Study of human origin and migration from a novel genomic footprint derived from mitochondrial sequences

Aritra Mahapatra[1*] and Jayanta Mukherjee[2*]

Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur, India, 721302

1 aritra.mhp@iitkgp.ac.in
2 jay@cse.iitkgp.ac.in
*** corresponding author**

Abstract

The origin of modern human and their migration across the world is one of the most debated topics for the decades. There exist two different hypotheses, recent African origin and multi-regional evolution, based on the genomic studies, haplogroups, archaeological records, cultural behaviors, palaeontology studies, etc. Various studies placed the modern humans in a phylogenetic tree to depict the relationships among them. However, the conflicts between the results obtained from the molecular data and the archaeological and palaeontological reports still exist. We adopt a novel genomic features derived from the whole mitochondrial sequence, and using them phylogenetic trees are constructed providing a new insight on human migration. The results we derived from the genomic feature is more consistent with the archaeological findings based on the time of origin of different communities. We find that some Asian communities are placed at the basal point with a very high bootstrap score. This study roughly estimates the existence of the archaic human at 4-5 million years ago and presence of human in Africa at 800 kilo years ago. The basal position of the Asian communities show some close relationships between the modern Asians and the archaic humans along with the Africans.

Introduction

The origin of the anatomically modern humans (AMH) and their migration into the world is being studied extensively. The multiple genetic admixtures due to the migration, intermarriage, slavery, human trafficking etc. played an important role to make the human history very complex [1, 2]. Apart from that, different assumptions such as gene flow [3], make the study of human evolution and migration more complex [4].

Broadly, there exist two hypotheses of the human evolution and migration. The "recent African origin" or "replacement" hypothesis [5–7] states that there exist the anatomically modern humans which subsequently replace the ancient humans and spread all over the world [8–10]. There are various methods incorporated to describe the history of the humans. Several genomic studies [11–13] supported the "replacement" hypothesis. This hypothesis proposes that there is no or very less admixture or genetic mixing between AMH and the archaic humans throughout the world. But this hypothesis is disproved by some recent studies [14, 15]. Alternatively, the "multi-regional" hypothesis [16] states that the evolution of the archaic human occurred in different parts of the world. This evolution ended up with the AMH. Several linguistic analysis [17, 18], cultural analysis [19, 20] support the "multi-regional" hypotheses.

Though various genomic studies were conducted for searching the origin of humans, most of the studies focused on the phylogenetic tree derived from alignment based methods [11, 21–23], SNP genotyping [24, 25], or supertree based approach [1]. Since then, the conflicts between the genomic studies and the archaeological and palaeontological studies exist in the study of the origin of humans [26]. Many archaeologists think the ancient DNA as the "devil's work" [26]. In this work, we consider the mitochondrial genome (mtDNA) for studying the origin of humans. The mtDNA is haploid, inherited maternally [27], and recombination is very rare event in it [28]. So the changes of mtDNA sequence occur mainly due to mutations. Our study shows a consistency with the archaeological and palaeontological studies based on the age of origin of human in Africa.

In our present study, we use a novel feature representation of the whole mitochondrial genome (mtDNA) based on a graphical encoding method, named GRAFree [29]. We convert a whole mtDNA sequence into a fixed dimensional feature vector. This feature vector is used for studying the phylogeny of human representatives of different regions of this modern world. It has been hypothesized that hereditary relationship among humans of different regions is reflected by their closeness in genomic signatures, in this case, in mitochondrial genomes [30]. Discovering this relationship and building a phylogenetic hierarchy among representatives of different parts of this world is the key motivation behind the mathematical representation of a genome. We plot the mtDNAs in a two dimensional coordinate plane by considering successive shifts in x and y coordinates using three different encodings of DNA alphabets (see Methods). The 2-D distribution of points are represented by a 5-D vector, formed by the mean, eigen values, and the direction of the eigen vector corresponding to the major eigen value. We adopt a distance function to compute the pair-wise distances using the feature vectors. This comparative study finally gives us a distance matrix. We apply a hierarchical clustering method, UPGMA [31], to derive the phylogenetic tree. We consider three different encodings of shifts of coordinates corresponding to the nucleotides in the genomic sequence, and generate three phylogenetic trees independently. Finally, we combine the trees using a quartet based supertree construction method, ASTRAL (v4.10) [32].

We apply this method on a large set of mitochondrial DNA sequences of humans. In this study, we consider 371 mitochondrial genomes from more than 120 countries covering six continents (except Antarctica). The data are sequenced and published by various researchers [33–43]. They are accumulated and stored in a database server named *mtdb* [44]. Apart from that we include six outgroup species in our study (see **Supplementary S1**). We found that GRAFree distinguishes the outgroups from the humans with very high bootstrap supports. It is to be mentioned that the output shows the time of appearance of human in Africa is consistent with the archaeological and palaeontological studies. Interestingly, we found that a small population of Asia in middle east and Indian subcontinent are placed at a distinct clade than the rest of the humans with a very high bootstrap support indicating very early migration of human to this part of world. This period is roughly estimated at 4-5 million years ago (mya). This may be earlier than the period presently being considered as the period of traceable human migration to these regions, which is after 200 kya. This study also indicates presence of lineages of migrated humans to this part of world which took place much ahead of 65 kya, what is considered to be the present time line for the Out of Africa hypothesis [45]. Our finding is consistent with archaeological findings and discoveries of ancient human skeletons in Middle East, which predate the estimated periods from analysis of different mtDNA and Y-Chromosomal haplogroups of human population. It provides a new insight to the "out of Africa" hypothesis.

