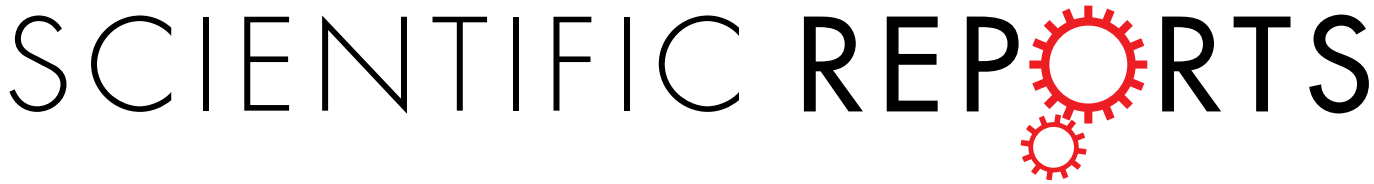

OPEN

Y-chromosome diversity suggests southern origin and Paleolithic backwave migration of Austro-Asiatic speakers from eastern Asia to the Indian subcontinent

Received: 18 May 2015
Accepted: 28 September 2015
Published: 20 October 2015

Xiaoming Zhang[1,*], Shiyu Liao[2,*], Xuebin Qi[1,*], Jiewei Liu[1,8], Jatupol Kampuansai[5], Hui Zhang[1], Zhaohui Yang[3,4], Bun Serey[6], Tuot Sovannary[6], Long Bunnath[6], Hong Seang Aun[6], Ham Samnom[7], Daoroong Kangwanpong[5], Hong Shi[3,4] & Bing Su[1,4]

Analyses of an Asian-specific Y-chromosome lineage (O2a1-M95)—the dominant paternal lineage in Austro-Asiatic (AA) speaking populations, who are found on both sides of the Bay of Bengal—led to two competing hypothesis of this group's geographic origin and migratory routes. One hypothesis posits the origin of the AA speakers in India and an eastward dispersal to Southeast Asia, while the other places an origin in Southeast Asia with westward dispersal to India. Here, we collected samples of AA-speaking populations from mainland Southeast Asia (MSEA) and southern China, and genotyped 16 Y-STRs of 343 males who belong to the O2a1-M95 lineage. Combining our samples with previous data, we analyzed both the Y-chromosome and mtDNA diversities. We generated a comprehensive picture of the O2a1-M95 lineage in Asia. We demonstrated that the O2a1-M95 lineage originated in the southern East Asia among the Daic-speaking populations ~20–40 thousand years ago and then dispersed southward to Southeast Asia after the Last Glacial Maximum before moving westward to the Indian subcontinent. This migration resulted in the current distribution of this Y-chromosome lineage in the AA-speaking populations. Further analysis of mtDNA diversity showed a different pattern, supporting a previously proposed sex-biased admixture of the AA-speaking populations in India.

There is a broad consensus that modern humans originated in Africa and then migrated to Asia along a coastal route by way of the Indian subcontinent as early as 60 thousand years ago (KYA)[1–7]. However, the later dispersion of this ancestral population across Asia is far less clear. Linguistic analyses have grouped Asian populations across eight language families in eastern Asia and South Asia: Altaic, Sino-Tibetan

[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China. [2]School of Life Sciences, Anhui University, Hefei 230039, China. [3]Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming 650500, China. [4]Yunnan Key Laboratory of Primate Biomedical Research, Kunming 650500, China. [5]Department of Biology, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand. [6]Department of Geography and Land Management, Royal University of Phnom Penh, Phnom Penh 12000, Cambodia. [7]Capacity Development Facilitator for Handicap International Federation and Freelance Research, Battambang 02358, Cambodia. [8]Kunming College of Life Science, University of Chinese Academy of Sciences, Beijing 100101, China. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to H.S. (email: shih@kmust.edu.cn) or B.Su. (email: sub@mail.kiz.ac.cn)

(ST, split into Han and Tibeto-Burman (TB) sub-branches), Daic, Hmong-Mien (HM), Austro-Asiatic (AA), Austronesian (AU), Dravidian (DR) and Indo-European (IE). With wide distribution in mainland China and Siberia, both Altaic and ST form two northern language families, DR and IE comprise the two main language families of the Indian subcontinent, while Daic, HM, AA and AU make up the southern language families that are primarily distributed in southern China and Southeast Asia.

Trying to use linguistic families to map out the origin and migration patterns of human populations in Asia has resulted in far less consensus. For example, of the southern language families, AA has a somewhat unique geographic distribution, with a wide distribution not only in southern China and Southeast Asia, but also in India. Subsequently, AA is the eighth largest language family in the world in terms of population size (104 millions)[8] with two major branches: Munda in eastern, northeastern and central India and Mon-Khmer, which stretches from northeastern India to the Andaman-Nicobar islands, Malay Peninsula and vast Mekong delta in MSEA. AA is the first language of many ethnic groups in Cambodia, Vietnam, Laos, Thailand, Burma and Malaysia, and serves as the main official language in Cambodia and Vietnam. Taking these realities into account, decades of research has resulted in long-standing debate about the geographic origin and prehistoric migratory route of the AA-speaking populations.

