Human origin and migration deciphered from a novel genomic footprints of mitochondrial sequences

Aritra Mahapatra[1], 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

1 Abstract

2 The origin of modern human and their migration across the world is one of the most
3 debated topics for the decades. There exist two different hypotheses, recent African
4 origin and multi-regional evolution, based on the genomic studies, haplogroups, archae-
5 ological records, cultural behaviors, palaeontology studies, etc. Various studies placed
6 the modern humans in a phylogenetic tree to depict the relationships among them.
7 The debate for determining those regions of Africa which witnessed the first origin
8 of humans still exists. The conflicts between the results obtained from the molecular
9 data and the archaeological and palaeontological reports still exist. We adopt a novel
10 genomic feature derived from the whole mitochondrial sequence, and using a novel
11 distance function the phylogenetic trees are constructed based on the feature which
12 provide a new insight on human migration. We propose a new method to derive the
13 bootstrap replica from the genome sequences by considering the genetic variance to
14 demonstrate the robustness of the obtained trees. The results derived from the ge-
15 nomic feature are more consistent with the archaeological findings based on the time
16 of origin of different communities. We find that west and central African communities
17 are placed at the basal point with a very high bootstrap score. This study roughly
18 estimates the existence of the archaic human at 800-900 kilo years ago and presence of
19 human in Africa at 600-700 kilo years ago. This supports the presence of an ancestor
20 in the west and central Africa much earlier than that of the fossils identified.

1   Introduction

24 The origin of the anatomically modern humans (AMH) and their migration into the world is being

25 studied extensively. The multiple genetic admixtures due to migration, intermarriage, slavery,

26 human trafficking etc. played an important role to make the human history very complex (Duda

27 and Zrzavỳ, 2016; Mellars, 2006). Apart from that, different assumptions such as gene flow (Shriner

28 et al., 2016), make the study of human evolution and migration more complex (Bae et al., 2017).

29   Broadly, there exist two hypotheses of human evolution and migration. The "recent African

30 origin" or "replacement" hypothesis (Chan et al., 2019; Cann et al., 1987; Vigilant et al., 1991)

31 states that the already existing anatomically modern humans subsequently replace the ancient hu-

32 mans and spread all over the world (Liu et al., 2006; Stringer, 2003; Wolpoff et al., 2000). There

33 are various methods incorporated to describe the history of the humans. Several genomic stud-

34 ies (Mondal et al., 2019; Browning et al., 2018; Mondal et al., 2016) supported the "replacement"

35 hypothesis. This hypothesis proposes that there is no or very less admixture or genetic mixing be-

36 tween AMH and the archaic humans throughout the world. But this hypothesis is negated by some

37 recent studies (Mafessoni, 2019; Prüfer et al., 2014). Alternatively, the "multi-regional" hypothe-

38 sis (Cartmill and Smith, 2009) states that the evolution of the archaic human occurred in different

39 parts of the world. This evolution ended up with the AMH. Several linguistic analysis (Baker et al.,

40 2017; Balter, 2019), and cultural analysis (Bhugra and Becker, 2005; Mesoudi, 2018) support the

41 "multi-regional" hypothesis.

42   Though various genomic studies were conducted for searching the origin of humans, most of

43 the studies focused on the phylogenetic tree derived from alignment based methods (Mondal et al.,

44 2019; Jinam et al., 2017; Töpf et al., 2005; Basu et al., 2003), SNP genotyping (Juyal et al., 2014;

45 Moorjani et al., 2013), or supertree based approach (Duda and Zrzavỳ, 2016). Since then, the

46 conflicts between the genomic studies and the archaeological and palaeontological studies exist in

47 the question of the origin of humans (Callaway, 2018). As a result many archaeologists think

48 the ancient DNA as the "devil's work" (Callaway, 2018). Though the origin of the human in the

49 eastern and southern parts of Africa has long been accepted (Sahle et al., 2018; Behar et al., 2008;

2

50 Chan et al., 2015, 2019), a few recent studies documented the presence of humans in the western

51 part of Africa much earlier than the other fossils identified till date (Stringer and Galway-Witham,

52 2017; Durvasula and Sankararaman, 2020). In this work, we consider the mitochondrial genome

53 (mtDNA) for studying the origin of humans. The mtDNA is haploid, inherited maternally (Hurles

54 and Jobling, 2001), and recombination is very rare event in it (Eyre-Walker and Awadalla, 2001).

55 So the changes of mtDNA sequence occur mainly due to mutations. Our study shows a consistency

56 with the archaeological and palaeontological studies based on the age of origin of human in Africa.

57 Moreover, this study also supports the recently published studies on the origin of human in Africa.

58 In our present study, we use a novel feature representation of the whole mitochondrial genome

59 (mtDNA) based on a graphical encoding method, named GRAFree (Mahapatra and Mukherjee,

60 2019). We convert a whole mtDNA sequence into a fixed dimensional feature vector. This feature

61 vector is used for studying the phylogeny of human representatives of different regions of this mod-

62 ern world. It has been hypothesized that hereditary relationship among humans of different regions

63 is reflected by their closeness in genomic signatures, in this case, in mitochondrial genomes (Langille

64 et al., 2010). Discovering this relationship and building a phylogenetic hierarchy among representa-

65 tives of different parts of this world is the key motivation behind the mathematical representation

66 of a genome. We plot the mtDNAs in a two dimensional coordinate plane by considering successive

67 shifts in x and y coordinates using three different encodings of DNA alphabets (see Methods). The

68 2-D distribution of points are represented by a 5-D vector, formed by the mean, eigen values, and

69 the direction of the eigen vector corresponding to the major eigen value. We adopt a distance func-

70 tion to compute the pair-wise distances using the feature vectors. This comparative study finally

71 gives us a distance matrix. We apply a hierarchical clustering method, UPGMA (Sneath and Sokal,

72 1973), to derive the phylogenetic tree. We consider three different encodings of shifts of coordinates

73 corresponding to the nucleotides in the genomic sequence, and generate three phylogenetic trees

74 independently. Finally, we combine the trees using a quartet based supertree construction method,

75 ASTRAL (v4.10) (Sayyari and Mirarab, 2016). We believe that, this is the first study of human

76 origin based on the graphical representation of genome.

