Polarity and Temporality of High-Resolution Y-Chromosome Distributions in India Identify Both Indigenous and Exogenous Expansions and Reveal Minor Genetic Influence of Central Asian Pastoralists

Sanghamitra Sengupta,[1] Lev A. Zhivotovsky,[2] Roy King,[3] S. Q. Mehdi,[4] Christopher A. Edmonds,[3] Cheryl-Emiliane T. Chow,[3] Alice A. Lin,[3] Mitashree Mitra,[5] Samir K. Sil,[6] A. Ramesh,[7] M. V. Usha Rani,[8] Chitra M. Thakur,[9] L. Luca Cavalli-Sforza,[3] Partha P. Majumder,[1] and Peter A. Underhill[3]

[1]Human Genetics Unit, Indian Statistical Institute, Kolkata, India; [2]N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow; [3]Department of Genetics, Stanford University, Stanford; [4]Biomedical and Genetic Engineering Division, Dr. A. Q. Khan Research Laboratories, Islamabad; [5]School of Studies in Anthropology, Pandit Ravishankar Shukla University, Raipur, India; [6]University of Tripura, Tripura, India; [7]Department of Genetics, University of Madras, Chennai, India; [8]Department of Environmental Sciences, Bharathiar University, Coimbatore, India; and [9]B. J. Wadia Hospital for Children, Mumbai, India

Although considerable cultural impact on social hierarchy and language in South Asia is attributable to the arrival of nomadic Central Asian pastoralists, genetic data (mitochondrial and Y chromosomal) have yielded dramatically conflicting inferences on the genetic origins of tribes and castes of South Asia. We sought to resolve this conflict, using high-resolution data on 69 informative Y-chromosome binary markers and 10 microsatellite markers from a large set of geographically, socially, and linguistically representative ethnic groups of South Asia. We found that the influence of Central Asia on the pre-existing gene pool was minor. The ages of accumulated microsatellite variation in the majority of Indian haplogroups exceed 10,000–15,000 years, which attests to the antiquity of regional differentiation. Therefore, our data do not support models that invoke a pronounced recent genetic input from Central Asia to explain the observed genetic variation in South Asia. R1a1 and R2 haplogroups indicate demographic complexity that is inconsistent with a recent single history. Associated microsatellite analyses of the high-frequency R1a1 haplogroup chromosomes indicate independent recent histories of the Indus Valley and the peninsular Indian region. Our data are also more consistent with a peninsular origin of Dravidian speakers than a source with proximity to the Indus and with significant genetic input resulting from demic diffusion associated with agriculture. Our results underscore the importance of marker ascertainment for distinguishing phylogenetic terminal branches from basal nodes when attributing ancestral composition and temporality to either indigenous or exogenous sources. Our reappraisal indicates that pre-Holocene and Holocene-era—not Indo-European—expansions have shaped the distinctive South Asian Y-chromosome landscape.

Two rival models based primarily on mtDNA and Y-chromosome data have been proposed to explain the genetic origins of tribes and castes in South Asia. One model suggests that the tribes and castes share considerable Pleistocene heritage, with limited recent gene flow between them (Kivisild et al. 2003*a*), whereas an exact opposite view concludes that caste and tribes have independent origins (Cordaux et al. 2004). Another analogous debate concerns the origins of the hypothetical proto–Elamo-Dravidian language, which is thought to be the precursor of Tamil. It has been proposed that the proto–Elamo-Dravidian language spread eastward from southwest Persia into South Asia with agriculture (McAlpin 1974, 1981), and the argument is bolstered by the existence of a solitary Dravidian-speaking group, the Brahui, in Pakistan (Renfrew 1996). The linguistic evidence, however, is compromised by uncertainty regarding whether word borrowing is responsible for the observed linguistic affinities (Blazek and Boisson 1992). A study of mtDNA-haplogroup frequencies in southwestern and central Asia reported that the Brahui gene pool was more similar to that of Indo-Iranian speakers from southwest Asia than to that of Dravidian populations of India (Quintana-Murci et al. 2004). The signature of low-frequency western Eurasian mtDNA lineages in India was interpreted (Kivisild et al. 1999) to support the hypothesis that Dravidian farmers arrived in India from the Middle East (Renfrew 1996). Since most Dravidians are clustered, like the Indo-Iranian speakers, within South Asian–derived mtDNA haplogroups (HGs) (Bamshad et al. 2001), this scenario would imply, in terms of maternal lineages, a farming-associated language shift in India. A competing alternative model based on both archaeobotanical material evidence and

Received July 26, 2005; accepted for publication November 3, 2005; electronically published December 16, 2005.

Address for correspondence and reprints: Dr. Partha P. Majumder, Human Genetics Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700108, India. E-mail: ppm@isical.ac.in

Am. J. Hum. Genet. 2006;78:202–221. 0002-9297/2006/7802-0004$15.00

colloquial agricultural terms, however, more parsimoniously postulates that early Dravidian has a epipaleolithic preagricultural heritage with origins near a South Asian core region, suggesting possible independent centers of plant domestication within the Indian peninsula by indigenous peoples (Fuller 2003).

The usually strong correlation of haploid Y-chromosome diversity with geography makes it a particularly effective gauge of both range and population expansions (Hammer et al. 2001; Underhill et al. 2001*b*). As the global architecture of the Y-chromosome phylogeny has become well defined (Underhill et al. 2000) and refined (Jobling and Tyler-Smith 2003), the complex Y-chromosome binary HG diversity of southern Asia has begun to be described (Qamar et al. 2002; Basu et al. 2003; Kivisild et al. 2003*a,* 2003*b*; Cordaux et al. 2004), revealing an intricate composition and phylogeography concomitant with the initial and subsequent repeated colonization events of South Asia, generally consistent with that reported for mtDNA (Kivisild et al. 1999; Quintana-Murci et al. 2004). To a considerable extent, molecular studies of haploid mtDNA and Y-chromosome structure in southwestern Asians have been framed in the context of the contemporary social hierarchy and/or linguistic fabric of various groups (Bamshad et al. 1998, 2001; Ramana et al. 2001; Roychoudhury et al. 2001; Basu et al. 2003; Wooding et al. 2004). Although such studies have provided interesting insights into the genetic history of paternal lineages in India, there have been two main limitations arising from a combination of ethnically ill-defined populations, limited geographic sampling, inadequate molecular resolution, and inappropriate statistical methods. Another underappreciated factor in some studies is the innate complexity of caste origins, including the recognition that some castes have tribal origins (Majumder 2001). The present study addresses these deficiencies by using a large set of well-defined ethnic groups, wider geographic sampling, high molecular resolution to define HGs, comprehensive haplotyping with use of a large set of microsatellite markers, evaluation of phylogeographic patterns of HG frequencies, and haplotype diversification, as well as estimation of both HG-based expansion and population-divergence times. The major issues addressed include the origins of castes and tribes, the extent of the impact of recent Central Asian agriculturists on the South Asian gene pool, and the geographic source of Dravidian speakers. We attempt to assess both external and internal demic expansions via a detailed phylogeographic approach and microsatellite-based analyses using samples assessed at similar comprehensive levels of Y-chromosome genetic resolution, comprising groups from the Indian peninsula, the Indus Valley, and East Asia. Additional insights regarding the origins of HGs are revealed from an analysis of genetic variability among social categories and linguistic clusters of populations.

Material and Methods

Population Samples

The social structure of the Indian population is dominated by the Hindu caste system. Most contemporary Indian populations belong to the Hindu religious fold and are hierarchically arranged in four main caste clusters; that is, Brahmin (priestly class, upper), Kshatriya (warrior class, middle), Vysya (business class, middle), and Sudra (menial labor class, lower). There are many tribal groups that are predominantly ancestor worshippers and are believed to be autochthones of India. In addition, there are several religious communities that practice different religions; that is, Islam, Christianity, Sikhism, Judaism, etc.

High-resolution assessment of Y-chromosome binary-HG composition was conducted on 728 Indian samples representing 36 populations, including 17 tribal populations, from six geographic regions and different social and linguistic categories. They comprise (Austro-Asiatic) Ho, Lodha, Santal, (Tibeto-Burman) Chakma, Jamatia, Mog, Mizo, Tripuri, (Dravidian) Irula, Koya Dora, Kamar, Kota, Konda Reddy, Kurumba, Muria, Toda, and (Indo-European) Halba. The 18 castes include (Dravidian) Iyer, Iyengar, Ambalakarar, Vanniyar, Vellalar, Pallan and (Indo-European) Koknasth Brahmin, Uttar Pradhesh Brahmin, West Bengal Brahmin, Rajput, Agharia, Gaud, Mahishya, Maratha, Bagdi, Chamar, Nav Buddha, and Tanti. With the exception of the Koya Dora and Konda Reddy groups, these samples have been described elsewhere (Basu et al. 2003). The samples were collected after receipt of informed consent. In addition, 176 samples representing 8 Pakistani populations and 175 East Asian samples from 18 population groups were also studied. Those samples are from the Human Genome Diversity Cell Line Panel (Cann et al. 2002). A description of relevant population characteristics and sample sizes are given in table 1.

Polymorphisms and Haplotyping

Binary polymorphisms were genotyped using either denaturing high-performance liquid chromatography, RFLP, or direct sequencing, following the hierarchy of the Y-chromosome phylogeny. The majority of the binary markers have been described elsewhere (Underhill et al. 2001*b*; Cinnioğlu et al. 2004). Figure 1 shows the phylogenetic relationships and Y Chromosome Consortium (YCC) nomenclature assignments of informative binary markers. All 351 non-Indian and 724 of the 728 Indian samples were also genotyped at 10 Y-microsatellite loci described elsewhere (Cinnioğlu et al. 2004), to estimate haplotype variation within an HG defined by binary markers. In addition, the dinucleotide *DYS413* microsatellite locus was analyzed in all J2-M172–related lineages (Malaspina et al. 1997).

We introduce eight new informative binary polymorphisms (M346, M356, M357, M377, M378, M379, M407, and M410), a deletion, and seven single-nucleotide substitutions, which are described in table 2. The M377, M378, and M379 markers occur within the same PCR-amplified fragment. The

Table 1

Geographic, Social, and Linguistic Description of Populations Studied

Country, Region, and Population (Code)	Social Category	Linguistic Category	No. of Chromosomes
South Asia:			
India:			
North:			
Chamar (cha)	Caste, low	Indo-European	18
Muslim (mus)	Religious group	Indo-European	19
Rajput (raj)	Caste, high	Indo-European	29
Uttar Pradesh Brahmin (ubr)	Caste, high	Indo-European	14
Northeast:			
Chakma (chk)	Tribe	Tibeto-Burman	4
Jamatia (jam)	Tribe	Tibeto-Burman	30
Mog (mog)	Tribe	Tibeto-Burman	5
Mizo (mzo)	Tribe	Tibeto-Burman	27
Tripuri (tri)	Tribe	Tibeto-Burman	21
East:			
Agharia (agh)	Caste, middle	Indo-European	10
Bagdi (bag)	Caste, low	Indo-European	11
Gaud (gau)	Caste, middle	Indo-European	5
Ho (ho)	Tribe	Austro-Asiatic	30
Lodha (lod)	Tribe	Austro-Asiatic	20
Mahishya (mah)	Caste, middle	Indo-European	13
Santal (san)	Tribe	Austro-Asiatic	14
Tanti (tan)	Caste, low	Indo-European	7
West Bengal Brahmin (wbr)	Caste, high	Indo-European	18
South:			
Ambalakarar (amb)	Caste, middle	Dravidian	29
Irula (ila)	Tribe	Dravidian	30
Iyengar (iyn)	Caste, high	Dravidian	30
Iyer (iyr)	Caste, high	Dravidian	29
Koya Dora (kdr)	Tribe	Dravidian	27
Kota (kot)	Tribe	Dravidian	16
Konda Reddy (krd)	Tribe	Dravidian	30
Kurumba (kur)	Tribe	Dravidian	19
Pallan (pln)	Caste, low	Dravidian	29
Toda (tod)	Tribe	Dravidian	8
Vanniyar (van)	Caste, middle	Dravidian	25
Vellalar (vlr)	Caste, middle	Dravidian	31
Central:			
Halba (hal)	Tribe	Indo-European	21
Kamar (kmr)	Tribe	Dravidian	30
Muria (mur)	Tribe	Dravidian	20
West:			
Koknasth Brahmin (kbr)	Caste, high	Indo-European	25
Maratha (mrt)	Caste, middle	Indo-European	20
Nav Buddha (nbh)	Caste, low	Indo-European	14
Total (India)			728
Pakistan[a]:			
North:			
Burusho		Burushaski[a]	20
Hazara		Indo-European	25
Kalash		Indo-European	20
Pathan		Indo-European	20
South:			
Balochi		Indo-European	25
Brahui		Akin to Dravidian	25
Makrani		Indo-European	20
Sindhi		Indo-European	21
Total (Pakistan)			176
East Asia:			
Cambodia:			
Cambodian		Austro-Asiatic	6
China:			
Dai		Sino-Tibetan	7
Daur		Altaic	7
Han		Sino-Tibetan	24
Hezhen		Altaic	6
Lahu		Sino-Tibetan	7
Miaozu		Sino-Tibetan	7
Mongola		Altaic	7
Naxi		Sino-Tibetan	8
Oroqen		Altaic	7
She		Sino-Tibetan	7
Tu		Altaic	7
Tujia		Sino-Tibetan	9
Uygur		Altaic	8
Xibo		Altaic	8
Yizu		Sino-Tibetan	9
Japan:			
Japanese		Altaic	23
Siberia:			
Yakut		Altaic	18
Total (East Asia)			175

[a] Isolated language; remains unclassified.

