Am. J. Hum. Genet. 68:432–443, 2001

A Predominantly Indigenous Paternal Heritage for the Austronesian-Speaking Peoples of Insular Southeast Asia and Oceania

Cristian Capelli,[1,2] James F. Wilson,[1,3] Martin Richards,[1] Michael P. H. Stumpf,[1,3] Fiona Gratrix,[1] Stephen Oppenheimer,[4] Peter Underhill,[5] Vincenzo L. Pascali,[2] Tsang-Ming Ko,[6] and David B. Goldstein[1]

[1]Department of Biology, University College London, London; [2]Istituto di Medicina Legale, Universitá Cattolica del Sacro Cuore, Rome; [3]Department of Zoology, University of Oxford, and [4] Green College, University of Oxford, Oxford; [5]Department of Genetics, Stanford University, Stanford, CA; and [6]Department of Obstetrics and Gynecology, National Taiwan University, Taipei

Modern humans reached Southeast Asia and Oceania in one of the first dispersals out of Africa. The resulting temporal overlap of modern and archaic humans—and the apparent morphological continuity between them—has led to claims of gene flow between *Homo sapiens* and *H. erectus*. Much more recently, an agricultural technology from mainland Asia spread into the region, possibly in association with Austronesian languages. Using detailed genealogical study of Y chromosome variation, we show that the majority of current Austronesian speakers trace their paternal heritage to Pleistocene settlers in the region, as opposed to more-recent agricultural immigrants. A fraction of the paternal heritage, however, appears to be associated with more-recent immigrants from northern populations. We also show that the northern Neolithic component is very unevenly dispersed through the region, with a higher contribution in Southeast Asia and a nearly complete absence in Melanesia. Contrary to claims of gene flow (under regional continuity) between *H. erectus* and *H. sapiens*, we found no ancestral Y chromosome lineages in a set of 1,209 samples. The finding excludes the possibility that early hominids contributed significantly to the paternal heritage of the region.

Introduction

The record of human activity is extremely rich and varied in insular Southeast Asia and Australasia. *Homo erectus* first reached this part of the world >1 million years before present (BP) and may have survived in the region as recently as 25,000 years ago (Swisher et al. 1996). Fully modern humans were present at sites dated 50,000–60,000 years BP in Australia (Roberts 1990) and at sites in Melanesia dated as early as 33,000 years BP (Pavlides and Gosden 1994; Allen 1996). With the exception of the essentially modern forms from the Levant, dated at ~100,000 years BP, the Australasian finds may be the oldest known outside Africa (Klein 1999). Before the arrival of a Neolithic culture from the north, the only evidence of intensive agricultural activity in the region is among the Papuan speakers of New Guinea and nearby islands. Beyond this broad and vague outline, however, little is known, with any certainty, about the region. In fact, it is not known whether modern humans who colonized the region came from populations first established in eastern Asia or came more directly from Africa along southern coastal routes (Lahr and Foley 1994).

This region is also home to the Austronesian language family. The most widespread language family before the age of European expansion, it is spoken from Madagascar to the remote Pacific. Archaeological evidence demonstrates a technological connection between the farming cultures of the south (Southeast Asia and Melanesia) and sites that are first known from mainland China, whereas a combination of archaeological and linguistic evidence has been interpreted as supporting a northern (southern China and Taiwan) origin for the Austronesian language family. In a recent treatment, all Austronesian languages were classified into 10 subfamilies, with all the extra-Formosan languages grouped in one subfamily and with representatives of the remaining 9 known only in Taiwan (Blust 1999). It has been argued that these patterns are best explained by dispersal of an agricultural people from Taiwan into insular Southeast Asia, Melanesia, and, ultimately, the remote Pacific. Although this model—termed the "express train to Polynesia" (Diamond 1988)—is broadly consistent with available data, concerns have been raised (Richards et al. 1998). Alternatives to this model posit an indigenous origin for the Austronesian languages in Mela-

Received November 6, 2000; accepted for publication December 15, 2000; electronically published January 22, 2001.

Address for correspondence and reprints: Dr. David B. Goldstein, Department of Biology (Galton Laboratory), University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, United Kingdom. E-mail: d.goldstein@ucl.ac.uk

0002-9297/2001/6802-0014$02.00

nesia or Southeast Asia (Dyen 1962, 1965; Oppenheimer 1998).

Here we characterize Y chromosome variation in 1,209 samples drawn from Southeast Asia, Oceania, southern China, and Taiwan. The combination of a geographically extensive sample set with high-resolution Y chromosome haplotype data, employing both unique-event polymorphisms (UEPs), which define haplogroups, and quickly evolving microsatellites (Thomas et al. 1998, 1999), which define haplotypes, provides information about the length of time that lineages have been in residence in defined geographic areas. This allows identification of lineages that appear to have been in residence in the south, long before any possible input from the north associated with agriculture, and other lineages that were more recently dispersed and that could have been dispersed in association with agriculture. The identification of such lineages suggests (1) that the paternal heritage of many East Asian populations has an origin in insular Southeast Asia and Melanesia and (2) that the genetic heritage of Austronesian agriculturalists throughout Southeast Asia and Melanesia has a conspicuous indigenous origin and that, in Melanesia in particular, the dispersal of the Austronesian languages was mainly a cultural process, in contradiction to the express-train model.

Populations and Methods

Samples

For simplicity, "East Asia" refers to the entire set of populations sampled. Where appropriate, Formosan and mainland-Chinese samples have been clustered as northern, and insular Southeast Asians and Melanesians have been clustered as southern. A set of >2,000 samples from these regions were sex-typed using published primers (Sullivan et al. 1993). DNA samples of Southeast Asians, Melanesians, Polynesians, and a subset of Formosans were provided by John Clegg, from the Institute of Molecular Medicine collection, Oxford. The majority of the aboriginal Formosans (Atayal, Ami, Paiwan, Bunun, and Yami) and the southern Chinese were collected by Tsang-Ming Ko, who provided information about group memberships. The Irian Jaya samples were collected by Martin Richards in the Ayamaru district, Bird's Head, New Guinea, Indonesia. The following samples were also analyzed: (1) Southeast Asia—Toraja and Palu (Sulawesi), Kota Kinabalu and Banjarmasin (Borneo), Pekanbaru (Sumatra), Mataram, and Alor (Lesser Sunda); (2) Melanesia—Fiji, Vanuatu, New Ireland, Madang (Papua New Guinea), and Irian Jaya; and (3) Polynesia—Tonga, Western Samoa, French Polynesia (mainly from Tahiti), and Atiu (southern Cooks). In addition, 28 individuals from the Philippines were genotyped. Geographic locations and sample sizes are shown in table 1 and figure 1.

