Chapter 17

The Genetics of Language and Farming Spread in India

Toomas Kivisild, Siiri Rootsi, Mait Metspalu, Ene Metspalu, Juri Parik, Katrin Kaldma, Esien Usanga, Sarabjit Mastana, Surinder S. Papiha & Richard Villems

Most maternal lineages of present-day Indians derive from a common ancestor in mtDNA haplogroup M that split into Indian, eastern Asian, Papuan, and Australian subsets 40,000–60,000 mtDNA-years ago. The second major component in Indian maternal heredity lines traces back to the split of haplogroup U into Indian, western Eurasian and northern African variants approximately at the same time. The variation in these two ancient Indian-specific sets of lineages is the main modifier in the heterogeneity landscape of Indian populations, defining the genetic differences between caste groups and geographic regions in the sub-continent. The difference between regional caste groups is accentuated furthermore by the presence of a northwest to south decline of a minor package of lineages of western Asian or European origin.

In contrast, the majority of Indian paternal lineages do not share recent ancestors with eastern Asian populations but stem from haplogroups common to (eastern) European or western Asian populations. This finding has recently been interpreted in favour of the classical Indo-Aryan invasion hypothesis. Here, we show that this interpretation is probably caused by a phylogeographically-limited view of the Indian Y-chromosome pool, amplified because of current inconsistencies in the interpretation of the temporal scale of the variability in the non-recombining part of the Y chromosome (NRY).It appears to us that the high variability of STRs in the background of NRY variants in India is consistent with the view of largely autochthonous pre-Holocene genetic diversification — a conclusion reached earlier for the Indian maternal lineages (Kivisild *et al.* 1999a).

While interpreting the genetic aspects of farming/language dispersal in the Indian context, it is easy to get lost in its 'multitude of endogamous pockets' (Cavalli-Sforza *et al.* 1994). Yet a forest can hopefully be seen behind the trees, provided that the conclusions to be drawn derive from a phylogeographically representative analysis of the people of the sub-continent.Perhaps new ideas, analogous to the recently introduced 'SPIWA' model for Europe (see Renfrew this volume), are needed when developing new farmingl/anguage dispersal models for India.

The earliest 'agricultural package' in the Indian subcontinent — a combined presence of wheat, barley, cattle, sheep and goat domestication — is found in Mehrgarh, Baluchistan, and dates to about 9000 years before present (BP). It spread first into an area extending from the Punjab in the northwest to Uttar Pradesh in the east and to Gujarat in the south. It took another 4000 years before it eventually reached southern Peninsular India (Chakrabarti 1999). In this northwestern early agricultural region lie the roots of the Indus Civilization, and any later cultural influence or human migration from the northwest or west had to pass through this area in order to reach the rest of India.

Neolithic communities in India did not start on empty ground. Cultural complexes belonging to a comparatively short Mesolithic episode developed from the preceding Middle and Upper Palaeolithic cultures and continued to exist through the Neolithic, Bronze and Iron Ages, with microlithic tools continuing in use here and there in some communities even today. The advent of agriculture in India, although largely reflecting local developments, is to be understood against the background of agricultural growth in its geographic neighbourhood, encompassing the Iranian plains and the Fertile Crescent in the west, and Southeast Asia — as far as rice is concerned — in the east (Chakrabarti 1999).

Three quarters of the Indian population today speak Indo-European (IE) languages. Next, in terms of the number of speakers, is the Dravidian lan-

guage family, spread now mostly over the southern parts of the Indian peninsula, with Telugu, Tamil, Kannada and Malayalam being the most important languages spoken today. Besides these two major groups, Austroasiatic and Tibeto-Burman languages are spoken in the central and eastern parts of India. The introduction of all these four major language families to India is thought to be related to relatively recent immigration episodes.

IE languages in India are commonly thought to originate from either the invasion of Indo-Aryan tribes during the post-Harappan period, or possibly from the spread of Neolithic populations (Renfrew 1987). Supposedly, Dravidian too had its closest linguistic relatives in western Asia (ancient Elamite?) (Ruhlen 1991) and was brought into India before the IE languages, together with or before the spread of farming. Alternatively, the Dravidian languages may turn out to be native to South India, as argued by Fuller (this volume). Neolithic origins, albeit not from the northwest but from the northeast, are claimed also for the spreads of the Austroasiatic and Tibeto-Burman languages.

Thus, according to these linguistic hypotheses, the ancestors of close to 100 per cent of the indigenous languages spoken in India today came to India during the Holocene. consequently, all the preceding pre-Neolithic languages were totally replaced. If this is indeed so, how extensive was the genetic replacement caused by these events?

Linguistic affiliations, although suggestive of some level of gene flow (as argued by Bellwood this volume), do not always or necessarily correlate with the genetic affinities of the populations. A well-known example of language change by elite dominance is the represented by the Turkish population in Anatolia, which clusters in genetic analyses with populations from the Middle East and Europe, rather than with the linguistically-related Altai populations. Analyses of mtDNA and Y chromosomes reveal that Turks share only about 5–10 per cent of their maternal and paternal lineages with their linguistic relatives in Altai, while the rest of their lineages belong to western Eurasian lineage families (Rolf *et al.* 1999; Tambets *et al.* 2000).

