to island Southeast Asia. Interestingly, the B4c2 haplogroup found in the Seletar was also extracted from ancient Negrito hair samples (Ricaut et al. 2006), indicating a diffusion from the mainland during the late Pleistocene. Haplogroup F is another common clade in Southeast Asia, with F1a1a previously reported to be frequent in the Temiar, a Senoi group (Hill et al. 2006). Haplogroup F1a'c shares the same basal mutations as F1a except at nucleotide position 4086 and is present in the Bidayuh, as well as the Besemah and Semende of Sumatra (Gunnarsdottir et al. 2011b).

Changes in Effective Population Size

The BSP, which were generated using coding-region sequences, are shown in supplementary figure S6, Supplementary Material online. When all 86 sequences were analyzed together (supplementary fig. S6A, Supplementary Material online), the observed pattern is that of an increase in population size from approximately 60,000 to 40,000 YBP. What appears to be a stable population size from 30,000 to 10,000 YBP was then followed by a decline which lasted until several hundred YBP. A similar pattern was observed when the Jehai (Negrito) was omitted (supplementary fig. S6B, Supplementary Material online) or when all four populations were analyzed separately (supplementary fig. S6C–F, Supplementary Material online). However, in the case of Temuan (supplementary fig. S6D, Supplementary Material online) and Seletar (supplementary fig. S6E, Supplementary Material online), there were no indications of population size increase from 40,000 YBP or older. Although the pattern of population increase from 60,000 to 40,000 YBP may suggest signals of population expansion as demonstrated in other worldwide populations (Atkinson et al. 2008; Fagundes et al. 2008), any interpretations should be taken with caution given the large confidence intervals for the estimates. A consistent pattern that appeared was that of a population size decrease from 10,000 YBP, and similar patterns were also observed in some Philippine populations (Gunnarsdottir et al. 2011a). The BSP plots also showed a trend of increasing population size in all four groups ~1,000 YBP. The underlying cause for the observed patterns can only be speculated and as such would warrant further investigation.

Genome-Wide SNP analysis

The results of PCA analysis using genome-wide autosomal SNP markers are shown in figure 4. In figure 4A where all Southeast Asian populations were analyzed, the first PC (PC1) distinguishes between the Melanesians and Southeast Asians. The Indonesians from the Alor Island and surrounding islands seem to be intermediate between the Melanesians and other Southeast Asians. The second PC (PC2) separates the Malaysian Negritos from other populations, whereas the third PC (PC3) in supplementary figure S7, Supplementary Material online, showed that PC3 separates the Philippine Negritos from the other populations. A recurring pattern observed in the Malaysian Negritos, Philippine Negritos, and Alorese is that the individuals are spread apart in a gradient, or comet-like pattern, suggesting recent admixture between these groups with Thai, Chinese, or other Austronesians who form a tight cluster in the PCA plot. This comet-like pattern was also observed in PCA analyses of Patterson et al. (2006), Bryc et al. (2010), and McEvoy et al. (2010), who all interpreted the comet-like patterns as results of recent admixtures.

