JSE Journal of Systematics and Evolution

doi: 10.1111/jse.12147

Research Article

A late Neolithic expansion of Y chromosomal haplogroup O2a1-M95 from east to west

GaneshPrasad Arunkumar[1,2], Lan-Hai Wei[3], Valampuri John Kavitha[1,4], Adhikarla Syama[1], Varatharajan Santhakumari Arun[1], Surendra Sathua[5], Raghunath Sahoo[6], R. Balakrishnan[7], Tomo Riba[8], Jharna Chakravarthy[9], Bapukan Chaudhury[10], Premanada Panda[11], Pradipta K. Das[12], Prasanna K. Nayak[13], Hui Li[3], Ramasamy Pitchappan[1,14]*, and The Genographic Consortium

[1]The Genographic Laboratory, School of Biological Sciences, Madurai Kamaraj University, Madurai 625021, India
[2]School of Chemical and Biotechnology, SASTRA University, Thanjavur 613401, India
[3]MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China
[4]Department of Biotechnology, Mother Theresa University, Kodaikanal 624102, India
[5]Project Coordinator, Sarba Sikhya Abjhijan, Rayagada, Orissa 765001, India
[6]DND Agency-Bondo Project Leader (Retd), Rayagada Directorate, Bhubaneswar, Orissa 751004, India
[7]Roja Muthiah Research Library, Taramani, Chennai 600113, India
[8]Department of Geography, Rajiv Ghandhi University, Rnono Hills, Itanagar, Arunachal Pradesh 791112, India
[9]Department of Zoology, Rajiv Ghandhi University, Rnono Hills, Itanagar, Arunachal Pradesh 791112, India
[10]Department of Anthropology, Gauwhati University, Guwahati, Assam 781014, India
[11]Department Anthropology, Sambalpur University, Jyoti Vihar, Sambalpur, Orissa 768019, India
[12]Department of Anthropology, Guru Ghasidas Central University, Bilaspur, Chattisgarh 495009, India
[13]Department of Anthropology, Utkal University, Vani Vihar, Bhubaneswar, Orissa 751004, India
[14]Genomics Laboratory, Chettinad Academy of Research and Education, Kelampakkam, Chennai 603103, India
The participants of the Genographic Consortium: Christina J. Adlera[a], Elena Balanovska[b], Oleg Balanovsky[b], Jaume Bertranpetit[c], Andrew C. Clarke[d], David Comas[c], Alan Cooper[a], Clio S. I. Der Sarkissian[a], Matthew C. Dulik[e], Jill B. Gaieski[e], Wolfgang Haak[a], Marc Haber[c,f], Angela Hobbs[g], Asif Javed[h], Li Jin[i], Matthew E. Kaplan[j], Shilin Li[i], Begoña Martínez-Cruz[c], Elizabeth A. Matisoo-Smith[d], Marta Melé[c], Nirav C. Merchant[j], R. John Mitchell[k], Amanda C. Owings[e], Laxmi Parida[h], Daniel E. Platt[h], Lluis Quintana-Murci[l], Colin Renfrew[m], Daniela R. Lacerda[n], Ajay K. Royyuru[h], Theodore G. Schurr[e], Fabrício R. Santos[n], Himla Soodyall[g], David F. Soria Hernanz[o], Pandikumar Swamikrishnan[p], Chris Tyler-Smith[q], Pedro Paulo Vieira[r], Miguel G. Vilar[e], R. Spencer Wells[o], Pierre A. Zalloua[f], and Janet S. Ziegle[s]
[a]University of Adelaide, South Australia, Australia; [b]Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow, Russia; [c]Universitat Pompeu Fabra, Barcelona, Spain; [d]University of Otago, Dunedin, New Zealand; [e]University of Pennsylvania, Philadelphia, PA, USA; [f]Lebanese American University, Chouran, Beirut, Lebanon; [g]National Health Laboratory Service, Johannesburg, South Africa; [h]IBM, Yorktown Heights, NY, USA; [i]Fudan University, Shanghai, China; [j]University of Arizona, Tucson, AZ, USA; [k]La Trobe University, Melbourne, Victoria, Australia; [l]Institut Pasteur, Paris, France; [m]University of Cambridge, Cambridge, United Kingdom; [n]Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; [o]National Geographic Society, Washington, DC, USA; [p]IBM, Somers, NY, USA; [q]The Wellcome Trust Sanger Institute, Hinxton, United Kingdom; [r]Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil; and [s]Applied Biosystems, Foster City, CA, USA.
*Author for correspondence. E-mail: pitchappanrm@yahoo.co.uk. Tel.: 044-47429050.
Received 3 October 2014; Accepted 31 January 2015; Article first published online 31 March 2015

Abstract The origin and dispersal of Y-Chromosomal haplogroup O2a1-M95, distributed across the Austro Asiatic speaking belt of East and South Asia, are yet to be fully understood. Various studies have suggested either an East Indian or Southeast Asian origin of O2a1-M95. We addressed the issue of antiquity and dispersal of O2a1-M95 by sampling 8748 men from India, Laos, and China and compared them to 3307 samples from other intervening regions taken from the literature. Analyses of haplogroup frequency and Y-STR data on a total 2413 O2a1-M95 chromosomes revealed that the Laos samples possessed the highest frequencies of O2a1-M95 (74% with >0.5) and its ancestral haplogroups (O2*-P31, O*-M175) as well as a higher proportion of samples with 14STR-median haplotype (17 samples in 14 populations), deep coalescence time (5.7 ± 0.3 Kya) and consorted O2a1-M95 expansion evidenced from STR evolution. All these suggested Laos to carry a deep antiquity of O2a1-M95 among the study regions. A serial decrease in expansion time from east to west: 5.7 ± 0.3 Kya in Laos, 5.2 ± 0.6 in Northeast India, and 4.3 ± 0.2 in East India, suggested a late Neolithic east to west spread of the lineage O2a1-M95 from Laos.

Key words: Laos, migration, Neolithic, NRY, O2a1-M95.

Non recombinant Y-chromosomal (NRY) DNA markers, inherited through paternal lineage, provide a robust tool to decipher male mediated migration, expansion, and dispersal. The origin and dispersal of NRY haplogroup O2a1-M95, seen in ~58% of Southeast Asian males, has been a subject of intense debate (Karafet et al., 2010; Cai et al., 2011; Zhang et al., 2014). In India, 15% of NRY variation has been attributed to the haplogroup O2a1-M95 (Trivedi et al., 2008). The highest

frequencies of this haplogroup in India is found among the Austro Asiatic (Munda branch) speakers, who are geographically restricted to Eastern and to parts of Central India (Sahoo et al., 2006; Sengupta et al., 2006). The origins of Indian Austro Asiatic speakers had earlier been correlated to the origin of O2a1-M95 (Kumar et al., 2007).

There are different schools of thoughts on the origin of Indian O2a1-M95. One proposal argues for an origin within India during the mid-Pleistocene (~65 kya), followed by spread to East and South East (SE) Asia via the North East (NE) Indian corridor (Kumar et al., 2007; Reddy et al., 2007). Another proposal based on both Y-Chromosomal and autosomal markers has proposed an origin in SE. Asia and a recent late Pleistocene spread (~15 Kya) to India (Chaubey et al., 2011). Since expansion of Austro Asiatic speakers has been interpreted to be congruent with O2a1-M95 lineage, Y-chromosomal analysis coupled with linguistic evidences may help decipher the mechanism of the spread of Austro Asiatic languages.

Linguists however differ in their opinion on the origin of Austro Asiatic language family. Studies based on linguistic cognates and archaeological evidences have suggested an origin of Austro Asiatic languages in the mid-Yangtze river basin of Southern China (Sichuan state) followed by a spread to India along the banks of the river Brahmaputra (Higham, 2002; Sidwell, 2009; Peiros, 2011). On the other hand, lexicostatistical analysis suggested an origin in the Mekong river valley that spreads across Thailand, Vietnam, Lao, Cambodia (Sidwell, 2010).

In the present study we investigated the antiquity and expansion of NRY haplogroup O2a1-M95, by studying the distribution and diversity of 2413 O2a1-M95 chromosomes covering the expanse from India to China and Southeast Asia. The analysis supports a deep ancestry of this clade in Laos and its late Neolithic spread to India.

Material and Methods

Samples and methods

A total of 8748 samples from 104 populations were collected from India (4092 samples/54 populations) and China (4656/50), between 2006 and 2010. Of these, 1633 were O2a1-M95 (Table 1) and to this we included data on 780 O2a1-M95 chromosomes from 27 populations available in literature encompassing geographic regions not covered by The Genographic Project (Table 1) (Kantang et al., 2010; Karafet et al., 2010; Siriboonpiputtana et al., 2010; Chaubey et al., 2011; Wu et al., 2011). Populations were categorized based on their geography. Populations from Guangxi province of China were categorized separately from other Chinese (referred as China continental in the manuscript) due to their geographic proximity to Laos populations (Mekong valley). No sample or data from Myanmar, either in The Genographic or from literature was available for inclusion and comparison.