Mitochondrial DNA continuity of Indian populations: identifying and quantifying ancient and recent gene flow

In India, where palaeoanthropologica1 findings from the Middle and Upper Palaeolithic are very rare and no ancient DNA study has proven successful (Kumar *et al.* 2000), evidence for the beginning of the settlement of modern humans comes from mtDNA and Y-chromosomal studies (Mountain *et al.* 1995; Passarino *et al.* 1996; Kivisild *et al.* 1999a; Quintana-Murci *et al.* 1999; Underhill *et al.* 2000; 2001; Bamshad *et al.* 2001).

Haplogroup M is the most frequent mtDNA cluster in present-day Indian populations and, because it is nearly absent in western Eurasia, it stands out as a separate cluster from the African haplogroup L3. It has been suggested that M represents the earliest wave of the migration of anatomically modern humans (AMH) out of Africa (Kivisild *et al.* 1999a; 2000; Quintana-Murci *et al.* 1999), following the suggested earlier 'southern route' (Cavalli-Sforza *et al.* 1994; Lahr & Foley 1994). The Indian haplogroup M lineages differ substantially from those found in eastern and central Asian populations and most likely represent *in situ* diversification in the sub-continent since the Palaeolithic (Kivisild *et al.* 1999b; Bamshad *et al.* 2001). It is important to note that this ancient cluster is present at frequencies above 50 per cent all over India. Its sub-clusters, as with the Indian mtDNA tree in general (Mountain *et al.* 1995), are not subdivided according to linguistic (Indo-European, Dravidian) or caste affiliations (Kivisild *et al.* 2000; Bamshad *et al.* 2001), although there may occur (sometimes drastic) population-wise differences in frequencies of particular sub-clusters.

Another profound peculiarity of the Indian mtDNA pool lies in the high frequency (~14 per cent on average in India) and great diversity of narrowly Indian-specific lineages deriving directly from the phylogenetic node R, otherwise ancestral to HV, JT, and U found in western Eurasia, and B and F in eastern Eurasia (Kivisild *et al.* 1999b). The coalescence age of this node is similar to that for haplogroup M and its presence in India suggests, once again, *in situ* differentiation of maternal lineages since the Upper Palaeolithic.

Furthermore, two sub-clusters of western Eurasian specific haplogroup U, these being U2 and U7, occur in India in relatively high frequencies. Because of their deep coalescence times, their presence was interpreted as testifying another Palaeolithic human migration to the Indian subcontinent from the west (Kivisild *et al.* 1999a). However, patterns of the spread of U2 and U7 differ in an interesting way. The lineages of the first group are restricted mainly to the Indian subcontinent and form a set of an Indian-specific sub-cluster U2i (Kivisild *et al.* 1999a). Although U2i is spread in a decreasing frequency from northwest to south and east in India, its presence

in nearby western Asian populations is marginal (< 1 per cent).

Quite a different phylogeographic picture emerges for U7 lineages. Its world-wide frequency is the highest among Iranians (we lack reliable data on Afghani populations) and U7 is also common among the Arabic-speaking western Asian populations (Table 17.1). On the one hand, its frequency in India is about four to five times lower than in Iran. On the other hand, considering the frequency of U7 among the subset of the western Eurasian lineages present in the Indian mtDNA pool, i.e. excluding from comparisons Indian-specific lineages of M, R, and U2i clusters, U7 appears over twice as frequently in Indians as in Iranians (Table 17.1). Most importantly however, we have observed that haplotype sharing between the Indian and western Asian U7 lineages occurs only through common founder motifs. On average, Indian U7 HVS-I sequences differ from the common founder motifs by one transition (~20,000 years), suggesting their split around the time of the Last Glacial Maximum (LGM). Hence, one may speculate that global cooling and the accompanying extensive spread of deserts separated mtDNA haplogroup U7 carriers into two branches — the western and the eastern. There is an analogy with U2 (U2e and U2i), except that this split occurred about twice as early. An analogy can be seen also in the spread and diversity of haplogroup W lineages (Kivisild *et al.* 1999b).

Thus, what we see as specific to Indian subcontinent is the presence of diverse sub-clusters of haplogroups M, R, and U that are virtually absent elsewhere. All these sub-clusters show coalescent times at around 50,000 BP. Given their high overall frequency in India this suggests a very limited gene flow — at least as far as maternal lineages are concerned — beyond the subcontinent over a long time span, likely since its initial colonization.

The Holocene

To focus on lineages that might be tentatively associated with the spread of Neolithic and Bronze Age cultures, Dravidian and Indo-European languages in West and South Asia, we ignore the lineage clusters that reveal clear signs of an 'early' (~15,000 years or longer) differentiation between Indians and western Asians. Our task becomes less sophisticated than it was in separating 'autochthonous' European mtDNA lineages from those present in the Levant (Richards *et al.* 2000).