Y Chromosome Genotyping

Ten Y-linked microsatellites (DYS388, DYS393, DYS392, DYS19, DYS390, DYS391, DYS425, DYS426, and DYS389 I and II) were genotyped, as described by Thomas et al. (1999), on an ABI 377 automated sequencer. A further nine UEPs were typed: M9 and M4 (Underhill et al. 1997), 92R7 (Mathias et al. 1994), RPS4Y 711 (Bergen et al. 1999), M175, M122, M119, and M95 (Underhill et al. 2000), and SRY 10831 (Whitfield et al. 1995), identifying 10 haplogroups (fig. 2). Genotyping was performed using multiplex kits. 92R7, M4, and M9 were coamplified using the following primers: M4U (5′-GTTATGATTACAGAGCGAGGATT-3′ and M4R 5′-TTGGGCAGAAAATTCAAAGT-3′), M9U (5′-TTGAAGCTCGTGAAACAGATTAG-3′ and M9R 5′-CTTTCAGGACCCTGAAATACAG-3′), and 92R7U (5′-TCAGAAAGATAGTAAGAGGAACACTTC-3′ and 92R7R 5′-GCATTGTTAAATATGACCAGCA-3′). PCR conditions were 94°C for 4 min, followed by 38 cycles at 94°C for 40 s, 58°C for 40 s, and 72°C for 40 s, with a final extension step at 72°C for 10 min. Reactions were run in a final volume of 10 μl containing 200 μM of each dNTP, 1 × Super*Taq* PCR Buffer (HT Biotech), 3.5 pmol of each primer, 0.13 units of *Taq* polymerase enzyme (HT Biotech), and 9.3 nM of *Taq*Start antibody (Clontech). RPS4Y was amplified in singleplex, with the same PCR conditions and the following primers: RPS4YU (5′-AATCATGAGCAAATTCAAACC-3′) and RPS4YR (5′-CTCTCCTCCCTTTCTTTCTGT-3′). From each of the 10-μl PCR reactions, 2 μl was multiplex digested with the following enzymes *Nde*I, *Hin*fI, and *Hin*dIII (New England Biolabs) in a 96-well plate overnight at 37°C. The RPS4Y amplicon was digested at 55°C for 2 h, with *Bsl*I (New England Biolabs). All the restriction reactions were performed in a final volume of 8 μl, and 2 μl was then loaded on an ABI 377 sequencer, and sizes were inferred, with TAMRA 350 as an internal-sizing ladder, using GENESCAN software (PE Biosystems). A second multiplex amplified the following markers: M175, M122, M119, and M95. Primers, PCR condition, and digestion protocol may be requested from C.C. RPS4Y was genotyped only on chromosomes bearing the M9 ancestral state (fig. 2). Similarly M175, M122, M119, and M95 were genotyped only on M9-derived chromosomes (haplogroup M9) (fig. 2). In addition, the marker SRY10831 was genotyped on chromosomes showing the ancestral allele both at RPS4Y and at M9, with a singleplex version of the conditions described by Thomas et al. (1999). Forty-five samples of haplogroup C, from Mongolia and India, were included in the analysis.

Table 1

Sample Sizes and Haplogroup Frequencies

Region (N[a])	Haplogroup[b]								
	B	C	D	E	F	G	H	I	L
Southeast Asia (312):									
Alor (50)	0	44	0	22	12	2	6	0	14
Mataram (43)	0	12	2	5	14	4	17	0	1
Kota Kinabalu (51)	2	13	0	0	2	17	38	0	28
Banjarmasin (34)	3	24	0	0	18	27	12	0	15
Palu (37)	3	6	3	3	17	14	17	0	37
Toraja (52)	2	8	2	6	13	15	38	0	15
Pekanbaru (45)	9	2	9	0	0	23	35	0	21
Melanesia (342):									
Irian Jaya (19)	0	58	0	16	26	0	0	0	0
Madang (90)	0	16	0	41	42	1	0	0	0
New Ireland (86)	0	17	0	29	45	0	3	0	5
Vanuatu (92)	0	18	0	15	6	0	3	0	3
Fiji (55)	0	3	0	15	41	0	6	0	9
Polynesia (200):									
Tonga (55)	0	23	0	8	1	0	2	0	58
Western Samoa (16)	0	69	0	0	6	6	6	0	13
French Polynesia (87)	0	53	2	0	8	0	2	0	35
Atiu (42)	0	84	3	0	1	0	0	0	03
Philippines (28)	3	0	0	0	0	0	41	0	56
Taiwan (246):									
Paiwan (53)	0	0	0	0	0	0	96	4	0
Bunun (50)	0	0	0	0	0	24	74	0	02
Atayal (50)	0	0	0	0	0	0	98	0	02
Ami (53)	0	0	0	0	0	6	43	4	47
Yami (40)	0	2	0	0	0	28	7	0	0
Southern China (80)	0	4	1	0	2	4	14	1	65

[a] European chromosomes were identified in 25 samples from French Polynesia, 5 from Vanuatu, 4 from Kota Kinabalu, 3 from Tonga and Atiu, 2 from Palu and Pekanbaru, and 1 each from Fiji, Mataram, Banjamarasin, and the Philippines.

[b] Frequencies are stated as percentages and were estimated with European chromosomes excluded.

Results and Discussion

European Chromosomes in Southeast Asia and Oceania

The 10 Y chromosome microsatellites and 9 UEPs confirmed the importance of male-mediated European gene flow into Polynesia (Hurles et al. 1998). Using a collection of European samples typed in our laboratory for six markers (DYS388, DYS393, DYS392, DYS19, DYS390, and DYS391), we identified as European those haplogroup B and haplogroup D (fig. 2) haplotypes that either matched a European haplotype or were separated by one or two mutational events. This must be taken as approximate, because homoplasy could lead to false identification of a haploytpe as European, and our data set may be missing European haplotypes present in the region. Using these criteria, we identify a total of 48 chromosomes as European, ranging in frequency from 0% in Western Samoa to 29% in French Polynesia. Some European introgression was also evident in Southeast Asia (2.3%–7.8%) and the Philippines (3.6%).

