Y Chromosomal Evidence for the Origins of Oceanic-Speaking Peoples

Matthew E. Hurles,*[,1] Jayne Nicholson,[†] Elena Bosch,[‡] Colin Renfrew,* Bryan C. Sykes[†] and Mark A. Jobling[‡]

**McDonald Institute for Archaeological Research, University of Cambridge, Cambridge CB2 3ER, United Kingdom, [†]Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, United Kingdom and [‡]Department of Genetics, University of Leicester, Leicester LE1 7RH, United Kingdom*

Manuscript received May 23, 2001
Accepted for publication October 12, 2001

ABSTRACT

A number of alternative hypotheses seek to explain the origins of the three groups of Pacific populations—Melanesians, Micronesians, and Polynesians—who speak languages belonging to the Oceanic subfamily of Austronesian languages. To test these various hypotheses at the genetic level, we assayed diversity within the nonrecombining portion of the Y chromosome, which contains within it a relatively simple record of the human past and represents the most informative haplotypic system in the human genome. High-resolution haplotypes combining binary, microsatellite, and minisatellite markers were generated for 390 Y chromosomes from 17 Austronesian-speaking populations in southeast Asia and the Pacific. Nineteen paternal lineages were defined and a Bayesian analysis of coalescent simulations was performed upon the microsatellite diversity within lineages to provide a temporal aspect to their geographical distribution. The ages and distributions of these lineages provide little support for the dominant archeo-linguistic model of the origins of Oceanic populations that suggests that these peoples represent the Eastern fringe of an agriculturally driven expansion initiated in southeast China and Taiwan. Rather, most Micronesian and Polynesian Y chromosomes appear to originate from different source populations within Melanesia and Eastern Indonesia. The Polynesian outlier, Kapingamarangi, is demonstrated to be an admixed Micronesian/Polynesian population. Furthermore, it is demonstrated that a geographical rather than linguistic classification of Oceanic populations best accounts for their extant Y chromosomal diversity.

THE island populations of the Pacific Ocean have historically been divided, on the basis of geography and culture, into Polynesians, Micronesians, and Melanesians (Bellwood 1989). According to this system Polynesians occupy islands within a triangle defined by apices at New Zealand, Hawaii, and Rapanui (Easter Island). Melanesians occupy the islands farther to the west (including Papua New Guinea), and Micronesians occupy the coral atolls that lie to the north of Melanesia. Within Melanesia and Micronesia lie a number of islands whose populations seem to share more cultural (including linguistic) features with Polynesians than with their geographical neighbors; these "Polynesian outliers" are thought to originate from recent back migrations from Polynesia (Bellwood 1989).

The settlement history of the Pacific islands divides into two distinct phases. An early phase lasting until 28,000 YBP saw the first colonization of Papua New Guinea and some of the neighboring more easterly islands that make up the western part of present-day Island Melanesia. The second phase was initiated by a rapid occupation of islands farther to the east associated with the Lapita ceramic culture, whose sites range from New Britain to the Polynesian islands of Tonga and Samoa between 3300 and 2700 YBP (Spriggs 1989, 1999). After a time lag of at least 1000 years, colonization of the more remote islands of Central and Eastern Polynesia began (Spriggs and Anderson 1993; Spriggs 1999). The prehistory of the islands of central and eastern Micronesia is less well known. Human occupation of these islands dates back at least 2000 years but the pottery found thus far gives little clue as to the ultimate ancestry of these populations (Irwin 1992) although eastern Melanesia has been suggested as a potential source population (Davidson 1988).

An alternative way of distinguishing Pacific populations has been proposed; it focuses on the linguistic and settlement histories of the islands and divides the region into those areas first occupied pre-Lapita, "Near Oceania," and those occupied post-Lapita, "Remote Oceania" (Kirch and Green 1992). The genetic validity of these alternative systems has not yet been tested.

Polynesian languages are closely related to each other and belong to the Oceanic subgroup of the Austronesian language family (Pawley and Ross 1993). The Oceanic subgrouping also includes the nuclear Micronesian languages of central-eastern Micronesia and the Austronesian languages spoken throughout Island Melanesia and the eastern half of coastal Papua New Guinea. The branching order of these various subgroups is unresolved (Green

[1]*Corresponding author:* McDonald Institute for Archaeological Research, University of Cambridge, Downing St., Cambridge CB2 3ER, United Kingdom. E-mail: meh32@cam.ac.uk

1999). The 1000–1200 languages belonging to the Austronesian language family are spoken in a continuum throughout Island Southeast (SE) Asia into Island Melanesia (as distinct from Papua New Guinea) and Micronesia and out into the remote Pacific Islands (Bellwood 1991). Austronesian languages are not the only ones spoken in these regions. Another group of highly diverse languages is also spoken, mainly in Melanesia. Dubbed "Papuan," this group is distinguished more through not being Austronesian than through shared characteristics (Pawley and Ross 1993). In Melanesia, Austronesian languages are largely restricted to coastal regions of New Guinea and the islands. The greatest diversity within Austronesian languages is apparent in Taiwan (Blust 1999). This, and the phylogenetic arrangement of Austronesian languages (Gray and Jordan 2000), has led to the hypothesis of a rapid movement of a relatively homogenous people through Melanesia and into Polynesia, fueled by the expansions of a Neolithic culture out of southeast China and Taiwan ~6000 years ago (Bellwood 1997). This remains the current dominant archeo-linguistic model for the origins of Pacific islanders.

Setting aside an American origin for the Polynesians (Heyerdahl 1950), there remain alternative hypotheses for the SE Asian origins of Pacific peoples (Oppenheimer 1998). Solheim argues on the basis of pottery typology for an Austronesian homeland in the islands of northeastern Indonesia and southern Philippines (Solheim 1996). Meacham argues for a more diffuse homeland covering the entirety of Island SE Asia (Meacham 1985).

Prior to recent Y chromosomal work, the best genetic evidence for the origins of Pacific peoples has come from the maternally inherited mitochondrial DNA (mtDNA), which clearly indicates a SE Asian origin with little Melanesian admixture into Polynesians (Redd *et al.* 1995; Sykes *et al.* 1995). Genetic evidence for the location of the Austronesian homeland more specifically within SE Asia has proved contentious, with phylogenetic topological evidence supporting Taiwan (Melton *et al.* 1995, 1998) but considerations of mtDNA intralineage diversity highlighting eastern Indonesia (Richards *et al.* 1998).

The human Y chromosome is nonrecombining over most of its length and thus contains potentially the most informative haplotypic system within the human genome (Jobling and Tyler-Smith 1995). By revealing the record of paternal ancestry, the Y chromosome complements the maternal history of a population gathered from mtDNA. The observed high degree of geographic differentiation of Y chromosomal diversity has been explained by mating practices, the cultural phenomenon of patrilocality, and the small effective population size of the Y chromosome (Seielstad *et al.* 1998) and has been utilized to investigate prehistoric migrations (*e.g.*, Zerjal *et al.* 1997; Santos *et al.* 1999).

The only known hypervariable minisatellite on the nonrecombining portion of the human Y chromosome, MSY1, is particularly informative in Oceania (Hurles *et al.* 1998). This locus comprises an array of 50–100 tandem repeats of a 25-bp palindromic sequence. Three common repeat sequence variants are generally found in blocks of different sizes within arrays. The order of blocks along an array defines its modular structure, which normally consists of three to six blocks. This locus mutates at a rate of ~6% per generation, mostly through single-step changes of repeat numbers within such blocks (Jobling *et al.* 1998). Consequently, these blocks of different repeat sequence variants can be analyzed in a fashion analogous to microsatellites (Hurles *et al.* 1999).

MSY1 is also capable of undergoing saltatory mutations and it is these much rarer events that allow us to define monophyletic subgroups (Hurles *et al.* 1998; Jobling *et al.* 1998; Kalaydjieva *et al.* 2001).

A recent study used a genealogical approach to analyzing paternal lineages in Island SE Asia and the Pacific by defining lineages within the Y chromosome by using binary markers and subsequently assaying intralineage diversity with more mutable microsatellites to provide a temporal framework to the geographical patterns of lineage distributions (Kayser *et al.* 2000a). This study contended, in contrast to a prior study that used only binary markers (Su *et al.* 2000), that the majority of Polynesian Y chromosomes, characterized by a unique deletion within the *DYS390* microsatellite, originated in Melanesia or eastern Indonesia. However, both of these publications assayed only diversity within a single true Polynesian population.

Here, MSY1 is assayed, together with Y chromosomal binary markers and microsatellites, in all three of the groups of Pacific populations and in other Austronesian-speaking populations from Island SE Asia, to address some of the issues identified above.

MATERIALS AND METHODS

Samples: The DNA samples used in this study were provided by 390 individuals from 17 locations in the Pacific, all of whom had agreed to take part in a genetic survey. Taiwanese samples were from four aboriginal groups: Ami, Atayal, Bunumi, and Paiwan. The Filipino sample came from Luzon. Northern Borneo samples were from Kota Kinabalu and southern Borneo samples from Banjarmasin. Micronesian samples came from Majuro in the Marshall Islands of eastern Micronesia. Polynesian samples came from Western Samoa, Rarotonga in the Cook Islands, Tonga, and the outlier population on Kapingamarangi. The Tongan sample was composed of two different general Tongan samples and a third sample from Vavua. Melanesian samples came from Port Moresby in Papua New Guinea and two populations in Vanuatu from Maewo and Port Olry. Some of the data on Cook Islanders and Papua New Guineans were described previously (Hurles *et al.* 1998, 2001).

