frontiers in Genetics

ORIGINAL RESEARCH
published: 02 February 2022
doi: 10.3389/fgene.2021.828805

Full Chromosomal Relationships Between Populations and the Origin of Humans

*Rui Dong[1,2], Shaojun Pei[3], Mengcen Guan[3], Shek-Chung Yau[4], Changchuan Yin[5], Rong L. He[6] and Stephen S.-T. Yau[3,2]**

[1]Yau Mathematical Sciences Center, Tsinghua University, Beijing, China, [2]Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing, China, [3]Department of Mathematical Sciences, Tsinghua University, Beijing, China, [4]Information Technology Services Center, The Hong Kong University of Science and Technology, Kowloon, Hong Kong, China, [5]Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, IL, United States, [6]Department of Biological Sciences, Chicago State University, Chicago, IL, United States

A comprehensive description of human genomes is essential for understanding human evolution and relationships between modern populations. However, most published literature focuses on local alignment comparison of several genes rather than the complete evolutionary record of individual genomes. Combining with data from the 1,000 Genomes Project, we successfully reconstructed 2,504 individual genomes and propose Divided Natural Vector method to analyze the distribution of nucleotides in the genomes. Comparisons based on autosomes, sex chromosomes and mitochondrial genomes reveal the genetic relationships between populations, and different inheritance pattern leads to different phylogenetic results. Results based on mitochondrial genomes confirm the "out-of-Africa" hypothesis and assert that humans, at least females, most likely originated in eastern Africa. The reconstructed genomes are stored on our server and can be further used for any genome-scale analysis of humans (http://yaulab.math.tsinghua.edu.cn/2022_1000genomesprojectdata/). This project provides the complete genomes of thousands of individuals and lays the groundwork for genome-level analyses of the genetic relationships between populations and the origin of humans.

Keywords: 1000 genomes project, reconstructed sequences, natural vector, divided natural vector, population genetics

OPEN ACCESS

Edited by:
Guanglin He,
Nanyang Technological University,
Singapore

Reviewed by:
Zheng Wang,
Sichuan University, China
Jiang Huang,
Guizhou Medical University, China

****Correspondence:***
Stephen S.-T. Yau
yau@uic.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *04 December 2021*
Accepted: *22 December 2021*
Published: *02 February 2022*

Citation:
Dong R, Pei S, Guan M, Yau S-C, Yin C, He RL and Yau SS-T (2022) Full Chromosomal Relationships Between Populations and the Origin of Humans. Front. Genet. 12:828805. doi: 10.3389/fgene.2021.828805

INTRODUCTION

Genetic analysis of hominins has yielded insights about modern populations, superpopulations and their histories. Meanwhile, the relationship between modern and ancient humans remains controversial and of great importance for understanding our phylogenetic position in the tree of life. As the largest public catalog of human variation and genotypes, the 1,000 Genomes Project, initiated by the International Genome Sample Resource (IGSR), aims to discover genotypes and provides accurate haplotype information for all forms of human DNA polymorphism based on 2,504 individuals from 26 populations all over the world (The 1000 Genomes Project Consortium, 2010). Besides the regular study about the relationships between genotype and phenotypes such as diseases, this provides an ideal material for phylogenetic analysis about relationships among populations on genome level. On the subject of human origin, no unanimous conclusion has been reached so far due to the conflict results mainly caused by different choices of genes, incomplete data and different data types (Wickett et al., 2014). Though most researchers acknowledge the "out-of-Africa" assumption, most studies are performed based on the neutral theory of molecular evolution, whose theoretical proof hasn't not rarely examined (Leffler et al., 2012). Counterexamples of neutral theory include genome-wide constraints such as fold pressure and GC-pressure, mtDNA and nuclear genome compatibility (Huang, 2016). The neutral theory, which largely sounds as a null model and framework of pre-saturation evolutionary processes, has met with great difficulty as an explanatory framework for most molecular evolutionary phenomena and as such should not have been so freely used to account for genetic diversity patterns (Yuan et al., 2019). Thus, the "out-of-Africa" assumption requires a reasonable model and should be conducted on full chromosome data rather than a few genes or local region. The relationships between populations should be inferred at the genomic level, which reflects a complete view of individuals in a population. A comprehensive description of human evolutionary patterns is essential for studying the relationships between modern populations and, if combined with the information on ancient hominins, can be used to study the origin of modern humans.

On the other hand, fossil records of ancient hominins have been found recently and are the most direct material with which to study human origins in archeology. Sequencing results for such fossils allow us to analyze them at the genome level. The Neandertals were recognized as a distinct group of hominids with numerous fossils as well as stone tool assemblages (Krings et al., 1997; Heyes et al., 2016; Douka et al., 2019). Nevertheless, few fossils of the Denisovans, another distinct member of the *Homo* genus, have been found. Analysis of the DNA from the Neandertal and Denisovan fossils has great potential to provide insight into their population histories and relationships with modern humans, but progress has been limited due to the rarity of samples and damaged state of the DNA (Kalvin et al., 1995). Although complete genomes of ancient hominins are extremely difficult to obtain, mitochondrial genomes have already been sequenced by primer extension capture (PEC) and other techniques (Briggs et al., 2009). However, most research on ancient mitochondrial genomes focuses on local gene information instead of the global feature of the genome, and suffers from the dispute of neutral theory.

Besides the traditional alignment-based methods which requires the neutral theory assumption, many alignment-free methods have been proposed in recent years, such as feature frequency profiles (FFP) (Jun et al., 2009), power spectrum method (Yin et al., 2014; Dong et al., 2018; Pei et al., 2019), the Natural Vector (NV) method and its extensions (Deng et al., 2011; Wen et al., 2014; Dong et al., 2019) with successful applications (Huang et al., 2014; Zheng et al., 2015; Dong et al., 2017). In contrast to alignment-based methods, NV doesn't rely on human intervention or arguable assumptions such as neutral theory, and the theoretical proof of NV is solid as fully described in (Deng et al., 2011). In this project, we propose an improved version of NV method named Divided Natural Vector (DNV), which integrates local nucleotide information together with global distribution data. A parameter, k, predetermined by the algorithm, indicates how many segments the original sequence should be divided into. With an appropriate value of k, the divided natural vector method achieves the same or better results compared with those from alignment-based methods, using less computation time, as shown in **Supplementary Figure S1**. Another advantage of both NV and DNV is high time-efficiency and the compatibility of mathematical techniques derived from the conversion from biological sequences to mathematical vectors. This allows a simple but effective idea of taking average to extract features on populational level without variability from personal genomes, but difficult to conduct on DNA sequences directly. As illustrated in *Materials and Methods* and **Supplementary Figure S2**, the first step of our project is to reconstruct 2,504 individual genomes from variant calling results of the 1,000 Genomes Project and reference genomes. Then we applied DNV on all reconstructed genome sequences and extract the feature for each population, to measure the genetic similarity among populations. We also performed the DNV approach on a larger dataset including mitochondrial genomes from the 1,000 Genomes Project, Human Mitochondrial database, L0 mitogenomes and sequencing results from fossils of ancient humans, to gain insight into our origin. The workflow of this project is summarized in **Supplementary Figure S3**.