Results

Computation of GFP, and Drift sequence.

In this study, we consider three sets of structural groups of nucleotides (purine, pyrimidine), (amino, keto), and (weak H-bond, strong H-bond) separately for representing DNA by a sequence of points in a $2D$ integral coordinate space. We derive the GFP using these three encodings. We consider them as GFP_{RY}, GFP_{SW}, and GFP_{MK}, respectively. The drift sequence of a GFP is the coordinate

difference between those pair of points which are spanned by L number of points in between. Our proposed method, GRAFree, depends on two parameters such as the size of the window (L) and number of fragments of the sequence of drifts ($\mathcal{F}$). Each fragment is represented by a $5-D$ vector as mentioned above and concatenation of $\mathcal{F}$ vectors provide a feature vector of dimension $5\mathcal{F}$.

We compute the Shannon entropy [46] of the drift sequence of each GFP for different values of L (starting from 50 to 2000 with the difference of 50). It is observed in Fig. 1 that for all of the GFPs initially by increasing the value of L, the entropy increases till a certain limit. This incident represents that by increasing the value of L, the number of distinct points of the sequence of the drifts also increases. After a certain value of L, e.g, at $L = 550$, the entropy of all mtDNA become stabilized (changes $< 1\%$). By increasing the value of L after that does not change the entropy at any significant level. It is observed that for $L \geq 550$, the $\Delta^{(L)}$ contains significantly large number of distinct point coordinates than that of $L < 550$. Hence, we choose the value of L as 550.

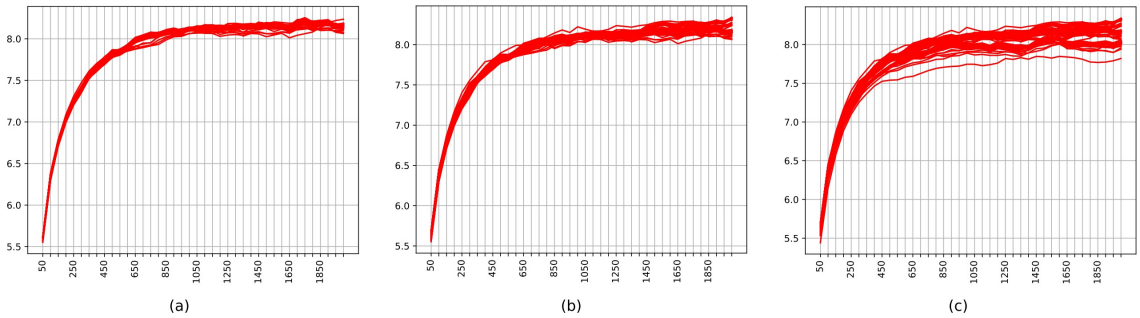

Figure 1: Shannon entropy of (a) GFP_{RY}, (b) GFP_{SW}, and (c) GFP_{MK} of different human mtDNA for all the value of L from 50 to 2000.

For getting more precise signature of a drift sequence, we make 15 non-overlapping partitions of the sequence of drift sequence. Hence, in this study, we consider $\mathcal{F}$ as 15. Each fragment is represented by the $5D$ feature vector which consists of the mean of the point coordinates of the corresponding fragment, eigen (both major and minor) values, and the angle formed by the eigen vector corresponding to the major eigen value with respect to the x-axis. Hence, a complete drift is represented by 75 dimensional feature vector.

We compute the Graphical Foot Print (GFP or ϕ) of each mtDNA of human taken from different parts of the world. For $L = 550$, we also compute the sequence of the drifts of the corresponding GFP. It is observed that the species which are closely related, have similar patterns in both of their GFP and drift sequence. Typical examples of drift sequences of humans from different regions of the world and the outgroup species, such as baboon, gorilla and chimpanzee are shown in Fig. 2.

Deriving trees.

We propose a novel distance function to compute the pair-wise distances. Our proposed distance function is a metric. We apply a hierarchical clustering method, Unweighted Pair Group Method with Arithmetic mean (UPGMA) [31], to form the phylogenetic tree from the distance matrix.

We apply this same technique on three different cases such as GFP_{RY}, GFP_{SW}, and GFP_{MK}, separately. Thus we derive three different phylogenetic trees by following three different hypotheses. These trees are presenting in the **Supplementary S2 Fig. 1-3 in Section 1**. There are various methods to combine multiple trees and form a supertree. We apply a quartet based method, ASTRAL [32], to combine three trees. Please refer to Fig. 3 and also refer to the **Supplementary S2 Fig. 4 in Section 1**.

