



# University of HUDDERSFIELD

## University of Huddersfield Repository

Silva, Marina, Oliveira, Marisa, Vieira, Daniel, Brandão, Andreia, Rito, Teresa, Pereira, Joana B., Fraser, Ross M., Hudson, Bob, Gandini, Francesca, Edwards, Ceiridwen J., Pala, Maria, Koch, John, Wilson, James F., Pereira, Luísa, Richards, Martin B. and Soares, Pedro

A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals

### Original Citation

Silva, Marina, Oliveira, Marisa, Vieira, Daniel, Brandão, Andreia, Rito, Teresa, Pereira, Joana B., Fraser, Ross M., Hudson, Bob, Gandini, Francesca, Edwards, Ceiridwen J., Pala, Maria, Koch, John, Wilson, James F., Pereira, Luísa, Richards, Martin B. and Soares, Pedro (2017) A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals. *BMC Evolutionary Biology*, 17 (1). ISSN 1471-2148

This version is available at <http://eprints.hud.ac.uk/id/eprint/31666/>

The University Repository is a digital collection of the research output of the University, available on Open Access. Copyright and Moral Rights for the items on this site are retained by the individual author and/or other copyright owners. Users may access full items free of charge; copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational or not-for-profit purposes without prior permission or charge, provided:

- The authors, title and full bibliographic details is credited in any copy;
- A hyperlink and/or URL is included for the original metadata page; and
- The content is not changed in any way.

For more information, including our policy and submission procedure, please contact the Repository Team at: [E.mailbox@hud.ac.uk](mailto:E.mailbox@hud.ac.uk).

<http://eprints.hud.ac.uk/>

RESEARCH ARTICLE

Open Access



# A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals

Marina Silva<sup>1†</sup>, Marisa Oliveira<sup>2,3†</sup>, Daniel Vieira<sup>4,5</sup>, Andreia Brandão<sup>2,3</sup>, Teresa Rito<sup>2,6,7</sup>, Joana B. Pereira<sup>2,3</sup>, Ross M. Fraser<sup>8,9</sup>, Bob Hudson<sup>10</sup>, Francesca Gandini<sup>1</sup>, Ceiridwen Edwards<sup>1</sup>, Maria Pala<sup>1</sup>, John Koch<sup>11</sup>, James F. Wilson<sup>8,12</sup>, Luísa Pereira<sup>2,3</sup>, Martin B. Richards<sup>1\*†</sup> and Pedro Soares<sup>3,5\*†</sup>

## Abstract

**Background:** India is a patchwork of tribal and non-tribal populations that speak many different languages from various language families. Indo-European, spoken across northern and central India, and also in Pakistan and Bangladesh, has been frequently connected to the so-called “Indo-Aryan invasions” from Central Asia ~3.5 ka and the establishment of the caste system, but the extent of immigration at this time remains extremely controversial. South India, on the other hand, is dominated by Dravidian languages. India displays a high level of endogamy due to its strict social boundaries, and high genetic drift as a result of long-term isolation which, together with a very complex history, makes the genetic study of Indian populations challenging.

**Results:** We have combined a detailed, high-resolution mitogenome analysis with summaries of autosomal data and Y-chromosome lineages to establish a settlement chronology for the Indian Subcontinent. Maternal lineages document the earliest settlement ~55–65 ka (thousand years ago), and major population shifts in the later Pleistocene that explain previous dating discrepancies and neutrality violation. Whilst current genome-wide analyses conflate all dispersals from Southwest and Central Asia, we were able to tease out from the mitogenome data distinct dispersal episodes dating from between the Last Glacial Maximum to the Bronze Age. Moreover, we found an extremely marked sex bias by comparing the different genetic systems.

**Conclusions:** Maternal lineages primarily reflect earlier, pre-Holocene processes, and paternal lineages predominantly episodes within the last 10 ka. In particular, genetic influx from Central Asia in the Bronze Age was strongly male-driven, consistent with the patriarchal, patrilocal and patrilineal social structure attributed to the inferred pastoralist early Indo-European society. This was part of a much wider process of Indo-European expansion, with an ultimate source in the Pontic-Caspian region, which carried closely related Y-chromosome lineages, a smaller fraction of autosomal genome-wide variation and an even smaller fraction of mitogenomes across a vast swathe of Eurasia between 5 and 3.5 ka.

**Keywords:** Mitochondrial DNA, Indian Subcontinent, Genome-wide, Y chromosome, Neolithic, Indo-European

\* Correspondence: m.b.richards@hud.ac.uk; pedrosoares@bio.uminho.pt

†Equal contributors

<sup>1</sup>Department of Biological Sciences, School of Applied Sciences, University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK

<sup>3</sup>IPATIMUP (Instituto de Patologia e Imunologia Molecular da Universidade do Porto), Rua Júlio Amaral de Carvalho 45, 4200-135 Porto, Portugal

Full list of author information is available at the end of the article



## Background

Following the out-of-Africa (OOA) migration, South Asia (or the Indian Subcontinent, here comprising India, Pakistan, Bangladesh, Sri Lanka, Nepal and Bhutan) was probably one of the earliest corridors of dispersal taken by anatomically modern humans (AMH) [1–3]. A remarkable genetic diversity, probably the second highest after sub-Saharan populations [1, 4] supports this view. Although the oldest modern human fossils in South Asia (in Sri Lanka) date to only ~36–28 thousand years ago (ka) [5, 6], genetic and archaeological evidence suggest an arrival of AMH over 50 ka (discussed extensively in Mellars et al. [2]) but after the eruption of Mount Toba in Sumatra ~74 ka, contrary to some suggestions [7]. Whilst some argue for a hint of an earlier dispersal [8], the trace is restricted to Australia/New Guinea, where it amounts to only ~2% of the data, and its significance remains unclear [9, 10].

India, the second most populous country worldwide, includes a patchwork of different religions and languages, including tribal groups (~8% of the population, speaking over 700 different dialects of the Austro-Asiatic, Dravidian and Tibeto-Burman families) and non-tribal populations, who mostly practice Hinduism, grounded in a strictly hierarchical caste system, and speak Indo-European or Dravidian languages. Indo-European is often associated with northern Indian populations, Pakistan and Bangladesh, and a putative arrival in South Asia from Southwest Asia ~3.5 ka (the so-called “Indo-Aryan invasions”) has been frequently connected with the origins of the caste system [11, 12]. Although some studies suggested a greater affinity of upper castes to European and Southwest Asian populations than lower castes [13, 14], genetic data have provided no clear evidence for the “Indo-Aryan invasions” so far [15], and their very existence is challenged by many archaeologists [16].

South India, on the other hand, is dominated by Dravidian languages, which have been connected to Neolithic dispersals from Southwest Asia [1, 12, 17], although the South Asian situation is complex and others have argued for indigenous development of agriculture within the Dravidian heartland [18, 19]. Generally, India displays a high level of endogamy, a result of its strict social boundaries, and high genetic drift due to long-term isolation [20] which, combined with a very complex history, makes the genetic study of Indian populations challenging. Many recent genetic studies explored different layers of South Asian genetic diversity and population structure [2, 13–15, 17, 21–26], but they have tended to focus on one or other marker system and, as a result, decisive results on the details of the settlement process are still lacking.

In the last few years, genome-wide (GW) studies have been employed [27–29]. However, it remains difficult to

make inferences concerning the timing and direction of migrations from GW results, without including ancient DNA (aDNA) data (still lacking for South Asia), and for India the results have been contradictory, especially for differentiating amongst various migration waves at greater time depths.

There is a way forward, despite the current lack of aDNA. The maternally inherited mitochondrial DNA (mtDNA) allows researchers to identify specific lineage clusters (clades or haplogroups) and to correlate them with geography. By applying a reliable mitogenome molecular clock [30], it is then possible to date migration events and uncover fine demographic patterns that would otherwise be missed. Previous studies [2, 31, 32] revealed that South Asian mtDNA diversity consists largely of basal autochthonous lineages of the OOA founder haplogroups M and N (the latter mostly from the derived haplogroup R) [20]. Moreover, similar analyses can be carried out for the paternally inherited Y-chromosome variation, and comparisons of the two systems can detect sex bias in dispersal patterns.

To assess the phylogeographic patterns of South Asian mtDNA lineages, we compiled mitogenomes from South Asia and neighbouring regions available in the literature, complemented with samples from the 1000 Genomes Project (1KGP) [33] and the Human Genome Diversity Project (HGDP) [34], including understudied populations from Pakistan, Sri Lanka and Bangladesh, combined with several newly sequenced samples. We aimed to provide a refined mtDNA phylogeographic portrait of South Asia, including most crucially an assessment of the extent of genetic influx from other regions (primarily Southwest and Central Asia), in order to assess the impact of immigration during the Late Glacial, postglacial, Neolithic and Bronze Age periods in shaping genetic diversity and structure in South Asia. For a comprehensive overview across the genome, we have also carried out several fresh analyses of GW patterns across the regions of Southwest, Central and South Asia, and assessed sex-biased gene flow in the region by direct comparison across the same sample sets, using the 1KGP data now available for GW, mtDNA and Y-chromosome diversity.

## Methods

### Mitogenome dataset

In order to clarify the phylogeny of haplogroups M, N and R in South Asia, we focused our study on the lineages with recognized or potential likely origin in the Subcontinent, belonging to macrohaplogroups M (M2, M3, M4'67, M5, M6, M13'46'61, M31, M32'56, M33, M34'57, M35, M36, M39, M40, M41, M42b, M44, M48, M49, M50, M52, M53, M58, M62), R (R5, R6, R7, R8, R30 and R31) and N (N1'5). We also studied U2 (excluding U2e due to its West Eurasian origin) in a complementary analysis. We obtained

381 whole-mtDNA sequences from the 1KGP [33] (although we note that these were collected from caste families from India and lack tribal groups) and 51 from the HGDP [34]. In addition, we generated 13 new sequences (accession numbers: KY686204 -KY686216) belonging to the aforementioned haplogroups from Southeast Asia: seven from Myanmar, one from Vietnam, one from Thailand and four from Indonesia. We combined these with other published data from South Asia and neighbouring areas, including a total of 1478 samples (Additional file 1: Table S1). The additional sequences increased substantially the sample size particularly in the West of the Indian Subcontinent, necessitating a re-evaluation of previously inferred phylogeographic patterns [2, 35].

In order to discern migrations into the Subcontinent at different time periods, we also performed a complementary analysis of several “non-autochthonous” N lineages present in South Asia (H2b, H7b, H13, H15a, H29, HV, I1, J1b, J1d, K1a, K2a, N1a, R0a, R1a, R2, T1a, T2, U1, U7, V2a, W and X2—all subclades of West Eurasian haplogroups), amounting to a total of 635 mtDNA sequences (Additional file 1: Table S2). We assigned haplogroups using HaploGrep [36], in accordance with the nomenclature in PhyloTree (Build 17, February 2016) [37].

#### Phylogenetic reconstruction and statistical analyses of mtDNA

We reconstructed the mitogenome phylogenetic tree manually, based on a preliminary reduced-median network analysis [38] with Network v.4.611, checked considering the frequency of each mutation [30] and the nomenclature of PhyloTree (Build 17) [37]. We estimated coalescence ages within haplogroups M and N using both the  $\rho$  statistic [39] and maximum likelihood (ML). We calculated  $\rho$  estimates with standard errors estimated as in Saillard et al. [40] using a synonymous clock of one substitution in every 7884 years and a mitogenome clock of one substitution every 3624 years further corrected for purifying selection [30]. We assessed ML estimations using PAML 4 and the same mitogenome clock assuming the REV mutation model with gamma-distributed rates (discrete distribution of 32 categories) and two partitions, in order to distinguish hypervariable segments I and II (HVS-I and HVS-II) from the rest of the molecule. We performed runs both assuming and not assuming a molecular clock, in order to perform likelihood ratio tests (LRT) [41].

Since haplogroup M displays a peculiar phylogeographic pattern in South Asia [2], we additionally estimated node ages in different sub-regions of the Subcontinent (west, south, central and east) with two different approaches: (1) considering all samples from a given region, regardless of the putative geographical origin of the clade and (2) considering the most probable origin of each major

haplogroup (by considering branching structure, number of main branches, and centre of gravity) and including only basal lineages of each region [2]. To evaluate the effective population size ( $N_e$ ) of haplogroup M in each region, we computed Bayesian Skyline Plots (BSPs) [42] using BEAST 1.8.0 [43]. Although haplogroups do not equate to populations, BSPs applied to specific lineages can provide insights into the size variations of the populations that include them [44–47]. We used a relaxed molecular clock (lognormal in distribution across branches and uncorrected between them), a two-parameter nucleotide evolution model and a mutation rate of  $2.514 \times 10^{-8}$  mutations per site per year [48].