RESULTS

Total Autosome Nucleotide Tree

An alignment-free approach, the divided natural vector approach, is proposed in **Supplementary Material**. We applied this approach to the reconstructed sequences. Each individual has a reconstructed genome, consisting of 22 pairs of autosomes, one pair of sex chromosomes (XY for male, and XX for female), and a mitochondrial genome. Considering the maternal inheritance of mitochondria, that sex chromosomes differ between males and females, and that autosomes comprise over 90% of the human genome, we reconstructed phylogenetic trees by integrating all autosomes. The sum of distance matrices of 22 pairs of autosomes reflects the dissimilarities between populations derived from all autosomes and is an ideal metric for phylogenetic analysis. The corresponding tree generated from this matric is therefore called the "total autosome nucleotide tree" (TANT). Among the 26 populations from the 1,000 Genomes Project, the six current American populations are actually combinations of humans from all over the world, which introduces noise into studies on the origin of humans. The phylogenetic tree in **Figure 1** shows that all the other four superpopulations, i.e., African, European, East Asian, and South Asian, are monophyletic groups. This contrasts with the model of sequence evolution, in which only non-African individuals are monophyletic (Briggs et al., 2009). Besides, we also constructed phylogenetic trees using UPGMA/Neighbor-Joining and DNV/NV with results shown in **Supplementary Figures S4, S5**. Results based on single chromosomes can be found in **Supplementary Figures S6–S8**, and we also calculated

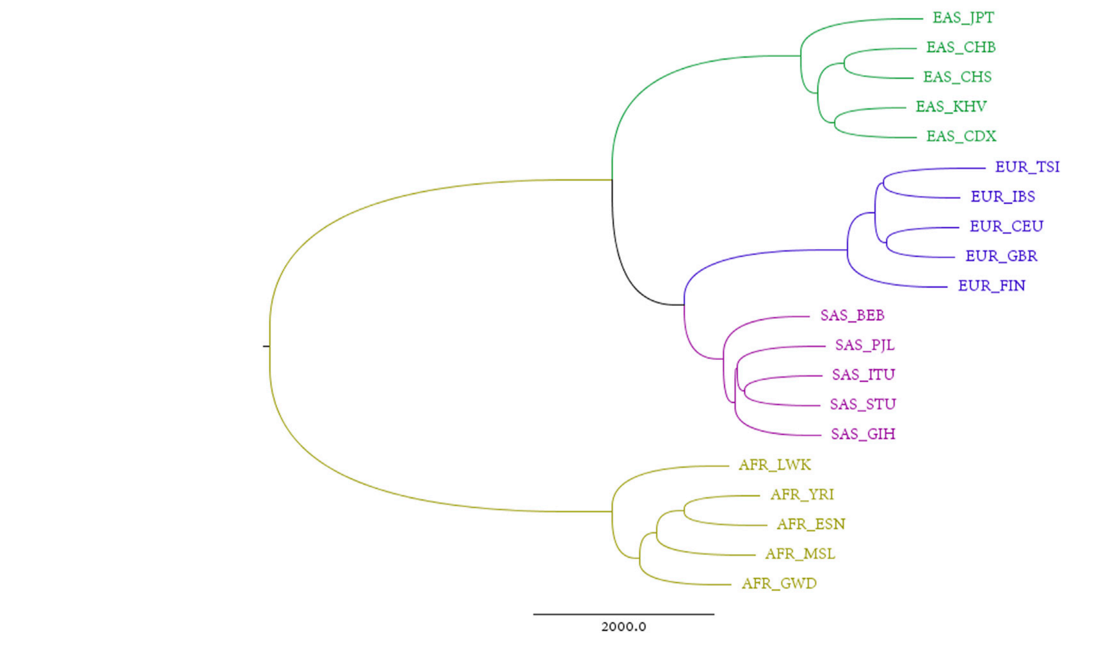

FIGURE 1 | Total autosome nucleotide tree (TANT) of 20 non-American populations based on the DNV (k = 4) approach.

the Robinson-Foulds distances for chromosome X and each autosome in **Supplementary Table S1**.

Natural Graph

Another novel way to study the relationship between units is *via* the natural graph approach. By finding every unit's closest neighbor based on layers, the relationships between populations can be intuitively observed. The natural graph for the DNV approach is presented in **Figure 2**. The first layer, which classifies each population with its closest neighbor, is shown as thin blue lines, and the second layer, which reflects the relationships between groups generated from the first layer, is shown as thick red lines. **Figure 2** shows that Americans are closely connected to other superpopulations, while African populations are more distinct from other populations (largest average distance between Africans and other monophyletic superpopulations). The distance within each subpopulation is presented in **Supplementary Table S2**, also supporting the "out

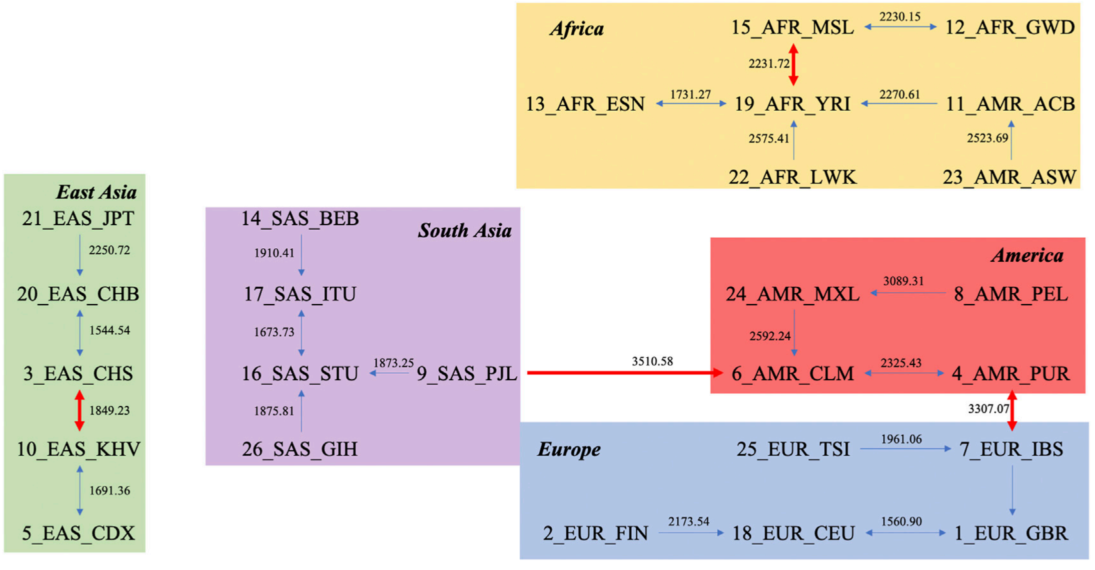

FIGURE 2 | Natural graph based on the distance matrix obtained by the DNV method.