First, it would be interesting to compare whether 'recent western' lineages among the present-day Indians correlate in their founder frequencies with those postulated to have been imported to Europe during and since the Neolithic. As shown in Table 17.1, a fraction of about 8 per cent of the Indian maternal gene pool can be assigned to a status of 'a putative recent import from the west' (PRIFW) — i.e. in time scale throughout the Neolithic, Bronze and Iron Ages to more recent times. In space, it begins with a possible impact from Iran and Afghanistan, extending to potential migrations from further west and north.

Of course, here we should mention a possibility that some of the lineage clusters we have marked above as 'pre-Holocene', such as U7 and W, could have contributed to more recent gene flow(s) into India as well. Yet any possible difference such a scenario would make is likely to be rather small because of the comparatively low overall frequency of such lineages in the extant Indian mtDNA pool.

It would be extremely difficult if not impossible to discern Neolithic and Bronze Age migrations to India apart from each other, especially if they originated from the same source population/geographically restricted pool of mtDNAs. Compared to the composition of the mtDNA pool of Europeans, the PRIFW component in Indians differs in higher frequencies of HV and U1 and lower frequencies of H and U5 — a pattern similar to that observed for Anatolians and Iranians (Table 17.1). Nevertheless, some differences between Indians and western Asians can be noted, like the absence of U3 and the significantly lower frequency of K ($p < 0.05$) in the former, while the Indian share of PRIFW lineages comprises haplogroups I and U4. These are frequent also in Eastern Europe and present in Central Asian populations at higher frequencies than in Iran, Anatolia and in Arabians (Table 17.1). A notable difference lies also in the pre-HV lineages, characterized by 16217C, which reveal similar diversity and patterns of spread as haplogroup U7 lineages in India and western Asia (Table 17.1).

Currently, the western Asian Neolithic component assigned to the present-day mtDNA pool of Europeans is thought to consist mostly of haplogroups J, U3 and T1. It has been suggested that female carriers of these lineage clusters migrated from the Near East to Europe, probably at the same time and possibly as a consequence of the spread of farming (Richards *et al.* 2000). In this context, it is specifically interesting to note that, like U3 and J, T1 lineages are also found in a comparatively low frequency in Indians (compare a T1 to T ratio of 2/14 in India versus 15/38 in Iran). It is possible, then, that Iranians ob-

Table 17.1. *MtDNA haplogroup composition in Indians compared to western Asian populations.*

	India				Iran			Anatolia			Arabia		
	n	All	Fr.l	Fr.2	*n*	All	Fr.2	*n*	All	Fr.2	*n*	All	Fr.2
A-G, M, N9	862	66.3%	–	–	28	6.2%	–	20	5.2%	–	30	7.7%	–
I	8	0.6%	5.1%	7.7%	9	2.0%	2.6%	9	2.3%	2.7%	3	0.8%	1.0%
N1a	1	0.1%	0.6%	1.0%	2	0.4%	0.6%	5	1.3%	1.5%	1	0.3%	0.3%
N1b	0	0.0%	0.0%	0.0%	2	0.4%	0.6%	3	0.8%	0.9%	11	2.8%	3.7%
W	19	1.5%	12.2%	–	9	2.0%	–	15	3.9%	–	7	1.8%	–
X	2	0.2%	1.3%	1.9%	13	2.9%	3.8%	17	4.4%	5.0%	7	1.8%	2.4%
N*	3	0.2%	1.9%	2.9%	10	2.2%	2.9%	0	0.0%	0.0%	8	2.1%	2.7%
pre-HV[1]	1	0.1%	0.6%	1.0%	6	1.3%	1.8%	9	2.3%	2.7%	58	14.9%	19.7%
pre-HV[2]	5	0.4%	3.2%	4.8%	5	1.1%	1.5%	2	0.5%	0.6%	1	0.3%	0.3%
HV	8	0.6%	5.1%	7.7%	25	5.5%	7.4%	14	3.6%	4.1%	14	3.6%	4.8%
H	31	2.4%	19.9%	29.8%	77	17.1%	22.6%	97	25.0%	28.7%	50	12.9%	17.0%
(pre-)V	0	0.0%	0.0%	0.0%	3	0.7%	0.9%	1	0.3%	0.3%	1	0.3%	0.3%
J	10	0.8%	6.4%	9.6%	61	13.5%	17.9%	42	10.8%	12.4%	81	20.8%	27.6%
T	14	1.1%	9.0%	13.5%	38	8.4%	11.2%	46	11.9%	13.6%	18	4.6%	6.1%
Ul	7	0.5%	4.5%	6.7%	12	2.7%	3.5%	17	4.4%	5.0%	6	1.5%	2.0%
U2I	101	7.8%	–	–	2	0.4%	–	1	0.3%	–	5	1.3%	–
U2e	2	0.2%	1.3%	1.9%	5	1.1%	1.5%	3	0.8%	0.9%	2	0.5%	0.7%
U3	0	0.0%	0.0%	0.0%,	12	2.7%	3.5%	21	5.4%	6.2%	5	1.3%	1.7%
U4	6	0.5%	3.8%	5.8%	5	1.1%	1.5%	4	1.0%	1.2%	2	0.5%	0.7%
U5	7	0.5%	4.5%	6.7%	15	3.3%	4.4%	21	5.4%	6.2%	2	0.5%	0.7%
U6	0	0.0%	0.0%	0.0%	1	0.2%	0.3%	0	0.0%	0.0%	4	1.0%	1.4%
U7	33	2.5%	21.2%	–	40	8.9%	–	6	1.5%	–	9	2.3%	–
K	2	0.2%	1.3%	1.9%	34	7.5%	10.0%	25	6.4%	7.4%	14	3.6%	4.8%
U*	0	0.0%	0.0%	0.0%	5	1.1%	1.5%	2	0.5%	0.6%	6	1.5%	2.0%
R*	178	13.7%	–	–	22	4.9%	–	7	1.8%	–	3	0.8%	–
L1-L3	0	0.0%	–	–	10	2.2%	–	1	0.3%	–	41	10.5%	–
	1300	100.0%	12.0%	8.0%	451	100.0%	75.4%	388	100.0%	87.1%	389	100.0%	75.6%