77 We apply this method on a large set of mitochondrial DNA sequences of humans. In this study,

78 we consider 369 mitochondrial genomes from more than 120 countries covering six continents (except

79 Antarctica). The data are sequenced and published by various researchers (Friedlaender et al., 2005;

80 Ingman et al., 2000; Roostalu et al., 2006; Hill et al., 2006; Achilli et al., 2005; Olivieri et al., 2006;

3

Maca-Meyer et al., 2001; Starikovskaya et al., 2005; Behar et al., 2006; Ingman and Gyllensten, 2003, 2007). They are accumulated and stored in a database server named *mtDB* (Ingman and Gyllensten, 2006). 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. The output shows that the time of appearance of human in Africa is consistent with the archaeological and palaeontological studies. We found that a significantly large population from the western and central parts of Africa are placed at a distinct clade as compared to the rest of the humans with a very high bootstrap support indicating a divergence of the genomic signatures of the western and central African population from the rest of the humans. This phenomena also indicates an early origin of human from these parts of Africa to the world. This period is roughly estimated to be 600-700 kilo years ago (kya). This may be earlier than the period presently being considered as the period of traceable human origin to these regions, which is between 230-300 kya (Andrews and Johnson, 2019; Smith and Ahern, 2013). 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 (Veeramah and Hammer, 2014). Our finding is consistent with archaeological findings and discoveries of ancient human skeletons, 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. In addition, this study also indicates the ancestor of the African populations of western and the central parts appeared much earlier than the period being documented as the origin of Neanderthals and Denisovans (Veeramah and Hammer, 2014; Green et al., 2010; Mendez et al., 2012).

2   Materials and 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 (Mahapatra and Mukherjee, 2019). 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

4

110   called ASTRAL (Mirarab et al., 2014; Sayyari and Mirarab, 2016).

2.1   Feature space

112   **Definition 1 *Graphical Foot Print (GFP).*** *Let $\mathcal{S}$ be a DNA sequence, such that, $\mathcal{S} \in \Sigma^{+}$,*

113   *$\Sigma = \{A, T, G, C\}$. For each combination of Purine (R)/Pyrimidine (Y), Amino (M)/Keto (K), and*

114   *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*

115   *coordinate space, such that (x_i, y_i) is the coordinate of the alphabet s_i, $\forall s_i \in \mathcal{S}$, for $i = 1, 2, \ldots, n$,*

116   *and $x_0 = y_0 = 0$.*

117   *Case-1: for Purine/Pyrimidine*

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

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

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

5

119   *Case-3: for Amino/Keto*

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

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

121   (Φ_{MK}), respectively.

122   **Definition 2 *Drift of GFP.*** *Let $\mathcal{S}$ be a DNA sequence and s_i be the alphabet ($s_i \in \{A, T, G, C\}$)*

123   *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 $\Phi(\mathcal{S})$.*

124   *Then for length L, drift at the ith position is defined as,*

125     $\delta_i^{(L)} = \phi_{i+L}(\mathcal{S}) - \phi_i(\mathcal{S})$, *where $(i + L) \leq |\mathcal{S}|$*

126   *Considering the drifts for every i^{th} location of the whole sequence, the sequence of drifts is*

127   *denoted by*

128     $\Delta^{(L)} = [\delta_0^{(L)}, \delta_1^{(L)}, \delta_2^{(L)}, \delta_3^{(L)}, \ldots, \delta_m^{(L)}]$, *where $(m + L) = |\mathcal{S}|$*

129   For GFP-RY (refer to Definition 1), an element (x_i, y_i) in $\Delta_{RY}^{(L)}$ provides excess numbers of G

130   from A and C from T in segment of length L starting from the ith location, respectively. Similarly,

131   in GFP-SW, they are the excess numbers of C from G and T from A (represents as $\Delta_{SW}^{(L)}$), and in

132   GFP-MK, they correspond to the excess numbers of A from C and T from G (represents as $\Delta_{MK}^{(L)}$),

133   respectively.

134   We also call the elements of $\Delta^{(L)}$ as points, as they can be plotted on a 2-D coordinate system.

135   We call this plot as the scatter plot of the drift sequence. Similarly, we get a scatter plot of a GFP.

136   It has been observed that in many cases the scatter plots of Δ have similar structure for closely

137   spaced species mentioned in literature. It can also be observed that differences between two GFPs

138   get reflected in their respective drifts.

139   We represent spatial distribution of these points of Δ by an elliptical model using a five dimen-

140   sional feature descriptor: $(\mu_x, \mu_y, \Lambda, \lambda, \theta)$, where (μ_x, μ_y) is the center of the coordinates, Λ and λ

6

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.

2.2 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,

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

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

3 Results

3.1 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 $5D$ vector as mentioned above and concatenation of $\mathcal{F}$ vectors provide a feature vector of dimension $5\mathcal{F}$.

