Deposited research article

Phylogeny of the M superhaplogroup inferred from complete mitochondrial genome sequence of Indian specific lineages

Revathi Rajkumar, Jheelam Banerjee, Hima Bindu Gunturi, R Trivedi and VK Kashyap

Address: National DNA Analysis Centre, Central Forensic Science Laboratory, 30 Gorachand Road, Kolkata- 70014, India.

Correspondence: VK Kashyap. E-mail: cflslkolkata@indiatimes.com

Posted: 23 December 2004

Genome ***Biology*** 2004, **6:**P3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/6/2/P3

Received: 14 December 2004

This is the first version of this article to be made available publicly.
A modified version is now available in full in *BMC Evolutionary Biology* at http://www.biomedcentral.com/1471-2148/5/26/abstract

Phylogeny of the M superhaplogroup inferred from complete mitochondrial genome sequence of Indian specific lineages.

Revathi Rajkumar, Jheelam Banerjee, Hima Bindu Gunturi, R Trivedi and VK Kashyap*

National DNA Analysis Centre, Central Forensic Science Laboratory, 30 Gorachand Road, Kolkata- 70014, India.

Abbreviated title: A phylogenetic analysis of M haplogroup.

***Address of Corresponding Author:**

National DNA Analysis Centre,
Central Forensic Science laboratory,
30 Gorachand Road, Kolkata 700014, INDIA
E-Mail address: cflslkolkata@indiatimes.com
Tel.: +91-33-2284-1638

Abstract

Background:

Phylogenetic analysis of human complete mitochondrial DNA sequences has largely contributed to resolving phylogenies and antiquity of different lineages belonging to the majorhaplogroups L, N and M (East-Asian lineages). In the absence of whole mtDNA sequence information of M lineages reported in India that exhibits highest diversity within the sub-continent, the present study was undertaken to provide a detailed analysis of this haplogroup to precisely characterize the lineages and unravel their intricate phylogeny.

Results:

The phylogenetic tree constructed from sequencing information of twenty four whole mtDNA genome revealed novel substitutions in the previously defined M2a and M6 lineages. The most striking feature of this phylogenetic tree is the formulation of a new lineage M30, distinguished by the presence of 12007 transition, and comprises of the recently defined M18 and a potential new sub-lineage possessing substitution at 16223 and 16300. M30 further branches into M30a sub-lineage, defined by 15431 and 195A substitution. The age of M30 lineage was estimated at 33,042 YBP, indicating a more recent expansion time than M2 (49,686 YBP). Contradictory to earlier reports, the M5 lineage does not always include a 12477 substitution, and is more appropriately defined by a transversion at 10986A. The phylogenetic tree also identifies a potential new lineage M* with HVSI sequence 16223,16325. No new substitutions were found in M25 and the M3 mt DNA genome could only be tentatively rooted by 16126 mutation. M4 and M*(16251, 16267) lineages could not be resolved distinctly.

Conclusion:

This study describes seven new basal mutations and fourteen lineages that substantially contribute to the present understanding of superhaplogroup M. The phylogenetic tree supported by median-joining network helps in distinctly identifying the genetic relation between different M lineages that could not be achieved solely by control region sequence information. Although high control region diversity has been reported in the different M lineages distributed in India, complete sequencing of M* and defined lineages suggests that these mt DNA genomes emerged from a limited number of branches arising from the M trunk.

Background

A surge of mitochondrial DNA control region sequence information has been generated along with information from coding region substitutions in diverse populations of the world to understand their genetic diversity, structuring and origins [1-11]. More recently, mt DNA sequence data has been used to lay emphasis on peopling of Asia and to comprehend various population demographic parameters [12-20]. From a genetic perspective India assumes importance in Asia because the (1) the contemporary ethnic populations residing here are both biologically and culturally well differentiated, (2) a clear distinction exists between non-tribal and tribal populations, autochthonous to the sub-continent [15, 21], lastly and more importantly this expanse is believed to be one of the initial regions of settlement of modern humans [20]. Of the four major matrilines identified till date L, M, N and R, about 60% of Indians trace their maternal roots in Indian specific branches of haplogroup M that is reported to have emerged from the African haplogroup L3. With the exception of the M1 lineage that is confined to Ethiopia [22], all the other branches of this superhaplogroup including M*, C, D, G, E and Z haplogroups are observed in Asia [12, 14, 23, 24]. The lineages M2, M3, M4, M5, M6, M18 and M25 are exclusive to India, with M2 reported to be the oldest lineage in the sub-continent with an age estimation of 60,000yrs-75,000 yrs. Furthermore, the frequencies of these clades among the different geographic, linguistic phyla and social strata have been investigated in detail, yet the fundamental question regarding origin of this super-haplogroup remains unanswered [15, 20]. While some authors have suggested a southwest Asian origin of M superhaplogroup, followed by a back migration to Africa [15], others support its African ancestry [25]. One major drawback in arriving to a

conclusion is the limitation of control region sequences, which provide useful information for forensic purposes but does not provide reliable estimate of phylogeny owing to homoplasy and recurrent mutations [23, 26].

