Assessing the potential of ancient protein sequences in the study of hominid evolution

Ioannis Patramanis[1], Enrico Cappellini[1], and Fernando Racimo[1]

[1]Globe Institute, University of Copenhagen

April 8, 2025

Summary

Palaeoproteomic data can provide invaluable insights into hominid evolution over long timescales. Yet, the potential and limitations of ancient protein sequences to resolve evolutionary relations between species remains largely unexplored. In this study, we aim to quantify how much information about these relations can be obtained from limited ancient protein data, at the scale that is currently available or will be available in the near future. We harness sequence alignments of 12 enamel and collagen proteins that have been previously reported in fossil material that is at least 1 million years old. We utilise in silico translations of hominid DNA sequences of these proteins and highlight their differential sequence conservation, indicating some of them contain much larger amounts of information than others. We also evaluate the extent to which inferred topologies from protein data differ from inferred topologies from the more informationally-dense DNA data. We show that the former may sometimes lead to inferences of the wrong tree topology due to the informational loss that comes when working with peptide data. Additionally, we determine the number of concatenated proteins necessary to confidently reconstruct the population / species tree summarizing the relations between humans, chimpanzees and gorillas, as well as those between modern humans, Neanderthals and Denisovans. As expected, increasing the number of proteins in a concatenation enhances resolution, but we note that trees inferred from the full set of collagen and enamel proteins do not necessarily correspond to population trees inferred from genome-wide data, especially in closely related groups. Our study underscores the potential and limitations of utilising palaeoproteomic data in deep time phylogenetic reconstructions, indicating that these will be aided not only by increased recovery of proteins in the future, but also by more careful modeling of evolutionary relations across the genome, beyond simply building single phylogenetic trees.

Introduction

Understanding the evolutionary relationships between extinct humans and other hominin groups is a fundamental question in paleoanthropology. This problem has largely been approached via methodologies derived from comparative morphology, phylogenetics and, more recently, the study of ancient DNA [1, 2, 3, 4]. In the last decade, improvements in the extraction and sequencing of ancient peptides have provided researchers with yet another source of evidence to tackle this problem. Paleoproteomics is an emerging field that addresses the deep-time limitations of ancient DNA, which degrades faster than the peptides of specific proteins [5]. Thus, ancient peptide sequencing has enabled the study of evolutionary relationships between organisms that lived from tens of thousands [6] to millions of years in the past [7, 8]. It has also yielded valuable data in areas where DNA tends to be poorly preserved, due to humid and warm climates [9]. These include regions of the world such as southern Europe [10], southern Asia [11, 12] and Africa [13], which are rich in archaic hominid fossil material. Palaeoproteomics thus

holds the potential to explore questions that were previously impossible to address using morphological or DNA data alone, including resolving the species or population identity of hundreds of fragmentary fossil specimens for which limited or no DNA sequences are available. Although the study of ancient proteins holds great promise, it also harbors limitations, caused by the number of proteins that can be retrieved from ancient material, and by the nature of protein data.

While some studies have managed to recover tens or even hundreds of proteins from relatively young samples [14, 15, 16, 17, 18, 19, 20], only collagen type I, enamel-specific proteins and a few others have so far been retrieved from million years old mammalian fossil material [8, 11, 13, 21, 22]. As only a handful of protein sequences are so far recoverable, only small amount of useful genetic information can be obtained from them. The sequences of these proteins are also not complete; degradation over time breaks down the original proteins into smaller and smaller peptides. After thousands of years, many peptides do not survive this process and are therefore unrecoverable, while peptides that do, can incorporate ambiguous amino acids due to post-mortem chemical modifications [5].

Even when fully preserved, proteins inherently contain less information than DNA. This is both due to the degenerate nature of the genetic code [23], and to the fact that negative selection tends to act upon protein coding sequences more strongly than in other regions of the genome, reducing sequence variation [24]. Consequently, ortholog proteins of closely related species often show limited or no variation. Amino acid mutations may also occur in the same position multiple times in different lineages, due to functional molecular constraints [25], leading to molecular convergence or homoplasy [26]. This convergence can create the illusion of close phylogenetic affinities and has been observed in multiple taxa [25, 27, 28, 29, 30, 31]. Furthermore, when only a few proteins sequences from a given species are available, it is difficult to establish whether observed variants are fixed or polymorphic within that taxon [12, 13]. All of the above reduce the utility of amino acid polymorphisms for reconstructing evolutionary relationships.

Moreover, small sets of ancient peptides can provide very limited information about the full ancestral recombination graph, i.e. the graph structure that describes the full genealogical relationships of a set of individual genomes [32, 33]. Reconstructions based on these peptides tend to focus on individual gene trees, which in turn provide only partial knowledge about overall population relationships. This may be because gene trees are affected by incomplete lineage sorting (ILS) in ancestral populations [34, 35], or because admixture events between populations may not be represented in such trees [36]. Both ILS and admixture are of concern in species or populations that are closely related to each other [37, 38, 39, 40, 41]. In African great apes, for example, multiple studies have detected both high levels of ILS (amounting to up to 30% percent of the genome) [35, 42, 43, 44], as well as possible past admixture episodes, in all 3 extant genera [45, 46, 47, 48].

Although most of these limitations have been previously acknowledged [10, 13, 49, 50], little quantitative work has been done to assess the potential and limitations of ancient protein data at resolving population relationships [51, 52, 53, 54]. In this study, we specifically focus on 12 collagen and enamel proteins and their ability to resolve evolutionary relationships between species. We focus on these proteins because they have been previously recovered from biological material older than a million years. We center our analysis on the Hominidae family and its well-studied genetic history, based on DNA evidence [55, 56, 57, 58], and leverage the high availability of genomic and protein data that exists for all four of its extant genera and some extinct populations [55, 59, 60, 61]. Lastly we investigate whether the recovery of peptides from a richer proteome, such as that of dentin or bone, would enhance evolutionary resolution (and if so, how much).

First, we measure the entropy and evolutionary conservation rates of these 12 proteins, using hominid alignments, ranking them based on the amount of information they provide and comparing them to other known conserved proteins. We further use the entropy metric to measure the informational loss that occurs when comparing intron-containing DNA alignments, to exon and to protein alignments of the same genetic locus. We additionally evaluate how inferred topologies differ between these alignments of different data type.

While ILS leads to topological mismatches between gene trees and the population tree, topological mismatches can also occur between the true and the inferred gene tree at a given locus, either when using DNA or protein sequences (see figure 1). DNA gene tree misinference can occur due to the inherent difficulties of reconstructing topologies from mutations, as well as from the information loss and errors that can take place during the sequencing process. Protein gene tree misinference, can occur due to the same reasons, but also the additional information loss

that occurs in protein translation. Local tree reconstructions are also necessarily affected by difficulties in inferring recombination events and resulting topological changes along the genome. For simplicity in our empirical analysis, and given our focus on information loss in ancient peptides specifically, we assume that gene trees reconstructed from DNA sequences are accurate and use them as a proxy for the true gene trees. We also assume that the sequences we study are small enough to be characterized by a single gene tree. However, we note that both of these are strong assumptions that might not hold in reality.