Similarly, analysis of genetic data to characterize the origin and migration history of AA-speaking populations has led to two rival hypotheses[9–15]. Data from the maternal lineage (mtDNA) makes a clear distinction between Munda-speakers in India and Mon-Khmer speakers in Southeast Asia, with a lack of shared mtDNA haplogroups[9,15–17]. By contrast, data from the paternal lineage (Y-chromosome) indicates a shared Asian-specific haplogroup (O2a1-M95) between the AA speakers from India (66.44% on average) and from Southeast Asia (56.55% on average)[9,10,12,13,18]. Given the relatively young age ($<$10 KYA) of the O2a1-M95 lineage estimated from the Y-chromosome short tandem repeats (Y-STRs) variation in India, the migratory route of the AA speakers would likely begin in Southeast Asia and then move to India[11,12]. However, the high mtDNA haplotype diversity in Munda-speaking populations[14,15] and an independent estimate of an old coalescence age (~65 KYA) of the O2a1-M95 lineage in the Indian AA-speaking populations[10] suggests an Indian origin followed by a dispersal to Southeast Asia, possibly before the Last Glacial Maximum (LGM, 19.0–26.5 KYA)[19]. This latter hypothesis seems to cope better with the more widely agreed upon costal migration of modern humans from Africa to Asia by way of the Indian subcontinent.

While both theories have certain peculiar merits, neither has dealt well with the large discrepancy of the estimated ages of the O2a1-M95 lineage from different studies. One explanation for the marked differences in the estimate may be limited samplings of the AA speakers in India and/or different genotyping approaches[10,12]. Fortunately, a recent study with a more extensive sampling of the AA speakers in India and a few samples from Southeast Asia[9] has clarified some of these inconsistencies. Through a genome-wide screening of 610K autosomal sequence variations and uniparental loci, Chaubey *et al.* demonstrated an older coalescent time (average 22.4 ± 4.9 KYA) of the O2a1-M95 lineage in Southeast Asia than that in India (average 15.9 ± 1.6 KYA), lending greater credence to the proposed westward migration of the AA speakers from Southeast Asia to India. Chaubey *et al.* also proposed a sex-specific admixture of the AA-speaking immigrants with local India populations by showing a different pattern in the mtDNA lineage[9].

Despite the data contributions from Chaubey *et al.* and numerous other studies on AA speakers, AA populations from MSEA and southern China continue to be under- sampled and represented. Similarly, no other southern populations have been included in these analysis to date, in spite of the high frequency of O2a1-M95 in certain populations, such as among Daic-speaking populations that have a ~45% frequency[20–23]. Complicating these oversights, existing genomic analysis also suffers from some deficiencies. For example, the Illumina Human Hap 610K Chips were developed by covering sequence variations identified in limited world populations, which in turn limits its power to detect genetic relationships among the hypothetically ancient AA populations. Given the sampling, methodology and technical limitations inherent in the existing literature, basic questions—where did the O2a1-M95 carrying AA-speakers originally emerge, or when did it begin expanding into Asia—remain unanswered.

In this study, we collected samples of 21 AA-speaking populations from Cambodia, Thailand and southern China (totally, 646 males)(Fig. 1). For individuals belonging to the O2a1-M95 lineage (343 of the 646)[18], we conducted genotyping of 16 Y-STRs. We also collected published data of the O2a1-M95 carriers from 107 populations (2,510 O2a1-M95 out of 7,693 male individuals in total) covering all the geographic distributions of the AA speakers as well as the other major language families in eastern Asia and India. To date, this data marks the most comprehensive collection of data of O2a1-M95 diversity. Our analysis showed that the O2a1-M95 lineage initially originated in the southern part of eastern Asia among the Daic-speaking populations around 20–40 KYA, followed by a southward dispersal to the heartland of MSEA ~16 KYA, and then a westward migration to India ~ 10 KYA. Furthermore, analysis of more than 20,000 mtDNA sequences, including these AA populations and other Asian populations, demonstrated that the maternal lineage has a different pattern from the Y-chromosome for these AA populations, supporting the proposed sex-biased admixture of the AA immigrants with local people in the Indian subcontinent.