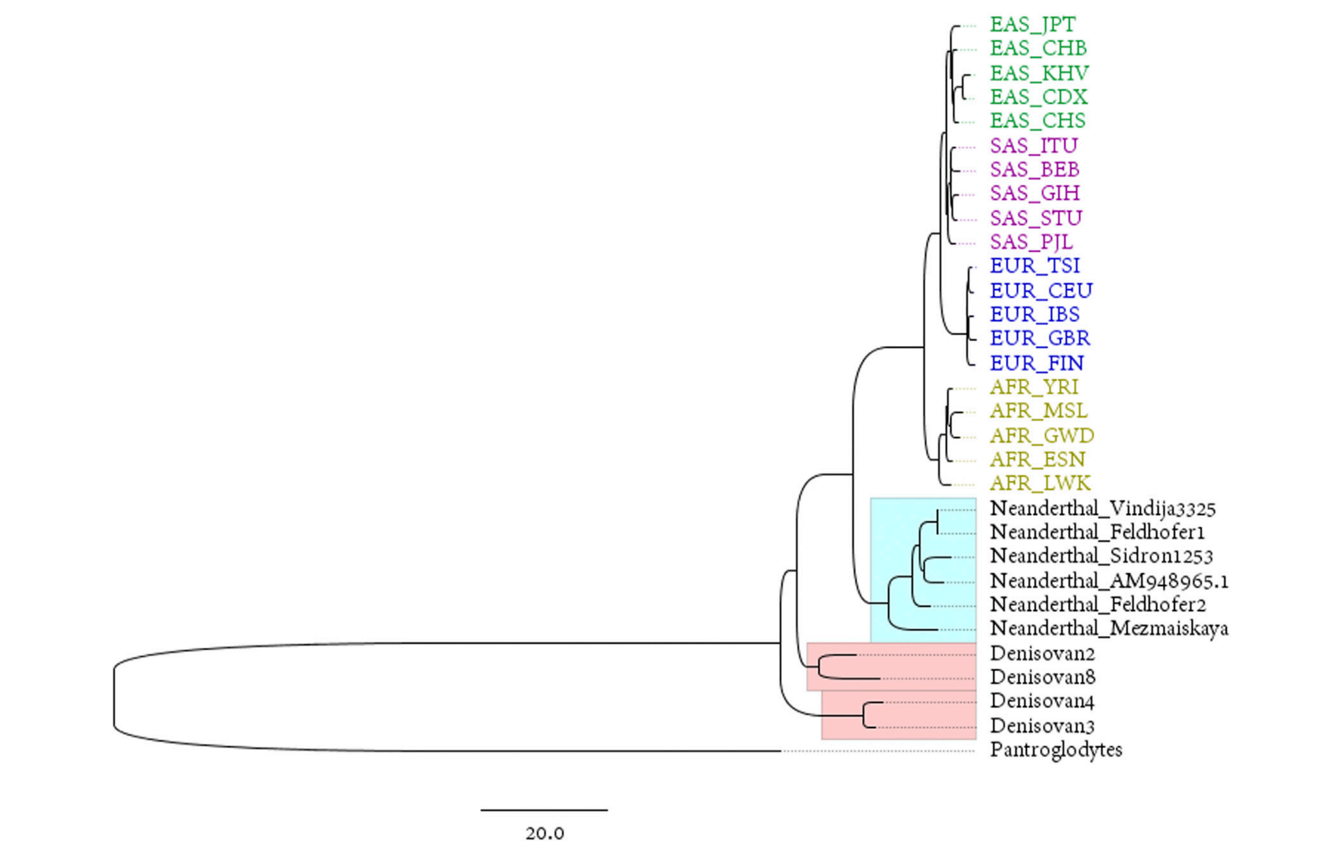

FIGURE 3 | BIONJ tree based on the distance matrix obtained with the DNV method from the mitochondrial data of 20 modern populations from 1,000 Genomes Project and 10 ancient hominins (with Pan troglodytes as the outgroup).

of Africa" hypothesis, given by the fact that larger genetic distances reflect higher diversity in genomes, and correspondingly, larger division among Africans compared to populations on other continents.

mtDNA Results Based on 1,000 Genomes Project Data

The Neanderthals and the Denisovans are two extinct species or subspecies of hominins in the genus Homo. Mitochondrial DNA was extracted from fossils of both taxa by PEC (Briggs et al., 2009) or other techniques. This provides an ideal material for studying the maternal inheritance patterns of ancient and modern humans. The reconstructed mitogenomes from the 1,000 Genomes Project represent modern populations in the dataset, while the complete mitogenomes of 4 Denisovans and 6 Neanderthals represent ancient hominins (the mitochondrial genome of Pan troglodytes is used as an outgroup to root the phylogenetic tree). The phylogenetic tree is shown in **Figure 3**.

The Luhya in Webuye, Kenya (LWK), population is the closest modern population to the root of the tree in **Figure 3**. This is the only population in the 1,000 Genomes dataset that are widely distributed in eastern Africa. Some of the earliest hominin skeletal remains have been found in this region, including those discovered in the Awash Valley of Ethiopia, as well as the Koobi Fora in Kenya. "Lucy", the well-known skeleton of a female from the hominin species Australopithecus afarensis, was discovered in 1974 in the Awash Valley and dated to approximately 3.2 million years ago (Ma). According to the predominantly held belief among most archeologists, East Africa is the area where anatomically modern humans, at least females, first appeared.

Inheritance Pattern and Relationships Among Super-Populations

In **Figure 1**, European populations show a closer relationship to South Asian, than to East Asian populations on the analysis of TANT which contains the information of all autosomes. It is easy to observe the same structure in Chromosome X for all females shown in **Supplementary Figure S7A**. This coincidence can be explained by the same parental inheritance for autosomes for all individuals and Chromosome X for females, i.e., offspring inherits one haploid from each parent.

For paternal inheritance of Chromosome Y of all males, **Supplementary Figure S7B** suggests that Europeans are more similar to East Asian compared to South Asian. Frequent wars between Europeans and East Asians in human history explain the reasons for this phenomenon as wars involves mainly men.