Polymorphic marker typing: All of the binary markers have been described previously and were typed using 10–20 ng of DNA in PCR protocols on an MJR PTC-200 thermocycler: YAP

(Hammer 1994) was typed according to Hammer and Horai (1995), *SRY*-1532 (Whitfield *et al.* 1995) according to Kwok *et al.* (1996), *SRY*-2627 according to Veitia *et al.* (1997), *DYS257*, which is phylogenetically equivalent to 92R7 (Rosser *et al.* 2000), according to Hammer *et al.* (1998), *DYS199* (Underhill *et al.* 1996), M4, and M9 (Underhill *et al.* 1997) according to Hurles *et al.* (1998), Tat according to Zerjal *et al.* (1997), and 12f2 (Casanova *et al.* 1985) according to Blanco *et al.* (2000). *RPS4Y* (Bergen *et al.* 1999) was typed as an allele-specific amplification in a touchdown protocol: the allele-specific primers 5′-TGGCAATAAACCTTGGATTTCT-3′ (specific for the A allele) and 5′-TGGCAATAAACCTTGGATT TCC-3′ (specific for the G allele) were used in conjunction with the nonspecific primer 5′-CACAAGGGGGAAAAAACAC-3′ to selectively amplify a fragment of 184 bp, the presence of which was ascertained by agarose electrophoresis. The PCR protocol was as follows: 4 min at 94° followed by 4 cycles of 94° for 30 sec, 68° for 30 sec (−1.0° per cycle), 72° for 30 sec, and then 30 cycles of 94° for 30 sec, 64° for 30 sec, and 72° for 30 sec. The LlY22g *Hin*dIII polymorphism was typed by a PCR-restriction fragment length polymorphism assay, which will be described elsewhere (E. Righetti and C. Tyler-Smith, unpublished results). The deep-rooting marker M9 was typed on all samples, and remaining markers were typed hierarchically according to the known phylogeny for these markers.

Three-state MSY1 MVR-PCR of repeat types 1, 3, and 4 was carried out according to Jobling *et al.* (1998). A code, for example, (1)20(3)35(4)20, represents the minisatellite array as blocks of different repeat unit variants; in this case, 20 type 1 repeats were followed by a block of 35 type 3 repeats and then 20 type 4 repeats. Modular structure nomenclature of, for example, the form (**1**, **3**, **4**) refers to a block of type 1 repeats followed by a block of type 3 repeats and then a block of type 4 repeats.

Six tetranucleotide repeat microsatellites (*DYS19*, *DYS389I*, *DYS389II*, *DYS390*, *DYS391*, and *DYS393*) and a single trinucleotide repeat microsatellite (*DYS392*) were typed on the majority of samples as described previously (Hurles *et al.* 1998). The remaining data were generated using multiplexes to be described elsewhere (E. Bosch and M. A. Jobling, unpublished results). The full data set is available from the authors on request.

Analysis: Neighbor-joining (NJ) and unweighted pair-group method using arithmetic averages (UPGMA) trees were constructed using the Neighbor program within the PHYLIP package (Felsenstein 1995). Weighted haplotypic distance matrices were generated as input for the PHYLIP programs by use of a program written by M. E. Hurles in Interactive Data Language 5.3 (IDL). Median-joining networks were constructed using the program Network 2.0c. The "*.mat" output file from the reduced median (RM) algorithm was used as input for the median-joining (MJ) algorithm. This reduces the ability of the median-joining algorithm to produce large, phylogenetically unrealistic cycles within the network (Forster *et al.* 2000; P. Forster, personal communication). Consequently, the loci were input into the RM algorithm in order of decreasing weight to ensure the stability of the least mutable loci. The weighting scheme of the loci was calculated on a lineage-by-lineage basis from the amount of intralineage variance displayed by each locus. The weights were apportioned relatively within a range of 1–10, with higher weights going to the least variable and thus slower mutating loci. An alternative form of weighting based on the observed numbers of mutations within pedigrees was not used, as this does not take account of the fact that the founder allele within some lineages is significantly smaller and thus less mutable than the alleles followed through pedigrees. For example, the *DYS390* locus has the highest pedigree mutation rate (Kayser *et al.* 2000b) and accordingly in most of the lineages under study here has the highest variance of all the microsatellite loci; however, in haplogroup (hg) 10 the allele lengths are, on average, 5.7 repeats smaller than those in which this pedigree rate was ascertained, and, correspondingly, *DYS390* has the lowest variance of all loci within this lineage. The weights assigned to each locus were supported by the posterior distributions for locus-specific mutation rates obtained for each lineage from the BATWING analysis, despite the fact that the prior probability distributions for these rates were based on the pedigree data. In the case of the MJ networks weights can be set for individual allelic transitions within a locus. This feature was used for blocks of MSY1 repeats that cover a large range of allele sizes; for example, type 4 repeats at the 3′ end of the repeat array range from 4 to 23 repeats within hg 26 chromosomes with the (**1, 3, 4**) or (**3, 1, 3, 4**) modular structures. Block size is closely correlated to mutability, and consequently when ranges exceeded a factor of 2 from largest to smallest, the shorter half of the range was given twofold greater weight than the longer half.

Sixty-four chromosomes belonging to lineages 26.1, 26.4, and 26.6 have been typed with binary markers M95, M119, and M122 in a previous study (Capelli *et al.* 2001). Lineage 26.1 is a sublineage of M95-derived chromosomes, 26.4 is a sublineage of M122-derived chromosomes, and 26.6 is a sublineage of M119-derived chromosomes. A single M122-derived chromosome has been assigned to lineage 26.6, indicating a lack of congruence between the prior study and the present one.

Bayesian coalescent analysis was performed using the program BATWING (Wilson *et al.* 2000), written by I. Wilson, M. Weale, and D. Balding, which uses a Markov-chain Monte Carlo method (Wilson and Balding 1998) to derive posterior distributions for a complete set of parameters that describe the relevant underlying model. The model used here is one that incorporates both population subdivision and a growth model that allows a period of constant size (N) prior to exponential growth. A total of 100,000 tree rearrangements were discarded as "burn in" and the posterior distributions for each parameter were estimated from 2000 sparse samplings from the subsequent 2×10^5 rearrangements. The median and equal-tailed 95% interval limits were calculated for each parameter. Prior distributions for the mutation rate at each locus used a gamma distribution conditioned on the observed pedigree mutation from Kayser *et al.* (2000b). The prior for the initial population size was a gamma distribution with a median of 49 and even-tailed 95% interval limits of 0.002–1266. Priors for the growth rate, age of expansion, and time of the first population split were exponential distributions with mean 1. These parameters varied widely in a number of test simulations to show that the resulting posterior distributions are robust to changing the priors and thus result from patterns within the data and not from restrictive prior distributions. Time is measured in units of $N \times$ generation time, and to generate absolute ages a generation time of 25 years was used.

Principal components were calculated using a program written by M. E. Hurles in IDL. Analysis of molecular variance (AMOVA), Mantel tests, genetic distances, and diversity indices were calculated using Arlequin 2.0 (Schneider *et al.* 2000).

RESULTS

The 10 binary markers typed here define 12 monophyletic lineages, or haplogroups, on the single most parsimonious phylogeny of Y haplotypes shown in Fig-

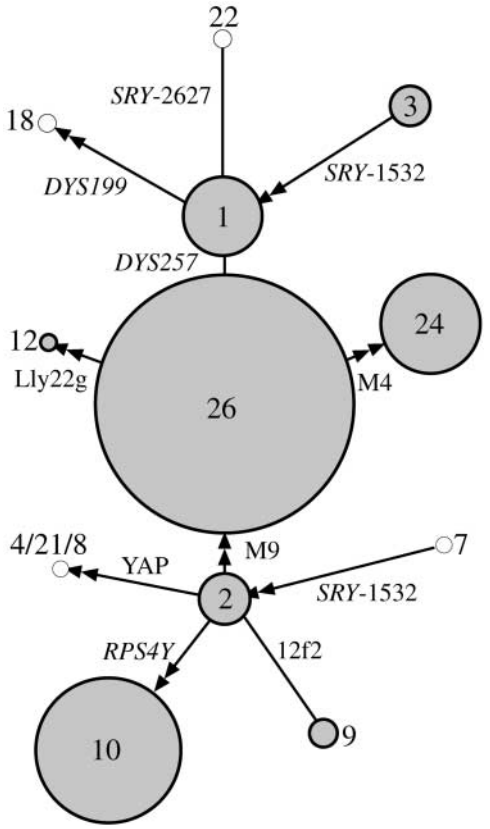

FIGURE 1.—Maximum parsimony tree of Y chromosomal binary marker haplotypes. Circles indicate haplogroups, which, if shaded, are found in the current data set. Circle area is proportional to frequency. Numbers next to circles indicate the nomenclature of JOBLING and TYLER-SMITH (2000). Labels next to the lines indicate the binary marker that distinguishes each haplogroup from its neighbor. The arrows point from ancestral to derived states of the markers where known.

ure 1. Eight of these 12 haplogroups are observed in our 390 samples. There are 227 different seven-locus microsatellite haplotypes and 291 different MSY1 codes among this same number of samples. Thus MSY1 codes are more variable than seven microsatellites, and combining MSY1 codes and microsatellites should give haplotypes that are at least as informative as 14 linked microsatellites of comparable allelic diversity. There are 323 such compound multiallelic haplotypes among these 390 chromosomes, none of which are shared between chromosomes of different haplogroups. Two haplogroups predominate in the Pacific, hg 10 and hg 26, which together account for 82% of the total, and it is within these two haplogroups that the Y chromosome ancestry of the region is to be read.

Haplogroup 26: Haplogroup 26 chromosomes comprise 63.3% of the total. They are defined by an ancient mutation, M9, the derived form of which is found all over Eurasia, and at highest frequencies in east Asia (UNDERHILL *et al.* 1997). A previous study has demonstrated the existence of a monophyletic sublineage within hg 26 Y chromosomes in Polynesia on the basis of a novel MSY1 repeat array structure. It is characterized by a large expansion within a block of type 3 repeats and a concomitant deletion within the block of type 4 repeats at the 3′ end of the array (HURLES *et al.* 1998). These chromosomes were named "26 (**3, 1, 3+, 4-**)."