Ethical clearances for the study were obtained from the Institutional Ethical Committees of respective institutions. Samplings were performed with written (signed or thumb impression) informed consent. The Indian centre collected mouth wash, while the Chinese centre used saliva, fingertip, or venous blood; DNA were then extracted by standard procedures (Ausubel et al., 2002). Y-Chromosomal SNP Genotyping was performed by custom made Taqman assays in ABI 7900HT Real-Time PCR System (Applied Biosystems, Foster City) in the respective Genographic centres. A set of 17 STRs and 6 Indels were genotyped using the Y-Filer Kit and Custom made Multiplex-2 kits (Applied Biosystems, Foster City) (Haber et al., 2011). But we could only compare and analyse a common set of 14 STRs (DYS458, DYS437, DYS438, DYS448, DYS-GATAH4, DYS635 plus set of eight STRs DYS389a, DYS389b, DYS390, DYS19, DYS391, DYS392, DYS393, and DYS439) available in the literature.

Statistical analyses

O2a1-M95 frequency was estimated by gene counting method. STR haplotype based Rst distances were calculated in Arlequin v3.5.1.2 using 1000 bootstraps (Excoffier et al., 2005), and this was used to compute a Multi Dimensional Scaling (MDS) plot (Kruskal, 1964) in ``R'' v2.12.1 (R Development Core Team, 2010). The phylogenetic relationships of various haplotypes were determined using the Network program 4.6.1.2 employing the reduced median algorithm with a reduction threshold of 1 (Bandelt et al., 1995). Contour plots of haplogroup frequency and STR variance distributions were computed using the Kriging algorithm employed and plotted in ``R'' v2.12.1 (R Development Core Team, 2010). Shape files for the maps were obtained from www.naturalearthdata.com. The Average Squared Distance (ASD) and the Sum of Squared Distance (SSD) were used to estimate the age and proportion of ancient haplotypes in a population respectively (Sengupta et al., 2006). Age estimates were obtained using genealogical mutation rates for 8-STR and 14-STR haplotype (Goedbloed et al., 2009). A generation time of 25 years was used for age estimates and populations with five or more samples only were used for STR based analyses.

Results and Discussion

Frequency and variance estimates

The higher frequency of an allele and high STR variance in majority of populations in a given region may indicate that the region was the probable place of origin and/or expansion of the NRY lineage in question (Sengupta et al., 2006; Myres et al., 2011). The present study revealed higher O2a1-M95 frequencies of >0.6 in 17 of 31 and >0.5 in 23 of 31 Laos populations with an average of 61.56%, but only in 4 (>0.6) and 6 (>0.5) of 21 E Indian populations (average of 34.09%) (Tables 1, 2). Guangxi (southern China, adjoining Mekong Valley) and China excluding Guangxi province (from now will be referred as China continental) showed much lower frequencies (0.03–0.5 and <0.06, respectively) with 1 and 8 of 25 populations showing >0.5 and >0.1 frequency, respectively, with a mean of 12.92% (Table 1; Fig. 1).

High STR variance at population level may represent long-standing expansion or sometimes admixture. In 8-STR variance analysis of various regional populations studied, the majority (22 of 28) of Laos populations showed higher STR variance (>0.3; median 8-STR variance of all samples in the present study $= 0.3$) compared to 8/21 E Indian and 7/13 SE Asian island populations (Fig. 2; Table S1). Considering 14 STRs, all the 7 SE Asian island populations studied showed >0.3

Table 1 The list of study populations from various regions sampled, numbers studied, language family and their NRY HG O2a-M95 frequency estimates

No.	Population	Country	Geographic region	Language family	Studied number	O2a count	O2a frequency%	Fishers *P* value^	Source
1	Vasava	India	W India	Indo European	87	3	3.45	**1.07E-05**	a
2	Siddi	India	W India	AfroAsiatic/IndoEuropean	50	1	2	**2.87E-04**	a
3	Korku	India	W India	AA-Munda North	66	31	46.97	**1.10E-06**	a
4	Kolam	India	W India	Indo European	37	5	13.51	4.13E-01	a
5	Gond Mah	India	W India	Dravidian Central	56	1	1.79	**7.93E-05**	a
6	Seharia	India	W India	AA-Munda/IndoEuropean	96	7	7.29	**7.88E-04**	a
7	Charan	India	W India	Indo European	69	1	1.45	**5.91E-06**	a
			W India Total		461	49	10.63		
8	Madiga	India	S India@	Dravidian South	49	2	4.08	**3.40E-03**	a
9	Sattibalija	India	S India	Dravidian South	100	3	3	**9.33E-07**	a
10	Mala	India	S India	Dravidian South	98	2	2.04	**1.58E-07**	a
11	Kamma	India	S India	Dravidian South	104	1	0.96	**3.45E-09**	a
12	Kapu	India	S India	Dravidian South	104	1	0.96	**3.45E-09**	a
13	Billava	India	S India	Dravidian South	78	1	1.28	**1.08E-06**	a
14	Saliyar	India	S India	Dravidian South	11	1	9.09	7.05E-01	a
15	Namboodiri	India	S India	Indo European	51	2	3.92	**2.27E-03**	a
16	Muslim	India	S India	Dravidian South	75	1	1.33	**1.60E-06**	a
17	Thoda	India	S India	Dravidian South	27	1	3.7	2.95E-02	a
18	Nattukottai Chettiar	India	S India	Dravidian South	174	4	2.3	**2.50E-12**	a
			S India Total		871	19	2.18		
19	Kamar	India	C India	Indo European	20	10	50	**2.59E-03**	a
20	Maria Gond	India	C India	Dravidian Central	2	1	50	3.60E-01	a
21	Gond MP	India	C India	Dravidian Central	40	8	20	1.00E+00	a
22	Halba	India	C India	Indo European	68	2	2.94	**9.09E-05**	a
23	Brahmin Saryupareen	India	C India	Indo European	106	2	1.89	**2.85E-08**	a
			C India Total		236	23	9.75		
24	Konda Kammara	India	E India	Dravidian South	51	12	23.53	4.87E-01	a
25	Konda Reddy	India	E India	Dravidian South	49	9	18.37	8.60E-01	a
26	Jalari	India	E India	Dravidian South	99	8	8.08	**1.52E-03**	a
27	Relli	India	E India	Indo European	109	6	5.5	**3.11E-05**	a
28	Ho	India	E India	AA-Munda North	115	72	62.61	**1.68E-23**	a
29	Munda	India	E India	AA-Munda North	104	51	49.04	**2.79E-11**	a
30	Santal	India	E India	AA-Munda North	121	52	42.98	**7.52E-09**	a
31	Binjhal	India	E India	Dravidian Central	109	43	39.45	**2.86E-06**	a
32	Gond Ori	India	E India	Dravidian Central	102	27	26.47	1.06E-01	a
33	Kisan	India	E India	AA-Munda/Dravidian	105	26	24.76	2.21E-01	a
34	Brahmin Jhadua	India	E India	Indo European	23	1	4.35	6.69E-02	a
35	Oraon	India	E India	Dravidian Central	116	41	35.34	**1.00E-04**	a

Continued

Table 1 Continued

No.	Population	Country	Geographic region	Language family	Studied number	O2a count	O2a frequency%	Fishers *P* value^	Source
36	Gadaba	India	E India	AA-Munda South	84	67	79.76	**4.97E-32**	a
37	Langia Soura	India	E India	AA-Munda South	64	44	68.75	**3.60E-17**	a
38	Dhongria Kondh	India	E India	Dravidian Central	86	53	61.63	**3.28E-17**	a
39	Bondo Lower	India	E India	AA-Munda South	28	13	46.43	**1.44E-03**	a
40	Gadaba Ollar	India	E India	Dravidian Central	9	3	33.33	3.96E-01	a
41	Bondo	India	E India	AA-Munda South	48	15	31.25	6.82E-02	a
42	Koya	India	E India	Dravidian Central	103	25	24.27	2.67E-01	a
43	Khandayat	India	E India	Indo European	48	2	4.17	**3.24E-03**	a
44	Brahmin Utkalya	India	E India	Indo European	102	1	0.98	**5.37E-09**	a
			E India Total		1675	571	34.09		
45	Garo	India	NE India	Tibeto-Burman Sal	107	22	20.56	9.03E-01	a
46	Khasi	India	NE India	AA-MonKhmer North	98	30	30.61	1.11E-02	a
47	Ahom	India	NE India	Daic-KamTai-Tai-Southwestern	36	1	2.78	**5.63E-03**	a
48	Kalita	India	NE India	Indo European	114	2	1.75	**5.13E-09**	a
49	Karbi	India	NE India	Tibeto-Burman Sal	109	21	19.27	9.05E-01	a
50	Meitei	India	NE India	Tibeto-Burman Sal	124	4	3.23	**5.38E-08**	a
51	Kuki	India	NE India	Tibeto-Burman Sal	77	34	44.16	**1.92E-06**	a
52	Naga	India	NE India	Tibeto-Burman Sal	67	1	1.49	**9.14E-06**	a
53	Galo	India	NE India	Tibeto-Burman Central	84	1	1.19	**2.88E-07**	a
54	Mishing	India	NE India	Tibeto-Burman Central	33	2	6.06	4.81E-02	a
			NE India Total		849	118	13.9		
		India Total			4092	780	19.06		
55	Shompen	India	Nicobar	AA-MonKhmer Aslian	12	12	100	**4.05E-09**	c
56	Nicobarese	India	Nicobar	AA-MonKhmer Nicobar	11	11	100	**2.03E-08**	c
			Nicobar Total		23	23	100		
57	Brau	Laos	Laos	AA-MonKhmer East	32	21	65.63	**2.55E-08**	b
58	Oy	Laos	Laos	AA-MonKhmer East	50	31	62	**1.01E-10**	b
59	Tai Mène	Laos	Laos	Daic-KamTai-Tai-Northern	24	20	83.33	**4.60E-11**	b
60	Aheu	Laos	Laos	AA-MonKhmer Viet-Muong	38	26	68.42	**1.31E-10**	b
61	Bo	Laos	Laos	AA-MonKhmer Viet-Muong	28	19	67.86	**5.30E-08**	b
62	Kang	Laos	Laos	Daic-KamTai-Tai	12	4	33.33	2.74E-01	b
63	Tai Deang	Laos	Laos	Daic-KamTai-Tai-Northern	44	36	81.82	**1.84E-18**	b
64	Tai Dam	Laos	Laos	Daic-KamTai-Tai-Southwestern	50	33	66	**1.95E-12**	b
65	Saek	Laos	Laos	Daic-KamTai-Tai-Northern	30	20	66.67	**3.70E-08**	b
66	PhuThai	Laos	Laos	Daic-KamTai-Tai-Southwestern	24	13	54.17	**2.15E-04**	b
67	Sô	Laos	Laos	AA-MonKhmer East	50	27	54	**1.01E-07**	b
68	Xinh mul	Laos	Laos	AA-MonKhmer North	29	25	86.21	**3.16E-14**	b
69	Hmong White	Laos	Laos	Hmong Mein	27	3	11.11	3.37E-01	b