Complete mitochondrial genome sequencing has gained importance in resolving phylogenies and understanding human evolution where control region motifs have failed. Extensive genome sequencing studies have been carried out in different lineages of L, N and M major- haplogroups across different global populations. Though the phylogenies of East Asian counter parts of M lineages: M7, M8a, M8C, M8Z, M9, E, D, G have been resolved in detail, but till date no similar studies have been attempted on the sub-lineages of the Indian M haplogroup [9, 27-33]. The complete mt DNA sequence information from Indian M lineages will not only help answer questions regarding the origin of this haplogroup, clarify the phylogeny to finer branches but would also be highly relevant in forensic work, studies pertaining to mitochondrial disorders and disease diagnosis [28 and references therein].

The present study was undertaken to construct an unambiguous phylogeny for the M superhaplogroup and infer precise ages for its sub-lineages. Mitochondrial genomes were initially classified on the basis of their HVSI and coding region motifs, followed by complete sequencing of twenty three samples representing different M matrilineals. A median-joining network was also constructed from the data generated to decipher the genetic relationships amongst these lineages.

Results

We have found seven group defining basal substitutions and described fourteen lineages in detail from complete mt DNA genome sequence information, which will help

in further resolving some of the Indian M lineages. The M trunk differs from revised Cambridge reference sequence (rCRS) by substitutions at A73G, A263G, A750G, A2706G, A1438G, A4769G, C7028T, A8701G, A8860G, T9540C, A10398G, C10400T, T10873C, G11719A, C12705T, C14766T, T14783C, G15043A, G15301A, A15326G and C16223T. The coding region mutation sites analyzed in the present study were different from those observed in the sister M1 lineage found in Ethiopia. The M phylogenetic tree constructed on whole mt sequence information of twenty four samples belonging to different M lineages and their sub-types including M1, M2, M2a, M2b, M30, M30a, M18, M*, M3, M4, M5, M6a, and M25 is summarized in Fig 1.

M2 lineage: The complete sequencing of five mt DNA genomes belonging to M2 and its sub-lineages, M2a and M2b, indicated that coding region mutations T477C, T1780C, A8502G were associated with HVSI motifs C16223T and G16319A, which formed the root of M2 lineages. The M2b sub-lineage containing the HVSI motif G16274A and T16357C, in addition to the M2 defining mutations sites did not share any coding region substitutions with M2a. In case of M2a, we report a novel basal substitution at site T9758C in addition to previously reported transitions at G5252A and A8396G. Screening for T9758C site in 27 Indian individuals possessing the M2a specific HVSI and coding region motifs, clearly established this as a marker of this sub-lineage. Furthermore, we propose that a sub-cluster, M2a1, be diversified from M2a to differentiate individuals that possess both the C16270T and G16274A control region substitutions (Fig 2).

M30 lineage: A new lineage M30 was differentiated from M superhaplogroup, comprising of seven mt genomes, six of whose HVSI motifs did not correspond to any of the earlier established M lineages and one that was identified as M18 lineage, represented

as shaded region in Fig 1. Since lineages of M have already been catalogued from M1 to M25, this potential new lineage is designated as M30 to avoid any ambiguity in classification of the M superhaplogroup. This branch arises from the main M trunk with transition at site G12007A. Finer resolution of this lineage was achieved by further clustering four complete sequences with mutation at two sites, T195A and G15431A into a sub-lineage designated as M30a. Three mt DNA genomes further branched out from the M30 lineage, possessing only the substitution at G12007A. Interestingly, the newly but not well-described M18 (C16223, A16318T) matriline is one of branches that directly emerges from the M30 lineage. Sequence analysis of ten M18 mt DNA genomes showed the presence of G12007A transition. Eighteen Indian individuals were identified from our mtDNA database as possessing a HVSI motif (C16223T, A16300G) similar to the "Sao" sample, which arose from M30 lineage. All the eighteen individuals tested positive for the 12007 transition, suggesting that it might be prudent to group this sequence type into a distinct sub-lineage within M30 (Fig 2).

M5 lineage: The basal motif T12477C, G16129A and C16223T describes the Indian M5 lineage of majorhaplogroup M. Whole genome sequencing of three samples with similar HVSI motif of G16129A and C16223T revealed that only one sample (I. B306) exhibited the T12477C mutation, and was designated as M5a in Fig 1. This site was nevertheless absent in the other two samples, one of which had a similar HVSI motif as the M5a mt DNA genome, while the other exhibited an additional site G16048A in its control region motif. Our study identifies a transversion at C10986A, shared by all the three samples, suggesting that these HVSI types branched out from a common root. Analysis of the C10986A substitution in 7 Indian samples possessing HVSI motif, G16048A, G16129A

and C16223T, confirms our finding that different branches emerged from the M5 lineage. It is, however, important to note that two similar HVSI motifs might not necessarily be belonging to the same M5 sub-type.