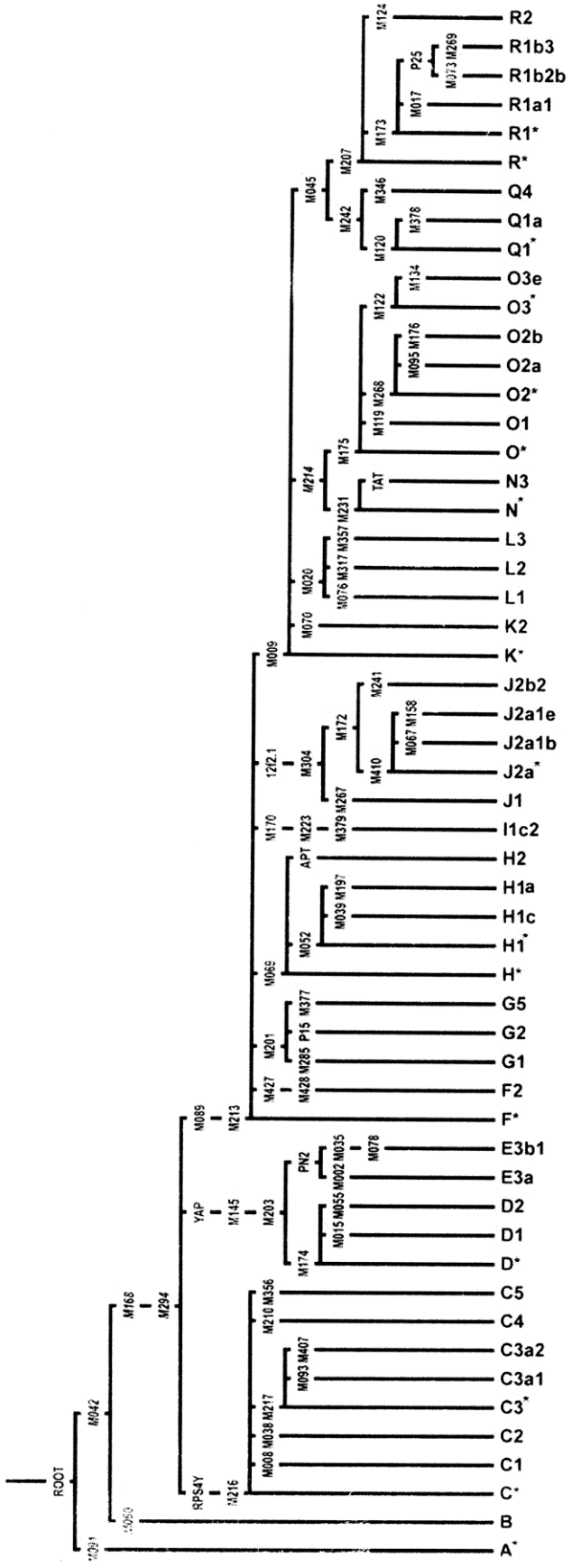


Figure 1 Phylogenetic relationships and nomenclature of Y-chromosome binary HGs observed in India, Pakistan, and East Asia. The 69 informative binary markers that were haplotyped are indicated in plain font. The markers in italics are shown to provide phylogenetic context. The following polymorphisms were typed but not observed in Indian populations: M8, M25, M38, M39, M56, M67, M70, M88, M97, M147, M170, M204, M210, M217, M282, M297, and M319.

Table 2

Primers and Specifications of New Y-Chromosome Binary Markers

Position (bp)	Primer (5′→3′)		Size (bp)	YCC HG
	Forward	Reverse		
33	ccccgtttttttcctctctgcc	aatctgccttccaacaaacc	419	Q4
185	aactcttgataaaccgtgctg	tccaatctcaattcatgcctc	215	C5
234	ccccgtttttttcctctctgcc	cacgtaacctgggatggtcata	458	L3
40	tatgcatttgttgagtatatgtc	gttctgaatgaaagttcaaacg	326	G5
114	tatgcatttgttgagtatatgtc	gttctgaatgaaagttcaaacg	326	Q1a
135–136	tatgcatttgttgagtatatgtc	gttctgaatgaaagttcaaacg	326	I1c2
149	gttataccctgctctaaagtgcttc	gtagagatggggcttcaccgtgttac	418	C3a2
115	caatcattgaccttaagtctgagtccc	actggatacctttctaggaagaattg	395	J2a

complete data set, with HG and repeat numbers at the 10 microsatellite loci for each individual, is provided in table 3.

Analysis of Haplotype Data

Microsatellite variances were estimated for each population with a sample size of at least five individuals over 10 loci for most binary HGs, to assess the relative amount of accumulated diversity as a function of geography. Spatial surfaces of both binary HG frequency and HG-associated microsatellite variance distributions for the India and Pakistan samples were computed using the Kriging procedure (Delfiner 1976) with Surfer Systems Golden software. The J2a-M410 frequency distribution across Eurasia was plotted using the Mapping Toolbox in The MathWorks MATLAB software package. This tool uses biharmonic spline interpolation to generate the surface data, as described by Sandwell (1987). In addition, frequency surface distributions for HGs J2a*-M410 were also generated, with use of 82 sets of data (sample size $\geq$20) spanning portions of Africa, Europe, the Near East, and West Asia (see table 4). The age of a binary HG was estimated from microsatellite variation within the HG with use of the method described by Zhivotovsky et al. (2004), with slight modifications (details provided in appendix A [online only]). Of note, there was no modal microsatellite haplotype within many HGs. Whereas the modal haplotype can reasonably be assumed to be the founding haplotype, the absence of such a haplotype prevented us from arbitrarily assuming any other existing haplotype as the founding one; instead, we used a logical and consistent procedure across all HGs. We used the median values of repeat scores at each microsatellite locus within each HG and assumed that the haplotype formed by the median values of the repeat scores is the founding haplotype. Although the median and the founding haplotypes coincide in the beginning of growth of microsatellite variation (L.A.Z., unpublished observations), working under that assumption would result in an underestimate of microsatellite variation if the median haplotype deviates far from an actual founder. However, since we used the median haplotype, irrespective of the presence of a modal haplotype, across all HGs, we believe that estimates of the relative ages of the HGs will be correct, even if their actual ages are underestimates. The Y-chromosome microsatellite diversity and variance in repeat number were calculated using the software package Microsat, version 1.5d. Equality of proportions of HGs among subsets of populations (e.g., social or linguistic categories) was tested using contingency χ^2 tests, with Yates's correction when required. Analysis of molecular variance (AMOVA) of microsatellite data was performed using the Arlequin package, version 2.000. Within specific HGs, median-joining networks of microsatellite haplotypes were constructed using the Network package, version 4.1.0.9 (Fluxus Engineering).

Results

The Phylogeny of South Asian Y-Chromosome Binary HGs

The phylogenetic relationships of Y-chromosome binary HGs determined for 1,079 samples are presented in figure 1. A total of 69 of 71 polymorphisms genotyped were informative and defined 52 HGs in the data set. The pooled frequency distributions of HGs for the 36 Indian, 8 Pakistani and 18 East Asian populations are given in table 5. (Additional detailed HG frequencies are given in tables 6 and 7.) Eight HGs display frequencies >5% within India and account for 95.8% of the samples. They are, in descending frequency order, HGs H and its subclades H1*, H1c, H1a, and H2 (26.4%); R1a1-M17 (15.8%); O2a-M95 (14.6%); R2-M124 (9.3%); J2-M172 (9.1%); O3e-M134 (8.0%); L1-M76 (6.3%); and F*-M89 (5.2%). In Pakistan, nine HGs exceed 5% frequency and account for 83.6% of the samples. They include HGs R1a1-M17 (24.4%), L*-M20 (13.1%), J2-M172 (11.9%), R2-M124 (7.4%), R1b-P25 (7.4%), G-M201 (6.3%), C3-M217 (6.8%), H-M69* (6.3%), and L1-M76 (5.1%). In the East Asian samples studied, the following seven HGs exceed 5% frequency and account

Table 3

Complete Data Set of Y-Chromosomal HGs, Numbers of Repeats at 10 Microsatellite Loci, and Descriptions of Populations

The table is available in its entirety in the online edition of *The American Journal of Human Genetics*.

Table 4

Frequencies of HG J2a*-M410 in Different Populations, with Latitudes and Longitudes

The table is available in its entirety in the online edition of *The American Journal of Human Genetics*.

for 75.4% of the total: O3e-M134 (15.4%), C3-M217 (13.1%), N3-TAT (12.0%), O2a-M95 (10.9%), O3-M122(xO3e) (11%), N-M231(xN3) (6.3%), and R1b2-M73 (6.3%). The HG frequency distributions among the three geographical regions and those between any pair of regions are all significantly different (all P values <.0001).

The J2-M172 Clade Is Composed of Two Sister Clades

New phylogenetic resolution has been achieved within the J2-M172 clade with the discovery of the M410 nucleotide A→G substitution (table 2). Now, all J2-M172–derived lineages can be assigned to one of two sister clades—namely, J2a*-M410 or J2b*-M12—which necessitates an updated revision of the previous "haplogroup by lineage" YCC nomenclature for J2 (Jobling and Tyler-Smith 2003). The J2*-M172 phylogenetic revisions are presented in figure 2. We include the *DYS413*≤18 allele-repeat node in the phylogeny, as suggested by Di Giacomo et al. (2004). It is notable that no J2*-M172 HG lacking both M410- and M12-derived alleles has yet been observed. The *DYS413* locus was typed in M410-derived samples from India, Pakistan, and Turkey. The vast majority displayed the ≤18 allele repeat, although 16/118 in Turkey had alleles ≥19, as did 5/17 in Pakistan and 5/28 in India, 4 of which were restricted to the Dravidian-speaking Iyengar and Iyer upper castes. The J2 clade is nearly absent among Indian tribals, except among Austro-Asiatic–speaking tribals (11%). Among the Austro-Asiatic tribals, the predominant J2b2 HG occurs only in the Lodha. J2 is present in significantly higher ($P < .001$) frequency among Dravidian castes (19%) than among Indo-European castes (11%). In Pakistan, the frequency (12%) is similar to that among Indian Indo-European castes, but this clade is nearly absent (1%) in East Asia.