A Melanesian Origin for East-Asian Y Chromosomes

If the European chromosomes are excluded, 99% of the chromosomes observed within East Asia are derived at either RPS4Y-711 (haplogroup C) or M9 (figs. 1 and 2 and table 1). Haplogroup C chromosomes are common south of the Philippines, with a significantly greater diversity, calculated as the average variance of the repeat score across loci, in the southern populations (Southeast Asia and Melanesia) than in the northern ones (southern China and Taiwan) ($P < .01$; 1,000 bootstraps over chromosomes) (tables 1 and 2). Southeast Asian and Melanesian groups have similar variances, whereas Polynesia is much less diverse. Diversity among the mainland samples (49 chromosomes from India, Mongolia, southern China, and Taiwan) is also substantially lower than in Southeast Asia and Melanesia, suggesting that haplogroup C may have originated in the south. The genealogy inferred on the basis of microsatellite variation (Ruiz-Linares et al. 1999) clearly shows deep splits between mainland, insular Southeast Asian, and Melanesian Y chromosomes, with Polynesians closely associated

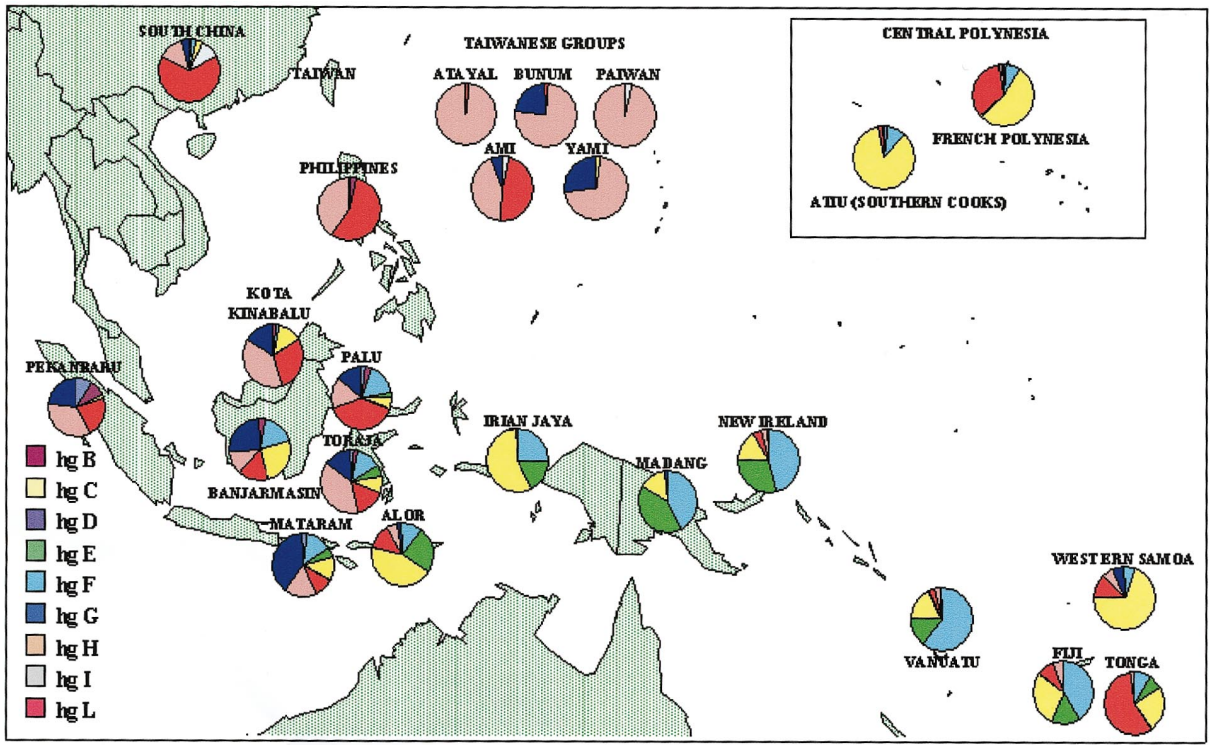

Figure 1 Geographic locations of the analyzed populations. Charts represent haplogroup frequencies. No haplogroup A chromosomes were found.

with the Melanesian clusters (fig. 3*a*). Because the variation in the mainland populations sampled is a subset of that in the south, we conclude that haplogroup C observed in the north is derived from that in the south. The correspondence between the genealogy and geography implies that the Southeast Asian/Melanesian split reflects differentiation in the region, as is also emphasized by the absence of shared haplotypes across these areas. Local antiquity of this haplogroup is indicated by its considerable genealogical depth, with the Southeast Asia clade estimated to coalesce with the nearest Melanesian clade at >12,000 years BP, when the average squared distance is used (Slatkin 1995; Goldstein et al. 1995*a*). Because haplogroup C is also present in Australia (Karafet et al. 1999), it appears to be a good candidate for having been included among the original or early settlers of the region.

Haplogroup F is mainly restricted to southern populations (appearing in only 2 of 327 samples in the north [i.e., Taiwan and southern China]) and has its greatest diversity and highest frequency in Melanesia (tables 1 and 2). The inferred genealogy has Melanesian types throughout, and no clear association with geography is apparent. Like haplogroup C, haplogroup F therefore appears to have an origin in the south—in particular, in Melanesia. Taken together, these two haplogroups represent 26% (23%–56%) of the Southeast Asian chromosomes and 68% (58%–84%) of the Melanesian ones.

It is known that haplogroup M9 is present in all the continents of the world, albeit at highly variable frequencies, ranging from >70% in the Americas and in East Asia to only 4% in Africa (Karafet et al. 1999). When only those samples without the derived markers 92R7 or Tat (Zerjal et al. 1997) are considered, this group has its maximum frequencies in Asian and Australasian regions and is generally rare elsewhere. Asian populations have a high proportion of haplogroup M9 chromosomes derived at M122, M119, and M95 (Su et al. 1999). Diversity data are not available outside the area investigated here. Nevertheless, Asia has by far the highest global frequencies of M9 chromosomes ancestral at those internal markers that have been genotyped; and, within Asia, Melanesia has the highest frequencies, as well as the greatest diversity, in haplogroup F. Unfortunately, our combination of UEP and microsatellite markers is insufficient to determine the branching order of lineages within the M9 group. This lack of resolution makes it impossible to determine the geographic location of the deepest lineages within M9. We do note, however, that the pattern of diversity would permit, though not prove, the existence of deep-branching lineages in Melanesia within the M9 haplogroup. This pattern raises