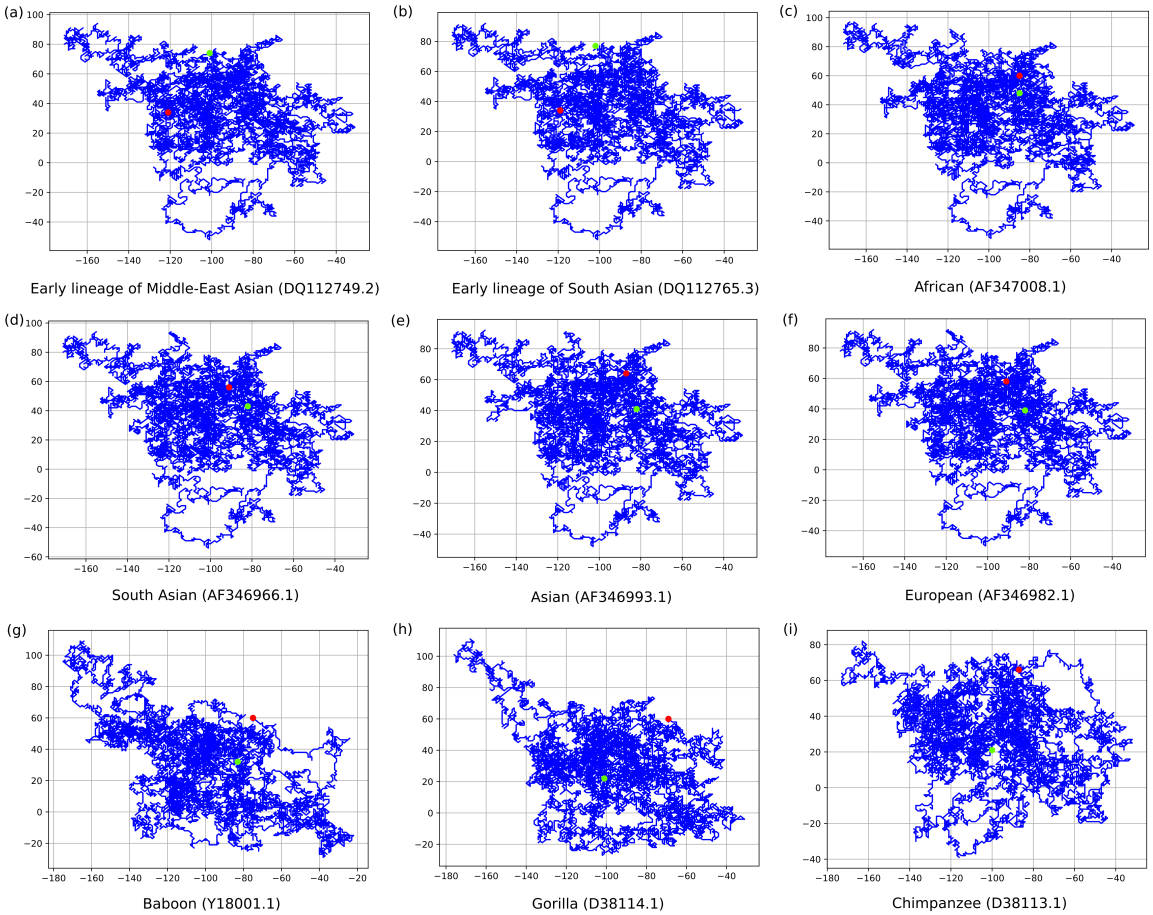

Figure 2: The sequence of drifts of different sample, mentioned under each graph, for the case when we consider the Purine and Pyrimidine as $2D$ coordinates space. The accession numbers of the corresponding samples are mentioned within the bracket. The start and end positions of the sequence of drifts are denoted by the green and red dots, respectively. The drift sequences of the Baboon, Gorilla, and Chimpanzee (g,h,i) are much different from the drift sequences of human (a-f). The drift sequences of the mtDNA of the early lineages of the Middle-East Asian and the South Asian (a,b) are slightly different from the drift sequences of the other humans (c-f).