Now, what is the optimal value of L and $\mathcal{F}$ which we should consider for the input dataset. For this, we compute the Shannon entropy (Shannon, 2001) 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

7

160 limit. This incident represents that by increasing the value of L, the number of distinct points of

161 the sequence of the drifts also increases. After a certain value of L, e.g, at $L = 550$, the entropy of

162 all mtDNA becomes stabilized (changes $< 1\%$). By increasing the value of L after that does not

163 change the entropy significantly. It is observed that for $L \geq 550$, the $\Delta^{(L)}$ contains significantly

164 large number of distinct point coordinates than that of $L < 550$. Hence, we choose the value of L

165 as 550.

166   For getting more precise signature of a drift sequence, we make 15 non-overlapping partitions of

167 the drift sequence. Hence, in this study, we consider $\mathcal{F}$ as 15. Each fragment is represented by the

168 $5D$ feature vector which consists of the mean of the point coordinates of the corresponding fragment,

169 eigen (both major and minor) values, and the angle formed by the eigen vector corresponding to

170 the major eigen value with respect to the x-axis. Hence, a complete drift is represented by 75

171 dimensional feature vector.

172   We compute the Graphical Foot Print (GFP or ϕ) of each mtDNA of human taken from different

173 parts of the world. For $L = 550$, we also compute the sequence of the drifts of the corresponding

174 GFP. It is observed that the species which are closely related, have similar patterns in both of their

175 GFP and drift sequence. Typical examples of drift sequences of humans from different regions of

176 the world and the outgroup species, such as baboon, gorilla, and chimpanzee are shown in Fig. 2.

3.2   Deriving trees

178 We propose a novel distance function to compute the pair-wise distances. Our proposed distance

179 function is a metric. We apply a hierarchical clustering method, Unweighted Pair Group Method

180 with Arithmetic mean (UPGMA) (Sneath and Sokal, 1973), to form the phylogenetic tree from the

181 distance matrix.

182   We apply this same technique on three different cases such as GFP_{RY}, GFP_{SW}, and GFP_{MK},

183 separately. Thus we derive three different phylogenetic trees by following three different hypotheses.

184 These trees are provided in the **Supplementary S2 Fig. 1-3 in Section 1**. There are various

185 methods to combine multiple trees and to form a supertree. We apply a quartet based method,

186 ASTRAL (Sayyari and Mirarab, 2016), to combine three trees. Please refer to Fig. 3 and also refer

187 to the **Supplementary S2 Fig. 4 in Section 1**.

3.3 Bootstrapping

The primary 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 sapiens* is within 1.2% (Castro et al., 1998). 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% (Jarczak et al., 2019). 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 (Felsenstein, 1985) 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 (Lemoine et al., 2018). So here we apply a modification of Felsenstein's bootstrapping, where the presence of a clade is quantified using the transfer distance proposed in (Lemoine et al., 2018). The transfer distance (Charon et al., 2006) is the minimum number of changes required to transform one partition to another. We compute the occurrence of each clade using the tool BOOSTER (Lemoine et al., 2018). 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**.

3.4 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**.

3.5 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) (Langergraber et al., 2012; Patterson et al., 2006; Reich, 2018). 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 groups 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 originated about 850 kya, while the African clade originated 650 kya. The derived trees also indicate that the "Out-of-Africa" migration appeared about 550 kya (Fig. 3). The trees with the edge lengths are provided in the **Supplementary S3**.

4 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. This method is effective in capturing broad distinct clades. This is evident from successful separation of the outgroup species from human in our study. However, this method identifies close relationship between chimpanzee and gorilla than human. It is also observed that the proposed distance function is more robust than the Euclidean distance function for the selected feature vector.

It is a well accepted hypothesis that the ancient human originated in Africa. The southern and eastern Africa is long been considered as the region where the human first originated (Behar et al., 2008; Chan et al., 2015, 2019). In recent years, the researchers found early *H. sapiens* fossils from Morocco which dates far earlier than the other fossils identified till date (Richter et al., 2017; Hublin et al., 2017). These findings raised the debate of the origin of human. Through the genetic study in very recent years, the researchers identified a common ancestor of the populations of the western parts of Africa (Durvasula and Sankararaman, 2020). This ancestor is mentioned as the "ghost archaic human" or "African Neanderthals" (Stringer and Galway-Witham, 2017; Durvasula and Sankararaman, 2020) and carries a significant distinct feature than the Neanderthals and Denisovans. It is also documented that $2-19\%$ of the west African DNA are inherited from the

10

244 ancestor, "African Neanderthals" (Durvasula and Sankararaman, 2020). There are some studies
245 which analyzed the genome-wide data and the artefacts, reported that the ancient humans were
246 more widespread across the globe (Callaway, 2018). Some impressions of the presence of ancient
247 humans earlier than the fossils record added that the ancient human were more widespread than we
248 think (Warren, 2019; Callaway, 2018). In the derived tree, the similar phenomena is also observed,
249 a major clade is formed from the populations of western and central parts of Africa. This clade
250 is placed at the most basal position of the phylogenetic tree with a very high bootstrap support
251 score (see Fig. 3). We estimated the time of origin of the ancestor of the clade as almost 600-700
252 kya which originated earlier than the other ancient humans, such as Neanderthals and Denisovans,
253 which is further concordance to the opinion of some recent studies (Steinrücken et al., 2018; Schaefer
254 et al., 2016). So this phenomena supports the early occurrence of our species, and this study also
255 supports that the western and central parts of Africa experienced the earliest occurrence of human.