We use an iterative analysis to compare protein-inferred phylogenetic trees with established tree topologies based on previously published genetic data. We estimate the number of combined proteins and amino acid variants required to reliably infer the population trees of 3 hominid genera and 3 hominin populations. We then repeat this iterative analysis by supplementing our 12 initial "deep time" proteins with 16 additional proteins, which have been experimentally recovered from bone or dentin [62, 63, 64], albeit so far, only from samples that are relatively younger.

Materials and Methods

Incomplete lineage sorting, DNA and proteins

We first selected 12 proteins that have previously been recovered from either tooth enamel (AHSG, ALB, AMBN, AMELX, AMELY, AMTN, COL17A1, ENAM, MMP20, ODAM) or bone material (COL1A1, COL1A2), from mammalian samples that are more than 1 million years old [8, 11, 13, 21, 22]. We then acquired the reference sequence for each of the proteins from Ensembl [65], for the following four hominid species: *Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla*, *Pongo abelii*. We aligned the ortholog sequences from the four hominid species using Mafft [66] and reconstructed gene trees with PhyML [67], using each of the 12 ortholog alignments separately. We rooted the 12 generated trees using *Pongo abelii* as the outgroup and compared them to the population tree that best represents the relationships between those 4 species [56]. To compare our protein tree results, we repeated the same process but using the reference DNA sequences (combined exons and introns) of the genes corresponding to the 12 proteins instead of the amino acid sequences.

Entropy and evolutionary conservation rates

To assess the conservation levels and the phylogenetic information of these 12 proteins, we calculated Shannon's information theoretic entropy [68] and an evolutionary rate score [69] for each amino acid position on the sequence alignments. We used these two scores as approximations to the information content in these alignments, while respectively ignoring and accounting for the evolutionary distance between each sequence. We used Bio3d [70] for the entropy calculation and Rate4Site[69] for the evolutionary rate computation on the alignments of the four hominid species. We obtained a score for each position of each of the 12 protein alignments. We aggregated the metrics across each protein to obtain a total score, and also divided them by the length of the alignment, to obtain a sequence-wide average score. To account for within-species diversity, we used multiple individuals as representatives from each of the four species, using previously published translated proteomes [71]. We randomly sampled 1 individual from each of the four species, repeated the calculations for 1000 repetitions and calculated the mean from all 1000 repetitions. To contextualize our results, we selected 5 proteins that have been previously reported as being either highly conserved or as containing hyper-variable segments [72] and included them in the entropy and evolutionary rate calculations. These included two highly conserved histones (H2BC3, H2BC9) and one ubiquitine (USP46) [73, 74], as well as two fibrinogen proteins (FGB, FGG) reportedly bearing a highly variable segment [72].

Informational content: exons, introns and proteins

We assessed the differences in informational content between a DNA sequence containing both introns and exons, a version of the same sequence containing only exons, and a peptide version of the sequence (translated amino acids)

for all 12 loci of our analysis. We applied Bio3d's entropy scoring to all three data types of the same gene, for all 12 genes, and then ranked the results (see Supplementary Material for details). We also divided each entropy metric by the length of the data type to compare the average information content per site, of each data type. Due to sequence length differences between DNA and protein data (with a 3 letter DNA codon corresponding to a single protein amino acid), we also applied a "length-correction" to this last measurement. In this correction we divided the entropy of each protein version of each gene by 3 (number of nucleotides that correspond to an amino acid), while keeping the introns-and-exons and exons-only version unaltered.

Iterative phylogenetic analysis

We investigated how the concatenation of different numbers and different combinations of the 12 proteins might affect the topology of the inferred "consensus" tree, which is often taken as an estimate of the population or species tree. For this analysis we utilised a "hominid dataset", consisting of *Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla* and *Pongo abelii* (as an outgroup) and a second "hominin dataset" consisting of *Homo sapiens*, Neanderthals, Denisovans and *Pan troglodytes* (outgroup). To assess how the recovery of additional proteins, from different tissues affects phylogenetic analyses, we expanded the "hominin dataset", creating a third "bone-dentin dataset". This dataset consisted in the protein sequences that are most often recovered from dentin or bone tissue. In choosing which proteins to include in this anaysis, we utilised the list provided by Ruther et al. 2022 [62], which includes 20 proteins utilised in species identification: COL1A1, COL1A2, COL2A1, COL3A1, COL4A4, AHSG, COL5A2, ALB, BGN, COL5A3, COL5A1, CHAD, COL22A1, COL11A2, SERPINF1, F2, COL11A1, LUM, COL12A1, POSTN. Four of these 20 proteins (COL1A1, COL1A2, AHSG and ALB) were already included in the original 12 proteins, leading to a final combined dataset of 28 proteins.

In each iteration of this analysis, we carry out a concatenation using a subset of proteins sampled from the full set of proteins, reflecting the fact that not all proteins in the full set might be available in practice. The subset ranges in size from 1 (a single protein recovered) to all proteins recovered (either 12 or 28, depending on the tested dataset). One representative individual per population or species is randomly chosen and included in the alignment, as ancient protein studies are often limited to single individuals that are made to represent an entire species. For each concatenation, we build a phylogenetic tree and record the resulting topology. We then compare it to the underlying population tree, as inferred from past DNA studies. In total, we do this over 1,000 iterations per each number of proteins, sampling different sets of proteins and different representative individuals, in each turn.

We performed the same iterative analysis on each of the three datasets ("hominid","hominin","bone-dentin"). The analysis for the "hominid" and "hominin" datasets was repeated 1000 times for each N, with N ranging from 1 to 12, resulting in a total of 12,000 generated trees for each of the two datasets. The same process was applied to the bone-dentin dataset, with N ranging from 1 to 28, resulting in a total of 28,000 trees. For each iteration, we first picked N proteins, without replacement, out of the maximum number of proteins for that dataset. For each protein, one sequence from each of the four taxa was randomly selected from the samples available to us [75] and then the four orthologous sequences were aligned using Mafft. Each iteration (out of a thousand) thus generated a total of N protein alignments. The N alignments were then concatenated into a single alignment which was used to generate a phylogenetic tree using PhyML. The generated tree was also trimmed for very short and unsupported branches (see Supplementary), which were transformed into polytomies. The tree was then rooted using an outgroup taxon (*Pongo* for hominid set, *Pan* for hominin and bone-dentin set) and compared to a model reference tree. The model reference tree is a simple 4-leaf tree that best describes the relations between the 4 taxa [57, 76]. In the case of the hominid set, the reference tree has the *Pan* and *Homo* nodes as the most closely related, followed by *Gorilla* as an outgroup to *Homo-Pan*. For the hominin set, the Neanderthal and Denisovan are the most closely related sister groups, with *Homo sapiens* as the outgroup to the Neanderthal-Denisovan clade.

For each comparison of a generated tree with the model reference tree, we assigned a label ("Topology #1", "Topology #2", "Topology #3" and "Topology #4"), each corresponding to one of the four possible topologies (including a polytomy). For both sets of taxons, the four topologies and their matching labels are shown in figure 5 and figure 6. We recorded the bootstrap support value of the node with the two most closely related

taxons of the generated tree, excluding polytomies. Additionally, we enumerated the number of variant sites in the concatenated alignment (excluding informative sites of the outgroup taxon) that were used to generate each tree. All protein sequences for the iterative analyses were acquired from the "Hominid Palaeoproteomic Reference Dataset", available on Zenodo [75]. All alignments, concatenation and phylogenetic trees were generated using Module 2 of PaleoProPhyler [71]. All downstream comparisons after generating the trees were done using scripts deposited on Github (see Supplementary Material).