In principle a number of different multivariate and phylogenetic approaches are capable of revealing the distinct clusters of related MSY1 codes that result from such saltatory mutations. Here, a median-joining network (not shown) was constructed on the set of MSY1 codes comprising the 224 hg 26 chromosomes with either (**1, 3, 4**) or (**3, 1, 3, 4**) MSY1 modular structures (91% of the total). Seven distinct clusters containing >5 related chromosomes that may represent monophyletic lineages were identified. One of these clusters contained all of the chromosomes belonging to the 26 (**3, 1, 3+, 4-**) lineage identified previously. It is necessary to test whether these clusters are indeed monophyletic or if they are composed of different lineages resulting from recurrent saltatory mutation. Recurrent saltatory mutation within such a deep-rooting lineage is likely to have occurred on different haplotypic backgrounds, as defined by Y microsatellites. In this case, when phylogenies are constructed from compound multiallelic haplotypes comprising both the microsatellite alleles and the MSY1 codes, the clusters of chromosomes based on MSY1 codes alone should not form single clades. To compensate for the high mutation rate of MSY1, which might bias such an analysis toward retaining MSY1 code clusters as clades, the blocks of MSY1 repeats were downweighted with respect to the microsatellite loci. Three different phylogenetic reconstruction methods were applied to the set of hg 26 chromosomes with either (**1, 3, 4**) or (**3, 1, 3, 4**) MSY1 modular structures. An NJ tree and a UPGMA tree were constructed from weighted haplotypic distance matrices. MJ networks were constructed from the output of the reduced median algorithm, as suggested by the authors of this method for reconstructing trees with longer branch lengths (PETER FORSTER, personal communication). The construction of the MJ network was also weighted so as to allow the microsatellite data to break up any polyphyletic MSY1 structures, should they exist (see MATERIALS AND METHODS for details).

All of the clusters formed by MSY1 codes alone were reconstructed as clades by all three phylogenetic methods when data from the microsatellite loci were incorporated, demonstrating that recurrent saltatory mutation of MSY1 had not occurred. The NJ tree is shown in Figure 2. It can be seen that all highlighted clades are characterized by short mean internal branch lengths relative to those that separate the clade from the rest of the tree. Diagnostic MSY1 codes associated with each lineage, labeled 26.1–26.7, are also shown in Figure 2. Lineage 26.4 is characterized by a massive expansion of type 3 repeats and a deletion of type 4 repeats and was previously known as "26 (**3, 1, 3+, 4-**)" (HURLES *et al.*

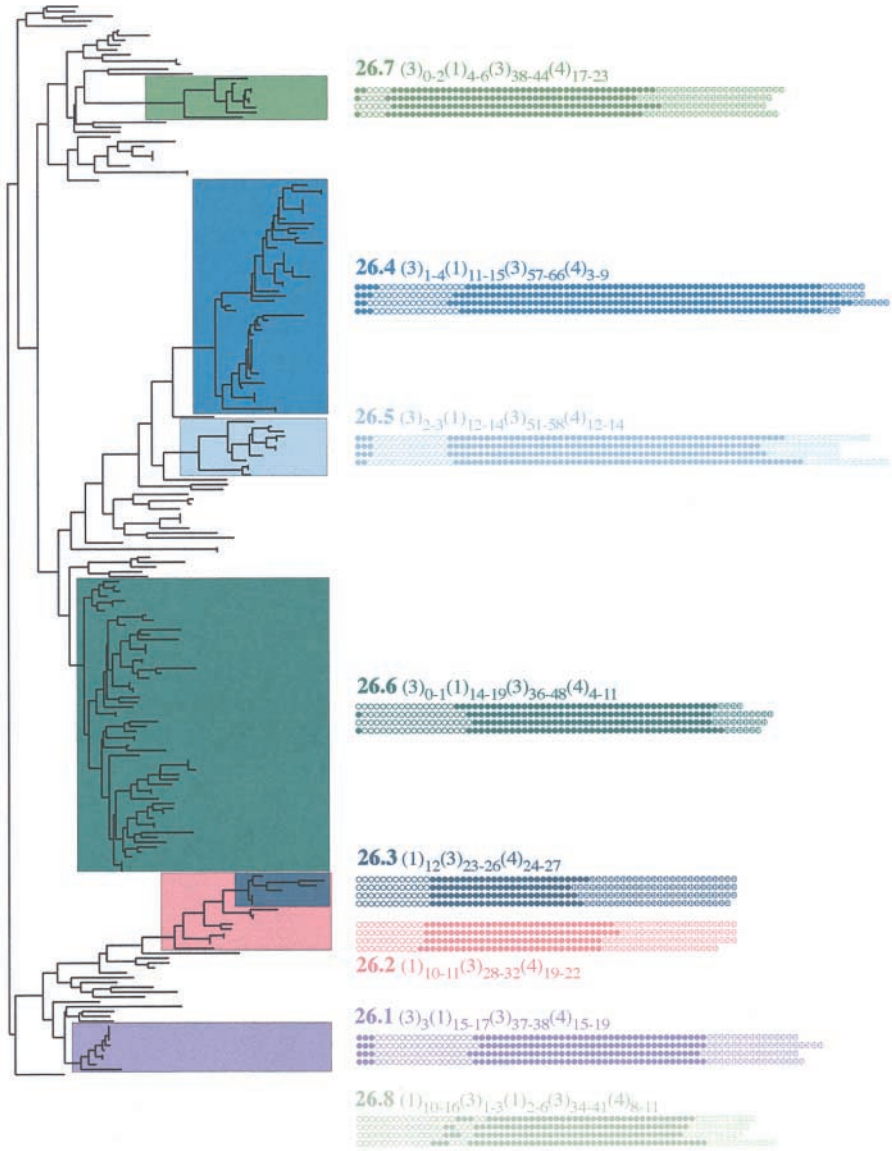

FIGURE 2.—A neighbor-joining tree of 224 chromosomes belonging to haplogroup 26. This unrooted tree was constructed from distances between haplotypes comprising seven microsatellites and MSY1 codes, weighted according to the mutation rate of each locus. Seven lineages defined by saltatory mutations in MSY1 form well-defined clades within the tree. These clades are labeled together with four diverse MSY1 codes from each lineage to indicate the diagnostic minisatellite structures. Open circles indicate type 1 repeats, solid circles indicate type 3 repeats, and shaded circles indicate type 4 repeats. An eighth lineage, 26.8, discussed in the text is included for comparison.

1998). A subset of chromosomes ($N = 64$) belonging to these lineages has been typed with additional binary markers in a published study (CAPELLI *et al.* 2001); with a single exception, these chromosomes have been assigned to lineages in a manner consistent with their being monophyletic clades (see MATERIALS AND METHODS). Removal of this aberrant chromosome from further calculations makes no change to the inferences drawn. These data provide an independent test for the validity of the lineage definitions above.

Eight different MSY1 modular structures are among the remaining 9% of hg 26 chromosomes. Six of these occur in only one to three chromosomes each. A further lineage (26.8) was defined on the basis of a cluster of six MSY1 codes within the seventh modular structure, namely, one with an insertion of two to six type 1 repeats within a central block of type 3 repeats, (**1, 3, 1, 3, 4**); see Figure 2. The final modular structure (**3, 1, 3, 1, 3, 4**) is found on eight chromosomes but, on the basis of unrelated MSY1 codes and microsatellite haplotypes, was not defined as a lineage because it seems to have arisen multiple times. All the monophyletic lineages defined within hg 26 have coherent geographical distributions, which are shown in Figure 3.

Haplogroup 10: In contrast to hg 26, hg 10 can be split qualitatively into monophyletic lineages on the basis of MSY1 modular structure alone. The insertion of a block of null repeats into the block of type 4 repeats at the 3′ end of the array has previously been identified as a monophyletic lineage (HURLES *et al.* 1998). These chromosomes are also distinguished by a single null repeat at the 5′ end of the array. All chromosomes within this lineage, named 10.2, have short alleles of 19–21 repeats at the *DYS390* locus and thus represent a sublineage of the *DYS390.3* deletion lineage identified by others (FORSTER *et al.* 1998; KAYSER *et al.* 2000a). An ancestral sublineage to 10.2, named 10.1, is defined here by the presence of short *DYS390* allele lengths