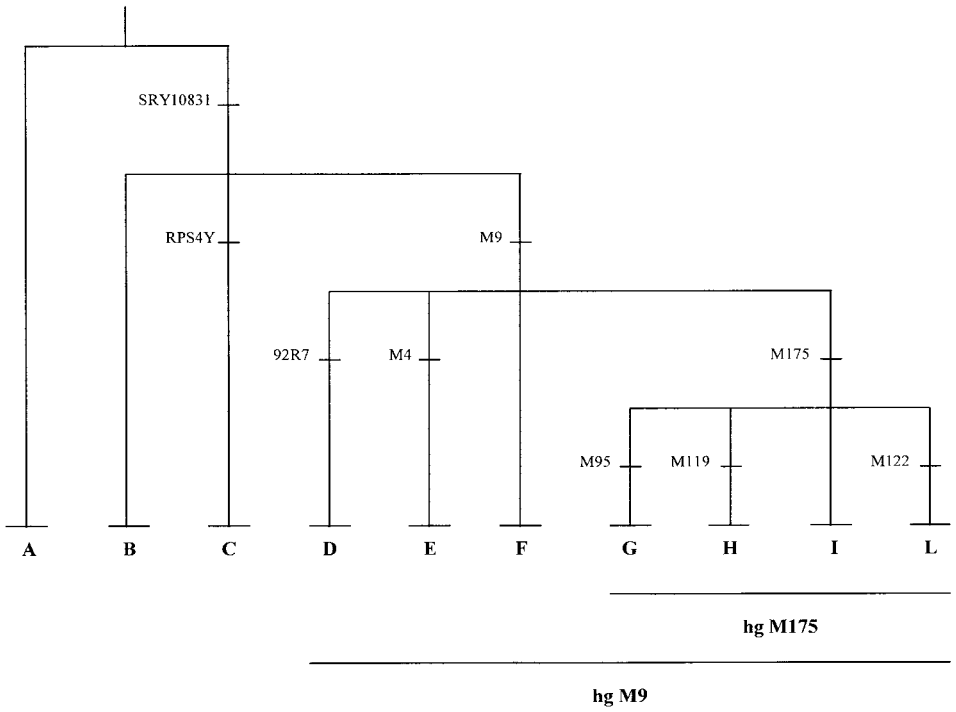

Figure 2 Genealogy showing haplogroup designations and the defining mutations. The coalescence time, in generations, of chromosomes within indicated haplogroups was estimated by inferring the founder haplotype, where possible, as the modal allele at each locus and by using methods described by Thomas et al. (1998), corrected for length dependence (Goldstein et al., in press). The applied mutation rate was 2.8×10^{-3} (Kayser et al. 2000), which resulted in the following estimates for indicated haplogroups: E, 304 generations; G, 344 generations; H, 179 generations; L, 524 generations; and C, 871 generations.

the possibility that non-African chromosomes carrying M9 could have resulted from an early expansion out of Melanesia that would now account for more than three-quarters of the paternal heritage of non-African populations (Karafet et al. 1999). Alternatively, this region could retain ancestral lineages within M9 that were once more widespread but now are lost or go unobserved in other regions. Determination of the geographic origin of this important haplogroup should be possible with only a modest increase of informative UEP markers within the M9 group, together with more-extensive sampling in mainland Asia.

If we consider the M9 chromosomes in our sample to have an origin in the south, then virtually all northern chromosomes can be traced back to the south. Within haplogroup M9, moreover, all but three northern chromosomes are within M175, and M175 itself appears to have an origin in Southeast Asia. Its diversity is greatest in this region, although haplogroup I, underived within haplogroup M175, appears in the north. Current evidence is consistent with a southern origin for virtually all the paternal heritage of the populations considered here.

Signatures of Mainland Asian Agriculturists

The two derived groups, L and H, within haplogroup M175 constitute the majority of the chromosomes observed in the north. Haplogroup L, in particular, shows a distribution of haplotypes that is dramatically different from the distribution observed in the groups discussed above. This group has its greatest diversity in the north, and, in particular, has extremely little variation in Melanesia. Shared microsatellite haplotypes were found across all the geographic regions, with China and Taiwan and the southern populations sharing nine types (representing 18 and 12 chromosomes, respectively).

The inferred genealogy for haplogroup L (fig. 3*b*) is even more striking, showing that chromosomes from a single Taiwanese group, the Ami, are distributed throughout all major groups in the tree. This observation is consistent with the variance within the Ami—which, at .49, is as high as the average across Southeast Asia. The simplest explanation for this pattern is that haplogroup L throughout the south is associated with dispersal from the north, mediated by the Ami or an Ami-like people. As expected from a recent expansion into

Table 2

Variance of the Repeat Score of Specific Haplogroups in Various Geographic Groupings

	Variance of Repeat Score[a]							
Haplogroup	All Northern[b]	Southern China	Taiwan	All Southern[c]	Southeast Asia	Melanesia	Polynesia	Overall
C	.36[d]			.88	.8	.7	.14	.86
E					.34	.43		.44
F					.57	.83	.39	.82
M175	.39	.46	.35	.49	.56	.42	.27	.43
G	.22	.43	.19	.28	.28			.28
H	.22	.13	.23	.34	.31			.27
I	.32							.32
L	.53	.42	.52	.39	.44	.13	.25	.43

[a] Averaged across loci estimated in geographic groups for the different haplogroups. Values are shown only when >10 chromosomes were present. Data from the Philippines are not shown.

[b] Includes Taiwan and southern China.

[c] Includes Southeast Asia and Melanesia.

[d] Includes samples from southern China and Taiwan and 45 chromosomes from India and Mongolia.

the south, the structure of the inferred genealogy for haplogroup L is not associated with the geographic origin of the samples. Haplogroup L is also present in Indochina (Su et al. 1999), and our data do not rule out the possibility of a parallel route of agricultural technology and people through the Malayan peninsula (Oppenheimer 1998). Additional samples from this region would help clarify the history of this lineage.

The evident geographic structure in the haplogroup C genealogy, and the absence of such structure for haplogroup L, suggests that haplogroup L dispersed through its geographic range more recently than did haplogroup C. To assess this difference quantitatively, we have compared the estimated pairwise coalescence times between chromosomes, within and between populations, for the two haplogroups. In particular, we have compared sampled chromosomes from the north and the south, for haplogroup L, and from Southeast Asia and Melanesia, for haplogroup C. To infer pairwise coalescence times, we use average squared distance (Slatkin et al. 1995; Goldstein et al. 1995*a*). The within- and between-populations distribution of distances for haplogroup L are indistinguishable, with similar means (within the north = 1.07; within the south = 0.79; and between north and south =1.0). For haplogroup C, however, the between-populations distribution has a higher mean value than the within-population distributions (1.39 and 1.61 for within populations, and 2.11 for between populations). The between-population distribution for haplogroup C is also clearly multimodal, whereas all other distributions are unimodal—except for haplogroup C in Melanesia, which has a much smaller minor peak. This second peak could result from the movement of types from Southeast Asia, because the distributions within populations otherwise appear to reflect the unimodal pattern associated with population growth.

To assess whether the different within- and between-population coalescence times reflect significantly different dispersal times for haplogroups C and L, we performed computer simulations focusing only on these haplogroups. In particular, for haplogroup L, we assumed that lineages within the south were separated from those in the north at some time t_1. For haplogroup C, we assumed that those in Melanesia were separated from those in Southeast Asia at time t_2. To estimate these times, we ignore all other haplogroups in the populations and use the genetic distance $\delta\mu^2$ (Goldstein et al. 1995*b*), which results in estimates for t_1 and t_2 of 22 and 158 generations, respectively. Although the calculation of a population separation time using only a subset of the chromosomes present in the populations is ad hoc and subject to bias, our intention is to eliminate, in making the calculations, those haplogroups that may have been dispersed at a different time. To assess whether t_1 and t_2 differ significantly, we performed simulations of 10 completely linked loci evolving in two independent populations that separated from one another at a time t in the past, with all populations at mutation-drift equilibrium. The populations were then evolved independently for 22 or 158 generations. When 500 replications are used for each of the two generation times, the resulting distributions of genetic distances are marginally significantly different ($P < .07$). The greater between-population value in comparison with within-population values observed for haplogroup C, compared with haplogroup L, therefore is consistent with its dispersal through its range earlier than was the case for haplogroup L.