#### GW dataset and analysis

We filtered a dataset comprising 1440 samples with 500,123 SNPs, combining data from the 1KGP and 8 independent studies (Additional file 1: Table S3) for linkage disequilibrium (LD) using PLINK v1.07 [49] ( $r^2 > 0.25$ , with a window size of 100 SNPs and step size of 1), yielding a subset containing 164,149 SNPs. We subjected these to principal component analysis (PCA) using the standard PCA tool provided in EIGENSOFT v6.0.1 [50], with which we calculated the first 10 principal components (PCs), from which we calculated the fraction of variance. We included three additional 1KGP populations—Han Chinese from Beijing, China (CHB), Tuscans from Italy (TSI) and Yoruba from Nigeria (YRI)—for ADMIXTURE v1.23 [51] and sNMF [52] analyses for cross-checking. We performed runs for values of  $K$  between 2 and 10, with 5-fold cross-validation in ADMIXTURE, and complementary analyses including Yamnaya aDNA samples [53]. The filtered datasets used ( $r^2 > 0.25$ , window size of 100 SNPs and step size of 1) included 66,245 SNPs, for ADMIXTURE analysis, and 64,926 SNPs for the PCA.

In order to assess potential sex-biased gene flow into the region, we compared uniparental (mtDNA and Y-chromosome) and autosomal ancestry in the five 1KGP South Asian populations: Bengali from Bangladesh (BEB), Gujarati Indian from Houston (GIH), Indian Telugu from the UK (ITU), Punjabi from Lahore, Pakistan (PJI) and Sri Lankan Tamil from the UK (STU). For the autosomal ancestry variation, we considered the mean of each component for the highest likelihood value. The putative origin of the uniparental lineages present in the populations is shown in Additional file 1: Table S4. Y-chromosome phylogeny was based on Yfull tree v4.10 (<https://www.yfull.com/tree/>) [54]. We considered as South Asian the Y-chromosome lineages that most likely entered the Subcontinent before the Last Glacial Maximum (LGM): H [55–57], K2a1\* [58] (this attribution on the basis of the early-branching lineage, and therefore uncertain, but only concerns a single sample and does not affect the results in

any way), and C5 [58]. Y-chromosome haplogroups G, J, L1, L3, Q, R1 and R2 seem to have entered South Asia more recently in the early to mid-Holocene from a West Eurasian source [17, 55–59]. C(xC5), O and N probably had a Holocene Eastern origin [55, 58, 60, 61].

## Results

### Indigenous South Asian mtDNA lineages: An explanation for the anomalous age of haplogroup M

The complete phylogeny for autochthonous South Asian M, N and R lineages is shown in Additional file 2 including age estimates for the main nodes (using  $\rho$  and ML age estimates). Age estimates for clades mentioned in the text are shown in Table 1 and a schematic phylogenetic tree scaled by ML age estimates is shown in Fig. 1.

Although haplogroup M in Asia has been shown to depart from a strict molecular clock [62], we found no evidence for a clock violation when performing a LRT ( $p > 0.05$ ). Curiously, however, we found violations to the molecular clock for South Asian R lineages ( $p < 0.00001$ ). Since ML analysis is partly based on the tree structure, it averages the branch lengths and provides similar estimates to a previous relaxed clock [63]. The values indicated throughout the text are therefore ML estimates (corrected for purifying selection). This is not observed in the global mtDNA tree [30, 64] and seems peculiar to the haplogroups in South Asia, due to demographic effects, as we argue below.

There are two major founder clades detected in South Asia (haplogroup N is very rare and its age does not correspond to a founder age). As previously, the age of haplogroup M, at 50.1 [44.8; 55.5] ka, and R, at 64.5 ka [55.9; 73.2] are younger than the Mount Toba eruption (~74 ka), suggesting a later arrival [2]. Haplogroup R and several of its subclades (R7, R30, R31) appear older than M, but this may be illusory—see below. The older clades in R predominate in the west and south of the Subcontinent, supporting a southern coastal route of primary colonization [1–3].

The phylogeography of haplogroup M is complex. While some older lineages (e.g. M2, M6, M32'56, M36, M39) originated in the western or southern regions of the Subcontinent (similarly to R), others trace to central India (M4'67, M35, M52) or the east (M13b, M31, M42b, M61, M49, M50 and M60). We need to tease out these more detailed patterns to explain the discrepancy in the age estimates.

If we perform regional estimates simply by considering all samples of each region, no discernible patterns are apparent, with M age estimates in the south and east showing similar ages (Table 2). However, when we take into account the inferred source for each clade and re-partition the data on that basis, the re-estimated age for M in the west becomes 55.3 [45.1; 65.9] ka—higher than across the rest of the Subcontinent (Table 2). This

suggests an early expansion in the west, similar to R, and a common origin and spread of both M and R along the southern coastal route, as also suggested recently from analyses of ancient DNA (aDNA) [65]. Although M has previously been dated to an earlier age in East Asia [30, 66], the lower age of M in the east of the Subcontinent *versus* the west argues against an eastern origin of M as recently proposed [35].

This result suggests that an ancient western ancestry may have been disguised by further re-expansions of haplogroup M in South Asia. Several branches of M (M38, M65, M45, M5b, M5c, M34, M57, M33a) display signals of dispersals from the east and the centre dating to ~45–35 ka, and M4'67 (which is only separated by a single mutation from the root of M), with a possible origin in central India, displays an extraordinary multi-branching structure dating to 38.0 [30.1; 46.0] ka, suggesting a major expansion at that time. If we consider that a root type of M could have survived for ~10,000 years after it arose (as is evident from modern clades within that age range), it is plausible that re-expansion created a secondary founder effect within M that decreased the overall age estimates. Such a scenario would impact even more on  $\rho$  than ML estimates, which is indeed what we see (Table 1). An expansion 45–35 ka would also fit well with the palaeoenvironmental and archaeological evidence [2, 67, 68], and is further supported by an increment in  $N_e$  associated with M across South Asia from ~40 ka (Additional file 1: Figure S1).

The next major discernible signal in indigenous lineages begins ~12 ka, at the Pleistocene/Holocene transition. Various star-like clades dating 12–9 ka suggest a rapid expansion across the Subcontinent, namely M6a1a (11.4 ka), M18a (9.2 ka), M30d (12.1 ka), R8b1 (11.6 ka) and U2b2 (9.2 ka), all from a southern source; and R30c + 373 (12.4 ka), from the west. An increment in  $N_e$  is also observed at this time in the BSP for haplogroup M in the west and south (Additional file 1: Figure S1).

We also see a further increment in the last few millennia. BSPs for M in the west and centre show an increment in the last 2.5 ka (Additional file 1: Figure S1), associated with the emergence of several subclades in the west (M2a3a + 4314, M2a1b, M2c + 1888 + 146, M30a2, M5a3b, M6a1 + 5585 + 146 + 1508) and centre (M2a1a1b, M3b, M3a1a, M63, M5a2a2 + 234, M5a3a and M61a + 5294).

### West Eurasian mtDNA lineages in South Asia: Multiple dispersals from the northwest since the LGM

Prehistoric West Eurasian lineages make up almost 20% of the South Asian genetic pool overall.

### LGM and Late Glacial arrivals

The earliest genetic evidence of movements into the Subcontinent after the first settlement is seen in



**Table 1** Age estimates (in ka) of the clades mentioned in the text. Node ages for haplogroup U2 were estimated in an independent analysis

Clade	ML	$\rho$ whole mtDNA	$\rho$ synonymous clock
N	67.7 [58.4–77.1]	63.5 [51.7–75.7]	71.5 [51.3–91.8]
R	64.5 [55.9–73.2]	57.0 [48.6–65.5]	63.5 [49.1–77.8]
R7	62.2 [52.9–71.7]	62.0 [43.0–81.6]	76.0 [42.2–109.8]
R8b1	12.0 [7.0–17.1]	11.1 [5.8–16.5]	5.1 [2.1–8.1]
R30	60.9 [49.6–72.5]	53.0 [40.6–65.8]	61.5 [40.5–82.6]
R30c + 373	8.6 [0.0–48.1]	9.0 [3.5–14.6]	6.3 [0.5–12.1]
R31	62.5 [53.0–72.1]	70.8 [50.4–92.0]	75.2 [43.3–107.1]
M	50.1 [44.8–55.5]	41.2 [37.0–45.4]	41.3 [34.6–48.0]
M2	43.2 [34.7–52.0]	51.2 [35.8–67.3]	44.5 [23.2–65.8]
M2a1a1b	22.0 [0.0–6.0]	3.3 [0.0–7.7]	3.4 [0.0–10.0]
M2a1b	0.7 [0.0–2.5]	0.6 [0.0–1.5]	1.0 [0.0–2.9]
M2a3a + 4314	0.9 [0.0–2.8]	0.9 [0.0–2.5]	–
M2c + 1888 + 146	2.5 [0.0–19.9]	3.5 [0.0–8.4]	10.5 [0.0–25.1]
M3a1 + 204 + 14476	1.2 [0.0–2.7]	1.0 [0.0–2.0]	2.4 [0.0–5.0]
M3a1 + 204 + 10845 + 13105	0.9 [0.0–3.3]	0.9 [0.0–2.6]	0.0
M3b	1.8 [0.0–4.5]	2.2 [0.0–5.7]	5.5 [0.0–15.6]
M4'67	38.0 [30.1–46.0]	27.8 [23.4–32.3]	22.7 [18.3–27.0]
M5a1b1a1 (M5a1b + 3954 + 9833 + 16298)	3.0 [1.0–5.0]	2.7 [1.4–4.1]	2.3 [0.0–4.7]
M5a2a + 8158 + 199	1.9 [0.7–3.2]	1.8 [0.7–2.8]	3.0 [0.6–5.3]
M5a2a2 + 234	1.5 [0.0–4.2]	1.4 [0.2–2.7]	2.6 [0.0–5.6]
M5a3a	0.7 [0.0–3.3]	–	–
M5a3b	1.6 [0.0–3.5]	1.5 [0.1–3.0]	1.6 [0.0–3.8]
M5b	33.0 [23.6–42.9]	30.7 [20.9–40.9]	36.9 [17.7–56.2]
M5c	35.2 [24.2–46.6]	41.5 [28.2–55.3]	49.3 [25.0–73.6]
M6	35.6 [25.9–45.7]	37.9 [23.4–53.2]	48.7 [19.6–77.9]
M6a1 + 5585 + 146 + 1508	1.3 [0.0–3.2]	1.1 [0.0–2.3]	0.9 [0.0–2.6]
M6a1a	11.4 [4.0–19.2]	10.6 [6.6–14.7]	10.3 [4.9–15.8]
M13b	32.8 [21.5–44.5]	30.7 [17.1–45.2]	33.8 [12.2–55.4]
M18a	9.2 [6.0–12.4]	8.1 [5.6–10.5]	6.0 [2.1–10.0]
M30a2	2.3 [0.0–8.5]	1.9 [0.0–4.8]	–
M30d	11.4 [4.6–18.5]	9.2 [4.1–14.3]	10.0 [2.8–17.2]
M31	38.0 [27.9–48.4]	38.4 [25.9–51.4]	43.6 [20.6–66.7]
M32'56	42.4 [25.8–60.0]	33.0 [16.7–50.4]	14.5 [0.5–28.4]
M33a	35.2 [24.5–46.3]	29.1 [21.2–37.2]	32.3 [19.3–45.3]
M34	29.7 [19.4–40.4]	28.1 [17.6–39.1]	39.4 [17.9–60.9]
M35	40.1 [25.4–55.5]	26.9 [18.5–35.6]	26.4 [15.5–37.3]
M36	36.4 [25.8–47.4]	26.9 [16.2–38.2]	30.6 [11.6–49.6]
M38	29.4 [20.4–38.7]	32.5 [23.6–41.7]	33.8 [19.4–48.2]
M39	36.8 [27.3–46.6]	23.7 [15.3–32.5]	21.2 [9.1–33.2]
M42b	42.5 [33.8–51.4]	43.5 [27.1–60.8]	49.7 [22.4–77.1]
M45	30.6 [19.0–42.8]	30.7 [18.5–43.6]	33.8 [14.1–53.5]
M49	31.0 [21.2–41.2]	26.3 [18.1–34.8]	25.6 [13.6–37.5]
M50	43.3 [30.6–56.6]	47.4 [32.3–63.3]	52.0 [26.4–77.7]

**Table 1** Age estimates (in ka) of the clades mentioned in the text. Node ages for haplogroup U2 were estimated in an independent analysis (*Continued*)