Discovery of Five New Clades That Improve HG Topology

We report five new clades that improve the HG topology within the Y-chromosome genealogy. The new subclade C5-M356 accounts for 85% of the former C* HGs. Although its overall frequency is only 1.4% in the Indian sample, it occurs in all linguistic groups and in both tribes and castes. It also occurs in one Dravidian Brahui in Pakistan (table 5). The new L3-M357 subclade accounts for 86% of L-M20(xL1xL2) chromosomes in Pakistan but occurs only sporadically (3/728) in India. All Indian HG Q representatives belong to the new M346-subclade. This new Q clade will aid in future studies attempting to narrow the candidate Asian/Siberian precursors of Native American chromosomes. The

Table 5

Y-Chromosome HG Frequencies in India, Pakistan, and East Asia

	No. (%) of Y-Chromosome HGs		
HG	India (n = 728)	Pakistan (n = 176)	East Asia (n = 175)
C*-M216/RPS4Y	2 (.27)		1 (.57)
C1-M008			1 (.57)
C3-M217		12 (6.82)	19 (10.86)
C3a1-M093			1 (.57)
C3a2-M407			3 (1.71)
C5-M356	11 (1.51)	1 (.57)	
D*-M174			1 (.57)
D1-M015			9 (5.14)
E3a-M002		1 (.57)	
E3b1-M035\M078		2 (1.14)	
F*-M089/M213	38 (5.22)		
F2-M427\M428			5 (2.86)
G1-M285		1 (.57)	
G2-P15	9 (1.24)	8 (4.55)	
G5-M377		2 (1.14)	
H-M069*	29 (3.98)	1 (.57)	
H1-M039			1 (.57)
H1-M052	146 (20.05)		
H1-M197	1 (.14)	1 (.57)	
H1*-M052		9 (5.11)	
H2-APT	16 (2.2)		
I1c2-M170\M223\M379		1 (.57)	
J1-M267	2 (.27)	6 (3.41)	
J2a-M410	26 (3.57	15 (8.52)	2 (1.14)
J2a1b-M067		2 (1.14)	
J2a1e-M158	2 (.27)		
J2b2-M241	38 (5.22)	4 (2.27)	
K*-M009		1 (.57)	
K2-M070			1 (.57)
L1-M076	46 (6.32)	9 (5.11)	
L2-M317		2 (1.14)	1 (.57)
L3-M357	3 (.41)	12 (6.82)	
N-M231			11 (6.29)
N3-TAT			21 (12)
O*-M175			1 (.57)
O1-M119			7 (4)
O2*-M268			5 (2.86)
O2a-M095	106 (14.56)		19 (10.85)
O2b-M176			8 (4.57)
O3-M122	3 (.41)	3 (1.7)	19 (10.86)
O3e-M134	58 (7.97)	1 (.57)	27 (15.43)
Q1-M120		1 (.57)	
Q1a-M378		2 (1.14)	
Q4-M346	3 (.41)	3 (1.7)	
R*-M207	2 (.27)	6 (3.41)	
R1*-M173		1 (.57)	
R1a1-M017	115 (15.8)	43 (24.43)	
R1b2b-M073		8 (4.55)	11 (6.29)
R1b3-M269	4 (.55)	5 (2.84)	1 (.57)
R2-M124	68 (9.34)	13 (7.39)	

Table 6

HG Frequencies in Social and Linguistic Subgroups of Indian Populations

	No. (%) of HGs									
	Tribe				Dravidian Caste			Indo-European Caste		
HG	Austro-Asiatic ($M = 3$; $n = 64$)	Dravidian ($M = 8$; $n = 180$)	Tibeto-Burman ($M = 5$; $n = 87$)	Indo-European ($M = 1$; $n = 21$)	Upper ($M = 2$; $n = 59$)	Middle ($M = 3$; $n = 85$)	Lower ($M = 1$; $n = 29$)	Upper ($M = 4$; $n = 86$)	Middle ($M = 4$; $n = 48$)	Lower ($M = 4$; $n = 50$)
C*-M216/RPS4Y		1 (.56)					1 (3.45)			
C5-M356	1 (1.56)	1 (.56)	1 (1.15)		2 (3.39)	1 (1.18)		1 (1.16)	2 (4.17)	2 (4.00)
F*-M089/M213		25 (13.89)		3 (14.29)		6 (7.06)	2 (6.90)	1 (1.16)	1 (2.08)	
G2-P15					7 (11.86)	1 (1.18)				1 (2.00)
H-M069*	5 (7.81)	2 (1.11)			1 (1.69)	12 (14.12)		1 (1.16)	1 (2.08)	7 (14.00)
H1-M052	9 (14.06)	66 (36.67)	2 (2.30)	5 (23.81)	4 (6.78)	17 (20.00)	4 (13.79)	9 (10.47)	16 (33.33)	14 (28.00)
H1-M197								1 (1.16)		
H2-APT	1 (1.56)	9 (5.00)		2 (9.52)			1 (3.45)	1 (1.16)	1 (2.08)	1 (2.00)
J1-M267		1 (.56)						1 (1.16)		
J2a-M410		3 (1.67)			8 (13.56)	2 (2.35)	1 (3.45)	8 (9.30)	2 (4.17)	1 (2.00)
J2a1e-M158					1 (1.69)					1 (2.00)
J2b2-M241	7 (10.94)				2 (3.39)	16 (18.82)	3 (10.34)	4 (4.65)	4 (8.33)	1 (2.00)
L1-M076		10 (5.56)			10 (16.95)	16 (18.82)	4 (13.79)	2 (2.33)	3 (6.25)	1 (2.00)
L3-M357							1 (3.45)	2 (2.33)		
O2a-M095	34 (53.13)	48 (26.67)	16 (18.39)	6 (28.57)					1 (2.08)	1 (2.00)
O3-M122			2 (2.30)							1 (2.00)
O3e-M134			57 (65.52)					1 (1.16)		
Q4-M346				1 (4.76)		1 (1.18)		1 (1.16)		
R*-M207					1 (1.69)					1(2.00)
R1a1-M017		5 (2.78)	4 (4.60)	4 (19.05)	17 (28.81)	10 (11.76)	7 (24.14)	39 (45.35)	5 (10.42)	13 (26.00)
R1b3-M269							1 (3.45)		3 (6.25)	
R2-M124	7 (10.94)	9 (5.00)	5 (5.75)		6 (10.17)	3 (3.53)	4 (13.79)	14 (16.28)	9 (18.75)	5 (10.00)

Note.—M = no. of population groups, n = no. of chromosomes.

Table 7

HG Frequencies in Countries Other Than India (Companion to Table 6)

The table is available in its entirety in the online edition of *The American Journal of Human Genetics*.

G5-M377 substitution is independent of G1-M285 and G2-P15 subclades (Cinnioğlu et al. 2004) and occurs in Pakistan. The M379 polymorphism defines the I1c2 subclade, which occurs only among the Pakistani in our study sample.

Phylogeography of J2a-M410 across Eurasia

Our new understanding of the phylogenetic structure concerning HG J2a-M410 has allowed us to combine both our new Indian and Pakistani data ($n = 905$) and relevant data from the literature (6,678 samples from 82 populations; n per population $\geqslant 20$) to construct a geographic map of its distribution based on 76 geographic regions (fig. 3). The J2a-M410 data used in the frequency surface plots are presented in table 4. The sister clade to J2a-M410 is J2b-M12. In India and Pakistan, all J2b members comprise the J2b2-M241 derivative HG. Notable is the high frequency (in 39 of 42 Indian and Pakistani samples only) of a 7-repeat motif at the A7.2 microsatellite locus, in contrast to the 8-repeat motif in Europe.

Inference of HG Expansions and Polarity of Spread

Geographic origins of HG expansions can be inferred from both frequency and associated diversity (Barbujani 2000), with spatial levels of accumulated microsatellite variability providing a metric for assessing directionality of movement and to help disentangle complexities associated with population stratification. It is important to recognize that regions of highest HG frequency are not necessarily representative of origin. An obvious example is HG C, which displays its highest frequency in Polynesia (Kayser et al. 2000), but Polynesia is one of the last regions known to be colonized by modern humans. Associated microsatellite diversification can detect clines not obvious in binary HG-frequency data. In Turkey, for example, the most frequent HG, J2-M172, showed no clinal pattern with geography, but its associated average Y-chromosome microsatellite variation was significantly inversely correlated with latitude (Cinnioğlu et al. 2004), consistent with a model of demic expansion (Edmonds et al. 2004). However, highly associated microsatellite variance may not always be a reflection of just in situ processes. In some cases, it could also be a consequence of repeated gene flow from diverse sources (Tambets et al. 2003), as evidenced in Central Asia, where high Y-chromosome diversity is attributed to multiple recent events (Zerjal et al. 2002). Thus, regions of high frequency and high variance need not always overlap. HGs H1, R1a1, and R2 (fig. 4) are examples of nonoverlap between regions of highest frequency and highest variance. The region of high variance can also reflect the area from which a subset of chromosomal diversity migrated and then underwent subsequent demographic growth. Despite the potential consequences of population stratification and language shift, additional important insights into geographic origins of Y chromosome HGs can be deduced (see the "Pre–Indo-European Expansions Shape the South Asian Y-Chromosome Landscape" section) from the proportions of HGs distributed among the different social and linguistic groups found in India (table 6). (HG frequencies in Pakistan and other countries are given in table 7.)

We computed F_{ST} values between linguistic subgroups of Indian populations and populations of Pakistan and East Asia. The trend of F_{ST} values with populations of Pakistan is that the Indo-European–speaking population groups of India show small values (0.079) compared with the Dravidian (0.121), Tibeto-Burman (0.229), or Austro-Asiatic (0.139) groups. In general, the F_{ST} values with populations of the southern Pakistan are lower than those of northern Pakistan. These F_{ST} values are consistent with known historical events. However, the F_{ST} values of Indian populations with those of East Asia do not follow an expected pattern; for example, the F_{ST} value between East Asian and Indian Tibeto-Burman populations (0.253) is higher than that between Indian Indo-European populations (0.163). This might be because of pooling East Asian populations with varied demographic histories. The pattern does not substantially change with the exclusion of the Japanese, who may have a considerably different population history from that of other East Asian populations. When the Japanese are excluded, the F_{ST} value between East Asians and Tibeto-Burmans reduces to 0.238, whereas that between East Asians and Indo-Europeans remains essentially unchanged (0.161).

We also performed AMOVA using the microsatellite haplotype data. The results are presented in table 8. The AMOVA results show that the percentage of between-population variation is much higher in India (18%) and East Asia (17%) than in Pakistan (6%). In India, the variation within a linguistic category is far greater for tribal populations (16%) than for castes (6%). In Pakistan, genetic differences between northern and southern regions is low (0.01%).

Figure 4 shows spatial maps of HG frequencies and microsatellite variance distributions across India and Pakistan for several HGs that are based on pooled variances of repeat numbers at the microsatellite loci within HGs. The variances at individual loci within HGs are presented in table 9. It is noteworthy that contributions of individual microsatellite loci to the pooled variances

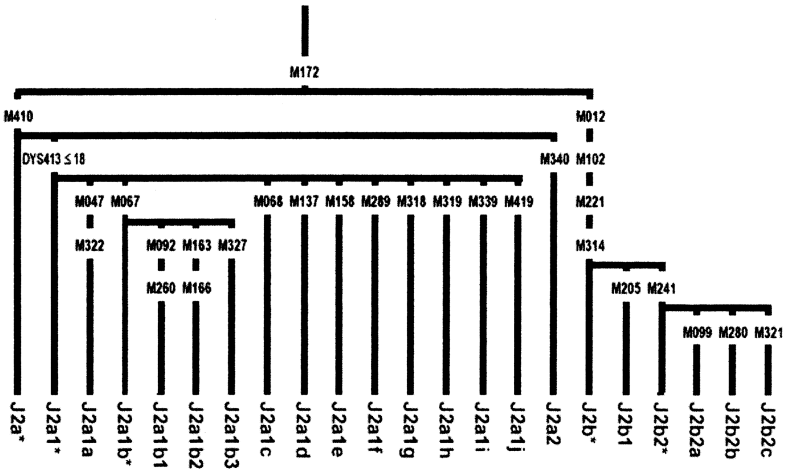


Figure 2 Revised phylogenetic relationships and nomenclature for Y-chromosome HG J2

are quite variable within and across HGs. These variations in the contribution of individual loci are certainly a reflection of founding events and also possibly of the variation in mutation rates across microsatellite loci. HGs L1-M76 and H1-M52 have peak variance distributions in the Maharashtra region in coastal western India. Other lineages (F*-M89; H2-APT) have higher variance patterns in Tamil Nadu and Andhra Pradesh near coastal eastern India. Alternatively, the H-M69* and O2a-M95 HGs display higher variances in northeastern India and the West Bengal region, respectively. These two HGs also display significant correlations with latitude and longitude, respectively (table 10). Lastly, HG J2b2-M241–related microsatellite variance is higher in Uttar Pradesh near the border of Nepal. It should be noted that numerous Mesolithic sites have been observed in this region (Kennedy 2000).