Besides, in **Supplementary Figure S7B**, Finland lies in the branch of East Asia and the long history connection between Finland and Russia since 1809 may contribute to this result.

For maternal inheritance, mitochondrial genomes lead to a conclusion different from either TANT or Chromosome Y. In **Figure 3** and **Supplementary Figure S8**, East Asians and South Asians share more genetics in comparison with Europeans. Geographical factors such as flat terrain, and similar language structures may contribute to the frequent communications within Asians, meanwhile Ural Mountains and Ural River separates Eurasia into two continents. On the contrary of paternal inheritance, wars play little role in the maternal inheritance and females are relatively conservative and subject to geography and linguistics due to historical reasons.

Therefore, parental, paternal and maternal inheritance patterns lead to different phylogenetic trees using the same DNV method in our work, and for more details and explanation on the results, please see **Supplementary Material Section 1.1.3**.

Human Origin Based on the Human Mitochondrial Database

Although the 1,000 Genomes Project covers most populations in the world, the five African populations are concentrated on the eastern and western coasts of Africa. The accurate prediction of human origin requires a dataset that includes most countries on the African continent. After filtering data from the Human Mitochondrial database (HmtDB) (https://www.hmtdb.uniba.it) (for more details, please see **Supplementary Material**), we analyzed 2932 high-quality mitochondrial genome sequences from 30 countries on African continent. The phylogenetic tree is shown in **Figure 4**. Eritrea lies on the branch closest to the root of the tree. A previous study by Italian researchers showed a possible link between *Homo erectus* and *Homo sapiens* (Macchiarelli et al., 2004; Zanolli et al., 2014; Medin et al., 2015). The link is related to one of the oldest hominids, named “Madam Buya” in Eritrea. Along with an adult cranium (UA 31), which displays a blend of *Homo erectus*-like and derived morpho-architectural features, and three pelvic remains, two isolated permanent incisors (UA 222 and UA 369) have also been recovered from the Homo-bearing outcrop of Uadi Aalad dated to 1 Ma. During the last interglacial period, the Red Sea coast of Eritrea was occupied by early anatomically modern humans. Eritrea also has a long and close affinity with Ethiopia, where a variety of fossils have been discovered. The second closest branch is the population of Uganda, another country in eastern Africa. The phylogenetic trees presented in **Figures 3**, **4** provide genetic evidence of the origin of humans, which is inferred to be in eastern Africa.

Phylogenetic Analysis Focusing on L0 Mitogenomes

Recent research (Chan et al., 2019) claimed that human origins in a southern African palaeo-wetland, and we explored the dataset they used in the literature as well. Two datasets, containing 6,334 and 1217 L0 mitogenomes respectively, were both firstly filtered under the same condition as illustrated in **Supplementary Material**. After filtration, we have 1733 and 563 sequences in two sets respectively. Clustal Omega algorithm and FASTME were used to construct the phylogenetic tree, and both trees suggest that an individual with accession number KF672800.1 is the deepest branch in the tree. This sample is from Kenya, therefore supports our conclusion that human, at least females, originated from eastern Africa.

In (Chan et al., 2019), there are very few mitogenomes from eastern Africa compared to the mitogenomes form southern Africa. Therefore, we also generate another set covering both the qualified sequences published in the literature and additional sequences from eastern Africa to achieve balance. This new set contains 3,495 individuals and the result also suggests the eastern Africa as the origin of human based on mitochondrial genomes. A thorough study was performed on the two specific sequences that are close to the root of the phylogenetic tree (EU092699.1 from Mozambique and EU092943.1 from Ethiopia), and we compare the similarity of these two sequences to mtDNA of gorilla (KM242275.1) using NCBI BLAST method. The identity between Ethiopia and gorilla is 13944/15584, higher than between Mozambique and gorilla 13939/15584. Besides, the third to sixth sequences closest to the root are all sequenced from samples in eastern Africa (Chad, Kenya, Ethiopia and Ethiopia). Therefore, based on the evidence from mitochondrial genomes, the origin of human should lie in eastern Africa rather than southern Africa.

The tree files can all be found on GitHub (https://github.com/YaulabTsinghua/Human-Origin-1kGP).

DISCUSSION

This project proposes the DNV approach and applies it on the data from the 1,000 Genomes Project, providing insight into the genetic relationships among the main human populations worldwide. Alignment-free methods have largely been ignored as a powerful tool for studying the human genomes and population relationships, and our research fills this gap. With the successful application of DNV described in *Natural Vector and Divided Natural Vector* and **Supplementary Material**, we prove that DNV outperforms NV on the test set and enjoys higher time-efficiency compared to alignment methods. DNV offers greater speed as recorded in **Supplementary Table S3**, which is significant advantage over alignment-based approaches, especially dealing with genomes data.

The relationship among populations is studied under different inheritance patterns based on whole-genome data of 2,504 individuals, and the results show that parental, paternal and maternal inheritances lead to different evolutionary results in the corresponding phylogenetic trees. Similarity between East Asians and Europeans can be explained by the frequent wars in history, mainly conducted and participated by males. Asian females share higher genetics within Asia compared to them with European females, reasons of which includes geographical and linguistics factors. Autosomes and Chromosome X for females, however, generate a