Continued

Table 1 Continued

No.	Population	Country	Geographic region	Language family	Studied number	O2a count	O2a frequency%	Fishers *P* value^	Source
70	Lao	Laos	Laos	Daic-KamTai-Tai-Southwestern	59	32	54.24	**5.80E-09**	b
71	Khmu	Laos	Laos	AA-MonKhmer North	51	34	66.67	**5.77E-13**	b
72	Bit	Laos	Laos	AA-MonKhmer North	28	15	53.57	**8.30E-05**	b
73	Phunoi	Laos	Laos	Tibeto-Burman Ngwi-Burmese	27	13	48.15	**9.46E-04**	b
74	Suy	Laos	Laos	AA-MonKhmer East	39	34	87.18	**2.90E-19**	b
75	Inh	Laos	Laos	AA-MonKhmer East	34	27	79.41	**1.49E-13**	b
76	Kataang	Laos	Laos	AA-MonKhmer East	38	15	39.47	**6.67E-03**	b
77	Alak	Laos	Laos	AA-MonKhmer East	31	22	70.97	**1.20E-09**	b
78	Ngeq	Laos	Laos	AA-MonKhmer East	35	17	48.57	**1.39E-04**	b
79	Laven	Laos	Laos	AA-MonKhmer East	50	22	44	**1.14E-04**	b
80	Talieng	Laos	Laos	AA-MonKhmer East	35	23	65.71	**5.23E-09**	b
81	Jeh	Laos	Laos	AA-MonKhmer East	32	15	46.88	**5.56E-04**	b
82	Katu	Laos	Laos	AA-MonKhmer East	45	31	68.89	**1.60E-12**	b
83	Lamet	Laos	Laos	AA-MonKhmer North	35	30	85.71	**1.07E-16**	b
84	Rien	Laos	Laos	Daic-KamTai-Tai	50	27	54	**1.01E-07**	b
85	Mal	Laos	Laos	AA-MonKhmer North	50	37	74	**2.50E-16**	b
86	Hmong	Laos	Laos	Hmong Mein	17	2	11.76	5.51E-01	b
87	Phuan	Laos	Laos	Daic-KamTai-Tai-Southwestern	22	13	59.09	**6.42E-05**	b
			Laos Total		1116	687	61.56		
88	Thai Central	Thailand	Thailand	Daic-KamTai	501	109	21.76	3.32E-01	f
89	Thai Yala	Thailand	Thailand	Daic-KamTai	99	22	22.22	6.14E-01	e
90	Vietnamese	Vietnam	Vietnam	AA-MonKhmer Viet-Muong	80	23	28.75	6.67E-02	c,d
91	Cambodian	Cambodia	Cambodia	AA-MonKhmer East	6	3	50	9.90E-02	c
			Thai region		686	157	22.89		
92	Pekanbaru	Indonesia	SE Asian Islands	Austronesian	31	8	25.81	3.77E-01	c
93	Sumatra	Indonesia	SE Asian Islands	Austronesian	95	14	14.74	2.46E-01	c,d
94	Kota Kinabalu	Malaysia	SE Asian Islands	Austronesian	44	7	15.91	5.76E-01	c
95	Malay	Malysia	SE Asian Islands	Austronesian	50	16	32	4.89E-02	c,d
96	Java	Indonesia	SE Asian Islands	Austronesian	114	52	45.61	**7.43E-10**	c,d
97	Borneo	Borneo	SE Asian Islands	Austronesian	126	33	26.19	9.25E-02	c,d
98	Palu	Indonesia	SE Asian Islands	Austronesian	28	5	17.86	1.00E+00	c
99	Banjarmasin	Indonesia	SE Asian Islands	Austronesian	27	9	33.33	9.23E-02	c
100	Bali	Indonesia	SE Asian Islands	Austronesian	641	367	57.25	**5.06E-102**	d
101	Mataran	Indonesia	SE Asian Islands	Austronesian	40	17	42.5	**1.11E-03**	c
102	Toraja	Indonesia	SE Asian Islands	Austronesian	49	8	16.33	5.96E-01	c
103	Sumba	Indonesia	SE Asian Islands	Austronesian	350	1	0.29	**4.69E-33**	d
104	Sulawesi	Indonesia	SE Asian Islands	Austronesian	54	7	12.96	2.34E-01	d
105	Flores	Indonesia	SE Asian Islands	Austronesian	394	18	4.57	**4.59E-19**	d

Continued

Table 1 Continued

No.	Population	Country	Geographic region	Language family	Studied number	O2a count	O2a frequency%	Fishers *P* value^	Source
106	Philippines	Phillipines	SE Asian Islands	Austronesian	87	3	3.45	**1.07E-05**	c,d
			SE Asian Total		2130	565	26.53		
107	Paiwan	China	China Continental*	Austronesian	208	6	2.88	**2.65E-13**	g
108	Taiwanese	China	China Continental	Austronesian	43	2	4.65	**7.09E-03**	c
109	Taiwanese Aboriginals	China	China Continental	Austronesian	48	1	2.08	**4.40E-04**	d
110	Tibetan Amdo	China	China Continental	Tibeto-Burman Western	260	4	1.54	**5.44E-20**	b
111	Baima	China	China Continental	Tibeto-Burman Northeastern	136	3	2.21	**8.09E-10**	b
112	Mongolian	China	China Continental	Altaic	73	5	6.85	**2.99E-03**	b
113	Manchu	China	China Continental	Altaic	219	4	1.83	**3.38E-16**	b
114	Han	China	China Continental	Sino-Tibetan Chinese	890	13	1.46	**2.08E-69**	b
115	Qiang	China	China Continental	Tibeto-Burman Northeastern	170	3	1.76	**5.37E-13**	b
116	Jiarong	China	China Continental	Tibeto-Burman Northeastern	104	1	0.96	**3.45E-09**	b
117	Muyag	China	China Continental	Sino-Tibetan Chinese	158	1	0.63	**2.07E-14**	b
118	Queyu	China	China Continental	Tibeto-Burman Northeastern	148	2	1.35	**5.18E-12**	b
119	Tibetan Central	China	China Continental	Tibeto-Burman Western	843	1	0.12	**3.17E-83**	b
120	Tibetan Khams	China	China Continental	Tibeto-Burman Western	200	3	1.5	**1.32E-15**	b
			China Continental Total		3500	49	1.4		
121	Pinghua-Han	China	Guangxi	Sino-Tibetan Chinese	101	43	42.57	**2.35E-07**	b
122	Mien	China	Guangxi	Hmong Mein	11	4	36.36	2.47E-01	b
123	Kam	China	Guangxi	Daic-KamTai-Kam Sui	27	7	25.93	4.69E-01	b
124	Yao	China	Guangxi	Sino-Tibetan Chinese	60	2	3.33	**2.89E-04**	d
125	Laka	China	Guangxi	Daic-KamTai-Lakkja	23	2	8.7	2.93E-01	b
126	Mulam	China	Guangxi	Daic-KamTai-Kam Sui	11	1	9.09	7.05E-01	b
127	NaPo-Han	China	Guangxi	Sino-Tibetan Chinese	82	45	54.88	**2.75E-12**	b
128	Minz-Zhuang	China	Guangxi	Daic-KamTai-Tai-Central	63	15	23.81	4.32E-01	b
129	Zhuang-GB	China	Guangxi	Daic-KamTai-Tai-Central	21	9	42.86	2.36E-02	b
130	She	China	Guangxi	Hmong Mein	51	18	35.29	1.25E-02	d
131	Miao	China	Guangxi	Hmong Mein	58	6	10.34	7.01E-02	d
			Guangxi Total		508	152	29.92		
			Total		12055	2413	20.02		

AA, Austro Asiatic; @, Andhra Pradesh samples bordering Orissa (>82E:17N) were considered as East India and Below this as South India. a, GIC: Present study: Genographic Indian Center; b, GCC: Present study: Genographic Chinese center; c, Chaubey et al. (2011); d, Karafet et al. (2010); e, Kantang et al. (2010); f, Siriboonpiputtana et al. (2010); g, Wu et al. (2011). ^, Bold represents *P* value < 0.01. *China continental refers to China excluding Guangxi province.