M6a lineage: The two M6a matrilines completely sequenced, harbors the characteristic group defining mutations at site T16231C, T16362C and C3539T. Our analysis identified another novel substitution at site A5301G in this lineage. This lineage could not be further resolved owing to the absence of similar sites found in the other analyzed lineages.

M25 lineage: The M25 lineage has been recently described by the presence of G15928A and T16304C. It differs from the M halogroup by only five coding region substitutions and arises directly from the M trunk with no additional group defining motifs.

M3 lineage: The M3 lineage having the HVSI motif T16126C, C16223T was one of the branches whose position in the phylogeny could not be well established. Whole sequencing of two such genomes demonstrated a total lack of sharing in substitution sites between the two members of this lineage or with any other M sub-lineage analyzed in the present study. The M3 lineage was, hence, erected on the HVSI substitution T16126C.

Two branches arose directly from the trunk of M. One of the matrilineal type possessing C16223T and T16325C as HVSI motif has been observed in relatively high frequency in Indian populations. Contrary to our expectation, full sequencing of this mtDNA (Ho69), did not exhibit the presence of G12007A mutation site that was observed in other unidentified M lineages analyzed in this study. A similar result was observed after complete sequencing of the HVSI motif type C16223T and C16251T. Both these lineages were designated as M*. The controversial M4 lineage, with

diagnostic markers C16223T and the fast mutating site T16311C shared a substitution at A5319G with one of the M6a sub-lineage. Nevertheless, it was placed directly under the trunk of M owing to the absence of M6a diagnostic markers.

Discussion

Analysis of short stretches of mt DNA HVSI and HVSII region have significantly aided in distinctly distinguishing some of the M lineages. With the aim of understanding migration routes of the diverse Indian people, more control region sequences are being generated without much support from coding region sites, resulting in an increasing number of conflicts within the classification of its lineages. We report here a phylogenetic tree constructed from whole genome sequencing of twenty three Indian and one Ethiopian M lineage to resolve some of the anomalies occurring due to recurrent mutations in the control region.

The control region sequences have exhibited the presence of an array of M lineages in India [12, 16, 20], despite which, complete mt DNA sequencing suggests that most of these lineages arose as limited offshoots from the main M trunk. The M2 genome has been widely characterized in the Indian populations, yet complete sequencing of M2, M2a and M2b demonstrated the presence of a novel site T9758C, which is characterized as a diagnostic marker for M2a sub-lineage, in addition to G5252A and A8396G which were previous reported [15]. Since the frequency of M2b in the Indian populations is found to be very low (authors unpublished data), therefore, only one genome of this sub-type was sequenced. The study was, however, unable to trace any specific marker for this sub-lineage. We, however, propose that a cluster, M2a1, be formed within the M2a sub-lineage to include samples that contain two substitutions G16274A and C16270T in their

HVSI, instead of transition only at C16270T. Although HVSI sequences are not very reliable for constructing phylogenies, this cluster can well differentiate individuals with only one or both the mutations and in turn resolve the phylogeny to its finer sub-lineages. Age of M2 lineages using only coding region motifs estimated 49,686+/- 10,903 years before present, Fig 2, opposed to the expansion date of 60,000-75,000 yrs calculated from control region sequence information [15]. Although the 12007 substitution has been previously identified in other haplogroups, besides the M lineages [29], this study presents a novel lineage M30 that was constructed to include mitochondrial genomes possessing the G12007A substitution. The assembling of the M30a sub-lineage with its root at T195A and G15431A will help in further classifying M* samples that have not been identified till date owing to the absence of any charecteristic HVSI motif. An important contribution of this study is the placement of M18 lineage in the M phylogeny. In the absence of a coding region marker for this lineage [20], the G12007A substitution provides a stable root to the M18 type, which is defined only on basis of the HVSI motif A16318T. Mitochondrial genomes possessing the 16223, 16300 motif appears to be a promising new sub-lineage arising from M30. Additional complete mtDNA sequencing of similar sub-types will help in precisely defining this branch. The M30 lineage was relatively younger than the M2 lineage and has an expansion age of 33,042+/- 7,840 YBP, calculated on the basis of its coding region sequence information.

