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

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. The debate for determining those regions of Africa which witnessed the first origin of humans still exists. 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 are more consistent with the archaeological findings based on the time of origin of different communities. We find that west and central African communities are placed at the basal point with a very high bootstrap score. This study roughly estimates the existence of the archaic human at 800-900 kilo years ago and presence of human in Africa at 600-700 kilo years ago which supports the presence of an ancestor of the west and central African earlier than that of the fossils identified.

1 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 human evolution and migration. The "recent African origin" or "replacement" hypothesis [5, 6, 7] states that there exist the anatomically modern humans which subsequently replace the ancient humans and spread all over the world [8, 9, 10]. There are various methods incorporated to describe the history of the humans. Several genomic studies [11, 12, 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 negated 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], and 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, 22, 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]. As a result many archaeologists think the ancient DNA as the "devil's work" [26]. Though the origin of the human in the southern part of Africa has long been accepted [27, 28, 5], a few recent studies documented the presence of humans in the western part of Africa much earlier than the other fossils identified till date [29, 30]. In this work, we consider the mitochondrial genome (mtDNA) for studying the origin of humans. The mtDNA is haploid, inherited maternally [31], and recombination is very rare event in it [32]. 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. Moreover, this study also supports the recently published studies on the 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 [33]. 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 [34]. 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 [35], 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) [36].

We apply this method on a large set of mitochondrial DNA sequences of humans. In this study, we consider 369 mitochondrial genomes from more than 120 countries covering six continents (except Antarctica). The data are sequenced and published by various researchers [37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47]. They are accumulated and stored in a database server named *mtDB* [48]. 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. We found that a significantly large populations from the western and central parts of Africa are placed at a distinct clade than the rest of the humans with a very high bootstrap support indicating a divergence of the genomic signatures of the western and central African populations from the rest of the humans. This phenomena also indicates an early migration of human from these parts of Africa to the world. This period is roughly estimated at 800-900 kilo years ago (kya). 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 [49]. 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 [49, 50, 51].

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 [33]. 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[52, 36].

2.1 Feature space

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

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,

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

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

We compute the Shannon entropy [53] 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 becomes stabilized (changes $< 1\%$). By increasing the value of L after that does not change the entropy significantly. 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.

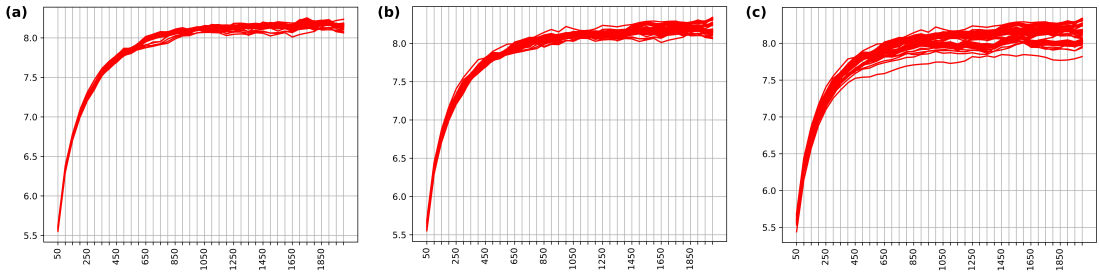

Figure 1: Shannon entropy of (a) GFP_{RY}, (b) GFP_{SW}, and (c) GFP_{MK} of different human mtDNA for all the value of L from 50 to 2000.

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.