Results

Incomplete lineage sorting, DNA and proteins

We inferred gene trees from each of the 12 loci of interest and observed that different topologies were recovered, depending on which data type we utilized (figure 3). For the protein set, 5 out of the 12 gene trees displayed an estimated topology that was different from the population tree (as inferred from genome-wide data[57]). In contrast, when utilizing the DNA sequences, corresponding to those 12 genes, only 2 out of the 12 gene trees differed in topology from the population tree. In all cases except two, when the protein data of a locus supported a topology different from the DNA data, the DNA data of that same locus supported the population tree topology. The two exceptions are: a) AHSG, where both DNA and protein data supported an alternative topology, which also differed from each other, and b) ENAM, where the DNA data supported an alternative topology, while the protein data supported the population tree topology.

Entropy and evolutionary conservation rates

We observe notable differences between the entropy levels and evolutionary rate scores of the proteins in question (figure 4). While the aggregated entropy and evolutionary rate rankings showed slight differences, when accounting for the length of each protein, both metrics showed a nearly identical arrangement of the proteins. When not accounting for protein length, ENAM, followed by COL17A1, were found to contain the highest amount of entropy as well as the highest evolutionary rate scores. When the entropy and evolutionary rate scores were divided by the length of each protein, however, ODAM was found to be the most variable, while proteins like AMELX, COL1A1 and COL1A2, fell on the lower end of the spectrum. When compared to known conserved proteins, almost all proteins showed substantially higher entropy and evolutionary rate scores than the ubiquitin and histone proteins that we compared them with, while the fibrin proteins fall within the range of the enamel proteins.

Informational content: exons, introns and proteins

When utilizing the entropy metric to examine the different data types, we observed that information decreased when comparing segments of mixed introns and exons to segments of pure exons and to those of amino acids (Sup. figure 1-A,B). While the difference in entropy from the mixed intron & exon data to pure exons was substantial, the difference between the exon and the amino acid dataset was much smaller in comparison. When we normalized each entropy measurement by the length of the sequence that generated it, this drop in information content did not hold. In certain cases the mixed intron & exon dataset contained less entropy than the pure exon dataset. In other cases, the amino acid alignment was the one with the highest entropy per site (Sup. figure 1-C). Differences associated with each data type's length (triplets of nucleotides corresponding to single amino acids) may be influencing these results. To correct for this, we divided the average per-site information of the introns & exons and pure exons (seen in Sup. figure 1-C) by 3, while keeping the protein measurement intact, in order to normalise this difference in datatype length. After applying this "length-correction" to the normalised entropies, we once again observe the original pattern of informational decrease for all genes (Sup. figure 1-D), with the exception of AMELX. For AMELX, the pure exon dataset showed more information content than the mixed intron & exon dataset, even after the length-correction.

Iterative phylogenetic analysis

Our consensus tree analysis yielded different results for the three different datasets it was applied to. For the hominid dataset, as the number of proteins and variants utilized in the concatenation increased, the number of consensus trees agreeing with the population tree topology (topology #1) increased as well, in an almost linear fashion (figure 5). At a lower number of concatenated proteins (N between 1-6), even if the majority of the consensus trees generated did agree with the population tree topology #1, large percentages of alternative topologies (#2, #3, #4) were also inferred. As an example, at N=1, 40% percent of the consensus trees had an estimated topology that was different from the population tree (though note here that the "consensus" is just a single gene tree for N=1). Topologies #2 and #3 were inferred in 20% of the iterations, with the other 20% corresponding to topology #4 (uninformative polytomy). This percentage of alternative topologies steadily decreased as N increased. At N=9, less than 5% of the consensus topologies differed from the population tree topology. When examining each of the discordant consensus topologies, unresolved polytomies between the 3 African great apes (topology #4) were represented at lower Ns (N=1-3) but were absent at higher Ns than 4. Above N=10 almost all consensus topologies converged into Topology #1, the topology in agreement with the genome-wide DNA-inferred population tree. The number of informative sites (amino acid variants) that were used to generate each tree ranged from a mean of around 9 variants for 1 protein, to a mean of 113 variants for 12 proteins.

In contrast to the hominid dataset, increasing the number of proteins in the concatenation of the hominin dataset did not lead to an overall convergence to the genome-wide inferred population tree topology (topology #1) (figure 6). Instead, the number of trees supporting topology #1 remained roughly stagnant past the N=5 mark, while the trees supporting topology #2 (Neanderthal and modern humans as sister lineages) steadily increased. Furthermore at N=1, roughly 80% of the trees supported topology #4 (the polytomy). Although the percentage of topology #4 trees steadily dropped as the number of proteins increased, it did not completely disappear even with the use of 12 proteins (with 10% of iterations still leading to topology #4). For this dataset the number of variants was roughly 10 times lower than that of the "hominid" and ranged from a mean of 1.2 variants for 1 protein, to 14.3 for 12 proteins.

When including additional proteins (the dentin-bone dataset), we observed that the greater set of available proteins decreased the proportion of trees supporting topology #2 and increased the proportion supporting topology #1. In this more protein-diverse dataset, a total of 12 proteins supported either of these two topologies (#1 and #2) with roughly equal representation. Further increases in the number of proteins did not appear to change this proportion, but we did observe a continual decrease of the cases where topology #4 (polytomy) was inferred. This polytomy effectively disappeared at around 20 proteins or a mean of roughly 30 variants. At 28 proteins, or roughly 40 variants, both topology #1 and topology #2 were still equally supported.

To better understand the relationship between the results of our analysis and the confidence metrics generated by the phylogenetic software itself, we extracted and plotted the bootstrap support of each tree from all 3 datasets. We grouped the bootstrap support scores according to the data set, the number of proteins used to generate them, and the tree topology they supported. We then plotted them as boxplots using python's 3 Matplotlib [77] package, selecting the option to not plot outliers for visual clarity.

In the hominid data set, we observe that the median bootstrap support of topology #1 steadily increases, from a median of 75% (N=1) to more than 90% (N¿8), as we increase the number of proteins (figure 8). The median support of topology #2, while initially very similar to the support for topology #1, rapidly drops with the addition of more protein sequences. Topology #3, already being weakly supported when using a few proteins, drops even further in support with the addition of more proteins.

For the hominin dataset, although a few trees supported topologies #1 and #3 when using a single protein, their median bootstrap support is 0, while topology #2 presented a median of 75% (Sup. figure 3). This median support for topology #2 slightly drops as the number of proteins increases, but remains roughly the same (support is 69% when using all 12 proteins). Topology #3 shows very low bootstrap support (¡60%), dropping further when higher numbers of proteins were used. Finally topology #1, the topology in agreement with the population tree, although initially unsupported, displayed similar bootstrap supports as topology #2 when using more than 3 proteins. Although the bone-dentin dataset utilized a larger number of available proteins (n=28) the resulting

bootstrap support remained roughly stagnant: Topologies #1 and #2 show almost identical distributions of values, with a fluctuating median of 70%, while topology #3 varies between 45% and 55% (Sup. figure 4).