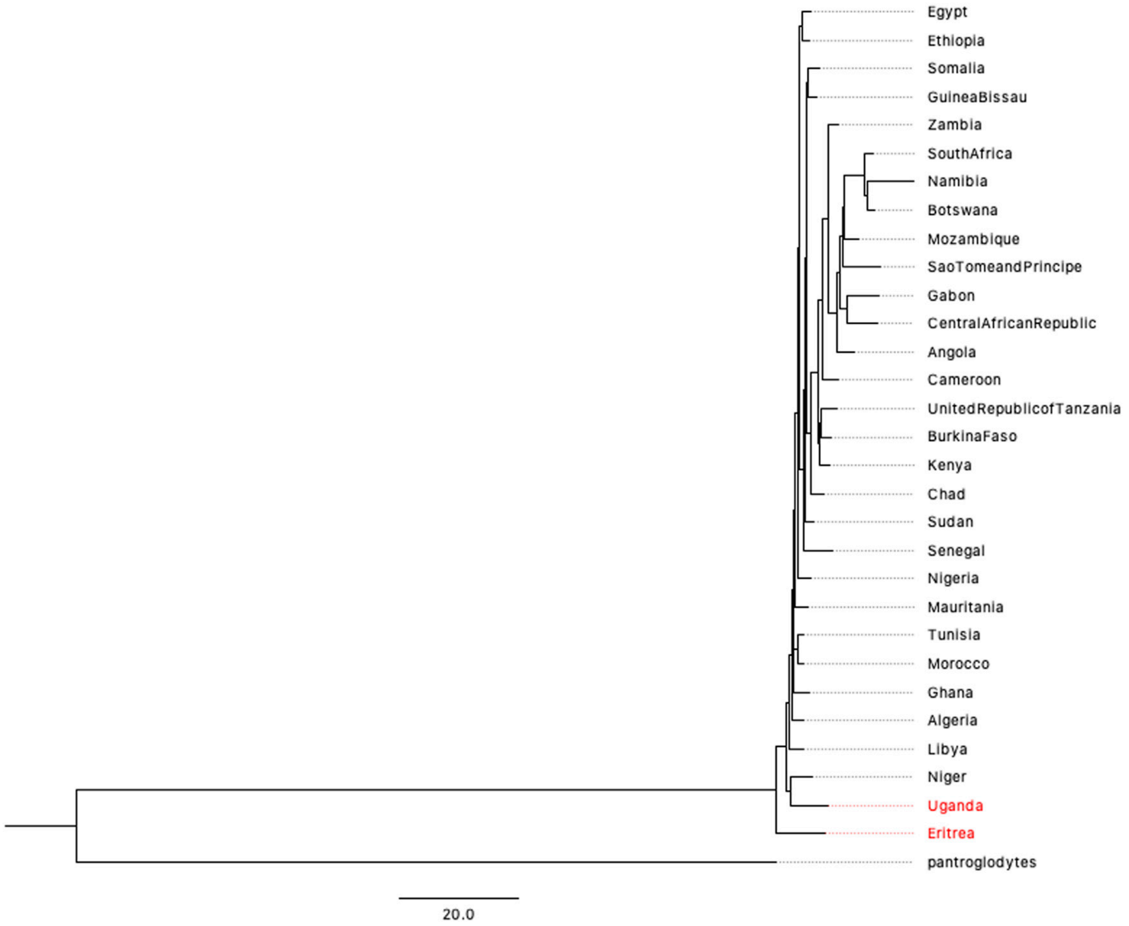

FIGURE 4 | BIONJ tree of 30 countries on African continent based on the mitochondrial data of 2932 individuals (with Pan troglodytes as the outgroup).

result different from both, but can be viewed as a balanced average of the paternal and maternal results.

As to the topic of origin of human, the large average distance within African populations suggests a high diversity on genome level, and phylogenetic analysis approves the origin of Africa as well. This is the solid evidence for the 'out-of-Africa' model because the distance is calculated based on all autosomes instead of a local region or gene. To explore more details about the location of origin, we collect external data source from other mitochondrial databases to cover the whole African continent, and compared them all with the ancient mitogenomes extracted from fossils. Results in **Figures 3**, **4** both suggest that eastern African is more likely to be the origin of human and Eritrea is with the highest possibility. Results based on the 1,000 Genomes Project suggests that Kenya population is the closest which also supports the conclusion of eastern Africa, though the data source covers only five populations in Africa. The assertion depends on the appropriate outgroup, in this project, chosen as the mitogenome of pan troglodytes. However, due to the inconsistency of number of autosomes of human and other primates, root of phylogenetic tree for autosomes is difficult to identify based on our method. Research results based on mitochondrial genomes reflect the maternal inheritance, and if we consider parental as a mixture of paternal and maternal, plus the fact that early migrations are less frequent than nowadays, the conclusion of eastern Africa origin still stands. Our project provides a satisfying answer to the debate generated by the use of different datasets and analytical methods. Resolving such phylogenetic relationships is important for comparative genomics, which may be further applied to the study of human diseases. Most studies published to date have applied alignment algorithms to local variants, which neglects the key information hidden in whole-genome data.

In summary, the proposed DNV method resolves the issue of alignment methods caused by insufficient data and our whole-genome data source improves phylogenetic reconstructions using the complete evolutionary record within each individual's genome (Javis et al., 2015). The reconstructed genomes are stored on our server (http://yaulab.math.tsinghua.edu.cn/2022_1000genomesprojectdata/) and can be further used for any genome-scale analysis of humans. The project lays the groundwork for genome-level analyses of the genetic relationships between populations and the origin of humans.

MATERIALS AND METHODS

Reconstruct the Genomes Data From VCF File

The 1,000 Genomes Project performs alignment using the Burrows-Wheeler transformation (BWA) mapping algorithms for low coverage sequencing in phase 3 of the whole project. Variant calls are always released in variant call format (VCF) (Li and Durbin 2009; Danecek et al., 2011). Based on the reference sequence and the VCF record data, it is possible to reconstruct the sequences for each individual, by replacing the nucleotides on the reference sequence with the variants at the corresponding positions. Although the current coverage of sequencing technology is insufficient to really 'reconstruct' the genome of each individual, it is still a promising approach for studying the relationships among populations, by using averaging to decrease the noises in the reconstructed dataset.

BCFtools is a set of utilities that manipulate variant calls in the VCF and its binary counterpart BCF. The command 'consensus' in the set of BCFtools can create a consensus sequence by applying VCF variants to a reference fasta file (Danecek and McCarthy, 2017). An example of a VCF file with five individuals is shown in **Supplementary Figure S1A**. The first individual, named as "HG00096", has no variants at these positions, therefore, its corresponding sequence remains the same as the reference sequence. The second individual "HG0770" has a variant of "C" at POS = 10, while the reference nucleotide is a "T". Therefore, we change the 10th nucleotide on the reference sequence from "T" to "C", so as the 26th nucleotide from "C" to "T". Any positive integer in the main data means the corresponding alternative in the ALT column. For "HG01992" and "HG02230", they both have the second alternatives on position 55, then we locate the 55th nucleotide, and because there is more than one nucleotide in the REF column, the corresponding segment to be replaced has more than one nucleotide. "POS = 55" refers to the first nucleotide in this segment, and the second alternative, "CATTTT", is used to fill in the segment in this window. If we only consider the data in **Supplementary Figure S1A**, individuals "HG00096" and "HG02231" should have the same reconstructed sequences as the reference sequence, and "HG01992" and "HG02230" should share the same reconstructed sequences. The corresponding reconstructed sequences are shown in **Supplementary Figure S1B**.

In the 1,000 Genomes Project, both short variations and structural variations are detected. Most of the complicated variations consist of copy number variations (CNV). CNV is a phenomenon in which sections of the genome are repeated and the number of repeats in the genome varies between individuals in the human population. It often becomes unclear whether or not a region is an overlap or a duplicated region. Usually, the POS coordinate is based on the leftmost possible position of the variant, which indicates that the accuracy of estimation could strongly affect the duplicated region. Therefore, in our study, the structural variants are not taken into consideration, and could be implemented into the current results in further analysis.