3.2 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,

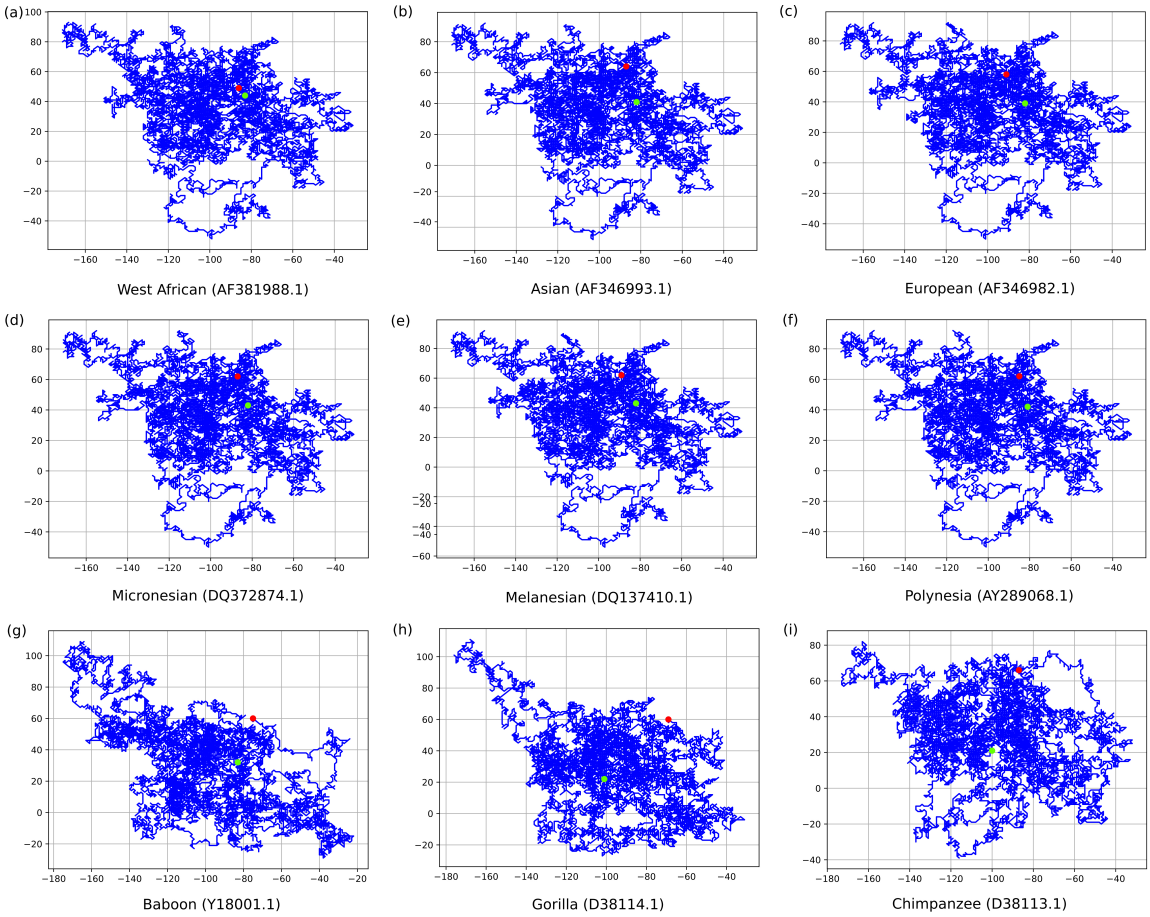

Figure 2: The sequence of drifts of different sample, mentioned under each graph, for the case when we consider the Purine and Pyrimidine as $2D$ coordinates space. The accession numbers of the corresponding samples are mentioned within the bracket. The start and end positions of the sequence of drifts 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 can be visible in naked eye at the starting (green dot) and end points (red dot) of the drift sequences.

Unweighted Pair Group Method with Arithmetic mean (UPGMA) [35], 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 provided in the **Supplementary S2 Fig. 1-3 in Section 1**. There are various methods to combine multiple trees and to form a supertree. We apply a quartet based method, ASTRAL [36], to combine three trees. Please refer to Fig. 3 and also refer 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% [54]. 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% [55]. 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 [56] 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 [57]. So here we apply a modification of Felsenstein's bootstrapping, where the presence of a clade is quantified using the transfer distance proposed in [57]. The transfer distance [58] is the minimum number of changes required to transform one partition to other. We compute the occurrence of each clade using the tool BOOSTER [57]. 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**.