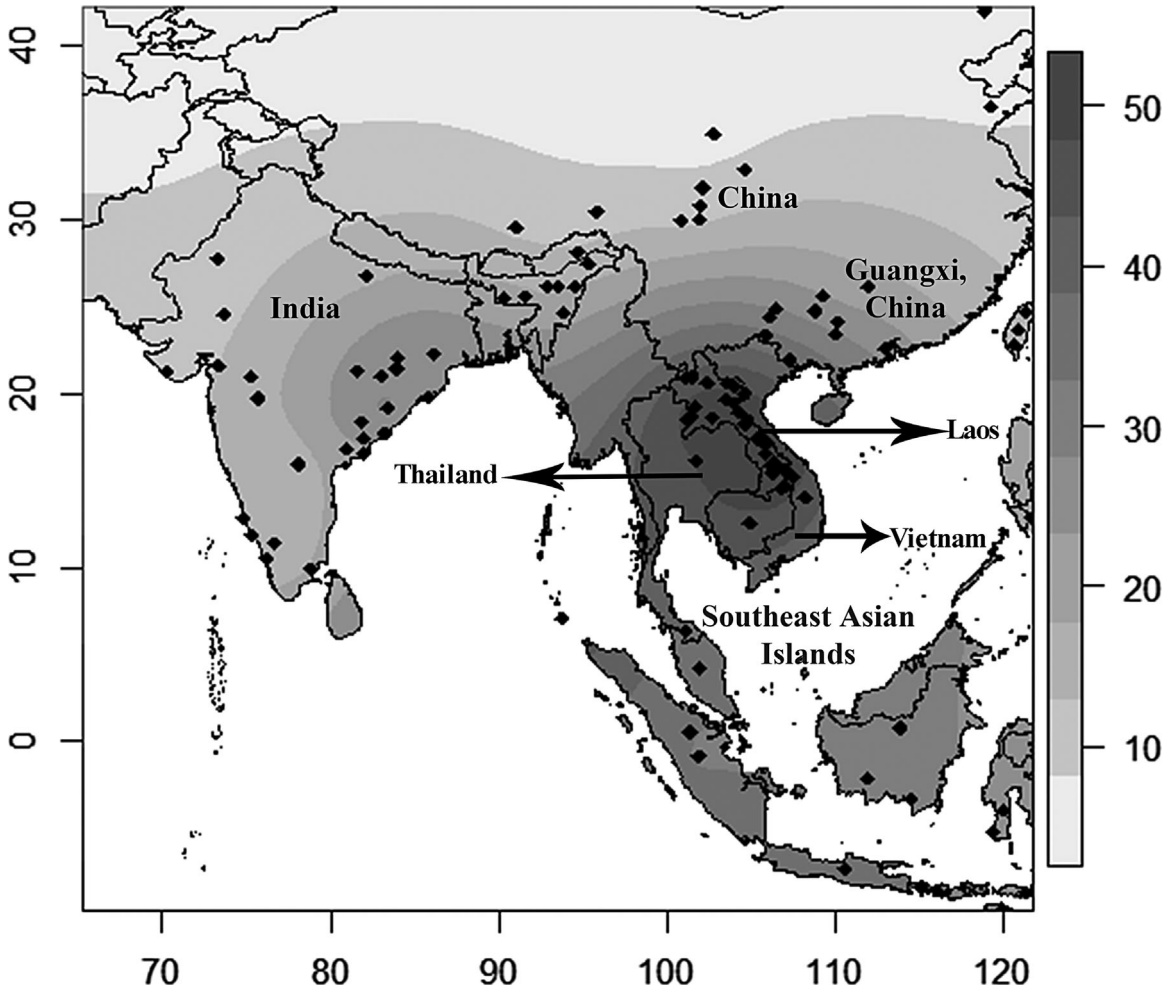

Fig. 1. Contour map of O2a1-M95 frequency distribution in the study region. Dots represent areas of sampling. X axis represents longitude and Y axis represents latitude. The highest frequency of O2a1-M95 in many populations was seen in Laos region. The frequency decreased as a function of distance in all the directions with Laos as the centroid. Shape file for the map is obtained from www.naturalearthdata.com.

variance while, 19/21 Laos and 2/17 E Indian populations showed this value. Although the STR variance found in Guangxi province was the highest, the haplogroup frequency was not consistently high in many populations from this region. Moreover, O2a1-M95 variance in China continental (0.55 ± 0.14) and Guangxi, China (0.57 ± 0.07) were higher than the overall variance estimate of all O2a1-M95 haplotypes put together (0.49 ± 0.01) (Tables 2, S1). A similar picture was obtained for the SE Asian island region as well. This implied that high STR variance combined with low frequency of the haplogroup in these regions were probably not due to long standing expansion. This may be an effect of drift, repeated gene flow from different sources or a combination of these two (Jobling et al., 2004; Tambets et al., 2004). Such a scenario of high diversity coupled with low frequency was identified for haplogroup E-M123 in Turkey (Semino et al., 2004). An expanding agricultural population tends to assimilate newer lineages resulting in convergence of the haplogroup founders and an increase in diversity (Semino et al., 2004). A similar mechanism has operated in South Indian agriculturists (ArunKumar et al., 2012). To test this hypothesis and understand the observation, a study encompassing other Y-chromosomal markers in these Chinese and SE Asian island populations along with samples from Myanmar is required.

O2a1-M95 STR haplotype estimates

High Y-STR variance compounded with samples in consecutive mutational steps from an identified median haplotype in the Y-STR network suggested a long standing evolution of the O2a1-M95 lineages. The 8-STR reduced median network analysis (Fig. S1) revealed the median haplotype (O2a1-M95: 13, 16, 25, 15, 10, 13, 14, 12) in 74 samples distributed in 33 populations, a majority being from Laos (42.4%), and E India (27.3%) (Table S2). In the 14-STR network, the median consisted 45 samples of which 42.2% and 37.7% were from Laos and E India. In the 8-STR-Rst based MDS plot the Laos populations clustered tightly in the centre surrounded by Indian populations while many of the SE Asian island and Chinese populations were seen in the periphery (Fig. 3).

Assuming that the median haplotype is the founding haplotype (Sengupta et al., 2006), haplotypes with shortest SSD from the median may indicate ancient population