We conclude that high bootstrap support is a metric indicative of the robustness of the underlying protein data in supporting a topology. This is based on the fact that high bootrap support (¿90%) was generated only in cases where the number of proteins in the alignment consistently supported the reference population tree.

Discussion

Palaeoproteomic data hold immense potential for the study of hominid evolution. Ancient proteins are already being used to track the distribution of different species through space and time [63, 64, 78, 79, 80, 81] and to assess the taxonomic placement of specimens [10, 11, 12]. Yet so far, only a few studies have attempted to investigate the phylogenetic potential of the various recoverable ancient proteins. Buckley et al. 2014[51] and Froment et al. 2021[52] have previously approached this question experimentally, by sequencing bone and tooth proteomes respectively, and identifying which proteins and peptides were recoverable from the investigated tissue. They used these recovered peptides to reconstruct phylogenetic trees of different taxa, inspecting those phylogenies for accuracy (by comparing the resulting topology with known species relations), confidence (by examining the bootstrap support of the trees) and resolution (by enumerating generated polytomies). The informativeness of those proteins was also assessed, by counting the number of species-informative variants. More recently, Fong et al. 2025 [53] delved deeper into this question, using a combination of dry and wet-lab methodologies. In this work, missingness was introduced into in-silico predicted enamel protein sequences, based on experimentally observed patterns of degradation. The in-silico degraded sequence alignments were then used to generate phylogenetic trees, assessing their topological distance from a DNA-supported tree. Lastly, in a recent preprint, Codlin et al. 2025 [54] translated avian egg-shell and collagen type I proteins, assessing their phylogenetic resolution and accuracy, but also highlighting the existence and possible effects of within-taxon amino acid variability.

The above studies have demonstrated some of the potential - as well as the limitations - of using palaeoproteomic data for reconstructing evolutionary relations. Questions centered around the scarcity of data, as a result of post-mortem degradation, have understandably stood at the forefront. Yet, the works described above have either focused on individual proteins or groups of proteins as a whole. In our work, we acknowledged the fact that researchers rarely have the option of choosing which proteins to utilise in their analysis, as this is usually simply the result of which peptides are recovered. As a result, we investigated the effect of incremental additions of protein data on the confidence and topology of the generated trees, exploring all possible numbers and combinations from a group of proteins. We also explored how well these preserving proteins perform in reconstructing the relations of very closely related and recently admixed populations. While more challenging, these relations are often of great interest to evolutionary biology and to hominid evolution in particular.

Protein informational reduction and incomplete lineage sorting

Multiple studies have previously identified high levels of ILS among great apes. For example, up to 30% of gene trees in a comparison between humans, chimpanzees and gorillas have been shown to be under ILS [35, 42, 43]. Yet, the proportion of sequences under ILS drops significantly when investigating coding regions exclusively. Moreover, gene families displaying high levels of selection are even less likely to be undergoing ILS [44]. Of the 12 genes investigated here, only 2 of them (16%) showcased apparent ILS (as inferred from DNA sequences) figure 3. As the entropy and evolutionary rate results showcased, some of these genes are under high constraint due to selection, which may be the main driver for the relatively low level of observed ILS among them. However, when the corresponding proteins were investigated, 5 out of 12 (41%) led to the inference of a topology that was different from the population tree. A likely explanation of this higher level of observed ILS when analyzing proteins is misestimation of the underlying gene trees, as a consequence of the reduced informational content of proteins relative to DNA. This "informational drop" between DNA and protein data types is supported by our entropy calculations, although admittedly, the question of "how much information" is contained within each data type, largely depends on how one measures said information: absolutely, per individual site or per individual codon (See

supplementary material section 2.1). These results suggest that inferences based on protein data alone, may lead to higher apparent levels of ILS than what is inferred when working with DNA data. Consequentially, protein-based phylogenetic inferences may distort evolutionary conclusions and overestimate the true amount of ILS present between taxons. Previous publications on modern proteomes have made similar arguments, based on the fact that closely related species showcase 'apparent molecular convergence' due to inference errors from very conserved sequences [26].

Enamel and collagen conservation

Our combined entropy and evolutionary rate results indicate that the hominid enamel proteome consist of proteins of variable levels of conservation and information content (figure 4). All protein sequences, including Collagen type I, appear more variable than the hyper-conserved ubiquitin and histone protein sequences that were used for comparison. When accounting for protein length, although some proteins like AMELX appear very conserved, all other enamel-related proteins yield greater average site variation, within hominids, than collagen type I proteins do figure 4. This pattern holds regardless of whether one measures variation through entropy or through evolutionary rate scores. As an example, ODAM displays more than 10 times the entropy score per amino acid than COL1A1 or COL1A2 and roughly three times their evolutionary rate score. These results are in line with the knowledge that collagen genes are heavily conserved in humans [82, 83]. They also agree with Codlin et al.[54], who showcased a higher conservation rate in avian collagen sequences compared to eggshell proteins, and Krueger et al.[53], who also noted the high conservation of collagens compared to enamel proteins in primates. A notable exception to this is AMELX: while being the most abundant enamel protein [84], it shows a similarly conserved sequence to collagen type I. Yet, while AMELX tends to display low sequence variation, its Y chromosome isoform, AMELY, is much more variable, especially when accounting for its short length. This should not be unexpected: AMELY is located on the non-recombining region of the Y chromosome where it is evolving under less selective constraint and under a faster local mutation rate [85, 86, 87, 88].

Our work is an initial investigation into the informativeness of AMELY, a protein that has recently become of great interest [89, 90, 91, 92, 93, 94, 95] due to its ability to identify the biological sex of heavily degraded samples [96, 97, 98]. Although it is expressed at lower concentration than AMELX [99], AMELY can provide useful information for species identification and phylogenetic inferences, due to its high variability. However, analyses based on AMELY are also fraught with difficulties. For most vertebrate species, even when the sequence of AMELX is well characterized, the amino acid sequence of AMELY and the location of its coding gene are unknown [65, 100]. In some taxonomic groups, the gene responsible for expressing AMELY is missing in its entirety [101]. Finally as noted by Fong et al. 2025 [53] (who for practical reasons chose not include AMELY in their enamel protein investigation), in some taxons the genes of AMELX and AMELY are not acting as independent loci [102, 103], limiting their phylogenetic utility. As a result, while our anaysis here showcases the high informativeness of AMELY - in contrast with that of AMELX - we recommend caution when working with this protein for evolutionary inferences.

Today, protein-based archeological species identification primarily relies on collagen (type I) mass fingerprinting [5, 104]. Our results indicate that, as a whole, the enamel proteome evolves faster than collagen type I and thus, when the appropriate tooth tissue is available, could differentiate between more closely related populations or species. This is especially important given the micro-destructive techniques, such as acid etching, that have successfully been applied to tooth enamel and bone material [105, 106, 107, 108]. These methods can extract useful amino acid sequence information using a minimal amount of material, inflicting only minor surface damage but preserving morphological information.