Humans Diploidy and Pairs of Chromosomes

Besides mitochondria genomes, the human genome includes the 23 chromosome pairs, 22 of which are called autosomes, with the other pair commonly known as the "sex chromosome". The diploidy of human individuals results in the possibility of different sequencing results on the same chromosome pair and it is impossible to detect if a chromosome is patrilineal or matrilineal with current sequencing techniques. For an individual, we picked one chromosome of each pair to reconstruct two separate chromosome sequences, named "a" and "b", one from the father and the other from the mother. Thus, regarding autosomes, every individual has Chr1a, Chr1b; Chr2a, Chr2b; ...; Chr22a, Chr22b in the reconstructed autosomes. Please note that it is possible that "a" means matrilineal in the Chromosome 1, but means patrilineal in the Chromosome 2.

For the 23rd chromosome pair, i.e., the sex chromosome pair, females have a pair of X chromosomes, and males have one X chromosome and one Y chromosome. Using female data to study the phylogeny of the X chromosome, and male data to study Y chromosome, makes it possible to separate any possible noise in the pseudo-autosomal region, although the number of samples are decreased by half. Pseudo-autosomal regions are homologous sequences of nucleotides on the X and Y chromosomes, but on different positions, partly due to their markedly different scales. Therefore, the existence of pseudo-autosomal regions increases the difficulty of studying the phylogenetic relationships based on gender. In the situation of female data, procedures are identical to the autosome cases, while for male data, it degenerates to a simpler case: one sequence (of a Y chromosome) represents a male individual.

Natural Vector and Divided Natural Vector

For each chromosome (except the Y chromosome), each individual's pairs of chromosomes are reconstructed. Then we applied the alignment-free natural vector method to this dataset.

The natural vector method was first proposed by Yau and his team (Deng et al., 2011). The natural vector of a sequence encodes the distribution of four nucleotides. Yau has proven mathematically that there is a one-to-one mapping between a sequence and its natural vector (Deng et al., 2011). Given any nucleotide sequence, the natural vector method maps it into a vector/point in real Euclidean space, where for each type of nucleotide, the first parameter describes its quantities in the sequence, the second gives the mean values of total distances to the first nucleotide in the sequence, and the last consists of the normalized central moments. The normalized central moments are defined as follows:

$$D_j^k = \sum_{i=1}^{n_k} \frac{(s[k][i] - \mu_k)^j}{n_k^{j-1} n^{j-1}}, j = 2, \ldots, n_k$$

where k = A, C, G, T. Here, n_k denotes the number of nucleotide k in the DNA sequence and n is the length of the sequence. s[k][i] is the distance from the first nucleotide (regarded as origin) to the *i*th nucleotide k in the sequence. $T_k = \sum_{i=1}^{n_k} s[k][i]$ denotes the total distance for each set of A, C, G, T from the origin, k = A, C, G, T. $\mu_k = \frac{T_k}{n_k}$ is the mean value of the distances of nucleotide k. The central moments can be written as $< D_2^A, D_2^C, D_2^G, D_2^T, \ldots, D_{n_A}^A, D_{n_C}^C, D_{n_G}^G, D_{n_T}^T >$. It is obvious that higher moments converge to 0. In common practice, the first 12 dimensions of the natural vector: $< n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_2^A, D_2^C, D_2^G, D_2^T >$ are used, and have been found to perform well. Once their natural vectors have been calculated, the similarity between sequences can be measured by the Euclidean distance between the corresponding vectors.

Consequently, each pair of chromosomes, i.e., sequences, can be converted to a pair of natural vectors, and then those natural vectors can be used for further analysis. Based on the uncertainty of which one in the pair is patrilineal, they should be assigned equal weight in the study. In mathematics, one of the best ways to describe a pair of vectors is to take their average position. Consider two points in n-dimensional space, $(x_1, x_2, .., x_n)$ and $(y_1, y_2, .., y_n)$, their average position is $(\frac{x_1+y_1}{2}, \frac{x_2+y_2}{2}, \ldots, \frac{x_n+y_n}{2})$. In the case of the traditional natural vector method, n = 12. Thus, we can use an individual's 22 autosomes to get 22 average personal natural vectors. For a female individual, extra average natural vector can be generated from her pair of Chromosome X and the Robinson-Foulds distance between BIONJ trees from chromosome X to each autosome is shown in **Supplementary Table S2**.

In the final phase of 1,000 Genomes Project, data from 2,504 individuals was made available to the worldwide scientific community through freely accessible public databases. Therefore, regarding Chromosome 1, we have 2,504 pairs of Chromosome 1, corresponding to 2,504 individuals. The 2,504 pairs of chromosomes are represented by 2,504 pairs of natural vectors, also by 2,504 average personal natural vectors, which reveal mathematically the hidden information in the original reconstructed sequence, i.e., from reference sequence and the SNP (single nucleotide polymorphism) and other variants.

Each population contains about 100 individuals in the dataset, for example, "ACB" is short for "African Caribbean in Barbados", which consists of 96 individuals, corresponding to 96 average personal natural vectors of Chromosome 1. The idea of taking an average is applied again here to find the representative point of each population, fixed one chromosome, which is the center of all the average personal natural vectors of the individuals that come from this population. Likewise, the other 25 populations can be represented by 25 points in the 12-dimensional space for each chromosome, and further analysis on genetic relationships among population are based on the average natural vectors for each chromosome.

The idea of taking an average plays a crucial role in our project, since it solves the problem of different sequences leading to different phylogenetic analyses. For alignment approaches, taking an 'average' alignment is not realizable, therefore further phylogenetic analysis does not lead to a convincing conclusion. However, alignment-free approaches are able to address this limitation by introducing mathematical concepts to the biological world, which is also one of the most common techniques in engineering. Averaging is both intuitively clear, and it is something easy to compute in practice.

We have also proposed an extension of the natural vector method, named Divided Natural Vector (DNV), based primarily on the idea of combining the natural vector method together with alignment algorithms. Alignment methods aim at finding the homologous positions among sequences. In contrast with the natural vector method, the alignment results are determined by local correspondence, rather than the global statistical information. The divided natural vector approach evaluates a sequence in two steps: first, we divide the sequence into k segments, where k is a positive integer; second, the 12-dimensional traditional natural vectors for each segment are concatenated into a single vector, with length of 12*k. When k = 1, the DNV approach degenerates to the case of traditional natural vector method. The computational complexity increases as k becomes larger, since more segments provide more direct evidence of local information than the rough statistics. The results of DNV could be the same as alignment approaches when an appropriate value of k is chosen.