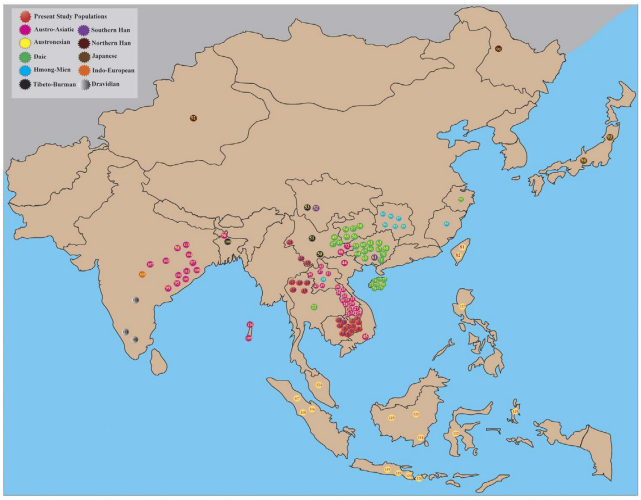

Figure 1. Geographic locations of the studied populations in Asia that contain the O2a1-M95 lineage. Populations are color-coded based on their language families. The figure was modified from our previous report[43] using Microsoft Powerpoint 2011 (Microsoft Corporation, USA).

Results

High O2a1-M95 frequencies in the AA populations from MSEA and southern China. The O2a1-M95 lineage was reported to be highly prevalent in some AA populations in India, e.g., as high as 67.53% and 74.00% respectively in Munda and Mon-Khmer populations[9,10]. We observed high O2a1-M95 frequencies in AA populations not analyzed in previous studies from Cambodia (70.67%), Thailand (52.51%) and Southern China (30.00%) (Fig. 2A, Table 1 and supplementary Table S2)[10–12,15]. In the Andaman-Nicobar islands, O2a1-M95 was also widespread (~45.18% on average) and is fixed (100%) in several populations, such as the Shompen and Onge[9,10], likely due to a strong bottleneck effect in these island populations, which is reflected in other major Y-chromosome lineages (e.g. DE-YAP and O3-M22)[24–26]. (Fig. 2A and supplementary Table S2). Consistent with previous results, the collective data shows that O2a1-M95 lineage is dominant in almost all AA populations, including those from MSEA and southern China, making it an informative genetic marker for tracing the patrilineal prehistory of the AA populations.

Dating the O2a1-M95 lineages of different Asian populations based on Y-STRs variations. Previous studies have sampled few AA populations from MSEA and Southern China[9,10,12]. To fill the sampling gap, we sampled a wide range of AA-speaking populations from Cambodia, Thailand and southern China[18] and genotyped 16 Y-STRs loci for those samples belonging to the O2a1-M95 lineage (Fig. 1, Table 1 and supplementary Table S1). Integrating these samples with the previous data, we dated the O2a1-M95 lineages among different regional populations (Fig. 3, supplementary Tables S3 and S4) and observed that the O2a1-M95 lineage has the oldest time of most recent common ancestor (TMRCA) among the populations in the southern part of mainland China and Taiwan (~20–40 KYA), most of which are Daic speaking (Fig. 3, Supplementary Tables S3 and S4). The average TMRCA for these Daic and Austronesian populations from southern China is ~30 KYA, markedly older than those in MSEA (~16 KYA), India (~10 KYA) or Island Southeast Asia (ISEA, ~11 KYA) (Fig. 3, supplementary Tables S3 and S4). The estimated coalescence ages for the AA speakers from MSEA, ISEA and India are similar to those reported by Chaubey *et al.*[9]. At the same time, the estimated ages of O2a1-M95 lineages in the Daic populations was consistent with the estimated ages of its sister lineages (O3-M122 and C-M130) in the same geographic regions[3,27], supporting the proposed antiquity of the Daic populations. These lines of evidence suggest that the O2a1-M95 lineage initially originated in the Daic populations living in southern China, prior to a southward expansion to MSEA and later migrations to India and ISEA after the LGM (19.0–26.5 KYA)[19].

Comparison of haplotype diversity of the O2a1-M95 lineages among different geographic populations. In line with the estimated TMRCAs, the unbiased Y-STRs haplotype diversity of the O2a1-M95 lineage are the highest in populations from southern China (~0.5017 on average), particularly among the Daic populations, followed by those in MSEA (~0.3858), ISEA (~0.3680) and then India (~0.3168) (Fig. 2B and supplementary Table S5), which together match the proposed migratory routes from southern China to MSEA, and then to ISEA and India. We further calculated the pairwise genetic distances measured by *Fst* (supplementary Table S6) and constructed an un-rooted neighbor-joining (NJ) tree based on the Y-STRs variations that showed populations clustered primarily along their respective language families and not by geographic regions. This tree structure suggests a within language family