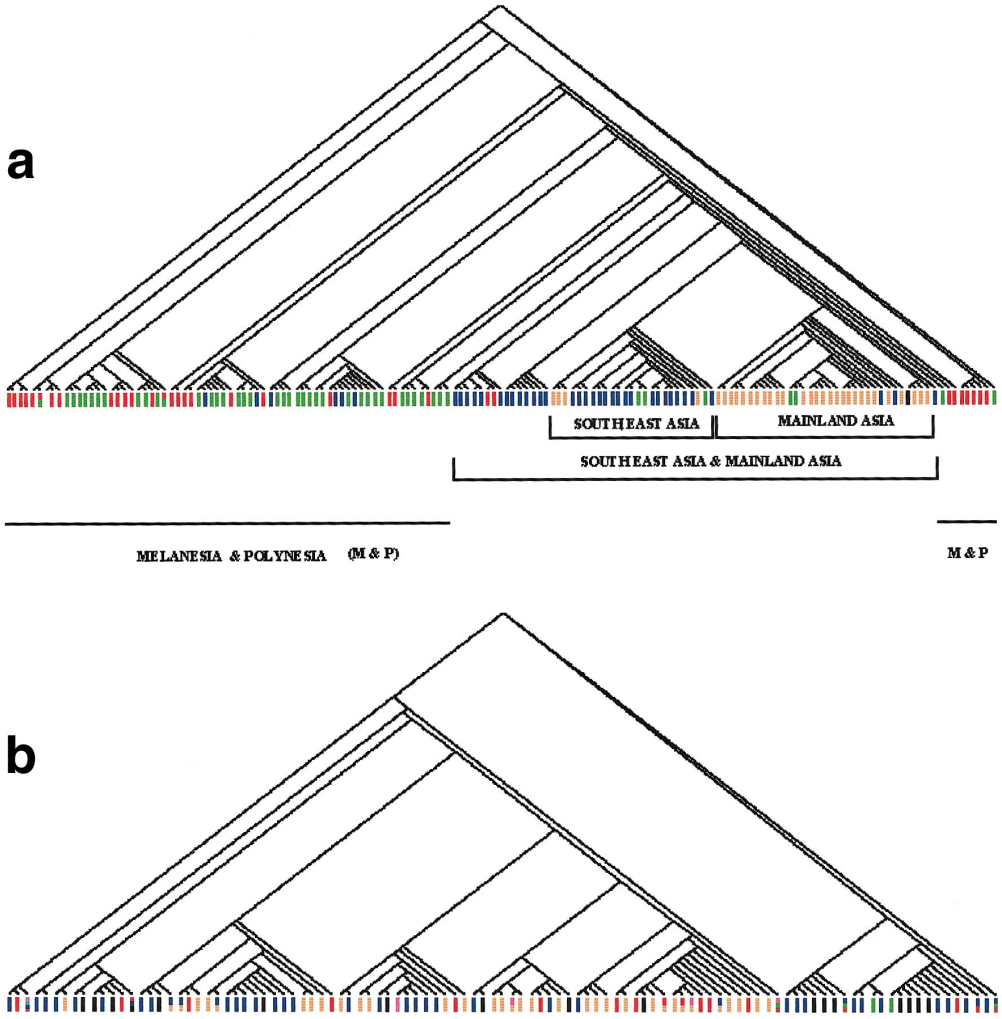

Figure 3 Genealogical tree of Y chromosome haplotypes in haplogroups C (*a*) and L (*b*). Neighbor-joining trees were reconstructed using the program NEIGHBOR (PHYLIP), on the basis of the average squared distance calculated by MICROSAT. Colored bars indicate haplotype geographic association as follows: red indicates Polynesia; green indicates Melanesia; blue indicates Southeast Asia; orange indicates southern China, India, and Mongolia; black indicates Taiwan; and purple indicates Philippines. Shared haplotypes across geographic regions are indicated by multicolored bars.

Our analysis of haplogroup L contradicts a recent claim, based on frequency, that the haplogroup was included among the founders of East Asia (Su et al. 1999). Analysis of the diversity within this haplogroup, however, indicates a clear association with one part of Asia and suggests that haplogroup L reached high frequency because of rapid population growth in the region. This makes clear that, in any effort to infer their ages, it is essential to consider diversity within lineages, because regionally varying patterns of population growth can have a dramatic influence on frequencies.

Haplogroup H is also at high frequency in the north, and it may therefore have been included in any dispersals to the south (table 1 and fig. 2). As for haplogroup L, there was extensive sharing of haplotypes within haplogroup H, across populations. In particular, 10 lineages were present in both the north and the south, representing 67 (33%) and 33 (39%) of the Taiwan/Chinese and southern chromosomes, respectively. However, the history of this lineage is more ambiguous than that of haplogroup L, because we do not observe the striking diversity, in any single group, that is seen, for the Ami, within haplogroup L. It is not clear, therefore, whether representatives of haplogroup H in the region are indigenous or associated with southward dispersal.

The Paternal Heritage of Southeast Asia, Melanesia, and Polynesia

The genealogies of haplogroups L and H are in sharp contrast to that of haplogroup C (fig. 3*a* and 3*b*), with

