CHAPTER FOUR


RESULTS AND DISCUSSION

Results and Discussion:

Part-I:

According to the maternally inherited mitochondrial DNA (mtDNA) phylogeny, the world population is divided into two broad categories - African and non-African (Ingman *et al.* 2000). All non-African lineages derive from just two limbs - M and N branching out from the root of African haplogroup L3. The contemporary distribution of non-African mtDNA haplogroups M and N, and their chronological and demographic expansion, reveal that there was an early single rapid dispersal event from the Horn of Africa via the southern coastal route which took to peopling South Asia, East Asia and Southeast Asia/Island Southeast Asia (Macaulay *et al.* 2005; Sun *et al.* 2006; Hill *et al.* 2006; Hill *et al.*2007). The exodus of non-African founding group from East Africa occurred after the appearance of haplogroup L3 ~ 85 thousand years ago (kya) (Macaulay *et al.* 2005). However, the precise geographic origins of non-African founders M and N are unknown and are subject to continuing controversy as to whether they arose in Africa just prior to the exodus, just after the dispersal into west Asia, or even further east into South Asia. The lack of ancient mtDNA lineages in Arabia and Southwest Asia (Abu-Amero K *et al.* 2008), and the highest diversity of basal haplogroups radiating from the root of non-African founding groups M and N found in South Asia, suggest that it was the locale of the first major expansion outside Africa and perhaps acted as an incubator in the genetic differentiation of Eastern Asian gene pools (Kivisild *et al.*2003). However, no population or geographic region has been identified to date from South Asia that has played an initial role in the genetic differentiation of Eastern Asian mtDNA gene pools.

In the last 10 years, numerous studies have evaluated the distribution and extent of variation of haplogroups in South Asia populations. The evidence of deep-rooting macro-haplogroups M and N lineages suggests that South Asians are likely to have been settled in this region since the middle Paleolithic period. The progress towards complete mtDNA sequences has provided a number of South Asian autochthonous haplogroups (Sun C *et al.* 2006; Palanichamy *et al.* 2004 ; Thangaraj *et al.* 2006 ; Reddy *et al.* 2007) . These haplogroups include M2-M6, M18, M30-M41, M43, M48-M53, N5, U2a-U2c, R5-R8, R30 and R31. Among these, the origin and spatial distribution have been described for only five haplogroups thus far - M2, M31, M32, R7, and R8 (Kumar

et al. 2008 ; Barik *et al.* 2008; Chaubey *et al.* 2008 ; Thangaraj *et al.* 2009). From the phylogenetic hierarchical evidence (except Andaman Island specific-M32), haplogroups M2, M31, R7, and R8 were found to have originated from the eastern part of India (Kumar *et al.* 2008; Thangaraj *et al.* 2009). The remaining South Asian haplogroups expansion processes remain poorly understood.

The mtDNA phylogeny of Southeast/Island Southeast Asia populations shows that most of the lineages branched directly from the Eurasian founder haplogroups (viz., M and N) ~ 60 kya, and are indigenous and unique to this region (Macaulay *et al.*2005; Hill *et al.* 200 6 ; Hill *et al.* 2007). These findings were interpreted as a very fast southern dispersal of modern humans to Asia (Macaulay *et al.*2005) . Recently two Southeast/Island Southeast Asia indigenous lineages (M21b and M21c) were identified from the northeast region of India-Nepal and Myanmar region (Dancause *et al.*2009 ; Fornarino *et al.* 2009). These findings provided the first direct phylogenetic evidence for the common ancestry among the Southeast/Island Southeast Asia and northeast Indian populations. The many defined autochthonous haplogroups in Southeast/Island Southeast Asians and the lack of close matches in the extant South and East Asia populations do not overthrow the previous rapid settlement interpretation (Macaulay *et al.* 2005). However, it should be noted that the populations' coverage from northeast India and neighboring regions, and the range of phylogenetic mtDNA data obtained so far are insufficient for a definitive conclusion. An additional issue is the excessive emphasis placed on searching world databases for control region motifs unique to Southeast/Island Southeast Asia populations, instead of analyzing information from coding region sequences.

In this study, intensive efforts had been made to show the evolutionary relationships among the South Asia, East Asia and Southeast/Island Southeast Asia lineages and in areas where these Eastern Asian populations were initially differentiated. The complete genome-sequencing approach was used, in addition to control-region sequence information, to study mtDNA variation in a large number of populations representing South Asia (India, Nepal, Bhutan, and Bangladesh) and Southeast/Island Southeast Asia (Thailand and Malaysia).

Results:

Updating South Asian mtDNA haplogroup M phylogeny

The phylogenetic analysis of south Asian-specific haplogroup M mtDNAs confirmed eighteen previously described independent basal haplogroup branches and a subclade (M4'30) that has seven basal haplogroup branches (Sun *et al.* 2006; Thangaraj *et al.* 2006; Reddy *et al.* 2007; Fornarino *et al.* 2009 ; Thangaraj *et al.* 2008)(**Figures S1-S5**). We have checked the mutations in the control or coding region of several unassigned lineages relative to previous Indian mtDNA classification system in our data set, and have characterized three new haplogroups M54, M55, and M56 (**Figures S1 and S3**). Haplogroup M54 was characterized by four control (240, 390T, 16051, and 16316) and six coding (572, 5493, 5821, 9302, 11167, and 11560) region mutations. Haplogroup M55 was characterized by five control region (199, 204, 16086, 16218, and 16362) and six coding region mutations (1313, 10754, 11908, 12366, 14299 and 15930). Haplogroup M56 was defined by the presence of a 9-bp deletion in the COII/$tRNA^{Lys}$ intergenic region with six transitions motifs (4943, 8270, 8623, 14552, 15924, and 16290).

Results from this study had helped to define several subhaplogroups based on information obtained from coding and control region motifs. Subhaplogroups M3b and M3c were defined by coding and control region transitions (11827-16344 and 195-3394-11167- 11914-15951-16185, respectively) (**Figure S2**). The definition of subhaplogroup M5a was revised here: there are now only two coding region mutations (709 and 3921) which define this haplogroup. In addition, subhaplogroup M5b was defined by three coding region mutations (4851, 5319, and 6413), while M5c was defined by two coding region mutations (8784T and 13368) (**Figure S1**). M18b was defined by a single control (16325) and four coding region (5153, 8277, 8279, and 14587) mutations (**Figure S1**). Subhaplogroups M30e and M30f were characterized by control and control region motifs, 15314-16111 and 8405-9053-16293, respectively (**Figure S5**). Using additional sequence information from Southeast Asia, we revised the definition of subhaplogroup M33b, which was characterized by a coding and control region mutation, 3221 and 16324, respectively (**Figure S1**). M37b was characterized by a single coding region (11050) variant and two control region (16111 and 16189) mutations (**Figure S1**).

Based on data from complete mtDNA sequences, we redefined the haplogroups M2, M6,

M34, M35, and M43. Haplogroup M2 was broadened by requiring only five mutations, 447G, 1780, 11083, 15670, and 16274. M2 could be divided into two major branches, M2a and M2b. M2a was characterized by two mutations (8502 and 16319), while M2b was characterized by four coding region mutations (1598, 4216, 13212, and 14800). Additionally, our phylogenetic analysis confirmed that two major subclades existed within subhaplogroup M2a, which we have designated as M2a1 and M2a2 (corresponding to the former haplogroups M2a and M2b), respectively (Sun *et al.*2006; Kivisild *et al.*(2003 ; Kumar *et al.* 2008). Additionally, we demonstrate here a number of M2a1 and M2a2 subclusters (M2a1a, M2a1b, M2a1c, M2a2a, and M2a2b), which are illustrated in figure S3.

Comprehensive analysis of complete mtDNA sequences in this present study from different control region motifs of M6 lineages broadened the definition of haplogroup M6. The phylogenetic tree pointed to 461, 5301, 10640, 14128, and 16362 as the basal mutations for haplogroup M6, whereas 3537, 5082, and 16231 were common only to the M6a subhaplogroup. Three additional transitions (3486, 9329, and 13966) characterized the subgroup M6a1, which included previously identified haplogroups M6a and M6b. One novel subgroup had been defined, designated M6b, by seven coding region mutations (593, 3745, 4380, 4418, 4592, 6746, and 13998) (**Figure S1**).

The phylogeny of haplogroup M34 was considerably refined with the aid of our newly collected sequences. This haplogroup could now be defined by a single coding region mutation (11101). Three previously reported individuals (C39, C56, and M42) (Sun *et al.* 2006; Thangaraj *et al.* 2006) were placed into a subcluster M34a1 as defined by four transitions at positions 569, 6794, 15865, and 16249. Another newly designated subhaplogroup M34b was defined by three coding region mutations (3483, 4020, and 13651) (**Figure S1**). Haplogroup M35 was defined by mutations 199 and 12561 (Sun *et al.* 2006). It was found that a considerable number of samples that possess coding region mutation 12561 but do not have a control region mutation 199. This suggested that 12561 was the only defining mutation for the haplogroup M35. Consequently, a novel lineage characterized by five coding and four control region mutations (2442A-5027-6866-7954G-13500, 16129-16284- 16327-16398) was defined and designated as M35c (**Figure S2**). It was also observed that haplogroup M43 had only four

characteristic mutations (709, 10316, 11696, and 12636), and its main subbranch, formerly identified as M43 by Thangaraj *et al.* 2008, was designated as M43a here (**Figure S5**). Haplogroup M52, which was defined recently by an array of five coding region mutations (1462T, 1598, 5460, 6020, and 10750; (Fornarino *et al.* 2009) was revised as M52a here. We have found a considerable number of samples that had coding region mutations 1462T, 1598, and 5460, but not 6020 and 10750 mutations. This suggested that the new M52 was defined by only three mutations (1462T, 1598, and 5460) (**Figure S2**).