[1] 73A, 11719G, 14766T, 16126C, 16362C
[2] 73G, 11719G, 14766C, 16217C
Fractions 1 and 2 (Fr.l; Fr.2) are defined excluding haplogroups indicated by dash (-), see text for details.

tained most of their U3, K, J, Tl and also X lineages only after a substantial diffusion of Proto-Iranian' lineages to the Indian mtDNA pool had taken place.

One should not forget that India is large

As has already been observed for classical markers (Cavalli-Sforza *et al.* 1994), changes in the Indian genetic landscape do not occur gradually but are structured as a 'multitude of endogamous pockets'. At first glance, the same might seem to be the case for mtDNA and Y-chromosomal markers as well, because of strong founder effects and drift in genetically semi-isolated communities. Therefore, the frequency of any specific lineage in India can vary profoundly. Yet, there are a few general patterns of change.

The geographic distribution of lineages belonging to the 'western loan' fraction is concentrated mainly toward the north and west, declining from a high of 25 per cent in the Punjab (northwest) and 15 per cent in Gujarat (west) to a low of 4 per cent in western Bengal and Andhra Pradesh. There is no significant difference, however, in the frequency of this fraction of maternal lineages between Hindi speakers from Uttar Pradesh (6 per cent) and Dravidian speakers (4 per cent) from Andhra Pradesh (Kivisild *et al.* 1999a). Contrary to a prediction, deriving from a hypothesis that a higher frequency of 'western' gene lineages should discriminate higher castes from lower castes, it was found (Bamshad *et al.* 2001) that the difference between 'upper', 'middle' and 'lower' caste Dravidian-speaking Telugus is more strongly stratified in terms of the two Indian-specific maternal lineage clusters M3 (19 per cent in 'upper', 4 per cent in 'middle', and 1 per cent in 'lower' castes) and U2i (17 per cent - 10 per cent - 6 per cent, respectively), rather than by those of recent western Asian origin (5 per cent - 2 per cent - 1 per cent, respectively). The five-fold frequency difference for the latter can be interpreted in terms of a selective western impact on the mtDNA pools of upper castes (Bamshad *et al.* 2001). However, the fact that just the two autochthonous Indian mtDNA clusters, out of a much larger variety, comprise about a third of all maternal lineages of the upper castes of

Dravidian-speaking Telugus suggests strongly that the origin of the endogamous caste system should not be traced to a simple model of a putative Indo-Aryan invasion some 4700 years ago.

If one wants to maintain an Aryan invasion scenario, then one must at least assume that the incoming female lineages were absorbed selectively into an *already existing* profound stratification. One should also keep in mind possible differences in sizes of migrant/local populations: for example, if the entire population of the British Isles would *in corpore* emigrate today to India, it would, assuming random admixture, leave a genetic impact of no more than 5 per cent on average.

Table 17.2. *Some Y-chromosomal haplogroup frequencies in India, western Asia and Europe.*

		DYS257 92R7 HG1	M89 HG2	SRY1532 HG3	12f2 HG9	M130 HG10	M9 HG26	M20 HG28	YAP HG21
Punjab	67	9.0	4.5	50.7	20.9	3.0	0	11.9	0
Gujarat	29	10.3	13.8	24.1	20.7	17.2	3.4	10.3	0
Andhra Pradesh	3 6	41.7	11.1	8.3	5.6	16.7	0	0	0
Western Bengal	31	29.0	16.1	38.7	9.7	3.2	3.2	0	0
Sri Lanka	87	24.1	20.7	23.0	16.1	0	0	16.1	0
India[1]	250	21.6	12.4	30.4	15.6	6.8	0.8	12.4	0
AP, higher castes[2]	55	9.1	n.d.	45.5	9.1	1.8	12.7	0	
AP, middle castes[2]	111	12.6	n.d.	16.2	12.6	2.7	21.6	0	
AP, lower castes[2]	74	12.2	n.d.	20.3	5.4	5.4	13.5	0	
Iran[1]	83	8.4	13.3	10.8	41.0	1.2	3.6	7.2	14.5
Anatolia[1] and Caucasus[3]	323	24.8	26.3	4.6	32.2	0.9	3.7	1.2	4.3
Eastern Europe[4]	302	10.9	n.d.	47.0	3.3	n.d.	0.6	5.0	
Western Europe[5]	327	66.4	n.d.	3.7	3.9	n.d.	0.9	5.5	

[1] - this study
[2] - Andhra Pradesh (Bamshad *et al.* 2001)
[3] - including Armenians and Georgians from (Rosser *et al.* 2000).
[4] - including Polish, Russian, Byelorussian, and Ukrainian populations (Rosser *et al.* 2000).
[5] - including French, Belgian, Scottish, Basque, and Spanish populations (Rosser *et al.* 2000).