In this project, we set k = 4 based on the testing of DNV, NV and MUSCLE on a published dataset (Briggs et al., 2009). The dataset including 55 samples was tested to perform the analysis, including the whole genomes of 46 modern individuals, one reference human genome, 6 Neanderthals and 2 other primates. This is a subset of the dataset used in (Briggs et al., 2009). The validation criterion was that the good approach should distinguish between the following groups: human and other primates, Neanderthals and modern individuals, 15 modern African and 31 non-African individuals. The UPGMA trees for the natural vector method, the divided natural vector (k = 4) method and MUSCLE are shown in **Supplementary Figures S2A–C**, respectively. All three approaches can distinguish between human and other primates, except the natural vector method fails to distinguish between Africans and non-Africans in **Supplementary Figure S2A**. Both the divided natural vector (k = 4) method and MUSCLE get satisfactory results, but MUSCLE takes about 1 hour 18 minutes, while DNV takes only seconds. Therefore, we have proved that the divided natural vector method extracts the information hidden in the original sequences more accurately than the traditional Natural Vector method, and is much more time-efficient compared to the alignment approach.

The DNV approach was also conducted on the same reconstructed sequences of all chromosomes and the results indicate that it does capture more information than the traditional natural vector method and the results are more stable.

Euclidean Distance and the Phylogenetic Analysis

The workflow of this project is summarized in **Supplementary Figure S3**. After the original reconstructed sequences are converted to natural vectors (which correspond to points in real Euclidean space), the Euclidean distance between the points can be easily calculated to measure the difference between the sequences. For a specific chromosome, an average natural vector of each population is represented by the center of average personal natural vectors of chromosome pairs from all individuals who belong to this population. Thus, the distance between two populations can be depicted by the distance between two centers, i.e., two average natural vectors.

After the pairwise-distance matrix between populations was obtained, phylogenetic analysis was constructed in several ways in our project. FASTME provides distance algorithms to infer phylogenies based on balanced minimum evolution, which is the principle behind Neighbor-Joining (NJ). FASTME is an improvement over NJ because it performs topological moves using fast, sophisticated algorithms (Saitou and Nei 1987; Gascuel et al., 1997; Desper and Gascuel 2002; Desper and Gascuel 2004). We applied FASTME to the dataset analyzed with default parameters, and BIONJ, an improved version of the Neighbor-Joining algorithm based on a simple model of sequence data. In addition, the Unweighted Pair Group Method with Arithmetic mean (UPGMA) and Neighbor-Joining (NJ) are also straightforward ways installed in MEGA X (Sneath and Sokal 1973; Saitou and Nei 1987; Kumar et al., 2018) to construct phylogenetic trees and study the evolutionary relationships based on the distance matrix of all units involved. These three algorithms were all tested to construct reliable phylogenetic trees in our project.

Any distance matrix can be the input to these tree construction algorithms, as long as it is symmetric, diagonal elements are zero, and the remaining elements are positive. In order to present a comprehensive phylogenetic tree, all the 22 distance matrices derived from single autosome are summed together to obtain a Total Autosome Nucleotide Tree (TANT), containing information from all autosomes. The sum here is also an application of the idea of taking an average, since the sum and average only differs in one constant, and this substantial equivalence will lead to same results.

Another novel method to visualize the relationships between units is Natural Graph (Zheng et al., 2015). By finding the closest neighbor of all units in several layers (in most cases, two layers are enough), it successfully classifies units into several groups and also reveals the relationships between groups. Evidently, different chromosomes contribute to different average natural vectors, which leads to different phylogenetic results, and this principle applies to both phylogenetic trees and Natural Graphs.

Filtration of Data From HmtDB and L0 Mitogenomes

We downloaded all mitochondrial sequences of each African country from HmtDB (https://www.hmtdb.uniba.it/query). Only complete genomes from healthy individuals on African continent are selected in this step. Since the length of human reference mitochondrial sequence is 16,569 bp, we deleted the sequences that are too short (<16,564 bp) or too long (>16,574 bp), and also the sequences that contain more than 1 "N". Besides, each country contains at least two sequences that satisfy the criteria above. After the filtration, we obtained 2932 sequences from 30 African countries which can be used for evolutionary analysis.

For the L0 mitogenomes, we used the same criteria to select the reliable sequences, based on the length and the number of "N" in each sequence. Because the trees were constructed on the individual level for L0 mitogenomes, no country information is required in the phylogenetic analysis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the **Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

SS-TY and RH conceived the idea of reconstructing genomes. SS-TY proposed Divided Natural Vector method and taking average to remove noise. RD implemented the idea and wrote the first draft of the manuscript. SP, MG, and CY discussed and revised the first draft. S-CY helped on the programming on servers and testing. All authors agreed with the manuscript results and conclusions. They jointly developed the structure and arguments for the paper, made critical revisions and approved final version, and reviewed and approved the final manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China grant (12171275), Tsinghua University Spring Breeze Fund (2020Z99CFY044), Tsinghua University start-up fund, and Tsinghua University Education Foundation fund (042202008).

ACKNOWLEDGMENTS

The corresponding author would like to thank National Center for Theoretical Sciences (NCTS) for providing excellent research environment while part of this research was done. We would like to thank Prof. Max Benson for his help on the revision of our manuscript, and Prof. Shi Huang for helpful discussions on his doubt of African origin of modern humans.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.828805/full#supplementary-material

REFERENCES

Briggs, A. W., Good, J. M., Green, R. E., Krause, J., Maricic, T., Stenzel, U., et al. (2009). Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. *Science* 325, 318–321. doi:10.1126/science.1174462

Chan, E. K. F., Timmermann, A., Baldi, B. F., Moore, A. E., Lyons, R. J., Lee, S-S., et al. (2019). Human Origins in a Southern African Palaeo-Wetland and First Migrations. *Nature* 575, 185–189. doi:10.1038/s41586-019-1714-1

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The Variant Call Format and VCFtools. *Bioinformatics* 27 (15), 2156–2168. doi:10.1093/bioinformatics/btr330

Danecek, P., and McCarthy, S. A. (2017). BCFtools/Csq: Haplotype-Aware Variant Consequences. *Bioinformatics* 33 (13), 2037–2039. doi:10.1093/bioinformatics/btx100

Deng, M., Yu, C., Liang, Q., He, R. L., and Yau, S. S-T. (2011). A Novel Method of Characterizing Genetic Sequences: Genome Space with Biological Distance and Applications. *PLoS One* 6 (3), e17293. doi:10.1371/journal.pone.0017293

Desper, R., and Gascuel, O. (2002). Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution. *J. Comput. Biol.* 9 (5), 687–705. doi:10.1089/106652702761034136

