NOVEL INSIGHTS FROM NON-ULTRAMETRIC PHYLOGENETIC TREES: THE CASE OF NEANDERTHAL INTROGRESSION

Arturo Tozzi
Center for Nonlinear Science, Department of Physics, University of North Texas, Denton, Texas, USA
1155 Union Circle, #311427Denton, TX 76203-5017 USA
tozziarturo@libero.it

ABSTRACT

Ultrametric spaces are widely used to depict evolutionary times in phylogenetic trees, since they assume that every population/species is located at the tips of divergent branches. The discrete branching of ultrametric trees permits the measurement of distances between pairs of individuals that are proportional to their divergence time. Here we overturn the traditional ultrametric concept of divergent phylogenetic tree and introduce a new type of non-ultrametric diagram to describe gene flows in terms of convergent branches. To provide an operational example, we examine the paleoanthropological issue of Neanderthal genome's introgression in non-African humans. Neanderthals and ancient humans are not anymore two species that exchange chunks of DNA, rather become a novel cluster that must be considered by itself. Our converging, non-ultrametric phylogenetic trees permit the calibration of molecular clocks with a twofold benefit. When the date of the branching of two population/species from a common ancestor is known, our method allows to calculate the time of subsequent introgressions. On the contrary, when the date of the introgression between two population/species is known, our method allows to detect the time of their previous branching from a common ancestor.

KEYWORDS: triangle inequality; sets; molecular clock; hominins.

INTRODUCTION

Phylogenetic analysis protocols entail various steps, from the identification of sets of homologous sequences to the download of the achieved sequences, ending up with the production of a phylogenetic tree (PT) inferred from the aligned sequences (Tamura et al., 2011). PTs depict morphological/genetic evolutionary relationships among populations or species (Woese 2000). Unlike other evolutionary diagrams such as, e.g., additive trees and cladograms, the PT abscissa describes the amount of change of a given character occurred over time, while the ordinate describes the evolutionary time. PT makes use of diverging branches in which the tips stand for groups of descendants' populations/species and the nodes for their common ancestors. The branch lengths are proportional to the inferred evolutionary distance among the specimens under investigation (Hall 2013). Rooted PT of clustering sequences are generally set in ultrametric spaces, i.e., non-Euclidean spaces in which all the tips are equidistant from the root. To provide an example, the widely used UPGMA and WPGMA methods (i.e., Unweighted and Weighted Pair Group Methods with Arithmetic mean) generate ultrametric trees characterized by the constant-rate assumption that the distances from the root to every branch tip are equal (Sokal and Michener, 1958). Therefore, ultrametric trees are very useful to describe a peculiar aspect of the evolutionary history, i.e., the time of divergence among populations/species (Page and Holmes, 1998).

Here we propose an alternative approach to ultrametric trees, i.e., a novel non-ultrametric phylogenetic tree (NUPT). We illustrate our model and discuss its theoretical advantages and operational implications, providing an example drawn from the paleoanthropological issue of ancient hominins' DNA introgression in Late Pleistocene Homo sapiens. Putative gene flows from archaic hominins to Homo sapiens point towards a very intricate evolutionary history characterized by a single of multiple episodes of gene flow between populations/species (Prüfer et al., 2017). For example, genome sequencing of various hominins' specimens points towards gene flows among Neanderthals, Denisovans, unknown archaic groups and early modern humans (Reich et al., 2010; Meyer et al., 2012; Prüfer et al. 2014; Hajdinjak et al., 2021).