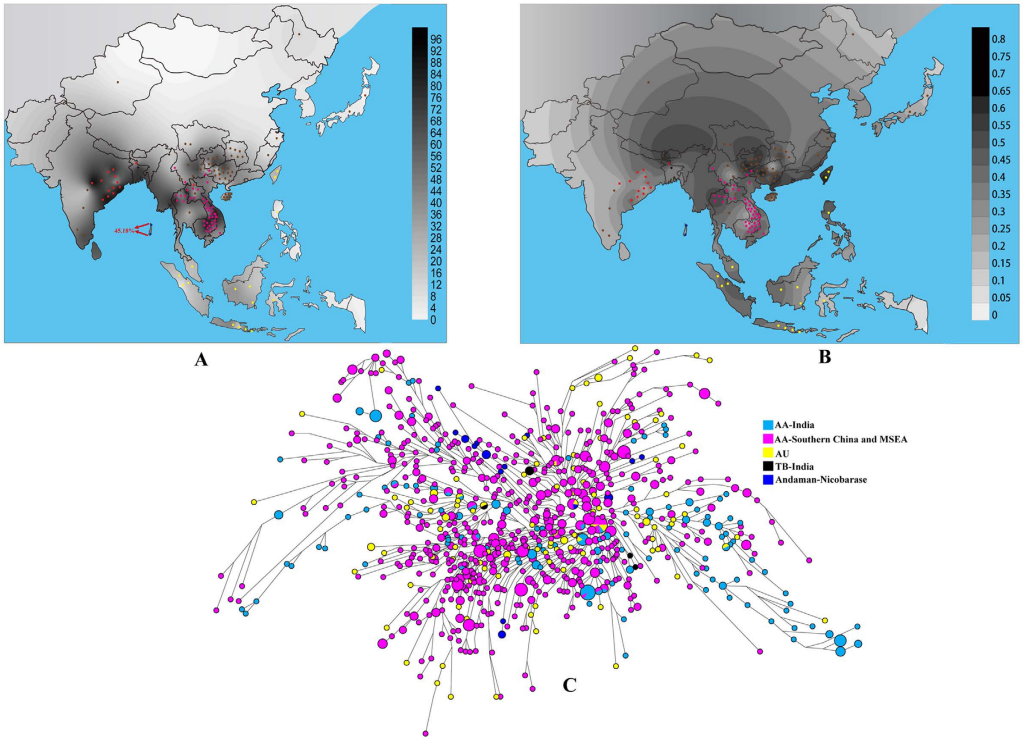

Figure 2. Frequency distribution, Uh diversity and phylogenetic structure of the O2a1-M95 lineages among Asian populations. Contour map shows the frequency **(A)** and Y-STRs Uh diversity **(B)** of lineage O2a1-M95 in Asia. Colored dots indicate the geographic locations of the analysed populations that correspond with Fig. 1; Bars indicate the frequency and Uh diversity spectrum respectively. (**C**) Phylogenetic network of Y-STRs haplotypes among O2a1-M95 populations generated from the following 14 Y-STRs: DYS19, DYS389 I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS458, DYS635 and GATA H4; Circles size is proportional to the number of samples. The contour maps were generated using Surfer10 (Golden Software Inc., Golden, USA), and the network was constructed using the Network package 4.6.1.3 (www.fluxus-engineering.com).

genetic affinity, though there were several interesting exceptions (Fig. 4). The AA populations from India clustered with the AA populations from Cambodia, not with the Dravidian and Indo-European speakers from India. This grouping strongly supports the hypothesized shared genetic ancestry among the AA populations, consistent with the previous observation by Chaubey *et al.*[9]. We also observed a lack of clear geographic clustering in the Y-STRs based phylogenetic network of O2a1-M95 (Fig. 2C), likely due to continuous gene flows among the regional AA speakers[9].

Interestingly, our analysis departed from several previous observations that had found a clear divergence of the Andaman-Nicobar Island populations from the other AA speakers[9]. Here, we detected shared Y-STRs haplotypes in these isolated island populations with some MSEA populations (Fig. 2C), which may not have been apparent in earlier studies due to a generalized under representation of MSEA populations.

mtDNA diversity suggests a sex-biased migration of AA-speakers from MSEA to the Indian subcontinent. To check the maternal side of the AA populations, we collected 21,470 mtDNA sequences from 545 populations distributed in East Asia, Southeast Asia and South Asia (supplementary Table S7) and analyzed the patterns of mtDNA diversity. Compared with the dominant occurrence of the O2a1-M95 lineage (65.53% on average) and the high frequency (e.g., 44.57%) of other East Asian specific lineages (NO, N, O, P and Q, supplementary Table S8) in the South Asia populations, we found only ~16.46% mtDNA sequences belonging to the East Asian specific lineages (A, B, C, D, F, G, M9 and M12) in South Asia (supplementary Figures S1, S2 and supplementary Table S9). PCA analysis using mtDNA haplotype frequencies indicated a clustering pattern of geographic locations and not language families, which is different from that of the Y-chromosome data. For example, AA and TB populations from India clustered with Dravidian and Indo-European populations from India and not the other AA populations from southern China and Southeast Asia (Fig. 5). This discrepancy supports the notion that the prehistoric migration of the AA-speakers from MSEA to India was likely sex-biased, confirming the hypothetical sex-biased admixture of the India AA populations posited by Chaubey *et al.*[9].