Updating phylogeny of South Asia autochthonous R lineages

Complete mtDNA sequence analysis of samples from Indian populations indicated the existence of seven autochthonous lineages within macrohaplogroup R (Palanichamy *et al.* 2004). Recently, Chaubey *et al.* 2008 has updated the phylogeny of R with additional complete mtDNA sequences. Consequently, 10 subhaplogroups (R5a1, R5a2, R6a, R6a1, R7a, R7b, R30a, R30b, R31a and R31b) were identified within the existing haplogroup R (**Figures S1, S3, and S4**). Our extensive typing of coding region and control region information and complete sequencing for twelve mtDNAs have allowed us to update the definition of previously defined haplogroups and to identify several new haplogroups viz., R6b, R6c, R30a1, R30c, and R30d. Haplogroup R6 was subdivided into three subclades R6a, R6b, and R6c. They were defined by coding and control region variants 228-11075-14058, 146-459-2386A -3637-9449-10084-11620-14470-16172-16266, and 246-4991-16179 -16227-16245 -16278, respectively (Figure S3). The addition of a single R haplotype sequence to the existing R30a haplogroup tree resulted in a deep branching point at the root by sharing the mutations 3316, 5442, 6764, 9242, 11047A, and 15055. Taking into account these modifications, the previously defined R30a haplogroup was revised into R30a1, and characterized by four coding region transition (2056, 4232, 9156, and 12714) (Chaubey *et al.* 2008). Subgroup R30c was characterized by five coding (2331A, 4025, 5074, 8508, and 15924) and HVS I region motifs (16291-16311), while R30d was defined by a single control (16292) and four coding region (6290, 7280, 7843, and 13539) mutations (**Figure S1**).

In this study, we have identified another two novel R haplogroups (R32, and R33) from

South Asia. Haplogroup R32 was defined by three coding (8784, 8887, and 15712) and three control region mutations (16145, 16185, and 16239). Haplogroup R33 was characterized by four coding region (1531, 7981, 13557, and 13768A) and one control region mutation (16304) (**Figure S1**). It should be noted that the previously reported haplogroup R31 lineages (samples C110, Rj25, and Rj48) actually belonged to haplogroup R33 (Palanichamy *et al.* 2004; Chaubey *et al.* 2008).

We have similarly updated the Indian U2 haplogroups with the aid of additional complete sequences, leading to the identification of many subhaplogroups -U2a1, U2a2, U2b1, U2b2, U2c1a and U2c2. Subhaplogroups, U2a1 and U2a2, were characterized by control and coding region motifs, 8572-16154 and 11383, respectively (Figure S5). Based on control region mutations (234 and 16353), subhaplogroup U2b1 was characterized. Two coding region mutations (1888 and 9094) further defined this haplogroup. The second subcluster U2b2 was defined by coding region polymorphisms 3915 and 4093 (**Figure S1**). Subhaplogroup U2c1a was determined by three coding region mutations, 8023-8676-9767. We have also identified a novel deep rooting mtDNA U2c clade, U2c2, which was defined by six (195, 4062, 4080, 15244, 16129, and 16239) transitions. Based on reconstructed phylogeny, the formerly defined U2c was redesignated as U2c1 here (Palanichamy *et al.* 2004) to avoid confusion (**Figure S1**).

Major expansions of macrohaplogroup M and N-R lineages from Eastern region of India

Using the phylogenetic hierarchical approach, we identified ten macrohaplogroup M lineages (M5, M6, M18, M31, M33, M34, M37, M38, M51, M53, and M55) as having originated from the eastern part of India (refer Dataset S1) and which have contributed to approximately half of the observed South Asia M lineages (Figure 1a and S1). Among these, six haplogroups (M5, M6, M18, M33, M38, and M53) shared ~40% of M haplogroup variations.

The paragroup M5 included three subclades, M5a, M5b, and M5c. The main subclade, M5a, was further divided into two subgroups M5a1 and M5a2. Subgroup M5a1 was quite star-like, with a predominantly southern, eastern and northern distribution. In contrast, the sister haplogroup M5a2 was restricted to the eastern part of India (Jharkhand, West Bengal, and Bihar). Other M5 subclades, M5b and M5c, were found predominantly in the

eastern region of India. These data indicated that the eastern part of India might have been the center of haplogroup M5 origin. Haplogroup M6 is common (~7-9%) in southern and eastern regions of India, but is very rare in northern and western parts of India. From the updated phylogenetic branching patterns, M6 haplotypes were classified into two major clusters, M6a and M6b. The majority belonged to M6a, which was the most frequent and widespread in the Indian subcontinent. There were two M6a clusters - M6a1 and M6a2. Both were well represented in the eastern part of India, and appear to have an origin in eastern India. The subgroup of M6a1 (M6a1a and M6a1b) initiated major expansion in the southern part of India. Here, we would emphasize that if it were not for the coding region information, subgroup M6b could not have been distinguished. This was due to M6b sharing control region variations (16223-16362; 16223-16274-16362) with other Indian M lineages (M30, M35, and M36) as well as East Asian haplogroups D and G. It was concentrated in the eastern part of India (West Bengal, Jharkhand and the geographically adjacent area Adilapad, Andhra Pradesh (Mittal *et al.* 2008) Haplogroups M6a and M6b appeared to be largely confined to eastern and adjacent Andhra Pradesh, thereby suggesting an origin of M6 haplogroup in the eastern region of India.

An east India origin of haplogroup M18 appeared likely, given that subclades M18a and M18b were common in the eastern region of India (West Bengal, Jharkhand, and Bihar) while being virtually absent in northern and western parts of India. Indeed the presence of M18a lineages in the southern India (Andhra Pradesh and Tamil Nadu) corroborated the eastern stream of expansion. The paragroup M33 was subdivided into two main subclades: M33a and M33b. Subgroup M33a could not be distinguished from M* without analysis of coding region mutations, although some of its subclades (M33a1 and M33a2) had distinctive control region motifs. M33a1 was frequent in the south eastern region (Andhra Pradesh and Tamil Nadu) but only sporadically observed in the eastern region. Interestingly, our study revealed a single M33a1 lineage in Southeast Asia (Thailand) population. In contrast, the sister clade M33a2 was found at higher frequencies in northeastern India (Garo, Nongtrai, and Lyngngam), suggesting a possible founder effect in these populations (Reddy *et al.* 2007). Additionally, the higher variances of M33a as well as subclade M33b in west Bengal and Jharkhand, coupled with their absence in

northern/western regions, again suggested an eastern stream of expansion of haplogroup M33.

The definitions of haplogroup M38 revised by Fornarino *et al.* 2009, included a subclade M38a. It should be noted that M38a was defined by a nucleotide change at 199 and a deletion at 522-523 position. In our sample collection, some M38 haplotypes had a deletion at 522-523 position but not the control region mutation 199. In addition both the sites occurred independently in different M and N-R haplogroup backgrounds. This raised a question about M38a classification. In contrast, our earlier published two M38 sequences shared five mutations (189-1808-6267-6899-9966) and pointed towards a new definition of subclade M38a to replace the Fornarino *et al.* (Fornarino *et al.* 2009) definition. Subgroup M38a lineages have been found in northern, eastern, and southern regions, but the elevated frequency in the southern Indian tribal groups was most likely the result of drift (Cordaux *et al.*2003). However, a group of unclassified M38 lineages occurred at considerable frequency in west Bengal, Jharkhand, Nepal, and Bihar, with some gene flow into the adjacent area of Andhra Pradesh but was virtually absent in northern, western, and southern India. This suggested that the origin of haplogroup M38 could be placed in the eastern region of India. Haplogroup M53 was very common and spilled over ~8% of the M lineages in the eastern part of India (Orissa, west Bengal, Jharkhand, and Bihar), with little dispersal into southern India.

The distributions of another four haplogroups (M31, M34, M37, M51, and M55) completely overlapped with other eastern India origin M haplogroups. On the basis of branching patterns of haplogroup M31 tree, the subclade M31a1 occurred in Andaman Islanders, whereas its sister clade M31a2 and the deepest branch M31b were found in the eastern/ northeastern regions, which suggested that haplogroup M31 most likely had an eastern India origin (Palanichamy *et al.* 2004; Reddy *et al.* 2007; Fornarino *et al.* 2009). Furthermore, the eastern origin of haplogroup M34 was indicated by the distribution of its subclades M34a1 (M34a1a and M34a1b), M34a2, and M34b, in west Bengal, Orissa, Nepal, Bangladesh, Jharkhand and an adjacent area Uttar Pradesh, but virtual absence in southern, western and northern regions. However, M34b lineage was observed in southern India and probably derived from the eastern region. We also inferred an eastern origin of haplogroup M37, as subclades M37b and M37* haplotypes were concentrated

in eastern/ northeastern India. In addition, a sister clade M37a was also found in an adjacent area Andhra Pradesh but its phylogeographic usefulness was limited due to its apparent extreme rarity. Haplogroup M51 occurred at moderate frequency among the eastern populations of Nepal, Assam, West Bengal, Jharkhand, Orissa, and extending to north Uttar Pradesh, Delhi, Rajasthan, Pakistan, and south regions Andhra Pradesh, Tamil Nadu and Sri Lanka. M55 lineage was very rare, but it occurred in eastern India (west Bengal, Meghalaya, Bangladesh); a matching type was found in Andhra Pradesh, thereby suggesting gene flow from the eastern region.