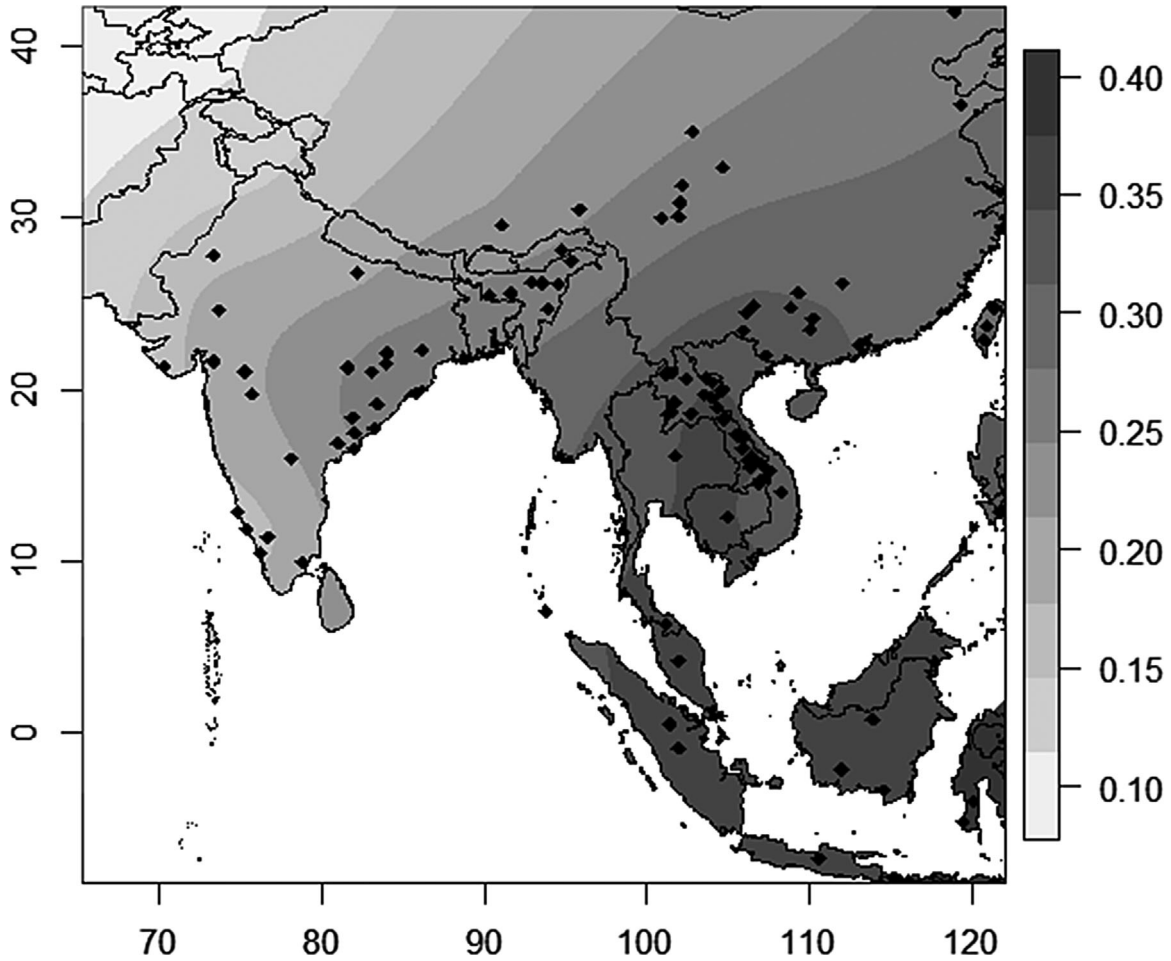

Fig. 2. Contour map of O2a1-M95—8 STR variance distribution in the study region. The contours of O1a1-M95 variance were similar to the frequency map showing a serial decrease in variance from East to West, but with populations of Laos and Southern China showing the highest variances. Shape file for the map is obtained from www.naturalearthdata.com.

expansion in the given region. In this analysis, samples from Laos and SE Asian islands showed 8-STR SSD ranging from 0 to 16 with samples in every step while the E India showed a continuous range from 0 to 11 (Table 3). On the contrary, the Chinese and Nicobarese populations did not possess the median haplotype and presented a discontinuous distribution. The 14-STR SSD analysis also gave a similar picture (Table 3). The observation of 0–19 SSDs (8-STRs) in SE Asian islands samples (Table 3) proposing SE Asian islands as a candidate of place of origin, was negated by the observation that the Bali samples constituted 641/2130 SE Asian island samples with a O2a1-M95 frequency of 0.65 (367/641) (Table 1). This population also possessed higher frequencies of samples with 0–4 SSD amongst the SE Asian island samples. Additionally, the 8-STR variance of SE Asian island samples was 0.4 ± 0.02 (variance $\pm$ SE) which was marginally lower compared to the total O2a1-M95 variance (0.48 ± 0.01) and the overall O2a1-M95 frequency was lower (26.53%) compared to their counterparts in Laos (61.56%) or East India (34.09%) (Table S1). Further, the O2a1-M95 frequency of SE Asian island samples excluding Bali dropped from 26.53% to 13.30% and the 8-STR variance shot up to 0.59 ± 0.05, which was significantly higher compared to the total O2a1-M95 variance (0.48 ± 0.01). In all, the observed high frequency of samples in early mutational steps, considerable O2a1-M95 frequency and appreciable STR variance of SE Asian island samples were all contributed by a single population from Bali. The geographic origin of a paternal lineage has a higher probability of originating in a region with many populations showing high frequency and high STR variance of the lineage than a region with a single population with similar attributes. We do not negate the possibility of O2a1-M95 origin in Bali population completely. A study encompassing complete Y-Chromosomal profile and whole genome markers in this population and a comparison with Laos populations is required to shed further insights on the possible origin of O2a1-M95 in Bali. Thus, the analyses of present dataset exclude the SE Asian islands as the birthplace of O2a1-M95.

The possibility that the median haplotype may change due to sampling bias was also explored. The sampling points were higher in Laos and this could have shifted the median haplotype towards Laos samples. However, a calculation of the median haplotype excluding the Laos samples, showed no change (Table S1). This suggests that the median haplotype and other estimates based on it (network, SSD, ASD, etc) were not biased due to the large number of samples from Laos.

Table 2 O2a1-M95 frequency, 8 and 14 STR variance and age estimates of regional and linguistic population clusters

Grouping	Population group	Populations number	Total number	O2a count	O2a frequency %	8 STR HT			14 STR HT		
						Samples number	Variance ± SE	Age ± SD	Samples number	Variance ± SE	Age ± SD
Region	W India	7	461	49	10.63	49	0.369 ± 0.070	4540 ± 852	49	0.328 ± 0.055	4060 ± 665
	C India	5	236	23	9.75	24	0.293 ± 0.070	3695 ± 872	23	0.347 ± 0.076	4200 ± 931
	S India	11	871	19	2.18	19	0.564 ± 0.158	7005 ± 2003	19	0.419 ± 0.100	5392 ± 1289
	E India	21	1675	571	34.09	571	0.357 ± 0.020	4355 ± 244	571	0.341 ± 0.018	4137 ± 213
	NE India	10	849	118	13.90	118	0.397 ± 0.048	5260 ± 639	118	0.422 ± 0.044	5893 ± 603
	Laos	31	1116	687	61.56	632	0.434 ± 0.024	5665 ± 307	632	0.424 ± 0.021	5447 ± 272
	Thailand	2	600	131	21.83	131	0.506 ± 0.061	8023 ± 951			
	Vietnam	1	80	23	28.75	20	0.338 ± 0.087	4328 ± 1131			
	Cambodia	1	6	3	50.00						
	Guangxi, China	11	508	152	29.92	84	0.577 ± 0.074	7266 ± 933	61	0.604 ± 0.090	7531 ± 1136
	China Continental (excluding Guangxi province)	14	3500	49	1.40	29	0.553 ± 0.140	6902 ± 1772	28	0.658 ± 0.161	7879 ± 1,949
	SE Asian Islands	15	2130	565	26.53	548	0.400 ± 0.022	6537 ± 342	95	0.578 ± 0.067	7293 ± 847
	Nicobar	2	23	23	100.00	11	0.511 ± 0.167	7127 ± 2330	11	0.419 ± 0.114	5590 ± 1538
Language family	MonKhmer East	13	477	288	60.38	278	0.434 ± 0.035	5631 ± 448	278	0.397 ± 0.030	5163 ± 378
	MonKhmer Nicobar	1	11	11	100.00	11	0.511 ± 0.167	7127 ± 2330	11	0.419 ± 0.114	5590 ± 1538
	MonKhmer North	6	291	171	58.76	153	0.365 ± 0.040	4716 ± 516	153	0.365 ± 0.036	4720 ± 459
	MonKhmer Viet-Muong	3	146	68	46.58	62	0.548 ± 0.092	7684 ± 1307	43	0.501 ± 0.096	7117 ± 1349
	MonKhmer Aslian	1	12	12	100.00						
	MonKhmer Total	24	937	550	58.70	504	0.447 ± 0.027	5679 ± 345	485	0.414 ± 0.024	5340 ± 304
	Munda North	4	406	206	50.74	206	0.311 ± 0.029	3864 ± 351	206	0.272 ± 0.023	3268 ± 275
	Munda South	4	224	139	62.05	139	0.412 ± 0.044	5184 ± 549	139	0.435 ± 0.043	5514 ± 540
	Munda/Dravidian	1	105	26	24.76	26	0.370 ± 0.088	4825 ± 1140	26	0.310 ± 0.064	4068 ± 834
	Munda/IndoEuropean	1	96	7	7.29	7	0.667 ± 0.262	8909 ± 3513	7	0.422 ± 0.123	5823 ± 1741
	Munda Total	10	831	378	45.49	378	0.367 ± 0.026	4464 ± 309	378	0.355 ± 0.023	4275 ± 274
	Tibeto-Burman Central	2	117	3	2.56	13	0.274 ± 0.099	3427 ± 1253	13	0.277 ± 0.096	3364 ± 1180
	Tibeto-Burman Ngwi-Burmese	1	27	13	48.15	8	0.522 ± 0.231	6266 ± 2812	8	0.522 ± 0.182	6211 ± 2171
	Tibeto-Burman Sal	5	484	82	16.94	82	0.376 ± 0.054	5339 ± 754	82	0.386 ± 0.049	4960 ± 629
	Tibeto-Burman Western	3	1303	8	0.61	5	0.688 ± 0.426	8909 ± 5769	5	0.914 ± 0.536	12228 ± 7256
	Tibeto-Burman Northeastern	4	558	9	1.61						
	Tibeto Burman Total	15	2489	115	4.62	111	0.414 ± 0.053	5495 ± 696	111	0.486 ± 0.058	6190 ± 732
	Austronesian	18	2429	574	23.63	555	0.402 ± 0.022	6551 ± 341	101	0.586 ± 0.065	7531 ± 841

Continued

Table 2 Continued

Grouping	Population group	Populations number	Total number	O2a count	O2a frequency %	8 STR HT			14 STR HT		
						Samples number	Variance ± SE	Age ± SD	Samples number	Variance ± SE	Age ± SD
	Daic	17	1096	364	33.21	324	0.461 ± 0.035	6488 ± 493	193	0.476 ± 0.043	5734 ± 518
	Dravidian Central	9	623	202	32.42	201	0.343 ± 0.031	4489 ± 404	201	0.291 ± 0.025	3781 ± 314
	Dravidian South	13	1019	46	4.51	46	0.387 ± 0.071	4782 ± 889	46	0.341 ± 0.051	4210 ± 626
	Sino-Tibetan Chinese	5	1291	104	8.06	40	0.515 ± 0.089	6642 ± 1210	39	0.536 ± 0.097	6661 ± 1244
	Indo European	12	834	37	4.44	38	0.335 ± 0.074	3927 ± 868	38	0.339 ± 0.065	4113 ± 790
	Hmong Mein	5	164	33	20.12	31	0.314 ± 0.059	4386 ± 813	9	0.671 ± 0.188	9317 ± 2556
	Altaic	2	292	9	3.08	9	0.455 ± 0.21	6168 ± 2928	9	0.526 ± 0.236	6988 ± 3184
	Afro Asiatic/Indo European	1	50	1	2.00						
	Grand Total O2a1-M95		12055	2413	20.02	2239	0.489 ± 0.014	6434 ± 183	1611	0.451 ± 0.014	5627 ± 176

HT, Haplotype; SE, Standard error; SD, Standard deviation.