NO.	Population	Region	Location	Linguistic Family	Sub-Branch	N	O2a1-M95 Counts	%
1	Brao	Cambodia	Ratanakri	Austro-Asiatic	West Bahnaric	37	24	64.86
2	Jarai	Cambodia	Ratanakri	Austronesian	Chamic	45	34	75.56
3	Kachac	Cambodia	Ratanakri	Austro-Asiatic	North Bahnaric	17	13	76.47
4	Khmer	Cambodia	Kratie	Austro-Asiatic	Khmer	34	18	52.94
5	Kravet	Cambodia	Ratanakri	Austro-Asiatic	West Bahnaric	24	12	50.00
6	Kreung	Cambodia	Ratanakri	Austro-Asiatic	West Bahnaric	22	14	63.64
7	Kuy	Cambodia	Stung Treng	Austro-Asiatic	Katuic	37	34	91.89
8	Lao	Cambodia	Stung Treng	Daic	Kadai	27	14	51.85
9	Lun	Cambodia	Ratanakri	Austro-Asiatic	West Bahnaric	13	12	92.31
10	Mel	Cambodia	Kratie	Austro-Asiatic	Monic	19	15	78.95
11	Phnong	Cambodia	Kratie	Austro-Asiatic	South Bahnaric	26	20	76.92
12	Stieng	Cambodia	Kratie	Austro-Asiatic	South Bahnaric	12	8	66.67
13	Tompoun	Cambodia	Ratanakri	Austro-Asiatic	South Bahnaric	51	37	72.55
14	Kraol	Cambodia	Ratanakri	Austro-Asiatic	South Bahnaric	1	1	100%
14	Blang	Thailand	Chiang Rai	Austro-Asiatic	Waic	7	5	71.43
15	Htin	Thailand	Nan	Austro-Asiatic	Mal-Phrai	35	30	85.71
16	Lawa	Thailand	Chiang Mai	Austro-Asiatic	Waic	41	14	34.15
17	Palaung	Thailand	Chiang Mai	Austro-Asiatic	Palaung-Riang	16	3	18.75
18	Mon	Thailand	Chiang Mai	Austro-Asiatic	Monic	2	0	0
19	Bulang	China	Yunnan	Austro-Asiatic	Waic	55	17	30.91
20	Wa	China	Yunnan	Austro-Asiatic	Palaung-Riang	57	5	8.77
21	De'ang	China	Yunnan	Austro-Asiatic	Waic	68	13	19.12
					Total	**646**	**343**	**53.10**

Table 1. Sampled populations from MSEA and southern China.

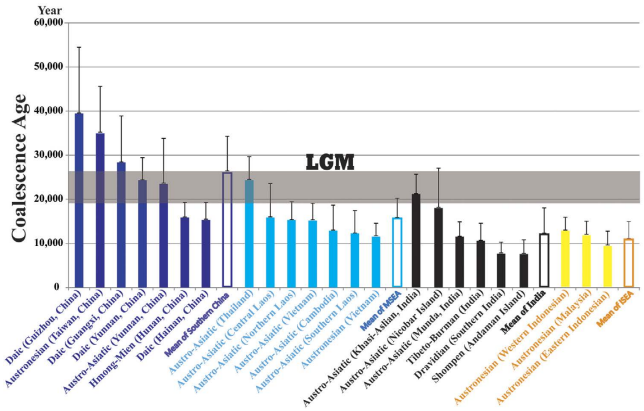

Figure 3. Comparison of coalescence ages of the O2a1-M95 lineages among diffenent geographic populations. The age of each geographic or linguistic group was calculated by taking the average of respective populations from supplementary Table S3.

Discussion

Throughout the previous studies, the two primary competing conceptions on the origin and prehistoric migratory pattern of the AA populations have left considerable debate, in part due to not including a wider geographic sampling. Here, we tested these rival hypotheses by systematically collecting AA samples from MSEA and southern China, and observed high frequencies of the O2a1-M95 lineages across all the studied AA populations. This broader survey confirmed that this Y-chromosome lineage represents a genetic signature of all AA populations, and can serve as an effective genetic marker for tracing the prehistoric movements and origins of these populations.