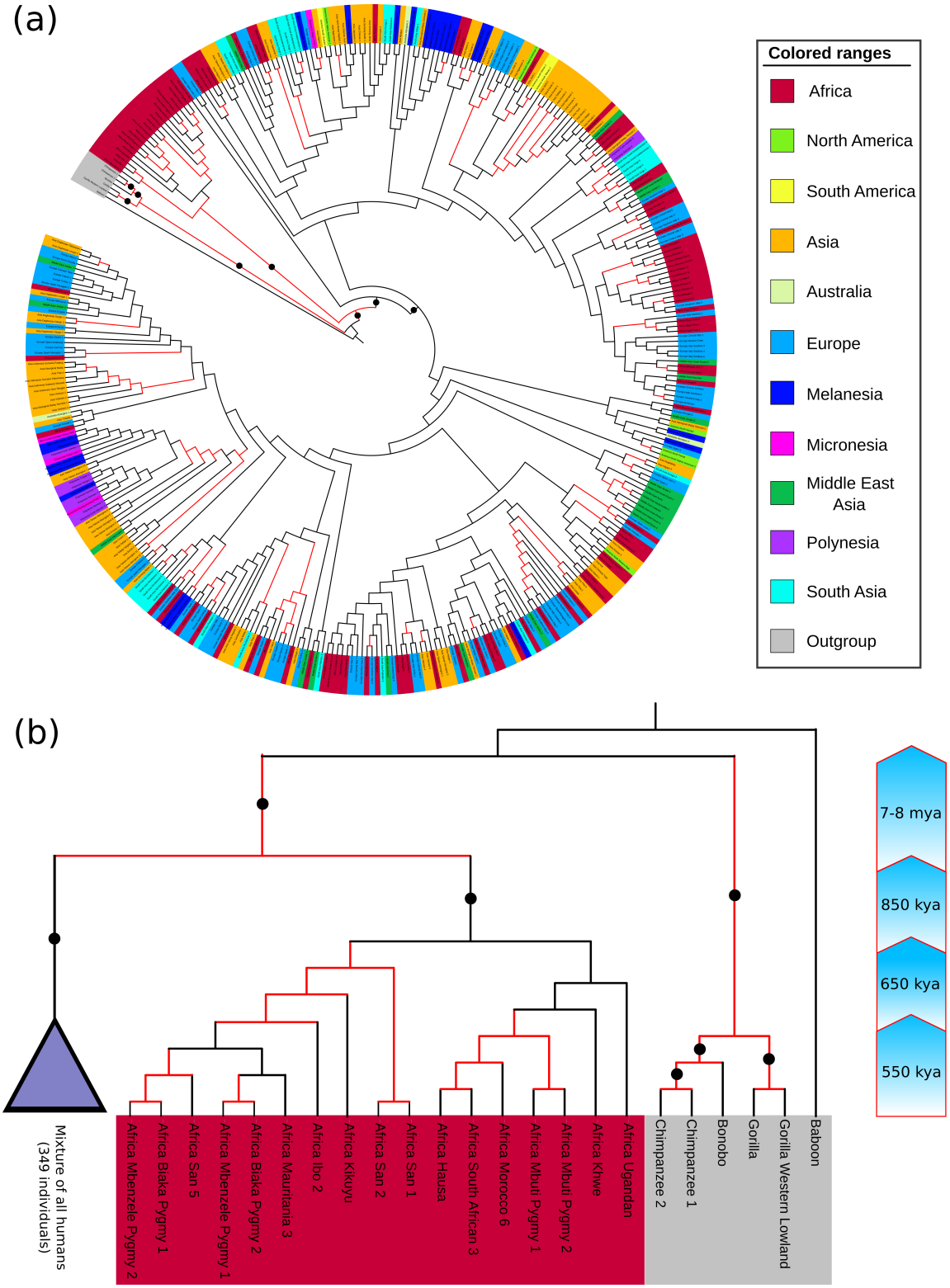

Figure 3: The derived trees using GRAFree. The black dots on the edges denote a high bootstrap support ($\geq 75\%$) of the corresponding clade. The edge marked as red color denotes that the robustness of the clade with respect to the hyper parameters is greater than 80%. (a) Tree derived after merging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