We focus here on the genetic relationships between Neanderthals and modern humans. The inferred timing of their last common ancestor is a hotly debated matter, since small variations in the weighed parameters lead to very different estimates of their divergence time (Meyer et al., 2016). Mixed results from studies of cranial variation, ancient DNA, human mutation rate, mitochondrial sequence comparisons, average value of linkage disequilibrium suggest that the lineages leading to modern humans and Neanderthals diverged between 300,000 and 600,000 years ago, with little subsequent admixture (Krings et al. 1997; Serre et al. 2004; Green et al. 2006, 2008; Noonan et al. 2006). More radical suggestions date back the last common ancestor to pre-800,000 years ago (Gómez-Robles A, 2019). It has been estimated that an introgression from Neanderthals to non-African human ancestor occurred 37,000–86,000 years ago, most likely between 47,000 and 65,000 years ago (Sankararaman et al., 2012; Posth et al., 2017). In the sequel, we make clear why NUPT sheds new light on the times of divergence and introgression of Neanderthals and non-African human populations.

ULTRAMETRIC DIVERGING TREES: WHAT, WHEN AND WHY

The ultrametric trees rely on the mathematical concept of ultrametricity, which is closely related to the concept of triangle inequality. In plane geometry, triangle inequality states that the length of every side of the triangle is lower than or equal to the sum of the other two, such that:

$$d\ (x,z) \leq d\ (x,y) + d\ (y,z)$$

where x, y, and z are the three vertices of a triangle and d is the distance between every pair of vertices (**Figure 1A**). In this case, the triangle lies in a classical Euclidean space.
Yet, there exist a strongest version of triangle inequality, such that:

$$d\ (x,z) \leq \max\ [d\ (x,y),\ (y,z)]$$

In plain words, the rule of strong triangle inequality takes into account just peculiar triangles, i.e., equilateral or isosceles with a basis smaller than the other sides (**Figure 1B**). Those triangles do not lie in Euclidean space, rather in a space termed ultrametric. This non-Euclidean space can be depicted either as a triangular grid where the distances are always preserved (**Figure 1C**), or as a tree-like hierarchy of triangles/balls (**Figure D**).
The ultrametric assumption requires that an ultrametric tree must display the following unusual properties:

1) The ultrametric distances are described by steps that are unavoidably discrete: continuum is not allowed.
2) The ultrametric-based clocks rely on the assumption that the temporal distance always corresponds to the total time multiplied by a constant.
3) Different ultrametric balls do not have points in common. This leads to counterintuitive outcomes:
    a) In an ultrametric space, two balls cannot overlap.
    b) In an ultrametric space, two balls always keep the same spacing and fixed distance.

Scientists use DNA sequence data gathered from hominin samples to test models of archaic admixture such as, in our case, Neanderthal gene flow in modern non-African humans. To evaluate recombination events and build PT, scientists infer the date of the last genetic exchange between the ancestral populations/species by measuring the extent of admixture linkage disequilibrium (Sankararaman et al., 2016). A widely used procedure for dating gene flow in target populations/species assesses a set of single nucleotide polymorphisms at which a single randomly chosen allele is derived relative to an ancestor (Wall 2000; Sankararaman et al., 2016). Traditionally, scientists have preferred ultrametric trees for ancestral state reconstruction and phylogenetic inference because of their unvaluable practical advantages. Discrete branchings of ultrametric trees are characterized by a distance between any pair of (modern) sequences that is plainly proportional to the time of their divergence (Gavryushkin et al., 2016). In ultrametric trees, the amount of change in living populations/species is related to the amount of time (Cusimano and Renner, 2014). This allows the building of molecular clocks able to correlate phenotypic evolution and genomic/molecular/proteic sequence data with branch lengths. Indeed, the molecular clock assumption suggests that the mutation rates are always constant on all the branches. This means that the mutation distances are proportional to the divergence time and equal between any pair of modern sequences and their most recent common ancestor (Page and Holmes, 1998).