256 All the derived trees show a very mixed representation of the humans from different communities.
257 Researchers found some signatures of the admixture with modern humans in as recent as 15-30
258 kya, such that almost $3 - 5\%$ of the Asian DNA are inherited from Denisovans (Gibbons, 2019)
259 and almost $1.5 - 2.1\%$ of the non-African genome are inherited from Neandarthals (Schaefer et al.,
260 2016). Though the bootstrap supports of all of these occurrences are lower than the accepted
261 significant level, but appearance of the same event in a repetitive manner shows a high possibility
262 of the admixture in modern humans.

263 There are still some small groups placed together in all the derived trees. It is also observed that
264 the Asian populations are closely related to the African population. There are some archaeological
265 studies that report the similarities in the ancient instruments in African and Asian population (Mel-
266 lars, 2006). Various studies proposed the close relation between north African and Levantine based
267 on the haplogroups of both mtDNA and Y-chromosomes (Olivieri et al., 2006; van de Loosdrecht
268 et al., 2018; Elkamel et al., 2018). It is also noticed that the eastern and the north-eastern Africans
269 are closely placed with both the Asian and the European populations. All the derived trees also
270 show the close relationships among the Pacific islanders (Oceania), such as Micronesia, Melanesia,
271 and Polynesia. It is also observed that DNA of these populations are also close to the DNA of the
272 Taiwanese. This phenomena also supports the existing studies where the close link between the
273 Pacific islanders and the Taiwanese is described (Trejaut, 2005; Friedlaender et al., 2008). Though
274 the bootstrap score is less than that of the accepted level, but its repeated occurrence in the derived

11

275 trees implies the higher chance of these relationships being true in the phylogenetic study.

276 All the observations of this study have been summarized in Table 1.

5 Conclusion

278 In this study we reported a study of human origin based on the graphical representation where
279 we inspect the origin and the migration of human based on a novel feature vector. The proposed
280 method may be extended to study evolution of other types of genomic sequences. This method
281 captures some unique signature of the whole mtDNA. It is observed that this method efficiently
282 captures the broad relationships rather than the precise relationships near the leaf nodes. In this
283 study, we present such broad divergence among the human population. Our result supports some
284 existing hypotheses like the Africa witnessed the first human in the world and the close relationship
285 among the Eurasian and the African population. The origin of the African clade in our study is
286 also close to the group of archaeological reports. Apart from that, this study provides an evidence
287 of the presence of the "ghost ancestor" of the population of the west and central Africa earlier than
288 that of the fossils identified. This phenomena also demonstrates that the west and central Africa
289 are very important for the early evolution of human.

Code availability

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

Data availability

297 All the data are publicly available in the online database, NCBI, with their accession number.
298 The list of accession numbers we consider in this study are provided in **Supplementary S1**. The
299 human mitochondrial sequences are further annotated and stored in public database server, mtDB.

Ethics declarations

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Approval

The ethical approval from the institute is not applicable for this study.

Funding

Informed Consent

All the data and personal information are collected from a public database. These data have been generated by various researchers since a decade.

References

Achilli, A., Rengo, C., Battaglia, V., Pala, M., Olivieri, A., Fornarino, S., Magri, C., Scozzari, R., Babudri, N., Santachiara-Benerecetti, A. S., Bandelt, H.-J., Semino, O., and Torroni, A. (2005). Saami and Berbers-An Unexpected Mitochondrial DNA Link. *The American Journal of Human Genetics*, 76(5):883–886.

Andrews, P. and Johnson, R. J. (2019). Evolutionary basis for the human diet: consequences for human health. *Journal of Internal Medicine*.

Bae, C. J., Douka, K., and Petraglia, M. D. (2017). On the origin of modern humans: Asian perspectives. *Science*, 358(6368):eaai9067.

Baker, J. L., Rotimi, C. N., and Shriner, D. (2017). Human ancestry correlates with language and reveals that race is not an objective genomic classifier. *Scientific Reports*, 7(1):1572.

Balter, M. (2019). Language May Have Helped Early Humans Spread Out of Africa. *Science*.

Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., Dey, B., Roy, M., Roy, B., Bhattacharyya, N. P., et al. (2003). Ethnic India: A Genomic View, With Special Reference to Peopling and Structure. *Genome Research*, 13(10):2277–2290.

Behar, D. M., Metspalu, E., Kivisild, T., Achilli, A., Hadid, Y., Tzur, S., Pereira, L., Amorim, A., Quintana-Murci, L., Majamaa, K., Herrnstadt, C., Howell, N., Balanovsky, O., Kutuev, I., Pshenichnov, A., Gurwitz, D., Bonne-Tamir, B., Torroni, A., Villems, R., and Skorecki, K. (2006). The Matrilineal Ancestry of Ashkenazi Jewry: Portrait of a Recent Founder Event. *The American Journal of Human Genetics*, 78(3):487–497.

Behar, D. M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., Comas, D., et al. (2008). The dawn of human matrilineal diversity. *The American Journal of Human Genetics*, 82(5):1130–1140.

Bhugra, D. and Becker, M. A. (2005). Migration, cultural bereavement and cultural identity. *World Psychiatry*, 4(1):18.

13

Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S., and Akey, J. M. (2018). Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell*, 173(1):53–61.

Callaway, E. (2018). Divided by DNA: The uneasy relationship between archaeology and ancient genomics. *Nature*, 555(7698).

Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, 325(6099):31.

Cartmill, M. and Smith, F. H. (2009). *The Human Lineage*. Wiley-Blackwell.

Castro, J. A., Picornell, A., and Ramon, M. (1998). Mitochondrial DNA: a tool for populational genetics studies. *International Microbiology*, 1(4):327–332.