Number of proteins and phylogenetic resolution

The 12 proteins investigated here have previously been used in different combinations and have been shown to discriminate between the 4 extant genera of the hominidae family [10, 11, 13]. Nevertheless the exact number of proteins required to reliably infer relations between these species is not yet clear. Our iterative analysis on

the hominid dataset showed that, as one might expect, the number of consensus trees supporting an alternative topology to that of the population tree, drops significantly with the inclusion of additional proteins in the analysis. An increasing number of proteins sees an overall drop in discordant trees, in a linear decay, up to around 9 or 10 proteins, which we conclude to be sufficient for consistently recovering the reference population tree. Additionally, a combination of any two of the twelve proteins drops the percentage of polytomies from the population tree from 30% (when using only a single protein) down to less than 20%, and a combination of four protein, to less than 1%. Thus, simply distinguishing between these 3 groups (without accurately inferring their phylogenetic relationships), should be possible with the recovery of around 4 of any of these proteins. We expect other taxa with similar genetic distances, as those between the extant genera of hominidae, to have similar phylogenetic resolution and power using palaeoproteomic data.

For our hominin dataset instead, our analysis revealed that the number of consensus trees agreeing with the topology of the population tree, did not increase with the inclusion of additional protein sequences in the analysis. Increasing the number of proteins from this set led to an increase in support for one of the 3 alternative topologies, topology #2, the one with modern humans and Neanderthals as sister lineages. Indeed, this one particular topology (itself discordant relative to the population tree) is supported by the fully concatenated 12-protein dataset we chose for this analysis (also shown previously by Welker et al. 2020 [10]). The inclusion of additional proteins from the bone-dentin dataset led to slightly different results. Instead of a single topology that doesn't match the population tree becoming increasingly supported, two competing topologies (topology #1 and topology #2), equally represented, become the most commonly observed when using between 10 and 28 proteins. Similarly, a polytomy between these three groups was still present when using up to 20-24 proteins, although admittedly in extremely low frequency. Our analysis reveals that when examining very closely related populations with limited protein data, the addition of a few more protein sequences may not always increase support for the tree that is closest to the population tree of said taxons, as inferred from genome-wide data. In such cases, trees with medium to high bootstrap values (70-80%) may be obscuring incongruence in the underlying data. Likewise, distinguishing between these groups, a process necessary for species or population identification, is also a difficult task, largely dependent on recovering some of the few informative sites that exist.

Closing remarks and future prospects

Currently, enamel and collagen type I proteins remain the only phylogenetically informative biomolecules that are recoverable for fossil taxa in deep time (samples that are more than 1 million years old). Although this unique resource is unparalleled in terms of preservation[8, 21], its phylogenetic potential may be more limited than previously thought [10]. In homininds, many studies have already noted a lack of resolution at finer taxonomic levels: unresolved polytomies generated using the enamel proteome have been identified inside the genus of *Pongo* (between the 3 extant species)[11, 12], the genus *Gorilla* (between the 2 extant species), as well as within the genus *Homo* (between present-day humans, Neanderthals and Denisovans) [6, 10, 81]. Polytomies have also been observed at a subspecies level, such as the divisions between subspecies of *Gorilla gorilla* and of *Pan troglodytes*. Nevertheless, the two species of the genus *Pan* (*P. troglodytes* and *P. paniscus*) can be confidently distinguished from one another using a concatenation of enamel proteins [10, 11]. The reason why some evolutionary relationships are easier to resolve than others needs to be further investigated, but probable causes include differences in split times, differences in the effective population size of ancestral populations and different levels of post-divergence migration [38].

Both results from previous publications and the present study suggest that phylogenetic analyses of archaic hominid taxa based on palaeoproteomic data should be taken with a degree of caution. Overall, protein data may lead to higher amounts of gene tree misestimation, as a result of the data type used for tree estimation. Here we have shown that the number of currently recoverable, deep-time proteins allows for the reconstruction of species relations at the level of genera in the hominid clade. This is very encouraging, given this particular clade's genetic history of recent splits, high levels of ILS and past admixture events. However, our results also indicate that the same data have limited power to resolve the population trees of more closely related groups, such as those within the hominin clade.

Increases in overall ancient peptide acquisition through novel lab methodologies [63, 109, 110, 111] may lead to more confident phylogenetic placements and enhanced evolutionary resolution for these taxa. Studies extracting the bone proteome of younger samples have so far delivered a greater number of proteins [14, 15, 17, 18, 20]. Future advances, such as targeted proteomic approaches [94, 112], the identification of better preserving bones [113] or improvements in downstream spectra identification [114, 115] may allow for similar recoveries in older samples. However, we believe that an even greater number of proteins than the ones investigated here (n=28) will be necessary for the accurate resolution of evolutionary relations for very closely related populations or species. Alternatively, more computationally-intensive methods of inference beyond concatenation, such as the multispecies coalescent (MSC), could result in better resolution by better accounting for the evolutionary processees that lead to different gene trees along a sequence [116, 117]. Given that the protein data examined here features high incongruence and a low number of informative loci, it is possible that tools like *Beast [118] might provide results that better agree with the evolutionary relationships inferred from DNA.

On top of this, the population tree itself may be a poor representation of the overall relationships between closely related groups, due to admixture events [45, 59, 60, 61]. Indeed, when investigating the relationships between organisms that are as closely related as the ones investigated here, concepts such as 'species' or 'trees' lose some of their utility. A growing body of work from the field of ancient population genetics has shown that admixture between even distantly related groups of hominids might be the standard rather than the exception [46, 47, 119]. This seems to be especially true within the confines of hominin evolution during the Late Pleistocene [4, 120, 121, 122]. In light of these discoveries, the field of paleoanthropology is also changing. Past quests for a single population tree are now slowly being replaced by the concept of a 'braided stream', a network of reticulating lineages that can split as much as they can merge [3, 122, 123]. Reconstruction of hominin evolutionary relations using other topological objects beyond trees (like admixture graphs) is still in its infancy, and so far, non-existent in the field of paleoproteomics. This, in turn, suggests a fruitful avenue for future methodological developments.

Acknowledgements

We thank Alberto John Taurozzi, Evan Irving-Pease, Graham Gower, Martin Petr, Johanna Krueger, Ryan Sinclair Paterson and other members of the Racimo and Cappellini groups, who provided valuable help, suggestions and feedback throughout the project.

Funding

The project was funded by the European Union's EU Framework Programme for Research and Innovation Horizon 2020, under Grant Agreement No. 861389 - PUSHH. F.R. was also supported by a Novo Nordisk Fonden Data Science Ascending Investigator Award (NNF22OC0076816) and by the European Research Council (ERC) under the European Union's Horizon Europe programme (grant agreements No. 101077592 and 951385). E.C. was additionally supported by the European Research Council (ERC) through the ERC Advanced Grant "BACKWARD", under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101021361).