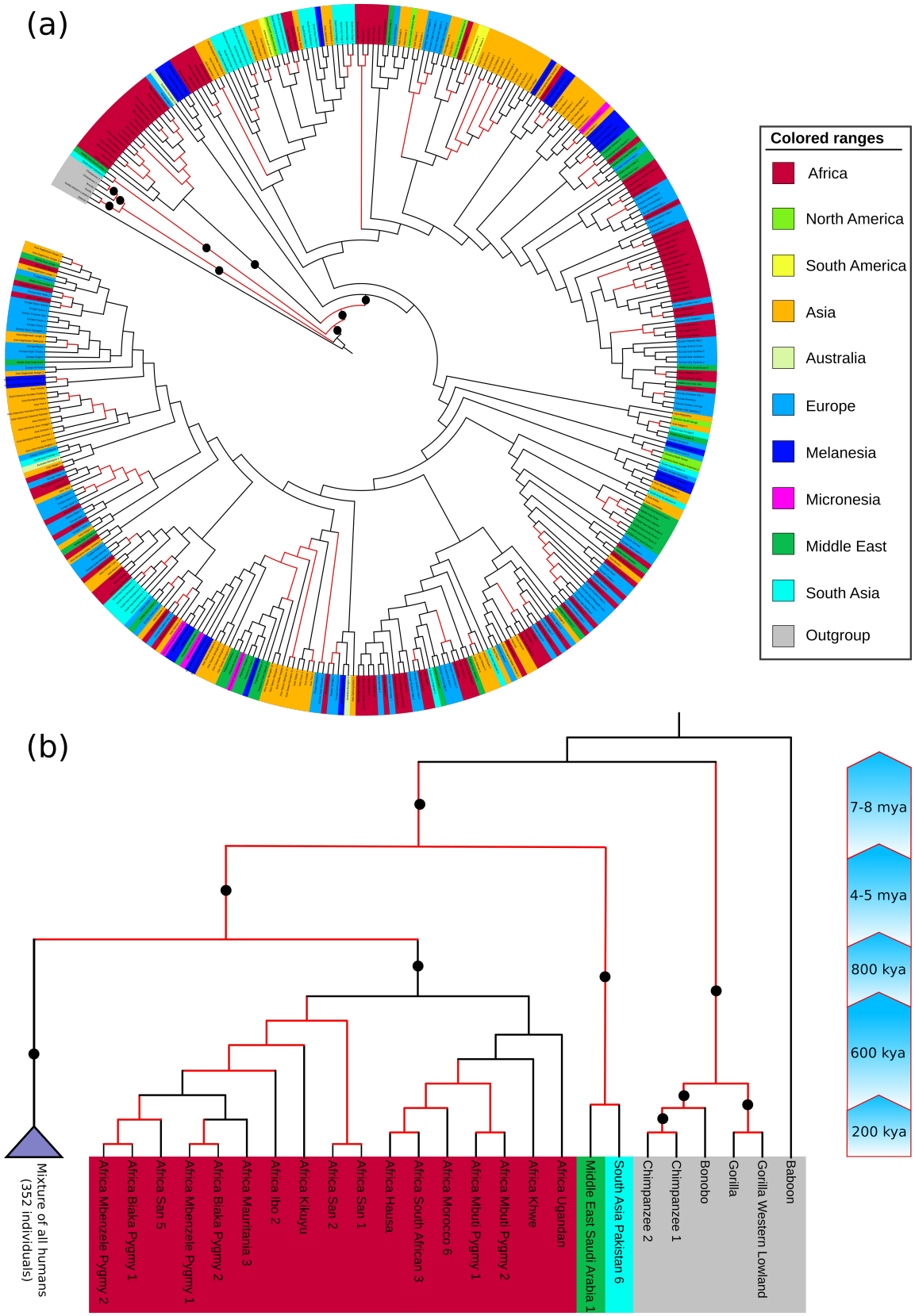

Figure 3: The derived trees using GRAFree. The black dots on the edges denote a high bootstrap support ($\geq$ 75%) of the corresponding clade. The edge marked as red color denotes that the robustness of the clade with respect to the hyper parameters is greater than 80%. (a) Tree derived after merging all three trees using ASTRAL. (b) The summarization of the tree having bootstrap support $\geq$ 75%.

Bootstrapping.

The main motivation of bootstrapping is to generate the population from a single genome. It is observed that the average intraspecies variation in mitochondrial genome of *Homo sapeince* is within 1.2% [47]. But in this study, we are trying to derive the relationships between different communities of human. Hence, we consider the average genetic variation of mtDNA for a particular community as 0.3% [48]. Here we propose a bootstrapping technique which considers the genetic variance of a sample space within 0.3%. For that we apply insertion, deletion, and mutation at each location with an equal probabilty of 0.1% and consider an unbiased selection of the nucleotides at each location for the insertion and the mutaion. We generate 100 replicas using this proposed bootstrapping method and construct trees from each set of sequences using GRAFree method by setting values of L and $\mathcal{F}$ as discussed in the previous section. Felsenstein's bootstrapping method [49] assesses the robustness of phylogenetic trees using the presence and absence of clades. For the large scale genomics Felsenstein's bootstrap is not efficient to sum up the replicas. For the hundred of species this method is inclined to produce low bootstrap support [50]. So here we apply a modification of Felsenstein's bootstrapping, where the presence of a clade is quantified using the transfer distance proposed in [50]. The transfer distance [51] is the minimum number of changes required to transform one partition to other. We compute the occurrence of each clade using the tool BOOSTER [50]. The clades having a high bootstrap support ($\geq 75\%$) are denoted as the black dot on the edge in Fig 3 and also refer to the **Supplementary S2 Fig. 1-4 in Section 1**.

Robustness against variation of hyper parameters.

Here we test the consistency of the results for different values of L and $\mathcal{F}$. For that we generate a large number of trees by considering the values of L from 500 to 1000 and $\mathcal{F}$ from 5 to 50. We consider the tree derived for $L = 550$ and $\mathcal{F} = 15$ as the reference tree. It is observed that most of the clades of the reference tree is consistent (occurrences more than 80%) for different parameter values (Fig. 3) and also refer to the **Supplementary S2 Fig. 1-4 in Section 1**.

Estimation of time of origin of clades.

We applied UPGMA technique to get the edge lengths of the derived trees. Most of the studies proposed that the speciation of chimpanzee and human occurred 7-8 million years ago (mya) [52–54]. To calibrate the distances of the derived trees we consider the speciation age of chimpanzee and human and derived the age of the origin of different group of humans. We have used this calibration to estimate time of origins of clades which have strong bootstrap supports. It is observed that the most recent common ancestor of all the humans is originated about 4-5 mya. Most of the human population along with the Africans are originated about 800 kya, while the African clade was originated 600 kya. Interestingly, all the trees derived by considering three different encodings, show a small group of human population separated 4-5 mya ago which finally distributed within the Middle East of the Asia and the Indian subcontinent almost 200 kya (Fig. 3). The trees with the edge lengths are provided in the **Supplementary S3**.