Ancestral haplogroups and origin of O2a1-M95

The place of origin of a haplogroup may also need to harbour immediate ancestors and subtypes of the given haplogroup (Sahoo et al., 2006). Samples from Laos harbored the highest frequency of O2a1-M95, considerable frequencies of ancestral haplogroups, O*-M175 (2%), O2*-P31 (2%) and the subtype O2a1a-M88 (6%) (Hai, Pers. Comm.; Cai et al., 2011). These haplogroups were negligible in China continental and Guangxi, China (Cai et al., 2011), nil O2*-P31 and negligible O*-M175 (0.2% in Munda) in E India, and nil O*-M175 and O2*-P31 in SE Asian islands (Karafet et al., 2010). But the recently discovered O2*-PK4 may need to be genotyped in the M*-M175 samples. This left Laos as a candidate for deepest O2a1-M95 antiquity.

We also observed high frequency but lower STR diversity of O2a1-M95 chromosomes in Indian Austro Asiatic speakers. This can be attributed to a more recent founder effect and expansion in isolation. The mean O2a1-M95 frequency in Indian Austro Asiatic (Munda) speaking tribes in the present study was 44% (range 7.29–79.76%). This estimate is lower than that of Kumar et al. (2007) (mean = 54.04%). In the absence of samples from Laos region and based on higher O2a1-M95 frequency and STR diversity in Munda speaking populations, previous studies have suggested the origin of this lineage in E India (Kumar et al., 2007; Reddy et al., 2007).

A recent study has described very high frequencies of O2a1-M95 ranging from 51.85% to 100% in North-eastern Cambodia, although the STR data for these samples were not available for comparison (Zhang et al., 2014). We note that these populations are only at a radius of ~300 km from the regions we had sampled in Laos, implying it does not differ from our conclusion of a deep antiquity of O2a-M95 in the region of Laos.

We hasten to add that neither The Genographic (we) nor other researchers have so far extensively sampled Myanmar, a vast expanse between North-East India and Thailand. Nonetheless, only 7% of the Myanmarese population speaks Austro Asiatic languages (Lewis et al., 2014) and hence we believe that the inclusion of samples from this region would not have significantly altered the observed picture, unless the Myanmar region contains appreciable O2a1-M95 in other language speakers. A study on samples from this interim region is thus essential before coming to a conclusion on the deep antiquity of O2a1-M95 in Laos.

O2a1-M95 age estimates

The expansion time of a NRY haplogroup is normally estimated using Average Squared Distance (ASD) of STRs, that reflects the STR variance accumulated over a period of time (Sengupta et al., 2006). We calculated the expansion time based on genealogical mutation rates on both the 8 and 14-STR datasets (Goedbloed et al., 2009).

An overall O2a1-M95 expansion time of 6.4 ± 0.1 Kya for 8-STR and 5.6 ± 0.1 Kya for 14-STR datasets was obtained (Table 2). Among the various regions studied, samples from Laos showed a comparable estimate of 5.7 ± 0.3 and 5.4 ± 0.2 Kya with 8 and 14 STRs, while those of China continental, Guangxi of China and SE Asian islands showed much higher estimates of >6.5 and 7.2 Kya with larger confidence intervals (Table 2). Considering the fact that ASD age estimates are directly related to STR diversity

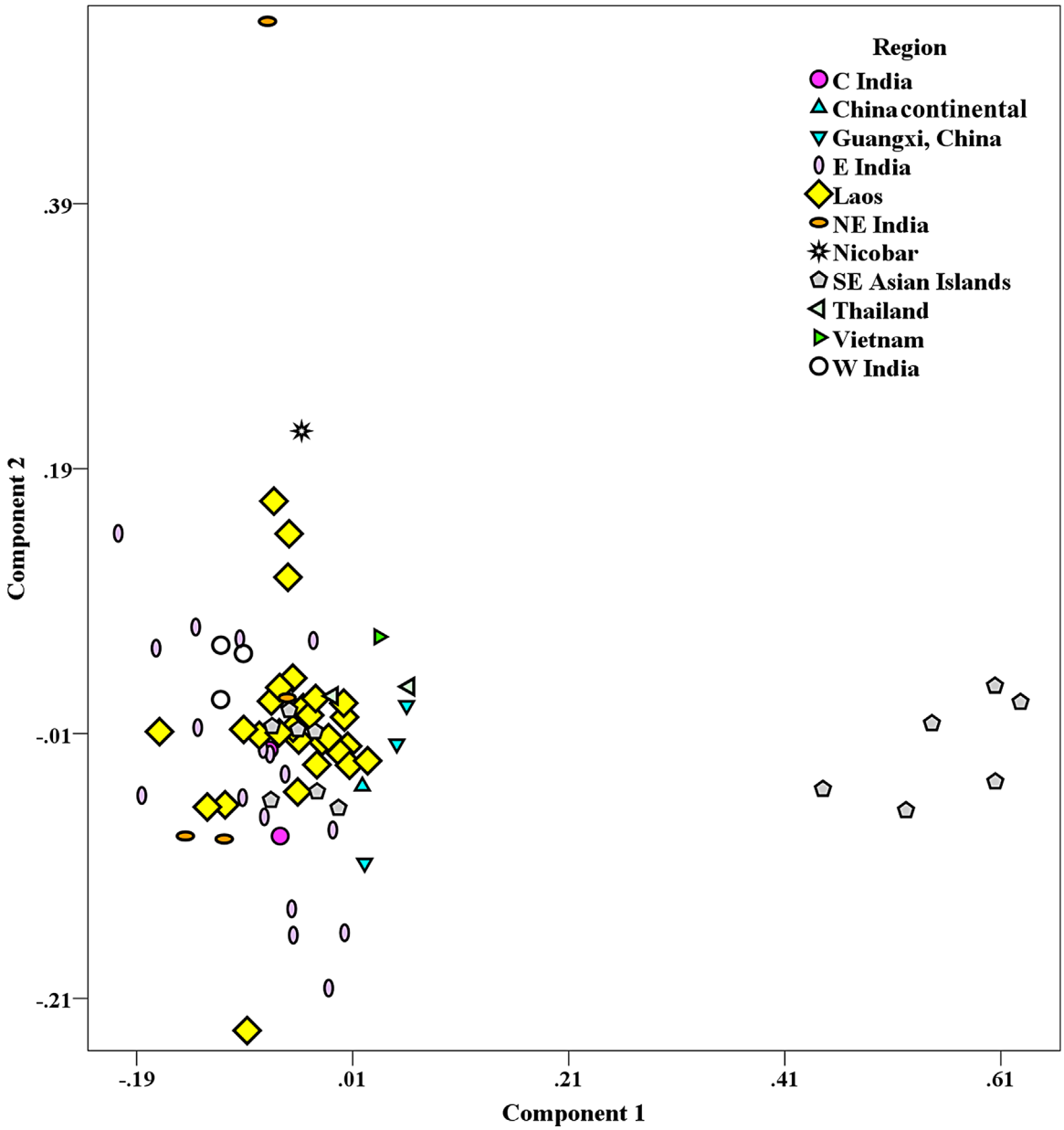

Fig. 3. Multidimensional scaling plot based on 8 STR Rst distance. The MDS was based on 8 STR Rst distance of O2a1-M95 samples studied. Stress value = 16.84. Most of the Laos populations clustered tight in the middle while E Indian populations were loosely distributed around this. SE Asian island populations formed two distinct clusters one lying with Laos and other distantly in the periphery.

(Zhivotovsky et al., 2004) the explanation that the higher variance in China was not due to long standing concerted evolution may hold good here. The scenario of drift leading to such an effect may need to be investigated considering all the NRY haplogroups present in the Chinese samples.

Further, one may also need to look at the even distribution of STR alleles in various geographic regions studied. The STRs of 54 SE Asian island samples from Banjarmasin, Palu, Kota Kinabalu, Mataran, Pekanbaru, and Toraja populations, possessed DYS389a allele sizes 9-11, constituting 9.8% of SE Asian island samples. Interestingly all other O2a1-M95 chromosomes reported here showed a higher allele size (10–16), with the exception of only 12 non SE Asian island samples falling within the lower range (Chaubey et al., 2011) (Table S2). This difference led to the formation of a separate cluster of the six SE Asian island populations in the MDS plot attributing these events to a founder and expansion. The other SE Asian island populations (Sulawesi, Borneo, Java, Bali, Malay, Sumatra, and Flores) clustered with the major cluster of other regions, indicating two different migrations of O2a1-M95 into SE Asian islands.