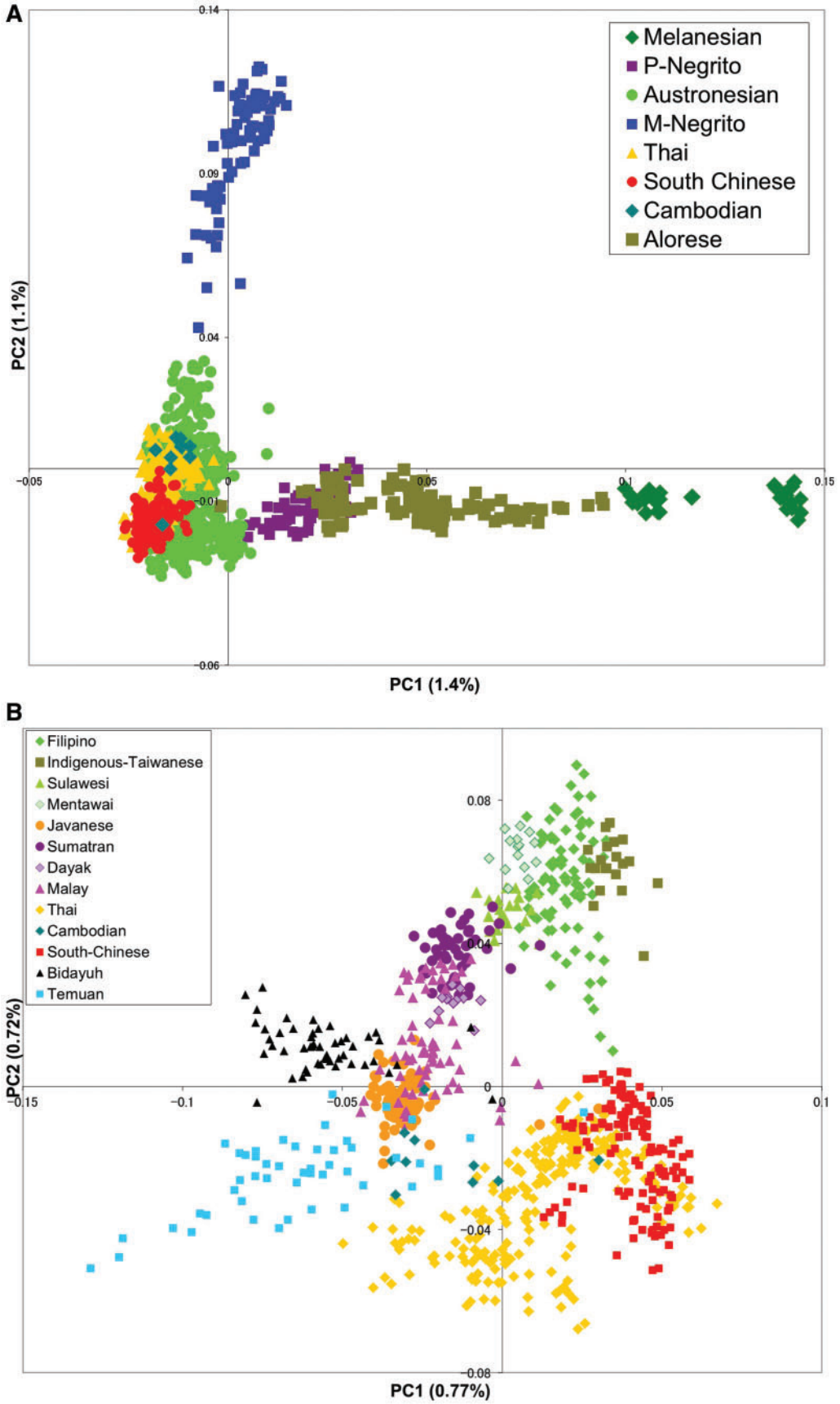

Fig. 4. Individual-based PCA using SNP genotype data. (*A*) PCA plot using all individuals listed in supplementary table S2, Supplementary Material online. (*B*) PCA plot after excluding Malaysian Negrito, Philippine Negrito, Melanesian, and Alorese individuals.