The M phylogenetic tree has largely aided in clarifying the position of the M5 lineage. Until recently, transition at G16129A along with basal motif for M, was used to characterize this lineage [34] and currently it is described by the presence of coding region mutation at T12477C [20]. The phylogenetic tree constructed in this study

provides evidence to support our finding that at least two sub-lineages arose from M5 that share a transversion at site C10986A and may or may not possess the T12477C transition. The presence of T12477C transition in only one of the two M5 mt DNA genomes sharing an identical HVSI motif, C16223T and G16129A, further substantiates the importance of coding region markers in precisely identifying mitochondrial phylogenies. Even though the G16048A, HVSI motif has not been included under M5 lineage owing to absence of T12477C, this study places it under M5. However, prior to defining the G16048A, G16129A and C16223T cluster, it is imperative that more samples representing this HVSI motif be completely sequenced. The age of the M5 lineage is estimated at 34,095+/- 6,425 YBP, indicating that M5 and its sister lineage M30 probably branched out from the M haplogroup around the same time.

Alike the M2 lineage that had been described in detail by previous studies [15], the present analysis identifies the presence of one novel mutation at position A5301G in the M6 lineage, with no further diversification of this lineage. In the absence of a coding region marker, the M3 lineage had to be tentatively erected on an HVSI substitution site, 16126, which makes this branch less stable than the others. Another interesting finding of the present study was the almost similar expansion dates calculated for M3 and M4 at 10,280+/- 3,801 and 15,420+/- 6,295 YBP, respectively. These two lineages are perhaps the youngest branches to emerge from superhaplogroup M. The newly defined M25 lineage did not share any common mutation sites with any other lineage and independently arose from M trunk with well established G15928A and T16304C substitutions. The moderately high frequency of the C16223T, T16325C HVSI motif types in the Indian samples suggest that there might be a potential new lineage that might

be more accurately described once additional genomes possessing this motif are fully sequenced. In the absence of this information, no attempt was made to classify this sequence type in to a lineage and hence, it was designated as M*. The other M* lineage bearing the control region motif, C16223T, C16251T, C16267T could not also be resolved further for similar reasons. Complete sequencing of more such undefined genomes would definitely help to trace the roots of these lineages.

Conclusions:

Our study presents the first phylogenetic tree constructed for the M lineages that are widely distributed in India. The emerging mt DNA phylogenetic tree constructed for the M lineages comprises of ten branches that identifies one new lineage and describes four new sub-lineages with seven group- defining mutations. This information can serve as a foundation for precisely describing the superhaplogroup M, as more complete mt DNA sequence information is generated. One of the most significant contributions of this study is the construction of a novel M30 lineage, which not only encompasses some undesignated HVSI motifs but also includes M18 as a sub-lineage within it. The M2 lineage was the oldest while the M3 and M4 were found to be the youngest branches that emerged out from the main trunk of haplogroup M. Although the HVSI and II sequences can be used to identify clusters with similar motifs, this study reinforces the importance of screening lineage defining coding region substitutions before defining a new lineage. The present analysis of twenty-four mt DNA genomes undermines the control region diversity reported in the M superhaplogroup by suggesting that most of the M lineages might in fact be derived from limited basal branches.

Methods:

Amplification and sequencing of HVSI and coding region motifs:

Genomic DNA was extracted from whole blood by standard Phenol/chloroform method. Amplification and sequencing of control region were performed in 1258 samples (authors unpublished data) as described in our earlier studies [35, 36]. Samples were initially analyzed for substitutions at site 10397 and 10400 via RFLP protocol before characterizing them under M haplogroup and subsequently clustered into different lineages of M via sequencing of the required coding region fragments- T447C, T1780C and A8502G for M2, G5252A for M2a, T12477C for M5 and C3539T for M6 [15, 20].

Complete mtDNA sequencing:

Twenty three complete mt DNA genomes belonging to M* (n-8), M2 (n-1), M2a (n-3), M2b (n-1), M3 (n-2), M4 (n-1), M5 (n-3), M6 (n-2), M18 (n-1) and M25 (n-1) sub-lineages of Indian superhaplogroup M were sequenced in the current study. DNA amplification and sequencing was carried out using primers described elsewhere [37]. New group defining substitutions were re-sequenced, and their frequency was determined in other Indian samples with similar control region motifs. Every sample was completely sequenced twice to remove any ambiguity. Additionally, since several segments of the same mt DNA had to be screened, care was taken to avoid artificial recombination caused by potential crossovers.

Phylogenetic analysis:

The sequence information generated by whole mt genome sequencing of twenty Indian specific M lineages was used to construct a phylogenetic tree of superhaplogroup M. Substitutions were reported with respect to the revised Cambridge Reference

Sequence (rCRS) [38]. Only sequence changes that occurred in at least two members of the same branch were defined as lineage specific. The complete sequence of the Ethiopian M1 genome [25] was included in the M Phylogenetic tree to determine its relationship with its sister lineages. Median-joining network algorithm was constructed based on sequence information of different M lineages [39].

Time estimate:

Age estimates for the M lineages were computed using only coding region substitutions identified from the complete mt DNA genome sequencing. The mean number of mutations per site to the most recent common ancestor was estimated and converted to real time using a substitution rate of $1.26X10^{-8}$ per site per year [40].