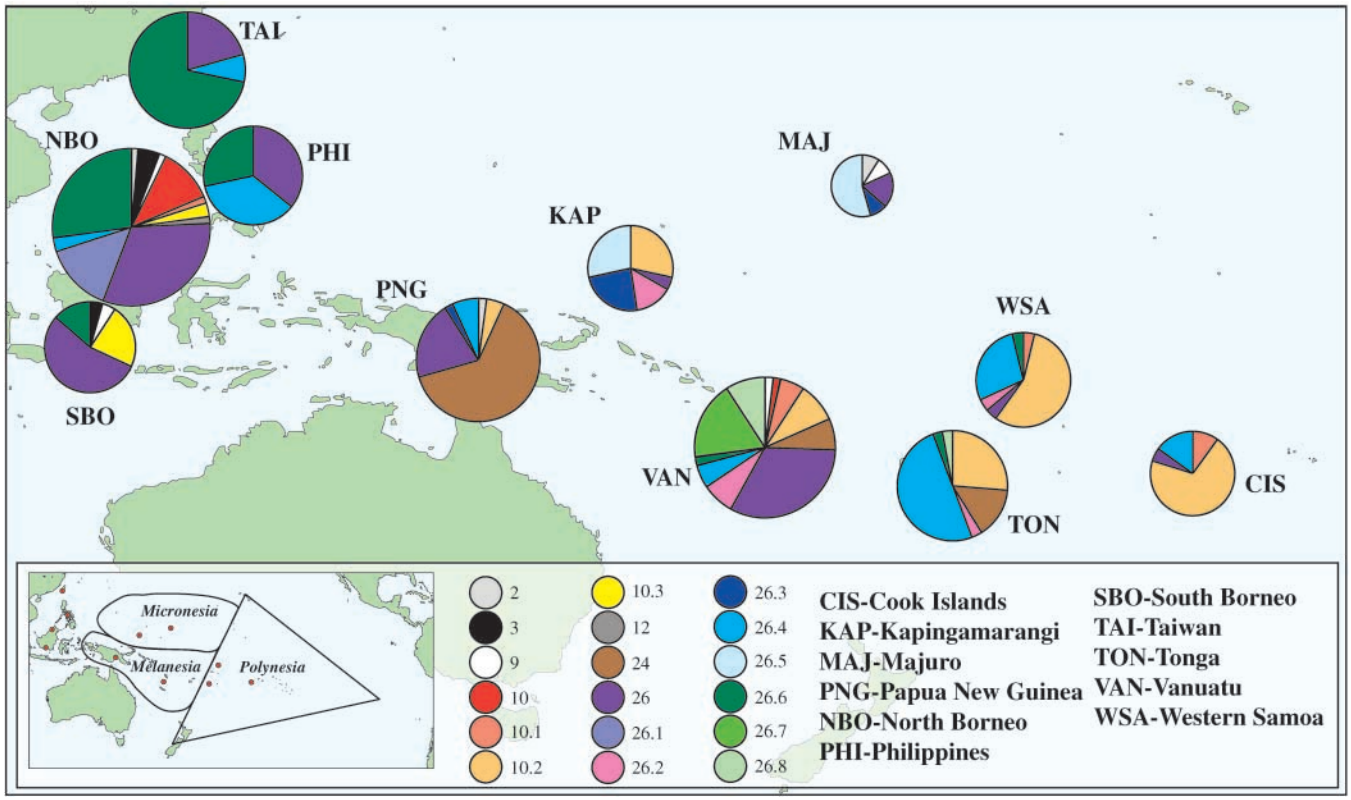

FIGURE 3.—Map of Oceania and SE Asia indicating Y chromosomal lineage frequencies in each of the 11 populations. Circle area is proportional to sample size. The inset map indicates the three geographical regions of the Pacific into which each population falls.

(19–21 repeats) and the null repeat at the 5′ end of the MSY1 repeat array, but the absence of the block of null repeats within the block of type 4 repeats at the 3′ end of the MSY1 repeat array. The lineage 10.1 chromosomes exhibit greater multiallelic diversity than those in lineage 10.2. It is likely that the *DYS390*.3 deletion is ancestral to the divergence of 10.1 and 10.2 as more chromosomes within Melanesia and Indonesia have short *DYS390* alleles (20–22 repeats) with a variety of MSY1 modular structures. A third lineage within hg 10, named 10.3, is defined by another MSY1 modular structure, (**1, 3, 4**), and all chromosomes have closely related microsatellite haplotypes and MSY1 block sizes.

Lineage 10.2 is the most frequent single lineage found in Polynesia. It extends, at much lower frequencies, westward into Melanesia but not into Indonesia. Lineage 10.1, the ancestral lineage to 10.2, is much less frequent in Polynesia than 10.2 although it is found at similar frequencies to 10.2 in Melanesia. A single representative is in northern Borneo. Lineage 10.3 is found only in Borneo, in both the northern and southern populations. Haplogroup 10 is completely absent from both the Filipino and Taiwanese samples.

Haplogroup 24: Haplogroup 24 is defined by the derived state of the M4 binary marker and has previously been found at high frequencies in Papua New Guinea and at lower frequencies in Island Melanesia and eastern Indonesia (HURLES *et al.* 1998; KAYSER *et al.* 2001). In this study haplogroup 24 is found in three populations: Papua New Guinea [64% (28/44)], Vanuatu [7% (4/55)], and Tonga [15% (5/34)].

Identifying admixture: Prior to making prehistorical inferences it is necessary to exclude chromosomes that originate from recent admixture with exogenous populations and that have been observed at high frequency in some Oceanic samples (HURLES *et al.* 1998). Since European contact in the 16th century there has been considerable introgression of distinctively European Y chromosomes into the Pacific Islands. Three lineages predominate in northwestern Europe (ROSSER *et al.* 2000): hg 1 chromosomes with the MSY1 modular structure (**1, 3, 4**), hg 2 chromosomes with the MSY1 modular structure (**3, 1, 3, 4**), and hg 3 chromosomes (HURLES *et al.* 1998; JOBLING *et al.* 1998). Of these three lineages, hg 1 is present at highest frequencies and hg 3 at the lowest. These lineages are also know to occur outside Europe, notably on the Indian subcontinent (HURLES *et al.* 1999; ZERJAL *et al.* 1999), although here other MSY1 subtypes predominate within these binary haplogroups (HURLES *et al.* 1998). We adopted a stringent approach to identifying admixed chromosomes by removing from future analysis all hg 1 (**1, 3, 4**) chromosomes, all hg 2 (**3, 1, 3, 4**) chromosomes, and hg 3 chromosomes when found in the same location as hg 1 and 2 chromosomes

TABLE 1

Lineage frequencies for each of the 11 populations

		Haplogroup lineage																		
		1	2	3	9	10				12	24	26								
Population	*N*					10	10.1	10.2	10.3			26	26.1	26.2	26.3	26.4	26.5	26.6	26.7	26.8
Cook Islands	20	0.0	0.0	0.0	0.0	0.0	10.0	70.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	15.0	0.0	0.0	0.0	0.0
	(32)	(25.0)	(9.4)	(3.1)	(0.0)	(0.0)	(6.3)	(43.8)	(0.0)	(0.0)	(0.0)	(3.1)	(0.0)	(0.0)	(0.0)	(9.4)	(0.0)	(0.0)	(0.0)	(0.0)
Western Samoa	25	0.0	0.0	0.0	0.0	0.0	4.0	56.0	0.0	0.0	0.0	4.0	0.0	4.0	0.0	28.0	0.0	4.0	0.0	0.0
Tonga	34	0.0	0.0	0.0	0.0	0.0	0.0	26.5	0.0	0.0	14.7	0.0	0.0	2.9	0.0	50.0	0.0	2.9	0.0	2.9
	(35)	(0.0)	(2.9)	(0.0)	(0.0)	(0.0)	(0.0)	(25.7)	(0.0)	(0.0)	(14.3)	(0.0)	(0.0)	(2.9)	(0.0)	(48.6)	(0.0)	(2.9)	(0.0)	(2.9)
Kapingamarangi	21	0.0	0.0	0.0	0.0	0.0	0.0	28.6	0.0	0.0	0.0	4.8	0.0	14.3	23.8	0.0	28.6	0.0	0.0	0.0
Majuro	11	0.0	9.1	0.0	9.1	0.0	0.0	0.0	0.0	0.0	0.0	18.2	0.0	0.0	9.1	0.0	54.5	0.0	0.0	0.0
Vanuatu	55	0.0	0.0	0.0	1.8	1.8	5.5	9.1	0.0	0.0	7.3	32.7	0.0	7.3	0.0	5.5	0.0	1.8	18.2	9.1
	(58)	(5.2)	(0.0)	(0.0)	(1.7)	(1.7)	(5.2)	(8.6)	(0.0)	(0.0)	(6.9)	(31.0)	(0.0)	(6.9)	(0.0)	(5.2)	(0.0)	(1.7)	(17.2)	(8.6)
Papua New Guinea	44	0.0	2.3	0.0	0.0	0.0	0.0	4.5	0.0	0.0	63.6	20.5	0.0	0.0	2.3	6.8	0.0	0.0	0.0	0.0
South Borneo	22	0.0	0.0	4.5	4.5	0.0	0.0	0.0	22.7	0.0	0.0	54.5	0.0	0.0	0.0	0.0	0.0	13.6	0.0	0.0
	(23)	(0.0)	(4.3)	(4.3)	(4.3)	(0.0)	(0.0)	(0.0)	(21.7)	(0.0)	(0.0)	(52.2)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(13.0)	(0.0)	(0.0)
North Borneo	70	0.0	1.4	4.3	1.4	11.4	1.4	0.0	2.9	1.4	0.0	31.4	14.3	0.0	0.0	2.9	0.0	27.1	0.0	0.0
	(72)	(1.4)	(2.8)	(4.2)	(1.4)	(11.1)	(1.4)	(0.0)	(2.8)	(1.4)	(0.0)	(30.6)	(13.9)	(0.0)	(0.0)	(2.8)	(0.0)	(26.4)	(0.0)	(0.0)
Philippines	28	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	35.7	0.0	0.0	0.0	35.7	0.0	28.6	0.0	0.0
	(30)	(3.3)	(3.3)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(33.3)	(0.0)	(0.0)	(0.0)	(33.3)	(0.0)	(26.7)	(0.0)	(0.0)
Taiwan	39	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.5	0.0	0.0	0.0	7.7	0.0	71.8	0.0	0.0
Lineage total	369	0.0	0.8	1.1	1.1	2.4	1.9	13.6	1.9	0.3	10.0	22.8	2.7	2.4	1.9	13.0	3.3	16.5	2.7	1.6
	(390)	(3.3)	(2.6)	(1.3)	(1.0)	(2.3)	(1.8)	(12.8)	(1.8)	(0.3)	(9.5)	(21.5)	(2.6)	(2.3)	(1.8)	(12.3)	(3.1)	(15.6)	(2.6)	(1.5)
Haplogroup total	369	0.0	0.8	1.1	1.1			19.8		0.3	10.0					66.9				
	(390)	(3.3)	(2.6)	(1.3)	(1.0)			(18.7)		(0.3)	(9.5)					(63.3)				

The four Taiwanese aboriginal populations have been pooled, as have the two population samples from Vanuatu and the three samples from Tonga. The numbers in parentheses denote the lineage frequencies before the admixed chromosomes were removed from the data set.