Discussion

We develop a new approach, GRAFree, to derive the phylogenetic relationships among a large number of mitochondrial genome of AMH of different regions of this world. We also include six outgroup species (such as other than human) in our study. The proposed approach is quite different from mitochondrial haplogroup based approaches. Though this method does not provide resolve phylogenetic relations of very closely related groups, it is effective in capturing broad distinct clades. This is evident from successful identification of the phylogentic relationship of outgroups in out study. We have applied the Euclidean metric in the same 75-D feature space while forming the phylogenetic by the UPGMA algorithm. It is observed that the Euclidean distance function shows

some discrepancies in the relationships between the outgroups and humans. It is also observed that the proposed distance function is more robust than the Euclidean distance function for the selected feature vector.

In the human clade, in contrary to present understanding of human migration, it is observed that a clade of Middle East Asian and the South Asian is placed at the most out side of the human clade with a very high bootstrap support score of 100%. This incident is very consistent for all the combinations of the parameter values (Fig. 3). The same phenomena is also found when the Euclidean distance function is used for building the phylogenetic tree (see **Supplementary S2 Fig. 5 in section 2**). This is to be mentioned that the two mtDNA of these early lineages of Middle East Asian and the South Asian were sequenced from the sample of Arab and Balochisthan, respectively, by Kivisild, et al. [55]. These sequences are submitted to the GenBank with the accession numbers of *DQ112749.2* and *DQ112765.3*, respectively [55]. Though various genetic studies and studies on SNPs and haplogroups do not support this formation of a clade, but the robustness and the consistency of the occurrence of this clade pledge us to rethink the occurrence of the particularly two members of our dataset that show the early speciation of humans. In this study, as we are dealing with intraspecies mitochondrial genome, hence, the differences in the GFP and the drift are very less. This is observed that the drifts of the two GFPs of early lineage of the Southern Asia and Middle East Asia are different from the drifts of the other humans (Fig. 2). This phenomena denotes that there exists a hypothetical ancestor almost 4-5 mya whose genome signature is carried by some of the Middle East Asian and the South Asian population and the group of people are less affected due to the admixture. Few studies agreed the existance of *"ghost archaic hominin"* which have diverged from the ancestors of AMHs, Neanderthals and Denisovans almost 4 mya [45, 56, 57]. There are some studies which analysed the genome-wide data and the artefacts, reported that the ancient human was widespread across the globe [26]. Some impressions of the presence of the Denisovans earlier than the fossils record added that the human were more widespread than we think [58]. But they all are extinct after a time being [26, 58]. Researchers found some signatures of the admixture with modern humans in as recent as 15-30 kya, such that almost $3-5\%$ of the Asian DNA are inherited from Denisovans [59]. The haplogroup study on our data does not corroborate the incident of the early migration of Southern Asia and Middle East Asia but our study using derived feature vectors from whole mtDNAs shows a strong uphold of this phenomena.

It is also observed that the Middle East Asian and the South Asian population are closely related to the African population. There are some archaeological studies that report the similarities in the ancient instruments in African and Asian population [2]. The archaeologists indicated the appearance of the first modern humans in India are almost 120 kya at Bhimbetka in the state Madhya Pradesh which contradicts the study of the geneticist [60]. Various studies proposed the close relation between north African and Levantine based on the haplogroups of both mtDNA and Y-chromosomes [38, 61, 62]. All the derived trees show a very mixed representation of the humans from different communities. Though the bootstrap supports of all of these occurrences is lower than the accepted significant level, but appearance of the same event in a repetitive manner shows the high possibility of the admixture in modern humans.

In conclusion, we would like to report that as per our knowledge goes, this is the first study of human origin based on the graphical based method where we inspect the origin and the migration of human based on a novel feature vector. This is to be mentioned that the proposed method can be extended to other types of genomic sequences. This method captures some unique signature of the whole mtDNA. It is observed that this method efficiently captures the broad relationships rather than the precise relationships near the leaf nodes. In this study, we present such broad divergence among the human population. Our result supports some existing hypotheses like the close relationship among the Asian and the African population. The origin of the African clade in our study is also close to the group of archaeological reports. Apart from that, this study provides new evidences of appearance of a small Asian population out side of the African population and the association of all the human communities together.

Methods

We consider three sets of structural groups of nucleotides (purine, pyrimidine), (amino, keto), and (weak H-bond, strong H-bond) separately for representing DNA by a sequence of points in a 2-D integral coordinate space. This point set is called **Graphical Foot Print (GFP)** of a DNA sequence. We adopt a technique for extracting features from GFPs and use them for constructing phylogenetic trees [29]. As there are three different types of numerical representation of nucleotides, we could form three different hypotheses for the phylogeny. Finally, we combine the trees generated from these three hypotheses by applying a tree merging algorithm called ASTRAL [32, 63].

Feature space.

Definition 1. ***Graphical Foot Print (GFP).***

Let a sequence, $\mathcal{S} \in \Sigma^+$, $\Sigma = \{A, T, G, C\}$. For each combination of Purine (R)/Pyrimidine (Y), Amino (M)/Keto (K), and Strong H-bond (S)/Weak H-bond (W),the GFP of $\mathcal{S}$, $\phi(\mathcal{S})$, is the locus of 2-D points in an integral coordinate space, such that (x_i, y_i) is the coordinate of the alphabet s_i, $\forall\ s_i \in \mathcal{S}$, for $i = 1, 2, \ldots, n$, and $x_0 = y_0 = 0$.