Authors' contributions

RR carried out most of the sequencing experiments, did phylogenetic analysis and drafted the manuscript. JB and HB also carried out extensive sequencing and RFLP experiments and analyzed the genetic data. RT provided critical and valuable information during processing of data. VKK is responsible for conceiving and designing of the study and contributed significantly in interpretation of data and shaping of the manuscript. All authors read and approved the final manuscript.

Acknowledgement

We are deeply indebted to Richa Ashma, Sonali Gaikwad, Sanghamitra Sahoo and Anamika Singh for allowing us to use the sequence information of complete mt DNA genomes processed by them. The present study would not have been possible without their contribution. A special thanks is extended to T Sitalaxhmi for reconfirming mutation sites. We express our appreciation to the blood donors analyzed in the present work. This work was supported by a research grant under the IX Five Year Plan to CFSL, Kolkata. RR and JB were assisted with DFS, and HB with CSIR (New Delhi), research fellowships.

References:

1. Kolman C, Sambuughin N, Bermingham: Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. Genetics 1996, 142: 1321-1334.
2. Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonne-Tamir B, Sykes B, Torroni A: The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 1999, 64: 232-249.
3. Tommaseo-Ponzetta M, Attimonelli M, De Robertis M, Tanzariello F, Saccone C: Mitochondrial DNA variability of West Guinea populations. Am J Phy Anthropol 2002, 117: 49-67.
4. Salas A, Richaqrds M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo A: The making of the African mtDNA landscape. Am J Hum Genet 2002, 71: 1082-1111.
5. Torroni A, Bandelt H-J, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, Villems R, Kivisild T, Metspalu E, Parik J, Tolk H-V, Tambets K, Forster P, Karger B, Francalacci P, Rudan P, Janicijevic B, Richards O, Savontaus M-L, Huoponen K, Laitinen V, Koivumaki S, Sykes B, Hickey E, Novelletto A, Moral P, Sellitto D, Coppa A, Al-Zaheri N, Santachiara-Benerecetti AS, Semino O, Scozzari R: A signal, from human mtDNA of postglacial recolonization in Europe. Am J Hum Genet 2001, 69: 844-852 .

6. Lalueza-Fox C, Sampierto ML, Gilbert MTP, Castri L, Facchini F, Pettener D, Bertranpetit: Unravelling migrations in the steppe: mitochondrial DNA sequences from ancient Central Asians. Proc R Soc Lond B 2004, 271: 941-947.
7. Wen B, Xie X, Gao S, Li H, Shi H, Song X, Qian T, Xiao C, Jin J, Su B, Lu D, Chakraborty R, Jin L. Analyses of genetic structure of Tibeto-Burman populations reveals admixture in southern Tibeto-Burmans: Am J Hum Genet 2004, 74: 856-865.
8. Behar DM, Hammer MF, Garrigan D, Villems R, Bonne-Tamir B, Richards M, Gurwitz D, Rosengarten D, Kaplan M, Pergola SD, Quintana-Murci L, Skorecki K: MtDNA evidence for a genetic bottleneck in the early history of the Ashkenazi Jewish population. Eur J Hum Genet 2004, 12:355-364.
9. Tanaka M, Cabrera VM, Gonzalez AM, Larruga JM, Takeyasu T, Fuku N, Guo L-J, Hirose R, Fujita Y, Kurata M, Shinoda K-I, Umetsu K, Yamada Y, oshida Y, Sato Y, Hattori N, Mizuno Y, Arai Y, Hirose N, Ohta S, Ogawa O, Tanaka Y, Kawamori R, Shamoto-Nagai M, Maruyama W, Shimokata H, Suzuki R, Shimodaira H: Mitochondrial genome variation in Eastern Asia and the peopling of Japan. Genome Res 2004, 14: 1832-1850.
10. Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, Scozzari R, Rengo C, Al-Zaheri N, Semino O, Santachiara- Benerecetti AS, Coppa A, Ayub Q, Mohyuddin a, Tyler-Smith C, Mehendi AQ, Torroni A, McElreavey K: Where west meets east: the complex mtDNA landscape of the southwest and central Asian corridor. Am J Hum Genet 2004, 74: 827-845.