Indigenous and Exogenous HGs Represented in India

On the basis of the combined phylogeographic distributions of haplotypes observed among populations defined by social and linguistic criteria, candidate HGs that most plausibly arose in situ within the boundaries of present-day India include C5-M356, F*-M89, H-M69* (and its sub-clades H1-M52 and H2-APT), R2-M124, and L1-M76. The congruent geographic distribution of H-M69* and potentially paraphyletic F*-M89 Y chromosomes in India suggests that they might share a common demographic history.

A median-joining network analysis (not shown) of F*-M89 microsatellite haplotypes in Indians and East Asian Lahu suggested divergence between these two populations. This is confirmed by the discovery of the linked HG F2-M427 and M428 markers (table 2) that are restricted to the Lahu in our data set. HG R2-M124 occurs with a frequency of 9.3% in India, consistent with 8%–10% reported elsewhere (Kivisild et al. 2003*a*; Cordaux et al. 2004). The decreasing frequency of R2—from 7.4% in Pakistan to 3.8% in Central Asia (Wells et al. 2001) to 1% in Turkey (Cinnioğlu et al. 2004)—is consistent with the pattern observed for the autochthonous Indian H1-M52 HG.

It is noteworthy that no C3-M217–derived lineages typical of East and Central Asia have been observed in the Indian samples reported thus far (Redd et al. 2002; Kivisild et al. 2003*a*; the present study). Conversely, HGs of likely exogenous origin include J2a-M410 and J2b-M12 in the Indus Valley, whereas HGs O2a-M95 and O3e-M134 have their most likely origin in Southeast Asia, judging from the fact that the HG O lineages observed in India represent a minor subset of East Asian variation (Su et al. 1999). Interestingly, within India, R1a1-M17, R2-M124, and L1-M76 display considerable frequency and HG-associated microsatellite variance (table 9). The widespread geographic distribution of HG R1a1-M17 across Eurasia and the current absence of informative subdivisions defined by binary markers leave uncertain the geographic origin of HG R1a1-M17. However, the contour map of R1a1-M17 variance shows the highest variance in the northwestern region of India (fig. 4).

The Age of Microsatellite Variation in the Majority of Indian HGs Exceeds 10,000–15,000 Years

Variances in repeat numbers of Y-chromosome microsatellites, pooled over the 10 loci, within HGs with sample size of $n \geq 5$ are given in table 9. Table 11 presents the estimates of the age of microsatellite variation by HG, computed in India overall as well as by language and social hierarchy.

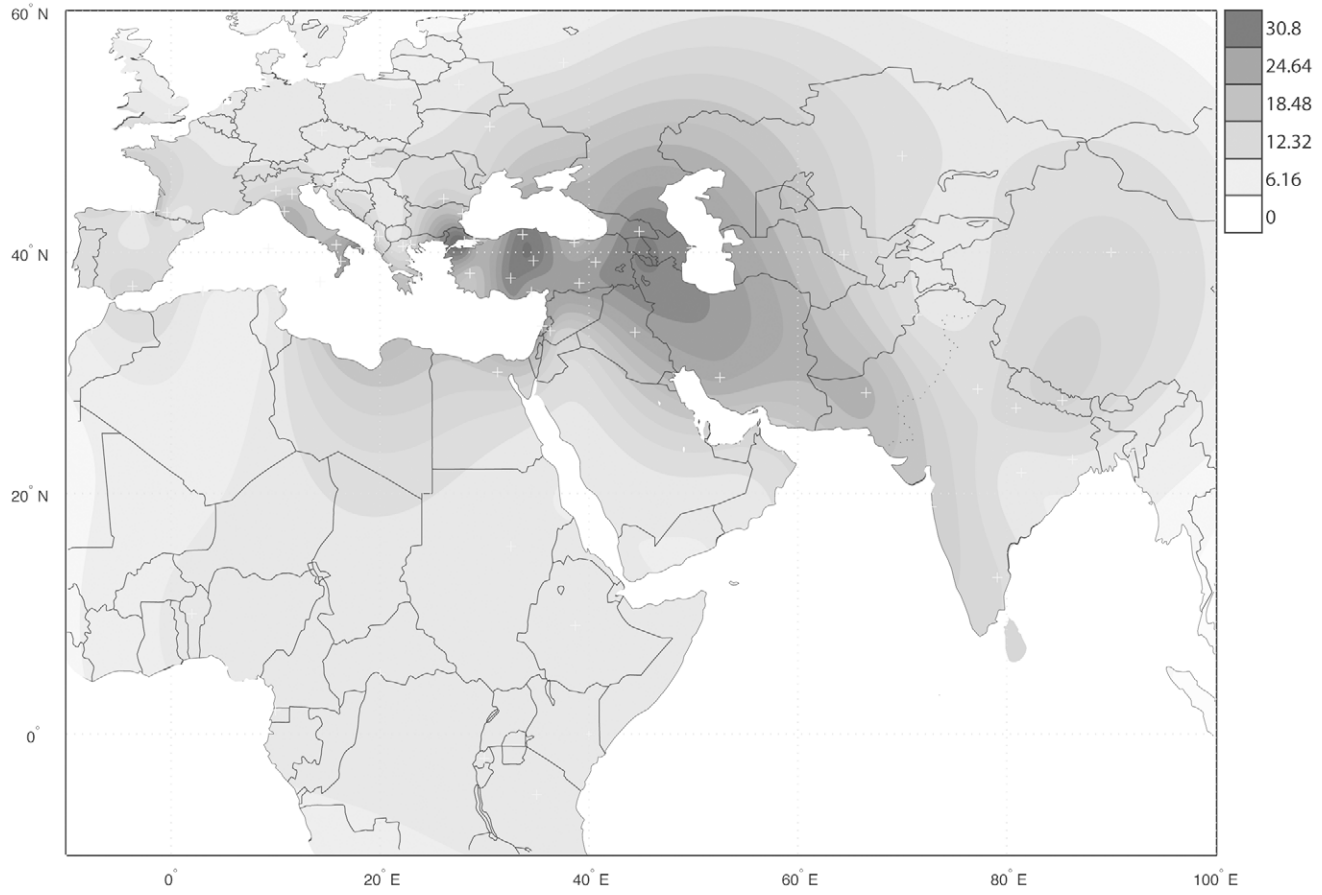


Figure 3 Spatial frequency distribution of Eurasian HG J2a-M410–related lineages. Plus signs (+) indicate geographic locations of population samples listed in table 4.

Discussion

The HG composition for southwestern Asia is complex and represents HGs that likely originated in situ as well as those that arrived from external sources. Although it is generally accepted that Indo-European languages were introduced to present-day India relatively recently (~3.6 thousand years ago [KYA]) from the northwest (Beekes 1995), interpretations differ regarding their influence on the pre-existing gene pool (Kivisild et al. 2003*a*; Cordaux et al. 2004). We investigated this issue, using a combination of more-extensive population sampling and higher molecular resolution, and determined both the polarity and temporality of different Y-chromosome HGs, using not only HG frequencies but also the extent of microsatellite variability within HGs.

Pre–Indo-European Expansions Shape the South Asian Y-Chromosome Landscape

On the basis of a broad distribution—involving all social and linguistic categories in India—and relatively high diversification patterns, it can be concluded that representatives of HGs C5-M356 H-M69*, F*, L1, and R2 have ancestry indigenous to the Asian subcontinent. The relatively high frequency of O2a-M95 lineages centered in Orissa and neighboring regions (fig. 4) across all linguistic categories of tribes (table 6)—coupled with its virtual absence in all caste groups—supports a model of independent origins (Cordaux et al. 2004). However, the considerable age of Y-microsatellite variation of R1a1, R2, and L1 (table 11) and the spatial frequency plots of these lineages (fig. 4) do not support the claim that the presence of these lineages among tribes could be due to occasional recent admixture only (Cordaux et al. 2004). Interestingly, a median-joining network analysis (fig. 5) of O2a-M95 microsatellite haplotypes in Indians suggests a partition between (1) Tibeto-Burman speakers and (2) Austro-Asiatic and Dravidian speakers who display more Y-microsatellite diversification than the Tibeto-Burmans.

Admixture analysis cannot distinguish between recent and ancient gene flow or directionality of flow. In addi-