M52	33.4 [23.4–43.9]	31.0 [22.1–40.2]	33.4 [19.0–47.9]
M57	32.4 [18.2–47.3]	28.8 [19.0–38.9]	24.5 [11.5–37.6]
M60	36.5 [23.3–50.4]	24.8 [15.8–34.2]	21.0 [8.9–33.2]
M61	24.6 [13.6–36.2]	11.8 [6.0–17.8]	12.4 [1.4–23.4]
M61 + 5294	1.6 [0.0–5.1]	1.9 [0.0–4.8]	2.0 [0.0–5.8]
M63	1.4 [0.0–3.8]	1.3 [0.0–2.8]	1.3 [0.0–3.9]
M65	29.3 [14.7–44.8]	20.6 [12.6–29.0]	21.3 [8.4–34.1]
N1a2	12.5 [2.9–22.6]	6.5 [2.1–11.2]	7.9 [0.2–15.6]
N1a1b1	20.9 [11.4–30.8]	19.0 [10.4–27.9]	22.1 [7.6–36.6]
H2b	6.2 [3.8–8.7]	5.2 [3.4–7.1]	4.8 [1.7–7.9]
H13a2a + 8952	6.6 [1.3–12.1]	7.2 [1.0–13.6]	2.0 [0.0–5.8]
H29 + 9156 + 4689	1.6 [0.0–4.7]	1.3 [0.0–3.8]	3.9 [0.0–11.7]
HV + 73	23.7 [17.1–30.4]	30.1 [19.6–41.0]	29.8 [12.1–47.5]
HV + 146	23.9 [10.3–38.4]	19.0 [8.8–29.8]	11.8 [0.0–25.2]
HV + 9716	19.6 [8.1–31.8]	13.4 [5.0–22.2]	3.9 [0.0–11.7]
HV + 16311	15.6 [9.9–21.5]	15.5 [7.6–23.8]	19.3 [3.4–35.1]
HV2	21.9 [15.1–28.9]	30.7 [17.9–44.2]	38.1 [12.2–64.0]
HV12b	13.3 [5.3–21.6]	12.6 [5.7–19.8]	5.6 [0.7–10.6]
HV14 + 150	6.9 [2.9–11.0]	6.7 [1.0–12.6]	11.4 [0.0–25.7]
I1	13.8 [8.5–19.2]	10.6 [6.3–15.0]	11.8 [4.1–19.6]
J1b1b1	13.9 [8.6–19.3]	12.6 [7.9–17.4]	12.4 [5.1–19.7]
J1d	24.1 [14.9–33.7]	16.2 [10.2–22.3]	17.3 [7.1–27.6]
K1a1b2a	10.4 [4.0–17.0]	12.0 [4.1–20.3]	7.9 [0.0–18.8]
K2a5	7.6 [3.6–11.7]	8.2 [3.9–12.6]	5.3 [1.1–9.5]
K2a5 + 2831	6.8 [2.9–10.7]	8.4 [3.5–13.5]	4.7 [0.0–10.1]
K2a5 + 2831 + 189	5.9 [2.1–9.8]	10.6 [3.2–18.4]	7.9 [0.0–18.8]
R0a2 + 11152	7.1 [1.1–13.3]	6.5 [0.8–12.5]	7.9 [0.0–18.8]
R2a + 7142	3.2 [0.0–6.9]	2.9 [0.0–5.9]	1.8 [0.0–4.2]
T2 + 195 + 4225	9.7 [2.9–16.8]	6.8 [2.3–11.5]	3.2 [0.0–7.5]
T2b	10.6 [5.3–16.0]	7.1 [3.6–10.8]	3.4 [0.0–7.2]
T2d1a	12.0 [5.0–19.3]	10.6 [4.5–16.9]	7.9 [0.0–16.8]
T2e2	10.6 [3.4–18.1]	12.0 [4.1–20.3]	11.8 [0.0–25.2]
U1a1	20.0 [14.4–25.7]	15.2 [10.4–20.1]	15.2 [6.2–24.3]
U1a1a2a	2.5 [0.0–7.3]	1.9 [0.0–4.8]	5.9 [0.0–14.6]
U1a3 + 10253	10.3 [4.6–16.2]	8.9 [4.6–13.3]	10.8 [2.9–18.8]
U1a3a	5.2 [0.0–11.0]	3.9 [0.0–8.4]	3.9 [0.0–11.7]
Pre-U1c	21.4 [9.1–34.5]	14.3 [6.7–22.2]	13.1 [1.6–24.7]
U2	52.3 [41.6–63.3]	53.8 [41.8–66.2]	54.1 [36.6–71.6]
U2b2	9.2 [6.3–12.2]	8.6 [6.1–11.1]	9.9 [5.3–14.4]
U2c1 + 146	1.4 [0.0–24.8]	1.7 [0.0–5.1]	–
U7a	18.1 [14.4–22.0]	18.8 [14.5–23.2]	19.7 [11.5–27.9]
U7a + 12373	10.2 [3.0–17.6]	8.8 [2.8–15.0]	10.5 [0.0–23.1]
U7a3a + 6150	9.8 [4.4–15.4]	8.6 [3.5–13.8]	2.0 [0.0–5.8]
U7b + 16309!	10.9 [6.1–15.9]	8.6 [3.6–13.8]	8.4 [0.0–18.1]

**Table 1** Age estimates (in ka) of the clades mentioned in the text. Node ages for haplogroup U2 were estimated in an independent analysis (*Continued*)

W3a1 + 143	9.8 [3.0–16.8]	7.9 [1.5–14.5]	19.7 [2.4–37.0]
W3a1 + 1709	8.1 [1.6–15.0]	6.5 [0.8–12.5]	–
W3a1b	11.4 [6.3–16.6]	11.2 [6.1–16.3]	7.1 [1.1–13.1]
W4	15.8 [9.5–22.3]	15.5 [8.7–22.5]	11.8 [2.4–21.3]
W6	11.5 [5.0–18.3]	10.9 [5.7–16.3]	13.1 [6.5–19.8]
X2 + 153 + 7109	7.7 [0.0–17.0]	4.3 [0.0–9.0]	2.6 [0.0–7.8]

haplogroup N1a1b1, which dates to ~21 ka (Additional file 1: Figure S2), with a probable source in the Near East [69]. Other haplogroups with similar age estimates and a Near Eastern source (pre-HV2, HV + 146!, HV + 9716, HV + 73!, pre-U1c, U1a1, J1d and a basal clade within T2) may have moved eastwards in the same time frame (Table 1, Additional file 1: Figure S2), corresponding to 2.6% in the overall South Asian 1KGP data. Further Near Eastern clades (W4, HV + 16311!, HV12b, I1, U7a and J1b1b1) spread to South Asia in the Late Glacial period, 16–13 ka (Table 1, Additional file 1: Figure S2), with frequencies of 4.5% in the South Asian 1KGP data.

**Early postglacial arrivals**

At ~12 ka, when various indigenous lineages show signals of expansion, we also observe further lineages arriving from Southwest Asia with exclusively South Asian branches (T2e2, T2 + 195 + 4225, W3a1 + 143, W3a1b, U1a3 + 10253, N1a2, U7a + 12373 and U7a3a + 6150) (Table 1, Additional file 1: Figure S2). Furthermore, South Asian lineages are nested within numerous other branches

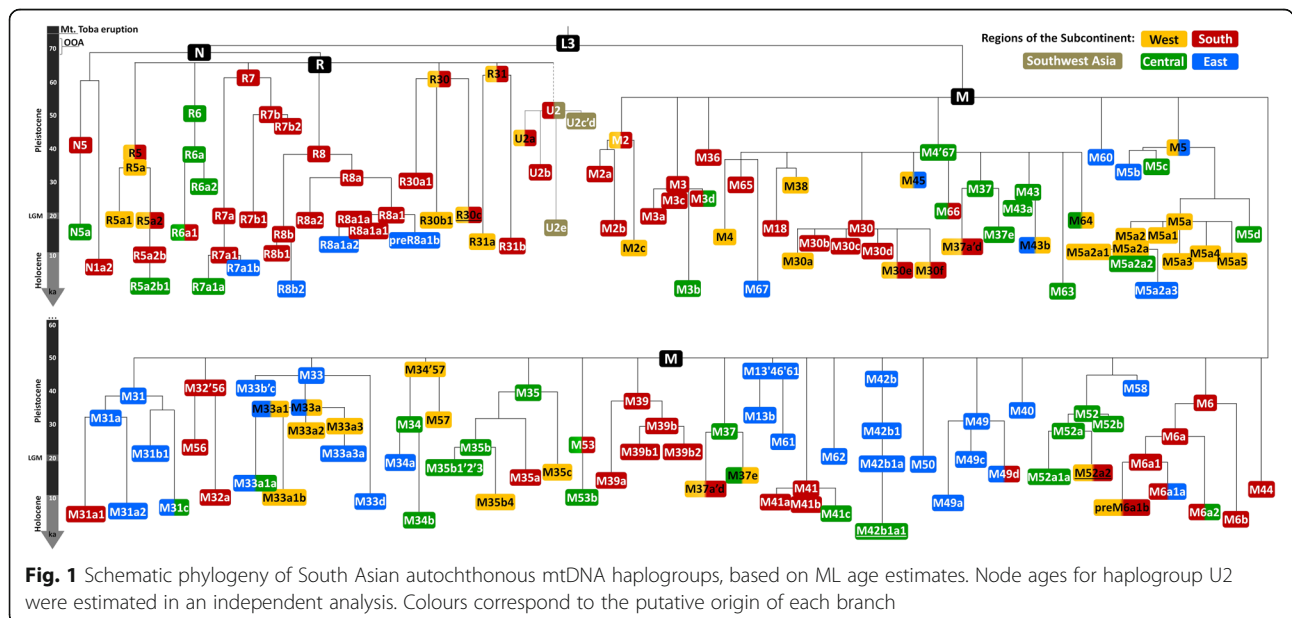
with similar node age estimates (W6, T2b, T2d1a, U7b + 16309! and K1a1b2a), allowing us to circumscribe the arrival times (Table 1, Additional file 1: Figure S2). These lineages represent a frequency of 4.7% in the South Asian 1KGP dataset.

**Neolithic arrivals**

More lineages entered the Subcontinent ~9–5 ka, representing putative Neolithic markers with a distinct origin in Anatolia, the Caucasus and Iran, again harbouring distinctive nested South Asian subclades (K2a5 + 2831 + 189, HV14 + 150, H13a2a + 8952, K2a5 + 2831, X2 + 153! + 7109 and U1a3a) (Table 1, Additional file 1: Figure S2) (3.4%). There is also evidence of movements from the Arabian Peninsula/Near East; the branch R0a2 + 11152 (~7.1 ka) is the most striking example. One case, H2b, might trace its source to Eastern Europe and may have entered South Asia through Central Asia a little later, as we discuss below.

**Bronze Age arrivals**

In the last 4 ka, most genetic influx on the maternal line was restricted to Pakistan and traces mostly to Iran



**Fig. 1** Schematic phylogeny of South Asian autochthonous mtDNA haplogroups, based on ML age estimates. Node ages for haplogroup U2 were estimated in an independent analysis. Colours correspond to the putative origin of each branch



**Table 2** Age estimates (in ka) of haplogroup M in different regions of South Asia: (1) using the raw modern geographic distribution and (2) considering the most probable origin of each major haplogroup and including only basal lineages of each region

	ML	$\rho$ whole mtDNA	$\rho$ synonymous clock
(1) West	47.7 [41.3–54.2]	37.4 [31.6–43.2]	39.0 [28.8–49.2]
South	47.2 [41.5–53.1]	42.4 [36.7–48.3]	40.0 [31.4–48.6]
East	47.7 [42.5–53.0]	42.4 [38.4–46.6]	43.9 [37.1–50.8]
Central	43.6 [38.1–49.1]	40.8 [35.4–46.3]	41.4 [33.0–49.7]
(2) West	55.3 [45.1–65.9]	44.5 [32.5–57.0]	50.6 [29.7–71.4]
South	48.9 [42.1–55.8]	47.5 [39.2–56.0]	41.1 [29.6–52.6]
East	45.2 [38.8–51.8]	40.8 [34.6–47.0]	40.1 [31.3–48.9]
Central	39.5 [31.9–47.2]	33.0 [26.8–39.3]	34.80 [23.2–46.5]

(H29 + 9156 + 4689, R2a + 7142 and U1a1a2a) (2.4% in South Asia, reaching 5.4% in the western populations). Gene flow at this time was clearly bi-directional, as seen in the expansion west of lineages M5a2a4, U2c1b + 146 and M3a1b + 13105). This is reflected in the genome-wide ADMIXTURE analysis (below), where the autochthonous South Asian component (green in Fig. 2a) appears at low levels in Iran. As an aside, the bulk of Romani lineages belongs to the branch M5a1b1a1 [70] at 3.0 ka, supporting previous linguistic and genetic evidence for a South Asian origin for the Romani diaspora [70, 71] in the west of the Subcontinent.