A recent massive western Y-chromosomal invasion of India?

Phylogeography of the mtDNA haplogroup M suggests that it spread during the Palaeolithic by the southern route taken by modern humans during their initial colonization of Eurasia (Quintana-Murci *et al.* 1999; Kivisild *et al.* 1999a). Because haplogroup M makes up the largest fraction (>50 per cent) of maternal lineages, both in India and eastern Asia, in population-wise comparisons, Indian maternal lineages cluster more closely with populations of East Asia (Bamshad *et al.* 2001). In the paternal history of present-day Indian populations, RPS4Y (M130) has been suggested to have been carried by the southern route migrants (Underhill *et al.* 2001). Yet its frequency in India is quite low (7 per cent). In fact, most Indian Y chromosomes cluster in haplogroups that are typical of European and western Asian populations (Rosser *et al.* 2000), but infrequent or even absent in eastern Asia (Su *et al.* 1999). Another NRY cluster, characterized by M52 and M69 mutations, has been suggested to accompany an early (likely pre-LGM) eastward expansion of Levantine mtDNA sub-cluster(s) of haplogroup U to India (Underhill *et al.* 2001).

There are differences in caste affinities for European Y-chromosomal varieties - - in Telugus, higher castes reveal shorter distances from Europeans (Bamshad *et al.* 2001). This sex-specific difference may be interpreted as resulting from a predominantly male-specific recent gene flow into the upper castes, not necessarily from Europe as such, but perhaps from western and/or central Asia. More specifically, Quintana-Murci *et al.* (2001) suggested that NRY marker 12f2 (haplogroup 9) indicates a Neolithic spread of farmers into India that is, with a short tandem repeat (STR) diversity in the background of $M9^{G}$-$SRY1532^{A}$ (haplogroup 3), consistent with an Indo-Aryan migration from Central Asia. Thus, both these studies suggest a substantial western male-specific gene flow to India during the Holocene.

However, several aspects of these genetic distance and haplogroup-wise comparisons should be considered with caution. First, the affinities of higher caste Telugus to European populations are not informative alone in telling from which source and when a putative migration took place. When comparing the Y-chromosomal affinities of Indian, western Asian and European populations in detail (Bamshad *et al.* 2001), it becomes apparent that 'higher' caste Telugus have, in contrast to 'lower' and 'middle' castes, a higher frequency (45.5 per cent) of haplogroup 3. Further typing of NRY markers in Indian populations has now revealed that a high frequency of this haplogroup is, however, characteristic not only of (eastern) European populations, but also of northwest India, where haplogroup 3 is

characteristic of about half of the male population and is also frequent among western Bengalis (Table 17.2). Therefore, the Y-chromosomal origin of 'higher' caste Telugus (i.e. high frequency of this particular NRY lineage among them) is not necessarily related to migration to India from outside and least likely from Iran and/or Anatolia, where haplogroup 3 is apparently much less frequent than among most of the Indian populations investigated in this respect.

Second, great caution is required when interpreting the dates deriving from Y-chromosomal STR coalescent calculations. Table 17.3 reveals that profoundly inconsistent time estimates can be reached when different calibration methods are used. Hence, it seems safer to operate with raw diversity estimates — to determine the polarity of the movement — leaving the time of origin question unanswered until reliable dating methods for Y-chromosomal STR diversity are worked out. Yet, even if time estimates are avoided, there are some problems introduced by sampling strategies and differences in demographic histories. For example, in the study by Quintana-Murci *et al.* (2001), a decline in diversity stretching from Iran to India was observed in haplogroups 3 and 9 and the authors rushed to interpret this empirical observation in favour of directional gene flow to India during Neolithic period (haplogroup 9). They linked this finding to the introduction of Indo-European languages (haplogroup 3) to India. Time estimates for their spread were derived from the STR clock.

Here, however, the clock is just a secondary problem — the first being 'the Indian reference sample' used. Indeed, the Indians included in this study consisted of a (limited) sample from Gujarat — one of the western maritime provinces of India. When extending the Indian sample with collections from different states, a quite different, even opposite, pattern emerges (Table 17.3). Indians appear to display the higher diversity both in haplogroups 3 and 9 — even if a pooled sample of eastern and southern European populations was considered. If we were to use the same arithmetic and logic *(sensu* haplogroup 9 is Neolithic) to give an interpretation of this table, then the *straightforward suggestion would be that both Neolithic (agriculture) and Indo-European languages arose in India* and from there, spread to Europe. We would also have to add that inconsistencies with the archaeological evidence would appear and disappear as we change rate estimates (Table 17.3).