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) [59, 60, 61]. 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 was originated about 850 kya, while the African clade was 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. Though this method does not resolve phylogenetic relations of very closely related groups, it 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 was originated in Africa. The southern Africa is long been considered as the region where the human was originated [27, 28, 5]. In recent years, the researchers found early *H. sapiens* fossils from Morocco which dates far earlier than the other fossils identified till date [62, 63]. 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 the Africa [30]. This ancestor is mentioned as the "ghost archaic human" or "African Neanderthals" [29, 30] and carries a significant distinct features than the Neanderthals and Denisovans. It is also documented that $2-19\%$ of the west African DNA are inherited from the ancestor, "African Neanderthals" [30]. 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 ancient humans earlier than the fossils record added that the ancient human were more widespread than we think [64, 26]. In the derived tree, the similar phenomena is also observed, a major clade is formed from the populations of western and central parts of the Africa. This clade is placed at the most basal position of the phylogenetic tree with a very high bootstrap support score (see Fig. 3). We estimated the time of origin of the ancestor of the clade is almost 650 kya which originated earlier than the other ancient humans, such as Neanderthals and Denisovans. So this

phenomena supports the early occurrence of our species, and this study also supports that the western and central parts of the Africa experienced the earliest occurrence of human.

All the derived trees show a very mixed representation of the humans from different communities. 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 [65]. 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 a high possibility of the admixture in modern humans.

There are still some small groups placed together in all the derived trees. It is also observed that the Asian populations 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]. Various studies proposed the close relation between north African and Levantine based on the haplogroups of both mtDNA and Y-chromosomes [42, 66, 67]. It is also noticed that the eastern and the north-eastern Africans are closely placed with both the Asian and the European populations. All the derived trees also show the close relationships among the Pacific islanders (Oceania), such as Micronesia, Melanesia, and Polynesia. It is also observed that DNA of these populations are also close to the DNA of the Taiwanese. This phenomena also supports the existing studies where the close link between the Pacific islanders and the Taiwanese is described [68, 69]. Though the bootstrap score is less than that of the accepted level, but its repeated occurrence in the derived trees implies the higher chance of these relationships being true in the phylogenetic study.

All the observations of this study is summarized in Table 1.

5 Conclusion

In this study we reported a 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. The proposed method may be extended to study evolution of 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 an evidences of the presence of the "ghost ancestor" of the population of the west and central Africa earlier than that of the fossils identified. This phenomena also demonstrates that the west and central Africa have been important in the early evolution of human.

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 which 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 northeastern 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	

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 S1**. The 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.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

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

[1] Pavel Duda and Jan Zrzavỳ. Human population history revealedby a supertree approach. *Scientific Reports*, 6:29890, 2016.

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

[3] Daniel Shriner, Fasil Tekola-Ayele, Adebowale Adeyemo, and Charles N. Rotimi. Ancient Human Migration after Out-of-Africa. *Scientific Reports*, 6(26565), 2016.

[4] Christopher J Bae, Katerina Douka, and Michael D Petraglia. On the origin of modern humans: Asian perspectives. *Science*, 358(6368):eaai9067, 2017.

[5] Eva KF Chan, Axel Timmermann, Benedetta F Baldi, Andy E Moore, Ruth J Lyons, Sun-Seon Lee, Anton MF Kalsbeek, Desiree C Petersen, Hannes Rautenbach, Hagen EA Förtsch, et al. Human origins in a southern African palaeo-wetland and first migrations. *Nature*, 575(7781):185–189, 2019.

[6] Rebecca L Cann, Mark Stoneking, and Allan C Wilson. Mitochondrial DNA and human evolution. *Nature*, 325(6099):31, 1987.

[7] Linda Vigilant, Mark Stoneking, Henry Harpending, Kristen Hawkes, and Allan C Wilson. African populations and the evolution of human mitochondrial DNA. *Science*, 253(5027):1503–1507, 1991.

[8] Hua Liu, Franck Prugnolle, Andrea Manica, and François Balloux. A Geographically Explicit Genetic Model of Worldwide Human-Settlement History. *The American Journal of Human Genetics*, 79(2):230–237, 2006.

[9] Chris Stringer. Out of Ethiopia. *Nature*, 423(6941):692, 2003.

[10] Milford H. Wolpoff, John Hawks, and Rachel Caspari. Multiregional, not multiple origins. *American Journal of Physical Anthropology*, 112(1):129–136, 2000.

[11] Mayukh Mondal, Jaume Bertranpetit, and Oscar Lao. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications*, 10(1):246, 2019.