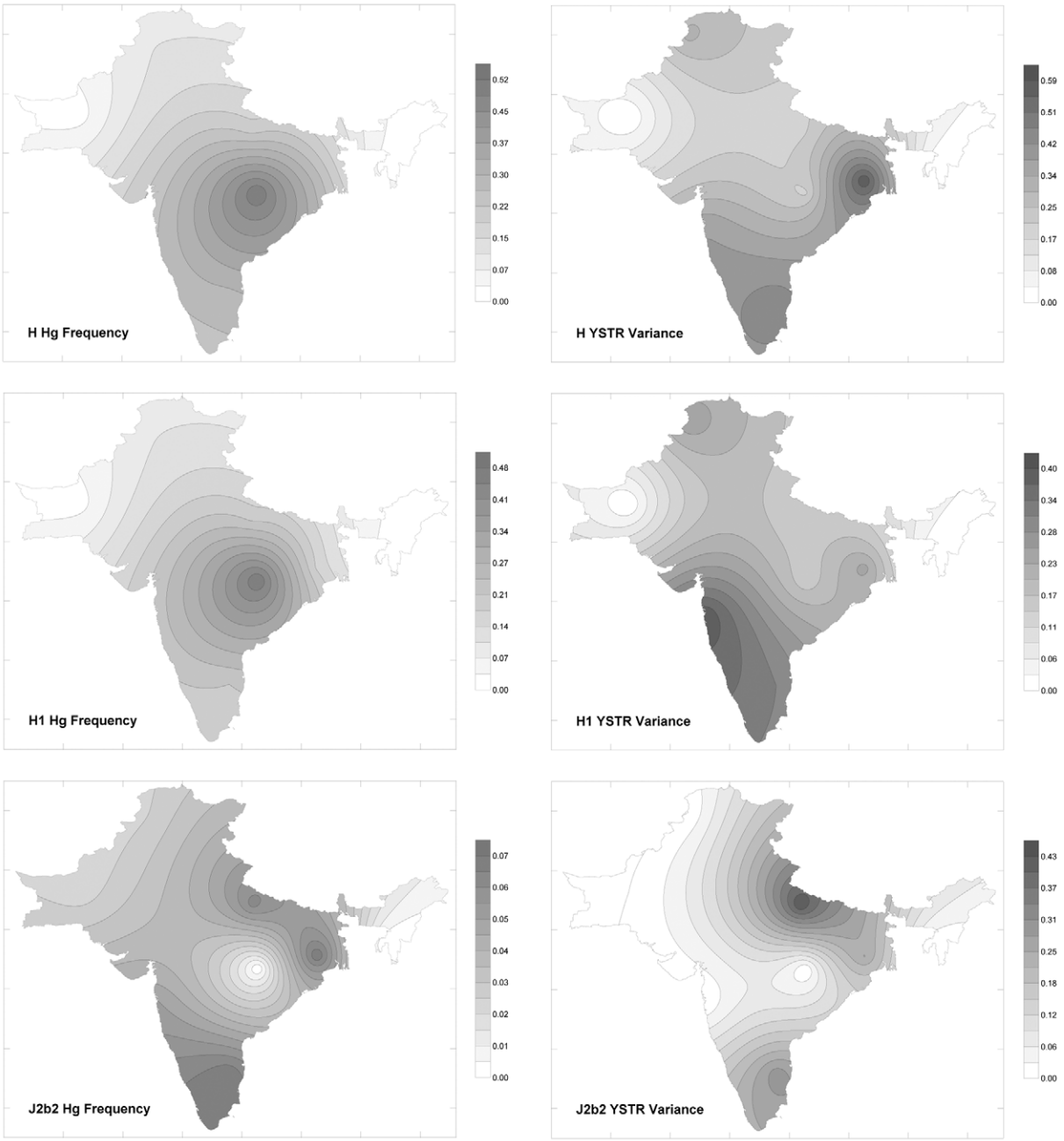
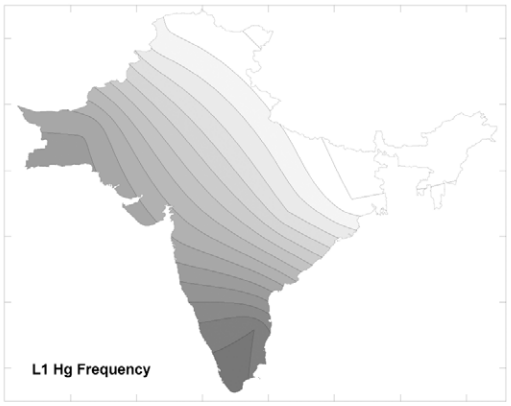
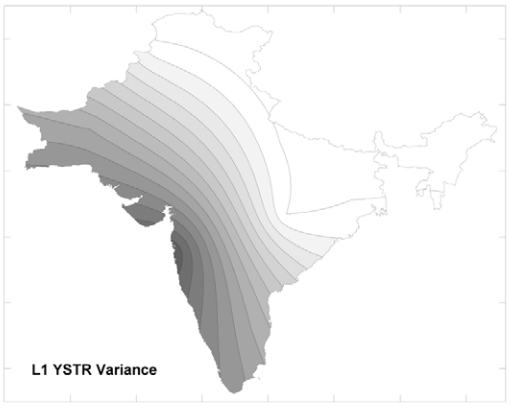
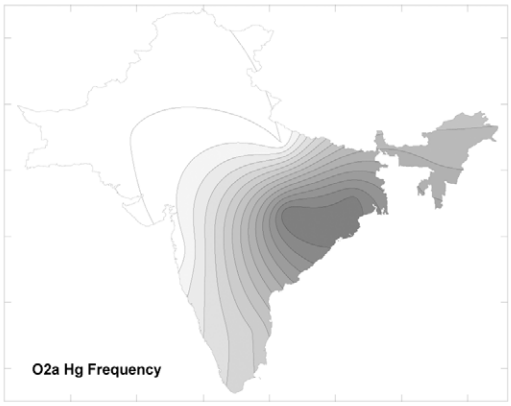
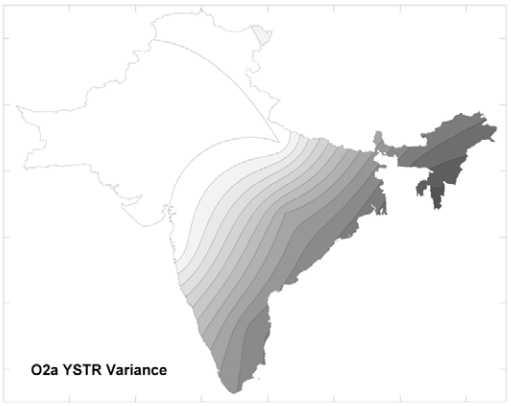


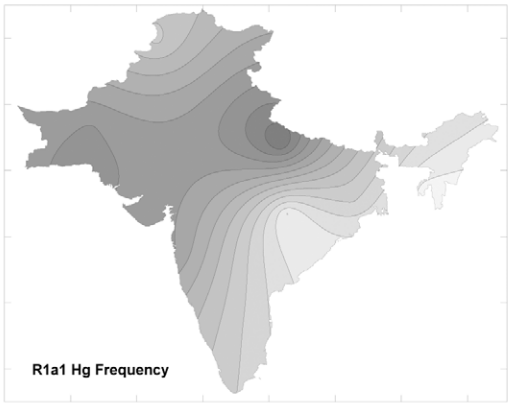
Figure 4 Spatial frequency and mean microsatellite variance distributions of Y-chromosome HGs in present-day India and Pakistan

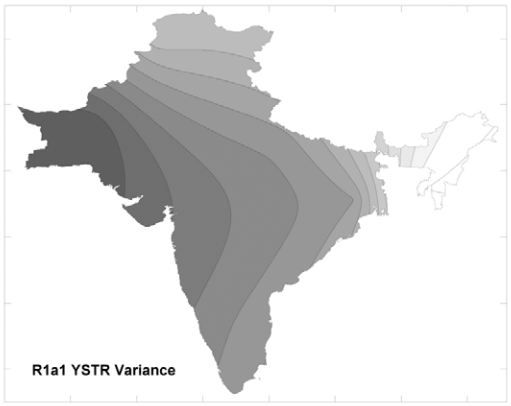

0.13
0.11
0.09
0.07
0.05
0.04
0.02
0.00
L1 Hg Frequency

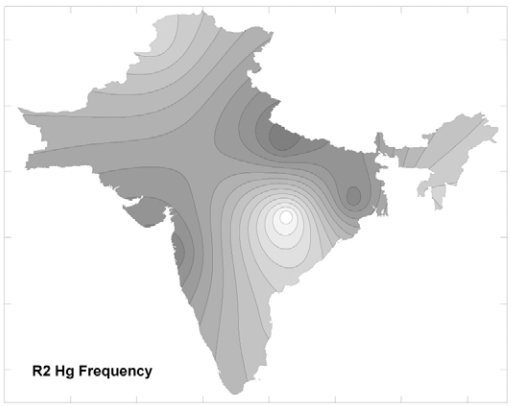

0.34
0.29
0.24
0.19
0.15
0.10
0.05
0.00
L1 YSTR Variance


0.30
0.25
0.21
0.17
0.13
0.08
0.04
0.00
O2a Hg Frequency


0.29
0.25
0.21
0.17
0.13
0.08
0.04
0.00
O2a YSTR Variance


0.40
0.34
0.29
0.23
0.17
0.11
0.06
0.00
R1a1 Hg Frequency


0.40
0.34
0.28
0.23
0.17
0.11
0.06
0.00
R1a1 YSTR Variance


0.16
0.14
0.12
0.09
0.07
0.05
0.02
0.00
R2 Hg Frequency

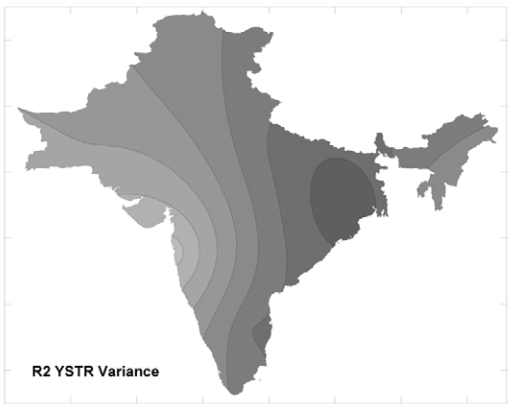

0.33
0.28
0.23
0.19
0.14
0.09
0.05
0.00
R2 YSTR Variance

Table 8

AMOVA Results Based on Y-Microsatellite Haplotype Frequencies

Region and Group	Total Variation (%)		
	Within Populations	Between Populations (within Groups)	Between Groups
India:			
All (36 populations considered one group)	82.38	17.62	...
Caste:			
Two linguistic groups (Indo-European and Dravidian)	91.59	6.38	2.03
Tribal:			
Four linguistic groups	71.40	15.90	12.70
Pakistan:			
All (8 populations considered one group)	93.86	6.14	...
Two geographical groups[a]	93.75	6.24	.01
East Asia:			
All (18 populations considered one group)	82.62	17.38	...

[a] South: Sindhi, Balochi, Brahui, and Makrani; north: Burusho, Hazara, Kalash, and Pathan.

tion, inferences based on basal HG frequencies alone and limited geographic sampling can be misleading and potentially incorrect. This shortcoming is especially magnified when coupled with inadequately resolved HG structure and small sample sizes. In HGs R1a1 and R2, the associated mean microsatellite variance is highest in tribes (table 12), not castes. This is a clear contradiction of what would be expected from an explanation involving a model of recent occasional admixture. Beyond taking advantage of highly resolved phylogenetic hierarchy as just an efficient genotyping convenience, a comprehensive approach that leverages the phylogeography of Y-chromosome diversification by using a combination of HG diversification with geography and expansion-time estimates provides a more insightful and accurate perspective to the complex human history of South Asia. The data regarding HG L are particularly instructive in this regard. Tables 5 and 6 indicate that HGs L1-M76, R1a1-M17, and R2-M124 occur in all Dravidian-speaking castes but are rare in tribes (with the exception of R2 in the Austro-Asiatic Lodha tribe and R1a1 in the Indo-European–speaking Halba). Such limited exceptions may be a result of drift. The temptation is to view R1a1 and R2 as terminal HGs with recent shared ancestry, without consideration of levels of subsequent diversification at either the sub-HG or microsatellite haplotype level. Knowledge of HG L diversification reveals a potentially significant flaw in this regard. HG L is defined by the M20 mutation with numerous companion markers (e.g., M11, M27, and M61) that reinforce the definition of this clade. When considered at the general HG level, L, R1a, and R2 all display approximate similarity with respect to population-category apportionment and frequency (Cordaux et al. 2004). With the exception of that of Kivisild et al. (2003*a*), who showed that all Indians were M27-M76 derived, other previous studies of HG L in India have not considered binary diversification beyond the HG L-M20 node. What is distinctive about HG L relative to the other two HGs and other relevant previous studies (Wells et al. 2001; Zerjal et al. 2002), however, is new knowledge concerning the subsequent binary diversification observed in the present study. Specifically, all HG L chromosomes studied apportion to one of three informative sub-HGs—namely, L1-M76, L2-M317, and L3-M357, each with distinctive geographic affiliation and polarity of spread based on microsatellite diversity that help considerably to localize their origins, especially L1 and L3. Noteworthy is the fact that, in India, virtually all members of HG L are L1 derivatives. When analyzed at more-informative levels of HG diversification, previous conclusions obtained at the coarser level become suspect. This critical knowledge imbues HG L with considerably more insight than simple frequency analysis for R1a1 and R2. Although it would be convenient to assume that R1a1 and R2 representatives reflect a recent common demography (Cordaux et al. 2004), it is entirely plausible that they harbor as-yet-undiscovered subsequent haplogroup diversification that approximates the phylogeographic patterns revealed for HG L. The frequency of R1a varies widely in Central Asia and is highest (63%) in the eastern Kyrgyz (Altaic) and Tajiks (Indo-European), both of whom display quite recent times to the most recent common ancestor (Zerjal et al. 2002). The question remains of how distinctive is the history of L1 relative to some or all of R1a1 and R2 representatives. This uncertainty neutralizes previous conclusions that the intrusion of HGs R1a1 and R2 from the northwest in Dravidian-speaking southern tribes is attributable to a single recent event. Rather, these HGs contain considerable demographic complexity, as implied by their high haplotype diversity. Specifically, they could have actually arrived in southern India from a southwestern Asian source region multiple times, with some episodes considerably earlier than others. Considerable archeological evidence exists regarding the presence of Mesolithic

Table 9

SDs of Repeat Numbers at 10 Microsatellite Loci within HGs with Sample Sizes >20, in Various Regions of India and Pakistan

Region and HG (*n*)	SD by Marker									
	DYS19	*DYS388*	*DYS389AB*	*DYS389CD*	*DYS390*	*DYS391*	*DYS392*	*DYS393*	*DYS439*	*DYSA7.2*
Northern India:										
R1a1 (32)	.762	.369	.803	.508	.581	.543	.390	.309	.246	.354
Eastern India:										
All H (36)	.692	1.158	.893	.898	.841	.525	.167	.766	.756	.586
O2a (36)	.398	.000	.681	.319	.401	.467	.232	.560	.618	.554
Southern India:										
F* (30)	.844	.664	.740	.740	1.095	.484	.640	.820	.973	.858
F* and all H (116)	.720	.705	.858	.657	1.010	.486	.399	.773	.844	.700
All H (86)	.649	.260	.891	.629	.937	.490	.275	.569	.709	.563
H1 (62)	.459	.000	.459	.704	.658	.409	.319	.505	.696	.522
J2b2 (21)	.218	.498	.590	.359	.928	.632	.000	.512	.680	.436
L1 (38)	.434	.453	.569	.000	.162	.226	.311	.453	.788	.437
O2a (33)	.394	.174	.500	.561	.242	.489	.000	.517	.781	.545
R1a1 (39)	.683	.223	.785	.595	.686	.486	.160	.354	.451	.320
R2 (22)	.492	.294	.610	.581	.526	.213	.000	.716	.703	.685
Central India:										
F* and all H (43)	.575	.413	.374	.764	.814	.413	.457	.708	.796	.674
All H (37)	.229	.363	.277	.702	.468	.229	.000	.277	.689	.676
H1 (34)	.442	.247	.625	.666	.466	.254	.146	.247	.605	.682
O2a (21)	.000	.000	.598	.402	.316	.463	.000	.784	.301	.402
Southern Pakistan:										
R1a1 (29)	.978	.186	.736	.817	.636	.509	.186	.680	.622	.455

peoples in India (Kennedy 2000), some of whom could have entered the subcontinent from the northwest during the late Pleistocene epoch. The high variance of R1a1 in India (table 12), the spatial frequency distribution of R1a1 microsatellite variance clines (fig. 4), and expansion time (table 11) support this view.