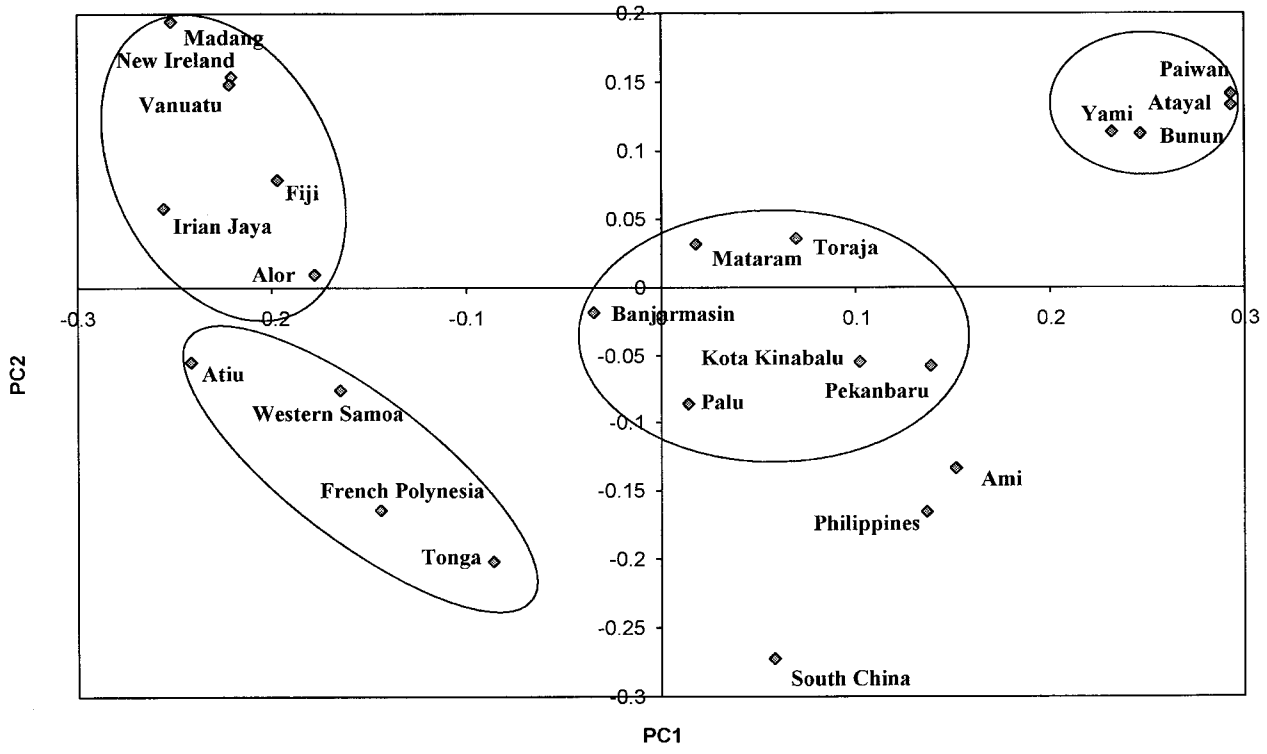

Figure 4 Graphic representation of the principal-component analysis of haplogroup frequencies

the likely explanation that the former haplogroups (especially L) have a northern origin, whereas the latter is indigenous. The picture for haplogroup G is less clear, although the genealogy does indicate that haplotypes from the north form a clearly identified subclade of the southern types (data not shown). Haplotypes outside this subset therefore appear to have an indigenous southern origin, consistent with the marginally greater diversity in Southeast Asia (table 2).

If we consider (1) haplogroups C, F, and E as indigenous in the south, (2) haplogroup L as introduced from northern populations, and (3) haplogroups H, G, I, D, and B as unassigned, then the genetic heritage of the investigated area can be estimated as follows: in Melanesia 93% of the observed chromosomes are indigenous, whereas only 3.6% have an origin in the north. Southeast Asia has a higher northern input, at 19.5%, with 32% identified as indigenous. Finally, Polynesia has 64% and 32% southern and northern contributions, respectively. It is clear, however, that these numbers are highly unrepresentative of the region, because of extensive genetic drift.

In short, more than half (519 of 807 [64%]) of the Y chromosomes present in the region appear to trace to the pre-Neolithic settlers who originally peopled the area. In insular Southeast Asia, 51% of the chromosomes can be assigned using the framework described above. Because haplogroup G shows suggestions of an indigenous origin in the south, information is completely lacking only for haplogroup H. Even if we consider haplogroup H as originating in the north, the chromosomes of northern origin that appear in insular Southeast Asia remain a minority at 43.5%. This indicates that the movement of people from the north was associated with extensive interbreeding with the previous inhabitants of the region. More striking is the result in Melanesia. If we consider only the Austronesian-speaking population included in our sample (New Ireland, Vanuatu, and Fiji), the northern types have a frequency of 9.25%, even when haplogroup H types are included. The pattern clearly indicates that in Melanesia the distribution of Austronesian languages was mainly a cultural process, with little or no northern genetic input; in other words, if there was a "train," it carried languages and technologies but few people.

In Southeast Asia, where the indigenous people may have been less densely distributed, the Austronesian expansion appears to have been both cultural and genetic.

A principal-component analysis (performed using POPSTR software by H. Harpending [personal communication]) is consistent with these interpretations (fig. 4). Along the first axis, an east-to-west difference dominates, with distinct clusters for Melanesian and Southeast Asian populations. This axis explains 46% of the

variation, consistent with our inference that differentiation within the region is very old. The haplogroups showing the greatest correlation coefficient with this axis were haplogroups H and C, at .64 and –.44, respectively. The second principal component, summarizing 20% of the total variation, follows a north-to-south gradient and appears to mark the influence of northern populations on the south. As expected, the haplogroups that have the most important influence on position on this axis are haplogroup L and haplogroup F (−.46 and −.22, respectively). Strikingly, it is only the southern Chinese and one Taiwanese aboriginal group that cluster at one end of this axis. The implication is that the northern influence in the south results from the spread of groups similar to the southern Chinese and the Ami, but not from the spread of other Taiwanese aboriginal groups, as is also indicated by the genealogical investigation of haplogroup L. It is interesting to note that a closer relationship between the language spoken by the Ami and the extra-Formosan Austronesian ones has been proposed (Ross 1994), although this suggestion is controversial (Blust 1999).

The Papuan Group

The Papuan group comprises >700 distinct languages that collectively are spoken by <3 million people (Ruhlen 1991), but their distribution is limited to New Guinea and nearby islands. The preservation of this group of languages may have resulted from population expansion associated with agricultural practice that developed in the region ⩾6,000 years ago (Ruhlen 1991). The distribution of the Papuan languages shows a striking correlation with haplogroup E. This haplogroup is mainly restricted to Melanesia, where its diversity is greater. The few types present in Polynesia or Southeast Asia are clearly associated with Melanesian ones, as is inferred by genealogical analysis. Outside Melanesia, the high frequency of haplogroup E in Alor is consistent with this interpretation, because Papuan languages are spoken in this region. The separation of Papuan and non-Papuan speakers is also partially mirrored by the cluster subdivision, in haplogroup C, between Melanesia and Southeast Asia and by the absence of shared types between Southeast Asia and Melanesia, within haplogroup C and haplogroup F. These observations suggest the presence of at least two differentiated indigenous groups in the region, one associated with Papuan speakers and another group centered to the west, in Southeast Asia. Superimposed on this structure is a modest and variable reintroduction of northern types that could be associated with a southward dispersal of agriculturalists. Interestingly, no northern types were observed in coastal New Guinea (Madang) in a sample including both Papuan and Austronesian speakers.