Haplogroups U2b, U2c, R7, R8, R30, R32, and R33, accounted for nearly half (~45%) of the macrohaplogroup R variations observed in the south Asia populations. On the basis of phylogenetic analysis, these haplogroups reflected their ancestral states of lineage divergence and expansion taking place in the eastern part of India. The haplogroup distribution is presented in the Dataset S1 and is discussed in detail below.

Two subclades of haplogroup U2b (U2b1 and U2b2) can be defined on the basis of complete sequence information. Subclade U2b1 further subdivided into two sub-branches - U2b1a and U2b1b, and both had wide geographic distribution. In contrast, sister clade U2b2 and its sublineages - U2b2a and U2b2b, were restricted to the eastern part of India. This data suggested that both the U2b subgroups arose in the eastern part of India and a subset of U2b1 lineages dispersed into other regions of India. Haplogroup U2c is subdivided into two major subclades - U2c1 and U2c2. Subclade U2c1 can be further subdivided into two principal subclades, U2c1a and U2c1b. The main subclade, U2c1a, was quite star-like having two derived sub-branches, U2c1a1 and U2c1a2. All were predominantly distributed in eastern/ southeastern India, with only minor occurrences in the north and west regions. The second principal subclade, U2c1b, was confined to eastern regions - west Bengal, Jharkhand, Orissa, an adjacent area Andhra Pradesh, and Tamil Nadu. In addition, the deepest branch U2c2 lineages were found in the eastern part shared with southern population, and were absent in north and west region; this further supported the eastern stream of U2c haplogroup expansion.

Two distinct subclades of R7, R7a and R7b, were observed (Chaubey *et al.* 2008). The first subclade R7a was common in East Indian states (Bihar, Jharkhand, and west Bengal) while being almost absent in southern, western, and Northern provinces. In contrast, R7b

was principally concentrated in southern regions, in particular Andhra Pradesh (Kivisild *et al.* 2003 ; Thanseem *et al.* 2006), which is geographically adjacent to the eastern region, although there was extension to Tamil Nadu and Kerala. This suggested that the origin of R7 can be placed in eastern India and the subhaplogroup lineages dispersed fairly recently to the east and south. Moreover, representatives of R7b lineages were also observed in eastern (Madhya Pradesh, Bihar and Bangladesh), northern (Uttar Pradesh) and western (Maharashtra) regions, adding support to this conclusion (Cordaux *et al.* 2003; Roychoudhury et al. 2001; Roy *et al.* 2003; Kumar *et al.* 2006).

Haplogroup R8 was also very common among the eastern and southeastern populations, but was very rare in northern and western populations. It was subdivided into two subhaplogroups - R8a and R8b. R8a could be further subdivided into two principal subclades - R8a1 and R8a2. By examining both the control region sequences obtained in this study and those published previously, a higher variation of R8a haplotype was noted in southern and eastern regions. The sister group R8a2 and the deep rooted R8b lineages had been found sporadically in southern, eastern, and the adjacent area of Uttar Pradesh, suggesting that the eastern part of India was the likely geographical source for the spread of R8 haplogroup. This finding was consistent with the recent work linking the Austro-Asiatic (Munda tribe) speaker and expansion of R8 haplogroup in eastern part of India (Thangaraj *et al.* 2009).

The major branches of R30 haplogroup - R30a, R30b, R30c and R30d, were linked together by a single coding region mutation 8584. Our new understanding of the phylogenetic structure of haplogroup R30a indicated that it most likely originated in the eastern part of India. There were two major subclades of this haplogroup, R30a1 and R30a2. Subclade R30a1 was further subdivided into two principal subclades R30a1a and R30a1b. There has been considerable drift in R30a1a lineage, and its distribution was predominantly in the southern region of India. R30a1b lineages were found in northern, eastern and southern regions, and occasionally in the western part of India. However, the deepest branch R30a2 was confined to the eastern region, which suggested that haplogroup R30a had an eastern India origin. R30b was found in the northern and eastern regions, and rare/absent in southern and western India. Haplogroups R30c and R30d which were confined to eastern and southeastern India were extremely rare in northern

and western regions of India. Taking into account the R30 subhaplogroup divergence and their geographic distribution, eastern India is believed to be the most likely location for R30 haplogroup origin.

The newly defined R32 haplogroup was more frequent in our studied east India populations but was rarely, if at all, observed in other regional populations. In fact, the close match of eastern group R32 haplotypes found in the geographically adjacent southern area of Andhra Pradesh suggested that it probably derived from the eastern region. We have corrected the phylogenetic position of previously identified R31 haplogroup lineages (C110, Palanichamy *et al.* 2004), Rajput25, and Rajput48 (Chaubey *et al.* (2008) into a newly assigned subhaplogroup of R33 - R33a1. This subclade has been sporadically found across northern and eastern provinces, pointing to a wider distribution in the past. However, the deepest branching lineages - R33a2 and R33b - found in the eastern parts of India, was sufficient to suggest haplogroup R33 originating in east India. It was also found that among the unclassified M* and R* haplotypes, two-thirds of the lineages were found to be unique to the eastern region of India. These findings, therefore, supported the proposition that these lineages very likely originated from the eastern region of India.

An eastern India expansion route of M subhaplogroups

Our improved phylogenetic analysis yielded numerous M subhaplogroups. Among the defined M subhaplogroups, M3b, M4c, M35c, M39a3, M40a, and M52a, encompassed ~11% of the south Asian M lineages. Based on detailed molecular surveys, these subhaplogroup lineages showed a clear phylogeographical structure as being expanded from the eastern region of India.

Paraphyletic haplogroup M3 represented ~10% of the south Asian M haplogroup and had four subgroups - M3a, M3b, and M3c. Amongst them, M3a lineages are widespread in the subcontinent and no phylogeographic structure was detected in this subclade. The subhaplogroup M3b, in contrast, showed clear phylogeographic structuring in their distribution. M3b subgroup could be readily distinguished from other unclassified M3* groups, and were confined to the eastern region of India. Paragroup M4 included the three major subclades M4a, M4b, and M4c, and all were distributed widely in south Asia populations. There was no phylogeographic structure evident within the subclades M4a

and M4b; both subclade sequence variants were observed in different regional populations. The phylogeographic analysis of M4c, however, showed that there was good structuring in the distribution of subclade M4c1 and M4c2 lineages. M4c1 haplotypes were distributed in elevated frequencies in the southern regions of Tamil Nadu and Andhra Pradesh, while M4c2 was confined to the eastern regions of West Bengal, Jharkhand, and Meghalaya. Indeed, unique M4c1 haplotypes found in the eastern and northern populations combined with the restricted distribution of M4c2 in the eastern populations suggested an eastern region expansion of M4c.

Paraphyletic haplogroup M35 had three subclades, M35a, M35b, and M35c, among which haplogroups M35a and M35b were most commonly found in all populations. Furthermore, phylogeographic analysis revealed there was no apparent structure existing for these clades. On the other hand, subclade M35c and an unclassified M35* haplotype showed clear phylogeography structures – they were concentrated in eastern and geographically adjacent areas. There was substantial haplotypes variation within the haplogroup M39 subclades - M39a1, M39a2, and M39a3; this variation, however, did not show strong geographical structuring. Generally, all subclade lineages were found throughout eastern coastal regional populations. However, representatives of M39a3 lineage showed some phylogeographic structure; they were confined to the eastern region (West Bengal, Jharkhand, Bihar and an adjacent area of Andhra Pradesh), while being rare in other regions.

Haplogroup M40 represents 3% of the south Asian M haplogroup and was represented by subgroup M40a, which occurred at high frequency among the eastern populations of West Bengal, Jharkhand, Bihar, and Orissa, and occasionally extending to the adjacent southern regions of Andhra Pradesh and Tamil Nadu. Additionally, our complete control sequence analysis revealed that a higher diversity of M40a lineages was found in the eastern regional populations. This suggested that M40a lineages might have expanded from the eastern region. Haplogroup M52 was primarily localized to the eastern and an adjacent southeastern region; in particular the internal sub-branches of M52a - M52a1 and M52a2 were found in eastern Indian populations. M52a1 lineages were observed in southern/western regional populations and shared ancestry with eastern Indian populations, thereby suggesting that subclade M52a spread from eastern India.

Limited phylogeographic structure in the M and N-R lineages: a possible eastern India range of expansion

M2 was the most frequent comprising one tenth of the South Asian M haplogroup and it had two major subclades, M2a and M2b. Subclade M2a could be further subdivided into two principal subclades, M2a1 and M2a2, which had very distinctive sub-branches - M2a1a, M2a1b, M2a1c, M2a2a, and M2a2b. Among these sisters groups, M2a1a, M2a1b, and M2a2a constituted 93% of the south Asian M2 sequences. The number of internal branches in subhaplogroup M2a1a was significantly higher than in other M2 sub-branches (Kumar *et al.* 2008), and was widespread in south Asia. Subhaplogroup M2a1b was more frequent in southeastern and eastern regions, while being virtually absent in western and northern regions of India. Like M2a1b, subhaplogroup M2a2a and its sublineages were found more frequently in eastern and southeastern regions, whereas sister haplogroup M2a2b was restricted to the southern states (Andhra Pradesh, Kerala, Tamil Nadu, and the neighboring area of Sri Lanka): this suggested that southern India might be a possible source population for M2a2. The present study showed, for the first time, a relatively small number of deep rooted M2b haplotypes in the southern and northern Indian caste populations. However, its low haplotype diversity in the southern Indian population indicated strong genetic drift and founder event influences. Overall, major clades M2a1, M2a2, and their subclades had their combined present range predominantly in the eastern and southern regions implying that these could have been the regions where the M2 clade started diversification (Kumar *et al.* 2008). Haplogroups of M41, M54, and M56, were sporadically spread at low frequency in southern and eastern states, and were almost absent in northern and western region.