Chan, E. K., Hardie, R.-A., Petersen, D. C., Beeson, K., Bornman, R. M., Smith, A. B., and Hayes, V. M. (2015). Revised Timeline and Distribution of the Earliest Diverged Human Maternal Lineages in Southern Africa. *PLoS One*, 10(3).

Chan, E. K., Timmermann, A., Baldi, B. F., Moore, A. E., Lyons, R. J., Lee, S.-S., Kalsbeek, A. M., Petersen, D. C., Rautenbach, H., Förtsch, H. E., et al. (2019). Human origins in a southern African palaeo-wetland and first migrations. *Nature*, 575(7781):185–189.

Charon, I., Denoeud, L., Guénoche, A., and Hudry, O. (2006). Maximum transfer distance between partitions. *Journal of Classification*, 23(1):103–121.

Duda, P. and Zrzavỳ, J. (2016). Human population history revealedby a supertree approach. *Scientific Reports*, 6:29890.

Durvasula, A. and Sankararaman, S. (2020). Recovering signals of ghost archaic introgression in african populations. *Science Advances*, 6(7).

Elkamel, S., Boussetta, S., Khodjet-El-Khil, H., Benammar Elgaaied, A., and Cherni, L. (2018). 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(3):e23100.

Eyre-Walker, A. and Awadalla, P. (2001). Does human mtdna recombine? *Journal of Molecular Evolution*, 53(4):430–435.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, 39(4):783–791.

Friedlaender, J., Schurr, T., Gentz, F., Koki, G., Friedlaender, F., Horvat, G., Babb, P., Cerchio, S., Kaestle, F., Schanfield, M., Deka, R., Yanagihara, R., and Merriwether, D. A. (2005). Expanding Southwest Pacific Mitochondrial Haplogroups P and Q. *Molecular Biology and Evolution*, 22(6):1506–1517.

Friedlaender, J. S., Friedlaender, F. R., Reed, F. A., Kidd, K. K., Kidd, J. R., Chambers, G. K., Lea, R. A., Loo, J.-H., Koki, G., Hodgson, J. A., Merriwether, D. A., and Weber, J. L. (2008). The genetic structure of pacific islanders. *PLOS Genetics*, 4(1):1–18.

Gibbons, A. (2019). Our mysterious cousinsthe Denisovansmay have mated with modern humans as recently as 15,000 years ago. *Science*.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., et al. (2010). A Draft Sequence of the Neandertal Genome. *Science*, 328(5979):710–722.

Hill, C., Soares, P., Mormina, M., Macaulay, V., Meehan, W., Blackburn, J., Clarke, D., Raja, J. M., Ismail, P., Bulbeck, D., Oppenheimer, S., and Richards, M. (2006). Phylogeography and Ethnogenesis of Aboriginal Southeast Asians. *Molecular Biology and Evolution*, 23(12):2480–2491.

Hublin, J.-J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., Bergmann, I., Le Cabec, A., Benazzi, S., Harvati, K., and Gunz, P. (2017). New fossils from Jebel Irhoud, Morocco and the pan-African origin of *Homo sapiens*. *Nature*, 546(7657):289–292.

Hurles, M. E. and Jobling, M. A. (2001). Haploid chromosomes in molecular ecology: lessons from the human Y. *Molecular Ecology*, 10(7):1599–1613.

Ingman, M. and Gyllensten, U. (2003). Mitochondrial Genome Variation and Evolutionary History of Australian and New Guinean Aborigines. *Genome Research*, 13(7):1600–1606.

Ingman, M. and Gyllensten, U. (2006). mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. *Nucleic Acids Research*, 34(suppl_1):D749–D751.

Ingman, M. and Gyllensten, U. (2007). A recent genetic link between Sami and the Volga-Ural region of Russia. *European Journal of Human Genetics*, 15(1):115.

Ingman, M., Kaessmann, H., Pääbo, S., and Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature*, 408(6813):708.

Jarczak, J., Grochowalski, Ł., Marciniak, B., Lach, J., Słomka, M., Sobalska-Kwapis, M., Lorkiewicz, W., Pułaski, Ł., and Strapagiel, D. (2019). Mitochondrial DNA variability of the Polish population. *European Journal of Human Genetics*, page 1.

Jinam, T. A., Phipps, M. E., Aghakhanian, F., Majumder, P. P., Datar, F., Stoneking, M., Sawai, H., Nishida, N., Tokunaga, K., Kawamura, S., et al. (2017). Discerning the Origins of the Negritos, First Sundaland People: Deep Divergence and Archaic Admixture. *Genome Biology and Evolution*, 9(8):2013–2022.

Juyal, G., Mondal, M., Luisi, P., Laayouni, H., Sood, A., Midha, V., Heutink, P., Bertranpetit, J., Thelma, B., and Casals, F. (2014). Population and genomic lessons from genetic analysis of two Indian populations. *Human Genetics*, 133(10):1273–1287.

Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., Inoue, E., Inoue-Muruyama, M., Mitani, J. C., Muller, M. N., Robbins, M. M., Schubert, G., Stoinski, T. S., Viola, B., Watts, D., Wittig, R. M., Wrangham, R. W., Zuberbühler, K., Pääbo, S., and Vigilant, L. (2012). 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(39):15716–15721.

Langille, M. G. I., Hsiao, W. W. L., and Brinkman, F. S. L. (2010). Detecting genomic islands using bioinformatics approaches. *Nature Reviews Microbiology*, 8(5):373–382.

Lemoine, F., Entfellner, J.-B. D., Wilkinson, E., Correia, D., Felipe, M. D., Oliveira, T., and Gascuel, O. (2018). Renewing Felsensteins phylogenetic bootstrap in the era of big data. *Nature*, page 1.