J2-M172 Clade and Holocene Expansions to Southwestern Asia

One interpretation of the presence of J2a-M410 chromosomes in North Africa and Eurasia is that it reflects the demographic spread of Neolithic farmers. This is consistent with previous interpretations of M172-associated HGs (Semino et al. 2000; King and Underhill 2002). Figure 3 demonstrates the eastward expansion of J2a-M410 to Iraq, Iran, and Central Asia coincident with painted pottery and ceramic figurines, well documented in the Neolithic archeological record (Cauvin 2000). Near the Indus Valley, the Neolithic site of Mehrgarh, estimated to have been founded ~7 KYA (Kenoyer 1998), displays the presence of these types of material culture correlated with the spread J2a-M410 in Pakistan. Although the association of agriculture with J2a-M410 is recognized, the spread of agriculture may not be the only explanation for the spread of this HG. Despite an apparent exogenous frequency spread pattern of HG J2a toward North and Central India from the west (fig. 3), it is premature to attribute the spread to a simplistic demic expansion of early agriculturalists and pastoralists from the Middle East. It reflects the overall net process of spread that may contain numerous as-yet-unrevealed movements embedded within the general pattern. It may also reflect a combination of elements of earlier prehistoric Holocene epi-Paleolithic peoples from the Middle East, subsequent Bronze Age Harappans of uncertain provenance, and succeeding Iron Age Indo-Aryans from Central Asia (Kennedy 2000). Although the overall age of J2a Y-microsatellite variation (table 11) exceeds the appearance of agriculture in the Indus Valley (~6 KYA),

Table 10

Spearman's Rank Correlation Coefficients between Y-Chromosome HG Frequency and Latitude and Longitude

HG	Spearman's Rank Correlation Coefficient	
	Latitude	Longitude
Entire H	−.76[a]	.33
H-M69*	−.37	.37
H1-M52	−.74[a]	.17
H2-APT	−.91[b]	.03
F and H	−.76[a]	.33
F-M89	−.79[a]	.00
O2a-M95	−.44	.77[a]
L1-M76	−.40	−.63
R1a1-M17	.40	−.64
R2-M124	.02	.00

[a] $P < .05$ (two-tailed test).
[b] $P < .01$ (two-tailed test).

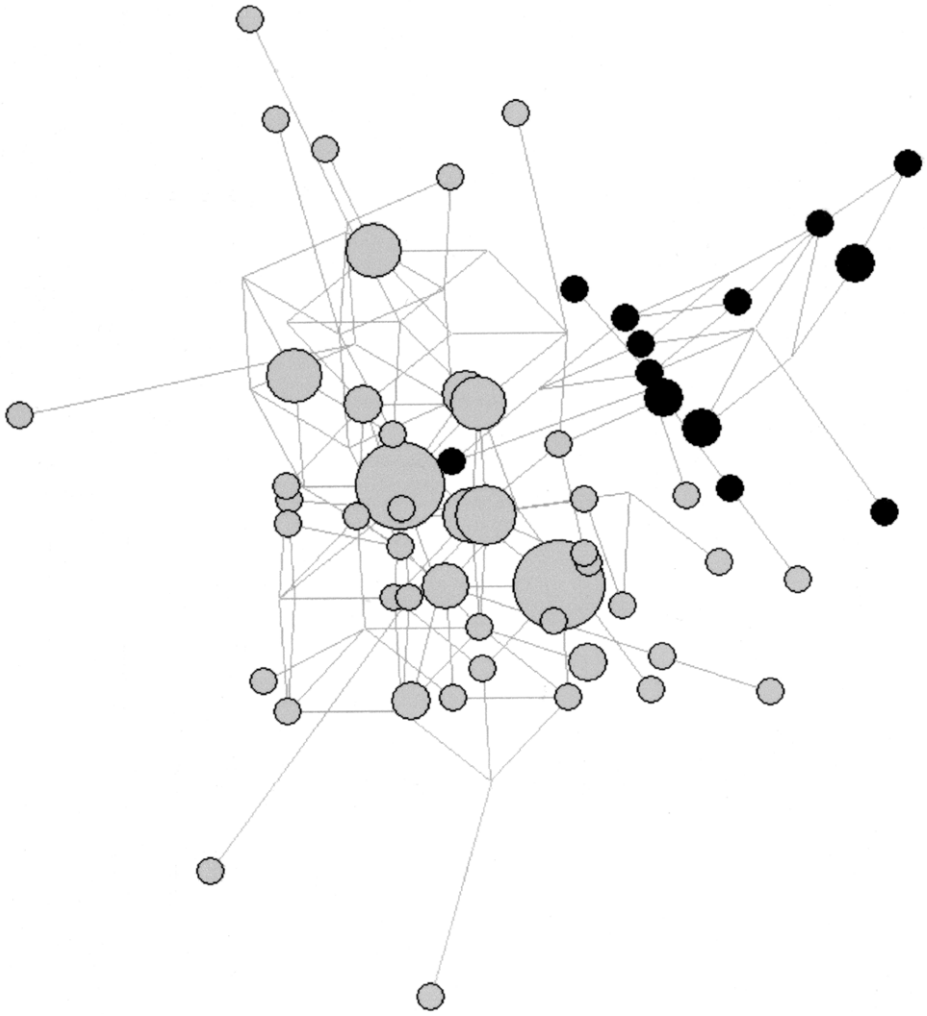


Figure 5 Median-joining network of Y-microsatellite haplotypes with HG O2a-M95 HG among Austro-Asiatic and Dravidian (*gray nodes*) and Tibto-Burman (*black nodes*) populations.

the current lack of informative subdivision within HG J2a in southwestern Asia prevents analysis of such potential layers, which are currently more evident in Anatolia, southeastern Europe, and the Mediterranean. In these regions, HGs J2a1b-M67(xM92) and J2a1b1-M92 have spatial and temporal characteristics consistent with the spread of early farmers and Bronze Age cultures (Di Giacomo et al. 2004). Besides the notable absence of J2a1b-M67(xM92) and J2a1b1-M92 in southwestern Asia, HGs J1-M267 and G-M201 that, respectively, occur at 9% and 10.9% in Turkey (Cinnioğlu et al. 2004), 33.1% and 2.2% in Iraq (Al-Zahery et al. 2003), and 3.4% and 6% in Pakistan are also virtually absent in India (table 5), indicating differential influences from the Middle East in southeastern Europe and southwestern Asia. Similarly, the presence of HG E lineages, thought to possibly be associated with the spread of agriculturalists in southeastern Europe (Hammer et al. 1998; Semino et al. 2004), are absent in India except in specific populations known to have recent African heritage (Thangaraj et al. 1999). Until the paraphyletic J2a-M410* with *DYS413* short-alleles chromosomes are better resolved molecularly in southeastern European, western Asian, and southwestern Asian regions, the magnitude of the contribution of agriculturalists within this HG remains uncertain. The mean variance for J2b2-M241 chromosomes is highest in southwestern Asia (0.33), in contrast with Turkey (0.24) (Cinnioğlu et al. 2004) and the Balkans (<0.2) (Pericic et al. 2005). Further, the mean expansion time of J2b2 in India is 13.8 KYA (table 11), clearly earlier than the appearance of agriculture. Perhaps the three J2b2 chromosomes with eight or more repeats for A7.2 are reflective of more-recent agriculturalists.

Table 11

Age of Microsatellite Variations within Various Y HGs in India

HG	Age[a] ± SD (*n*) of Microsatellite Variation							
				Dravidian		Indo-European		
	Pooled	Austro-Asiatic Tribe	Tibeto-Burman Tribe	Tribe	Caste	Tribe	Caste	Muslim
C5-M356	19.4 ± 4.2 (11)						24.6 ± 8.0 (5)	
F*-M89	28.9 ± 4.4 (38)			26.7 ± 5.7 (25)	28.1 ± 5.3 (8)			
G2-P15	10.5 ± 3.9 (9)				8.6 ± 3.7 (8)			
H1-M52	10.6 ± 1.8 (145)	9.7 ± 4.4 (9)		9.8 ± 2.0 (64)	12.6 ± 3.8 (25)	10.9 ± 3.6 (5)	10.0 ± 2.1 (40)	
H2-Apt	17.8 ± 4.3 (16)			11.3 ± 6.4 (9)				
H-M69*	30.4 ± 7.7 (29)	8.0 ± 4.0 (5)			20.9 ± 9.5 (13)		30.6 ± 12.5 (9)	
J2*-M410/M158	13.7 ± 2.9 (28)				15.8 ± 5.3 (12)		12.4 ± 2.0 (12)	
J2b2-M241	13.8 ± 3.8 (38)	3.6 ± 1.7 (7)			12.1 ± 3.4 (21)		12.0 ± 3.1 (9)	
L1-M76	9.1 ± 1.9 (46)			6.0 ± 2.4 (10)	7.1 ± 2.1 (30)		13.6 ± 4.3 (6)	
O2a-M95[b]	11.7 ± 1.6 (106)	8.8 ± 2.0 (34)	12.9 ± 3.1 (16)	8.2 ± 1.9 (48)		8.5 ± 3.8 (6)		
O3e-M134[c]	9.2 ± 2.7 (58)		9.3 ± 2.7 (57)					
R1a1-M17[d]	14.0 ± 3.1 (114)			10.9 ± 3.8 (5)	12.2 ± 3.0 (34)		14.6 ± 3.5 (57)	10.5 ± 3.0 (11)
R2-M124	11.6 ± 2.1 (68)	9.8 ± 2.4 (7)	10.1 ± 4.3 (5)	18.1 ± 4.6 (9)	8.9 ± 2.3 (13)		12.7 ± 3.5 (28)	8.5 ± 4.8 (6)

[a] In KYA.
[b] In Cambodia, 11.9 ± 3.5 (*n* = 7); in China, 17.9 ± 5.5 (*n* = 14).
[c] In the Tibeto-Burman tribe, excluding Mizo, 6.7 ± 2.1 (*n* = 44); in Mizo, 15.1 ± 7.0 (*n* = 13); in Southeast Asia, 26.2 ± 7.0 (*n* = 27).
[d] In Pakistan, 15.6 ± 3.0 (*n* = 50); in Oman, 12.5 ± 2.9 (*n* = 11); in western Eurasia, 12.8 ± 2.5 (*n* = 16); in Greece, 9.3 ± 2.8 (*n* = 19); in Turkey, 10.0 ± 2.6 (*n* = 36); in Central Asia, 11.2 ± 5.0 (*n* = 10).