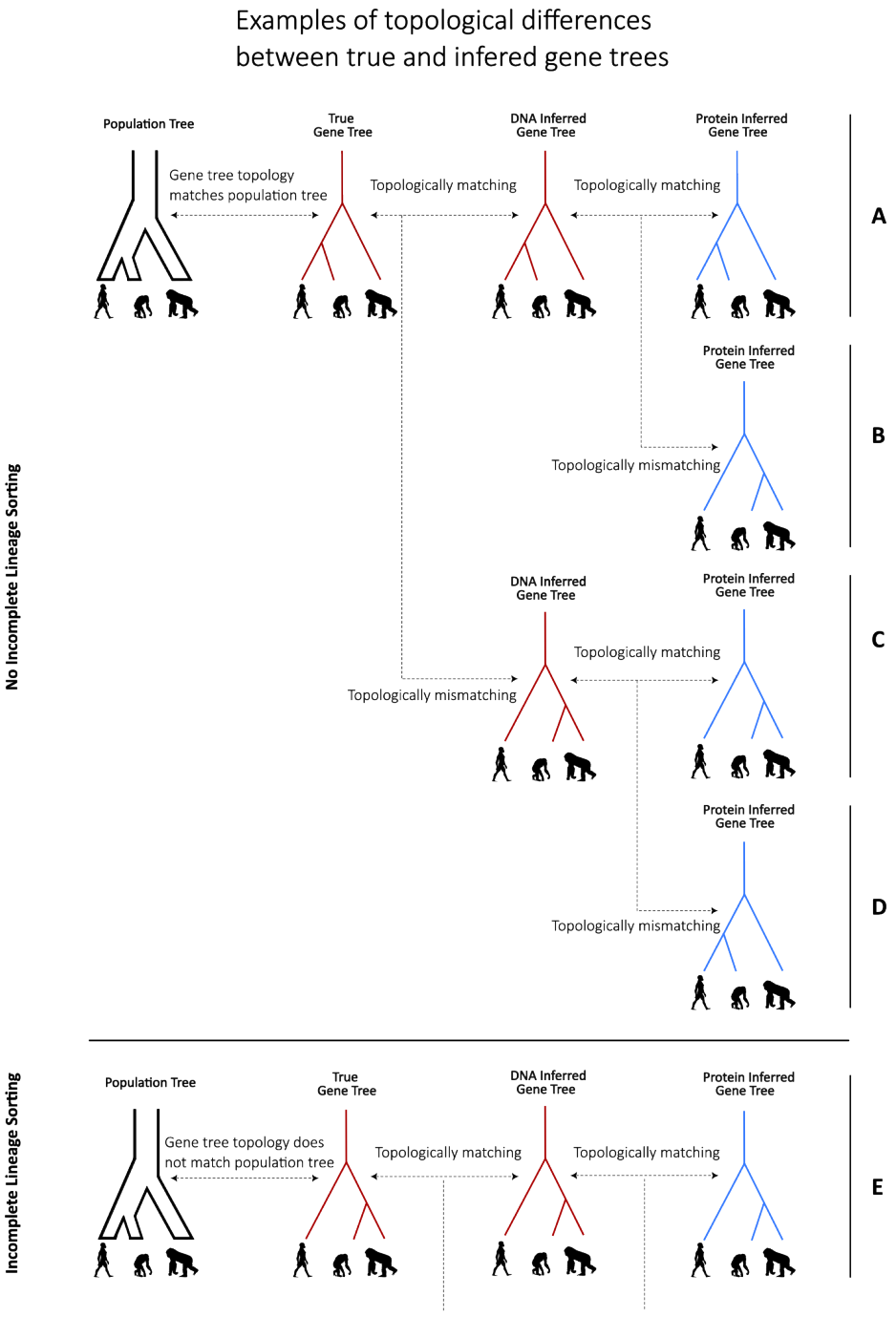

Figure 1: Examples of topological differences between true and inferred gene trees

Figure 2: Schematic explaining different scenarios of tree topology discordance and concordance. The first column depicts the evolutionary relations between three different species (humans, gorillas and chimpanzees), as described by a simple population tree. The second column depicts the tree representing the genetic relation between 3 homologous segments at a specific locus of genomes obtained from these three species (a "gene tree"). The third column depicts a tree inferred using DNA sequence data, obtained from said homologous region. The fourth column depicts the gene tree inferred using protein data from the same region. We show a (limited) set of the many possible scenarios that can arise when comparing all the above trees with one another: **A.** No incomplete lineage sorting and all four trees agree with one another: the use of either DNA or protein data leads to the correct inference of the gene tree, which happens to agree with the population tree. **B.** No incomplete lineage sorting, but the protein gene tree differs from the true gene tree at the locus under study (e.g. due to the reduced information contained in peptide sequences, relative to DNA). **C.** No incomplete lineage sorting, but both the DNA- and protein-reconstructed gene trees are misinferred (e.g. due to very few genetic variants present at both the DNA and protein levels for correct tree resolution). **D.** No incomplete lineage sorting, but while the DNA-reconstructed gene tree is misinferred, the topology of the protein-reconstructed tree happens to match the true gene tree. **E.** There is actual incomplete lineage sorting (mismatch in topology between the true population tree and the true gene tree under study). In this specific case, the gene tree is also correctly inferred using either DNA or protein data. Other scenarios involving incomplete lineage sorting but misinferred topologies from either DNA data, protein data or both have been omitted for brevity. All silhouette images are reused from https://www.phylopic.org/.

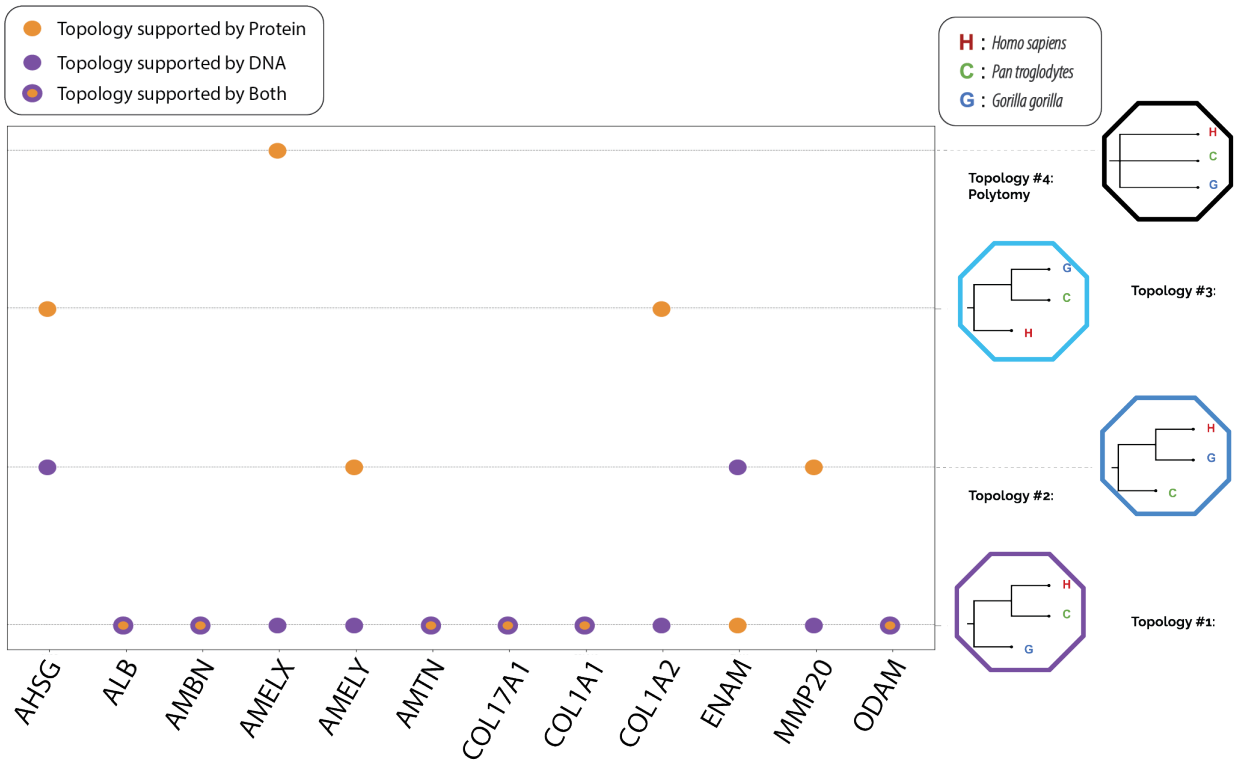

Figure 3: Comparison of topology supported by the 12 proteins (yellow points) under investigation and their corresponding DNA data (purple points).The 4 possible topologies are visible on the right side. For 6 loci (yellow points with purple circle), both DNA and protein data support the same topology.