R5 is the most frequent subclade of R in India and was subdivided into two subhaplogroups - R5a and R5b. In comparison with R5b, R5a was well represented across the broad range of populations and was further subdivided into two principal subhaplogroups - R5a1 and R5a2. Subhaplogroup R5a2 constituted the majority of R5 in the southern and eastern parts of its distribution range. However, the high frequency with a low haplotype diversity of R5, i.e., R5a1 and R5a2b, in the north Indian Bhargava population was most likely due to genetic drift and founder events. R5a2 was further subdivided into three subclades, R5a2a, R5a2b, and R5a2c. Among these subclades,

R5a2a and R5a2b contributed towards two-thirds of R5 sequence variation and they showed wide geographic distributions. There was no phylogeographic structure evident within the R5a2b. On the other hand, shallow phylogeographic population structure was observed in the sister clade R5a2a - R5a2a1; it was mainly found in the eastern regions of West Bengal, Jharkhand, Bihar, and adjacent areas of Uttar Pradesh and Andhra Pradesh. Furthermore, representatives of R5a2c as well as the higher variation of clade R5a1 lineages in the southern/eastern regions, in concert with the observed limited occurrence of R5b in the southeast region, suggested that R5 probably differentiated from the southeastern regions.

Haplogroup R6 is the second most common haplogroup in the macrohaplogroup R clade, and it was subdivided into three major well defined subclades - R6a, R6b and R6c. Among these subclades, R6a was well represented in eastern and southeastern regions, but was absent or rarely observed in western and northern regions suggesting that a possible demographical expansion occurred in the eastern region. The occurrence of sublineage R6a at its highest frequency in eastern India tribal populations was likely due to genetic drift and founder events. In addition, we also observed small subclades within R6a that exhibit shallow phylogeographic structures in southern populations. Subhaplogroup R6b had a much more limited geographical distribution, occurring sporadically in southern, eastern and northern populations. Subhaplogroup R6c sequences had been found exclusively in eastern, southeastern and neighboring areas including Bhutan and Nepal, suggesting a possible geographic diffusion of R6c along the eastern region of India. Taking into account the overall phylogeographic structure displayed by haplogroups M2, M41, M54, M56, R5, and R6 and their subhaplogroups, a progressively eastern and southeastern colonization by these haplogroups could be inferred.

On the basis of the phylogeographic structure of haplogroups M36, R31, and N5, it was concluded that the ancestor of these haplogroups might have emerged from the adjacent areas of the eastern region, i.e., Andhra Pradesh, and subsequently radiated southwards and occasionally to the northern region. It was possible to distinguish several small subclades in M30 (M30a-M30f), M43 (M43a and M43b), and U2a (U2a1 and U2a2), but they were distributed in all major geographic areas and did not represent the distinct phylogeographic structure.

The improved structure of aboriginal Southeast/Island Southeast Asian mtDNA clades

Haplogroups M21, M22, M45, M46, M47, N21, N22, R21, R22, R23 and unclassified M*/N* haplotypes comprised over 20% of the observed mtDNA variation in aboriginal Southeast and Island Southeast Asia populations. Amongst these, haplogroups M45, M46, M47, R22 and R23 had been characterized only by control region mutations (Hill *et al.* 2007).

The addition of new complete mtDNA sequences allowed the revision of the phylogenetic tree of previously defined subhaplogroups M21-M21a, M21b, and M21c (**Figure S6**). In the reconstructed phylogeny, the M21b branch showed close affinity with haplogroup M13 through sharing three coding region mutations (6023, 6253, and 15924). Furthermore, it shared six coding (709, 9182, 10172, 11665A, 15468, and 15930) and two control region (16263 and 16381) mutations with our new sequence that specifically defined the subhaplogroup M21b. We could not detect the coding region mutation 11482 in either our sample or in counterpart M13 haplogroup sequences, which had defined haplogroup M21. We, therefore, replaced the M21b lineage from the haplogroup M21 and formed a new clade M13'M21b. On the basis of the shared mutations in our data and the data of Macaulay *et al.* 2005, the subhaplogroups of M21-M21a and M21c could now be defined. Subhaplogroup M21a was characterized by five coding (709, 3796, 11287, 11611, and 15924) and three control region (152, 16129, and 16256) mutations, whereas subhaplogroup M21c was characterized by only three coding region mutations (3915, 5108, and 7861). It is noteworthy that the subhaplogroup M21d of Hill *et al.* 2007 is actually a subbranch of M21c, as demonstrated via complete sequence analysis (Dancause *et al.*2009).

To date, only one M22 complete sequence has been reported from the Southeast Asia Temuan population. The sequence data has allowed us to identify three coding region mutations (710, 4639, and 6776) to characterize this haplogroup M22 (**Figure S6**). We also detected two phylogenetic sub-branches, M22a and M22b, which were characterized by control and coding region motifs 6071-9316-16290 -16304 and 7444-13681, respectively.

Haplogroups M45 and M46 have been defined on the basis of control region mutations

16209 and 16278, respectively, and their phylogenetic positions have recently been confirmed with complete sequences from Cambodian samples (Hartmann *et al.* 2009). We have added additional complete sequences to each haplogroup, thus allowing us to detect the characteristic mutations for M45 and M46. Haplogroup M45 and M46 were defined by control and coding region variants (143, 930C, 1598, 12973, 13651, 16209, 16233, and 16304) and (4697, 4973, 9509, 14687, 15317, and 16278) respectively (**Figure S6**). Furthermore, we have detected several representatives of the previously defined M45a (Hill *et al.* 2007) lineage, which harbored characteristic HVS-I mutations 16129-16209-16223-16272. In order to confirm the phylogenetic position, we sequenced mitochondrial genomes from populations of northeast India and Malays in Singapore. The phylogenies of studied individuals fell into a separate clade-M20 (**Figure S6**).

To date, there has been only one report of a completely sequenced R21 individual of Southeast Asian origin (Macaulay *et al.* 2005). In the present study, we added another related complete sequence from Malays in Singapore, which allowed us to define the diagnostic mutations for haplogroup R21. The phylogenetic status of a previously identified haplogroup R22 based on control region data now receives clear support from our complete coding region information: the control (152, 329, 16249, 16288, 16304, and 16311) and coding region (1040, 6770, 10316, 10609, and 13359) mutations. We have also added a single N21 and N22 complete sequence to the existing Southeast Asia lineage and identified key characteristic mutations for haplogroups N21 and N22 (337del, 6752, 8701, 10583, 13437,14560, 16193; 942, 7158, 9254, 11365, 16168, 16249) (**Figure S6**).

The identification of novel lineages from Southeast/Island Southeast Asia

A rare haplotype 16223-16324-143-234 was observed in our Malays sample, and complete sequence analysis revealed that it had a M33 haplogroup coding region mutation 2361 while also sharing mutations 3221 and 16324 with subhaplogroup M33b (**Figure S1**). This result refined the previously defined M33b, which we have re-designated as M33b1. The new lineage has been tentatively assigned as M33b2. Based on the control region motif 16278, the four Island Southeast Asia haplotypes (16092-16209-16223 -16224-16263 -16278 -16319, 16093-16209-16223-16224-16263-16278- 16319, 16093-16209-16223-16224

-16263-16265 -16278-16319, and 16093-16209-16223-16224-16263-16274 -16278 -16319-16356) were grouped into haplogroup M46 (Hill *et al.* 2007). However, phylogenetic results from our complete sequences showed that the Island Southeast Asia haplotype shared coding (7226 and15663) and control region (16263 and 16319) mutations with a recently assigned South Asia haplogroup M50 (Reddy *et al.* 2007). Within haplogroup M50, a new subclade named M50a, with characteristic seven coding region (1383, 3417, 6092, 8802, 12281, 12486, and 15055) and two control region (16166C and 16309) mutations had been identified on the basis of shared mutations between our data and the data of Reddy et al. 2007. The sister Malay lineage has been named as M50b (**Figure S6**). Undefined N haplotypes which shared HVS-I motifs 16263-16274-16311 -16343 from northeast India, Thailand and Southeast Asia Malays samples were completely sequenced. The phylogenetic analysis showed that they all shared the coding region mutations (2760, 3027, 7885, 7961, 8188, 13710 and 15211); we have, therefore, proposed a new haplogroup, which we have designated as N23 (**Figure S6**).