### GW overview of South Asia

South Asian populations can be distinguished in both the ADMIXTURE and sNMF analyses from  $K=3$  (Additional file 1: Figure S3 and Figure S4), highlighting the distinctive genetic diversity of the region. At the highest likelihood value of  $K=7$  (Fig. 2a and Additional file 1: Figure S5a), the overall pattern is straightforward and clinal [72], with a substantial autochthonous component (shown in green) across the region, apart from the Kalash, which display a virtually exclusive component probably caused by localised genetic drift in a small, isolated population [72, 73].

A striking feature in both the ADMIXTURE and sNMF analyses (for  $K=7$ ) is the much higher fraction of West Eurasian components (brown, yellow and dark blue) in the western (especially Pakistani) South Asian populations. The main non-autochthonous component in the Subcontinent, the Iran/Caucasus/Steppe component (brown), exceeds 35% in Pakistan and Gujarat [23–25], although it reaches most of the Subcontinent. This component approaches ~100% in Late Palaeolithic and Mesolithic remains from the Caucasus, and was therefore dubbed the “Caucasus hunter–gatherer” (CHG) component [74], but it is seen at similarly high frequencies in remains from Mesolithic and Neolithic Iran [75]

and at ~50% in Early Bronze Age Yamnaya pastoralist remains from the Pontic-Caspian steppe [53, 76], as shown in Additional file 1: Figure S6 for  $K=7$  (lowest cross-validation error, Additional file 1: Figure S5b).

The Pakistani Muslim Balochi, Brahui and Makrani carry ~15% of the Near Eastern/Arabian component (yellow), which is carried across Europe with the spread of the Early Neolithic [75, 77]. However, this component is virtually absent in other South Asians (including Muslims) except for Jewish groups (supporting previous mtDNA evidence for little genetic input from Arabia into Indian Muslim populations [78]).

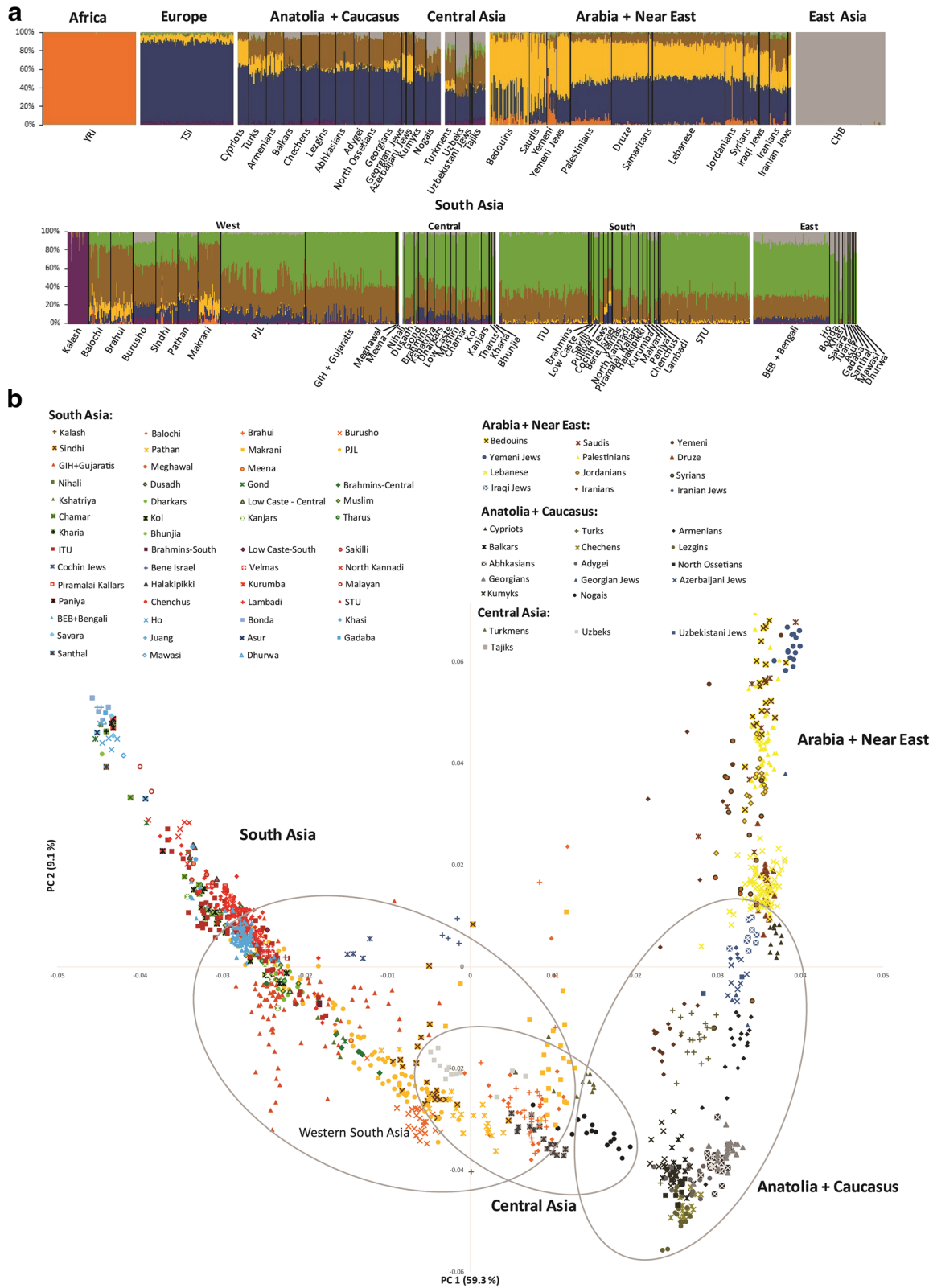
The PCA (Fig. 2b) portrays a complex gradient of affinities, but with South Asians closer to Central Asian and Caucasus groups than to those from the Near East or Arabia. Pakistani populations occupy an intermediate position, particularly close to the currently Turkic-speaking peoples of Central Asia (the Turkmens, the Nogais and the Uzbeks) and the Indo-Iranian-speaking Tajiks in PC1 (which accounts for 59.3% of the variation). Genetically, Turkic-speaking groups resemble their geographic neighbours, indicating deep local ancestry and recent language shift [79].

The current paradigm for explaining modern Indian population structure suggests that they derive from admixture between two main ancestral populations, Ancient North Indians (ANI) and Ancient South Indians (ASI) [25], with the proximity of Pakistani groups and Gujaratis to Southwest Asians due to high levels of ANI ancestry [25], which may have arrived in two waves [24]. However, our mtDNA results (and the current GW analysis) suggest that the process is likely to have been much more complex. The profile for Pakistani populations is likely the result of at least four waves of dispersal into the region, involving all three of the inferred ancestral West Eurasian components, from at least as far back as the LGM through into the Bronze Age.

The Yamnaya aDNA samples are scattered around the Central Asian and Pakistani groups (Additional file 1: Figure S8), confirming the ADMIXTURE results (Additional file 1: Figure S6), and suggesting links between the Bronze Age Steppe and today’s Central Asia and Indian Subcontinent. Pakistanis and Gujaratis appear much more scattered in PC1 than other South Asians, which only show substantial divergence in the lower-weight PC2 (9.1%) and PC3 (6.3%) (Fig. 2b, Additional file 1: Figure S7).

### Comparing marker systems: Massively different ancestry on the male and female lines of descent

The mtDNA patterns suggest much higher levels of autochthonous variation on the maternal line (~70–90%) compared to the overall GW estimate (about a half to two-thirds), the implications of which we further



**Fig. 2 a** ADMIXTURE analysis for  $K=7$ . **b** PCA of South Asian populations. Detailed information on the populations included in the Additional file 1: Table S3. Note that the three typical European components are not detected here in the Tuscans, probably due to the small overall European representation in the analysis

explored by studying Y-chromosome lineages. We used the five South Asian 1KGP populations, which comprise unbiased population data, and are the only available datasets that can be simultaneously analysed for GW, mtDNA and Y-chromosome variation.

A markedly higher proportion of male lineages of likely West Eurasian origin, of ~50–90%, is evident across the Subcontinent (Fig. 3c), in comparison with both the maternal line (Fig. 3b) and the GW pattern (Fig. 3d). A sex-biased pattern is also seen in the East Asian fraction, but is much less marked, with a much lower contribution overall and mainly focused on speakers of Tibeto-Burman and Austroasiatic language families [22].

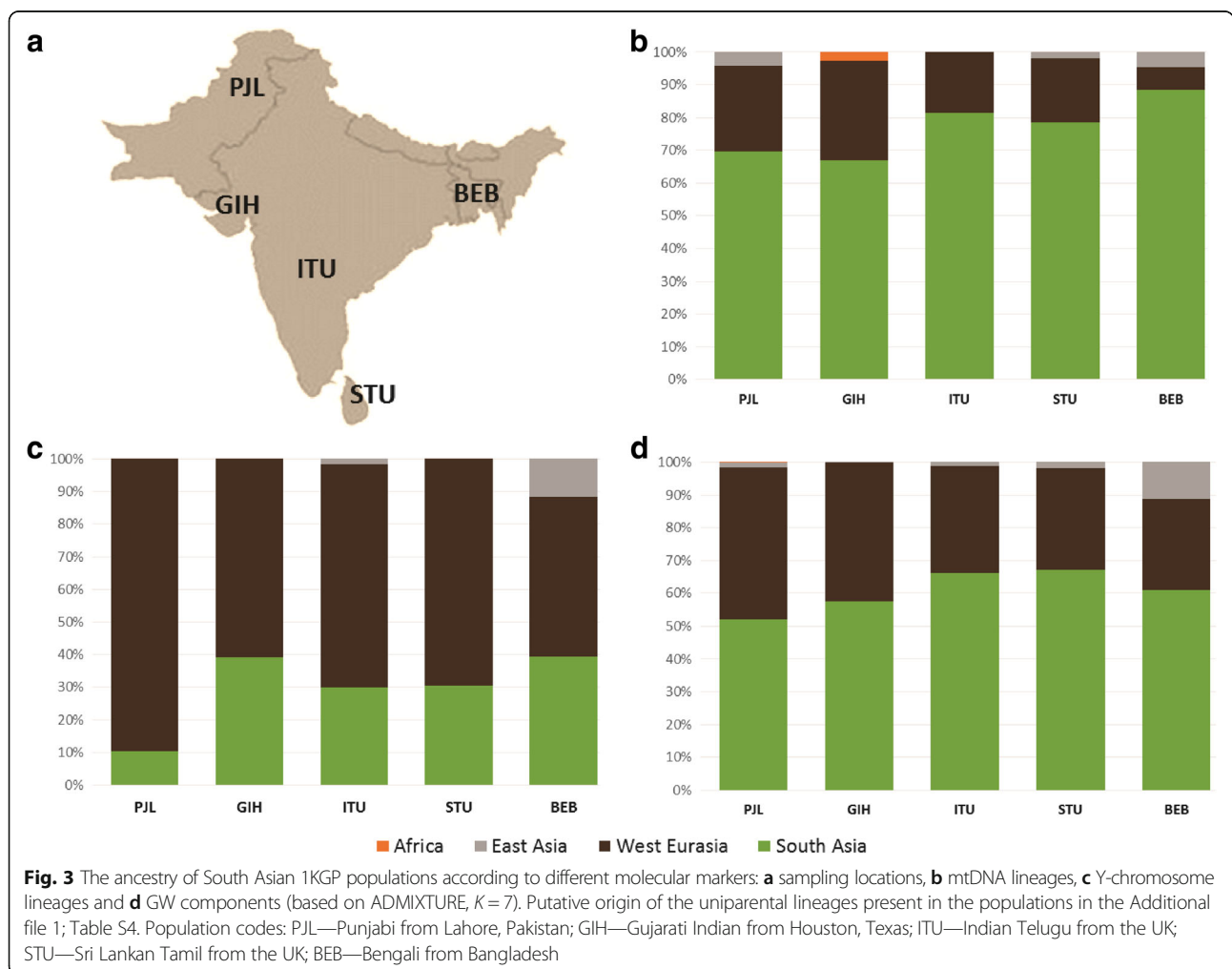
**Discussion**

**Towards a more fine-grained history of South Asian settlement**

The phylogeographic analysis of non-recombining marker systems offers certain strengths that can complement genome-wide analyses. In particular, the polarity of

gene trees allows us to identify the source of dispersals, and the increasing precision of molecular clocks for mtDNA and the Y chromosome allows us to date events during the ancestry of lineages with some confidence. However, the contribution of the two systems to the overall picture is not always the same, and South Asia is a case in point. Here it is clear from our analyses that there is a very strong sex bias in the ancestry of South Asians. The female line of descent is mostly autochthonous and traces back to the first settlement ~55 ka. However, the male line of descent emphasizes more recent ancestry, since the LGM, from Southwest Asia and Central Asia.