[12] Sharon R Browning, Brian L Browning, Ying Zhou, Serena Tucci, and Joshua M Akey. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell*, 173(1):53–61, 2018.

[13] Mayukh Mondal, Ferran Casals, Tina Xu, Giovanni M Dall'Olio, Marc Pybus, Mihai G Netea, David Comas, Hafid Laayouni, Qibin Li, Partha P Majumder, et al. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nature Genetics*, 48(9):1066, 2016.

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

[15] Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H Sudmant, Cesare De Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, Christoph Theunert, Arti Tandon, Priya Moorjani, Joseph Pickrell, James C. Mullikin, Samuel H. Vohr, Richard E. Green, Ines Hellmann, Philip L. F. Johnson, Hélène Blanche, Howard Cann, Jacob O. Kitzman, Jay Shendure, Evan E. Eichler, Ed S. Lein, Trygve E. Bakken, Liubov V. Golovanova, Vladimir B. Doronichev, Michael V. Shunkov, Anatoli P. Derevianko, Bence Viola, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Pääbo. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43, 2014.

[16] Matt Cartmill and Fred H. Smith. *The Human Lineage*. Wiley-Blackwell, 2009.

[17] Jennifer L Baker, Charles N Rotimi, and Daniel Shriner. Human ancestry correlates with language and reveals that race is not an objective genomic classifier. *Scientific Reports*, 7(1):1572, 2017.

[18] Michael Balter. Language May Have Helped Early Humans Spread Out of Africa. *Science*, 2019.

[19] Dinesh Bhugra and Matthew A Becker. Migration, cultural bereavement and cultural identity. *World Psychiatry*, 4(1):18, 2005.

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

[21] Timothy A Jinam, Maude E Phipps, Farhang Aghakhanian, Partha P Majumder, Francisco Datar, Mark Stoneking, Hiromi Sawai, Nao Nishida, Katsushi Tokunaga, Shoji Kawamura, et al. Discerning the Origins of the Negritos, First Sundaland People: Deep Divergence and Archaic Admixture. *Genome Biology and Evolution*, 9(8):2013–2022, 2017.

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

[23] Analabha Basu, Namita Mukherjee, Sangita Roy, Sanghamitra Sengupta, Sanat Banerjee, Madan Chakraborty, Badal Dey, Monami Roy, Bidyut Roy, Nitai P Bhattacharyya, et al. Ethnic India: A Genomic View, With Special Reference to Peopling and Structure. *Genome Research*, 13(10):2277–2290, 2003.

[24] Garima Juyal, Mayukh Mondal, Pierre Luisi, Hafid Laayouni, Ajit Sood, Vandana Midha, Peter Heutink, Jaume Bertranpetit, BK Thelma, and Ferran Casals. Population and genomic lessons from genetic analysis of two Indian populations. *Human Genetics*, 133(10):1273–1287, 2014.

[25] Priya Moorjani, Kumarasamy Thangaraj, Nick Patterson, Mark Lipson, Po-Ru Loh, Periyasamy Govindaraj, Bonnie Berger, David Reich, and Lalji Singh. Genetic Evidence for Recent Population Mixture in India. *The American Journal of Human Genetics*, 93(3):422–438, 2013.

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

[27] Doron M Behar, Richard Villems, Himla Soodyall, Jason Blue-Smith, Luisa Pereira, Ene Metspalu, Rosaria Scozzari, Heeran Makkan, Shay Tzur, David Comas, et al. The dawn of human matrilineal diversity. *The American Journal of Human Genetics*, 82(5):1130–1140, 2008.

[28] Eva KF Chan, Rae-Anne Hardie, Desiree C Petersen, Karen Beeson, Riana MS Bornman, Andrew B Smith, and Vanessa M Hayes. Revised Timeline and Distribution of the Earliest Diverged Human Maternal Lineages in Southern Africa. *PLoS One*, 10(3), 2015.

[29] Chris Stringer and Julia Galway-Witham. Palaeoanthropology: On the origin of our species. *Nature*, 546(7657):212–214, 2017.

[30] Arun Durvasula and Sriram Sankararaman. Recovering signals of ghost archaic introgression in african populations. *Science Advances*, 6(7), 2020.