Liu, H., Prugnolle, F., Manica, A., and Balloux, F. (2006). A Geographically Explicit Genetic Model of Worldwide Human-Settlement History. *The American Journal of Human Genetics*, 79(2):230–237.

Maca-Meyer, N., González, A. M., Larruga, J. M., Flores, C., and Cabrera, V. M. (2001). Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics*, 2(1):13.

Mafessoni, F. (2019). Encounters with archaic hominins. *Nature Ecology & Evolution*, 3(1):14–15.

Mahapatra, A. and Mukherjee, J. (2019). GRaphical footprint based Alignment-Free method (GRAFree) for classifying the species in Large-Scale Genomics. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 105–112. Springer.

Mellars, P. (2006). Going East: New Genetic and Archaeological Perspectives on the Modern Human Colonization of Eurasia. *Science*, 313(5788):796–800.

Mendez, F. L., Watkins, J. C., and Hammer, M. F. (2012). Global genetic variation at OAS1 provides evidence of archaic admixture in Melanesian populations. *Molecular Biology and Evolution*, 29(6):1513–1520.

Mesoudi, A. (2018). Migration, acculturation, and the maintenance of between-group cultural variation. *PLOS ONE*, 13(10):e0205573.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30:i541–i548.

Mondal, M., Bertranpetit, J., and Lao, O. (2019). Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications*, 10(1):246.

Mondal, M., Casals, F., Xu, T., Dall'Olio, G. M., Pybus, M., Netea, M. G., Comas, D., Laayouni, H., Li, Q., Majumder, P. P., et al. (2016). Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nature Genetics*, 48(9):1066.

Moorjani, P., Thangaraj, K., Patterson, N., Lipson, M., Loh, P.-R., Govindaraj, P., Berger, B., Reich, D., and Singh, L. (2013). Genetic Evidence for Recent Population Mixture in India. *The American Journal of Human Genetics*, 93(3):422–438.

Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N., Scozzari, R., Cruciani, F., Behar, D. M., Dugoujon, J.-M., Coudray, C., Santachiara-Benerecetti, A. S., Semino, O., Bandelt, H.-J., and Torroni, A. (2006). The mtDNA Legacy of the Levantine Early Upper Palaeolithic in Africa. *Science*, 314(5806):1767–1770.

Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., and Reich, D. (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097):1103.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., De Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J. C., Vohr, S. H., Green, R. E., Hellmann, I., Johnson, P. L. F., Blanche, H., Cann, H., Kitzman, J. O., Shendure, J., Eichler, E. E., Lein, E. S., Bakken, T. E., Golovanova, L. V., Doronichev, V. B., Shunkov, M. V., Derevianko, A. P., Viola, B., Slatkin, M., Reich, D., Kelso, J., and Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43.

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

Richter, D., Grün, R., Joannes-Boyau, R., Steele, T. E., Amani, F., Rué, M., Fernandes, P., Raynal, J.-P., Geraads, D., Ben-Ncer, A., Hublin, J.-J., and McPherron, S. P. (2017). The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. *Nature*, 546(7657):293–296.

Roostalu, U., Kutuev, I., Loogväli, E., Metspalu, E., Tambets, K., Reidla, M., Khusnutdinova, E., Usanga, E., Kivisild, T., and Villems, R. (2006). 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(2):436–448.

Sahle, Y., Reyes-Centeno, H., and Bentz, C. (2018). Modern human origins and dispersal: current state of knowledge and future directions. *Evolutionary Anthropology: Issues, News, and Reviews*, 27(2):64–67.

Sayyari, E. and Mirarab, S. (2016). Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular Biology and Evolution*, 33(7):1654–1668.

Schaefer, N. K., Shapiro, B., and Green, R. E. (2016). Detecting hybridization using ancient DNA. *Molecular Ecology*, 25(11):2398–2412.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

Shriner, D., Tekola-Ayele, F., Adeyemo, A., and Rotimi, C. N. (2016). Ancient Human Migration after Out-of-Africa. *Scientific Reports*, 6(26565).

Smith, F. H. and Ahern, J. C. (2013). *The origins of modern humans: biology reconsidered.* John Wiley & Sons.

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

Starikovskaya, E. B., Sukernik, R. I., Derbeneva, O. A., Volodko, N. V., Ruiz-Pesini, E., Torroni, A., Brown, M. D., Lott, M. T., Hosseini, S. H., Huoponen, K., and Wallace, D. C. (2005). Mitochondrial DNA Diversity in Indigenous Populations of the Southern Extent of Siberia, and the Origins of Native American Haplogroups. *Annals of Human Genetics*, 69(1):67–89.

Steinrücken, M., Spence, J. P., Kamm, J. A., Wieczorek, E., and Song, Y. S. (2018). Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. *Molecular Ecology*, 27(19):3873–3888.

Stringer, C. (2003). Out of Ethiopia. *Nature*, 423(6941):692.

Stringer, C. and Galway-Witham, J. (2017). Palaeoanthropology: On the origin of our species. *Nature*, 546(7657):212–214.

Töpf, A. L., Gilbert, M. T. P., Dumbacher, J. P., and Hoelzel, A. R. (2005). Tracing the Phylogeography of Human Populations in Britain Based on 4th11th Century mtDNA Genotypes. *Molecular Biology and Evolution*, 23(1):152–161.