The mtDNA is, therefore, at present a uniquely powerful tool for teasing out multiple settlement episodes and dating them, establishing a timeline for demographic events in South Asia. By combining that information with GW patterns and Y-chromosome data, and taking into account also archaeological, palaeontological and palaeoclimatological data, we can reconstruct an outline demographic history of human populations in South



Asia that captures some of the complexity of the region and moves beyond simplistic models of admixture between autochthonous Indians and invading Neolithic farmers or Indo-Aryan speakers (Fig. 4).

**Resolving the Pleistocene modern human settlement**

Evidence is mounting that haplogroups M, N and R had a common origin and entered South Asia together, following a southern coastal route from Eastern Africa after the Toba eruption [2, 3]. This is supported by their global (non-African) distribution [3], including the detection of basal M lineages, M0 and M1, in Europe and the Near East respectively [65, 80, 81], and their similarity in age elsewhere either using both a stipulated clock [30] and aDNA-driven estimation [65].

We have resolved the issue of the anomalously low age of haplogroup M in South Asia by showing that the discrepancy vanishes when we take into account the regional origin of each basal branch. In the west, M dates to 55.3 [45.1; 65.9] ka, overlapping with the founder age of R (Fig. 4). The anomaly is most likely a result of major expansions across the Subcontinent ~45–35 ka: there is an increment in  $N_e$  in M across the Subcontinent ~40 ka, coinciding with the appearance and spread of microlithic technology and greater aridity [67, 68]. The lower age of M is most striking in central India, which is also the centre of gravity of the dramatic radiation of M4'67, which dates to ~40 ka. Microlithic technology can be traced to ~45 ka in central India [82], supporting this region as the likely source of the re-expansion.

**Re-peopling after the Last Glacial Maximum**

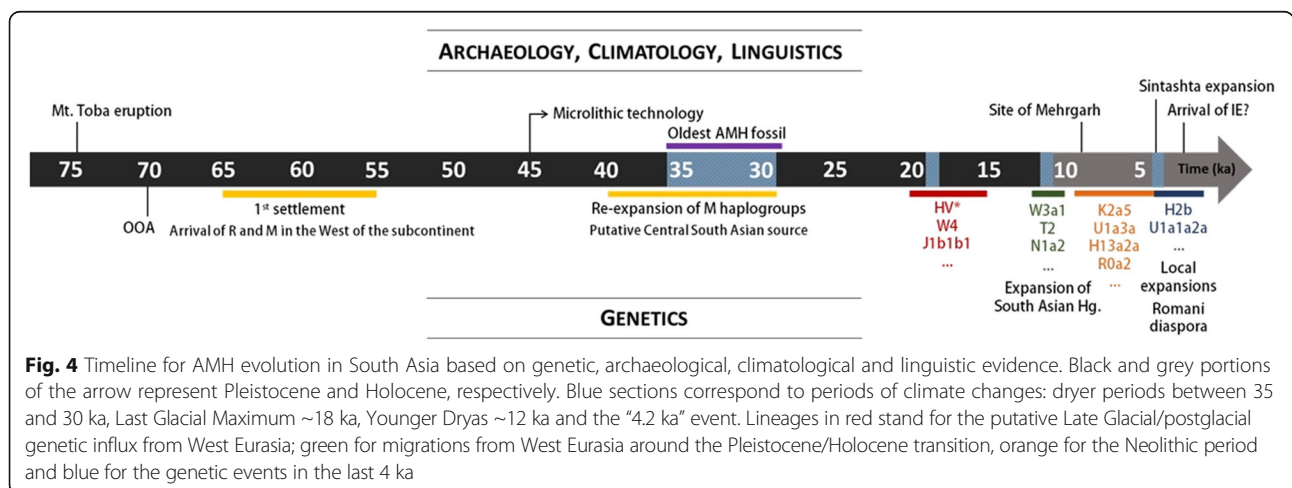
Although South Asia displays a very high level of indigenous variation, the region subsequently received substantial genetic input from both west and east, dramatically re-shaping its genetic structure. Broadly,

South Asian populations are closer to the Caucasus and Central Asian groups rather than to other West Eurasian populations. Pakistanis and Gujaratis in particular carry a preponderance of the “Ancestral North Indian” (ANI) gene pool, contrasting with the ASI or autochthonous population of the Subcontinent [25, 26]. However, our results suggest that this profile is due to multiple dispersals from the north-west, from several distinct sources, rather than just one or two major admixture events in the Neolithic/Bronze Age.

In fact, we see mtDNA lineages from Southwest Asia start to arrive as early as ~20 ka. This was a time of short-lived relative global warmth following the peak of the last glaciation, which might have triggered population movements in several regions [83]. Some lineages arrived in Late Glacial times, again from a Southwest Asian refugium, mirroring the situation in Europe [84]. After ~12 ka, with the end of the Younger Dryas glacial relapse, these movements intensified, with the arrival of yet more Southwest Asian lineages. This period also witnessed the expansion of several autochthonous mtDNA lineages across South Asia, in part from sources in the west (possibly carried alongside dispersing Southwest Asian lineages), but primarily from the south. Supporting this view,  $N_e$  increments at this period are visible in the west and the south, related to the expansion of indigenous M lineages.

**Disentangling Early Neolithic and Bronze Age dispersals into South Asia**

After the first settlement, most attention in genetic studies has been focused on the Neolithic and Bronze Age periods, in part due to potential implications for the spread of Indo-European languages. The earliest Neolithic sites, on the Indus Valley around Mehrgarh in Baluchistan, date to before 9 ka [85, 86], and the earliest crops in South Asia





derived from Southwest Asian founder crops from the Fertile Crescent [19, 87]. Numerous mtDNA lineages entered South Asia in this period from Anatolia, the Caucasus and Iran.

Although some have argued for co-dispersal of the Indo-Aryan languages with the earliest Neolithic from the Fertile Crescent [88, 89], others have argued that, if any language family dispersed with the Neolithic into South Asia, it was more likely to have been the Dravidian family now spoken across much of central and southern India [12]. Moreover, despite a largely imported suite of Near Eastern domesticates, there was also an indigenous component at Mehrgarh, including zebu cattle [85, 86, 90]. The more widely accepted “Steppe hypothesis” [91, 92] for the origins of Indo-European has recently received powerful support from aDNA evidence. Genome-wide, Y-chromosome and mtDNA analyses all suggest Late Neolithic dispersals into Europe, potentially originating amongst Indo-European-speaking Yamnaya pastoralists that arose in the Pontic-Caspian Steppe by ~5 ka, with expansions east and later south into Central Asia in the Bronze Age [53, 76, 93–95]. Given the difficulties with deriving the European Corded Ware directly from the Yamnaya [96], a plausible alternative (yet to be directly tested with genetic evidence) is an earlier Steppe origin amongst Copper Age Khavlyyn, Srednij Stog and Skelya pastoralists, ~7–5.5 ka, with an infiltration of southeast European Chalcolithic Tripolye communities ~6.4 ka, giving rise to both the Corded Ware and Yamnaya when it broke up ~5.4 ka [12].

An influx of such migrants into South Asia would likely have contributed to the CHG component in the GW analysis found across the Subcontinent, as this is seen at a high rate amongst samples from the putative Yamnaya source pool and descendant Central Asian Bronze Age groups. Archaeological evidence suggests that Middle Bronze Age Andronovo descendants of the Early Bronze Age horse-based, pastoralist and chariot-using Sintashta culture, located in the grasslands and river valleys to the east of the Southern Ural Mountains and likely speaking a proto-Indo-Iranian language, probably expanded east and south into Central Asia by ~3.8 ka. Andronovo groups, and potentially Sintashta groups before them, are thought to have infiltrated and dominated the *soma*-using Bactrian Margiana Archaeological Complex (BMAC) in Turkmenistan/northern Afghanistan by 3.5 ka and possibly as early as 4 ka. The BMAC came into contact with the Indus Valley civilisation in Baluchistan from ~4 ka onwards, around the beginning of the Indus Valley decline, with pastoralist dominated groups dispersing further into South Asia by ~3.5 ka, as well as westwards across northern Iran into Syria (which came under the sway of the Indo-Iranian-speaking Mitanni) and Anatolia [12, 95, 97, 98].

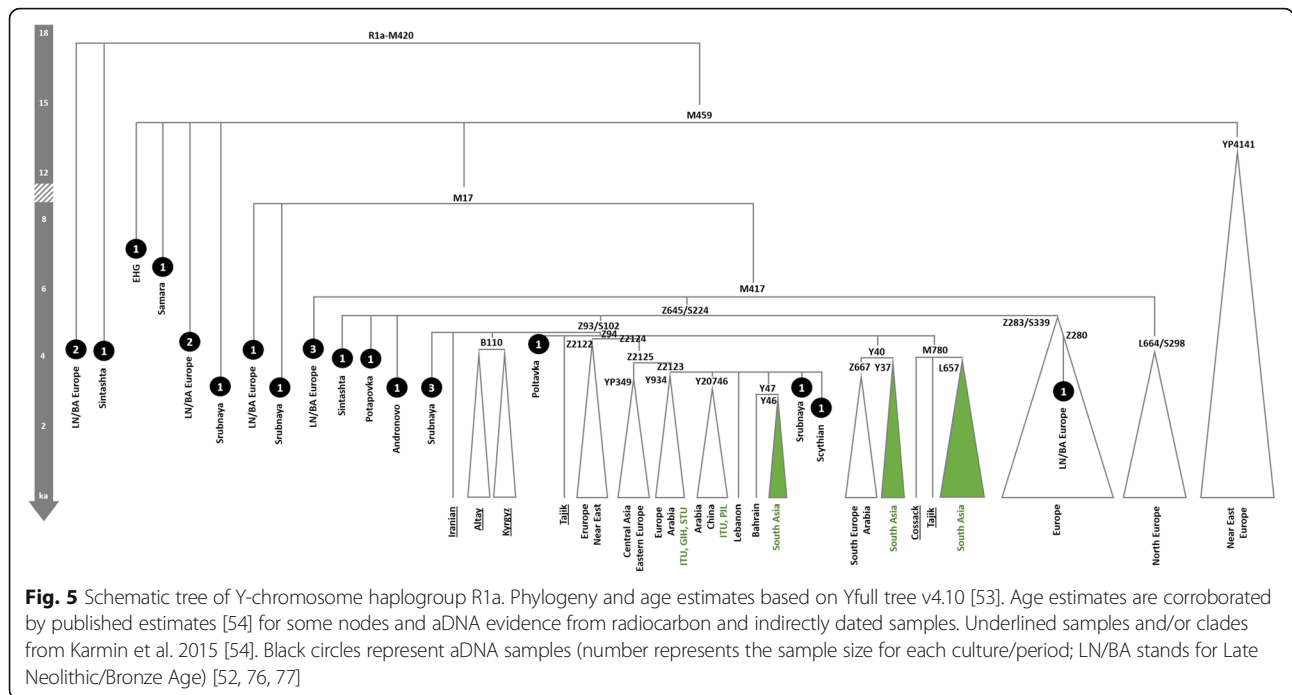
Although GW patterns have been broadly argued to support this view [24], there have also been arguments against. For example, Metspalu et al. [28] argued cogently that the GW pattern in South Asia was the result of a complex series of processes, but they also suggested that an East Asian component, common in extant Central Asians, should be evident in the Subcontinent if it had experienced large-scale Bronze Age immigration from Central Asia. In fact, however, aDNA evidence shows that this element was not present in the relevant source regions in the Early Bronze Age [76]. Moreover, whilst the dating and genealogical resolution of Y-chromosome lineages has been weak until recently, it is now clear that a very large fraction of Y-chromosome variation in South Asia has a recent West Eurasian source.

#### Genetic signals of Indo-European expansions

Contrary to earlier studies [99, 100], recent analyses of Y-chromosome sequence data [55, 58, 94] suggest that haplogroup R1a expanded both west and east across Eurasia during the Late Neolithic/Bronze Age. R1a-M17 (R1a-M198 or R1a1a) accounts for 17.5% of male lineages in Indian data overall, and it displays significantly higher frequencies in Indo-European than in Dravidian speakers [55].