Ultrametric trees have been used to build up PTs of gene flows among disparate hominins. Several tasks have been accomplished such as, e.g., the genome sequence of Neanderthals from the Altai Mountains (Prüfer et al. 2016) and Vindija Cave (Prüfer et al., 2017), the putative date of interbreeding between Neandertals and modern humans

(Sankararaman et al., 2012), the Neanderthal ancestry in Initial Upper palaeolithic European humans (Hajdinjak et al., 2021), the Denisovan and Neanderthal gene flow in Icelandic genomes (Skov et al., 2020), the ancient gene flow from early modern humans into Eastern Neanderthals (Kuhlwilm et al., 2016), the genetic evidence for archaic admixture in Africa (Hammer et al., 2011). It is noteworthy that the statistic across pairs of introgressed alleles is expected to have an exponential decay with genetic distance, since linkage-disequilibrium decays at a constant rate per generation. This means that recombination events are expected to break down the genetic segments shared by different populations/species such as, e.g., modern humans and Neandertals (Sankararaman et al., 2012). In the next chapter, we will suggest a non-ultrametric alternative to the widespread ultrametric trees, i.e., a novel PT able to assess recombination events in evolutionary times.

A NEW, CONVERGING TREE

We acknowledged that ultrametric trees are favoured by scientists to ascertain gene flows between populations/species. Ultrametric trees entail the philosophical and methodological implication that PTs are divergent. Every branch leads to a single population/species that is kept conceptually distinct from the others even in case of introgressive episodes (**Figure 2A**). Therefore, the tips (i.e., the population or the species under assessment) shall be regarded as separated sets in ultrametric trees. We propose a variant of PT in which the ultrametric assumption is partially lessened. In our new framework, the two population/species undergoing introgression give rise to a SINGLE population, so that the branching tree tends to converge in a cluster (**Figure 2B**), rather than being divergent. In mathematical terms, non-overlapping sets located at the numerous tips of divergent ultrametric trees become overlapping subsets located at the unique tip of convergent NUPT.

Given these theoretical premises, we operationally require a type of novel diagram which allows the calculation of gene flows in converging trees. Our diagram consists of a square. Indeed, all the sides and the angles must be identical to preserve translational symmetries and avoid not replicable distortions due to geometric transformations, projections, dispacements, shift operators, affine connexions, parallel transport, etc. Our square is illustrated in **Figure 2B**. The upper side corresponds to the temporal threshold of the populations/species split from a common ancestor. The left and the right sides of the square represent the two divergent population/species branching from a common ancestor (in our case, Neanderthals and non-African humans). The lower side of the square encompasses a percent scale from 0 to 100% which measures the amount of ancient human's DNA maintained in the current non-African human populations. Further, a grid portrays a temporal clock with time running from the past (top) to the present-day (bottom).

Once developed the proper square diagram, our procedure requires two further steps, depicted by the two numbered yellow circles in **Figure 2A**. The first step consists of drawing a line from the right upper vertex of the ancient humans to the 96% value of the lower side. In fact, according to the estimates, the modern exiting population of non-Africans preserves about the 96% of the DNA of the human ancestors, while the 4% is provided by the Neanderthal's DNA introgression. The operation of drawing the line leads to the formation of the angle β (see **Figure 2A**), which stands for the genomic divergence of the current humans from the Neanderthals/humans split. The second step consists of projecting the angle β to the lower side of the square, with the vertex located at the value of 96%. The side of the angle β meets the Neanderthal's left side of the square in a point corresponding to the date of the Neanderthal's DNA introgression in the archaic humans.

It is noteworthy that the lower the angle β, the closer to the present the introgression took place. Indeed, a low amount of genetic divergence between two introgressing species suggests that their merging occurred later after their split. Our method accomplishes a valuable result, i.e., the date of introgression changes according to the chosen temporal grid. **Figures 2B-C** describe this crucial notion: if the split between Neanderthals and archaic humans is estimated at 400,000 years ago, Neanderthal introgression took place about 40,000 years ago. As an alternative, if the split between Neanderthals and archaic humans is estimated at 600,000 years ago, a time of Neanderthal introgression of about 75,000 years ago is accomplished. This means that our approach has a twofold virtue. If the date of the split between two population/species from a common ancestor is known, our method allows to calculate the date of their subsequent introgression. In turn, if the date of the introgression between two population/species is known, our method allows to detect the date of their previous split from a common ancestor.