A reassessment of unclassified Southeast/Island Southeast Asia lineages

We have checked the mutations in the control/coding region of several unassigned lineages relative to previously reported Southeast/Island Southeast Asia in our current data set and their evolutionary relationships were inferred using phylogenetic analysis. Five haplotypes having linked HVS-I polymorphic sites 16223-16234-16290 observed in the Island Southeast Asia haplogroup M/R data were found to actually belong to the East Asian haplogroup M12 (Hill *et al.* 2007). We found a rare haplotypes bearing the mutations 16129-16140 -16223-16271, 16129-16136-16140-16223-16271, and 16182C-16183C-16189-16223- 16271-16311 from the Malays, northern Thailand and Bangladesh populations (**Figure S6**). We applied this phylogenetic information to the unclassified Island Southeast Asia M haplogroup sequences, and we found that seven sequences belonged to haplogroup M60. In addition, anther rare haplotype observed in the Malays had the control region mutation 16167-16223-16246- 16311-16362-195. Another unclassified N lineage from the southern parts of East Asia has been recognized by control region (HVS-I linked motif 16111-16172-16189-16223- 16311-16362) and partial coding region RFLP (Kivisild *et al.* 2002). Subsequent complete mtDNA sequence

analysis revealed that it shared coding region mutations with previously defined East Asia haplogroup N10 (**Figure S6**). Similar subhaplogroup sequences were observed in the unclassified Island Southeast Asia N lineage (Hill *et al.* 2007) and could be assigned into haplogroup N10 without ambiguity.

The aboriginal Southeast/Island Southeast Asians shared a common ancestry with mainland Indo-China regional populations

The haplotypes observed in the northeast India, southern East Asia, Thailand, and Malays in Singapore were compared, with the indigenous haplogroups previously reported in the Southeast/Island Southeast Asia populations and their evolutionary relationships are discussed here.

In our survey three lineages from northeast India and one lineage from the Malays in Singapore harbored haplogroup M21b mutations. Recently, additional two M21b lineages were reported from Nepal-Tharu population (Fornarino *et al.* 2009). Furthermore, these lineages shared similar control region motifs and had been reported in northern Thailand Yong population, distant southern aboriginal Malays (Semelai and Temuan), Sumatra (Madan), and Sulawesi (Manado) populations (Hill *et al.* 2006; Hill *et al.* 2007; Kampuansai *et al.* 2007). Another distinct M21b lineage was found in northern Semang (Jahai and Mendrik), Senoi (Temiar), and far distance Borneo (Banjarmasin) populations (Hill *et al.* 2006; Hill *et al.* 2007). Conversely M21b sister clade M13 lineages have been found in northeast India, Tibetan/ southern East Asia, and adjacent northern Thailand regions, but were virtually absent in Southeast/Island Southeast Asia (Oota 2001; Yao *et al.* 2002a; Wen *et al.* (2004a).

We noted that coding region information is obligatory for identifying the haplogroup M21c. On the basis of control plus complete coding region information, four major groups of sequence variants could be discerned for haplogroup M21c. Group I and II related sequences had shared control region motifs 16093-16223-16249-16301-16311 and 16145-16192-16223- 16291-16304, respectively. Group III and IV have 16223-16242-16319 and 16042-16070-16183C-16189-16209-16223-16233 HVS-I motifs, respectively. The phylogeographic analysis revealed that Group I sequences were restricted to East India populations while Group II sequences have been observed in northern Thailand, Korea and extremely distant Island Southeast Asia Bali populations

(Hill *et al.* 2007; Dancause *et al.*2009 ; Yao *et al.* 2002a ; Fucharoen 2001; Lee *et al.* 2006). Group III and IV lineages were restricted to Malay populations (Hill *et al.* 2006; Wong *et al.* 2007).

An updated phylogenetic analysis of haplogroup M22 found the subhaplogroup M22a lineage to be detected in Thailand and Malay populations (Hill *et al.* 2006; Fucharoen *et al.* 2001; Wong *et al.* 2007). As with the new phylogenetic structure, haplogroup M45 lineages were sporadically distributed in east India, eastern Thailand (Khmer), Cambodia, Malays, Sumatra, Alor, and Ambon populations (Hill *et al.* 2007; Hartmann *et al.* 2009 ; Wong *et al.* 2007; Black *et al.* 2006; Lertrit *et al.* 2008). Haplogroup M46 was most divergent and frequent in aboriginal Southeast Asia (Thailand, Cambodia, and Malaysia) and Island Southeast Asia populations (Hill *et al.* 2007; Hartmann *et al.* 2009; Fucharoen *et al.* 2001; Wong *et al.* 2007; Lertrit *et al.* 2008). Nevertheless, while M46 was restricted to Southeast/Island Southeast Asia, it shared a common ancestor with southern East Asia specific haplogroup M62 (Kampuansai *et al.* 2007; Yao *et al.* 2002a). In addition, we found another diverse haplogroup M20 lineage in east/northeast India (West Bengal, Bangladesh and Meghalaya), Southern East Asia (Yunnan and Quangxi), Southeast Asia (Thailand and Malaysia), and Island Southeast Asia (Sumatra, Java, Borneo, Bali, Sumba, and Sulawesi) populations (Hill *et al.* 2007; Reddy *et al.* 2007; Oota *et al.* 2001; Yao *et al.* 2002a; Wong *et al.* 2007; Lertrit *et al.* 2008; Li *et al.* 2007). Furthermore, these clades (M20'46'62) shared a basal coding region mutation 14110, which reinforced the common ancestry origin of these lineages.

The distribution of haplogroups R22 and N21 was also interesting: haplogroup R22 has been observed in the northeast India (Assam, Nagaland, Tripura, and Andaman-Nicobar island) (Cordaux et *al.* 2003; Prasad *et al.* (2001); Thangaraj *et al.* 2003), Southeast Asia (Thailand, Cambodia, and Malaysia) (Yao *et al.* 2002a; Wong *et al.* 2007; Black *et al.* 2006), and Island Southeast Asia (Sumatra, Java, Bali, Borneo, Sumba, Sulawesi, and Lombok) (Hill *et al.* 2007), but was not detected in the southern East Asia populations. Haplogroup N21 was detected in the southern East Asia (Yunnan) (Wen *et al.* 2004b) southern and east India (Tamil Nadu and Bangladesh), Thailand, Cambodia, Malaysia, Sumatra, Sulawesi, Bali, and Alor populations (Hill *et al.* 2006; Hill *et al.* 2007; Kampuansai et al. 2007; Oota et *al.* 2001; Yao *et al.* 2002a; Fucharoen *et al.* 2001; Wong

et al. 2007; Black *et al.* 2006; Lertrit *et al.* 2008). Haplogroup N22 distribution in the Southeast/Island Southeast Asia region was nonuniform. It was found only in the Malays, Temuan, and Island Southeast Asia Waingapu populations (Hill *et al.* 2006; Hill *et al.* 2007). However, similar HVS-I control region variants of haplogroup N22 lineage was observed in the east India (Orissa-Gope) (Sahoo *et al.*2006) and northern Thailand (Yuan) populations (Kampuansai *et al.* 2007).

In addition, our phylogenetic analysis (following the nomenclature of QPK *et al.* submitted a paper) showed that most of the unclassified mtDNA variants found in the Southeast/Island Southeast Asia shared their ancestry with the mainland Indo-China and adjacent regional populations, e.g., haplogroups M50b, M60, N10a, and N23. Haplogroup M60 was observed in east India (West Bengal and Bangladesh), southern East Asia (Yunnan, Quizhou, Quangxi, Hunan, and Hainan) (Yao *et al.* 2002a; Wen *et al.* 2004a; Li *et al.* 2007; Wen *et al.* 2004b), northern Thailand (Chiang Mai, Chiang Rai, and Mae Hong Son) (Kampuansai *et al.* 2007; Sahoo *et al.* 2006), Malaysia, Sumatra, Sumba and Philippines populations (Hill *et al.* 2007; Wong *et al.* (2007). Haplogroup M50 was first identified in northeast India populations (Reddy *et al.*2007). Variants of this haplogroup, i.e., subgroup M50a, was observed in Khasi populations and a related haplotype was found in aboriginal Sakai, Thailand (Fucharoen *et al.* 2001); yet another subgroup, M50b, was found in Malays, Sumatra, and Sulawesi populations (Hill *et al.* 2006; Hill *et al.* 2007; Wong *et al.* 2007). Subhaplogroup of N10, N10a, occurred in the populations of Southern/western East Asia (Yunnan, Guangxi, Guangdong, Guizhou, Tibet, and Uzbek) (Kivisild *et al.* 2002; Yao *et al.* 2002a; Li *et al.* 2007; Wen *et al.* 2004b; Li *et al.* 2007), northern Thailand, Malaysia, and Borneo (Hill *et al.* 2007; Fucharoen *et al.* 2001; Wong *et al.* 2007), but its sister clade N10b was restricted to southern East Asia (Hainan, Guangdong, and Fujian) (Kivisild et al. 2002; Li *et al.* 2007) and its nearest neighbor region Jiangsu/Shanghai (Wen *et al.* 2004b). In addition, northern Thailand Yuan ethnic group had a single haplotype of this haplogroup (Kivisild *et al.*2002). Haplogroup N23 has been observed in east India (West Bengal), southern East Asia (Guangxi) (Yao *et al.* 2002a), northern/eastern Thailand, Malaysia, Sumatra and Java populations (Hill *et al.* 2007; Wong *et al.* 2007; Lertrit *et al.* 2008; Besaggio *et al.* 2007). Additionally, we detected a Malay lineage that shared a common root with northeast

Indian subhaplogroup M33b.

East Asian haplogroups in northeast India

Phylogenetic analysis using control and coding region information clustered the observed northeast Indian haplotypes into major East Asian M haplogroups (C, D, G, M7, M9, M12, M13 and Z) and N-R haplogroups (A, B, F, R9, and R11).