[31] James M. Cummins, Teruhiko Wakayama, and Ryuzo Yanagimachi. Fate of microinjected spermatid mitochondria in the mouse oocyte and embryo. *Zygote*, 5(4):301308, 1997.

[32] Adam Eyre-Walker and Philip Awadalla. Does human mtdna recombine? *Journal of Molecular Evolution*, 53(4):430–435, 2001.

[33] Aritra Mahapatra and Jayanta Mukherjee. 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, 2019.

[34] Morgan G. I. Langille, William W. L. Hsiao, and Fiona S. L. Brinkman. Detecting genomic islands using bioinformatics approaches. *Nature Reviews Microbiology*, 8(5):373–382, 2010.

[35] Peter H. A. Sneath and Robert R. Sokal. *Numerical Taxonomy*. W. H. Freeman and Company, San Francisco, 1973.

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

[37] Jonathan Friedlaender, Theodore Schurr, Fred Gentz, George Koki, Francoise Friedlaender, Gisele Horvat, Paul Babb, Sal Cerchio, Frederika Kaestle, Moses Schanfield, Ranjan Deka, Ric Yanagihara, and D. Andrew Merriwether. Expanding Southwest Pacific Mitochondrial Haplogroups P and Q. *Molecular Biology and Evolution*, 22(6):1506–1517, 2005.

[38] Max Ingman, Henrik Kaessmann, Svante Pääbo, and Ulf Gyllensten. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408(6813):708, 2000.

[39] U Roostalu, I Kutuev, EL Loogväli, E Metspalu, K Tambets, M Reidla, EK Khusnutdinova, E Usanga, T Kivisild, and R Villems. 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, 2006.

[40] Catherine Hill, Pedro Soares, Maru Mormina, Vincent Macaulay, William Meehan, James Blackburn, Douglas Clarke, Joseph Maripa Raja, Patimah Ismail, David Bulbeck, Stephen Oppenheimer, and Martin Richards. Phylogeography and Ethnogenesis of Aboriginal Southeast Asians. *Molecular Biology and Evolution*, 23(12):2480–2491, 2006.

[41] Alessandro Achilli, Chiara Rengo, Vincenza Battaglia, Maria Pala, Anna Olivieri, Simona Fornarino, Chiara Magri, Rosaria Scozzari, Nora Babudri, A Silvana Santachiara-Benerecetti, Hans-Jrgen Bandelt, Ornella Semino, and Antonio Torroni. Saami and Berbers-An Unexpected Mitochondrial DNA Link. *The American Journal of Human Genetics*, 76(5):883–886, 2005.

[42] Anna Olivieri, Alessandro Achilli, Maria Pala, Vincenza Battaglia, Simona Fornarino, Nadia Al-Zahery, Rosaria Scozzari, Fulvio Cruciani, Doron M. Behar, Jean-Michel Dugoujon, Clotilde Coudray, A. Silvana Santachiara-Benerecetti, Ornella Semino, Hans-Jürgen Bandelt, and Antonio Torroni. The mtDNA Legacy of the Levantine Early Upper Palaeolithic in Africa. *Science*, 314(5806):1767–1770, 2006.

[43] Nicole Maca-Meyer, Ana M González, José M Larruga, Carlos Flores, and Vicente M Cabrera. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics*, 2(1):13, 2001.

[44] Elena B Starikovskaya, Rem I Sukernik, Olga A Derbeneva, Natalia V Volodko, Eduardo Ruiz-Pesini, Antonio Torroni, Michael D Brown, Marie T Lott, Seyed H Hosseini, Kirsi Huoponen, and Douglas C. Wallace. 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, 2005.

[45] Doron M Behar, Ene Metspalu, Toomas Kivisild, Alessandro Achilli, Yarin Hadid, Shay Tzur, Luisa Pereira, Antonio Amorim, Lluís Quintana-Murci, Kari Majamaa, Corinna Herrnstadt, Neil Howell, Oleg Balanovsky, Ildus Kutuev, Andrey Pshenichnov, David Gurwitz, Batsheva Bonne-Tamir, Antonio Torroni, Richard Villems, and Karl Skorecki. The Matrilineal Ancestry of Ashkenazi Jewry: Portrait of a Recent Founder Event. *The American Journal of Human Genetics*, 78(3):487–497, 2006.