There are now sufficient high-quality Y-chromosome data available (especially Poznik et al. [58]) to be able to draw clear conclusions about the timing and direction of dispersal of R1a (Fig. 5). The indigenous South Asian subclades are too young to signal Early Neolithic dispersals from Iran, and strongly support Bronze Age incursions from Central Asia. The derived R1a-Z93 and the further derived R1a-Z94 subclades harbour the bulk of Central and South Asian R1a lineages [55, 58], as well as including some Russian and European lineages, and have been variously dated to 5.6 [4.0;7.3] ka [55], 4.5–5.3 ka with expansions ~4.0–4.5 ka [58], or 4.7 [4.0;5.5] ka (Yfull tree v4.10 [54]). The South Asian R1a-L657, dated to ~4.2 ka [3.3;5.1] (Yfull tree v4.10 [54]), is the largest (in the 1KG dataset) of several closely related subclades within R1a-Z94 of very similar time depth. Moreover, not only has R1a been found in all Sintashta and Sintashta-derived Andronovo and Srubnaya remains analysed to date at the genome-wide level (nine in total) [76, 77], and been previously identified in a majority of Andronovo (2/3) and post-Andronovo Iron Age (Tagar and Tachtyk: 6/6) male samples from southern central Siberia tested using microsatellite analysis [101], it has also been identified in other remains across Europe and Central Asia ranging from the Mesolithic up until the Iron Age (Fig. 5).

The other major member of haplogroup R in South Asia, R2, shows a strikingly different pattern. It also has deep non-Subcontinental branches, nesting a South Asian specific subclade. But the deep lineages are mainly



seen in the eastern part of the Near East, rather than Central Asia or eastern Europe, and the Subcontinental specific subclade is older, dating to ~8 ka [55].

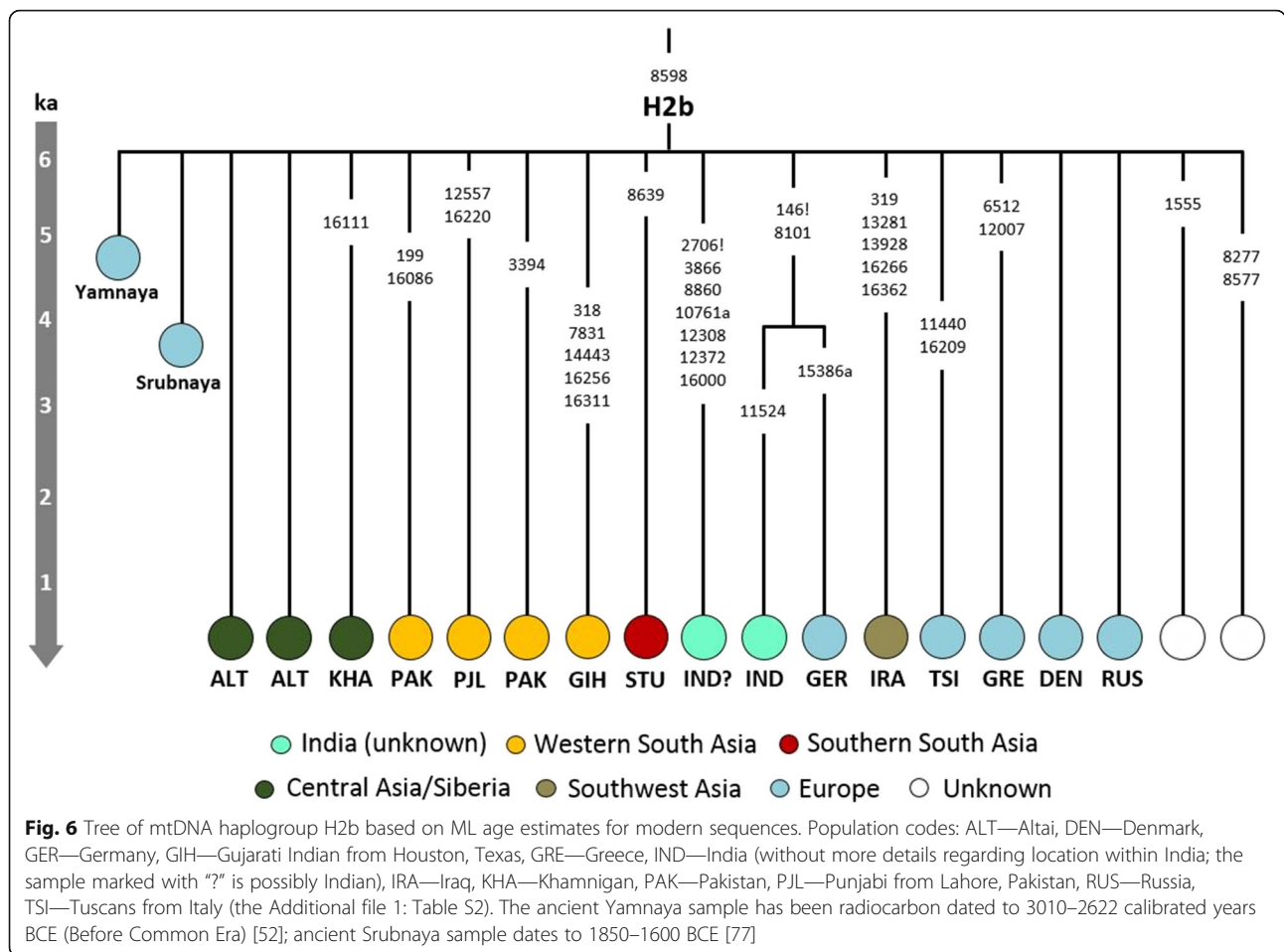
Altogether, therefore, the recently refined Y-chromosome tree strongly suggests that R1a is indeed a highly plausible marker for the long-contested Bronze Age spread of Indo-Aryan speakers into South Asia, although dated aDNA evidence will be needed for a precise estimate of its arrival in various parts of the Subcontinent. aDNA will also be needed to test the hypothesis that there were several streams of Indo-Aryan immigration (each with a different pantheon), for example with the earliest arriving ~3.4 ka and those following the Rigveda several centuries later [12]. Although they are closely related, suggesting they likely spread from a single Central Asian source pool, there do seem to be at least three and probably more R1a founder clades within the Subcontinent [58], consistent with multiple waves of arrival. Genomic Y-chromosome phylogeography is in its infancy compared to mitogenome analysis so it is of course likely that the picture will evolve with sequencing of further South Asian Y-chromosomes, but the picture is already sufficiently clear that we do not expect it to change drastically.

Although these migrations appear to have been male-driven, it might nevertheless be possible to detect a minor maternal signal. For example, haplogroup H2b (dating to 6.2 ka [3.8–8.7] ka; Fig. 6) is a starlike subclade with a probable ultimate ancestry in Eastern Europe, but includes several South Asian lineages (from Pakistan, India and Sri Lanka) that probably arrived more recently from Central

Asia. Tellingly, H2b also includes two aDNA samples (Fig. 6): one individual from the small number of Yamnaya sampled to date [53, 76] and another from the Late Bronze Age Srubnaya culture [77].

Even so, the spread of Indo-European within the Subcontinent seems to have been mainly male-mediated, in agreement with recent X-chromosome analyses [102] and as indicated by the high frequency of West Eurasian (mainly R1a) paternal lineages across the region—varying in the 1KG data from ~25% in the northwest and ~20% in the northeast to ~14% in the south, but much more dramatically when taking caste and language into account (from almost 50% in upper-caste Indo-European speakers to almost zero in eastern Austro-Asiatic speakers) [12, 56, 59]. This present-day distribution cannot be directly correlated with language replacement, however, since the signal is also strong in Dravidian-speaking populations (Fig. 3). The last four millennia witnessed major cultural changes in the Indian Subcontinent, with the decline of the Indus Valley civilisation and the rise of Vedic religion, based on a strict caste system, often associated with the arrival of Indo-Aryan speakers. The mix of autochthonous and immigrant genetic lineages seen across South Asia, however, suggests a gradual merging of male-dominated Andronovo/BMAC immigrants with the indigenous descendants of the Indus Valley civilisation [12], possibly associated with the spread of the Megalithic culture as far south as Sri Lanka in the first century Before Common Era (BCE), prior to the establishment of the full jāti caste system very roughly ~2 ka [12, 103]. Basu et al. [26] date the “freezing” of India’s population structure to ~1.5 ka.





Although the mtDNA does not suggest similar continent-wide dispersals involving women, the last ~4 ka nevertheless witnessed a profound impact on the demography of maternal lineages, with a population increment associated with the indigenous lineages which might have involved local movements and facilitated the diffusion of the Indo-Aryan languages. This expansion is mainly evident amongst the autochthonous lineages in west and central South Asia.

We see no evidence that the caste system emerged in the wake of the arrival of Indo-Aryan speakers from the north, in agreement with formal admixture analyses [24, 26]. Higher-ranking castes do seem closer genetically to Pakistan and ultimately Caucasus and Central Asian populations, but this proximity was most likely established over millennia, by several distinct migratory events—indeed, a sizeable fraction of the non-R1a West Eurasian Y-chromosome lineages (e.g. R2a-M124, J2-M241, L1a-M27, L1c-M357) were most likely associated with the spread of agriculture or even earlier expansions from Southwest Asia, as with the mtDNA lineages [55, 59]. The tribal groups are generally more divergent from other South Asian groups and in particular from western South Asians,

but the particular genetic diversity of tribal groups might have been due to isolation [20], and not necessarily because of more recent strict social boundaries enforced by newly-arriving groups imposing a new system, which in its historical form was likely established much more recently, not more than around 2000 years ago [12, 24, 26, 103].

### Conclusions

#### The trans-continental demographic impact of the Eurasian Bronze Age

In conclusion, analysis of the uniparental marker systems can provide complementary insight into the main genome-wide component that arrived in and spread throughout South Asia since the LGM. This “CHG” component is now known to reach almost 100% in both pre-Neolithic remains from the Caucasus [74] and pre-Neolithic and Early Neolithic remains from Iran [75], and to occur at ~50% in the Pontic-Caspian steppe zone [53, 76], north of the Caucasus, by ~5 ka. This component underwent of multiple dispersals into the Subcontinent, with chronologically distinct sources in the eastern Fertile Crescent and the Steppe, via Central Asia. Moreover, these dispersals involved not simply the

spread of early farming from Southwest Asia and the male-dominated arrival of Indo-Aryan speakers from Central Asia. The mtDNA signal suggests several streams of dispersal into the Subcontinent from the northwest since the LGM, and there were also more recent dispersals from the east, with a more limited impact [22].

In some ways, the overall picture for South Asia resembles the settlement history for a much smaller peninsula on the far side of the Near East with a similar sink status—Europe. Europe too was settled by early modern humans in the late Pleistocene, albeit suffering much greater impact from the LGM due to its latitude. Even so, Europe similarly experienced subsequent settlement episodes from the LGM onwards, culminating in the spread of agriculture from Southwest Asia ~9 ka, followed by the similarly male-dominated spread of pastoralism and, most likely, the Indo-European language family in the Late Neolithic/Early Bronze Age from the Pontic-Caspian steppe [65, 76, 77, 84, 104, 105].

Indeed, Y-chromosome haplogroup R1a, which spread with pastoralism and the Indo-European languages into South Asia, also seems to have been carried into Europe a millennium earlier, alongside a similar pastoral economy and language package and its sibling lineage, R1b [53, 58, 76, 94]. Notably, however, the extent to which the R1 lineages replaced earlier Y chromosomes was much greater across Europe than we see in South Asia. This corresponds to the greater impact of Indo-European languages in Europe, which ultimately left few relicts of earlier language families surviving—the only cases still extant being Basque and Finno-Ugric, with Etruscan and Iberian as well-attested but extinct examples. By comparison, almost a quarter of modern Indians speak the Dravidian languages that seem most likely to have been spread by the first farmers [12].

This greater impact in Europe is also reflected in the genome-wide picture. In Europe, although the CHG component is only 10–15% in most populations, it is thought to have been accompanied by a similar fraction of indigenous Mesolithic European lineages from the steppe, seen in Yamnaya samples [53]. This component does not seem to have spread significantly east and south into Central and South Asia, however [76].