On the basis of phylogenetic with phylogeographic analysis of the haplogroup M9 subclades, Soares *et al.* (Soares *et al.* 2008) suggested that Southeast Asia was likely to be a geographical source and that the spread of subhaplogroup M9a into East Asia mainland occurred after the last glacial maximum. In the present study, we completely sequenced three individuals of M9 from populations of northeast India. The phylogenetic analysis revealed the presence of subhaplogroup M9, which was phylogenetically most related to Southeast Asian pre-M9a. In view of the new sequence information as well as chronological criteria, we have replaced the Southeast Asian pre-M9a and pre-pre-M9a with new designations, i.e., M9a1b and M9a2, respectively. In our survey, M9a1b was present at high frequency with substantial diversity in northeast India with occasional spreading to southern East Asia (Yao *et al.* 2002a; Wen *et al.* 2004a; Li *et al.* 2007; Wen *et al.* 2004b) and Thailand (Kampuansai *et al.* 2007; Oota *et al.* 2001). This pointed towards the northeast India area being the most probable site of origin of the M9a1b. Despite using only the HVS-I motif, we could identify three M9a2 haplotypes: one each from Yunnan, Zhejiang, and Liaoning of mainland East Asia, and the northern Thailand populations (Oota *et al.* 2001); Wen *et al.* 2004a; Wen *et al.* 2004b). In addition, a new M9 lineage - M9a3 - was found in Southern East Asia (Yunnan), eastern region of China (Henan), and northern Thailand (Khuen) populations (Kampuansai *et al.* 2007; Li *et al.* 2007; Li *et al.* 2007). Finally, the deepest branch of M9 - M9b - found in southern China (Li *et al.* 2007; Wen *et al.* 2004c; Kong *et al.* 2006) and other unclassified M9* lineages in northeast India (Cordaux *et al.* 2003, matched with aboriginal Malays, suggesting that M9 as a whole was most likely to have had an origin in the Indo-China corridor.

We have partially sequenced rare East Asian M12 haplogroup lineages from northeast India and Bangladesh. The sequences from northeast India together with two published M12 sequences from southern China (Guangdong) clustered separately from the Bangladesh lineage. This cluster was designated as M12a (Kong *et al.* 2003a). Similar

matches to the Bangladesh lineage were found in southern China (Yunnan) (Wen *et al.* 2004a; Li *et al.* 2007; Wen *et al.* 2004c) and northern Thailand (Oota *et al.* 2001; Yao *et al.* 2002a; Fucharoen *et al.*2001; Wong *et al.* 2007; Besaggio *et al.* 2007), which we designated as M12b (Dataset S10). Although occurring in only low frequencies, the presence of haplogroup M12 in the Southeast/Island Southeast Asia pointed towards a genetic continuum between mainland southern East Asia and Southeast Asia (Hill *et al.* 2007).

Haplogroup M13 was present at low frequencies in different East Asian and Northern Asian populations (Wen *et al.* 2004a; Tanaka *et al.* 2004; Derenko *et al.* 2007). It should be noted that most of the Northern Asian, Japanese, and Koreans, belonged to the M13a, a subcluster of M13. In this study, we observed an additional M13 lineage in the northeast India area people harboring HVS1 motif (16145-16168-16188-16223-16311, and 16145-16168-16188-16223- 16257 -16311). To assess their phylogenetic position, we have compared our complete sequence data with published M13 northern Asian and Japanese data (Kong *et al.* 2006; Tanaka *et al.* 2004). In phylogenetic analysis, this lineage did not share any coding region mutations with the reported M13a subclade; hence, we designated this new lineage as M13b. This variant of mtDNA haplotypes was found in Tibetans/Yunnan (Yao *et al.* 2002a; Wen *et al.* 2004a), Thailanders (Oota *et al.* 2001) and northeast Indians, but was extremely rare in other East Asian, i.e., this lineage was overwhelmingly restricted to the Indo-China corridor region.

Our partial coding region sequence analysis demonstrated that the northeast Indian lineages were shared with members of the novel East Asian haplogroups M61, M64, and M65. Information on the designated novel East Asian haplogroups (M61, M64, and M65) was retrieved from QPK *et al.* submitted paper. We observed mtDNA types of haplogroups M61 and M65 in Uttar Pradesh, West Bengal, Tripura, Orissa, and in south Indian states Andhra Pradesh, Tamil Nadu, and Karnataka. The previously proposed M48 and M49 lineages from northeast India were actually a subclade of M64 (Reddy *et al.* 2007), which was present at high frequencies in the east and northeast India populations. Analyses of the overall distribution of mtDNA haplotypes showed that northeast Indian populations were closer to the southern East Asians.

The evidence from the present phylogenetic with phylogeographic study suggested that

aboriginal Southeast/Island Southeast Asians likely differentiated in the area around the Indo-China corridor, and migrated to the Malaya Peninsula and subsequently dispersed into Island Southeast Asia. Furthermore, we found high South Asian M and N-R haplogroups diversity in the samples from the eastern regions of West Bengal, Jharkhand, Bihar, Orissa, and the adjacent area Andhra Pradesh; this suggested that South Asian lineages might be diverged from the eastern India precursor population(s).

Coalescence Times

Table 14 and Table 13 show lineage diversity (ρand σ) and age estimates for the South, Southeast and Island Southeast Asian lineages. Coalescence ages of mtDNA haplogroups were calculated by using the coding region mutation rate of one substitution per 5,140 years (Mishmar *et al.* 2003), with an estimate of standard error following Saillard et al. (Saillard *et al.*2000). This method has been applied to many South, Southeast and Island Southeast relic populations mtDNA sequences (Macaulay *et al.* 2005; Hill *et al.* 2006; Thangaraj *et al.* 2009; Thangaraj *et al.* 2005b).

Discussion:

Mitochondrial DNA surveys in worldwide populations have shown a continent-specific distribution of mtDNA lineages. African populations are characterized by the oldest superhaplogroups, L0, L1, L2, L3, L4, L5, and L6, (Behar *et al.* 2008) but all detected non-African lineages belonged to two founder clusters, named M and N, which share a common root with African haplogroup L3. It appeared that only African L3 radiated out of Africa, mainly in the form of haplogroups M and N, giving rise to the extant non-African variation. The geographical distribution and similar founder ages of M and N in South, East, and Southeast Asia suggested that this wide geographic colonization could have happened in a single African migration within a relatively short time period (Macaulay *et al.* 2005; Sun *et al.* 2006).

The single southern coastal route rapid settlement hypothesis assumes that the ancestor to non-African haplogroup L3 carried to Arabia ~80,000 years ago, was followed by a period of diversification and drift leading to the appearance of haplogroups M and N and the loss of the ancestor L3 type (Richards *et al.* 2006) . This suggested that non-African founding haplogroups M and N originated somewhere between East Africa and the Persian Gulf (Torroni *et al.* 2006). Then both M and N root type carried rapidly along the

southern coastal route and the subsequent evolution of Asia region-specific haplogroups took place more gradually from the coastal settlement sites (Richards *et al.* 2006; Metspalu *et al.* 2006).

Unfortunately, there is no evidence for the diversification of M and N in the Arabia Peninsula; it lacked ancient mtDNA lineages (Abu-Amero *et al.* 2008). From the present study, phylogenetically hierarchical with phylogeographic evidence showed the spread of South Asian founding M and N, and N derived R lineages, to have originated from the eastern part of India. In addition, Southeast/Island Southeast Asian indigenous mtDNA haplogroups shared their ancestry with our Indo-China corridor samples. Those lineages were scattered across east/north east India and southern East Asia region including southern China and parts of northern Thailand. In addition to our analysis, the mtDNA data presented in literature showed a southern origin (i.e., South China) of East Asians. We didn't observe any major ancient lineage originating from the northern/northwestern, western and southern regions, other than the eastern region of India. Overall, it appears from our study that the north east Indo-China region must be a center for Asian founder differentiation.

The coalescence time of South Asia and shared Southeast/ Island Southeast Asia clades varied substantially, ranging from 9 to 69.4 thousand years. Ignoring the younger clades, the South, Southeast/Island Southeast Asians appear to have an ancestry in the Pleistocene. It seems likely that the precursor of the Asian founders (mtDNA- L3 lineages) left Africa in the last major Pleistocene wet phase somewhere between 82-78 kya to Asia (Pope *et al.*2008; Oppenheimer, 2009). This was followed by a tight bottleneck in the precursor group(s) living in Asia caused by the Toba volcano explosion (74 kya) which effectively buried the southern peninsular India with volcanic ash (Acharya SK and Basu PK ,1993), in association with the deteriorating climate in Oxygen isotope stage 4 (OIS 4, 71-59 kya) (Pope , 2008). These events are believed to have resulted in both the branching off of the founders M and N, and the extinction of ancestor haplogroup L3 in Asia. The high rainfall tropical forest area in the northeast Indo-China region is likely the only refugium where the surviving population(s) underwent a bottleneck which might have given rise to the founders of Asia (Ambrose, 1998). The distribution of two ancient Y-chromosome haplogroups C and D in Asia also showed evidence of a transition

occurring in the northeast Indo-China region (Underhill and Kivisild, 2007); Shi *et al.* 2008), lending further support to this hypothesis.