The coalescence time of O2a1-M95 in the present study dated to 6.4 ± 0.4 Kya for the 8-STR dataset (Table 2). This estimate, based on the variance, describes the time taken for the population to accumulate variance that is observed today and not the time of origin of the haplogroup. This time depth

Table 3 Distribution of sum of squared distances (8 and 14 STR) in various regional populations studied

			Frequency (×100) distribution of sum of squared distances																								
		No. of samples	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	>25
8-STR haplotype	W India	49	2.0	30.6	36.7	4.1	10.2	10.2									2.0					2.0						2.0
	S India	19	5.3	5.3	21.1	21.1	5.3		5.3	21.1		5.3			10.5												
	C India	23	4.4	13.0	21.7	47.8	8.7			4.4																	
	E India	571	3.9	15.8	25.0	24.2	11.7	6.5	4.6	6.1	0.5	0.2	0.2	0.9										0.2	0.2			0.2
	NE India	118	4.2	17.0	13.6	16.1	12.7	9.3	6.8	16.1	2.5		1.7															
	Nicobar	11		9.1	9.1	27.3	9.1							27.3	9.1			9.1										
	Laos	632	2.9	9.5	21.0	18.8	11.9	8.1	6.7	9.2	1.7	2.9	1.3	1.9	2.9	0.5	0.6	0.2	0.2									
	Thailand	131		6.1	13.7	22.1	17.6	4.6	4.6	8.4	1.5	3.8	1.5	2.3	5.3	3.8	1.5		0.8			1.5						0.8
	Vietnam	20		10.0	15.0	5.0		45.0	15.0	5.0		5.0																
	Cambodia	3			33.3					66.7																		
	SE Asian Islands	548	4.7	21.9	21.4	21.5	7.5	4.2	3.5	3.3	0.9	1.6	0.9	3.8	1.8	0.7	0.4	0.4	0.2	0.4		0.2				0.4		0.4
	Guangxi, China	84		2.4	6.0	26.2	9.5	9.5	3.6	4.8	3.6	1.2	9.5	16.7	3.6			1.2	2.4									
	China continental*	29		3.5	6.9	31.0	17.2	3.5	10.3		10.3	6.9		3.5	3.5				3.5									
14-STR haplotype	W India	49	2.0	8.2	24.5	18.4	6.1	14.3	6.1	4.1	2.0	2.0	2.0	2.0				2.0					4.1					2.0
	S India	19		5.3	10.5	15.8	10.5	15.8			15.8	10.5	5.3			5.3					5.3							
	C India	23		8.7	4.4	13.0	17.4		30.4	8.7	8.7	8.7																
	E India	571	0.5	4.7	11.4	17.7	17.7	13.1	9.6	7.7	4.9	3.9	2.3	0.9	3.5	0.2		0.5		0.2				0.2				1.1
	NE India	118		4.2	6.8	13.6	11.9	7.6	5.1	11.9	3.4	4.2	4.2	3.4	4.2	4.2	0.9	7.6	4.2		0.9		0.9	0.9				
	Nicobar	11		9.1			27.3		18.2						9.1	9.1		18.2	9.1									
	Laos	632	1.1	3.2	7.1	10.3	12.3	9.2	7.9	8.2	8.1	6.3	5.5	2.7	2.9	2.4	2.9	1.7	1.9	2.1	0.2	0.5	0.2	1.6	0.5	0.3	0.2	1.0
	Thailand																											
	Vietnam	1								100.0																		
	Cambodia	3				33.3						33.3														33.3		
	SE Asian Islands	95	1.1	2.1	7.4	4.2	8.4	1.1	3.2	4.2	2.1	2.1	4.2	6.3	4.2	13.7	9.5	6.3	3.2	3.2		4.2		1.1	1.1		1.1	6.3
	Guangxi, China	61					8.2	3.3	23.0		3.3		8.2		1.6	1.6	4.9	3.3	4.9	3.3	6.6	1.6	1.6	3.3	1.6	1.6	3.3	14.8
	China continental*	28				3.6	10.7		3.6	3.6	10.7	17.9			10.7	10.7					3.6		7.1	3.6				14.3

Presence of samples in appreciable frequencies in 0 and the earlier consecutive SSDs may indicate the presence of ancestral haplotypes in the population. Greater the SSD, higher the expansion provided we see samples in consecutive SSDs. Samples with sporadically and discontinuously distributed SSDs may reflect drift, influx into the populations, or insufficient sampling. *China continental refers to China excluding Guangxi province.

correlates to the time scale of putative Austro Asiatic language family chronology postulated by Diffloth (2005) and Peiros (cross reference Sidwell, 2009) based on historical linguistics and Glottochronological lexicostatistics, respectively (Diffloth, 2005; Sidwell, 2009; Peiros, 2011). Although the linguistic tree topology varies in these two studies, there is a consensus in the time scale of 7 Kya for the branching of Austro Asiatic languages. These two trees also have shown Mon-Khmer as an early off-shoot and Munda a relatively recent one. The coalescent time (ASD) estimates of all Mon-Khmer and Munda speaking populations in the present study (5.7 ± 0.3 Kya & 4.3 ± 0.3 Kya) corroborate these linguistic tree chronologies (Table 2). This correlation may need to be validated based on analysis including samples from Myanmar and other genomic markers.

Migration to India

Contrary to the Laos population, E Indian O2a1-M95 showed an expansion time of 4.3 ± 0.2 and 4.1 ± 0.2 Kya (8 and 14-STR, respectively), less than the NE Indian samples (5.2 ± 0.6 and 5.8 ± 0.6 Kya), and also much smaller than the Laos samples (5.7 ± 0.3 and 5.4 ± 0.2 Kya) (Table 2). This was also reflected in the decrease of STR variance from Laos to India (Fig. 2). This decrease in the age and variance from east to west (Laos to India) suggests a migratory path of O2a1-M95 lineage during the late Neolithic. Further, the Indian samples did not show a region specific clustering or a unique founder event in the reduced median network (Figs. S1, S2). Thus, the origin of O2a1-M95 lineages in India may not be attributed to a single founding event, but may be a result of multiple migratory events as suggested earlier by Chaubey et al. (2011).

Archaeological and Paleobotanical evidences suggest two Neolithic traditions in E India (Orissa) dating back to 5 Kya in the river valleys and coastal plains, and in the foothills and uplands (Harvey et al., 2006). The ``Munda'' branch of Austro Asiatic language family is spoken in the whole of Orissa, and the ``Khasi'' branch in isolated pockets of Meghalaya (North East India). Both the Munda and Khasi speakers are primarily agriculturists practising seasonal agriculture with less sedentary lifestyle (Harvey et al., 2006). Their coalescence time observed in the present study also dated back to the same time (4.3 ± 0.2 Kya and 3.3 ± 0.6 Kya) (Tables 2, S1). In the light of similar evidences obtained in most of the Munda speakers from Orissa, it is possible that the O2a1-M95 carrying people spread to this region from the East, along with agriculture, in multiple migrations and gave rise to the Munda languages.

We also note that the STR based time calibration was not affected by the choice of STR. The average STR variance and the ASD age estimates of the 8-STR and the 14-STR datasets were similar with overlapping confidence intervals in most of the cases (Tables 2, S1). In the light of the work by Busby et al. (2012) this could be attributed to similar average STR linearity of both the 8 and 14 STR datasets (average range of STR loci = 6.5 and 6.4 for 8 and 14 STR dataset, respectively) (Busby et al., 2012).

Conclusion

The present study has provided evidences of high O2a1-M95 frequency, associated STR variance and larger expansion time estimates in Laos among the sampled regions. This, along with the presence of ancestral haplogroups O*-M175 and O2*-P31 suggests the deep antiquity of this lineage in the Laos region. Further the serial decrease in the age estimates of O2a1-M95 from Laos to India suggests a late Neolithic east to west expansion of this lineage. A study on other polymorphisms and whole genomes scans with the inclusion of samples from Myanmar and other nearby regions is required to confirm the present study.

Acknowledgements

We gratefully acknowledge the volunteers from India, China and the Mekong Valley who participated in this study. We also thank the various community leaders and field work assistants who participated in sampling expeditions and helped in sample identification. A special acknowledgement to Mr. Sreekanth Patel of Laikara College for his help in sampling in Orissa. The Genographic Project was supported and funded by National Geographic Society, IBM, and The Ted Wait Family Foundation, USA.

Conflict of Interest

The authors declare no conflict of interest. The funding agencies (National Geographic Society, IBM, and The Ted Wait Family Foundation) had no role in the design of the study, the collection and analysis of data and the decision to publish.