[46] Max Ingman and Ulf Gyllensten. Mitochondrial Genome Variation and Evolutionary History of Australian and New Guinean Aborigines. *Genome Research*, 13(7):1600–1606, 2003.

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

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

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

[50] Richard E Green, Johannes Krause, Adrian W Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, et al. A Draft Sequence of the Neandertal Genome. *Science*, 328(5979):710–722, 2010.

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

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

[53] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[54] José A. Castro, Antònia Picornell, and Misericòrdia Ramon. Mitochondrial DNA: a tool for populational genetics studies. *International Microbiology*, 1(4):327–332, 1998.

[55] Justyna Jarczak, Łukasz Grochowalski, Błażej Marciniak, Jakub Lach, Marcin Słomka, Marta Sobalska-Kwapis, Wiesław Lorkiewicz, Łukasz Pułaski, and Dominik Strapagiel. Mitochondrial DNA variability of the Polish population. *European Journal of Human Genetics*, page 1, 2019.

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

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

[58] Irene Charon, Lucile Denoeud, Alain Guénoche, and Olivier Hudry. Maximum transfer distance between partitions. *Journal of Classification*, 23(1):103–121, 2006.

[59] Kevin E. Langergraber, Kay Prüfer, Carolyn Rowney, Christophe Boesch, Catherine Crockford, Katie Fawcett, Eiji Inoue, Miho Inoue-Muruyama, John C. Mitani, Martin N. Muller, Martha M. Robbins, Grit Schubert, Tara S. Stoinski, Bence Viola, David Watts, Roman M. Wittig, Richard W. Wrangham, Klaus Zuberbühler, Svante Pääbo, and Linda Vigilant. 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, 2012.

[60] Nick Patterson, Daniel J Richter, Sante Gnerre, Eric S Lander, and David Reich. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097):1103, 2006.

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

[62] Daniel Richter, Rainer Grün, Renaud Joannes-Boyau, Teresa E Steele, Fethi Amani, Mathieu Rué, Paul Fernandes, Jean-Paul Raynal, Denis Geraads, Abdelouahed Ben-Ncer, Jean-Jacques Hublin, and Shannon P. McPherron. The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. *Nature*, 546(7657):293–296, 2017.

[63] Jean-Jacques Hublin, Abdelouahed Ben-Ncer, Shara E Bailey, Sarah E Freidline, Simon Neubauer, Matthew M Skinner, Inga Bergmann, Adeline Le Cabec, Stefano Benazzi, Katerina Harvati, and Philipp Gunz. New fossils from Jebel Irhoud, Morocco and the pan-African origin of *Homo sapiens*. *Nature*, 546(7657):289–292, 2017.

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

[65] Ann Gibbons. Our mysterious cousinsthe Denisovansmay have mated with modern humans as recently as 15,000 years ago. *Science*, 2019.

[66] Marieke van de Loosdrecht, Abdeljalil Bouzouggar, Louise Humphrey, Cosimo Posth, Nick Barton, Ayinuer Aximu-Petri, Birgit Nickel, Sarah Nagel, El Hassan Talbi, Mohammed Abdeljalil El Hajraoui, Saaïd Amzazi, Jean-Jacques Hublin, Svante Pääbo, Stephan Schiffels, Matthias Meyer, Wolfgang Haak, Choongwon Jeong, and Johannes Krause. Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. *Science*, 360(6388):548–552, 2018.

[67] Sarra Elkamel, Sami Boussetta, Houssein Khodjet-El-Khil, Amel Benammar Elgaaied, and Lotfi Cherni. 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, 2018.

[68] Jean Trejaut. Mitochondrial dna provides a link between polynesians and indigenous taiwanese. *PLOS Biology*, 3(8), 07 2005.

[69] Jonathan S Friedlaender, Franoise R Friedlaender, Floyd A Reed, Kenneth K Kidd, Judith R Kidd, Geoffrey K Chambers, Rodney A Lea, Jun-Hun Loo, George Koki, Jason A Hodgson, D. Andrew Merriwether, and James L Weber. The genetic structure of pacific islanders. *PLOS Genetics*, 4(1):1–18, 01 2008.