Case-1: for Purine/Pyrimidine

$$\begin{aligned} x_i &= x_{i-1} + 1; \quad \textit{if } s_i = G \\ &= x_{i-1} - 1; \quad \textit{if } s_i = A \\ &= x_{i-1}; \quad \textit{otherwise} \\ y_i &= y_{i-1} + 1; \quad \textit{if } s_i = C \\ &= y_{i-1} - 1; \quad \textit{if } s_i = T \\ &= y_{i-1}; \quad \textit{otherwise} \end{aligned} \tag{1}$$

Case-2: for Strong H-bond/Weak H-bond

$$\begin{aligned} x_i &= x_{i-1} + 1; \quad \textit{if } s_i = C \\ &= x_{i-1} - 1; \quad \textit{if } s_i = G \\ &= x_{i-1}; \quad \textit{otherwise} \\ y_i &= y_{i-1} + 1; \quad \textit{if } s_i = T \\ &= y_{i-1} - 1; \quad \textit{if } s_i = A \\ &= y_{i-1}; \quad \textit{otherwise} \end{aligned} \tag{2}$$

Case-3: for Amino/Keto

$$\begin{aligned} x_i &= x_{i-1} + 1; \quad \textit{if } s_i = A \\ &= x_{i-1} - 1; \quad \textit{if } s_i = C \\ &= x_{i-1}; \quad \textit{otherwise} \\ y_i &= y_{i-1} + 1; \quad \textit{if } s_i = T \\ &= y_{i-1} - 1; \quad \textit{if } s_i = G \\ &= y_{i-1}; \quad \textit{otherwise} \end{aligned} \tag{3}$$

We denote GFPs of Case-1, Case-2 and Case-3, as GFP-RY (Φ_{RY}), GFP-SW (Φ_{SW}) and GFP-MK (Φ_{MK}), respectively.

Definition 2. ***Drift of GFP.***

Let $\mathcal{S}$ be a DNA sequence and s_i be the alphabet ($s_i \in \{A, T, G, C\}$) at the ith position of $\mathcal{S}$. Let $\phi_i(\mathcal{S})$ denote the corresponding (x_i,y_i) the coordinate of s_i in $\Phi(\mathcal{S})$.

Then for length L, drift at the ith position is defined as,
$\delta_i^{(L)} = \phi_{i+L}(\mathcal{S}) - \phi_i(\mathcal{S})$, *where* $(i + L) \leq |\mathcal{S}|$

Considering the drifts for every i^{th} location of the whole sequence, the sequence of drifts is denoted by
$\Delta^{(L)} = [\delta_0^{(L)}, \delta_1^{(L)}, \delta_2^{(L)}, \delta_3^{(L)}, \ldots, \delta_m^{(L)}]$, *where* $(m + L) = |\mathcal{S}|$

For GFP-RY (refer to Definition 1), an element (x_i, y_i) in $\Delta_{RY}^{(L)}$ provides excess numbers of G from A and C from T in segment of length L starting from the ith location, respectively. Similarly, in GFP-SW, they are the excess numbers of C from G and T from A (represents as $\Delta_{SW}^{(L)}$), and in GFP-MK, they correspond to the excess numbers of A from C and T from G (represents as $\Delta_{MK}^{(L)}$), respectively.

We also call the elements of $\Delta^{(L)}$ as points, as they can be plotted on a 2-D coordinate system. We call this plot as the scatter plot of the drift sequence. Similarly, we get a scatter plot of a GFP. It has been observed that in many cases the scatter plots of Δ have similar structure for closely spaced species mentioned in literature. It can also be observed that differences between two GFPs get reflected in their respective drifts.

We represent spatial distribution of these points of Δ by an elliptical model using a five dimensional feature descriptor: $(\mu_x, \mu_y, \Lambda, \lambda, \theta)$, where (μ_x, μ_y) is the center of the coordinates, Λ and λ are major and minor eigen values of the covariance matrix, and θ is the angle formed by the eigen vector corresponding to Λ with respect to the x-axis. We make $\mathcal{F}$ number of non overlapping equal length fragments from Δ and represent each fragment using the five dimensional feature descriptor.

Distance function.

For two sequences $\mathcal{P}$ and $\mathcal{Q}$ with the feature descriptors of ith fragments $(\mu_{x\mathcal{P}i}, \mu_{y\mathcal{P}_i}, \Lambda_{\mathcal{P}i}, \lambda_{\mathcal{P}i}, \theta_{\mathcal{P}i})$ and $(\mu_{x\mathcal{Q}i}, \mu_{y\mathcal{Q}_i}, \Lambda_{\mathcal{Q}i}, \lambda_{\mathcal{Q}i}, \theta_{\mathcal{Q}i})$, where $i \leq \mathcal{F}$, we use the following distance function between them,

$$D(\mathcal{P}, \mathcal{Q}) = \frac{1}{\mathcal{F}} \sum_{i=1}^{\mathcal{F}} \Bigg[\sqrt{{\mu_{\mathcal{P}i}}^T \mu_{\mathcal{P}i} + {\mu_{\mathcal{Q}i}}^T \mu_{\mathcal{Q}i} - 2{\mu_{\mathcal{P}i}}^T \mu_{\mathcal{Q}i} cos(\theta_{\mathcal{P}i} - \theta_{\mathcal{Q}i})} + \sqrt{(\Lambda_{\mathcal{P}i} - \Lambda_{\mathcal{Q}i})^2 + (\lambda_{\mathcal{P}i} - \lambda_{\mathcal{Q}i})^2} \Bigg]$$
$$\text{where, } \mu_{\mathcal{P}i} = [\mu_{x\mathcal{P}i}, \mu_{y\mathcal{P}_i}] \text{ and } \mu_{\mathcal{Q}i} = [\mu_{x\mathcal{Q}i}, \mu_{y\mathcal{Q}_i}] \quad (4)$$

It can be shown that the above distance function is a metric.