Thirdly, it has been suggested that the Neolithic spread of farmers to Europe included, above 12f2, also Y chromosomes carrying markers M35 (at the background of YAP+) and M201 (Semino *et al.* 2000; Underhill *et al.* 2001). But note that while in Europe, Anatolia and the South Caucasus as well as in Iran, both M35 (haplogroup 21) and 12f2 (haplogroup 9) are present — and could even be called friendly co-inhabitants of the corresponding Y-chromosomal pools (Table 17.2) — this does not hold for India (Table 17.2). Indians, in contrast to their neighbours, generally lack the Alu insertion in their Y chromosomes (Kivisild *et al.* 1999b, and references therein), while possessing haplogroup 9 Y chromosomes in a substantial frequency. Thus, here we observe a situation, analogous to that indicated above for mtDNA. One does not find a strong correlation between the identity of European and Indian (putatively) 'Neolithic' components, having supposedly spread out from Levant/Middle East. In particular, the lack of YAP+ chromosomes in India (although found in some Pakistani populations), contrary to their presence in Europe, suggests that Y chromosomes carrying the M35 marker arrived in the Near and Middle East (likely from northern Africa) only after a putative earlier gene flow from Iran to India had taken place — but obviously earlier than the spread of a certain fraction of the Near Eastern Y chromosomes to Europe. One may see here an obvious analogy to a certain set of maternal lineages, such as K, U3, T1 and J.

However, in general, this lack of symmetry of possible eastwards and westwards Neolithic spreads from the Fertile Crescent should not be seen as a contradiction. Indeed, why should one assume that the initial area of the beginning of agriculture was itself geographically narrow and genetically homogeneous (see e.g. Bar-Yosef this volume)?

Y chromosomes and mitochondrial DNA — not necessarily together, not necessarily apart

Besides the example of parallelism in the patterns of the spread of Y-chromosomal markers RPS4Y and M52 with mitochondrial haplogroups M and U (Underhill *et al.* 2001), noted above, there might be other links of interest and worth further exploring. One involves M20 in NRY and haplogroups U7 and pre-HV[2] (see Tables 17.1 & 17.2) in mtDNA, which seem to co-decrease in frequencies from India and Iran to the Caucasus and the southern Mediterranean.

Clear differences in the genetic impact of a (probable) Neolithic component in Europeans and Indians, both in their mtDNA and Y-chromosome pools, are not easily explained with the simplest model of a single narrow source region — be it Anatolia or the Fertile Crescent — from which

Table 17.3. *Variance and coalescent time estimates on Y-chromosomal STRs.*

	Variance[1]	Age estimates Pedigree rate[2]	Phylogenetic rate[3]
Haplogroup 9			
Europe	0.44	6100	42,200
India	0.51	7100	48,900
Haplogroup 3			
Europe	0.24	3300	23,100
India	0.37	5200	35,700

[1] - the variances were calculated using five STR loci (DYS19, DYS388, DYS390, DYS391, and DYS393)
[2] - using rate of 1.8 x 10^{-3} based on most recent pedigree studies (Quintana-Murci *et al.* 2001)
[3] - using rate of 2.6 x 10^{-4} based on phylogenetic calibration (Forster *et al.* 2000).
Note that each calibration involves large error margins.

'Neolithic genes' moved in European and Indian directions. Other models should be sought and tested for explanation. But given the relatively low frequency of recent western lineages in the Indian mtDNA pool, a great number of samples from a wide variety of diverse Indian populations should be analyzed to collect a representative sample of sufficient size for a rigorous founder analysis. Similarly, massive founder analysis is desirable for the Y-chromosome because, as we have demonstrated, interpretations deriving solely from haplogroup frequency and even from combined SNP-STR diversity distributions can be misleading. It appears likely that more informative markers in this context are needed as well.

Concludingremarks

When discussing the genetics of Indian populations, different authors have now and then stressed the enormous complexity of their social systems, perhaps dating back much longer than written evidence. While that is certainly true, it nevertheless seems to us that knowledge accumulated thus far allows us not only to draw the first reasonably well-supported conclusions concerning what one may call the basic time-and-space oriented landmarks of the Indian maternal and paternal lineages, but also to avoid the pitfalls so easily created by an obvious desire 'to tell an exciting tale'. Table 17.4 brings together our current understanding of the arrival of maternal lineages to India — as far as it can be deduced from the approximately 1300 extant mtDNAs analyzed.

Unfortunately, here we cannot provide an 'equally simple' table for the NRY markers for reasons given above (see Table 17.3), but it would be very surprising indeed if present-day Indians, possessing at least 90 per cent of what we think of as autochthonous Upper Palaeolithic maternal lineages, were to carry but a small fraction of equally old paternal lineages.