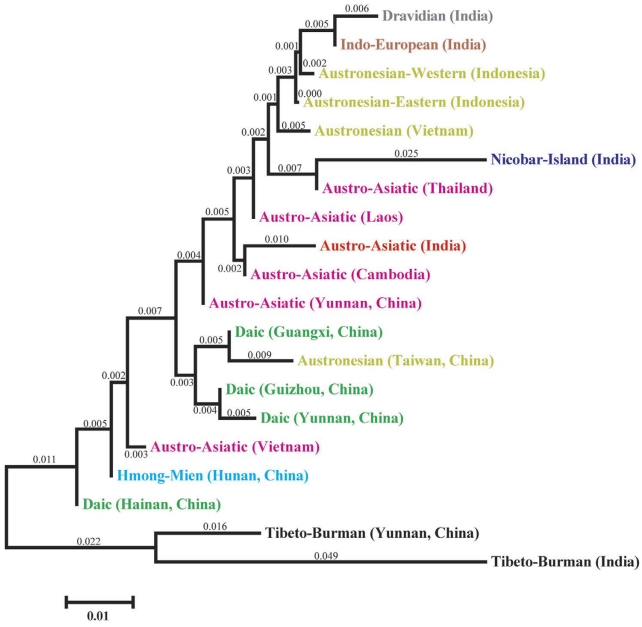

Figure 4. NJ-tree constructed of Y-STRs variations among different language family populations. Different linguistic families are shown using different colors. Branch length values are indicated above the branch.

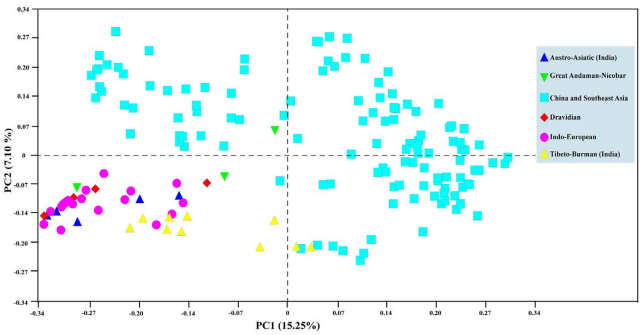

Figure 5. Map of principal component analysis (PCA) among Asian populations. Populations of East Asia and South Asia were grouped respectively by geograpghic region and language family. AA and TB-speaking populations closely clustered with DR anf IE populations in the lower left. The first and the second components explain 15.25% and 7.10% of the genetic variance, respectively.

The Y-chromosome data collected in this study does not support an Indian origin of the O2a1-M95 lineage, but instead shows that O2a1-M95 carriers in India originated in southern China and then migrated from MSEA to India around 10 KYA after the LGM. Moreover, our broader analysis that included Daic speaking populations from southern China showed that this population possessed the most diversified O2a1-M95 lineage with an average coalescence age of ~30 KYA, making it the oldest of all known O2a1-M95 carrying populations, and thereby supporting an initial origin of this Y-chromosome lineage in Daic speakers who migrated southward to MESA and a later westward to India ~10 KYA. During the preparation of this manuscript, Arunkumar *et al.* published a similar analysis of O2a1-M95 in Asian populations[28], and their data also favored an east-to-west migration although the estimated age of migration was much younger than ours due to different mutation rates and methods used for age estimation. Our analysis of mtDNA diversity suggests that after dispersal to India, the O2a1-M95 carrying populations widely absorbed the local maternal gene pool. In contrast to the well-known earliest migration of modern humans from Africa to eastern Asia by way of the Indian subcontinent, our data illustrates a back wave and sex-biased migration of the AA speakers from MSEA to India after the LGM, hinting at a far more complex prehistory of Paleolithic human populations.

Materials and Methods

Genotyping and data collection. For the 343 male samples that belong to the O2a1-M95 lineage from our previous study (Table 1)[18], we genotyped 16 Y-STRs (DYS19/394, DYS388, DYS389 I, DYS389 II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS458, DYS461, DYS635 and GATA H4) with the methods described previously[29,30]. DYS389I (DYS389cd) was subtracted from DYS389II and renamed 389ab because DYS389II contains the repeat number of DYS389I. To dissect the origin and migratory patterns of the O2a1-M95 lineage, we collected all available O2a1-M95 Y-STRs data, which covers 107 geographic populations (up to 2,510 samples carrying O2a1-M95) from East Asia, Southeast Asia and South Asia[9,12,20–24,26,31–38] (Fig. 1, supplementary Tables S1 and S2).

Data analysis. We estimated the time of most recent common ancestor (TMRCA) of the O2a1-M95 lineage using Y-STRs variation in each population as described previously, with a 25-year generation time and a mutation rate of 6.9×10^{-4} [12,39] (supplementary Table S3). For comparison, when calculating the ages we used three sets of loci for each population: a) the actual number of loci in the corresponding references, b) a 7-loci set (DYS19, DYS389 I, DYS389 II, DYS390, DYS391, DYS392 and DYS393) and c) a 6-loci set (DYS19, DYS389 I, DYS390, DYS391, DYS392 and DYS393), and the results from different calculations are very similar for most populations (supplementary Table S4). The mean TMRCAs of a geographic region are the average of its populations (Fig. 3). We also estimated the unbiased haplotype diversity of every population using GenAlEx 6.3. When estimating the age and diversity, O2a1-M95 populations with less than 10 samples were either excluded or merged to other closely related populations. In total, the coalescence ages and diversity of the O2a1-M95 lineages from 105 Asian populations were calculated (supplementary Tables S3 and S5).