The presence of two mtDNA haplogroups, the M1 and U6 originated from southwestern Asia (Olivieri *et al.* 2006), and the coalescence time ~40 to 45 kya suggests that one founding group dispersed east to Asia and another group remained in southwestern Asia or west of the Indian subcontinent, but after ~5 to 10 kya, its descendant populations dramatically expanded to western Eurasia (Goebel ,2007). However, our phylogeography studies and the coalescence time (~45-40 kya) indicated that Indian U lineages were differentiated along with M and N-R from the eastern part of India (**Table 14**). In addition, the southwestern Asia M1 clade (Olivieri *et al.* 2006) shared a common single coding region mutation (14110) with our newly defined Eastern Asia clade M20'46'62, suggesting a common ancestor. The differentiation of U, M and N founders from the east India and Indo-China region suggested that the southwestern Asia upper Paleolithic M1 and U lineages (Olivieri *et al.* 2006) may have resulted from the peripheral isolation of descendants of the founders from Eastern Asia rather than ancient break-up of the precursor founder group in Southwest Asia (Macaulay *et al.* 2005; Metspalu *et al.*2006).

(Note: See Appendix for Table 9,10,11,12)

Part-II:

Archeological and genetic studies have revealed that India occupies an important geographical location and has served as a major corridor for the dispersal of modern humans (Cavalli-Sforza *et al,* 1994; Majumder, 1998; Cann, 2001). From the available prehistoric, ethno-historic, linguistic and biological evidences, it appears that the Indian subcontinent was probably first peopled approximately 60,000 years BP (Misra, 1992; Kivisild *et al,* 2003). Though the entire Indian subcontinent shows diverse ethnic composition, Eastern and North eastern part of India covering the sub-Himalayan terrain has been a hotspot for linguistic (Ruhlen, 1991) and population diversity and thought to have witnessed and retained the signature many ancient and contemporary genetic event of population migration and settlement. The north-eastern and sub-Himalayan region of India is predominantly occupied by speakers of the Tibeto-Burman languages (of the Sino-Tibetan family) along with Austro-Asiatic(AA) and Indo European(IE) speakers along with sporadic presence of Dravidian group. There

are two contrasting and competing models based on the mtDNA and Y-chromosome studies in the NE populations regarding the possible role of NE India in the population migration as well as in settlement. One group based on their mitochondrial DNA (mtDNA) and Y-chromosome studies inferred that Northeast India might have acted as a barrier rather than the facilitator of the movement of populations both into and out of India (Courdax *et al,* 2004) where as the same view has summarily been rejected in a recent study by Reddy *et al*, 2007 who confirms that Northeast India has acted as a corridor for initial movement of populations.

It is now well defined that people belonging to the Indo-European, Dravidian, and Tibeto-Burman groups represent both caste and tribes, but the Austro-Asiatic speakers are exclusively tribal. Though the tribal populations of India possess higher genetic diversity than caste populations and are considered as the possible original inhabitants of India (Thapar 1966; Basu *et al*, 2003, The Indian Genome variation Consortium, 2008), but their evolutionary histories with respect to settlement, biological origin and relationships with caste populations have been debated comprehensivel and remained as an enigma till to date. The phylogenetic relationship between tribe and castes have been shown quite explicitly in the Southern part of India (Kivisild *et al,* 2003, Cordaux *et al*, 2004) however, the tribe-caste relationship with respect to uniparental

markers have not yet been studied in Eastern and North Eastern India which is the site where TB and AA speakers are abundant and thought to have made an important contribution in the sub-Himalayan gene pool. Therefore, it is crucial to carry out such genetic investigations in culturally and linguistically disparate and ethnically well-defined populations inhabiting the Eastern and North Eastern periphery of sub-Himalayan region.

The present investigation has been carried out in ten sub-Himalayan populations from North East India belonging to TB, AA, DR and IE group to further understand i) the role of North East Indian passage way in population migration, ii) to delineate to relationships between caste and tribes in sub-Himalayan and iii) to understand the genomic diversity amongst various groups in North east India and their relation with other Indian as well as Asian populations by employing high resolution SNP markers of NRY region of Y-chromosome. Total no. of 383 males who comprise the caste and

tribal populations of Sub-Himalayan range of East India were considered for this present study. **Table-1** lists the sample size, geography, and linguistic affiliation of the populations examined. All ethical guidelines were followed, as stipulated by the institutions involved in the study. A total number of 78 binary markers belonging to all the major haplogroups known to be present in Indian as well as Asian populations were considered for genotyping (**Table-2, see Appendix**). The nomenclature followed for the Y-SNP haplogroup is as recommended by the Karafet *et al,* 2008.

Table 1: Geographical distribution and the Linguistic Affiliations along with sample size of the Eight Studied Populations

Name of the Population	Sample Size	Social Category	Linguistic group	Sampling Area
Dhimal	**36**	Tribe	Tibeto-Burman	Darjeeling District
Rabha	**26**	Tribe	Tibeto-Burman	Jalpaiguri District
Mech	**19**	Tribe	Tibeto-Burman	Darjeeling District
Lachungpa	**11**	Tribe	Tibeto-Burman	Sikkim
Kol	**60**	Tribe	Austro-Asiatic	Darjeeling District
Santhal	**51**	Tribe	Austro-Asiatic	Jalpaiguri District
Kharia	**35**	Tribe	Austro-Asiatic	Jalpaiguri District
Oraon	**34**	Tribe	Dravidian	Jalpaiguri District
Rajbanshi	**57**	Caste	Indo-European	Darjeeling District
Bengali	**54**	Caste	Indo-European	Darjeeling District

Results:

Altogether, 34 polymorphic loci in these ten groups define 22 paternal haplogroups(**Fig Y-1-see Appendix**)

Distribution and frequency of Y-chromosome haplogroups in Austro-Asiatic Group:

H haplogroup was the most prevalent haplogroup in the three different AA tribes having a

combined frequency 43.33% of H1 in Kol, 43.13% in Santhal and 22.85% in Kharia. This was followed by the frequency of O2-M95 whose frequency was 38.33% in Kol, 37.25% in Santhal and 68.51 % in Kharia. Apart from these two haplogroups, J2, O-M175 and R1 were present but with low frequency.

Distribution and frequency of Y-chromosome haplogroups Tibeto-Burman Group:

In the TB populations except Lachungpa the differentiated O3 haplogroup i.e. M117 showed the highest frequency i.e. 41.66% in Dhimal, 57.69% in Rabha and 47.36% in Mech. However, the frequency of undifferentiated O3 i.e. O3-M122 was highest, 40% in Lachungpa. Dhimal had 27.7% of J2-M410 while in other three populations it was completely absent. On the other hand, Lachungpa and Mech showed 36.36% and 5.26% of D-M174 respectively while none of the individuals from Dhimal and Rabha had this marker. The frequency of H1 was 8.33% in Dhimal, 11.53% in Rabha, 10% in Lachungpa while it was completely absent in Mech. Apart from these only Dhimal had 8.33% of R1, Lachungpa had 10% of R-M207 and 10% of Q1-M378 was present in Lachungpa.

Distribution and frequency of Y-chromosome haplogroups Dravidian Group:

The only Dravidian tribe i.e. Oraon showed highest frequency of combined H1, 70.58%. Apart from this, the frequency of R2 was 20.58%. Apart from these two haplogroups, there was no other haplogroups in this population.

Distribution and frequency of Y-chromosome haplogroups Indo-European Group:

In the IE populations, many haplogroups were detected with low to moderately high frequency. Bengali showed 38.88% of R1 while Rajbanshi had 38.59% of differentiated O3 i.e. O3-M117. Apart from these, other HGs like H, J2, L, C, O2, R2 were also present but wit low frequency

Table-3: Some of the major HG Frequencies in Linguistic Subgroups of Sub-Himalayan Populations

HG	Number (%) of Haplogroups									
	Austro-Asiatic			Tibeto-Burman				Dravidian	Indo-European	
	Kol N=60	Santhal N=51	Kharia N=35	Dhimal N=36	Rabha N=26	Mech N=19	Lachungpa N=11	Oraon N=34	Bengali N=54	Rajbanshi N=57
C-M216									1(1.85)	
D-M174						1(5.26)	4(36.36)			
H1-M370	19 (31.66)	13(25.49)	6(17.14)		2(7.69)			16(47.05)		1(1.75)
H1-M82	7 (11.66)	9(17.64)	2(5.71)	3(8.33)	1(3.84)		1(10.00)	8(23.52)	6(11.11)	1 (1.75)
J2-M172									1(1.85)	
J2-M410	1(1.66)			10(27.77)					5(9.25)	3(5.26)
J2-M12	1(1.66)									
J2-M241									4(7.40)	2(3.50)
L-M11										1(1.75)
L1-M76									1(1.85)	1(1.75)
O-M175	3(5.00)		1(2.85)	2(5.55)	1(3.84)				2(3.70)	1(1.75)
O2-M95	23(38.33)	19(37.25)	24(68.51)	1(2.77)		4(21.05)			2(3.70)	1(1.75)
O3-M122					2(7.69)	3(15.78)	4(40.00)			
O3-M134					4(15.38)				4(7.40)	
O3-M117				15(41.66)	15(57.69)	9(47.36)				22(38.59)
R-M207		2 (3.92)	1(2.85)				1(10.00)		1(1.85)	1(1.75)
R1-M17	2(3.33)	1(1.96)		3(8.33)				1(2.94)	21(38.88)	9(15.78)
R2-M124		6(11.76)						7(20.58)	1(1.85)	2(3.50)
Q1-M378							1(10.00)			

Discussion:

Genetic Diversity:

Overall, Mech exhibited high homogeneity with only four haplogroups, whereas both the Indo-European Castes showed highest heterogeneity with 12 haplogroups, while rest of the populations showed moderate heterogeneity with 5-7 haplogroups. Considering our results, it can be said that in sub-Himalayan region, caste populations belonging to the Indo-European linguistic group shows more diversity at least in Y-haplogroups than the tribal populations. Amongst the tribal groups, both the AA and TB exhibited more diversity than the DR group.