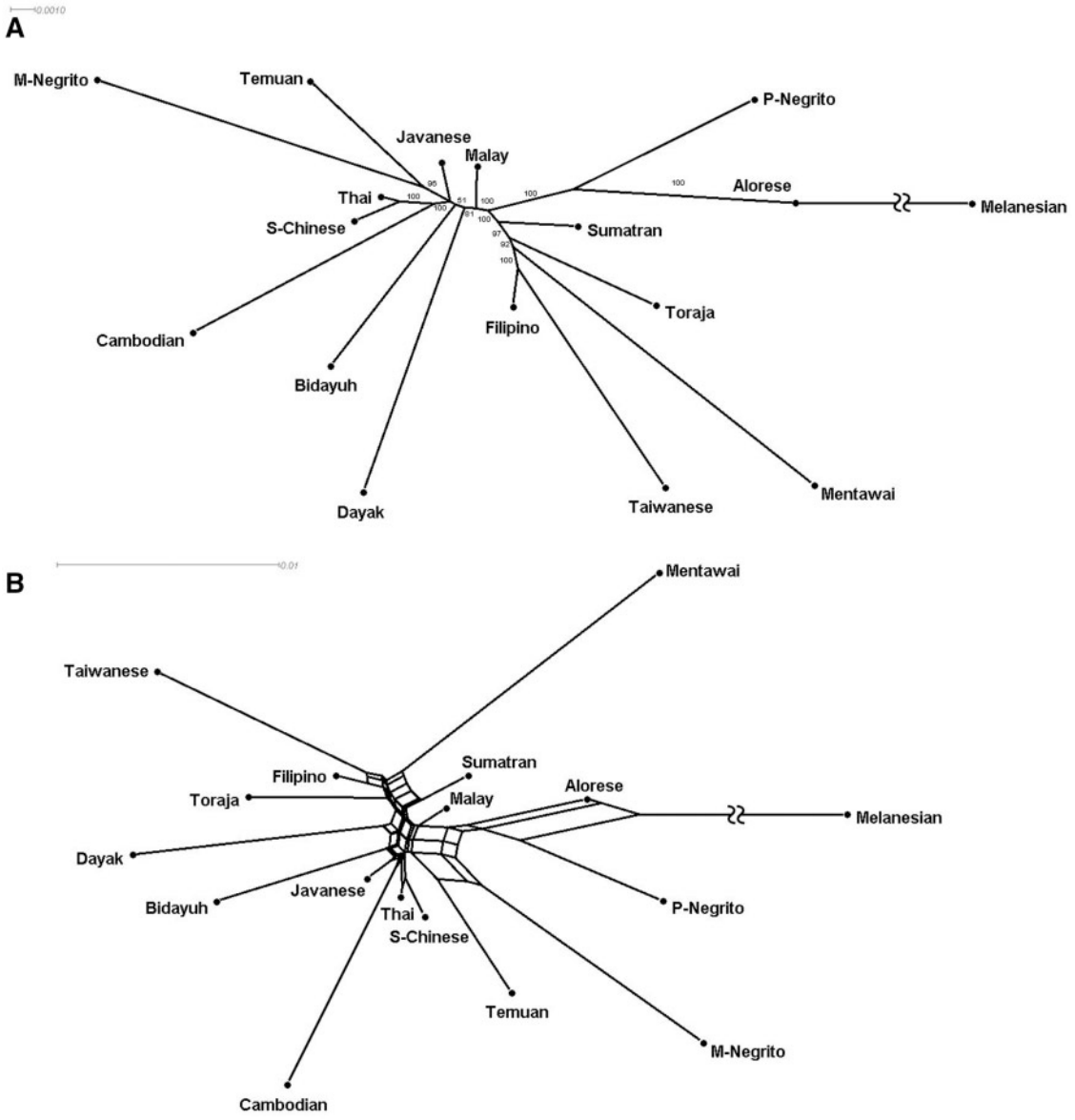

FIG. 5. (*A*) Neighbor-Joining tree and (*B*) Neighbor-Net network constructed from Nei's standard genetic distance matrix between populations listed in supplementary table S3, Supplementary Material online. SNP allele frequencies were used to generate the distance matrix. Bootstrap values above 90% are shown on the Neighbor-Joining tree.

To explore this tight clustering of groups, PCA was rerun after omitting Melanesian, Alorese, and Negrito (both Malaysian and Philippine) individuals, and the resulting PCA plot is shown in figure 4*B*. Generally, the Southern Chinese and Thai individuals tend to cluster together, whereas the other Austronesian populations appear to be spread out roughly according to geographic order, with indigenous Taiwanese, Filipinos, and Sulawesi individuals at one end and the Temuan, Bidayuh, and Javanese at the other. Further examination of other PCs (supplementary fig. S8, Supplementary Material online) show that the Temuan and Bidayuh individuals are also spread apart in a comet-like pattern, suggesting that these two groups also experienced recent admixture with other groups. Estimation of individual ancestry proportions using *frappe* software (Tang et al. 2005) in supplementary figure S10, Supplementary Material online, reveals a mosaic of Taiwanese (green), Southeast Asian (blue), and East Asian (orange) ancestry components at $k = 3$. The Taiwanese ancestry component decreases gradually according to the geographical order of populations, and the same pattern was observed for the Southeast Asian component in the opposite direction. At $k = 4$ and $k = 5$, the Temuan and Bidayuh appear to be distinguished from other Southeast Asian populations.

The Neighbor-Joining tree and Neighbor-Net network based on SNP allele frequencies of Southeast Asian groups (supplementary table S3, Supplementary Material online) were constructed. The Neighbor-Joining tree (fig. 5*A*) shows that populations that are geographically located east of the Wallace line (Filipino, Taiwanese, Toraja, Philippine Negrito,

Melanesian, and Alorese) form a cluster with 100% bootstrap probability. We refer to this cluster as "island" cluster. Populations from the west of the Wallace line (Bidayuh, Javanese, Malaysian Negrito, and Temuan) form another cluster with populations from mainland Asia (South Chinese, Thai, and Cambodian) and we refer to this cluster as "continental" cluster. There are exceptions to this general clustering, whereby the Sumatrans and Mentawai (west coast of Sumatra) that are geographically located west of the Wallace line but are clustered with the Filipino and Taiwanese on the "island" cluster. The Neighbor-Net network (fig. 5*B*) generally displays the same clustering pattern with the Neighbor-Joining tree and also highlights the division between "island" and "continental" clusters. The main difference is that the network shows the Malaysian Negritos as being closer to other "Australoid" populations (Philippine Negrito, Melanesian, and Alorese) than to other Austronesian groups, suggesting they may share some ancient commonalities.

Although the populations from the Indonesian islands of Alor and its vicinity speak Austronesian languages, they form a cluster with Melanesians. The reticulation observed between the Alorese and Melanesians on the network and the very short branch leading to the Alorese on the NJ tree suggests admixture between those two groups, and this was further confirmed by Xu et al. (2012). The reticulation observed between the Temuan and Malaysian Negrito may also be indicative of admixture events as suggested by the "comet-like" pattern in SNP-based PCA (fig. 4) and also the sharing of mtDNA haplogroups between the two groups. In general, the tree and the network appear to show the dichotomy between island and continental Southeast Asians as shown earlier by the mtDNA and SNP-based PCA results.

Discussion

This study includes the first description of mtDNA diversity in four indigenous Malaysian populations using complete sequence data from all individuals sampled. This is in contrast to most studies in which complete mtDNA sequencing was performed only on selected haplotypes based initially on control region diversity. Such biased sampling can lead to exaggerated results in some analyses as demonstrated in Gunnarsdottir et al. (2011a). A striking feature that we observed from our data was the limited mtDNA diversity in the Seletar. There were only four distinct haplogroups detected, reflected in the very low haplotype diversity statistic of 0.54, although not as extreme as the value of 0.167 observed in the Moken Sea Gypsies from Myanmar (Dancause et al. 2009). This limited mtDNA diversity in the Seletar may be the result of genetic drift, exacerbated by their small population size, which numbers approximately 800 individuals (Nicholas 2000). This may explain how haplogroup N9a6, which was reported at low frequencies in island Southeast Asia (Hill et al. 2007), rose to such high frequencies in the Selatar. We also took advantage of the availability of genome-wide SNP data from Southeast Asian populations reported in the PASNP (HUGO PASNP Consortium 2009) and the HGDP-CEPH (Li et al. 2008) studies to supplement the mtDNA data and to provide some insights into the migratory and demographic histories of Southeast Asian populations.