A view of the genetics of the spread of language and farming in India that emphasizes mtDNA data (Kivisild et al. 2003*b*) underscores an inherent indigenous and geographically constrained presence of HGs and the challenge of sieving out ancient from relatively more-recent gene flow. Population origins and demic expansions other than those attributable to the recent arrival of immigrants from the northwest have undoubtedly also influenced the genetic population structure of the region but, for the most part, have either been neglected as immaterial or treated as less important. The two least ambiguous HGs most likely to be attributable to having arisen in situ within India, on the basis of phylogeography and accumulated diversification, are F*-M89 and H-M69*; they constitute ~25%–31% of the Y-chromosome census (Kivisild et al. 2003*a*; Cordaux et al. 2004; the present study). Besides the diffusion of the agricultural economy eastward of the Indus Valley during pre-Harappan times (6–7 KYA) and subsequently into the Gangetic plain and the appearance of rice agriculture in northeastern India, the archaeobotanical evidence implicating the potential importance of independent centers of plant domestication within the Indian peninsula by local aboriginal peoples must be acknowledged (Fuller 2003).

Limited Impact of Central Asian Agriculturalists

Besides the two main limitations of the previous competing studies (inadequate samples and inadequate molecular resolution), some studies ignore information intrinsic to the unequivocal binary HG phylogenetic structure. Rather, an approach is applied in which frequencies of lineages are viewed as terminal rather than potentially paraphyletic interior nodes in the phylogeny. The best example is the case of HG L-M20. The potential pitfalls regarding miscomprehension regarding the internal diversification of HG L is reminiscent of that previously discussed concerning the African-versus-Asian origin of the YAP polymorphism that lies at the root of HGs D and E (Underhill and Roseman 2001) and issues relating to the phylogeography of rare deep-rooting basal lineages for HG E that support an African origin (Weale et al. 2003). At the basal M20 level of molecular resolution, the results show presumed equivalent affinity across the geographic expanse of Anatolia, southwestern Asia, and South Asia. However, the failure to recognize subsequent binary and associated microsatellite haplotype diversification patterns—coupled with a lack of expansion or divergence times or their erroneous underestimation—undercuts the apparent recentness of its spread and leads to a convenient but incorrect conclu-

Table 12

Y-Microsatellite Variances Pooled over 10 Loci within HGs R1a1 and R2

Sample	Variance (No.) of Y Microsatellite by HG	
	R1a1-M17	R2-M124
India:		
Tribal	.39 (12)	.34 (21)
Muslim	.22 (11)	.19 (6)
Upper caste	.26 (56)	.24 (20)
Middle caste	.22 (15)	.25 (12)
Lower caste	.36 (20)	.35 (9)
All castes	.28 (91)	.27 (41)
Pakistan	.36 (43)	.33 (13)
Turkey	.25 (36)	.34 (5)

sion. By comprehensively describing the next level of binary molecular resolution within the L clade, it becomes evident from the phylogeography that HG L1-M76 underwent early diversification in South India and subsequently expanded toward peripheral regions. The expansion time for L1-M76 spans at least the early Holocene period, well before the Neolithic. The phylogeography and the similarity of microsatellite variation of HGs R1a1 and R2 to L1-M76 in South Asian tribes argues that they likely share a common demographic history.

Expansion of HGs R1a1-M17 and R2-M124

The phylogeography of the HG R*-M207 spans Europe, the Caucasus, West Asia, Central Asia, and South Asia; therefore, the hypothesis that there is an HG R*-M207 expansion locus central to all these regions is both plausible and parsimonious. This is consistent with our observation that HG R*-M207 is observed at a maximum of 3.4% frequency in Baluchistan and Punjab regions, whereas, in inner India, it is 0.3%. HG R1a1 displays both high frequencies and widespread geography, ranging from India to Norway (Quintana-Murci et al. 2001; Passarino et al. 2002), but is rare in East Asia (Su et al. 1999). The distribution of HG R2-M124 is more circumscribed relative to R1a1, but it has been observed at informative levels in Central Asia, Turkey, Pakistan, and India. The distribution of R1a1 and R2 within India is similar, as are the levels of associated microsatellite variance (table 12). The ages of the Y-microsatellite variation (table 11) for R1a1 and R2 in India suggest that the prehistoric context of these HGs will likely be complex. A principal-components plot of R1a1-M17 Y-microsatellite data (fig. 6) shows several interesting features: (*a*) one tight population cluster comprising southern Pakistan, Turkey, Greece, Oman, and West Europe; (*b*) one loose cluster comprising all the Indian tribal and caste populations, with the tribal populations occupying an edge of this cluster; and (*c*) Central Asia and Turkey occupy intermediate positions. The divergence time between the two clusters was 8–12 KYA. The pattern of clustering does not support the model that the primary source of the R1a1-M17 chromosomes in India was Central Asia or the Indus Valley via Indo-European speakers. Further, the relative position of the Indian tribals (fig. 6), the high microsatellite variance among them (table 12), the estimated age (14 KYA) of microsatellite variation within R1a1 (table 11), and the variance peak in western Eurasia (fig. 4) are entirely inconsistent with a model of recent gene flow from castes to tribes and a large genetic impact of the Indo-Europeans on the autochthonous gene pool of India. Instead, our overall inference is that an early Holocene expansion in northwestern India (including the Indus Valley) contributed

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics*.

Figure 6 Plot of R1a1-M17–derived chromosomes against values for the first two principal components for 10 microsatellite loci. Indian population codes are as described in table 1. The Central Asian data include 10 chromosomes described by Underhill et al. (2000). The Turkish data are from Cinnioğlu et al. (2004). The Greek data (R.K., unpublished data) include 19 chromosomes.

R1a1-M17 chromosomes both to the Central Asian and South Asian tribes prior to the arrival of the Indo-Europeans. The results of our more comprehensive study of Y-chromosome diversity are in agreement with the caveat of Quintana-Murci et al. (2001, p. 541), that "more complex explanations are possible," rather than their simplistic conclusion that HGs J and R1a1 reflect demic expansions of southwestern Asian Dravidian-speaking farmers and Central Asian Indo-European–speaking pastorialists.

On the basis of frequency differences between castes and southern non–Indo-European–speaking tribes, Cordaux et al. (2004) concluded that the R HGs reflect the impact of Indo-European pastoralists from Central Asia, thus linking HG frequency to specific historical events. Although any recent immigration from Central Asia would have undoubtedly contributed some R HGs to the pre-existing gene pool (together with other lineages frequent in Central Asia, such as C3 and O sub groups), other potential events—such as range expansions of Ice Age hunter-gatherers into peninsular India from other source regions, not necessarily far from the mountains extending from Baluchistan to Hindu Kush, on both sides of which the R1a frequency is currently the highest—could have also contributed significantly to the observed distributions, both in India and in Central Asia (Kennedy 2000). In other words, there is no evidence whatsoever to conclude that Central Asia has been necessarily the recent donor and not the receptor of the R1a lineages. The current absence of additional informative binary subdivision within this HG obfuscates potential different histories hidden within this HG, making such interpretations as the sole and recent source area overly simplistic. The same can be said in respect to HG R2-M124.

C5-M356 Is an Ancient HG That Originated in India

The phylogeography, frequency (tables 5 and 6), and age of Y-microsatellite variation (table 11) of HG C5-M356 lineages are indicative of an in situ Indian origin and considerable antiquity. HG C(xC3) lineages occur at considerable frequency in Australia (Kayser et al. 2000). Some have interpreted that the paraphyletic HG C(xC3) reflects the Paleolithic colonization event con-

cerning the peopling of Australia via a southern coastal migration route involving India (Kivisild et al. 1999, 2003*a*; Underhill et al. 2001*a*), whereas others (Redd et al. 2002) argued that such C*-defined chromosomes reflects recent (≤5 KYA) genetic affinity between Indian tribes and Australian Aborigines. Interestingly, the Australian Aborigines with the C(xC3) HG display a large multistep deletion at the *DYS390* microsatellite locus (Forster et al. 1998) that is considered to be a unique mutational event (Kayser et al. 2000), whereas the Indians do not. These Australian *DYS390.1*-deletion HG C chromosomes lack the M356 mutation (M. Kayser, personal communication), which undermines claims of a recent common shared ancestry with India. Conversely, the C5 data support a more-ancient affinity that is consistent with that observed with respect to mtDNA polymorphisms (Macaulay et al. 2005).

Y-Chromosome Substructure and the Geographic Origins of Dravidian Speakers

The impact that Neolithic pastoralists had on the gene pool remains an outstanding topic, as is the origin of the Dravidian language. However, despite the potential consequences of population stratification and language shift, the irregular distributions of Y-chromosome HGs among the various social and linguistic groups provide important insights (table 6). HG L1-M76 has phylogeographic hallmarks consistent with an indigenous-origin model and an age of Y-microsatellite variation consistent with the early Holocene (~9 KYA). Fuller (2003), using archeobotanical evidence, suggests that Dravidian originated in India, in contrast to western Asia (McAlpin 1974, 1981). The Dravidian-speaking castes essentially display similar levels of HGs R1a1, R2-M124, and L1-M76. On the basis of linguistic and religious evidence, if pastoralists arrived recently on a track from the north via Bactria, southern Tajikistan, northern Afghanistan, and the Hindu Kush into the northern Pakistan plains (Witzel 2004), one would expect to see L3-M357 in India. Although this HG occurs with an intermediate frequency in Pakistan (6.8%), it is very rare in India (0.4%). Conversely, L1-M76 occurs at a frequency of 7.5% in India and 5.1% in Pakistan. Lastly, the L1-M76 mean microsatellite variance is higher in India (0.35) than in Pakistan (0.19).

Further, the distribution of J2a-M410 and J2b-M12 extends from Europe to India, and the virtual absence of HGs J1, E, and G in India indicates a complex scenario of movements not all associated with a single Neolithic and Indo-European spread from a common origin. Recently, mtDNA evidence was brought to bear on the model of an external northwest Indo-Iranian borderland as the Elamo-Dravidian source (Quintana-Murci et al. 2004). The result was cautiously interpreted as consistent with the proto-Elamite hypothesis. However, the possibility that the observed mtDNA HG frequency differences could also reflect the relocation of a more-ancient Indian Dravidian-speaking population that expanded toward the Indus and subsequently experienced gene flow from southwestern Asian sources could not be excluded (Quintana-Murci et al. 2004). The HG F*-M89 and H-M69* data are not in agreement with an exogenous-origin model, but the presence of these HGs in all subgroup classifications does not unequivocally support the indigenous model of Dravidian origins either. However, HG L1-M76, because it is clearly predominant in Dravidian speakers, corresponds most closely with the indigenous model. Further, the microsatellite variance within L1 is greater in southern India than in the Indus region (table 9). The spatial distributions of both L1 HG frequency and associated microsatellite variance (fig. 4) show a pattern of spread emanating from southern India. HG J2a-M410 is confined to upper-caste Dravidian and Indo-European speakers, with little occurrence in the middle and lower castes. This absence of even modest admixture of J2a in southern Indian tribes and middle and lower castes is inconsistent with the L1 data. Overall, therefore, our data provide overwhelming support for an Indian origin of Dravidian speakers.

Acknowledgments

We thank all the men who donated DNA used in this study. This study was supported by grants from the Department of Biotechnology (DBT), Government of India, and the Indian Statistical Institute (to P.P.M.); National Institutes of Health grant GM28428 (to L.L.C.-S.); and Russian Foundation for Basic Research grant 04-04-48639 (to L.A.Z.). We thank P. J. Oefner for calibrating *DYS413*-allele–repeats controls using Mass Spectrophotometric analyses. We thank DBT for providing permission for S.S. to genotype a fraction of the Indian samples in the Stanford laboratory, for purposes of protocol standardization.