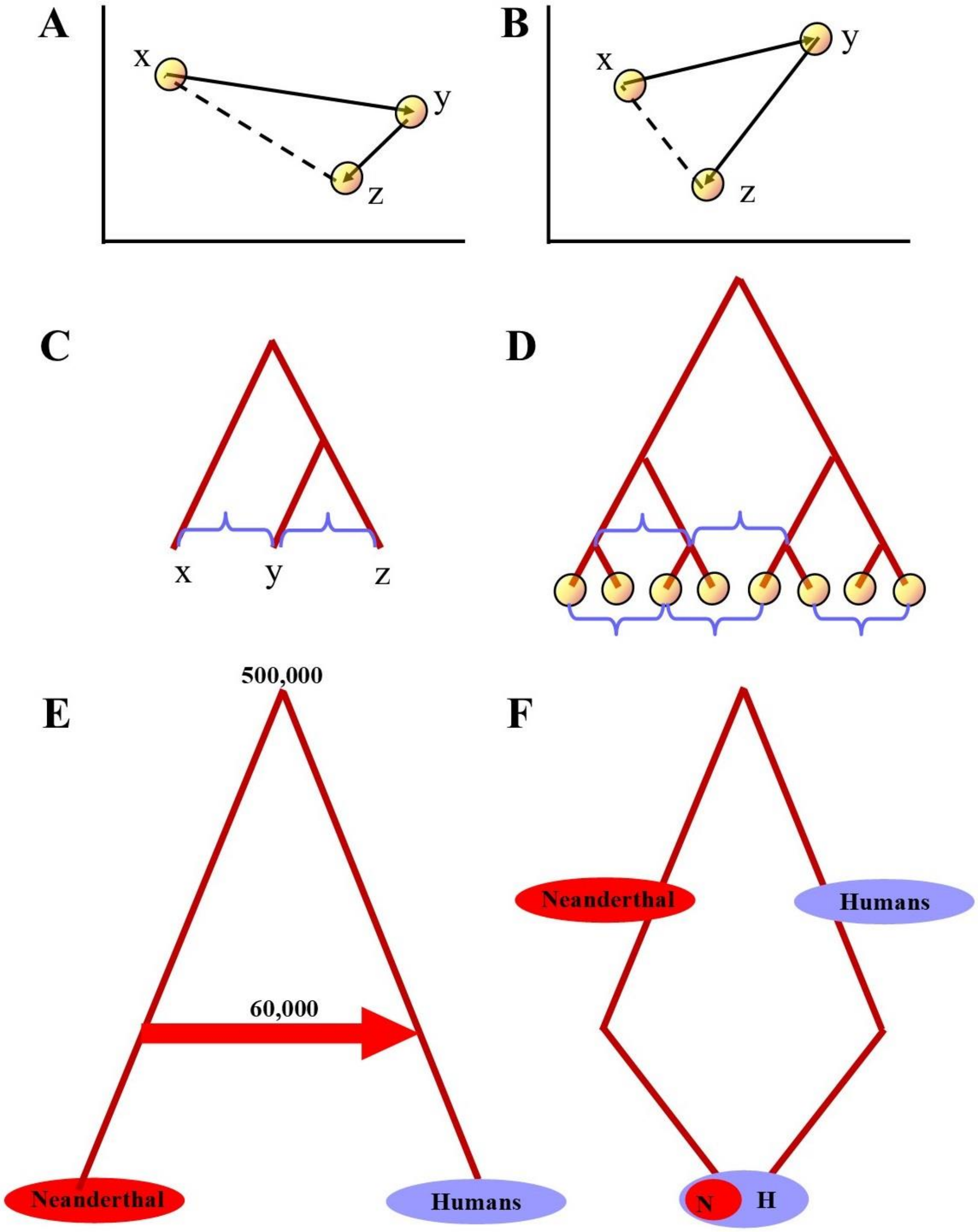

Figure 1-D. Building and arrangement of ultrametric spaces. **Figure 1A**. Triangle inequality according to the standard Euclidean geometry: every side of a triangle cannot be longer than the sum of the other two. **Figure 1B**. Strong triangle inequality. In this non-Euclidean case, just triangles with two or three sides of the same length are allowed. **Figure 1C**. An ultrametric tree is shaped as a triangle grid generated by the steps among x, y e z. Note that the distances between the branches are accurately preserved. The points x, y e z may stand not just for the vertices of equilateral or isosceles triangles, but also for other features, such as, e.g., biological taxa, populations, species. **Figure 1D**. Regular branching tree of a translational invariant ultrametric space. The population/species are represented by the yellow balls at the tips of the distal branches, while the distances among the population/species are provided by the tree itself according to the hierarchy of the balls' sets and subsets.

Figure 1E. Phylogenetic ultrametric tree of genetic flows from Neanderthals to Non-African humans (termed "humans" for sake of clarity). The dates of the split and introgression (500,000 and 60,000 years ago, respectively) are tentative and based on the current literature. If the two available sets are deemed to be separated, they can be located inside a diverging branching tree. **Figure 1F**. New NUPT of genetic flows from Neanderthals to Non-African humans. In this case, the two sets partially superimpose, giving rise to phylogenetic trees that are convergent instead of divergent. In sum, we achieve a large set of modern extant hominins which encompasses as subsets both introgressed Neanderthals and ancient humans.