Regarding the Australoid populations in Southeast Asia, our mtDNA data do not appear to show any similarities in the extant mtDNA lineages of the Negrito groups (Andaman, West Malaysia, and the Philippines), Melanesians, and Australian Aboriginals. The mtDNA diversity in each of these Australoid groups is characterized by distinct markers, namely M31 and M32 in the Andamanese (Thangaraj et al. 2003, 2005; Barik et al. 2008), N11b in the Mamanwa of the Philippines (Gunnarsdottir et al. 2011a; this haplotype was labeled as N* in their article), M21a and R21 in the Jehai and Kensiu from West Malaysia, and haplogroups P, Q, S, and O in the Melanesians and Australian Aboriginals. We found that those mtDNA lineages have a time depth ranging from 30,000 to 50,000 YBP and is consistent with earlier reports (Ingman et al. 2000; Macaulay et al. 2005; Thangaraj et al. 2005; Gunnarsdottir et al. 2011a). This suggests their long-term presence in the Southeast Asia, probably dating back to the original inhabitants of the region. The diversity of the mtDNA lineages within these Australoid groups and their branching pattern on the NJ tree (supplementary fig. S4, Supplementary Material online) may imply multiple founder effects followed by long time isolation but whether there was a single, rapid entry (Macaulay et al. 2005) is still open to debate.

It also appears that these Australoid groups have experienced substantial gene flow with their neighboring populations. For example, haplogroups such as F1a1a and N9a6a in the Malaysian Negritos or E1a1a1 and B4b1a2 in the Philippine Negritos (Gunnarsdottir et al. 2011a) may have been introduced by admixture with the neighboring Austronesian populations. The effect of this admixture is also demonstrated in the comet-like patterns in the PCA plots using SNP data (fig. 4) and the reticulations in the Neighbor-Net network (fig. 5). Although the relatedness between these geographically distinct Australoid groups is not apparent based on the mtDNA diversity, SNP PCA plots, and Neighbor-Joining tree, the NeighborNet network (fig. 5*B*) shows that the Malaysian Negritos may have some commonalities with other Australoid groups (Philippine Negrito and Melanesian). A study of archaic hominin (Denisovan) admixture in Southeast Asia (Reich et al. 2011) showed that the Australoid populations (Jehai, Mamanwa, Australians, and New Guineans) share an ancient common ancestry but have since experienced different admixture episodes with different populations. A more exhaustive survey involving more Australoid populations with a denser set of autosomal SNP markers would therefore be desirable to paint a clearer picture regarding their interesting past.

As for the history of Austronesians, our mtDNA data points to a more substantial influence from the mainland in shaping the haplotype diversity of the three Austronesian groups we studied: Temuan, Seletar, and Bidayuh. The putative markers for the "Out of Taiwan" expansion, B4a1a and M7c3c account for less than 10% of the mtDNA lineages in all three Austronesian groups combined. Furthermore, other markers such as Y2, D5, M7b3, F3b, and

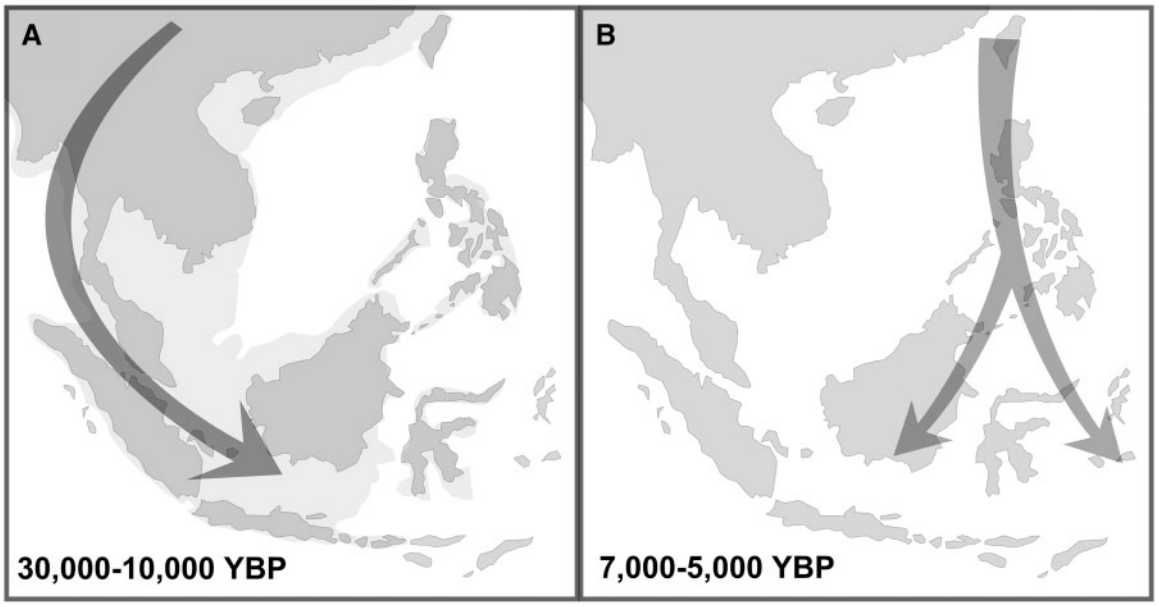

FIG. 6. Migration models relating to the origins of Austronesians. (*A*) "Early train" model originating from Indochina/South China ~30,000–10,000 years ago. (*B*) "Express train" model originating from Taiwan ~7,000–5,000 years ago.