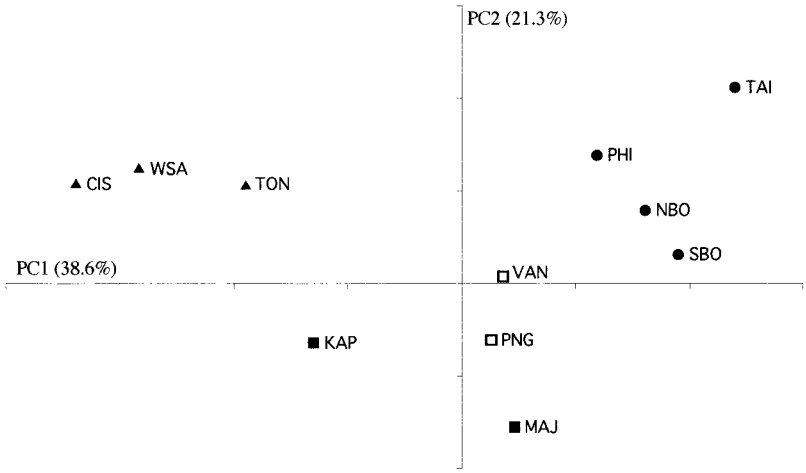

FIGURE 4.—A plot of the first two PCs within this data set. Polynesian populations (defined geographically) are indicated with solid triangles, Melanesian with open squares, Micronesian with solid squares, and Island Southeast Asian with solid circles. The abbreviations are explained in the inset legend to Figure 3. Axes are labeled with the percentage of the total variance summarized by that PC.

of the northwestern European subtypes. A total of 21 chromosomes (5.4%) were thus removed. The resulting data set of 369 chromosomes is detailed in Table 1.

Population clustering: Principal components (PC) analysis was used to explore the relationships between populations in a nonbifurcating manner. The first two PCs, calculated from lineage frequencies of nonadmixed chromosomes given in Table 1, account for 60% of the variance within the data and were plotted against one another in Figure 4. The first PC separates populations on the basis of Polynesian ancestry. The second PC separates the Polynesian outlier from the true Polynesian populations and the Micronesian population from the Melanesian ones. It can be seen from the PC analysis (PCA) plot that the true Polynesian populations form a cluster although notably Tonga is the closest to the Melanesian populations. Tonga shares hg 24 and lineage 26.8 with Melanesian populations. Kapingamarangi, the Polynesian outlier, lies between the Polynesian populations and the Micronesian one in the PCA, reflecting its mixed ancestry. This population contains the 10.2 lineage found in Polynesia but not Micronesia; however, it also contains the 26.3 and 26.5 lineages found in Micronesia but not Polynesia.

Population diversity: A number of different diversity indices were calculated for each of the 11 populations, and their performance is compared in Figure 5. Nei's estimator of diversity applied to lineage frequencies reveals considerable variance among the populations, with high diversities apparent in Borneo, Vanuatu, and Kapingamarangi, and less diversity in Polynesia and Taiwan. However, lineage-based diversity measures are prone to ascertainment bias due to a greater impact of founder effects in Oceania than in SE Asia, resulting in more clearly defined groups of related haplotypes. What is needed is an estimator that uses the unbiased diversity apparent in the multiallelic markers, which are polymorphic in all populations. However, the uninformative nature of Nei's estimator based on compound multiallelic haplotypes (comprising both MSY1 codes and microsatellite haplotypes; see Figure 5) additionally reveals a requirement for an estimator to take into account genetic distance between haplotypes rather than mere identity. The sometimes saltatory nature of MSY1 evolution may well bias such estimators and was excluded from further analyses. The mean pairwise difference (MPD) within populations based on the seven-locus microsatellite haplotypes reveals variance in population diversities similar to that of Nei's estimator based on the lineage frequencies, but will overemphasize diversity in populations that have gone through a bottleneck if more than one lineage survives. To overcome these limitations of existing estimators we calculated a new diversity measure. This measures the MPD *within each lineage* for a given population and averages them, weighted for the frequency of each lineage. Obviously such a measure will exclude lineages for which there is but a single representative in a given population. Consequently, the values displayed in Figure 5 are calculated from the haplogroups defined by the binary markers alone rather than the full set of lineages. As a result 98% (362/369) of the nonadmixed Y chromosomes in this data set contribute to these estimates. This diversity estimator, the weighted mean intralineage mean pairwise difference (WIMP), better captures the true reduction of diversity apparent in Polynesia. However, the properties of this novel diversity measure merit further investigation.

Bayesian coalescent analysis: Lineages comprising >30 chromosomes were dated using two different methods that relate the amount of intralineage diversity of seven-locus microsatellite haplotypes to the age of the lineage. The first calculates the average squared distance (ASD) between a root haplotype and all other chromosomes within the lineage and relates it to the age of the lineage (THOMAS *et al.* 1998). The root haplotype is obtained by combining the modal alleles at each locus together. The second method is a Bayesian-based coalescent analysis, called BATWING, that simulates the coalescence of haplotypes using a population model

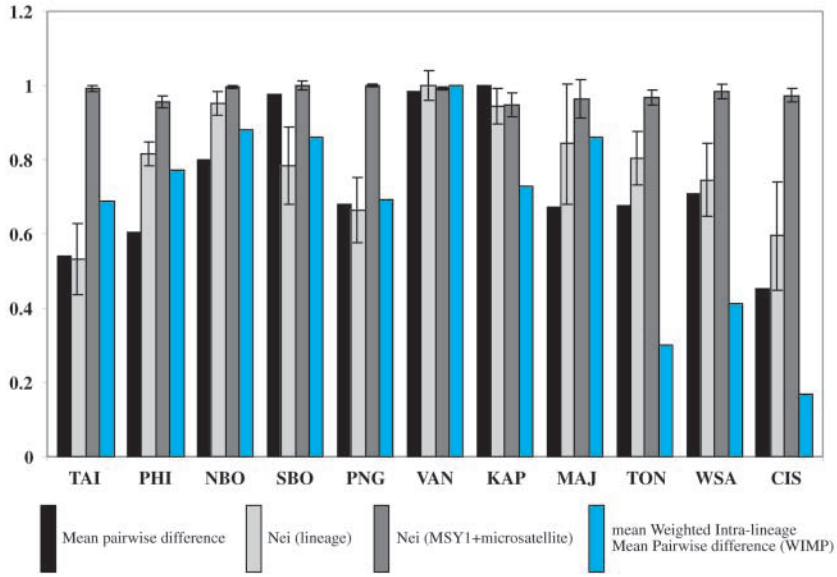

FIGURE 5.—Normalized diversity indices for each population in this study.

that incorporates both subdivision and a period of constant population size followed by a period of exponential growth (WILSON *et al.* 2000). The age of the most recent common ancestor (MRCA) of the lineage is only one of a number of model parameters that this analysis provides. Table 2 gives the ages obtained for five monophyletic lineages using both methods of analysis: for three of the lineages there is good agreement between the two methods. However, for the 10.2 lineage defined by the insertion of a block of null repeats near the 3′ end of the MSY1 array and the lineage (10.1 + 10.2) defined by the insertion of a null repeat at the 5′ end of the MSY1 repeat array, there is a large discrepancy between the estimates of the two analyses. The ASD method gives substantially younger ages for these two nested lineages. Figure 6 shows that the mismatch distributions for the three lineages whose ages agree well between the two analyses are smooth. However, the two discordant lineages show a bimodal distribution that might indicate that most of the chromosomes sampled from this lineage derive from a recent expansion of closely related haplotypes within a more diverse and ancient lineage. Similarly, the posterior distribution for the age of the population expansion for these lineages (from the BATWING analysis; see Table 2) also shows evidence of a much later population expansion relative to the age of the lineage. Studying the MJ network of compound multiallelic haplotypes from lineage 10.2 in Figure 7 indicates the likely source of this discordance. Polynesian chromosomes sampled from lineage 10.2 are closely related and appear to have expanded recently from a few related haplotypes, whereas the Melanesian examples of this lineage are much more diverse, indicating the true age of this lineage.

AMOVA classifications: To test which of the three approaches to distinguishing Pacific populations discussed in the Introduction best corresponds with the observed pattern of extant genetic diversity, an AMOVA

TABLE 2

Dating estimates for the five lineages with >35 representatives

		ASD		BATWING		
Lineage	*N*	Age estimate	95% limits	Age estimate	95% limits	Expansion
10.1 + 10.2	57	2820	1970–4220	6750	3170–16,640	6.3 (0.3–37.2)
10.2	50	2310	1620–3470	5780	2500–13,000	6.1 (1.8–16)
24	37	5840	4090–8760	5900	3270–12,150	34.7 (17.8–87.5)
26.4	48	1210	840–1810	1760	780–4660	54.5 (19.9–96.8)
26.6	61	3710	2600–5560	4430	2200–10,650	42.7 (14.1–80.4)

Two dating methods, ASD and BATWING, are compared. All estimates are rounded to the nearest 10 years. ASD confidence limits are based on uncertainty in the mutation rate. BATWING age estimates are based on the median of the posterior distribution for the TMRCA of each lineage and the limits of the 95% even-tailed interval. The final column gives the median of the posterior distribution for the time since the population started growing exponentially, together with the limits of the even-tailed 95% interval in parentheses, both expressed as percentages of the age of the lineage.