Code availability

We implement a Python based pipeline available at `http://www.facweb.iitkgp.ac.in/~jay/GraFree/HIIARG-GRAFree.html`. The mirror copy of this program is also available at `https://github.com/aritramhp/GRAFree.git`. This program (i) computes the tree based on the GRAFree algorithm, (ii) generates multiple number of bootstrap sequence of a target sequence, and (iii) derives the bootstrap tree from the bootstrap sequence.

Data availability

All the data are publicly available in the online database, NCBI, with their accession number. The list of accession numbers we consider in this study are provided in Supplementary Table 1.

References

[1] Duda, P. & Zrzavỳ, J. Human population history revealedby a supertree approach. *Scientific Reports* **6**, 29890 (2016).

[2] Mellars, P. Going East: New Genetic and Archaeological Perspectives on the Modern Human Colonization of Eurasia. *Science* **313**, 796–800 (2006).

[3] Shriner, D., Tekola-Ayele, F., Adeyemo, A. & Rotimi, C. N. Ancient Human Migration after Out-of-Africa. *Scientific Reports* **6** (2016).

[4] Bae, C. J., Douka, K. & Petraglia, M. D. On the origin of modern humans: Asian perspectives. *Science* **358**, eaai9067 (2017).

[5] Chan, E. K. *et al.* Human origins in a southern African palaeo-wetland and first migrations. *Nature* 1–5 (2019).

[6] Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31 (1987).

[7] Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507 (1991).

[8] Liu, H., Prugnolle, F., Manica, A. & Balloux, F. A Geographically Explicit Genetic Model of Worldwide Human-Settlement History. *The American Journal of Human Genetics* **79**, 230–237 (2006).

[9] Stringer, C. Out of Ethiopia. *Nature* **423**, 692 (2003).

[10] Wolpoff, M. H., Hawks, J. & Caspari, R. Multiregional, not multiple origins. *American Journal of Physical Anthropology* **112**, 129–136 (2000).

[11] Mondal, M., Bertranpetit, J. & Lao, O. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications* **10**, 246 (2019).

[12] Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173**, 53–61 (2018).

[13] Mondal, M. *et al.* Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nature Genetics* **48**, 1066 (2016).

[14] Mafessoni, F. Encounters with archaic hominins. *Nature Ecology & Evolution* **3**, 14–15 (2019).

[15] Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43 (2014).

[16] Cartmill, M. & Smith, F. H. *The Human Lineage* (Wiley-Blackwell, 2009).

[17] Baker, J. L., Rotimi, C. N. & Shriner, D. Human ancestry correlates with language and reveals that race is not an objective genomic classifier. *Scientific Reports* **7**, 1572 (2017).

[18] Balter, M. Language May Have Helped Early Humans Spread Out of Africa. *Science* (2019). URL `https://www.sciencemag.org/news/2011/04/language-may-have-helped-early-humans-spread-out-africa`.

[19] Bhugra, D. & Becker, M. A. Migration, cultural bereavement and cultural identity. *World Psychiatry* **4**, 18 (2005).

[20] Mesoudi, A. Migration, acculturation, and the maintenance of between-group cultural variation. *PLOS ONE* **13**, e0205573 (2018).

[21] Jinam, T. A. *et al.* Discerning the Origins of the Negritos, First Sundaland People: Deep Divergence and Archaic Admixture. *Genome Biology and Evolution* **9**, 2013–2022 (2017).

[22] Töpf, A. L., Gilbert, M. T. P., Dumbacher, J. P. & Hoelzel, A. R. Tracing the Phylogeography of Human Populations in Britain Based on 4th11th Century mtDNA Genotypes. *Molecular Biology and Evolution* **23**, 152–161 (2005).

[23] Basu, A. *et al.* Ethnic India: A Genomic View, With Special Reference to Peopling and Structure. *Genome Research* **13**, 2277–2290 (2003).

[24] Juyal, G. *et al.* Population and genomic lessons from genetic analysis of two Indian populations. *Human Genetics* **133**, 1273–1287 (2014).

[25] Moorjani, P. *et al.* Genetic Evidence for Recent Population Mixture in India. *The American Journal of Human Genetics* **93**, 422–438 (2013).

[26] Callaway, E. Divided by DNA: The uneasy relationship between archaeology and ancient genomics. *Nature* **555** (2018).

[27] Cummins, J. M., Wakayama, T. & Yanagimachi, R. Fate of microinjected spermatid mitochondria in the mouse oocyte and embryo. *Zygote* **5**, 301308 (1997).

[28] Eyre-Walker, A. & Awadalla, P. Does human mtdna recombine? *Journal of Molecular Evolution* **53**, 430–435 (2001).

[29] Mahapatra, A. & Mukherjee, J. GRaphical footprint based Alignment-Free method (GRAFree) for classifying the species in Large-Scale Genomics. In *8th International Conference on Pattern Recognition and Machine Intelligence, PReMI* (Springer, 2019). DOI: 10.1007/978-3-030-34872-4_12 (Preprint).