Desper, R., and Gascuel, O. (2004). Theoretical Foundation of the Balanced Minimum Evolution Method of Phylogenetic Inference and its Relationship to Weighted Least-Squares Tree Fitting. *Mol. Biol. Evol.* 21 (3), 587–598. doi:10.1093/molbev/msh049

Dong, R., He, L., He, R. L., and Yau, S. S-T. (2019). A Novel Approach to Clustering Genome Sequences Using Inter-nucleotide Covariance. *Front. Genet.* 10, 234. doi:10.3389/fgene.2019.00234

Dong, R., Zheng, H., Tian, K., Yau, S-C., Mao, W., Yu, W., et al. (2017). Virus Database and Online Inquiry System Based on Natural Vectors. *Evol. Bioinformatics* 13, 1–7. doi:10.1177/1176934317746667

Dong, R., Zhu, Z., Yin, C., He, R. L., and Yau, S. S-T. (2018). A New Method to Cluster Genomes Based on Cumulative Fourier Power Spectrum. *Gene* 673, 239–250. doi:10.1016/j.gene.2018.06.042

Douka, K., Slon, V., Jacobs, Z., Ramsey, C. B., Shunkov, M. V., Derevianko, A. P., et al. (2019). Age Estimates for Hominin Fossils and the Onset of the Upper Palaeolithic at Denisova Cave. *Nature* 565, 640–644. doi:10.1038/s41586-018-0870-z

Gascuel, O. (1997). BioNJ: an Improved Version of the NJ Algorithm Based on a Simple Model of Sequence Data. *Mol. Biol. Evol.* 14 (7), 685–695. doi:10.1093/oxfordjournals.molbev.a025808

Heyes, P. J., Anastasakis, K., de Jong, W., van Hoesel, A., Roebroeks, W., and Soressi, M. (2016). Selection and Use of Manganese Dioxide by Neanderthals. *Scientific Rep.* 6, 22159. doi:10.1038/srep22159

Huang, H-H., Yu, C., Zheng, H., Hernandez, T., Yau, S-C., He, R. L., et al. (2014). Global Comparison of Multiple-Segmented Viruses in 12-dimensional Genome Space. *Mol. Phylogenet. Evol.* 81, 29–36. doi:10.1016/j.ympev.2014.08.003

Huang, S. (2016). New Thoughts on an Old riddle: what Determines Genetic Diversity within and between Species. *Genomics* 108, 3–10. doi:10.1016/j.ygeno.2016.01.008

Javis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., et al. (2015). Whole-genome Analyses Resolve Early Branches in the Tree of Life of Modern Birds. *Science* 346 (6215), 1320–1331. doi:10.1126/science.1253451

Jun, S-R., Sims, G. E., Wu, G. A., and Kim, S-K. (2009). Whole-proteome Phylogeny of Prokaryotes by Feature Frequency Profiles: An Alignment-free Method with Optimal Feature Resolution. *Proc. Natl. Acad. Sci. United States America* 107 (1), 133–138. doi:10.1073/pnas.0913033107

Kalvin, A. D., Dean, D., and Hublin, J.-J. (1995). Reconstruction of Human Fossils. *IEEE Comp. Graphics Appl.* 15 (1), 12–15. doi:10.1109/38.364954

Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking, M., and Pääbo, S. (1997). Neandertal DNA Sequences and the Origin of Modern Humans. *Cell* 90 (1), 19–30. doi:10.1016/s0092-8674(00)80310-4

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi:10.1093/molbev/msy096

Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., et al. (2012). Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species. *Plos Biol.* 10 (9), e1001388. doi:10.1371/journal.pbio.1001388

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

Macchiarelli, R., Bondioli, L., Chech, M., Coppa, A., Fiore, I., Russom, R., et al. (2004). The Late Early Pleistocene Human Remains from Buia, Danakil Depression, Eritrea. *Rivista Italiana di Paleontologia e Stratigrafia* 110, 133–144. doi:10.13130/2039-4942/5768

Medin, T., Martínez-Navarro, B., Rivals, F., Libsekal, Y., and Rook, L. (2015). The Late Early Pleistocene Suid Remains from the Paleoanthropological Site of Buia (Eritrea): Systematics, Biochronology and Eco-Geographical Context. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 431, 26–42. doi:10.1016/j.palaeo.2015.04.020

Pei, S., Dong, R., He, R. L., and Yau, S. S-T. (2019). Large-scale Genome Comparison Based on Cumulative Fourier Power and Phase Spectra: central Moment and Covariance Vector. *Comput. Struct. Biotechnol. J.* 17, 982–994. doi:10.1016/j.csbj.2019.07.003

Saitou, N., and Nei, M. (1987). The Neighbor-Joining Method: a New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.* 4 (4), 406–425. doi:10.1093/oxfordjournals.molbev.a040454

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

The 1000 Genomes Project Consortium (2010). A Map of Human Genome Variation from Population-Scale Sequencing. *Nature* 467, 1061–1073. doi:10.1038/nature09534

Wen, J., Chan, R. H.-F., Yau, S-C., He, R. L., and Yau, S. S. T. (2014). K-mer Natural Vector and its Application to the Phylogenetic Analysis of Genetic Sequences. *Gene* 546, 25–34. doi:10.1016/j.gene.2014.05.043

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., et al. (2014). Phylotranscriptomic Analysis of the Origin and Early Diversification of Land Plants. *Proc. Natl. Acad. Sci. United States America* 111, E4859–E4868. doi:10.1073/pnas.1323926111

Yin, C., Chen, Y., and Yau, S. S.-T. (2014). A Measure of DNA Sequence Similarity by Fourier Transform with Applications on Hierarchical Clustering. *J. Theor. Biol.* 359, 18–28. doi:10.1016/j.jtbi.2014.05.043

Yuan, D., Lei, X., Gui, Y., Wang, M., Zhang, Y., Zhu, Z., et al. (2019). *Modern Human Origins: Multiregional Evolution of Autosomes and East Asia Origin of Y and mtDNA*. BioRxiv, 101410.

Zanolli, C., Bondioli, L., Coppa, A., Dean, C. M., Bayle, P., Candilio, F., et al. (2014). The Late Early Pleistocene Human Dental Remains from Uadi Aalad and Mulhuli-Amo (Buia), Eritrean Danakil: Macromorphology and Microstructure. *J. Hum. Evol.* 74, 96–113. doi:10.1016/j.jhevol.2014.04.005

Zheng, H., Yin, C., Hoang, T., He, R. L., Yang, J., and Yau, S. S-T. (2015). Ebolavirus Classification Based on Natural Vectors. *DNA Cel Biol.* 34 (6), 418–428. doi:10.1089/dna.2014.2678

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.