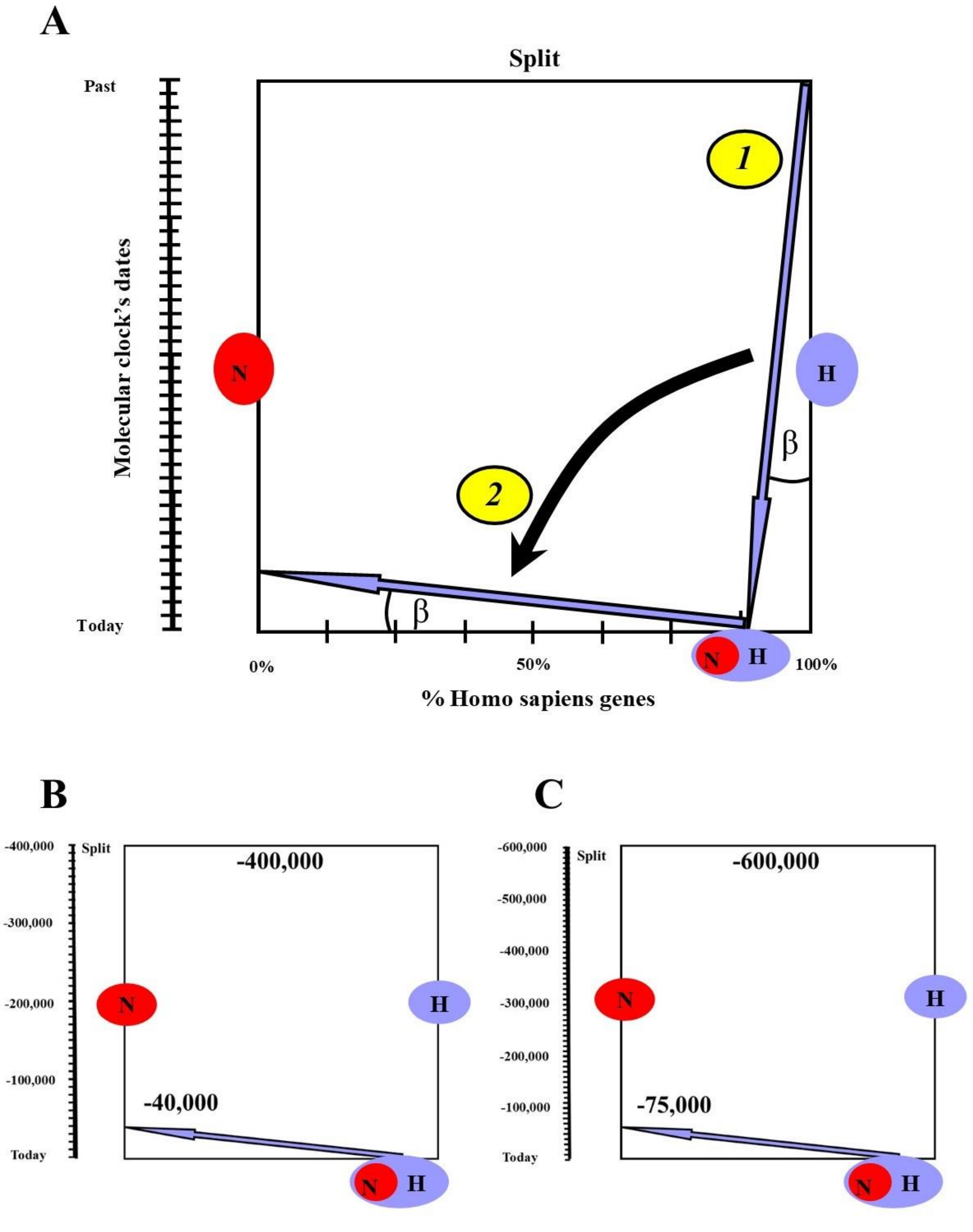

Figure 2. A novel non-ultrametric approach to detect the timing of genetic flows. After the split (upper side of the square) at a given time in the past, two branches are generated: the branch of the Neanderthals (left side of the square) and of the non-African humans (right side of the square). When the ancient human DNA is introgressed by Neanderthal DNA, the final result is a single modern population containing both the genetic materials. In the lower side of the square, the 96% of modern humans' DNA comes from ancient humans, while the 4% from Neanderthals. The arrow from the upper right vertex to the lower side of the square points towards the value of 96, giving rise to the angle β. The two numbered yellow circles illustrate the two steps of our operational procedure: for further details, see the main text.

Figures 2B-C. Molecular clocks reporting two hypothetical dates of the genetic flows between Neanderthals and humans. Note that different calibrations of the molecular clock lead to different temporal estimates of the introgressive event. Our method suggests what follows: if the split took place 400,000 years ago, the introgression occurred about 40,000 years ago; if the split took place 600,000 years ago, the introgression occurred about 75,000 years ago.

DISCUSSION

We proposed a simple procedure to build non-ultrametric phylogenetic trees and calculate the hypothetical date of interbreeding among different populations/species. Starting from two overlapping sets with a few features in common, we built a coalescent model of phylogenetic tree (Garrick et al., 2019) characterized by converging branches. We started from the premise that the widely used ultrametric trees do not describe exactly the state of matter (Balaban et al., 2019), since they favour the description of the differences between population/species, instead of their similarities. In turn, our NUPT considers the clustering of populations/species instead of their distinctions.

To provide an example, we went through the instance of Neanderthal's DNA introgression in modern non-African humans. Nevertheless, our procedure can be generalized to all the cases of genetic recombination between populations/species, ranging from introgression to hybridization, from reticulation to repeated lineage splitting and lineage fusion, i.e., the complete merging of two or more populations resulting in a single panmictic group (Garrick et al., 2014; Garrick et al., 2020). Our approach allows the prediction of the branching date between two populations/species when the amount of genetic material's introgression is known. This could be useful in the assessment of far-flung issues, such as the interbreeding of wild and domestic animal populations, the creation and the fitness of new artificial varieties of hybrid plants, the study of the evolution of SARS-Cov-2 after the split of common coronavirus ancestors from bat and pangolin strains (Touati et al., 2020).