[30] Langille, M. G. I., Hsiao, W. W. L. & Brinkman, F. S. L. Detecting genomic islands using bioinformatics approaches. *Nature Reviews Microbiology* **8**, 373–382 (2010).

[31] Sneath, P. H. A. & Sokal, R. R. *Numerical Taxonomy* (W. H. Freeman and Company, San Francisco, 1973).

[32] Sayyari, E. & Mirarab, S. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular Biology and Evolution* **33**, 1654–1668 (2016).

[33] Friedlaender, J. *et al.* Expanding Southwest Pacific Mitochondrial Haplogroups P and Q. *Molecular Biology and Evolution* **22**, 1506–1517 (2005).

[34] Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708 (2000).

[35] Roostalu, U. *et al.* Origin and Expansion of Haplogroup H, the Dominant Human Mitochondrial DNA Lineage in West Eurasia: The Near Eastern and Caucasian Perspective. *Molecular Biology and Evolution* **24**, 436–448 (2006).

[36] Hill, C. *et al.* Phylogeography and Ethnogenesis of Aboriginal Southeast Asians. *Molecular Biology and Evolution* **23**, 2480–2491 (2006).

[37] Achilli, A. *et al.* Saami and Berbers-An Unexpected Mitochondrial DNA Link. *The American Journal of Human Genetics* **76**, 883–886 (2005).

[38] Olivieri, A. *et al.* The mtDNA Legacy of the Levantine Early Upper Palaeolithic in Africa. *Science* **314**, 1767–1770 (2006).

[39] Maca-Meyer, N., González, A. M., Larruga, J. M., Flores, C. & Cabrera, V. M. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics* **2**, 13 (2001).

[40] Starikovskaya, E. B. *et al.* Mitochondrial DNA Diversity in Indigenous Populations of the Southern Extent of Siberia, and the Origins of Native American Haplogroups. *Annals of Human Genetics* **69**, 67–89 (2005).

[41] Behar, D. M. *et al.* The Matrilineal Ancestry of Ashkenazi Jewry: Portrait of a Recent Founder Event. *The American Journal of Human Genetics* **78**, 487–497 (2006).

[42] Ingman, M. & Gyllensten, U. Mitochondrial Genome Variation and Evolutionary History of Australian and New Guinean Aborigines. *Genome Research* **13**, 1600–1606 (2003).

[43] Ingman, M. & Gyllensten, U. A recent genetic link between Sami and the Volga-Ural region of Russia. *European Journal of Human Genetics* **15**, 115 (2007).

[44] Ingman, M. & Gyllensten, U. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. *Nucleic Acids Research* **34**, D749–D751 (2006).

[45] Veeramah, K. R. & Hammer, M. F. The impact of whole-genome sequencing on the reconstruction of human population history. *Nature Reviews Genetics* **15**, 149–162 (2014).

[46] Shannon, C. E. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* **5**, 3–55 (2001).

[47] Castro, J. A., Picornell, A. & Ramon, M. Mitochondrial DNA: a tool for populational genetics studies. *International Microbiology* **1**, 327–332 (1998).

[48] Jarczak, J. *et al.* Mitochondrial DNA variability of the Polish population. *European Journal of Human Genetics* 1 (2019).

[49] Felsenstein, J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution* **39**, 783–791 (1985).

[50] Lemoine, F. *et al.* Renewing Felsensteins phylogenetic bootstrap in the era of big data. *Nature* 1 (2018).

[51] Charon, I., Denoeud, L., Guénoche, A. & Hudry, O. Maximum transfer distance between partitions. *Journal of Classification* **23**, 103–121 (2006).

[52] Langergraber, K. E. *et al.* Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences* **109**, 15716–15721 (2012).

[53] Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103 (2006).

[54] Reich, D. *Who we are and how we got here: Ancient DNA and the new science of the human past* (Oxford University Press, 2018).

[55] Kivisild, T. *et al.* The role of selection in the evolution of human mitochondrial genomes. *Genetics* **172**, 373–387 (2006).

[56] Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328**, 710–722 (2010).

[57] Mendez, F. L., Watkins, J. C. & Hammer, M. F. Global genetic variation at OAS1 provides evidence of archaic admixture in Melanesian populations. *Molecular Biology and Evolution* **29**, 1513–1520 (2012).

[58] Warren, M. Biggest Denisovan fossil yet spills ancient humans secrets. *Nature* **569**, 16–17 (2019).

[59] Gibbons, A. Our mysterious cousinsthe Denisovansmay have mated with modern humans as recently as 15,000 years ago. *Science* (2019). URL https://www.sciencemag.org/news/2019/03/our-mysterious-cousins-denisovans-may-have-mated-modern-humans-recently-15000-years-ago.

[60] Joseph, T. *Early Indians: The Story of Our Ancestors and Where We Came From* (Juggernaut Publication, 2018).

[61] van de Loosdrecht, M. *et al.* Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. *Science* **360**, 548–552 (2018).

[62] Elkamel, S., Boussetta, S., Khodjet-El-Khil, H., Benammar Elgaaied, A. & Cherni, L. Ancient and recent Middle Eastern maternal genetic contribution to North Africa as viewed by mtDNA diversity in Tunisian Arab populations. *American Journal of Human Biology* **30**, e23100 (2018).

[63] Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–i548 (2014).