Polynesian Y Chromosome Composition

When European types are excluded, haplogroups C, F, and L account for 94% of the Polynesian chromosomes observed. The haplogroup C types were similar, showing a microsatellite variance of only .14. In particular, two main types, denoted "Polynesian modal haplotype A" (PMHA) and "Polynesian modal haplotype B" (PMHB) represented 56% of the total Polynesian haplogroup C chromosomes. (PMHA consists of 15, 14, 12, 16, 20, 10, 13, 11, 10, and 17 microsatellite-repeat counts and PMHB consists of 15, 14, 12, 16, 20, 10, 13, 11, 9, and 17 microsatellite-repeat counts; in both cases, the microsatellites are ordered as indicated in the Populations and Methods section). PMHA showed frequencies of 12.5% and 7% in Western Samoa and Tonga, respectively, and PMHB showed frequencies of 30% and 59% in French Polynesia and Atiu, respectively. It would appear, therefore, that much of Polynesia traces to very few founder chromosomes and that population growth in the region was subsequently very fast. This view is supported by the one-step network that connects 94% of Polynesian haplogroup C chromosomes with PMHA at the center (data not shown). The average square distance between current chromosomes and PMHA is 0.121 and, under a model of rapid growth, implies that growth began ~43 generations ago, or ~1,100 years BP (Kittles et al. 1998; Ruiz-Linares et al. 1999; Kayser et al. 2000). Similarly, haplogroup F showed types associated with Melanesian ones, but with a higher variance, 0.39. Again two types dominate, encompassing 60% of haplogroup F chromosomes observed in Polynesia. Finally haplogroup L showed wide sharing with Southeast Asian, Melanesian, and northern populations, with one haplotype present in all Polynesian islands, except in Atiu, and modal in Tonga (haplotype L85, consisting of 13, 13, 13, 15, 24, 10, 12, 11, 9, and 16).

In short, diversity within all three haplogroups is severely reduced, with a few haplotypes obtaining very high frequencies in one or more Polynesian population. The distribution of variation is consistent with a nested series of founder effects associated with the colonization of the Pacific, leading to enhanced frequencies of related types through much of Polynesia (e.g., the one-step cluster in haplogroup C) and also leading to enhanced frequencies of specific types in specific parts of Polynesia (e.g., the modal haplotype in Tonga haplogroup L and PMHB in French Polynesia and Atiu). This distribution implies that the detailed pattern of migration of the islands should be recoverable with a detailed study of a sufficiently large sample set; for example, any islands colonized from French Polynesia may be expected to include PMHB.

Given both the extensive drift in the history of Polyne-

sian populations and their proximate origin in the Melanesian islands, it is clear that models of the Austronesian expansion should be tested not by evaluation of Polynesian diversity but, rather, by the diversity present in the regions that served as staging grounds for the colonization of Polynesia; for example, the frequency of the northern-type haplogroup L is 57% in Tonga and only 3% in Atiu. The overall position of Polynesia, however, is summarized in the principal-component analysis (fig. 4). The first component associates Polynesians with Melanesian populations, whereas the northern agricultural contribution is highlighted by the second component, which associates Polynesians with Southeast Asians and the Ami and is closer to southern China. While there is a clear increase of northern types in Polynesia relative to Melanesia, the extensive genetic drift in Polynesia makes it difficult to draw any strong inferences based on this difference.

Finally, the only chromosomes not considered are four haplogroup E chromosomes from Tonga, two haplogroup D chromosomes, and one and three chromosomes of haplogroup H and haplogroup G, respectively. The low frequencies of these chromosomes in Polynesia make it difficult to infer their histories in this region.

Comparison with Other Genetic Systems

Although genetic analyses of the region have focused mainly on Polynesia, some data are available from other populations, allowing comparison between other marker systems and the Y chromosome data presented here. Analysis of mtDNA variation shows considerable homogeneity in Polynesia, with near fixation of the 9-bp deletion (Sykes et al. 1995). In this region the deletion is associated with a control-region motif (Polynesian motif) whose geographic distribution appears limited to Melanesia, Polynesia, eastern Indonesia, and Madagascar (Soodyall et al. 1995; Richards et al. 1998; Merriwether et al. 1999). More generally, the common types found in Polynesia are also observed in Melanesia.

Melton et al. (1995, 1998) have suggested a Taiwanese origin for the Polynesians, on the basis of mtDNA variation, but Richards et al. (1998) have argued that an East Indonesian origin is more likely. Autosomal data are largely consistent with the uniparental systems. Population specific alleles at the α-globin locus have revealed connections between Melanesia, Polynesia, and Southeast Asia (Hill et al. 1987), whereas HLA variation suggests a Polynesian connection to both Melanesia and mainland Asia (Hagelberg et al. 1999). Finally, analysis of autosomal microsatellites revealed a group including Polynesia, Melanesia and Australia and separated from East Asian populations, whereas mtDNA analysis of the same individuals grouped Melanesia, Australia, and East Asian populations (Lum et al. 1998), in contrast with distribution of the Polynesian motif.

In short, the overall picture supports an association between Polynesia and Melanesia and a distinction between Melanesia and Southeast Asia, possibly reflecting a long-term human presence in the region and at least partial independence, during this time, between Melanesia and Southeast Asia. Unfortunately, the current difficulty in determining the geographic origins of the more frequent mtDNA lineages makes it impossible to directly assess whether males and females have had different demographic roles in the peopling of the region. A more detailed genealogical study of mtDNA variation in the region should permit such a comparison.

Regional Continuity and Replacement Model

Genetic evaluation of the origins of anatomically modern humans have consistently supported the out-of-Africa model over the multiregional model, which posits genetic input from global archaic humans into the modern gene pool (Cavalli-Sforza et al. 1994). It has been pointed out, however, that most genetic studies would fail to detect a small input of archaic genetic material (Goldstein et al. 1996). Until now it has not been possible to systematically test for small levels of archaic input. Because of the security of Y chromosome genealogies, however, an archaic type would be easily identified by the absence of derived states that characterize non-African Y chromosomes. Of the chromosomes examined in this study, 48 were ancestral at M9 and RPS4Y and were further tested with SRY10381 (Whitfield et al. 1995). This marker defines a group that includes a large set of non-African types but that excludes ancestral lineages found only in Africa (Karafet et al. 1999). The 48 chromosomes were shown to be derived at this marker, definitively ruling out the inclusion of any archaic chromosomes in our sample. This result argues strongly against a significant genetic contribution of *H. erectus* to the current genetic composition of the region, at least on the paternal side.

Acknowledgments

We would like to thank all the people who donated their DNA for this work. We are grateful to John Clegg, for kindly providing samples, and to J. Jelsma, J. Pasveer, and G. J. Bartstra, for help with the collection of the Irian Jaya samples, and to J. Miedema and the Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Irian Jaya Studies program, for financial support. M.P.H.S. is grateful for a fellowship from the Wellcome Trust. We also thank Chris Tyler Smith and Vincent Macaulay, for helpful discussion and suggestions, and two anonymous reviewers, for comments on an earlier version of the manuscript.