We detected two new SNPs one in the H-haplogroup and have named it as M505 whereas other one is in the O Haplogroup which is a Tdel in P31 of O2 haplogroup. The identification of M505 and typing of all markers present in the Haplogroup H has helped to revise and update H-haplogroup tree constructed by Karafet *et al,* 2008. Regarding the nomenclature and positioning of Tdel mutation in P31, we need to carry out additional works on more samples from various linguistic groups from different regions of India as well as Asia.

Revised Resolution of Haplogroup-H

We detected a new mutation, M505, A-G transition at np 146 while typing M89 in 30 samples; 9 Kols, 2 Kharia, 5 Oraon, 8 Santhals, 3 Dhimals, 1 Rabha, 1 Lachungpa, 6 Bengali and 2 Rajbanshis. M505 was detected in 35 H-M82 samples while remaining 2 were present in M370 samples. However, one of our samples had H2-Apt chromosomes. After typing this M505 in all the M69, M52, M370 and M82 chromosomes, we consider M505 to be phylogenetically equivalent to M52. Our present findings on the H-haplogroups and its nomenclature also differ from the tree constructed by Karafet *et al*, 2008. Therefore, we revise and update the H-haplogroup tree as shown in **Fig.Y-2**.

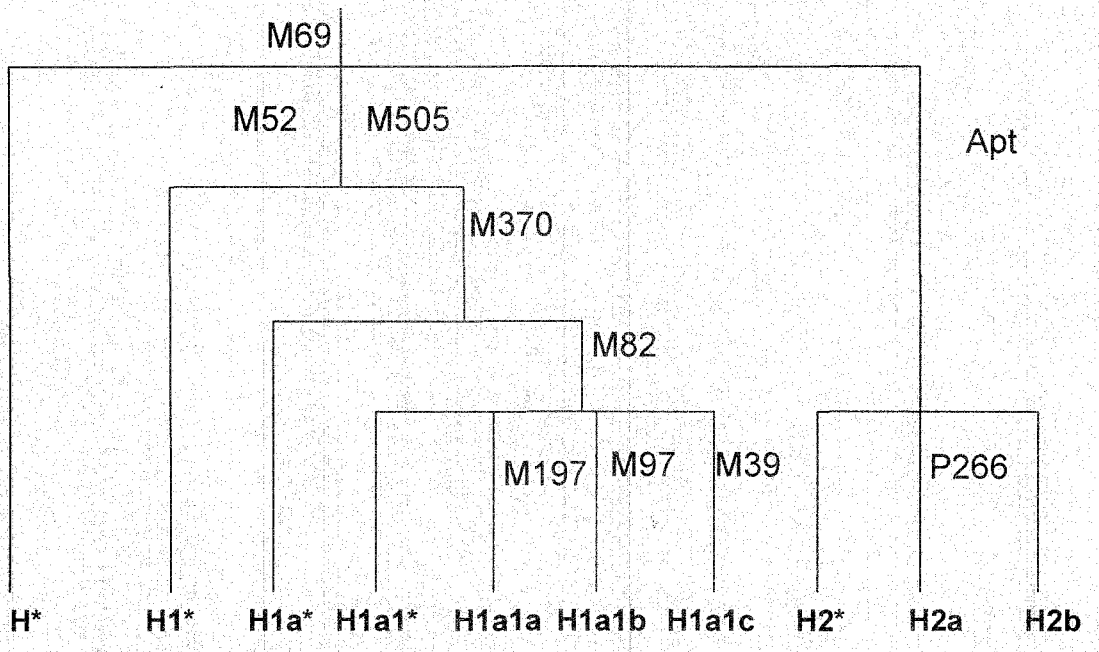

Fig-Y-2: Revised phylogenetic relationships and nomenclature of Y-chromosome Haplogroup-H

Previous studies have shown that haplogroup H*-M69 and its subclades H1-M52 and H2-APT has an indigenous Indian origin (Kivisild *et al*, 2003, Sengupta *et al,* 2006, Sahoo *et al*, 2006), however, the distribution and frequency of H-haplogroup in various populations across different provinces of India has not been investigated systematically. Trivedi *et al,* 2008 have shown that distribution of Haplogroup H showed a north-south gradient (24.4% to 27.4%), however geographically its total frequency was highest (44.4%) in populations of western India). However, our present findings differ from Trivedi *et al,* 2008 as the frequency of H1 in Oraon population in sub-Himalayan region is much more i.e. 70.58% than the frequency reported in West Indian populations. The present study also support the view of Kivisild *et al*, 2003 that in comparison with caste groups, tribal populations showed significantly higher frequencies of haplogroup H1. Therefore, it is quite difficult to say at this moment regarding the place of origin of H haplogroup in Indian subcontinent however, it can be taken for granted that H haplogroup

might have originated in tribal populations as it has uniformly high distribution in the tribal populations representing North, West, South, Central, East India with the exception of TB tribes in North east India.

Role of North eastern region of India in Population migration and settlement:

Austro-Asiatic group is considered to be the oldest linguistic family in all the linguistic families and has three sub-families: Mundari, Mon-Khmer and Khasi-Khmuic (Diffloth, 2005).Though it has been supported by many genetic studies that Austro-Asiatic groups had its origin in India but there are many competing theories on their origin and subsequent settlement in India. Our present analysis on three Austro-Asiatic Mundari speaking populations has shown that it has combined frequency of O2-M95 i.e.48.3% with complete absence of O3-M122. We also typed M88, the downstream SNP marker of M95 but none of the samples had this marker. The frequency of M95 in other Mundari populations representing different parts of India (East, West and Central) shows a combined frequency of 54% (Sengupta *et al,* 2006; Kumar *et al,* 2007) and in Khasi-Khmuic it is 30.3% (Reddy *et al,* 2007). Therefore, it can be deduced that O2-M95 might have originated in Mundari speaking Austro-Asiatic inhabiting either Central or East India which shows a gradient of distribution from East (Mundari) i.e. 53.13% (Sengupta *et al,* 2006 to North East (Khasi-Khmuic) i.e. 30.3% (Reddy *et al,* 2007) India and by using North East as a corridor it might have moved out of India to South East Asia. Though all the sub families of AA speaking populations might have common paternal origin as suggested by Kumar *et al,* 2007 but they might have undergone various waves of genetic changes in India as well as in South East Asia as evidence by the presence of O3-M122 and its sub clades with relatively higher frequencies in Khasi-Khmuic (Kumar *et al,* 2007) and other South East Asian Austro-Asiatics (Shi *et al,* 2005).

Therefore on the basis of Y-chromosome diversity studies it can be concluded that North East has acted as a corridor at least for the back migration of Austro-Asiatics people from out of India and possibly to the South East Asia.

Caste-tribe relationship in the Sub-Himalayan region:

The origin of caste system in India is a matter of debate. Previous genetic studies on Indian castes and tribes failed to achieve a consensus on their origins and affinities. A few studies reported closer affinity of Indian castes with either the Europeans or the

Asians. Studies of Bamshad *et al*, 2001 and Basu *et al*, 2003 support the genetic differentiation of caste and tribal populations, and the North Indian invasion of Indo-European speaking nomads, pushing the Dravidian tribes to southern peninsula. On the other hand, Kivisild *et al*, 2003 suggest that Indian tribal and caste populations derived largely from the same genetic heritage of Pleistocene southern and western Asians, receiving limited gene flow from external regions since Holocene. Further, Cordaux *et al*, 2004 reports that the paternal lineages of Indian castes are more closely related to the Central Asians than to the Indian tribal groups, thereby supporting the view that Indian caste groups are primarily the descendents of the Indo-European migrants. More studies are required for a better understanding of the genetic structure of the diverse Indian populations, where many questions remain unanswered.

The frequency of haplogroup distribution in eight tribal groups and two caste groups has been presented in **Table-3**. Our results show that almost in all the tribal groups the most frequent haplogroups was H1, O2 and O3 whereas the haplogroup distribution was quite different in two different castes. In Bengali, R1 was the most frequent haplogroup whereas in Rajbanshi differentiated O3 i.e. O3- M117 had the highest frequency of 38.59%. The frequency of R1 was much lower i.e. 15.78% than Bengali and other castes reported in other studies (Thanseem *et al*, 2006). The high frequency of O3 in Rajbanshi caste raises vital question about the tribe-caste relationships as HG-O3 has very low frequency amongst the caste populations across India (Sengupta *et al*, 2006; Sahoo *et al*, 2006). Moreover, this haplogroup has a very high frequency amongst the Tibeto-Burman populations found in North East India as well as in South East Asia. Historical clippings have revealed that Rajbanshis have a disputed history and many suggest that Rajbanshis had tribal origin. From the resent understanding and on the basis of haplogroup sharing, it can be said that Rajbanshis might have a Tibeto-Burman connection regarding its origin and therefore tender more supports to the long standing opinion that Indian lower castes had tribal origin (Kivsild *et al*, 2003, Thanseem *et al*, 2006).