F4, which were proposed to have followed the movement out of Taiwan (Hill et al. 2007; Tabbada et al. 2009), were not observed in any of the Austronesian groups in our study. An alternative explanation for the lack of "Out of Taiwan" haplogroups in the Austronesian groups we studied may be that the Austronesian expansion involved incorporation of females from existing populations, rather than replacing them. However, a study by Jordan et al. (2009) suggested that the ancestral Austronesian populations mostly practiced matrilocal postmarital residence. This would mean that the expanding Austronesian populations were more likely to retain their own mtDNA lineages. This casts some doubt on the incorporation of existing females by the expanding Austronesians, which would mask the Out of Taiwan signal. Instead we found a sizeable proportion of haplogroups with links to the mainland around the vicinity of Indochina or South China with ages predating the Austronesian expansion. This is characterized by haplogroups M21a, N9a6, N21, N22, and F1a'c, which account for more than 60% of the mtDNA lineages in the three Austronesian groups. This is in addition to haplogroups M74b, M22, G1c, M7b1, B5b2, M7c2, and B4c2, which also have roots in the mainland. These haplogroups are not found or very infrequent in populations around the vicinity of Taiwan and the ages range from approximately 30,000 to 10,000 YBP (fig. 3), corresponding to the late-Pleistocene to early-Holocene period.

The PCA plots using mtDNA (fig. 2) and SNP data (fig. 4) together with the Neighbor-Net network and Neighbor-Joining tree (fig. 5) appear to indicate a dichotomy in Austronesian populations. Populations from the mainland Asia (Thai, Southern Chinese, and Cambodian) tend to be closer to populations that were previously part of the Sundaland landmass (populations from Malaysia and Java and Borneo islands). This "continental" cluster of populations is separate from the "island" cluster of populations, which include Taiwanese, Filipino, and Sulawesi. The Malay from West Malaysia and Dayak from Borneo tend to be in intermediate positions of these two major clusters. This dichotomy is even clearer when Australoid populations were omitted from the network and tree analysis (supplementary fig. S9, Supplementary Material online). The high-bootstrap values lend support to the dichotomy of these two clusters (island and continental). Taken together, there results suggest that Austronesians might have originated from two or possibly more, separate migration events.

We therefore propose an "early train" hypothesis (fig. 6*A*) which differs from the "express train" (fig 6*B*), for explaining the observed results from mtDNA and SNP analysis in Austronesian groups. It essentially involved migration(s) originating from Indochina or South China, which spread south to West Malaysia, Sumatra, Java, and Borneo when they were still connected as Sundaland during the Last Glacial Maximum. The origin of this "early train" migration is inferred from the phylogenetic analysis of mtDNA lineages, which indicate that the ancestral types of lineages found in the continental cluster (West Malaysia, Borneo, and Java) tend to be found in Indochina or South China. The timing of this migration may have ranged from 30,000 to 10,000 YBP based on the age estimates of mtDNA haplogroups indicated in figure 3. Furthermore, the BSP plots in the three Austronesian groups studied (supplementary fig. S6, Supplementary Material online) do not indicate any signs of population expansion that might have taken place 5,000–7,000 YBP if they indeed originated from the "express train" expansion from Taiwan. The dichotomy between island and continental cluster of Austronesian populations based on PCA, network, and tree analyses lend further

support to our idea of a separate, earlier migration from the mainland which predates the "express train" from Taiwan.

Our proposed "early train" movement from the mainland does not preclude the episode of a Neolithic expansion from Taiwan involving Austronesian agriculturalists (Bellwood 2005), depicted in figure 6*B*. This is because the populations from Taiwan and its vicinity (the Philippines and Sulawesi) cluster together with high confidence based on the Neighbor-Joining tree, whereas the PCA plot in figure 4*B* also points to a relationship among the Taiwan, Philippine, and Sulawesi populations, which extend to the other island Austronesian groups such as the Mentawai and Sumatrans. The *frappe* analysis (supplementary fig. S10, Supplementary Material online) also shows a signal of Austronesian expansion at $k = 3$ to $k = 5$ as indicated by an ancestry component (in green) that is most frequent in Taiwan, which then decreases sequentially in the Philippine, Indonesian, and Malay groups who are all Austronesian speakers. Furthermore, the mtDNA lineages found in the Philippines and Taiwan tend to have time depths of less than 10,000 YBP (supplementary fig. S5, Supplementary Material online). The presence of haplogroups B4a1a and M7c3c in the Bidayuh from Borneo can be taken as an indication for the impact of the "Out of Taiwan" expansion on continental groups. It may be possible that Borneo was an intersection between the "early train" and "express train" movements given its location as the outer edge of the Sundaland landmass (fig. 1). Interestingly, although the Sumatrans and Mentawai are geographically located on the west of the Wallace line, thus on the continental side, they cluster with other populations on the "island" cluster on the NJ tree and network (fig. 5). These groups could have been part of the subsequent "express train" migration instead of the "early train" movement, hence their clustering with the "island" cluster of populations. Furthermore, the admixture between Asian and Melanesian ancestry in eastern Indonesian populations was dated back to ~4,000–5,000 YBP (Xu et al. 2012), which is consistent with the timing for the Austronesian expansion.