A median-joining network, resolved with the MP algorithm, was constructed using the Network package 4.6.1.3 (www.fluxus-engineering.com). The O2a1-M95 variance isofrequency maps based on frequency and unbiased haplotype diversity were generated using Surfer10 (Golden Software Inc., Golden, USA), following the Kriging procedure. Average number of pairwise difference of Y-STRs for the studied populations was calculated using the Arlequin 3.5[40], and NJ-tree was constructed with MEGA 6.0[41]. We performed principal component analysis (PCA) based on the frequencies of mtDNA haplogroups according to the method developed by Richards *et al.*[42] with MVSP 3.13.

To compare the paternal and maternal gene pool between populations from East Asia and South Asia, we analyzed ~21,470 mtDNA sequences among these populations published previously (Supplementary Table S7).

References

1. Macaulay, V. *et al.* Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* **308,** 1034–1036, doi: 10.1126/science.1109792 (2005).
2. Su, B. *et al.* Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. *Am. J. Hum. Genet.* **65,** 1718–1724, doi: 10.1086/302680 (1999).
3. Shi, H. *et al.* Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. *Am. J. Hum. Genet.* **77,** 408–419, doi: 10.1086/444436 (2005).
4. Consortium, H. P.-A. S. *et al.* Mapping human genetic diversity in Asia. *Science* **326,** 1541–1545, doi: 10.1126/science.1177074 (2009).
5. Su, B. *et al.* Polynesian origins: insights from the Y chromosome. *Proc Natl Acad Sci USA* **97,** 8225–8228 (2000).
6. Jin, L. & Su, B. Natives or immigrants: modern human origin in east Asia. *Nat. Rev. Genet.* **1,** 126–133, doi: 10.1038/35038565 (2000).
7. Zhang, X. *et al.* Analysis of mitochondrial genome diversity identifies new and ancient maternal lineages in Cambodian aborigines. *Nat Commun* **4,** 2599, doi: 10.1038/ncomms3599 (2013).
8. Lewis, M. P. Ethnologue: languages of the world. Dallas (TX): SIL International. [Internet; cited 2010 Sep],<http://www.ethnologue.com/>(2009).
9. Chaubey, G. *et al.* Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. *Mol. Biol. Evol.* **28,** 1013–1024, doi: 10.1093/molbev/msq288 (2011).
10. Kumar, V. *et al.* Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. *BMC Evol. Biol.* **7,** 47, doi: 10.1186/1471-2148-7-47 (2007).
11. Sahoo, S. *et al.* A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proc Natl Acad Sci USA* **103,** 843–848, doi: 10.1073/pnas.0507714103 (2006).
12. Sengupta, S. *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.* **78,** 202–221, doi: 10.1086/499411 (2006).
13. Kivisild, T. *et al.* The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am. J. Hum. Genet.* **72,** 313–332, doi: 10.1086/346068 (2003).
14. Chakravarti, A. Human genetics: Tracing India's invisible threads. *Nature* **461,** 487–488, doi: 10.1038/461487a (2009).
15. Basu, A. *et al.* Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.* **13,** 2277–2290, doi: 10.1101/gr.1413403 (2003).
16. Metspalu, M. *et al.* Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet.* **5,** 26, doi: 10.1186/1471-2156-5-26 (2004).
17. Reddy, B. M. *et al.* Austro-Asiatic tribes of Northeast India provide hitherto missing genetic link between South and Southeast Asia. *PLoS One* **2,** e1141, doi: 10.1371/journal.pone.0001141 (2007).
18. Zhang, X. *et al.* An updated phylogeny of the human Y-chromosome lineage O2a-M95 with novel SNPs. *PLoS One* **9,** e101020, doi: 10.1371/journal.pone.0101020 (2014).
19. Clark, P. U. *et al.* The Last Glacial Maximum. *Science* **325,** 710–714, doi: 10.1126/science.1172873 (2009).
20. Cai, X. *et al.* Human migration through bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum revealed by Y chromosomes. *PLoS One* **6,** e24282, doi: 10.1371/journal.pone.0024282 (2011).