Furthermore, in the case of Europe, the major stages are simpler to disentangle from the genome-wide evidence. This is because the distinctiveness of the Levantine source for the Early Neolithic, compared to the Pontic-Caspian steppe, gives most European populations a clear tripartite ancestry that is less evident in South Asia. In fact, even in Europe the situation may be more complex than it first appeared [80, 105, 106]. In the Subcontinent, the Levantine component is (like the European Mesolithic component) minor, due to a deep east–west separation across the Fertile Crescent prior to the spread of the Neolithic

[75]. As a result, both the Southwest Asian source for the Late Palaeolithic/Early Holocene and the Steppe/Central Asian source for the Bronze Age largely share the same ancestral pool, which may have arisen in the region of the Caucasus and eastern Fertile Crescent and expanded both north and south during the later Neolithic and Early Bronze Age [74, 75, 95].

Consequently, it may be that only a minor fraction of the CHG component represents Indo-Aryan arrivals in South Asia, perhaps helping to explain why Metspalu et al. [28] were unable to detect it. In any case, estimates of the putative ancestral contributions in clustering analyses such as ADMIXTURE vary considerably depending on the data used, as well as being confounded by other factors such as bottlenecks and unsampled source regions, and so need to be treated with considerable caution [107, 108].

However, an attempt to quantify the relative contribution of Iran/Caucasus *versus* the Steppe by formal admixture analyses was recently made by Lazaridis et al. [75], using ancient DNA data to identify Neolithic Iran and the Yamnaya as the most plausible sources. Like Y-chromosome evidence, this analysis has again emphasized a lack of a direct fit with modern languages—for example, the Iranian component contributes predominantly in several sampled populations in the northwest, both Iranian and Dravidian speaking. Even so, in most of the sampled populations, the Steppe contribution was estimated to equal or even exceed the Iranian fraction, in agreement with the picture from uniparental markers presented here.

## Additional files

**Additional file 1: Figure S1.** BSP for haplogroup M in different regions of the Subcontinent: (a) West, (b) South, (c) Central and (d) East South Asia. **Figure S2.** Putative origin and age ranges (95% confidence interval) for non-autochthonous mtDNA lineages found in South Asia. The colours represent the most likely source for each lineage; branches exclusively with South Asian complete sequences coloured in green, whereas branches that also harbour sequences from other regions in white and with green contour. Ages according to ML estimates. **Figure S3.** ADMIXTURE analysis for all *K* values. Information on the populations included in Additional file 1: Table S3. **Figure S4.** sNMF analysis of modern populations for all *K* values. Information on the populations included in Additional file 1: Table S3. **Figure S5.** Cross-validation errors for different values of *K* for ADMIXTURE: (a) considering only modern populations and (b) including the Yamnaya in the analysis. **Figure S6.** ADMIXTURE analysis including the Yamnaya for all *K* values. Information on the populations included in Additional file 1: Table S3. **Figure S7.** PCA (for PC1 and PC3) of modern populations. Detailed information on the populations included in Additional file 1: Table S3. **Figure S8.** PCA (for PC1 and PC2) including the Yamnaya. Information on the populations included in Additional file 1: Table S3. **Table S1.** List of complete mtDNA sequences belonging to South Asian autochthonous haplogroups. **Table S2.** List of non-autochthonous complete modern mtDNA sequences used in our analyses. **Table S3.** Dataset used for the GW analyses. (a) Modern dataset. Populations marked with three asterisks (\*\*\*) were added to the dataset exclusively for ADMIXTURE and sNMF runs, thereby being absent from the PCA. **Table S4.** Putative origin for the uniparental lineages found in the 1KGP South Asian populations. (PDF 7577 kb)

**Additional file 2:** Phylogenetic tree of South Asian mtDNA haplogroups. (XLSX 664 kb)

### Acknowledgements

We thank John Clegg, A. S. M. Sofro and Sean O'Riordan for samples, Asko Parpola and Gyaneshwer Chaubey for critical advice on the manuscript and Hans-Jürgen Bandelt for valuable discussions on Indo-European origins.

### Funding

This work was supported by FEDER funds through COMPETE (FCOMP-01-0124-FEDER-029291) and COMPETE 2020 (Project number 016899) and by national funds through the Portuguese Foundation for Science and Technology (FCT) through the research projects PTDC/IVC-ANT/4917/2012 and PTDC/EPH-ARQ/4164/2014. MS is supported by a Leverhulme Doctoral Scholarship awarded to MBR. AB, MO and TR are supported by FCT grants (SFRH/BD/78990/2011, SFRH/BD/95626/2013 and SFRH/BPD/108126/2015). PS is supported by FCT, ESF, POPH and the FCT Investigator Programme (IF/01641/2013) and acknowledges FCT I.P. and ERDF (through COMPETE2020—POCI) for CBMA's strategic programme UID/BIA/04050/2013 (POCI-01-0145-FEDER-007569). I3S is financed by FEDER funds through the COMPETE 2020—POCI, Portugal2020, and by Portuguese funds through FCT/Ministério da Ciência, Tecnologia e Inovação in the framework of the project "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274).

### Availability of data and materials

The datasets supporting this article have been uploaded as Additional files 1 and 2. Mitochondrial genomes newly sequenced in the study are deposited into GenBank (accession numbers: KY686204-KY686216).

### Authors' contributions

MS, MO, LP, MBR and PS designed the study. MO, AB and TR performed the experiments. MS, MO, DV, JBP, RMF performed the analysis. MS, MO, FG, CE, BH, MP, JK, JFW, LP, MBR and PS wrote and/or critically reviewed the manuscript. All authors read and approved the final manuscript.

### Competing interests

On behalf of all authors, the corresponding authors state that there is no conflict of interest.

### Ethics approval and consent to participate

The 13 samples sequenced were from archival samples collected for earlier projects on Southeast Asia. Informed consent was obtained from all individuals participants included in the study. The research was approved by the University of Huddersfield's School of Applied Sciences Ethics Committee, 26/3/2012, and no additional permissions were necessary for the present study.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Author details

<sup>1</sup>Department of Biological Sciences, School of Applied Sciences, University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK. <sup>2</sup>I3S (Instituto de Investigação e Inovação em Saúde, Universidade do Porto), R. Alfredo Allen 208, 4200-135 Porto, Portugal. <sup>3</sup>IPATIMUP (Instituto de Patologia e Imunologia Molecular da Universidade do Porto), Rua Júlio Amaral de Carvalho 45, 4200-135 Porto, Portugal. <sup>4</sup>Department of Informatics, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal. <sup>5</sup>CBMA (Centre of Molecular and Environmental Biology), Department of Biology, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal. <sup>6</sup>Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal. <sup>7</sup>ICVS/3B's—PT Government Associate Laboratory, Braga/Guimarães, Portugal. <sup>8</sup>Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, Scotland, UK. <sup>9</sup>Synpromics Ltd, Nine Edinburgh Bioquarter, Edinburgh EH16 4UX, UK. <sup>10</sup>Archaeology Department, University of Sydney, Sydney, NSW 2006, Australia. <sup>11</sup>University of Wales Centre for Advanced Welsh and Celtic Studies, National Library of Wales, Aberystwyth, SY23 3HH Wales, UK. <sup>12</sup>MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, Scotland, UK.

Received: 15 December 2016 Accepted: 14 March 2017

Published online: 23 March 2017

### References

- Majumder PP. The human genetic history of South Asia. *Curr Biol*. 2010;20:R184–7.
- Mellars P, Gori KC, Carr M, Soares PA, Richards MB. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. *Proc Natl Acad Sci U S A*. 2013;110:10699–704.
- Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, et al. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science*. 2005;308:1034–6.
- Xing J, Watkins WS, Hu Y, Huff CD, Sabo A, Muzny DM, et al. Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biol*. 2010;11:R113.
- Kennedy K, Deraniyagala S. Fossil remains of 28,000-year-old hominids from Sri Lanka. *Curr Anthropol*. 1989;80:139–44.
- Perera N, Kourampas N, Simpson IA, Deraniyagala SU, Bulbeck D, Kamminga J, et al. People of the ancient rainforest: late Pleistocene foragers at the Batadomba-lena rockshelter, Sri Lanka. *J Hum Evol*. 2011;61:254–69.
- Petraglia M, Korisettar R, Boivin N, Clarkson C, Ditchfield P, Jones S, et al. Middle Paleolithic assemblages from the Indian subcontinent before and after the Toba super-eruption. *Science*. 2007;317:114–6.
- Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, et al. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*. 2016;538:238–42.
- Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. *Nature*. 2016;538:207–14.
- Mallik S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. 2016;538:201–6.
- Witzel M. Autochthonous visions of ancient India. In: Bryant EF, Paton LL, editors. *The Indo-Aryan Controversy: Evidence and Inference in Indian History*. New York: Routledge; 2005. p. 341–404.
- Parpola A. *The Roots of Hinduism: The Early Aryans and the Indus Civilization*. New York: Oxford University Press; 2015.
- Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, et al. Genetic evidence on the origins of Indian caste populations. *Genome Res*. 2001;11:994–1004.
- Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, et al. Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res*. 2003;13:2277–90.
- Kivisild T, Rootsi S, Metspalu M. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet*. 2003;72:313–32.
- Coningham R, Young R. *The Archaeology of South Asia: From the Indus to Asoka, c. 6500 BCE–200 CE*. Cambridge: Cambridge University Press; 2015.
- Palanichamy MG, Mitra B, Zhang C-L, Debnath M, Li G-M, Wang H-W, et al. West Eurasian mtDNA lineages in India: an insight into the spread of the Dravidian language and the origins of the caste system. *Hum Genet*. 2015;134:637–47.
- Sengupta S, Zhivotovskiy LA, King R, Mehdi SQ, Edmonds CA, Chow C-ET, et al. 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. *Am J Hum Genet*. 2006;78:202–21.
- Fuller DQ. Non-human genetics, agricultural origins and historical linguistics in South Asia. In: Petraglia MD, Allchin B, editors. *The Evolution and History of Human Populations in South Asia*. Dordrecht: Springer Netherlands; 2007. p. 393–443.
- Chaubey G, Metspalu M, Kivisild T, Vilems R. Peopling of South Asia: investigating the caste-tribe continuum in India. *BioEssays*. 2007;29:91–100.
- Maruyama S, Nohira-Koike C, Minaguchi K, Nambiar P. MtDNA control region sequence polymorphisms and phylogenetic analysis of Malay population living in or around Kuala Lumpur in Malaysia. *Int J Legal Med*. 2010;124:165–70.
- Chaubey G, Metspalu M, Choi Y, Mägi R, Romero IG, Soares P, et al. Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. *Mol Biol Evol*. 2011;28:1013–24.
- Juyal G, Mondal M, Luisi P, Laayouni H, Sood A, Midha V, et al. Population and genomic lessons from genetic analysis of two Indian populations. *Hum Genet*. 2014;133:1273–87.

24. Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh P-R, Govindaraj P, et al. Genetic evidence for recent population mixture in India. *Am J Hum Genet.* 2013;93:422–38.
25. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. *Nature.* 2009;461:489–94.
26. Basu A, Sarkar-Roy N, Majumder PP. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proc Natl Acad Sci.* 2016;113:1594–9.
27. Novembre J, Ramachandran S. Perspectives on human population structure at the cusp of the sequencing era. *Annu Rev Genomics Hum Genet.* 2011;12:245–74.
28. Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, et al. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J Hum Genet.* 2011;89:731–44.
29. Fernandes V, Triska P, Pereira JB, Alshamali F, Rito T, Machado A, et al. Genetic stratigraphy of key demographic events in Arabia. *PLoS One.* 2015;10:e0118625.
30. Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, et al. Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet.* 2009;84:740–59.
31. Palanichamy MG, Sun C, Agrawal S, Bandelt H-J, Kong Q-P, Khan F, et al. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. *Am J Hum Genet.* 2004;75:966–78.
32. Sun C, Kong Q-P, Palanichamy MG, Agrawal S, Bandelt H-J, Yao Y-G, et al. The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. *Mol Biol Evol.* 2006;23:683–90.
33. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature.* 2015;526:75–81.
34. Cann HM, de Toma C, Cazes L, Legrand M-F, Morel V, Piouffre L, et al. A human genome diversity cell line panel. *Science.* 2002;296:261–2.
35. Marrero P, Abu-Amero KK, Larruga JM, Cabrera VM, Cann R, Stoneking M, et al. Carriers of human mitochondrial DNA macrohaplogroup M colonized India from southeastern Asia. *BMC Evol Biol.* 2016;16:246.
36. Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, et al. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum Mutat.* 2011;32:25–32.
37. van Oven M. PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Sci Int Genet.* 2015;5:e392–4.
38. Bandelt H-J, Forster P, Sykes BC, Richards MB. Mitochondrial portraits of human populations using median networks. *Genetics.* 1995;141:743–53.
39. Forster P, Harding R, Torroni A, Bandelt HJ. Origin and evolution of Native American mtDNA variation: a reappraisal. *Am J Hum Genet.* 1996;59:935–45.
40. Saillard J, Forster P, Lynnerup N, Nørby S. mtDNA variation among Greenland Eskimos: the edge of the Beringian Expansion. *Am J Hum Genet.* 2000;1:718–26.
41. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol.* 1981;17:368–76.
42. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 2005;22:1185–92.
43. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 2012;29:1969–73.
44. Atkinson QD, Gray RD, Drummond AJ. Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. *Proc Biol Sci.* 2009;276:367–73.
45. Rito T, Richards MB, Fernandes V, Alshamali F, Cerny V, Pereira L, et al. The first modern human dispersals across Africa. *PLoS One.* 2013;8:e80031.
46. Costa MD, Pereira JB, Pala M, Fernandes V, Olivieri A, Achilli A, et al. A substantial prehistoric European ancestry amongst Ashkenazi maternal lineages. *Nat Commun.* 2013;4:2543.
47. Soares PA, Trejaut JA, Rito T, Cavadas B, Hill C, Eng KK, et al. Resolving the ancestry of Austronesian-speaking populations. *Hum Genet.* 2016;135:309–26.
48. Pereira L, Silva NM, Franco-Duarte R, Fernandes V, Pereira JB, Costa MD, et al. Population expansion in the North African late Pleistocene signalled by mitochondrial DNA haplogroup U6. *BMC Evol Biol.* 2010;10:390.
49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559–75.
50. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet.* 2006;2:e190.
51. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 2009;19:1655–64.
52. Fricot E, Mathieu F, Trouillon T, Bouchard G, François O. Fast and efficient estimation of individual ancestry coefficients. *Genetics.* 2014;196(4):973–83.
53. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature.* 2015;522(7555):207–11.
54. Adamov D, Guryanov V, Karzhavin S, Tagankin V, Urasin V. Defining a new rate constant for Y-chromosome SNPs based on full sequencing data. *Russ J Genet Geneal (Русская версия).* 2015;7:68–89.
55. Karmin M, Saag L, Vicente M, Wilson Sayres MA, Järve M, Talas UG, et al. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res.* 2015;25:459–66.
56. Trivedi R, Sahoo S, Singh A, Bindu GH, Banerjee J, Manuj T, et al. Genetic imprints of pleistocene origin of Indian populations: a comprehensive Phylogeographic sketch of Indian Y-chromosomes. *Int J Hum Genet.* 2008;8:97–118.
57. Zhao Z, Khan F, Borkar M, Herrera R, Agrawal S. Presence of three different paternal lineages among North Indians: A study of 560 Y chromosomes. *Ann Hum Biol.* 2009;36:46–59.
58. Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, et al. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat Genet.* 2016;48:593–9.
59. Sahoo S, Singh A, Himabindu G, Banerjee J, Sitalaximi T, Gaikwad S, et al. A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proc Natl Acad Sci U S A.* 2006;103:843–8.
60. Zhong H, Shi H, Qi X-B, Xiao C-J, Jin L, Ma RZ, et al. Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. *J Hum Genet.* 2010;55:428–35.
61. Yan S, Wang C-C, Li H, Li S-L, Jin L. An updated tree of Y-chromosome Haplogroup O and revised phylogenetic positions of mutations P164 and PK4. *Eur J Hum Genet.* 2011;19:1013–5.
62. Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli E-L, Silva NM, et al. A “Copernican” reassessment of the human mitochondrial DNA tree from its root. *Am J Hum Genet.* 2012;90:675–84.
63. Silva M, Alshamali F, Silva P, Carrilho C, Mandlate F, Jesus Trovoada M, et al. 60,000 years of interactions between Central and Eastern Africa documented by major African mitochondrial haplogroup L2. *Sci Rep.* 2015;5:12526.
64. Soares P, Abrantes D, Rito T, Thomson N, Radivojac P, Li B, et al. Evaluating purifying selection in the mitochondrial DNA of various mammalian species. *PLoS One.* 2013;8:e58993.
65. Posth C, Renaud G, Mittnik A, Drucker DG, Rougier H, Cupillard C, et al. Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a Late Glacial population turnover in Europe. *Curr Biol.* 2016;26:1–7.
66. Kong Q-P, Sun C, Wang H-W, Zhao M, Wang W-Z, Zhong L, et al. Large-scale mtDNA screening reveals a surprising matrilineal complexity in east Asia and its implications to the peopling of the region. *Mol Biol Evol.* 2011;28:513–22.
67. Petraglia M, Clarkson C, Boivin N, Haslam M, Korisettar R, Chaubey G, et al. Population increase and environmental deterioration correspond with microlithic innovations in South Asia ca. 35,000 years ago. *Proc Natl Acad Sci U S A.* 2009;106:12261–6.
68. Clarkson C, Petraglia M, Korisettar R, Haslam M, Boivin N, Crowther A, et al. The oldest and longest enduring microlithic sequence in India: 35 000 years of modern human occupation and change at the Jwalapuram Locality 9 rockshelter. *Antiquity.* 2009;83:326–48.
69. Fernandes V, Alshamali F, Alves M, Costa MD, Pereira JB, Silva NM, et al. The Arabian cradle: mitochondrial relicts of the first steps along the southern route out of Africa. *Am J Hum Genet.* 2012;90:347–55.
70. Gómez-Carballa A, Pardo-Seco J, Fachal L, Vega A, Cebery M, Martínón-Torres N, et al. Indian signatures in the westernmost edge of the European Romani diaspora: new insight from mitogenomes. *PLoS One.* 2013;8:e75397.
71. Rai N, Chaubey G, Tamang R, Pathak AK, Singh VK, Karmin M, et al. The phylogeography of Y-chromosome haplogroup H1a1a-M82 reveals the



- likely Indian origin of the European Romani populations. *PLoS One*. 2012;7:e48477.
72. Rosenberg NA, Mahajan S, Gonzalez-Quevedo C, Blum MGB, Nino-Rosales L, Niniš V, et al. Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet*. 2006;2:e215.
  73. Ayub Q, Mezzavilla M, Pagani L, Haber M, Mohyuddin A, Khaliq S, et al. The Kalash genetic isolate: ancient divergence, drift, and selection. *Am J Hum Genet*. 2015;96:775–83.
  74. Jones ER, Gonzalez-Forbes G, Connell S, Siska V, Eriksson A, Martiniano R, et al. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat Commun*. 2015;6:8912.
  75. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights into the origin of farming in the ancient Near East. *Nature*. 2016;536:419–24.
  76. Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of Bronze Age Eurasia. *Nature*. 2015;522:167–72.
  77. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*. 2015;528:499–503.
  78. Eaaswarkhanth M, Haque I, Ravesh Z, Romero IG, Meganathan PR, Dubey B, et al. Traces of sub-Saharan and Middle Eastern lineages in Indian Muslim populations. *Eur J Hum Genet*. 2010;18:354–63.
  79. Yunusbayev B, Metspalu M, Metspalu E, Valeev A, Litvinov S, Valiev R, et al. The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet*. 2015;11:e1005068.
  80. Richards MB, Soares P, Torroni A. Palaeogenomics: mitogenomes and migrations in Europe's past. *Curr Biol*. 2016;26:R243–6.
  81. Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, et al. The mtDNA legacy of the Levantine Early Upper Palaeolithic in Africa. *Science*. 2006;314(5806):1767–70.
  82. Mishra S, Chauhan N, Singhvi AK. Continuity of microblade technology in the Indian Subcontinent since 45 ka: implications for the dispersal of modern humans. *PLoS One*. 2013;8:e69280.
  83. Perry CA, Hsu KJ. Geophysical, archaeological, and historical evidence support a solar-output model for climate change. *Proc Natl Acad Sci U S A*. 2000;97:12433–8.
  84. Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, Reidla M, et al. Mitochondrial DNA signals of Late Glacial recolonization of Europe from Near Eastern refugia. *Am J Hum Genet*. 2012;90:915–24.
  85. Jarrige J-F, Jarrige C. Premiers pasteurs et agriculteurs dans le sous-continent Indo-Pakistanaï. *Comptes Rendus Palevol*. 2006;5:463–72.
  86. Petrie CA. Case Study: Mehrgarh. In: Barker CA, Goucher C, editors. *The Cambridge World History: Volume 2, A World with Agriculture, 12,000 BCE–500 CE*. Cambridge: Cambridge University Press; 2015. p. 289–309.
  87. Kingwell-Banham E, Petrie CA, Fuller DQ. Early agriculture in South Asia. In: Barker CA, Goucher C, editors. *The Cambridge World History: Volume 2, A World with Agriculture, 12,000 BCE–500 CE*. Cambridge: Cambridge University Press; 2015. p. 261–88.
  88. Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, et al. Mapping the origins and expansion of the Indo-European language family. *Science*. 2012;337:957–60.
  89. Renfrew C. *Archaeology and Language: The Puzzle of Indo-European Origins*. London: Jonathan Cape; 1987.
  90. Chen S, Lin B-Z, Baig M, Mitra B, Lopes RJ, Santos AM, et al. Zebu cattle are an exclusive legacy of the South Asia neolithic. *Mol Biol Evol*. 2010;27:1–6.
  91. Pereltsvaig A, Lewis MW. *The Indo-European Controversy*. Cambridge: Cambridge University Press; 2015.
  92. Anthony DW, Ringe D. The Indo-European homeland from linguistic and archaeological perspectives. *Annu Rev Linguist*. 2015;1:199–219.
  93. Brandt G, Haak W, Adler CJ, Roth C, Szecsenyi-Nagy A, Karimnia S, et al. Ancient DNA reveals key stages in the formation of Central European mitochondrial genetic diversity. *Science*. 2013;342:257–61.
  94. Batini C, Hallast P, Zadić D, Delser PM, Benazzo A, Ghirotto S, et al. Large-scale recent expansion of European patrilineages shown by population resequencing. *Nat Commun*. 2015;6:7152.
  95. Anthony DW. *The horse, the wheel, and language: how Bronze-Age riders from the Eurasian steppes shaped the modern world*. New Jersey: Princeton University Press; 2010.
  96. Mallory JP. *In Search of the Indo-Europeans: Language, Archaeology and Myth*. London: Thames & Hudson; 1989.
  97. Kristiansen K. The decline of the Neolithic and the rise of Bronze Age society. In: Fowler C, Harding J, Hofmann D, editors. *The Oxford Handbook of Neolithic Europe*. Oxford: Oxford University Press; 2015.
  98. Anthony DW, Bogucki P, Comşa E, Gimbutas M, Jovanović B, Mallory JP, et al. The "Kurgan Culture", Indo-European Origins, and the domestication of the horse: a reconsideration. *Curr Anthropol*. 1986;27:291–313.
  99. Underhill PA, Myres NM, Rootsi S, Metspalu M, Zhivotovskiy LA, King RJ, et al. Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. *Eur J Hum Genet*. 2010;18:479–84.
  100. Wilson JF, Weiss DA, Richards M, Thomas MG, Bradman N, Goldstein DB. Genetic evidence for different male and female roles during cultural transitions in the British Isles. *Proc Natl Acad Sci U S A*. 2001;98:5078–83.
  101. Keyser C, Bouakaze C, Crubézy E, Nikolaev VG, Montagnon D, Reis T, et al. Ancient DNA provides new insights into the history of south Siberian Kurgan people. *Hum Genet*. 2009;126:395–410.
  102. Goldberg A, Günther T, Rosenberg NA, Jakobsson M. Ancient X chromosomes reveal contrasting sex bias in Neolithic and Bronze Age Eurasian migrations. *Proc Natl Acad Sci U S A*. 2017;114:2657–2662.
  103. Dumont L. *Homo Hierarchicus: The Caste System and Its Implications*. Chicago: University of Chicago Press; 1981.
  104. Haak W, Forster P, Bramanti B, Matsumura S, Brandt G, Tänzer M, et al. Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science*. 2005;310:1016–8.
  105. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, et al. The genetic history of Ice Age Europe. *Nature*. 2016;534:200–5.
  106. Pala M, Gyaneshwer C, Soares P, Richards MB. *The Archaeogenetics of European Ancestry*. *Encycl Life Sci*. Chichester: John Wiley & Sons, Ltd 2014.
  107. van Dorp L, Balding D, Myers S, Pagani L, Tyler-Smith C, Bekele E, et al. Evidence for a common origin of blacksmiths and cultivators in the Ethiopian Ari within the last 4500 years: lessons for clustering-based inference. *PLoS Genet*. 2015;11:e1005397. Di Rienzo A, editor.
  108. Falush D, van Dorp L, Lawson D. A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots. *bioRxiv*. 2016. <https://doi.org/10.1101/066431>.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