Trejaut, J. (2005). Mitochondrial dna provides a link between polynesians and indigenous taiwanese. *PLOS Biology*, 3(8).

van de Loosdrecht, M., Bouzouggar, A., Humphrey, L., Posth, C., Barton, N., Aximu-Petri, A., Nickel, B., Nagel, S., Talbi, E. H., El Hajraoui, M. A., Amzazi, S., Hublin, J.-J., Pääbo, S., Schiffels, S., Meyer, M., Haak, W., Jeong, C., and Krause, J. (2018). Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. *Science*, 360(6388):548–552.

17

Veeramah, K. R. and Hammer, M. F. (2014). The impact of whole-genome sequencing on the reconstruction of human population history. *Nature Reviews Genetics*, 15(3):149–162.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., and Wilson, A. C. (1991). African populations and the evolution of human mitochondrial DNA. *Science*, 253(5027):1503–1507.

Warren, M. (2019). Biggest Denisovan fossil yet spills ancient humans secrets. *Nature*, 569(7754):16–17.

Wolpoff, M. H., Hawks, J., and Caspari, R. (2000). Multiregional, not multiple origins. *American Journal of Physical Anthropology*, 112(1):129–136.

18

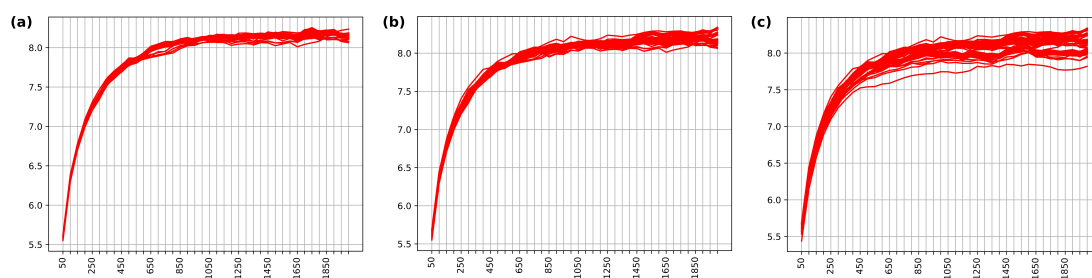Figure 1: Shannon entropy of (a) GFP_{RY}, (b) GFP_{SW}, and (c) GFP_{MK} of different human mtDNA for all the values of L from 50 to 2000.
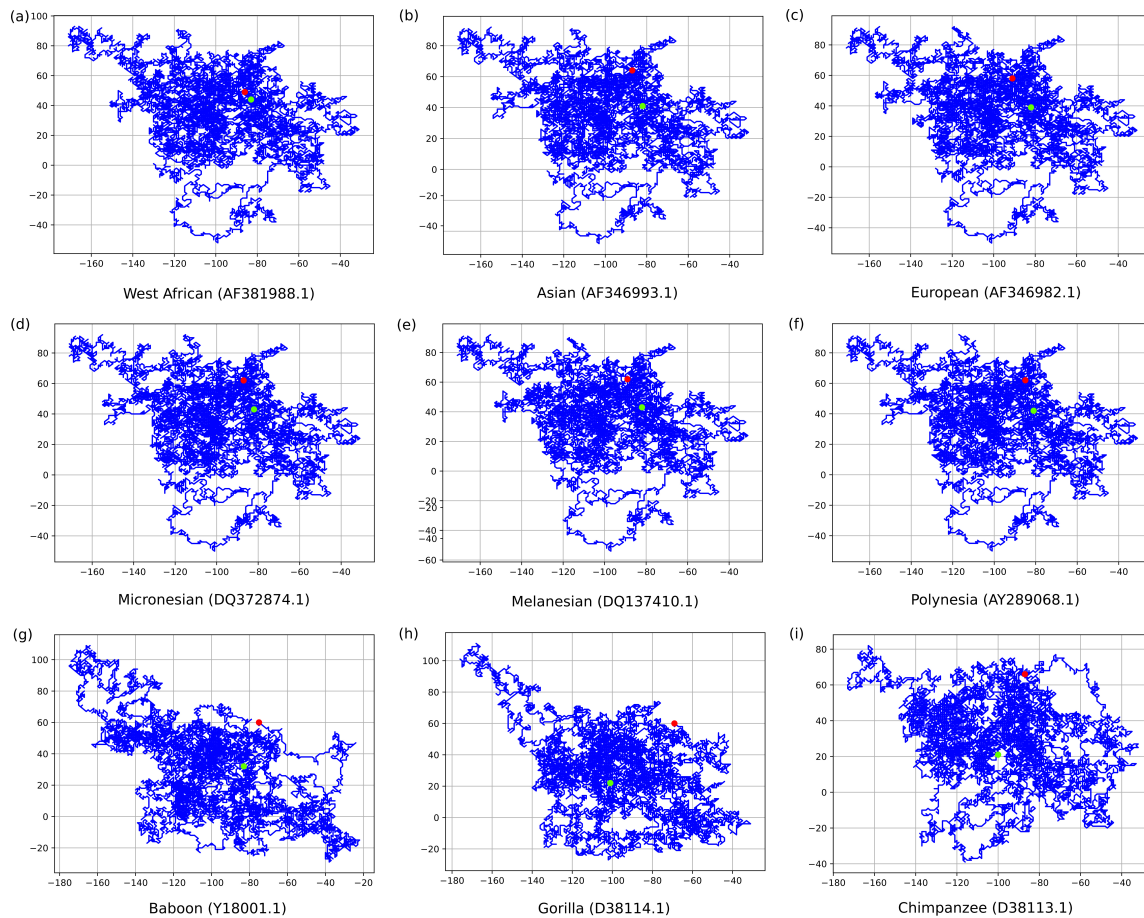
Figure 2: The sequence of drifts of different sample, mentioned under each graph, for the case when we consider the Purine and Pyrimidine as $2D$ coordinates space. The accession numbers of the corresponding samples are mentioned within the bracket. The start and end positions of the sequence of drifts 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 population of the western Africa (a) are slightly different from the drift sequences of the other humans (b-f). The differences are visible to the naked eye at the starting (green dot) and end points (red dot) of the drift sequences.