Also, NUPT leads to intriguing theoretical consequences. The hot disputes between the Out-Of-Africa and multiregional evolution models of human evolution in the Pleistocene (Wolpoff et al., 2000; Bräuer et al., 2004; Groucutt et al., 2015; Montinaro et al., 2021) can be tackled in terms of ancient populations, instead of species. These populations are able to diverge, converge and clustering in a worldwide network of genic exchanges that leads to "modern humans", i.e., individuals characterized by local genetic differences.

ACKNOWLEDGEMENTS

The Author would like to thank his intellectual brother Norbert Jausovec, who gave precious suggestions and instructive insights. Norbert, sadly, left us a few months ago.

REFERENCES

1) Balaban M, Moshiri N, Mai U, Jia X, Mirarab S. 2019. TreeCluster: Clustering biological sequences using phylogenetic trees. PLoS ONE 14(8): e0221068. https://doi.org/10.1371/journal.pone.0221068.
2) Bräuer G, Collard M, Stringer C. 2004. On the reliability of recent tests of the Out of Africa hypothesis for modern human origins. Anat Rec A Discov Mol Cell Evol Biol. 279(2):701-7. doi: 10.1002/ar.a.20064.
3) Cusimano N, Renner SS. 2014. Ultrametric trees or phylograms for ancestral state reconstruction: Does it matter? Taxon. Vol. 63, No. 4 (August 2014), pp. 721-726. https://www.jstor.org/stable/taxon.63.4.721.
4) Garrick RC, Benavides E, Russello MA, Hyseni C, Edwards DL, et al. 2014. Lineage fusion in Galápagos giant tortoises. Mol Ecol. 2014 Nov;23(21):5276-90. doi: 10.1111/mec.12919.
5) Garrick RC, Banusiewicz JD, Burgess S, Hyseni C , Symula RE. 2019. Extending phylogeography to account for lineage fusion. Journal of Biogeography. Volume46, Issue2, Pages 268-278. https://doi.org/10.1111/jbi.13503.
6) Garrick RC, Hyseni C, Arantes IC. 2020. Efficient summary statistics for detecting lineage fusion from phylogeographic datasets. Journal of Biogeography. Volume47, Issue10, Pages 2129-2140. https://doi.org/10.1111/jbi.13932.
7) Gavryushkin A, Drummond AJ. 2016. The space of ultrametric phylogenetic trees. Journal of Theoretical Biology. Volume 403, Pages 197-208. https://doi.org/10.1016/j.jtbi.2016.05.001.
8) Gómez-Robles A. 2019. Dental evolutionary rates and its implications for the Neanderthal–modern human divergence. Science Advances. Vol 5, Issue 5. DOI: 10.1126/sciadv.aaw1268.

9) Groucutt HS, Petraglia MD, Bailey G, Scerri EML, Parton A, et al. 2015. Rethinking the dispersal of Homo sapiens out of Africa. Evol Anthropol. 24(4):149-64. doi: 10.1002/evan.21455.

10) Hajdinjak M, Mafessoni F, Skov L, Vernot B, Hübner A, et al. 2021. Initial Upper Palaeolithic humans in Europe had recent Neanderthal ancestry. Nature volume 592, pages253–257.

11) Hall BG. 2013. Building Phylogenetic Trees from Molecular Data with MEGA. 2013. Molecular Biology and Evolution, Volume 30, Issue 5, Pages 1229–1235, https://doi.org/10.1093/molbev/mst012.

12) Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD. 2011. Genetic evidence for archaic admixture in Africa. PNAS September 13, 2011 108 (37) 15123-15128; https://doi.org/10.1073/pnas.1109300108.

13) Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M. 1997. Neandertal DNA Sequences and the Origin of Modern Humans. Cell. volume 90, issue 1, p19-30. DOI:https://doi.org/10.1016/S0092-8674(00)80310-4.

14) Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338 (6104): 222-226.

15) Meyer M, Arsuaga J-L, Nagel S, Aximu-Petri A, Martínez I. 2016. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature 531, 504–507.

16) Montinaro F, Pankratov V, Yelmen B, Pagani L, Mondal M- 2021. Revisiting the out of Africa event with a deep-learning approach. Am J Hum Genet. 108(11):2037-2051. doi: 10.1016/j.ajhg.2021.09.006.

17) Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S, Pritchard JK, et al. 2006. Sequencing and analysis of Neanderthal genomic DNA. Science 314: 1113–1118.

18) Page, R. & Holmes, E. 1998 “Molecular Evolution: A Phylogenetic Approach,” Blackwell Science ISBN 0-86542-889-1.

19) Posth C, Wißing C, Kitagawa K, Pagani L, van Holstein L, et al. 2017. Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. Nature Communications. volume 8, Article number: 16046.

20) Prüfer, K., Racimo, F., Patterson, N. et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505, 43–49. https://doi.org/10.1038/nature12886.

21) Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, et al. 2017. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 358(6363): 655–658. doi: 10.1126/science.aao1887.

22) Reich, D., Green, R., Kircher, M., Krause J, Patterson N, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468, 1053–1060 (2010). https://doi.org/10.1038/nature09710.

23) Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. 2012. The Date of Interbreeding between Neandertals and Modern Humans. PLOS Genetics. https://doi.org/10.1371/journal.pgen.1002947

24) Sankararaman S, Mallick S, Patterson N, Reich D. 2016. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. Curr Biol. 2016 May 9; 26(9): 1241–1247. Published online 2016 Mar 28. doi: 10.1016/j.cub.2016.03.037

25) Santamaria R, Theron R. 2009. Treevolution: visual analysis of phylogenetic trees. Bioinformatics. 25 (15): 1970–1971. Doi:10.1093/bioinformatics/btp333.

26) Serre D, Langaney A, Chech M, Teschler-Nicola M, Paunovic M, Mennecier P, Hofreiter M, Possnert G, Pääbo S 2004. No evidence of Neandertal mtDNA contribution to early modern humans. PLoS Biol 2: e57 doi: 10.1371/journal.pbio.0020057.

27) Skov, L., Coll Macià, M., Sveinbjörnsson, G. et al. 2020. The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes. Nature 582, 78–83.

28) Soares A, Râbelo R, Delbem A. 2017. Optimization based on phylogram analysis. Expert Systems with Applications. 78: 32–50. doi:10.1016/j.eswa.2017.02.012.

29) Sokal R, Michener C 1958. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin. 38: 1409–1438.

30) Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.

31) Touati R, Haddad-Boubaker S, Ferchichi I, Messaoudi I, Elloumi Ouesleti A, et al. 2020. Comparative genomic signature representations of the emerging COVID-19 coronavirus and other coronaviruses: High identity and possible recombination between Bat and Pangolin coronaviruses. Genomics. 112(6): 4189–4202. doi: 10.1016/j.ygeno.2020.07.003.

32) Wall JD. 2000. Detecting Ancient Admixture in Humans Using Sequence Polymorphism Data. Genetics. vol. 154 no. 3 1271-1279.

33) Woese CR. 2000. Interpreting the universal phylogenetic tree. PNAS. 97 (15) 8392-8396; https://doi.org/10.1073/pnas.97.15.8392.

34) Wolpoff MH, Hawks J, Caspari R. 2000. Multiregional, not multiple origins. Am J Phys Anthropol. 112(1):129-36. doi: 10.1002/(SICI)1096-8644(200005)112:1<129::AID-AJPA11>3.0.CO;2-K. DOI: 10.1002/(SICI)1096-8644(200005)112:1<129::AID-AJPA11>3.0.CO;2-K.