It should be noted that some caveats are in order when interpreting these results. The age of an mtDNA haplotype does not necessarily equate to its age in a population. It may well be possible that an "old" haplotype was introduced into a population by recent migrations, thus the haplogroup age estimates of 30,000 YBP may represent the upper limit for the possible time window of migration. However, even if we take age estimates of mtDNA lineages as the upper limit for the time of human migration events, we found that age estimates of mtDNA haplogroups associated with the Austronesian expansion (supplementary fig. S5, Supplementary Material online) are not too far off from the time estimated by archaeological data, which is 5,000 YBP. On the basis of our current analyses, we could not make any conclusions as to whether our "early train" migration involved a single entry or multiple waves of migration. It would therefore be desirable in future studies to vigorously test our "early train" model against other competing and plausible scenarios using demographic modeling, such as those reported by Batini et al. (2011).

Although we do not have direct archaeological evidence to support our "early train" dispersal from the mainland, some clues may lie within the existence of the Hoabinhian tradition, characterized by flaky pebble tools (Glover and Bellwood 2004). The Hoabinhian are thought to have emerged from Indochina during the early-Holocene period (14,000–10,000 YBP) based on dating of the stone tools and spread southward, as evidenced by the archaeological sites found in Sumatra and West Malaysia (Cavalli-Sforza et al. 1994; Glover and Bellwood 2004). We conjecture that the Hoabinhian may have made it all the way south to Java and Borneo. Linguistics may also provide some clues regarding human movement from continental Southeast Asia. Austro-Asiatic languages are mostly spoken in Indochina but are also used by the Senoi and the Negritos in Peninsular Malaysia. The ancestors of the Senoi are thought to have migrated southward from Indochina and introduced Austro-Asiatic languages to the extant Negrito populations in the Malay Peninsula (Bellwood 2005). However, the time of 4,000 YBP proposed by Bellwood (2005) for that migration is very recent compared with our "Early Train" model.

Other lines of evidence that corroborate our data include Y-chromosomal markers, which suggests a Paleolithic (30,000–15,000 YBP) contribution from mainland Asia (Karafet et al. 2010) and autosomal short tandem repeat markers which also show a dichotomy between Austronesians in Java and Samoa (Shepard et al. 2005). Other supporting data include that from domesticated animals which tend to accompany humans in their migrations. Genetic analysis of domesticated pig (Larson et al. 2007) and dog (Oskarsson et al. 2012) both propose a migration from mainland Asia by Sundaland and into the Pacific region, which may have been accompanied by humans.

Taken together, our results suggest an "early train" wave(s) of migration originating from South China or Indochina during late Pleistocene to early Holocene (30,000–10,000 YBP), predating the Neolithic expansion from Taiwan (Glover and Bellwood 2004; Bellwood 2005, 2007). We do not preclude the Out of Taiwan migration, but it appears improbable that it contributed wholly to the genetic diversity in all Austronesian groups, particularly those west of the Wallace line. In conclusion, our data suggest a more intricate migration history than the generally accepted, if not oversimplified, two-wave hypothesis regarding the peopling of island Southeast Asia.

Supplementary Material

Supplementary figures S1–S10 and tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Anuar Zaini, Khalid Kadir, Badariah Ahmad, Siti Harnida, Mah Yong Cheng, and other members of the Cardio-Metabolic RS at the Jeffrey Cheah School of Medicine and Health Sciences, MUSC; JAKOA officers who participated in the field study of the Orang Seletar; Roland

Schroeder, Anne Butthof, and Mingkun Li at MPI, Leipzig; Yoshimi Noaki at NIG, Mishima; and all participants/families who generously gave us their time. This work was supported by research grants # FP066-2007C from University of Malaya and # 5140060 from Monash University (Sunway Campus); SOKENDAI Strategic Research Project Grant; and The Max Planck Society.

References

Atkinson QD, Gray RD, Drummond AJ. 2008. mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Mol Biol Evol.* 25:468.

Barik SS, Sahani R, Prasad BVR, et al. (14 co-authors). 2008. Detailed mtDNA genotypes permit a reassessment of the settlement and population structure of the Andaman Islands. *Am J Phys Anthropol* 136:19–27.

Barker G, Barton H, Bird M, et al. (27 co-authors). 2007. The 'human revolution' in lowland tropical Southeast Asia: the antiquity and behavior of anatomically modern humans at Niah Cave (Sarawak, Borneo). *J Hum Evol.* 52:243–261.