11. Tambets K, Rootsi S, Kivisild T, Help H, Serk P, Loogväli E-L, Tolk H-V, 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 J-P, Barać L, Peričić M, Rudan P, Terzić R, Mikerezi I, Krumina A, Baumanis V, Koziel A, Richards O, De Stefano GF, Anagnou N, Pappa KI, Michalodimitrakis E, Ferák V, Freüdi S, Komel R, Beckman L, Villems R: The western and eastern roots of the Saami- the story of genetic "outliers" told by mitochondrial DNA and Y chromosomes. Am J Hum Genet 2004, 74: 661-682.
12. Kivisild T, Bamshad MJ, Kaldma K, Metspalu M, Reidla M, Laos S, Parik J, Waltkins WS, Dixon ME, Papiha SS, Mastana SS, Mir MR, Ferak V, Villems R: Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. Current Biol 1999, 9: 1331-1334.
13. Kaur I, Roy S, Chakrabarti S, Sarhadi KV, Majumder PP, Bhanwer AJS, Singh JR: Genomic diversities and affinities among four endogamous groups of Punjab (India) based on autosomal and mitochondrial DNA polymorphisms. Hum Biol 2002, 74: 819-836.
14. Kivisild T, Tolk HV, Parik J, Wang Y, Papiha SS, Bandelt HJ, Villems R: The emerging limbs and twigs of the east Asian mtDNA tree. Mol Biol Evol 2002, 19: 1737-1751.
15. Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, Parik J, Metspalu E, Adojaan M, Tolk HV, Stepanov V, Gölge M, Usanga E, Papiha SS, Cinnioğlu,

King R, Cavalli-Sforza LL, Underhill PA, Villems R: The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am J Hum Genet 2003, 72: 313-332.

16. Basu A, Mukherjee N, Roy S, Sengupta S, Benerjee S, Chakraborty M, Dey B, Roy M, Roy B, Bhattacharyya NP, Roychoudhury S, Majumder PP: Ethnic India: A genomic view, with special reference to peopling and structure. Genome Res 2003, 13: 2277-2290.
17. Cordeux R, Saha N, Bentley GR, Aunger R, Sirajuddin SM, Stoneking M: Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. Eur J Hum Genet 2003, 11: 253-264.
18. Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R: Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 2004, 75: 752-770.
19. Yao Y-G, Kong Q-P, Wang C_Y, Zhu C-L, Zhang Y_P. Different matrilineal contributions to genetic structure of ethnic groups in the silk road region in China: Mol Biol Evol 2004, 21: 2265-2280.
20. Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, Serk P, Karmin M, Behar DM, Gilbert MT, Endicott P, Mastana S, Papiha SS, Skorecki K, Torroni A, Villems R: Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genetics 2004, 5: 26.

21. Vishwanathan H, Deepa E, Cordeux R, Stoneking M, Usha Rani MV, Majumder PP: Genetic structure and affinities among tribal populations of southern India: a study of 24 autosomal DNA markers. Ann Hum Genet 2004, 68: 128-138.
22. Passarino G, semino O, Quintana-Murci L, Excoffier L, Hammer M, Santachiara-Benerecetti AS: Different genetic components in Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. Am J Hum Genet 1998, 62:420-434.
23. Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zheng YP. Phylogeographic differentiation of mitochondrial DNA in Han Chinese: Am J Hum Genet 2002, 70: 635-651.
24. Kong Q-P, Yao Y-G, Liu M, Shen S-P, Chen C, Zhu C-L, Palanichamy MG, Zhang Y-P: Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. Hum Genet 2003, 113:391-405.
25. Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, McElreavey K, Santachiara-Benerecetti AS: Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet 1999, 23: 437-441.
26. Howell N, Kubacka DA, Mackey DA: How rapidly does the human mitochondrial genome evolve? Am J Hum Genet 1996, 59: 501-509.
27. Maca-Meyer N, Gonzàlez AM, Larruga JM, Flores C, Cabrera VM: Major genomic mitochondrial lineages delineate early human expansions. BMC Genetics 2001, 2:13.
28. Palanichamy MG, Sun C, Agarwal S, Bandelt H-J, Kong Q-P, Khan F, Wang C-Y, Chaudhuri TK, Palla V, Zhang Y-P: Phylogeny of mtDNA superhaplogroup N

in India based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet 2004, 75:966-978.

29. Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, Anderson C, Ghosh SS, Olefsky JM, Beal MF, Davis RE, Howell N: Reduced –Median-Joining network analysis of complete mitochondrial DNA coding-region sequences for major African, Asian and European haplogroups. Am J Hum Genet 2002, 70:1152-1171.
30. Finnilä S, Hassinen IE, Al- Kokko L, Majamaa K: Phylogenetic network of the mtDNA haplogroup U in northern Finland based on sequence analysis of the complete coding region by conformation- sensitive gel electrophoresis. Am J Hum Genet 2000, 66: 1017- 1026.
31. Finnilä S, Lehtonenm MS, Majamma K: Phylogenetic networks for European mtDNA. Am J Hum Genet 2001, 68: 1475-1484.
32. Yao YG, Zhang YP: Phylogeographic analysis of mtDNA variation in four ethnic populations from Yunnan Province: new data and a reappraisal. J Hum Genet 2002, 47:311-318.
33. Kong QP, Yao YG, Sun C, Bandelt HJ, Zhu CL, Zhang YP: Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. Am J Hum Genet 2003, 73: 671-676.
34. Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad ME, Reddy PG, Rasanayagam A, Papiha SS, Villems R, Reddy AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB: Genetic evidence

on the origins of Indian caste populations. Genome Res 2001, 11: 994-1004.