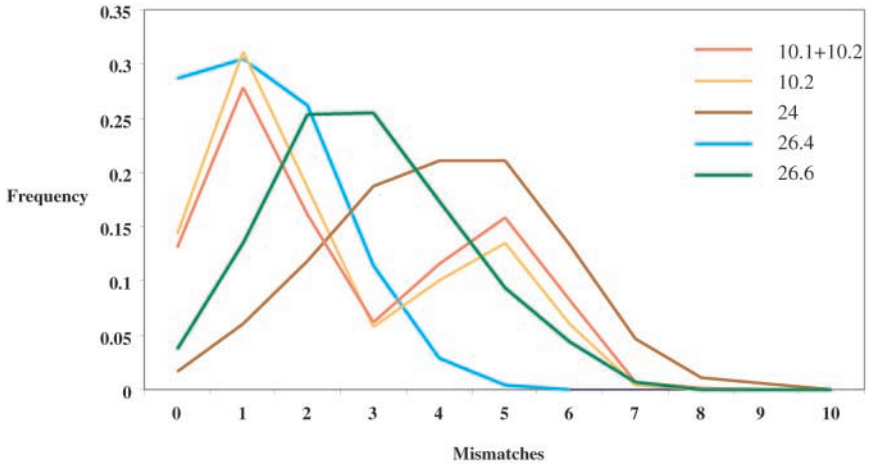

FIGURE 6.—Mismatch distributions for each dated lineage. Mismatch distributions based on relative rather than absolute frequencies are displayed for five lineages, color coded in accordance with Figures 2 and 3.

was performed on the lineage frequencies in the seven Pacific populations using three groupings based on similarities of geography, ethnology, and settlement history. This method apportions the total variance within the data between the three hierarchical levels apparent within any such classification, that is, within populations, between populations, within groups, and between groups. The best classification of these populations is expected to maximize the amount of variance that is apportioned between groups. The results (Table 3) demonstrate that the best grouping is obtained when populations are grouped geographically, rather than ethnologically or by settlement history.

Mantel testing: It has been suggested that when genetic distances correlate better with geographical than linguistic distances in Oceania a high level of post-settlement gene flow is implied (LUM *et al.* 1998). If the opposite is the case, then initial settlement patterns are thought to dominate the distribution of extant diversity. The relative correlation of geography, linguistics, and

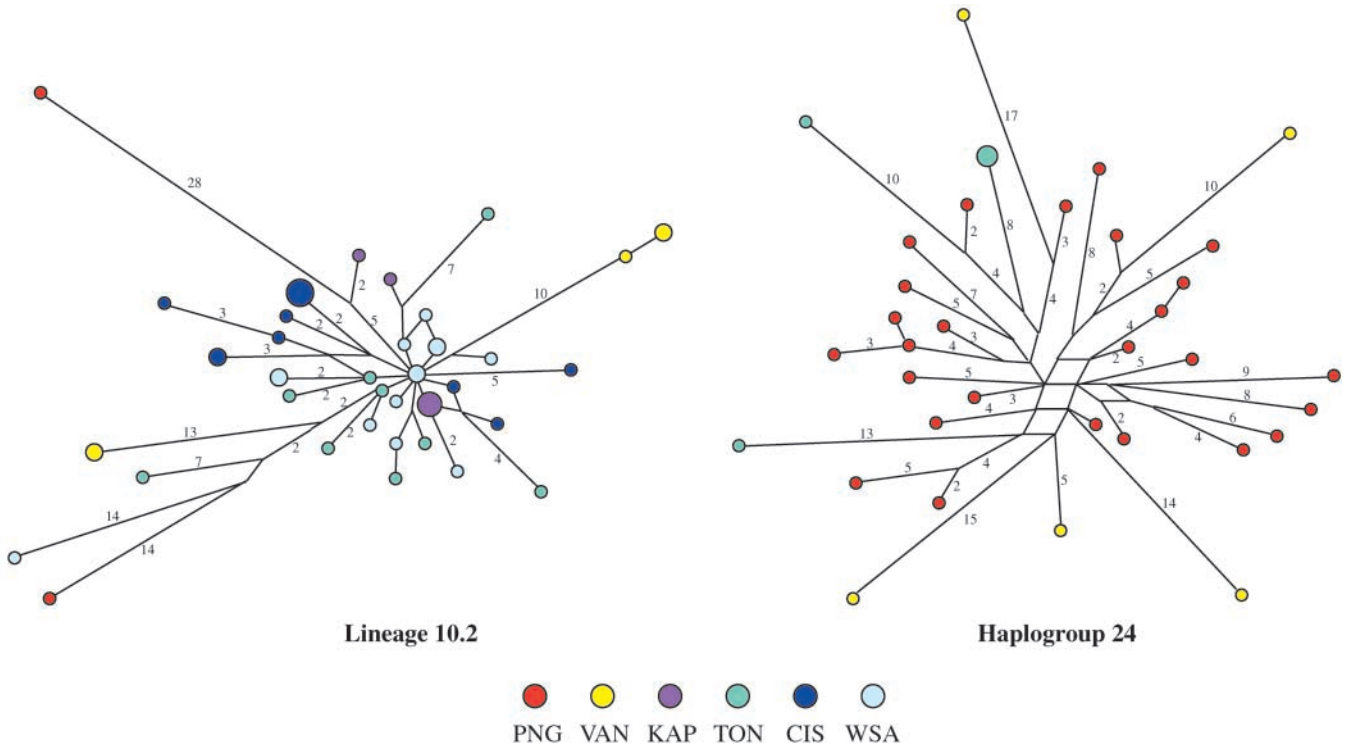

FIGURE 7.—Median-joining networks of lineage 10.2 and haplogroup 24. Networks are based on compound multiallelic haplotypes comprising both seven-locus microsatellite haplotypes and MSY1 codes and are weighted for the mutation rates of each locus. Mutational steps greater than a single repeat are labeled. Circles represent haplotypes, whose areas are proportional to the number of chromosomes with that haplotype, and color indicates the population in which each haplotype is found. The abbreviations are explained in the inset legend to Figure 3.

TABLE 3

AMOVA analysis of three classifications of Pacific populations

Grouping	Basis	% of total variance		
		Within populations	Between populations within groups	Between groups
(TON, WSA, CIS) (VAN, PNG) (MAJ, KAP)	Geography	74.17	12.58	13.25
(TON, WSA, CIS, KAP) (VAN, PNG) (MAJ)	Ethnology	73.86	14.79	11.35
(VAN, TON, WSA, CIS, MAJ, KAP) (PNG)	Settlement history	71.7	17.24	11.05

The amount of variance apportioned to each of the three levels of the classification is given for three different classifications of the seven Oceanic-speaking populations, based on geography, ethnology, and settlement history. The abbreviations are explained in the inset legend to Figure 3.

genetics can be processed by Mantel tests (Mantel 1967) of distance matrices between the populations in question. Previous work contrasting Mantel tests using genetic distances from biparentally inherited autosomal markers and maternally inherited mtDNA implied higher male than female gene flow in Oceania (Lum *et al.* 1998). Here, this methodology is followed to attempt to address this issue using the paternally inherited Y chromosome. The genetic distances used are F_{ST} values calculated from the lineage frequencies and geographical distances are great circle distances between the sample sites. Linguistic distances are from the tree shown in Figure 8, which is taken from the previous study (Lum *et al.* 1998), to maintain comparability between studies, although adapted slightly to include additional languages. Two sets of populations were studied by this method, the first being all Austronesian-speaking populations and the second being Oceanic-speaking populations (see Table 4). In every case, geographic and genetic distances are significantly correlated even when language is taken into account. However, while linguistic distances are not significantly correlated with genetic distances, when geographical distances are taken into account among Austronesian populations, they are sig-

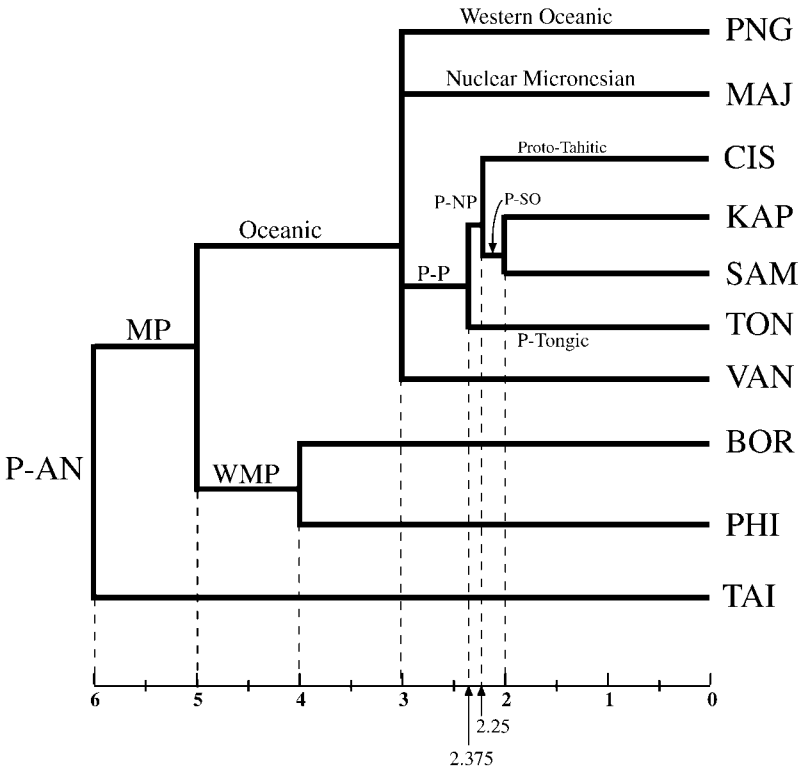

Figure 8.—Language tree relating the 10 language groups used for Mantel testing. The language tree is minimally adapted from that in Lum *et al.* (1998) to fit the languages spoken by the populations in this study. Language families and subfamilies are indicated on the branches of the tree. The abbreviations are explained in the inset legend to Figure 3, with the addition of BOR, Borneo; P-AN, proto-Austronesian; MP, Malayo-Polynesian; WMP, Western Malayo-Polynesian; P-P, proto-Polynesian; P-NP, proto-nuclear Polynesian; and P-SO, proto-Samoic outlier.

TABLE 4

Mantel tests of the correspondence between geography, genetics, and language

	Austronesian populations		Oceanic populations	
Correlation	*r*	*P* value	*r*	*P* value
F_{ST} *vs.* geography	0.46	0.003*	0.58	0.004*
F_{ST} *vs.* language	0.37	0.053	0.51	0.013*
F_{ST} *vs.* geography (language)	0.32	0.022*	0.68	0.001*
F_{ST} *vs.* language (geography)	0.15	0.273	0.64	0.000*

Two sets of populations were studied: 10 populations that speak Austronesian languages (the two sampled from Borneo were pooled) and 7 populations that speak Oceanic languages. The correlation is measured separately between genetic distances and linguistic distances and between genetic distances and geographical distances. Then the partial correlation is measured for each of the above relationships while controlling for the effect of the third matrix. Significance is measured by 1000 permutations. *P values <0.05.

nificantly correlated among Oceanic populations. Thus this test does not in general provide support for a higher rate of male compared to female gene flow among Oceanic populations.