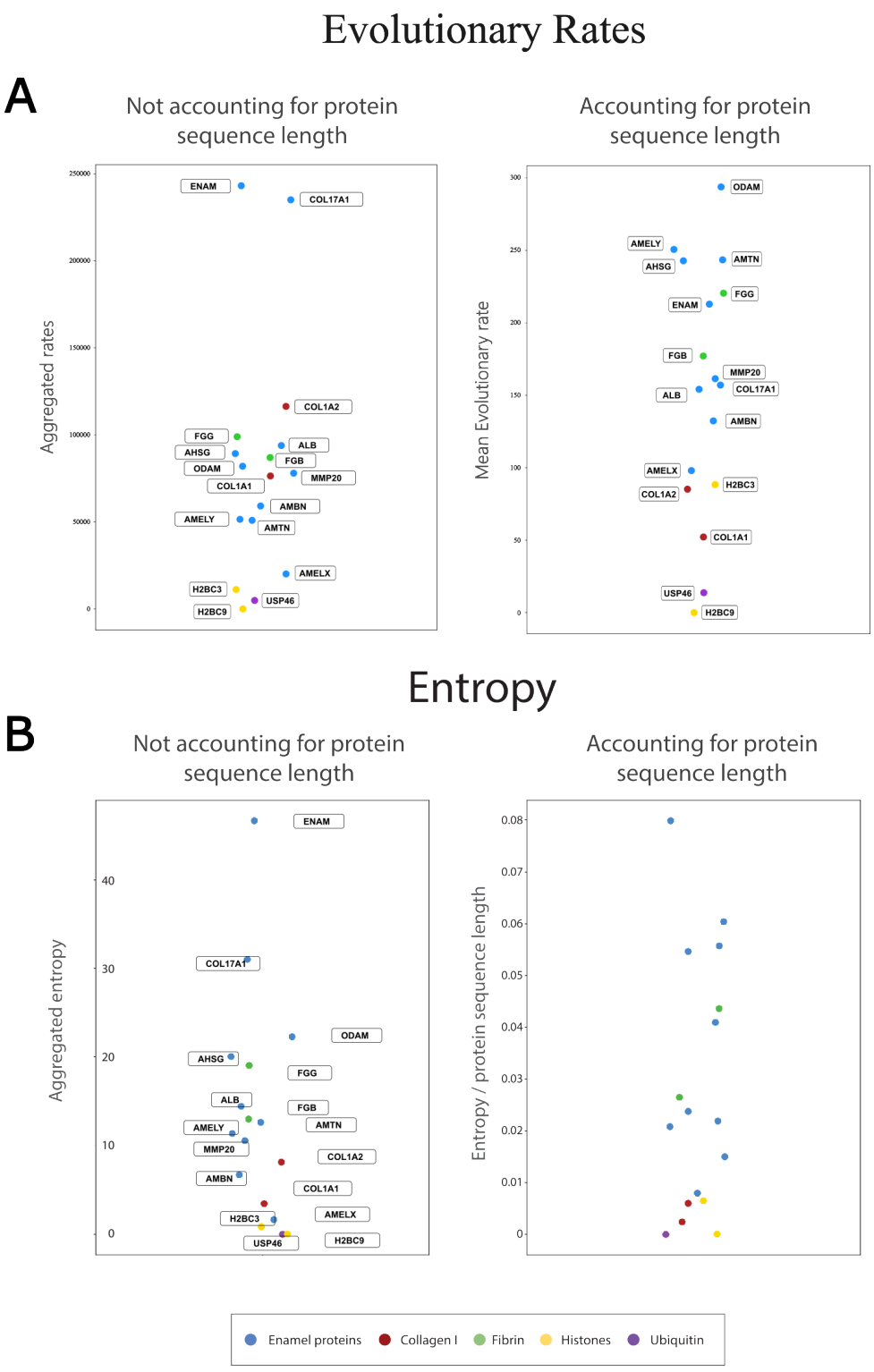

Figure 4: A: Protein entropy scoring comparison. Left: Each protein is ranked from highest to lowest based on the entropy scoring. Right: The entropy scoring is normalized based on the length of each protein, which causes some proteins to swap ranking. B: Protein evolutionary rates scoring comparison. Left: Each protein is ranked from highest to lowest based on the total evolutionary rate across all sites. Right: The total evolutionary rate is normalized based on the length of each protein, which causes some proteins to swap ranking.