35. Rajkumar R, Kashyap V.K: Mitochondrial DNA hypervariable region I and II sequence polymorphism in Dravidian linguistic group of India. J Forensic Sci 2003, 48: 227-237.

36. Rajkumar R, Kashyap V.K: Haplotype Diversity in Mitochondrial DNA Hypervariable Region I and II in Three Communities of Southern India. Forensic Sci Int. 2004, 136: 79-82.

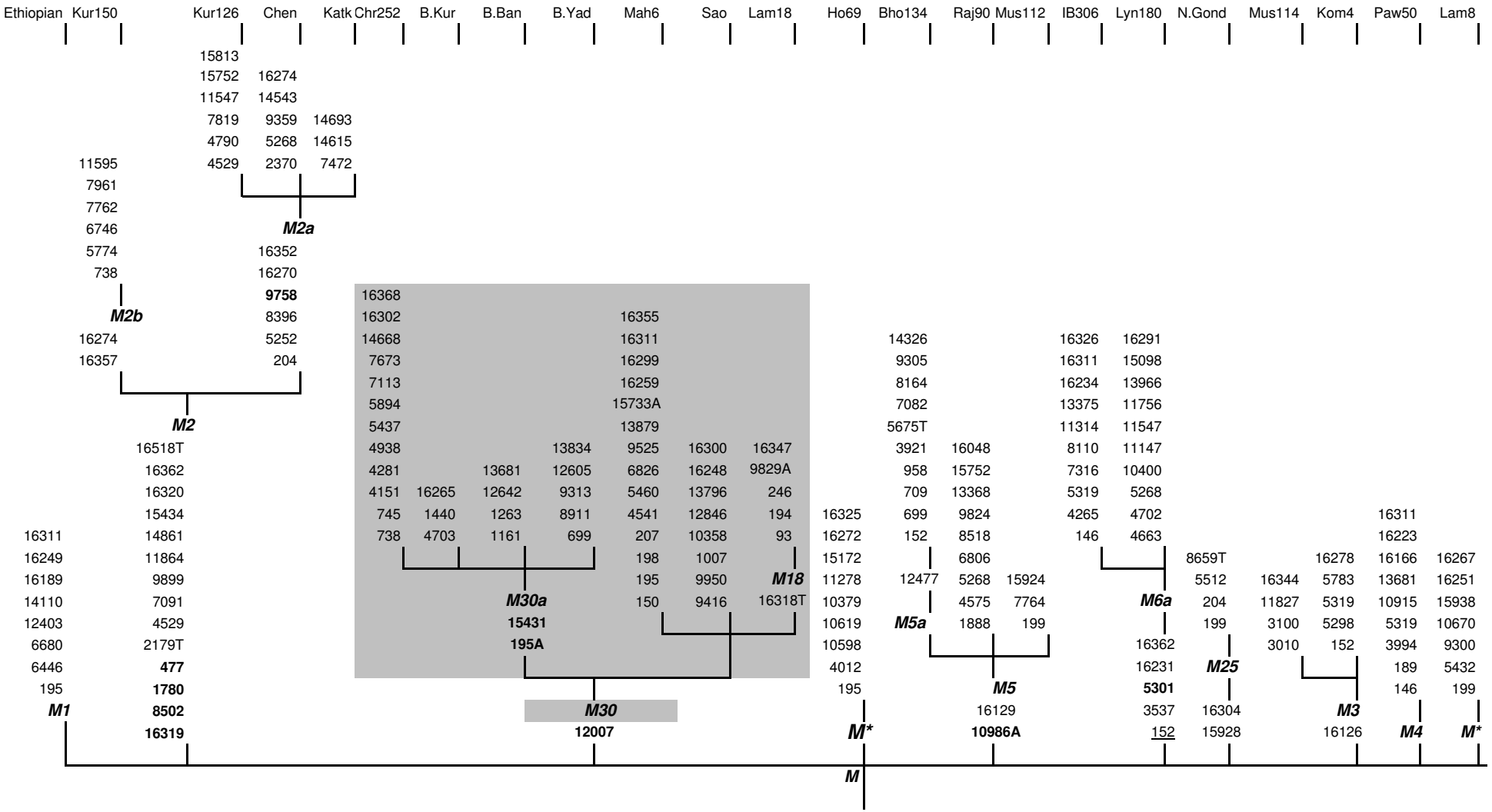
37. Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Valle G, Richards M, Maculay V, Scozari R: Do the four clades of the mtDNA haplogroups L2 evolve at different rates? Am J Hum Genet 2001, 69:1348-1356.

38. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N: Reanalysis and revisions of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 1999, 23:147.