Table 17.4. *Stratification of mtDNA lineages according to their probable sequence of appearance in the Indian sub-continent.*

Time period	MtDNA cluster	Frequency
Primarily early UP	M	66%
	U2i	8%
	K	14%
		88%
Primarily late UP	W	1-2%
	U7	2-3%
		~4%
Neolithic and later	H, T, J, I, HV, Ul, U5, U4, pre-HV, X, K, U2e, and Nl	~8%

References

Bamshad, M., T. Kivisild, W.S. Watkins, M.E. Dixon, C.E. Ricker, B.B. Rao, J.M. Naidu, B.V. Prasad, P.G. Reddy, A. Rasanayagam, S.S. Papiha, R. Villems, A.J. Redd, M.F. Hammer, S.V. Nguyen, M.L. Carroll, M.A. Batzer & L.B. Jorde, 2001. Genetic evidence on the origins of Indian caste populations. *Genome Research* 11(6), 994-1004.

Cavalli-Sforza, L.L., P. Menozzi & A. Piazza, 1994. *The History and Geography of Human Genes.* Princeton (NJ): Princeton University Press.

Chakrabarti, D.K., 1999. *India: an Archaeological History. Palaeolithic Beginnings to Early Historic Foundations.* Oxford: Oxford University Press.

Comas, D., F. Calafell, E. Mateu, A. Perez-Lezaun, E. Bosch, R. Martinez-Arias, J. Clarimon, F. Facchini, G. Fiori, D. Luiselli, D. Pettener & J. Bertranpetit, 1998. Trading genes along the silk road: mtDNA sequences and the origin of Central Asian populations. *American Journal of Human Genetics* 63(6), 1824-38.

Forster, P., A. Rohl, P. Lunnemann, C. Brinkmann, T. Zerjal, C. Tyler-Smith & B. Brinkmann, 2000. A short tandem repeat-based phylogeny for the human Y chromosome. *American Journal of Human Genetics* 67(1), 182-96.

Kivisild, T., M.J. Bamshad, K. Kaldma, M. Metspalu, E. Metspalu, M. Reidla, S. Laos, J. Parik, W.S. Watkins, M.E. Dixon, S.S. Papiha, S.S. Mastana, M.R. Mir, V. Ferak & R. Villems, 1999a. Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Current Biology* 9(22), 1331-4.

Kivisild, T., K. Kaldma, M. Metspalu, J. Parik, S.S. Papiha & R. Villems, 1999b. The place of the Indian mitochondrial DNA variants in the global network of maternal lineages and the peopling of the Old World,

in Papiha *et al.* (eds.), 135–52.

Kivisild, T., S.S. Papiha, S. Rootsi, J. Parik, K. Kaldma, M. Reidla, S. Laos, M. Metspalu, G. Pielberg, M. Adojaan, E. Metspalu, S.S. Mastana, Y. Wang, M. Golge, H. Demirtas, E. Schnekenberg, G.F. de Stefano, T. Geberhiwot, M. Claustres & R. Villems, 2000. An Indian ancestry: a key for understanding human diversity in Europe and beyond, in Renfrew & Boyle (eds.), 267–79.

Kumar, S.S., I. Nasidze, S.R. Walimbe & M. Stoneking, 2000. Discouraging prospects for ancient DNA from India. *American Journal of Physical Anthropology* 113(1), 129–33.

Lahr, M. & R. Foley, 1994. Multiple dispersals and modern human origins. *Evolutionary Anthropology* 3, 48–60.

Mountain, J.L., J.M. Hebert, S. Bhattacharyya, P.A. Underhill, C. Ottolenghi, M. Gadgil & L.L. Cavalli-Sforza, 1995. Demographic history of India and mtDNA-sequence diversity. *American Journal of Human Genetics* 56(4), 979–92.

Papiha, S., R. Deka & R. Chakraborty (eds.), 1999. *Genomic Diversity: Application in Human Population Genetics.* New York (NY): Kluwer Academic/Plenum Publishers.

Passarino, G., O. Semino, L.F. Bernini & A.S. Santachiara-Benerecetti, 1996. Pre-Caucasoid and Caucasoid genetic features of the Indian population, revealed by mtDNA polymorphisms. *American Journal of Human Genetics* 59(4), 927–34.

Quintana-Murci, L., O. Semino, H.-J. Bandelt, G. Passarino, K. McElreavey & A.S. Santachiara-Benerecetti, 1999. Genetic evidence of an early exit of *Homo sapiens* sapiens from Africa through eastern Africa. *Nature Genetics* 23(4), 43741.

Quintana-Murci, L., C. Krausz, T. Zerjal, S.H. Sayar, M.F. Hammer, S.Q. Mehdi, Q. Ayub, R. Qamar, A. Mohyuddin, U. Radhakrishna, M.A. Jobling, C. Tyler-Smith & K. McElreavey, 2001. Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *American Journal of Human Genetics* 68(2), 53742.

Renfrew, C., 1987. *Archaeology and Language: the Puzzle of Indo-European Origins.* London: Jonathan Cape.

Renfrew, C. & K. Boyle (eds.), 2000. *Archaeogenetics: DNA and the Population Prehistory of Europe.* (McDonald Institute Monographs.) Cambridge: McDonald Institute for Archaeological Research.