21. Li, H. *et al.* Paternal genetic affinity between Western Austronesians and Daic populations. *BMC Evol. Biol.* **8,** 146, doi: 10.1186/1471-2148-8-146 (2008).
22. Li, D. *et al.* Paternal genetic structure of Hainan aborigines isolated at the entrance to East Asia. *PLoS One* **3,** e2168, doi: 10.1371/journal.pone.0002168 (2008).
23. Gan, R. J. *et al.* Pinghua population as an exception of Han Chinese's coherent genetic structure. *J. Hum. Genet.* **53,** 303–313, doi: 10.1007/s10038-008-0250-x (2008).
24. Thangaraj, K. *et al.* Genetic affinities of the Andaman Islanders, a vanishing human population. *Curr. Biol.* **13,** 86–93 (2003).
25. Trivedi, R. *et al.* Molecular insights into the origins of the Shompen, a declining population of the Nicobar archipelago. *J. Hum. Genet.* **51,** 217–226, doi: DOI 10.1007/s10038-005-0349-2 (2006).
26. Chandrasekar, A. *et al.* YAP insertion signature in South Asia. *Ann Hum Biol* **34,** 582–586, doi: 10.1080/03014460701556262 (2007).
27. Zhong, H. *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.* **55,** 428–435, doi: 10.1038/jhg.2010.40 (2010).
28. GaneshPrasad, A. *et al.* A late Neolithic expansion of Y chromosomal haplogroup O2a1-M95 from east to west. *Journal of Systematics and Evolution*. 1–15 doi: 10.1111/jse.12147 (2015).
29. Butler, J. M. *et al.* A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. *Forensic Sci. Int.* **129,** 10–24 (2002).
30. Kayser, M. *et al.* Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Legal Med* **110,** 125–133, 141–129 (1997).
31. Delfin, F. *et al.* The Y-chromosome landscape of the Philippines: extensive heterogeneity and varying genetic affinities of Negrito and non-Negrito groups. *Eur J Hum Genet* **19,** 224–230, doi: 10.1038/ejhg.2010.162 (2011).
32. Hammer, M. F. *et al.* Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. *J. Hum. Genet.* **51,** 47–58, doi: 10.1007/s10038-005-0322-0 (2006).
33. He, J. D. *et al.* Patrilineal perspective on the Austronesian diffusion in Mainland Southeast Asia. *PLoS One* **7,** e36437, doi: 10.1371/journal.pone.0036437 (2012).
34. Karafet, T. M. *et al.* Major east-west division underlies Y chromosome stratification across Indonesia. *Mol. Biol. Evol.* **27,** 1833–1844, doi: 10.1093/molbev/msq063 (2010).
35. Kutanan, W. *et al.* Genetic structure of the Mon-Khmer speaking groups and their affinity to the neighbouring Tai populations in Northern Thailand. *BMC Genet.* **12,** 56, doi: 10.1186/1471-2156-12-56 (2011).
36. Nonaka, I., Minaguchi, K. & Takezaki, N. Y-chromosomal binary haplogroups in the Japanese population and their relationship to 16 Y-STR polymorphisms. *Ann Hum Genet* **71,** 480–495, doi: 10.1111/j.1469-1809.2006.00343.x (2007).
37. Xue, Y. *et al.* Male demography in East Asia: a north-south contrast in human population expansion times. *Genetics* **172,** 2431–2439, doi: 10.1534/genetics.105.054270 (2006).
38. Trivedi, R. *et al.* Molecular insights into the origins of the Shompen, a declining population of the Nicobar archipelago. *J. Hum. Genet.* **51,** 217–226, doi: 10.1007/s10038-005-0349-2 (2006).
39. Zhivotovsky, L. A. *et al.* The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am. J. Hum. Genet.* **74,** 50–61, doi: Doi 10.1086/380911 (2004).
40. Excoffier, L., Laval, G. & Schneider, S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online* **1,** 47–50 (2005).
41. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30,** 2725–2729, doi: 10.1093/molbev/mst197 (2013).
42. Richards, M., Macaulay, V., Torroni, A. & Bandelt, H. J. In search of geographical patterns in European mitochondrial DNA. *Am. J. Hum. Genet.* **71,** 1168–1174, doi: 10.1086/342930 (2002).
43. Shi, H. *et al.* Genetic evidence of an East Asian origin and paleolithic northward migration of Y-chromosome haplogroup N. *PLoS One* **8,** e66102, doi: 10.1371/journal.pone.0066102 (2013).

Acknowledgements

We are grateful to all the volunteers for providing blood samples, and to Andrew Willden for editing the manuscript. This study was supported by the National 973 Program of China (2012CB518202 to X.Q.), the National Natural Science Foundation of China (31130051 and 91231203 to B.S., 31371268 and 91131001 to H.S. and 31371269 to X.Q.) and the Natural Science Foundation of Yunnan Province (2010CI044 to H.S.).

Author Contributions

B.S. and H.S. designed the experiment; X.M.Z., X.B.Q., J.K., Z.H.Y., B.S., T.S., L.B., H.S.A., H.S., D.K. and H.S. collected the samples; X.M.Z., S.Y.L. and X.B.Q. collected the data and conducted data analysis; J.W.L. and H.Z. provided technical assistance in the experiments; X.M.Z., X.B.Q., H.S. and B.S. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Zhang, X. *et al.* Y-chromosome diversity suggests southern origin and Paleolithic backwave migration of Austro-Asiatic speakers from eastern Asia to the Indian subcontinent. *Sci. Rep.* **5**, 15486; doi: 10.1038/srep15486 (2015).