Batini C, Lopes J, Behar DM, Calafell F, Jorde LB, van der Veen L, Quintana-Murci L, Spedini G, Destro-Bisol G, Comas D. 2011. Insights into the demographic history of African Pygmies from complete mitochondrial genomes. *Mol Biol Evol.* 28(2): 1099–1110.

Bellwood P. 2005. The first farmers: the origins of agricultural societies. Victoria, Australia: Blackwell Publishing.

Bellwood P. 2007. Prehistory of the Indo-Malaysian Archipelago. Canberra, Australia: ANU E Press.

Brothwell DR. 1960. Upper Pleistocene human skull from Niah caves, Sarawak. *Sarawak Museum J.* 9:323–350.

Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, Hammer M, Bustamante CD, Ostrer H. 2010. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc Natl Acad Sci U S A.* 107:8954–8961.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton, NJ: Princeton University Press.

Dancause KN, Chan CW, Arunotai NH, Lum JK. 2009. Origins of the Moken Sea Gypsies inferred from mitochondrial hypervariable region and whole genome sequences. *J Hum Genet.* 54:86–93.

Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Dambueva I, Perkova M, Dorzhu C, Luzina F, Lee H, Vanecek T. 2007. Phylogeographic analysis of mitochondrial DNA in Northern Asian populations. *Am J Hum Genet.* 81:1025–1041.

Diamond JM. 1988. Express train to Polynesia. *Nature* 336:307–308.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 7:214.

Fagundes NJR, Kanitz R, Eckert R, Valls ACS, Bogo MR, Salzano FM, Smith DG, Silva WA Jr, Zago MA, Ribeiro-dos-Santos AK. 2008. Mitochondrial population genomics supports a single pre-clovis origin with a coastal route for the peopling of the Americas. *Am J Hum Genet.* 82:583–592.

Felsenstein J. 2005. PHYL0IP (Phylogeny Inference Package). Seattle: Department of Genome Sciences, University of Washington.

Fucharoen G, Fucharoen S, Horai S. 2001. Mitochondrial DNA polymorphisms in Thailand. *J Hum Genet.* 46: 115–125.

Glover I, Bellwood PS. 2004. Southeast Asia: from prehistory to history. Oxford, UK: Routledge.

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol.* 9:585–598.

Gray RD, Drummond AJ, Greenhill SJ. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323: 479–483.

Gray RD, Jordan FM. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405:1052–1055.

Gunnarsdottir ED, Li M, Bauchet M, Finstermeier K, Stoneking M. 2011a. High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Res.* 21:1–11.

Gunnarsdottir ED, Nandineni MR, Li M, Myles M, Gil D, Pakendorf B, Stoneking M. 2011b. Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. *Nat Commun.* 2:228.

Hill C, Soares P, Mormina M, et al. (12 co-authors). 2006. Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Mol Biol Evol.* 23:2480–2491.

Hill C, Soares P, Mormina M, et al. (11 co-authors). 2007. A mitochondrial stratigraphy for island southeast Asia. *Am J Hum Genet.* 80: 29–43.

Horai S, Satta Y, Hayasaka K, Kondo R, Inoue T, Ishida T, Hayashi S, Takahata N. 1992. Man's place in Hominoidea revealed by mitochondrial DNA genealogy. *J Mol Evol.* 35:32–43.

HUGO Pan-Asian SNP Consortium. 2009. Mapping human genetic diversity in Asia. *Science* 326:1541–1545.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 23:254–267.

Ingman M, Kaessmann H, Paabo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* 408: 708–713.

Jin H-J, Tyler-Smith C, Kim W. 2009. The peopling of Korea revealed by analyses of mitochondrial DNA and Y-chromosomal markers. *PLoS One.* 4:e4210.

Jinam T, Phipps M, Indran M, Kuppusamy U, Mahmood AA, Hong LC, Edo J. 2008. An update of the general health status in the indigenous populations of Malaysia. *Ethn Health.* 13:277–287.

Jinam T, Saitou N, Edo J, Mahmood A, Phipps M. 2010. Molecular analysis of HLA class I and class II genes in four indigenous Malaysian populations. *Tissue Antigens.* 75:151–158.

Jordan FM, Gray RD, Greenhill SJ, Mace R. 2009. Matrilocal residence is ancestral in Austronesian societies. *Proc R Soc B.* 276:1957–1964.

Karafet TM, Hallmark B, Cox MP, Sudoyo H, Downey S, Lansing JS, Hammer MF. 2010. Major east–west division underlies Y chromosome stratification across Indonesia. *Mol Biol Evol.* 27:1833–1844.

Kayser M, Brauer S, Weiss G, Underhill P, Roewer L, Schiefenhövel W, Stoneking M. 2000. Melanesian origin of Polynesian Y chromosomes. *Curr Biol.* 10:1237–1246.

Kayser M, Choi Y, van Oven M, Mona S, Brauer S, Trent RJ, Suarkia D, Schiefenhövel W, Stoneking M. 2008. The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. *Mol Biol Evol.* 25:1362–1374.

Kong QP, Sun C, Wang HW, et al. (16 co-authors). 2011. Large-scale mtDNA screening reveals a surprising matrilineal complexity in East Asia and its implications to the peopling of the region. *Mol Biol Evol.* 28:513–522.