DISCUSSION

The dominant archeo-linguistic model for the origins of Polynesian populations is that they represent the eastern fringe of an agriculturally driven expansion that originated in SE China and Taiwan some 6000 years ago (Bellwood 1997). If genetic data were to support the biological validity of this model, we would expect to find lineages in Polynesia that can be traced to this region of the world within this time scale. A number of analytical approaches relate observed intralineage diversity to time to MRCA (TMRCA). Some calculate model-free summary statistics that require the stipulation of a root, such as ASD dating (Thomas *et al.* 1998) and rho dating (Bertranpetit and Calafell 1996; Forster *et al.* 1996). The root haplotype can be estimated phylogenetically or statistically by combining modal alleles. It has been noted that these summary statistic methods often give more recent ages than expected from independent estimates (Bosch *et al.* 1999), and this has led to the questioning of the pedigree mutation rates for the multiallelic loci used to assay intralineage diversity (Bosch *et al.* 1999; Forster *et al.* 2000). A growing number of model-based coalescent simulation methods can be used to estimate a variety of parameters within the model, among others the BATWING (Wilson *et al.* 2000) method used here. The example of the 10.2 lineage in this study reveals one reason why summary statistic methods may underestimate TMRCA. The recent expansion of a subset of haplotypes within a more ancient lineage will lead the ASD and rho methods to specify a root haplotype that is in fact the ancestral haplotype of the expansion and not the lineage, as in the case of the apparent root haplotype in the MJ network of lineage 10.2 in Figure 7. Consequently, although root-based summary statistics are capable of producing unbiased estimators of lineage age, they are compromised by the difficulties in defining the root haplotype accurately. The new generation of coalescent-based methods that incorporate increasingly realistic population and growth models appears to be superior for estimating the ages of paternal lineages.

We found two dominant lineages in Polynesia, lineage 10.2 and lineage 26.4, together accounting for 81% of nonadmixed Polynesian Y chromosomes. Taking the coalescent estimates for the TMRCA of lineage 10.2 we obtain an age of ~6000 years old that should lead us to expect to find these chromosomes in Taiwan, should they have originated there. However, these chromosomes are found only in Melanesia and Polynesia. Diversity at multiallelic loci is restricted in Taiwan, suggestive of a recent population bottleneck or low long-term effective population size, both of which scenarios could have led to the local extinction of lineage 10.2. However, the absence of 10.2 chromosomes and their more ancient ancestors (lineage 10.1 and hg 10) from the Philippines as well suggests that this is not the case. It appears that lineage 10.2 owes its ancestry, much like that of its phylogenetic predecessor, the *DYS390.3* chromosomes (Kayser *et al.* 2000a), to a source population in Melanesia and/or eastern Indonesia.

By contrast, lineage 26.4 is shared between Island SE Asia, including Taiwan and Polynesia. These chromosomes demonstrate a striking lack of diversity given their wide distribution, and coalescent age estimates suggest a very recent origin for this lineage, within the past 4500 years. The site of maximal intralineage diversity is often taken to be the likely place of origin of a lineage (Richards *et al.* 1998; Kayser *et al.* 2000a), although it should be noted that when equating diversity to age, long-term effective population sizes are assumed not to be significantly different. The lineage 26.4 chromosomes in Island SE Asia are most diverse, as measured by their mean pairwise difference in compound multiallelic haplotypes (8.2 compared to 5.8 in Melanesia and 6.1 in Polynesia). There are too few chromosomes to attempt to define the likely origin at a finer geographical resolu-

tion although their higher frequency in Taiwan and the Philippines may indicate an origin in northern Island SE Asia. Y chromosomes exhibiting the derived form of the M122 binary marker, of which this lineage is a subgroup, have a been shown to have similar geographical distribution and a similar site of origin has been proposed (Kayser *et al.* 2001).

The origins of Micronesian populations are less well characterized archeologically and linguistically than those of Polynesians. Although only a single small population of Micronesians was analyzed here, the absence of both the 26.4 and 10.2 lineages is striking. The majority of Micronesian Y chromosomes (55%) belong to a single lineage, 26.5, that is found only in one other population in this study, Kapingamarangi. There are no clear ancestors to this set of chromosomes, although the most closely related chromosomes in the NJ tree are found in Borneo. Lineage 26.3 (9%) is also shared with Kapingamarangi but with no Polynesian populations, suggesting that it is restricted to Micronesia. A single chromosome belonging to this lineage is found in Papua New Guinea, suggesting an ultimately Melanesian origin for these chromosomes. Thus, Micronesian Y chromosomes appear to have a distinct ancestry to those in Polynesia. They seem to derive from Melanesia and SE Asia but from populations that are genetically distinct from those that subsequently colonized Polynesia. This pattern of a clear distinction between Polynesian and Micronesian Y chromosomes is mirrored in a recent study comparing mtDNA diversity in the same region (Lum and Cann 2000, p. 165), which concluded that Polynesian and Micronesian populations "were settled from a common source, via a similar route, but by distinct populations" and that subsequently they had "largely distinct prehistories."

The genetic ancestry of the Polynesian outliers is poorly resolved. It would appear from the present study that the island of Kapingamarangi has dual Polynesian and Micronesian ancestry. This explains its surprisingly high diversity, compared to other islands defined ethnologically as being Polynesian, and is in accordance with archeological evidence for population assimilation that suggests that Polynesian ancestry will be reflected less clearly in genetics than in language (Bellwood 1989).

What can we say of the patterns of genetic diversity within Polynesia? In accordance with previous studies (Flint *et al.* 1989; Sykes *et al.* 1995) there is a reduction of diversity across the Pacific from west to east. Most lineages found in the three Polynesian populations are shared by at least two of them, as would be expected from their common origin; however, two lineages in Tonga, hg 24 and lineage 26.8, are specific to that island group within Polynesia. Elsewhere both of these lineages are found only in Melanesia, suggesting gene flow from this region into Tonga but not to other Polynesian islands. We envisage two scenarios to explain the presence of these lineages. The first is that these chromosomes came into Tonga together during the initial settlement, and the second is that they had arrived more recently. The pattern of diversity within Tongan hg 24 chromosomes shown in Figure 7 does not suggest that these chromosomes expanded from a pool of closely related founder haplotypes as have the other two major Polynesian lineages, 10.2 and 26.4. It seems more likely that these chromosomes had arrived since the first settlement of Tonga, perhaps as a result of trading contacts between Melanesia and Polynesia and reflecting the geographical proximity of Tonga to Fiji.

This raises the wider issue of the degree of male gene flow throughout Oceania. Mantel testing provides no support for the contention of a prior study that male gene flow might be higher than female gene flow throughout Oceania. The previous findings may have more to do with the different effective population sizes and mutation dynamics of the mitochondrial and autosomal loci studied than they do with their different patterns of inheritance. While we do not discount the possibility of higher male than female gene flow in Oceania, the degree of differentiation between Melanesian, Micronesian, and Polynesian Y chromosomes does not fit with the description that higher male gene flow throughout Oceanic populations results in an "entangled bank" of diversity (Lum *et al.* 1998).

In conclusion, this study, while not strongly supporting the hypothesis of a rapid Austronesian expansion from Taiwan, is not necessarily incompatible with it. Biological and cultural origins can become uncoupled to varying degrees. Whereas the dominant model for the cultural evolution of Pacific peoples does not adequately explain the origins of the majority of Polynesian Y chromosomes, these populations may still retain a genetic signal of their cultural origins in a minority of their paternal lineages.

The authors thank John Clegg for kindly providing samples. The authors are also grateful to Manfred Kayser and Christian Capelli for providing access to their data, Victor Paz, Stephen Oppenheimer, and Peter Forster for helpful discussions, Chris Tyler-Smith for unpublished information, and Ian Wilson for advice with statistical analysis. M.E.H. was supported by the Medical Research Council and the McDonald Institute. M.A.J. is a Wellcome Trust Senior Fellow in Basic Biomedical Science (grant no. 057559). The research also received further support from the Medical Research Council and the Wellcome Trust.

Note added in proof: Studies of mitochondrial diversity on Kapingamarangi show a similar picture, with two common, closely related, mtDNA haplotypes. One of these haplotypes is dominant in Polynesia; the other is common in Micronesia (Sykes *et al.* 1995; Lum and Cann 2000).

LITERATURE CITED

Bellwood, P. S., 1989 The colonisation of the Pacific: some current hypotheses, pp. 1–60 in *The Colonisation of the Pacific: A Genetic Trail*, edited by A. V. S. Hill and S. W. Serjeantson. Clarendon Press, Oxford.

Bellwood, P., 1991 The Austronesian dispersal and the origin of languages. Sci. Am. **July:** 70–75.

Bellwood, P., 1997 *The Prehistory of the Indo-Malaysian Archipelago.* University of Hawaii Press, Honolulu.

Bergen, A. W., C. Y. Wang, J. Tsai, K. Jefferson, C. Dey *et al.*, 1999 An Asian-Native American paternal lineage identified by RPS4Y resequencing and by microsatellite haplotyping. Ann. Hum. Genet. **63:** 63–80.

Bertranpetit, J., and Calafell, F., 1996 Genetic and geographic variability in cystic fibrosis: evolutionary considerations, pp. 97–118 in *Variation in the Human Genome*, edited by D. Chadwick and G. Cardew. John Wiley & Sons, Chichester, UK.

Blanco, P., M. Shlumukova, C. A. Sargent, M. A. Jobling, N. Affara *et al.*, 2000 Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. J. Med. Genet. **37:** 752–758.