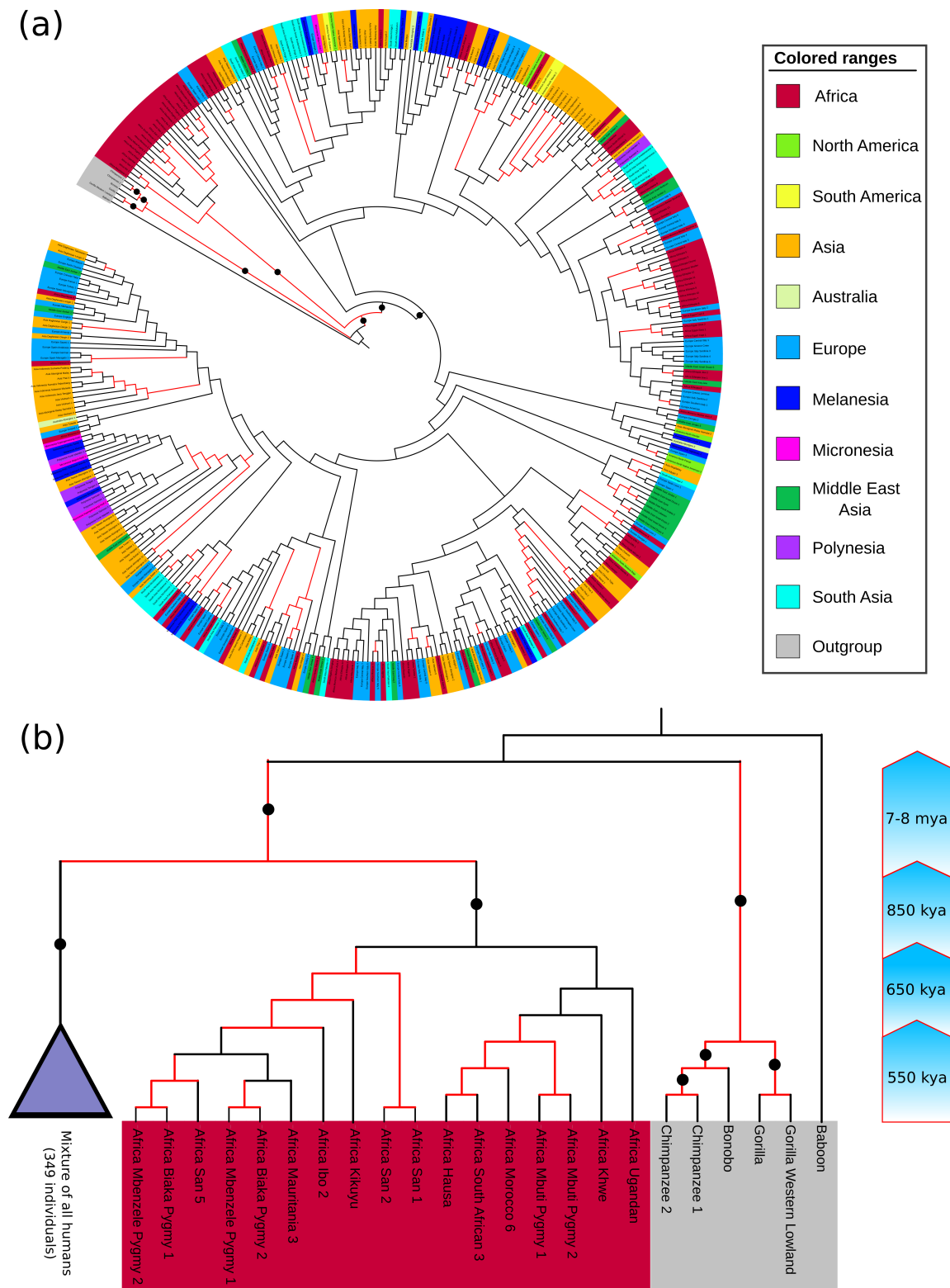
Figure 3: The derived trees using GRAFree. The black dots on the edges denote a high bootstrap support ($\geq$ 75%) of the corresponding clade. The edge marked as red color denotes that the robustness of the clade with respect to the hyper parameters is greater than 80%. (a) Tree derived after merging all three trees using ASTRAL. (b) The summarization of the tree having bootstrap support $\geq$ 75%.

Table 1: Summary of the observations

	Observations	Bootstrap support	Robustness	Conclusion
African population	• There is a common ancestor of the African population from western and central parts. • The origin of this ancestor is estimated to be much earlier than the fossils identified till date.	Very high	Very high	• Supports the presence of "African Neanderthals". • Origin of human was in the west and central parts of Africa.
Out of Africa migration	• Out-of-Africa migration occurred almost 550 kya.	Very high	Very high	• Humans were widespread than that of reported till date.
World populations	• Very mixed representation of humans from different community.	Low	Very high	• A very high admixture within different community of the world.
Small communities	• Asian populations are closely related to the African population. • The eastern and the north-eastern Africans are closely placed with both the Asian and the European populations. • The Pacific islanders (Oceania), such as Micronesia, Melanesia, and Polynesia are closely related.	Low	Very high	

22