Kong QP, Yao YG, Sun C, Bandelt HJ, Zhu CL, Zhang YP. 2003. Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. *Am J Hum Genet.* 73:671–676.

Larson G, Cucchi T, Fujita M, et al. (32 co-authors). 2007. Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in Island Southeast Asia and Oceania. *Proc Natl Acad Sci U S A.* 104: 4834–4839.

Leavesley M, Chappell J. 2004. Buang Merabak: additional early radiocarbon evidence of the colonisation of the Bismarck Archipelago, Papua New Guinea. *Antiquity* 78(301). Available from: http://antiquity.ac.uk/projgall/leavesley/index.html.

Li JZ, Absher DM, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Macaulay V, Hill C, Achilli A, et al. (21 co-authors). 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* 308:1034–1036.

McEvoy BP, Lind JM, Wang ET, Moyzis RK, Visscher PM, van Hoslt Pellekaan SM, Wilton AN. 2010. Whole-genome genetic diversity in a sample of Australians with deep Aboriginal ancestry. *Am J Hum Genet.* 87:297–305.

Mirabal S, Herrera KJ, Gayden T, Regueiro M, Underhill PA, Garcia-Bertrand RL, Herrera RJ. 2012. Increased Y-chromosome resolution of haplogroup O suggests genetic ties between the Ami aborigines of Taiwan and the Polynesian Islands of Samoa and Tonga. *Gene* 492: 339–348.

Mishmar D, Ruiz-Pesini E, Golik P, et al. (13 co-authors). 2003. Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci U S A.* 100:171–176.

Nei M. 1972. Genetic distance between populations. *Am Nat.* 106: 283–292.

Nicholas C. 2000. The Orang Asli and the contest for resources. Indigenous politics, development and identity in peninsular Malaysia. Denmark: IWGIA & Center for Orang Asli Concerns.

O'Connell JF, Allen J. 2004. Dating the colonization of Sahul (Pleistocene Australia-New Guinea): a review of recent research. *J Archaeol Sci.* 31:835–853.

Oppenheimer SJ, Richards M. 2001. Slow boat to Melanesia? *Nature* 410: 166.

Oskarsson MCR, Klütsch CFC, Boonyaprakob U, Wilton A, Tanabe Y, Savolainen P. 2012. Mitochondrial DNA data indicate an introduction through Mainland Southeast Asia for Australian dingoes and Polynesian domestic dogs. *Proc Biol Sci.* 279:967–974.

Patterson N, Price AL, Reich D. 2006. Population structure and Eigenanalysis. *PLoS Genet.* 2:e190.

Peng MS, Quang HH, Dang KP, Trieu AV, Wang HW, Yao YG, Kong QP, Zhang YP. 2010. Tracing the Austronesian footprint in mainland Southeast Asia: a perspective from mitochondrial DNA. *Mol Biol Evol.* 27:2417–2430.

Reich D, Patterson N, Kircher M, et al. (15 co-authors). 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Gen.* 89:516–528.

Ricaut FX, Bellatti M, Lahr MM. 2006. Ancient mitochondrial DNA from Malaysian hair samples: some indications of Southeast Asian population movements. *Am J Hum Biol.* 18:654–667.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4(4):406–425.

Shepard EM, Chow RA, Suafo'a E, Addison D, Pérez-Miranda AM, Garcia-Bertrand RL, Herrera RJ. 2005. Autosomal STR variation in five Austronesian populations. *Hum Biol.* 77:825–851.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet.* 84:740–759.

Soares P, Rito T, Trejaut J, et al. (16 co-authors). 2011. Ancient voyaging and polynesian origins. *Am J Hum Genet.* 88:239–247.

Soares P, Trejaut J, Loo JH, et al. (14 co-authors). 2008. Climate change and postglacial human dispersals in Southeast Asia. *Mol Biol Evol.* 25: 1209–1218.

Tabbada KA, Trejaut J, Loo JH, Chen YM, Lin M, Mirazon-Lahr M, Kivisild T, De Ungria MCA. 2009. Philippine mitochondrial DNA diversity: a populated viaduct between Taiwan and Indonesia? *Mol Biol Evol.* 27: 21–31.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol.* 10:512–526.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol.* 28: 289–301.

Terrell J. 1988. Prehistory in the Pacific Islands. Cambridge, UK: Cambridge University Press.

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, Singh L, Reddy AG, Rao VR, Sehgal SC, Underhill PA, Pierson M, Frame IG, Hagelberg E. 2003. Genetic affinities of the Andaman islanders, a vanishing human population. *Curr Biol.* 13: 86–93.

Torroni A, Rengo C, Guida V, et al. (12 co-authors). 2001. Do the four clades of the mtDNA haplogroup L2 evolve at different rates. *Am J Hum Genet.* 69:1348–1356.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat.* 30:E386–E394.

Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nürnberg P, Stoneking M, Kayser M. 2010. Demographic history of Oceania inferred from genome-wide data. *Curr Biol.* 20:1983–1992.

Xu S, Pugach I, Stoneking M, Kayser M, Jin L. 2012. Genetic dating indicates that the Asian–Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *PNAS.* 109: 4574–4579. Available from: http://www.pnas.org/content/early/2012/03/05/1118892109.