References

Arunkumar G, Soria-Hernanz DF, Kavitha VJ, Arun VS, Syama A, Ashokan KS, Gandhirajan KT, Vijayakumar K, Narayanan M, Jayalakshmi M, Ziegle JS, Royyuru AK, Parida L, Wells RS, Renfrew C, Schurr TG, Smith CT, Platt DE, Pitchappan R, Genographic Consortium. 2012. Population differentiation of Southern Indian male lineages correlates with agricultural expansions predating the caste system. *PLoS ONE* 7: e50269.

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K. 2002. *Short Protocols in Molecular Biology*. New York: Wiley.

Bandelt HJ, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* 141: 743–753.

Busby GB, Brisighelli F, Sanchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, Thomas MG, Bradley DG, Gusmao L, Winney B, Bodmer W, Vennemann M, Coia V, Scarnicci F, Tofanelli S, Vona G, Ploski R, Vecchiotti C, Zemunik T, Rudan I, Karachanak S, Toncheva D, Anagnostou P, Ferri G, Rapone C, Hervig T, Moen T, Wilson JF, Capelli C. 2012. The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proceedings of the Royal Society B: Biological Sciences* 279: 884–892.

Cai X, Qin Z, Wen B, Xu S, Wang Y, Lu Y, Wei L, Wang C, Li S, Huang X, Jin L, Li H. 2011. Human migration through bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum revealed by Y chromosomes. *PLoS ONE* 6: e24282.

Chaubey G, Metspalu M, Choi Y, Magi R, Romero IG, Soares P, van Oven M, Behar DM, Rootsi S, Hudjashov G, Mallick CB, Karmin M, Nelis M, Parik J, Reddy AG, Metspalu E, van Driem G, Xue Y, Tyler-Smith C, Thangaraj K, Singh L, Remm M, Richards MB, Lahr MM,

Kayser M, Villems R, Kivisild T. 2011. Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. *Molecular Biology and Evolution* 28: 1013–1024.

Diffloth G. 2005. The contribution of linguistic palaeontology to the homeland of Austroasiatic. In: Sagart L, Blench R, Sanchez-Mazas A eds. *The peopling of East Asia: putting together archaeology, linguistics and genetics*. London: Routledge Curzon. 77–80.

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 1: 47–50.

Goedbloed M, Vermeulen M, Fang RN, Lembring M, Wollstein A, Ballantyne K, Lao O, Brauer S, Kruger C, Roewer L, Lessig R, Ploski R, Dobosz T, Henke L, Henke J, Furtado MR, Kayser M. 2009. Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR Yfiler PCR amplification kit. *International Journal of Legal Medicine* 123: 471–482.

Haber M, Platt DE, Badro DA, Xue Y, El-Sibai M, Bonab MA, Youhanna SC, Saade S, Soria-Hernanz DF, Royyuru A, Wells RS, Tyler-Smith C, Zalloua PA. 2011. Influences of history, geography, and religion on genetic structure: the Maronites in Lebanon. *European Journal of Human Genetics* 19: 334–40.

Harvey EL, Fuller DQ, Mohanty RK, Basanta M. 2006. Early agriculture in Orissa: some archeobotanical results and field observations on the Neolithic. *Man and Environment* XXXI: 21–32.

Higham CFW. 2002. Languages and farming dispersals: Austroasiatic languages and rice cultivation. In: Bellwood PRC ed. *Examining the farming/language dispersal hypothesis*. Cambridge: McDonald Institute. 223–332.

Jobling M, Hurles M, Tyler-Smith C. 2004. *Human evolutionary genetics: origins, peoples & disease*. New York: Garland Science.

Kantang S, Kusamran T, Waiyawuth W. 2010. Allele Frequecncies and Haplotybe Diversity of 12 Loci Male-Specific Y-Chromosome (STRs) among Thai Population in Yala Province of Thailand. Proceedings of the 5th CIFS Academic Day. 1–3 September, Bangkok, Thailand.

Karafet TM, Hallmark B, Cox MP, Sudoyo H, Downey S, Lansing JS, Hammer MF. 2010. Major east-west division underlies Y chromosome stratification across Indonesia. *Molecular Biology and Evolution* 27: 1833–1844.

Kruskal JB. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29: 1–27.

Kumar V, Reddy AN, Babu JP, Rao TN, Langstieh BT, Thangaraj K, Reddy AG, Singh L, Reddy BM. 2007. Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. *BMC Evolutionary Biology* 7: 47.

Lewis MP, Gary FS, Charles DF. 2014. Ethnologue: Languages of the World. Dallas, Texas: SIL International [online]. Available from www.ethnologue.com [accessed 1 October 2014].

Myres NM, Rootsi S, Lin AA, Jarve M, King RJ, Kutuev I, Cabrera VM, Khusnutdinova EK, Pshenichnov A, Yunusbayev B, Balanovsky O, Balanovska E, Rudan P, Baldovic M, Herrera RJ, Chiaroni J, Di Cristofaro J, Villems R, Kivisild T, Underhill PA. 2011. A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *European Journal of Human Genetics* 19: 95–101.

Peiros I. 2011. Some thoughts on the problem of the Austro-Asiatic homeland. *Journal of Language Relationship* 6: 101–113.

R Development Core Team. 2010. R: a language and environment for statistical computing. Vienna: R foundation for statistical computing.

Reddy BM, Langstieh BT, Kumar V, Nagaraja T, Reddy AN, Meka A, Reddy AG, Thangaraj K, Singh L. 2007. Austro-Asiatic tribes of Northeast India provide hitherto missing genetic link between South and Southeast Asia. *PLoS ONE* 2: e1141.

Sahoo S, Singh A, Himabindu G, Banerjee J, Sitalaximi T, Gaikwad S, Trivedi R, Endicott P, Kivisild T, Metspalu M, Villems R, Kashyap VK. 2006. A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proceedings of the National Academy of Scinces USA* 103: 843–848.

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

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, Lin AA, Mitra M, Sil SK, Ramesh A, Usha Rani MV, Thakur CM, Cavalli-Sforza LL, Majumder PP, Underhill PA. 2006. 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. *American Journal of Human Genetics* 78: 202–221.

Sidwell P. 2009. *Classifying Austro Asiatic languages: history and state of the art*. Gmunderstr: LINCOM GmbH.

Sidwell P. 2010. The Austroasiatic central riverine hypothesis. *Journal of Language Relationship* 4: 117–134.

Siriboonpiputtana T, Jomsawat U, Rinthachai T, Thanakitgosate J, Shotivaranon J, Limsuwanachot N, Polyorat P, Rerkamnuaychoke B. 2010. Y-chromosomal STR haplotypes in Central Thai population. *Forensic Science International Genetics* 4: e71–e72.

Tambets K, Rootsi S, Kivisild T, Help H, Serk P, Loogvali EL, Tolk HV, Reidla M, Metspalu E, Pliss L, Balanovsky O, Pshenichnov A, Balanovska E, Gubina M, Zhadanov S, Osipova L, Damba L, Voevoda M, Kutuev I, Bermisheva M, Khusnutdinova E, Gusar V, Grechanina E, Parik J, Pennarun E, Richard C, Chaventre A, Moisan JP, Barac L, Pericic M, Rudan P, Terzic R, Mikerezi I, Krumina A, Baumanis V, Koziel S, Rickards O, De Stefano GF, Anagnou N, Pappa KI, Michalodimitrakis E, Ferak V, Furedi S, Komel R, Beckman L, Villems R. 2004. The western and eastern roots of the Saami-the story of genetic ``outliers`` told by mitochondrial DNA and Y chromosomes. *American Journal of Human Genetics* 74: 661–682.

Trivedi R, Sahoo S, Singh A, Bindu GH, Banerjee J, Tandon M, Gaikwad S, Rajkumar R, Sitalaximi T, Ashma R, Chainy GBN, Kashyap VK. 2008. Genetic imprints of Pleistocene origin of Indian populations: A comprehensive phylogeographic sketch of Indian Y-Chromosomes. *International Journal of Human Genetics* 8: 97–118.

Wu FC, Ho CW, Pu CE, Hu KY, Willuweit S, Roewer L, Liu DH. 2011. Y-chromosomal STRs haplotypes in the Taiwanese Paiwan population. *International Journal of Legal Medicine* 125: 39–43.

Zhang X, Kampuansai J, Qi X, Yan S, Yang Z, Serey B, Sovannary T, Bunnath L, Aun HS, Samnom H, Kutanan W, Luo X, Liao S, Kangwanpong D, Jin L, Shi H, Su B. 2014. An updated phylogeny of the human Y-chromosome lineage O2a-M95 with novel SNPs. *PLoS ONE* 9: e101020.

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

Supplementary Material

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.12147/suppinfo:

Fig. S1. Reduced median network of 8 STR O2a1-M95 haplotypes. Samples are colored based on their geographic location. The size of the circle is proportional to the number of samples with the same haplotype. The length of the branch is proportional to the number of mutational steps between haplotypes. The network showed clear expansion of the O2a1-M95 lineage. No region specific clusters were identified suggesting a recent expansion of the lineage.

Fig. S2. Reduced median network of 14 STR O2a1-M95 haplotypes. Samples are colored based on their geographic location. The Chinese and SE Asian Island samples were found in the periphery of the network.

Table S1. The list of study populations from various regions and their 8 and 14 STR Variance based Age estimate.

Table S2. List of samples studied by the Genographic centres of India and China included in the present analysis.