References

Allen J (1996) The pre-Austronesian settlement of island Melanesia. In: Goodenough WH (ed) Prehistoric settlement of the Pacific. American Philosophical Society, Philadelphia, pp 11–27

Bellwood P (1997) Prehistory of the Indo-Malaysian archipelago. University of Hawaii Press, Honolulu

Bergen AW, Wang CY, Tsai J, Jefferson K, Dey C, Smith KD, Park SC, Tsai SJ, Goldman D (1999) An Asian-Native American paternal lineage identified by RPS4Y resequencing and by microsatellite haplotyping. Ann Hum Genet 63:63–80

Blust R (1999) Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics. In: Zeitoun E, Jen-kuei Li, P (eds) Selected papers from the Eighth International Conference on Austronesian Linguistics. Academia Sinica, Taipei, pp 31–94

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and the geography of human genes. Princeton University Press, Princeton, pp 60–67

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

Dyen I (1962) The lexicostatistical classification of Malayapolynesian languages. Language 38:38–46

——— (1965) A lexicostatistical classification of the Austronesian language. Int J Am Ling Memoir 19

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995*a*) An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463–471

——— (1995*b*) Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92:6723–6727

Goldstein DB, Zerjal T, Wilson JF, Pandya A, Santos FR, Thomas M, Tyler-Smith C. The rate of differentiation at human Y chromosome microsatellites shows a linear dependence on repeat count. Genetics (in press)

Goldstein DB, Zhivotovsky LA, Nayar K, Linares AR, Cavalli-Sforza LL, Feldman MW (1996) Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol Biol Evol 13: 1213–1218

Hagelberg E, Kayser M, Nagy M, Roewer L, Zimdahl H, Krawczak M, Lio P, Schiefenhovel W (1999) Molecular genetic evidence for the human settlement of the Pacific: analysis of mitochondrial DNA, Y chromosome and HLA markers. Philos Trans R Soc Lond B Biol Sci 354:141–152

Hill AV, Gentile B, Bonnardot JM, Roux J, Weatherall DJ, Clegg JB. (1987) Polynesian origins and affinities: globin gene variants in eastern Polynesia. Am J Hum Genet 40: 453–463

Hurles ME, Irven C, Nicholson J, Taylor PG, Santos FR, Loughlin J, Jobling MA, Sykes BC (1998) European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. Am J Hum Genet 63: 1793–1806

Lum JK, Cann RL, Martinson JJ, Jorde LB (1998) Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. Am J Hum Genet 63:613–624

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, Klitz W, Harihara S, de Knijff P, Wiebe V, Griffiths RC, Templeton AR, Hammer MF (1999) Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. Am J Hum Genet 64:817–831

Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, Kruger C, Krawczak M, Nagy M, Dobosz T, Szibor R, de Knijff P, Stoneking M, Sajantila A (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet 66: 1580–1588

Kittles RA, Perola M, Peltonen L, Bergen AW, Aragon RA, Virkkunen M, Linnoila M, Goldman D, Long JC (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. Am J Hum Genet 62:1171–1179

Klein RG (1999) The human career: human biological and cultural origins, 2d ed. University of Chicago Press, Chicago

Lahr MM, Foley RA (1994) Multiple dispersals and modern human origins. Evol Anthropol 3:48–60

Mathias N, Bayes M, Tyler-Smith C (1994) Highly informative compound haplotypes for the human Y chromosome. Hum Mol Genet 3:115–123

Melton T, Clifford S, Martinson J, Batzer M, Stoneking M (1998) Genetic evidence for the proto-Austronesian homeland in Asia: mtDNA and nuclear DNA variation in Taiwanese aboriginal tribes. Am J Hum Genet 63:1807–1823

Melton T, Peterson R, Redd AJ, Saha N, Sofro AS, Martinson J, Stoneking M (1995) Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. Am J Hum Genet 57:403–414

Merriwether DA, Friedlaender JS, Mediavilla J, Mgone C, Gentz F, Ferrell RE (1999) Mitochondrial DNA variation is an indicator of Austronesian influence in island Melanesia. Am J Phys Anthropol 110:243–270

Oppenheimer S (1998) Eden in the east: the drowned continent of Southeast Asia. Weidenfield & Nicholson, London

Pavlides C, Gosden C (1994) 35,000-year-old sites in the rain forests of West New Britain, Papua New Guinea. Antiquity 69:604–610

Richards M, Oppenheimer S, Sykes B (1998) mtDNA suggests Polynesian origins in eastern Indonesia. Am J Hum Genet 63:1234–1236

Roberts RG, Jones R, Smith MA (1990) Thermoluminescence dating of a 50,000-year-old human occupation site in northern Australia. Nature 345:153–156

Ross MD (1994) Some current issues in Austronesian linguistics. In: Tyron DT (ed) Comparative Austronesian dictionary. Mouton de Gruyter, Berlin, p 69

Ruhlen M (1991) A guide to the world's languages. Vol. 1. Stanford University Press, Stanford, pp 172–184

Ruiz-Linares A, Ortiz-Barrientos D, Figueroa M, Mesa N, Munera JG, Bedoya G, Velez ID, Garcia LF, Perez-Lezaun A, Bertranpetit J, Feldman MW, Goldstein DB (1999) Microsatellites provide evidence for Y chromosome diversity among the founders of the New World. Proc Natl Acad Sci USA 96:6312–6317

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462

Soodyall H, Jenkins T, Stoneking M (1995) Polynesian mtDNA in the Malagasy. Nat Genet 10:377–378

Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, Huang

W, Shen D, Lu D, Luo J, Chu J, Tan J, Shen P, Davis R, Cavalli-Sforza L, Chakraborty R, Xiong M, Du R, Oefner P, Chen Z, Jin L (1999) Y-chromosome evidence for a northward migration of modern humans into eastern Asia during the last Ice Age. Am J Hum Genet 65:1718–1724

Sullivan KM, Mannucci A, Kimpton CP, Gill P (1993) A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin. Biotechniques 15:636–638, 640–641

Swisher CC III, Rink WJ, Anton SC, Schwarcz HP, Curtis GH, Suprijo A, Widiasmoro (1996) Latest *Homo erectus* of Java: potential contemporaneity with *Homo sapiens* in southeast Asia. Science 274:1870–1874

Sykes B, Leiboff A, Low-Beer J, Tetzner S, Richards M (1995) The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. Am J Hum Genet 57:1463–1475

Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB (1998) Origins of Old Testament priests. Nature 394:138–140

Thomas MG, Bradman N, Flinn HM (1999) High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. Hum Genet 105:577–581

Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res 7:996–1005

Underhill PA, Shen P, Lin, AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonné-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ (2000) Y chromosome sequence variation and the history of human populations. Nat Genet 26:358–361

Whitfield LS, Sulston JE, Goodfellow PN (1995) Sequence variation of the human Y chromosome. Nature 378:379–380

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhovel W, Fretwell N, Jobling MA, Harihara S, Shimizu K, Semjidmaa D, Sajantila A, Salo P, Crawford MH, Ginter EK, Evgrafov OV, Tyler-Smith C (1997) Genetic relationships of Asians and northern Europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet 60:1174–1183