Blust, R., 1999 Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics. Symp. Ser. Inst. Linguistics Acad. Sinica **1:** 31–94.

Bosch, E., F. Calafell, F. R. Santos, A. Perez-Lezaun, D. Comas *et al.*, 1999 Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. Am. J. Hum. Genet. **65:** 1623–1638.

Capelli, C., J. F. Wilson, M. Richards, M. P. H. Stumpf, F. Gratrix *et al.*, 2001 A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. Am. J. Hum. Genet. **68:** 432–443.

Casanova, M., P. Leroy, C. Boucekkine, J. Weissenbach, C. Bishop *et al.*, 1985 A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. Science **230:** 1403–1406.

Davidson, J. M., 1988 Archaeology in Micronesia since 1965: past achievements and future prospects. New Zealand J. Archaeol. **10:** 83–100.

Felsenstein, J., 1995 PHYLIP: Phylogeny Inference Package, Department of Genetics, University of Washington, Seattle.

Flint, J., A. J. Boyce, J. J. Martinson and J. B. Clegg, 1989 Population bottlenecks in Polynesia revealed by minisatellites. Hum. Genet. **83:** 257–263.

Forster, P., R. Harding, A. Torroni and H.-J. Bandelt, 1996 Origin and evolution of Native American mtDNA variation: a reappraisal. Am. J. Hum. Genet. **59:** 935–945.

Forster, P., M. Kayser, E. Meyer, L. Roewer, H. Pfeiffer *et al.*, 1998 Phylogenetic resolution of complex mutational features at Y-STR DYS390 in Aboriginal Australians and Papuans. Mol. Biol. Evol. **15:** 1108–1114.

Forster, P., A. Rohl, P. Lunnemann, C. Brinkmann, T. Zerjal *et al.*, 2000 A short tandem repeat-based phylogeny for the human Y chromosome. Am. J. Hum. Genet. **67:** 182–196.

Gray, R. D., and F. M. Jordan, 2000 Language trees support the express-train sequence of Austronesian expansion. Nature **405:** 1052–1055.

Green, R. C., 1999 Integrating historical linguistics with archaeology: insights from research in remote Oceania. Indo-Pacific Prehist. Assoc. Bull. **18:** 3–16.

Hammer, M. F., 1994 A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. Mol. Biol. Evol. **11:** 749–761.

Hammer, M. F., and S. Horai, 1995 Y-chromosomal DNA variation and the peopling of Japan. Am. J. Hum. Genet. **56:** 951–962.

Hammer, M. F., T. Karafet, A. Rasanayagam, E. T. Wood, T. K. Altheide *et al.*, 1998 Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol. Biol. Evol. **15:** 427–441.

Heyerdahl, T., 1950 *Kontiki: Across the Pacific by Raft.* Rand McNally, Chicago.

Hurles, M. E., C. Irven, J. Nicholson, P. G. Taylor, F. R. Santos *et al.*, 1998 European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. Am. J. Hum. Genet. **63:** 1793–1806.

Hurles, M. E., R. Veitia, E. Arroyo, M. Armenteros, J. Bertranpetit *et al.*, 1999 Recent male-mediated gene flow over a linguistic barrier in Iberia suggested by analysis of a Y-chromosomal DNA polymorphism. Am. J. Hum. Genet. **65:** 1437–1448.

Hurles, M. E., C. Irven, J. Nicholson, P. G. Taylor, F. R. Santos *et al.*, 2001 European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA (Erratum). Am. J. Hum. Genet. **68:** 298.

Irwin, G., 1992 *The Prehistoric Exploration and Colonisation of the Pacific.* Cambridge University Press, Cambridge, UK.

Jobling, M. A., and C. Tyler-Smith, 1995 Fathers and sons: the Y chromosome and human evolution. Trends Genet. **11:** 449–456.

Jobling, M. A., and C. Tyler-Smith, 2000 New uses for new haplotypes: the human Y chromosome, disease and selection. Trends Genet. **16:** 356–362.

Jobling, M. A., N. Bouzekri and P. G. Taylor, 1998 Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (*DYF155S1*). Hum. Mol. Genet. **7:** 643–653.

Kalaydjieva, L., F. Calafell, M. A. Jobling, D. Angelicheva, P. de Knijff *et al.*, 2001 Patterns of inter- and intra-group genetic diversity in the Vlax Roma as revealed by Y chromosome and mitochondrial DNA lineages. Eur. J. Hum. Genet. **9:** 97–104.

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

Kayser, M., L. Roewer, M. Hedman, L. Henke, J. Henke *et al.*, 2000b 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.

Kayser, M., S. Brauer, G. Weiss, W. Schiefenhovel, P. Underhill *et al.*, 2001 Independent histories of human Y chromosomes from Melanesia and Australia. Am. J. Hum. Genet. **68:** 173–190.

Kirch, P. V., and R. C. Green, 1992 History, phylogeny and evolution in Polynesia. Curr. Anthropol. **33:** 161–186.

Kwok, C., C. Tyler-Smith, B. B. Medonca, I. Hughes, G. D. Berkovitz *et al.*, 1996 Mutation analysis of 2kb 5′ to SRY in XY females and XX intersex subjects. J. Med. Genet. **33:** 465–468.

Lum, J. K., and R. L. Cann, 2000 mtDNA lineage analyses: origins and migrations of Micronesians and Polynesians. Am. J. Phys. Anthropol. **113:** 151–168.

Lum, J. K., R. L. Cann, J. J. Martinson and L. B. Jorde, 1998 Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. Am. J. Hum. Genet. **63:** 613–624.

Mantel, N., 1967 The detection of disease clustering and a generalised regression approach. Cancer Res. **27:** 209–220.

Meacham, W., 1985 On the improbability of Austronesian origins in South China. Asian Perspectives **25:** 100.

Melton, T., R. Peterson, A. J. Redd, N. Saha, A. S. M. Sofro *et al.*, 1995 Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. Am. J. Hum. Genet. **57:** 403–414.

Melton, T., S. Clifford, J. Martinson, M. Batzer and M. Stoneking, 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.

Oppenheimer, S., 1998 *Eden in the East: The Drowned Continent of Southeast Asia.* Phoenix, London.

Pawley, A., and M. Ross, 1993 Austronesian historical linguistics and culture history. Annu. Rev. Anthropol. **22:** 425–459.

Redd, A. J., N. Takezaki, S. T. Sherry, S. T. McGarvey, A. S. M. Sofro *et al.*, 1995 Evolutionary history of the COII/tRNAlys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. Mol. Biol. Evol. **12:** 604–615.

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

Rosser, Z., T. Zerjal, M. E. Hurles, M. Adojaan, D. Alavantic *et al.*, 2000 Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am. J. Hum. Genet. **67:** 1526–1543.

Santos, F. R., A. Pandya, C. Tyler-Smith, S. D. J. Pena, M. Schanfield *et al.*, 1999 The central Siberian origin for Native American Y chromosomes. Am. J. Hum. Genet. **64:** 619–628.

Schneider, S., D. Roessli and L. Excoffier, 2000 Arlequin: a software for population genetics data analysis, version 2.000. Genetics and Biometry Laboratory, University of Geneva.

Seielstad, M. T., E. Minch and L. L. Cavalli-Sforza, 1998 Genetic evidence for a higher female migration rate in humans. Nat. Genet. **20:** 278–280.

Solheim, W. G., 1996 The Nusantao and North-South dispersals. Indo-Pacific Prehistory Assoc. Bull. **15:** 106–107.

Spriggs, M., 1989 The dating of the Island Southeast Asian Neolithic: an attempt at chronometric hygiene and linguistic correlation. Antiquity **63:** 587–613.

Spriggs, M., 1999 Archaeological dates and linguistic sub-groups in the settlement of the Island Southeast Asian-Pacific region. Indo-Pacific Prehistory Assoc. Bull. **18:** 17–24.

Spriggs, M., and A. Anderson, 1993 Late colonisation of East Polynesia. Antiquity **67:** 200–217.

Su, B., L. Jin, P. Underhill, J. Martinson, N. Saha *et al.*, 2000 Polynesian origins: insights from the Y chromosome. Proc. Natl. Acad. Sci. USA **97:** 8225–8228.

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

Thomas, M. G., K. Skorecki, H. Ben-Ami, T. Parfitt, N. Bradman *et al.*, 1998 Origins of Old Testament priests. Nature **384:** 138–140.

Underhill, P. A., L. Jin, A. A. Lin, S. Q. Mehdi, T. Jenkins *et al.*, 1997 Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res. **7:** 996–1005.

Veitia, R., A. Ion, S. Barbaux, M. A. Jobling, N. Souleyreau *et al.*, 1997 Mutations and sequence variants in the testis-determining region of the Y chromosome in individuals with a 46,XY female phenotype. Hum. Genet. **99:** 648–652.

Whitfield, L. S., J. E. Sulston and P. N. Goodfellow, 1995 Sequence variation of the human Y chromosome. Nature **378:** 379–380.

Wilson, I. J., and D. J. Balding, 1998 Genealogical inference from microsatellite data. Genetics **150:** 499–510.

Wilson, I., M. Weale and D. Balding, 2000 BATWING: Bayesian Analysis of Trees With Internal Node Generation. Department of Mathematical Sciences, University of Aberdeen, U.K. (http://www.maths.abdn.ac.uk/~ijw).

Zerjal, T., B. Dashnyam, A. Pandya, M. Kayser, L. Roewer *et al.*, 1997 Genetic relationships of Asians and northern Europeans, revealed by Y-chromosomal DNA analysis. Am. J. Hum. Genet. **60:** 1174–1183.

Zerjal, T., A. Pandya, F. R. Santos, R. Adhikari, E. Tarazona *et al.*, 1999 The use of Y-chromosomal DNA variation to investigate population history: recent male spread in Asia and Europe, pp. 91–102 in *Genomic Diversity: Applications in Human Population Genetics*, edited by S. S. Papiha, R. Deka and R. Chakraborty. Plenum, New York.

Communicating editor: M. K. Uyenoyama