Web Resources

The URLs for data presented herein are as follows:

Arlequin, http://anthro.unige.ch/arlequin/ (for version 2.000)
Fluxus Engineering, http://www.fluxus-technology.com/sharenet.htm (for Network, version 4.1.0.9)
Microsat, http://hpgl.stanford.edu/projects/microsat/ (for version 1.5)

References

Al-Zahery N, Semino O, Benuzzi G, Magri C, Passarino G, Torroni A, Santachiara-Benerecetti AS (2003) Y-chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. Mol Phylogenet Evol 28:458–472

Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BVR, Reddy PG, Rasanayagam A, Papiha SS,

Villems R, Redd AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB (2001) Genetic evidence on the origins of Indian caste populations. Genome Res 11:994–1004

Bamshad MJ, Watkins WS, Dixon ME, Jorde LB, Rao BB, Naidu JM, Prasad BV, Rasanayagam A, Hammer MF (1998) Female gene flow stratifies Hindu castes. Nature 395:651–652

Barbujani G (2000) Geographic patterns: how to identify them and why. Hum Biol 72:133–153

Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, Dey B, Roy M, Roy B, Bhattacharyya NP, Roychoudhury S, Majumder PP (2003) Ethnic India: a genomic view, with special reference to peopling and structure. Genome Res 13:2277–2290

Beekes RSP (1995) Comparative Indo-European linguistics: an introduction. J Benjamins, Amsterdam/Philadelphia

Blazek V, Boisson C (1992) The diffusion of agricultural terms from Mesopotamia. Archiv Orientalni 60:16–23

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, et al (2002) A human genome diversity cell line panel. Science 296: 261–262

Cauvin J (2000) The birth of the gods and the origins of agriculture. Cambridge University Press, Cambridge, United Kingdom

Cinnioğlu C, King R, Kivisild T, Kalfoglu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K, Oefner PJ, Shen P, Semino O, Cavalli-Sforza LL, Underhill PA (2004) Excavating Y-chromosome haplotype strata in Anatolia. Hum Genet 114:127–148

Cordaux R, Aunger R, Bentley G, Nasidze I, Sirajuddin SM, Stoneking M (2004) Independent origins of Indian caste and tribal paternal lineages. Curr Biol 14:231–235

Delfiner P (1976) Linear estimation of non-stationary spatial phenomena. In: Guarasio M, David M, Haijbegts C (Ed.) Advanced geostatistics in the mining industry. Dordrecht, Reidel, Austria, pp 49–68

Di Giacomo F, Luca F, Popa LO, Akar N, Anagnou N, Banyko J, Brdicka R, et al (2004) Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. Hum Genet 115:357–371

Edmonds C, Lillie A, Cavalli-Sforza L (2004) Mutations arising in the wave front of an expanding population. Proc Natl Acad Sci USA 101:975–979

Forster P, Kayser M, Meyer E, Roewer L, Pfeiffer H, Benkmann H, Brinkmann B (1998) Phylogenetic resolution of complex mutational features at Y-STR DYS390 in aboriginal Australians and Papuans. Mol Biol Evol 15:1108–1114

Fuller D (2003) An agricultural perspective on Dravidian historical linguistics: archaeological crop packages, livestock and Dravidian crop vocabulary. In: Bellwood P, Renfrew C (eds) Examining the farming/language dispersal hypothesis. McDonald Institute for Archaeological Research, Cambridge, United Kingdom, pp 191–213

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, Templeton AR, Zegura SL (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol 15:427–441

Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, Zegura SL (2001) Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol 18:1189–1203

Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4:598–612

Kayser M, Brauer S, Weiss G, Underhill PA, Roewer L, Schiefenhovel W, Stoneking M (2000) Melanesian origin of Polynesian Y chromosomes. Curr Biol 10:1237–1246

Kennedy KAR (2000) God-apes and fossil men: palaeoanthropology of South Asia. University of Michigan Press, Ann Arbor, pp 165–171

Kenoyer J (1998) Ancient cities of the Indus Valley civilization. Oxford University Press, Oxford, United Kingdom

King R, Underhill PA (2002) Congruent distribution of Neolithic painted pottery and ceramic figurines with Y-chromosome lineages. Antiquity 76:707–714

Kivisild T, Bamshad MJ, Kaldma K, Metspalu M, Metspalu E, Reidla M, Laos S, Parik J, Watkins WS, Dixon ME, Papiha SS, Mastana SS, Mir MR, Ferak V, Villems R (1999) Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. Curr Biol 9:1331–1334

Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, Parik J, Metspalu E, Adojaan M, Tolk H-V, Stepanov V, Gölge M, Usanga E, Papiha SS, Cinnioğlu C, King R, Cavalli-Sforza L, Underhill PA, Villems R (2003*a*) The genetic heritage of earliest settlers persist in both the Indian tribal and caste populations. Am J Hum Genet 72: 313–332

Kivisild T, Rootsi S, Metspalu M, Metspalu E, Parik J, Kaldma K, Usanga E, Mastana S, Papiha SS, Villems R (2003*b*) The genetics of language and farming spread in India. In: Bellwood P, Renfrew C (eds) Examining the farming/language dispersal hypothesis. McDonald Institute for Archaeological Research, Cambridge, United Kingdom, pp 215–222

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt HJ, Oppenheimer S, Torroni A, Richards M (2005) Single rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308:1034–1036

Majumder PP (2001) Indian caste origins: genomic insights and future outlook. Genome Res 11:931–932

Malaspina P, Ciminelli BM, Viggiano L, Jodice C, Cruciani F, Santolamazza P, Sellitto D, Scozzari R, Terrenato L, Rocchi M, Novelletto A (1997) Characterization of a small family (CAIII) of microsatellite-containing sequences with X-Y homology. J Mol Evol 44:652–659

McAlpin DW (1974) Toward proto-Elamo-Dravidian. Language 50: 89–101

——— (1981) Proto-Elamo-Dravidian: the evidence and its implications. Trans Am Phil Soc 71:3–155

Passarino G, Cavalleri GL, Lin AA, Cavalli-Sforza LL, Børresen-Dale A-L, Underhill PA (2002) Different genetic components in the Norwegian population revealed by the analysis of mtDNA and Y chromosome polymorphisms. Eur J Hum Genet 10:521–529

Pericic M, Lauc LB, Klaric IM, Rootsi S, Janicijevic B, Rudan, I, Terzic R, Colak I, Kvesic A, Popovic D, Sijacki A, Behluli I, Dordevic D, Efemovska L, Bajec DD, Stefanovic BD, Villems R, Rudan P (2005) High-resolution phylogenetic analysis of southeastern Europe (SEE) traces major episodes of paternal gene flow among Slavic populations. Mol Biol Evol 22:1964–197522:1964–1975

Qamar R, Ayub Q, Mohyuddin A, Helgason A, Mazhar K, Mansoor A, Zerjal T, Tyler-Smith C, Mehdi SQ (2002) Y-chromosomal DNA variation in Pakistan. Am J Hum Genet 70:1107–1124

Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, Scozzari R, Rengo C, Al-Zahery N, Semino O, Santachiara-Benerecetti AS, Coppa A, Ayub Q, Mohyuddin A, Tyler-Smith C, Qasim Mehdi S, Torroni A, McElreavey K (2004) Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. Am J Hum Genet 74:827–845

Quintana-Murci L, Krausz C, Zerjal T, Sayar SH, Hammer MF, Mehdi SQ, Ayub Q, Qamar R, Mohyuddin A, Radhakrishna U, Jobling MA, Tyler-Smith C, McElreavey K (2001) Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. Am J Hum Genet 68:537–542

Ramana GV, Su B, Jin L, Singh L, Wang N, Underhill P, Chakraborty R (2001) Y-chromosome SNP haplotypes suggest evidence of gene flow among caste, tribe, and the migrant Siddi populations of Andhra Pradesh, South India. Eur J Hum Genet 9:695–700

Redd A, Roberts-Thomson J, Karafet T, Bamshad M, Jorde L, Naidu J, Walsh B, Hammer MF (2002) Gene flow from the Indian sub-

continent to Australia: evidence from the Y chromosome. Curr Biol 12:673–677

Renfrew C (1996) Language families and the spread of farming. In: Harris DR (ed) The origins and spread of agriculture and pastoralism in Eurasia. Smithsonian Institution Press, Washington, DC, pp 70–92

Roychoudhury S, Roy S, Basu A, Banerjee R, Vishwanathan H, Usha Rani MV, Sil SK, Mitra M, Majumder PP (2001) Genomic structures and population histories of linguistically distinct tribal groups of India. Hum Genet 109:339–350

Sandwell DT (1987) Biharmonic spline interpolation of GEOS-3 and SEASAT altimeter data. Geophys Res Letters 2:139–142

Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, Triantaphyllidis C, Shen P, Oefner PJ, Zhivotovsky LA, King R, Torroni A, Cavalli-Sforza LL, Underhill PA, Santachiara-Benerecetti AS (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the Neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet 74:1023–1034

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiæ M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti A, Cavalli-Sforza LL, Underhill PA (2000) The genetic legacy of Palaeolithic *Homo sapiens sapiens* in extant Europeans: a Y-chromosome perspective. Science 290:1155–1159

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 in East Asia during the last Ice Age. Am J Hum Genet 65: 1718–1724

Tambets K, Tolk H-V, Kivisild T, Metspalu E, Parik J, Reidla M, Voevoda M, Damba L, Bermisheva M, Khusnutdinova E, Golubenko M, Stepanov V, Puzyrev V, Usanga E, Rudan P, Beckmann, L, Villems R (2003) Complex signals for population expansions in Europe and beyond. In: Bellwood P, Renfrew C (eds) Examining the farming/language dispersal hypothesis. McDonald Institute for Archaeological Research, Cambridge, United Kingdom, pp 449–457

Thangaraj K, Ramana GV, Singh L (1999) Y-chromosome and mitochondrial DNA polymorphisms in Indian populations. Electrophoresis 20:1743–1747

Underhill PA, Passarino G, Lin AA, Marzuki S, Cavalli-Sforza LL, Chambers G (2001*a*) Maori origins, Y chromosome haplotypes and implications for human history in the Pacific. Hum Mutat 17:271–280

Underhill PA, Passarino G, Lin AA, Shen P, Foley RA, Mirazón Lahr M, Oefner PJ, Cavalli-Sforza LL (2001*b*) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet 65:43–62

Underhill PA, Roseman CC (2001) The case for an African rather than an Asian origin of the human Y-chromosome YAP insertion. In: Jin L, Seielstad M, Xiao C (eds) Recent advances in human biology. Vol. 8: Genetic, linguistic and archaeological perspectives on human diversity in Southeast Asia. World Scientific, New Jersey, pp 43–56

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 R S, 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

Weale ME, Shah T, Jones AL, Greenhalgh J, Wilson JF, Nymadawa P, Zeitlin D, Connell BA, Bradman N, Toomas MG (2003) Rare deep-rooting Y chromosome lineages in humans: lessons for phylogeography. Genetics 165:229–234

Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, Jin L, Su B, et al (2001) The Eurasian heartland: a continental perspective on Y-chromosome diversity. Proc Natl Acad Sci USA 98:10244–10249

Witzel M (2004) Central Asian roots and acculturation in South Asia: linguistic and archaeological evidence from western central Asia, the Hindukush and northwestern south Asia for early Indo-Aryan language and religion. In: Osada T (ed) Linguistics, archaeology and the human past. Indus Project, Research Institute for Humanity and Nature, Kyoto, pp 87–211

Wooding S, Ostler C, Prasad BVR, Watkins WS, Sung S, Bamshad M, Jorde LB (2004) Directional migration in the Hindu castes: inferences from mitochondrial, autosomal and Y-chromosomal data. Hum Genet 115:221–229

Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C (2002) A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia. Am J Hum Genet 71:466–482

Zhivotovsky LA (2001) Estimating divergence time with use of microsatellite genetic distances: impacts of population growth and gene flow. Mol Biol Evol 18:700–709

Zhivotovsky LA, Underhill PA (2005) On the evolutionary mutation rate at Y-chromosome STRs: comments on paper by Di Giacomo et al (2004). Hum Genet 116:529–532

Zhivotovsky LA, Underhill PA, Cinnioğlu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G, Chambers GK, Herrera RJ, Yong KK, Gresham D, Tournev I, Feldman MW, Kalaydjieva L (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet 74:50–61