Richards, M., V. Macaulay, E. Hickey, E. Vega, B. Sykes, V. Guida, C. Rengo, D. Sellitto, F. Cruciani, T. Kivisild, R. Villems, M. Thomas, S. Rychkov, O. Rychkov, Y. Rychkov, M. Golge, D. Dimitrov, E. Hill, D. Bradley, V. Romano, F. Cali, G. Vona, A. Demaine, S. Papiha, C. Triantaphyllidis, G. Stefanescu, J. Hatina, M. Belledi, A. di Rienzo, A. Novelletto, A. Oppenheim, S. Norby, N. Al-Zaheri, S. Santachiara-Benerecetti, R. Scozzari, A. Torroni & H.-J. Bandelt, 2000. Tracing European founder lineages in the near eastern mtDNA pool. *American Journal of Human Genetics* 67, 1251–76.

Rolf, B., A. Rohl, P. Forster & B. Brinkmann, 1999. On the origin of the Turks: study of six Y-chromosomal short tandem repeats, in Papiha *et al.* (eds.), 75–83.

Rosser, Z.H., T. Zerjal, M.E. Hurles, M. Adojaan, D. Alavantic, A. Amorim, W. Amos, M. Armenteros, E. Arroyo, G. Barbujani, G. Beckman, L. Beckman, J. Bertranpetit, E. Bosch, D.G. Bradley, G. Brede, G. Cooper, H. Corte-Real, P. de Knijff, R. Decorte, Y.E. Dubrova, O. Evgrafov, A. Gilissen, S. Glisic, M. Golge, E.W. Hill, A. Jeziorowska, L. Kalaydjieva, M. Kayser, T. Kivisild, S.A. Kravchenko, A. Krumina, V. Kucinskas, J. Lavinha, L.A. Livshits, P. Malaspina, S. Maria, K. McElreavey, T.A. Meitinger, A.V. Mikelsaar, R.J. Mitchell, K. Nafa, J. Nicholson, S. Norby, A. Pandya, J. Parik, P.C. Patsalis, L. Pereira, B. Peterlin, G. Pielberg, M.L. Prata, C. Previdere, L. Roewer, S. Rootsi, D.C. Rubinsztein, J. Saillard, F.R. Santos, G. Stefanescu, B.C. Sykes, A. Tolun, R. Villems, C. Tyler-Smith & M.A. Jobling, 2000. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *American Journal of Human Genetics* 67, 152643.

Ruhlen, M., 1991. *A Guide to the World's Languages.* London: Edward Arnold.

Semino, O., G. Passarino, P.J. Oefner, A.A. Lin, S. Arbuzova L.E. Beckman, G. de Benedictis, P. Francalacci, A. Kouvatsi, S. Limborska, M. Marcikiae, A. Mika, B. Mika, D. Primorac, A.S. Santachiara-Benerecetti, L.L. Cavalli-Sforza & P.A. Underhill, 2000. The genetic legacy of paleolithic *Homo sapiens sapiens* in extant Europeans: a Y-chromosome perspective. *Science* 290, 1155–9.

Su, B., J. Xiao, P. Underhill, R. Deka, W. Zhang, J. Akey, W. Huang, D. Shen, D. Lu, J. Luo, J. Chu, J. Tan, P. Shen, R. Davis, L. Cavalli-Sforza, R. Chakraborty, M. Xiong, R. Du, P. Oefner, Z. Chen & L. Jin, 1999. Y-chromosome evidence for a northward migration of modem humans into Eastern Asia during the last Ice Age. *American Journal of Human Genetics* 65(6), 1718–24.

Tambets, K., T. Kivisild, E. Metspalu, J. Parik, K. Kaldma, S. Laos, H.-V. Tolk, M. Golge, H. Demirtas, T. Geberhiwot, S.S. Papiha G.F. de Stefano & R. Villems, 2000. The topology of the maternal lineages of the Anatolian and Trans-Caucasus populations and the peopling of the Europe: some preliminary considerations, in Renfrew & Boyle (eds.), 219–35.

Underhill, P.A., P.D. Shen, A.A. Lin, L. Jin, G. Passarino, W.H. Yang, E. Kauffman, B. Bonné-Tamir, J. Bertranpetit, P. Francalacci, M. Ibrahim, T. Jenkins, J.R. Kidd, S.Q. Mehdi, M.T. Seielstad, R.S. Wells, A. Piazza, R.W. Davis, M.W. Feldman, L.L. Cavalli-Sforza & P.J. Oefner, 2000. Y-chromosome sequence variation and the history of human populations. *Nature Genetics* 26, 358–61.

Underhill, P.A., G. Passarino, A.A. Lin, P. Shen, M. Mirazon Lahr, R. Foley, P.J. Oefner & L.L. Cavalli-Sforza, 2001. The phylogeography of Y-chromosome binary haplotypes and the origins of modern human populations. *Annals of Human Genetics* 65(1), 43–62.

MCDONALD INSTITUTE MONOGRAPHS

Examining the farming/ language dispersal hypothesis

Edited by Peter Bellwood & Colin Renfrew