39. Bandelt H-J, Forster P, Sykes BC, Richards MB: Mitochondrial portraits of human populations using median networks. Genetics 1995, 141:743-753.

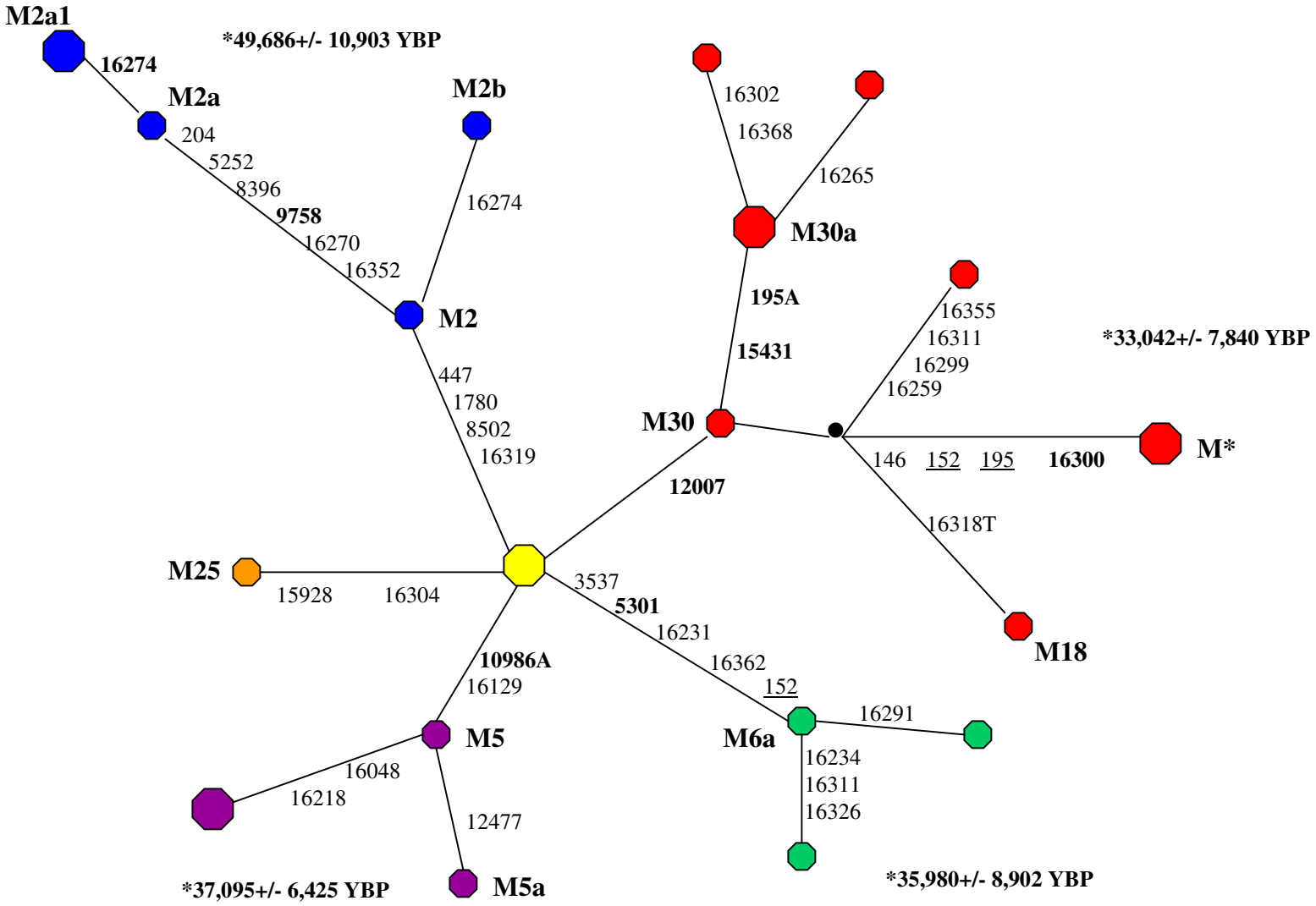
40. Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace D: Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci USA 2003, 100: 171-176.

Fig 1. Phylogenetic tree of superhaplogroup M based on compete mt DNA genome sequences.



Numbers along links refer to substitutions at nucleotide positions with respect to rCRS. Suffixes are transversions. Grey colour represents potential new lineage. Asterisk denotes unclasiffied lineage. The M haplogroup differs from the rCRS at sites: 73, 263, 750, 2706, 1438, 4769, 7028, 8701, 8860, 9540,10398, 10400, 10873, 11719, 12705, 14766, 14783, 15043, 15301, 15326, 16223.

Fig 2. Median-joining network relating HVSI sequences and coding region motifs of superhaplogroup M.



*Asterisk denotes the coalescence estimate of the lineages. Variant bases are numbered as in Fig1.