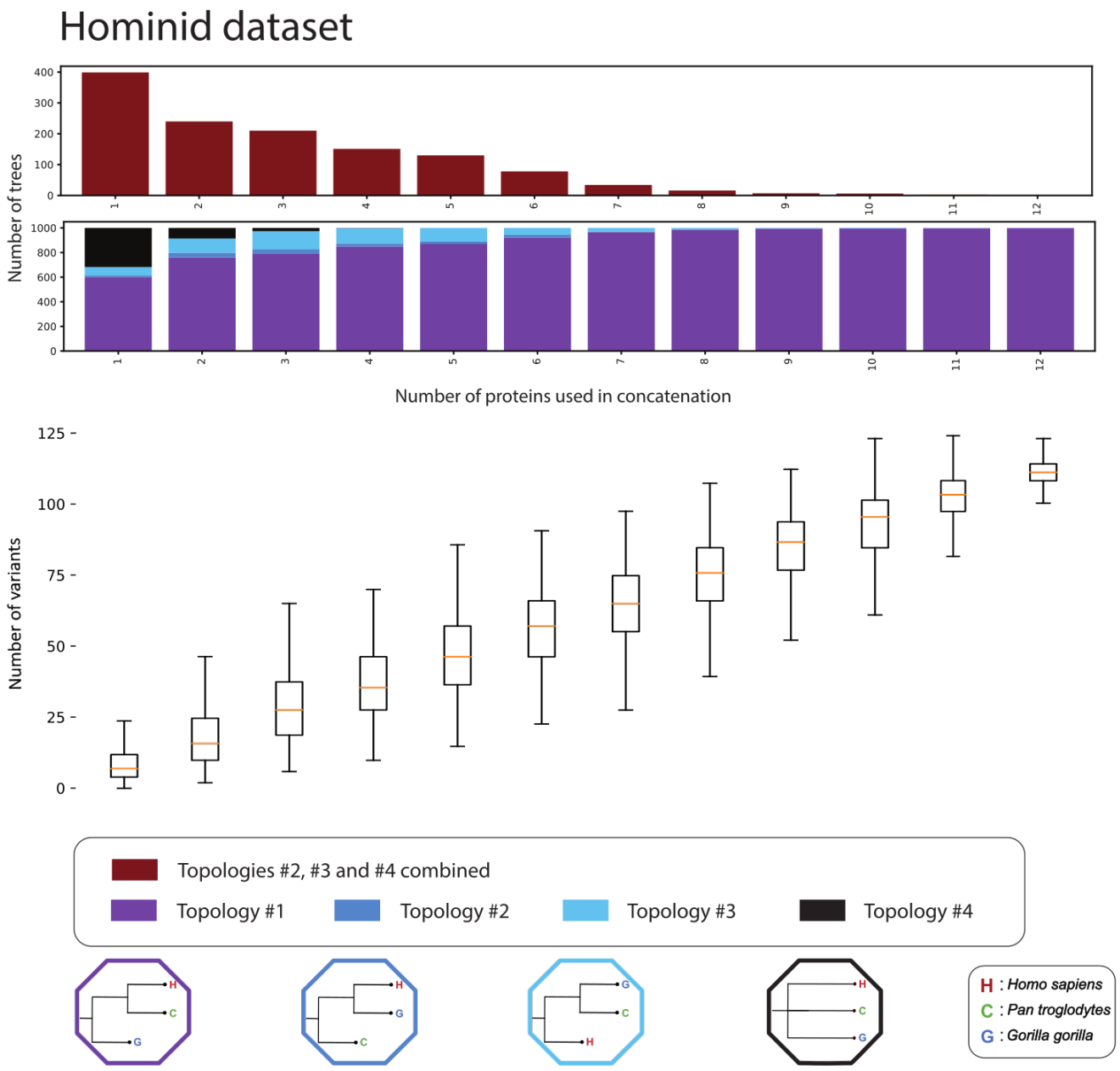

Figure 5: Iterative concatenation analysis for the hominid dataset. The upper barplot showcases the number of trees, out of 1000, differing from Topology #1 for an N number of proteins used in the concatenation. The lower barplot showcases the percentage of trees supporting each of the 4 possible topologies, out of 1000, for an N number of proteins used in the concatenation. The 4 possible topologies and their corresponding colours are visible at the bottom of the plot. The number of variants present in the dataset creating each tree are visible below the bar plot as a box plot. For each box the orange line denotes the median of the variants present in the N number of proteins concatenation. Each box plot denotes the 25%, 50% (the median, the line in the middle of the box) and 75% quantiles of the distribution. The whiskers of each box denote extremely low values (25% quantile - 1.5 * interquantile range) and extremely high values (75% quantile + 1.5 * interquantile range) for that distribution. The table containing the exact mean, median, maximum and minimum variants for each N proteins is available in Supplementary (S3.3).

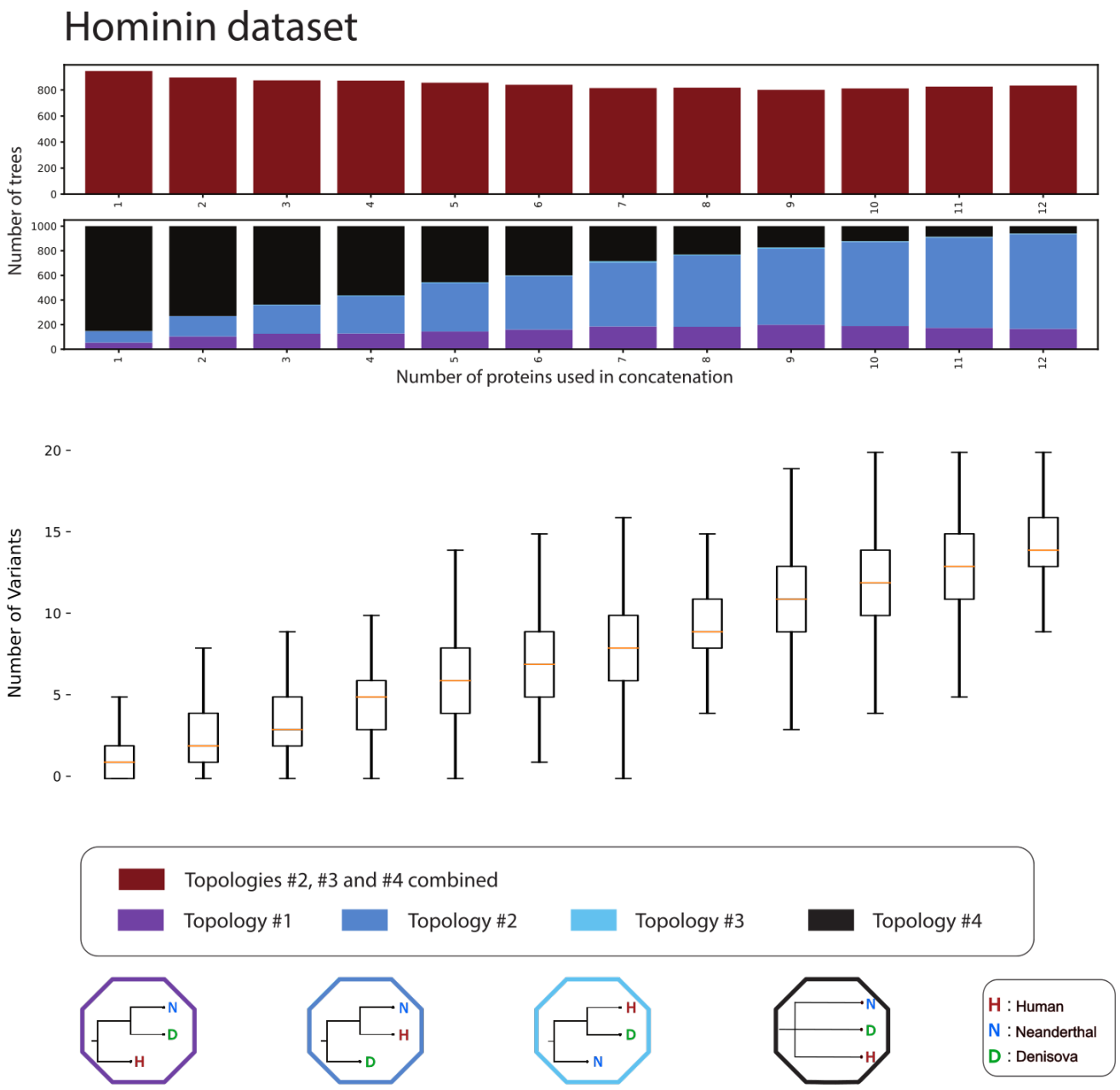

Figure 6: Iterative concatenation analysis for the hominin dataset. The lower barplot showcases the percentage of trees supporting each of the 4 possible topologies, out of 1000, for an N number of proteins used in the concatenation. The 4 possible topologies and their corresponding colours are visible at the bottom of the plot. The number of variants present in the dataset creating each tree are visible below the bar plot as a box plot. For each box the orange line denotes the median of the variants present in the N number of proteins concatenation. Each box plot denotes the 25%, 50% (the median, the line in the middle of the box) and 75% quantiles of the distribution. The whiskers of each box denote extremely low values (25% quantile - 1.5 * interquantile range) and extremely high values (75% quantile + 1.5 * interquantile range) for that distribution. The table containing the exact mean, median, maximum and minimum variants for each N proteins is available in Supplementary (S3.3).

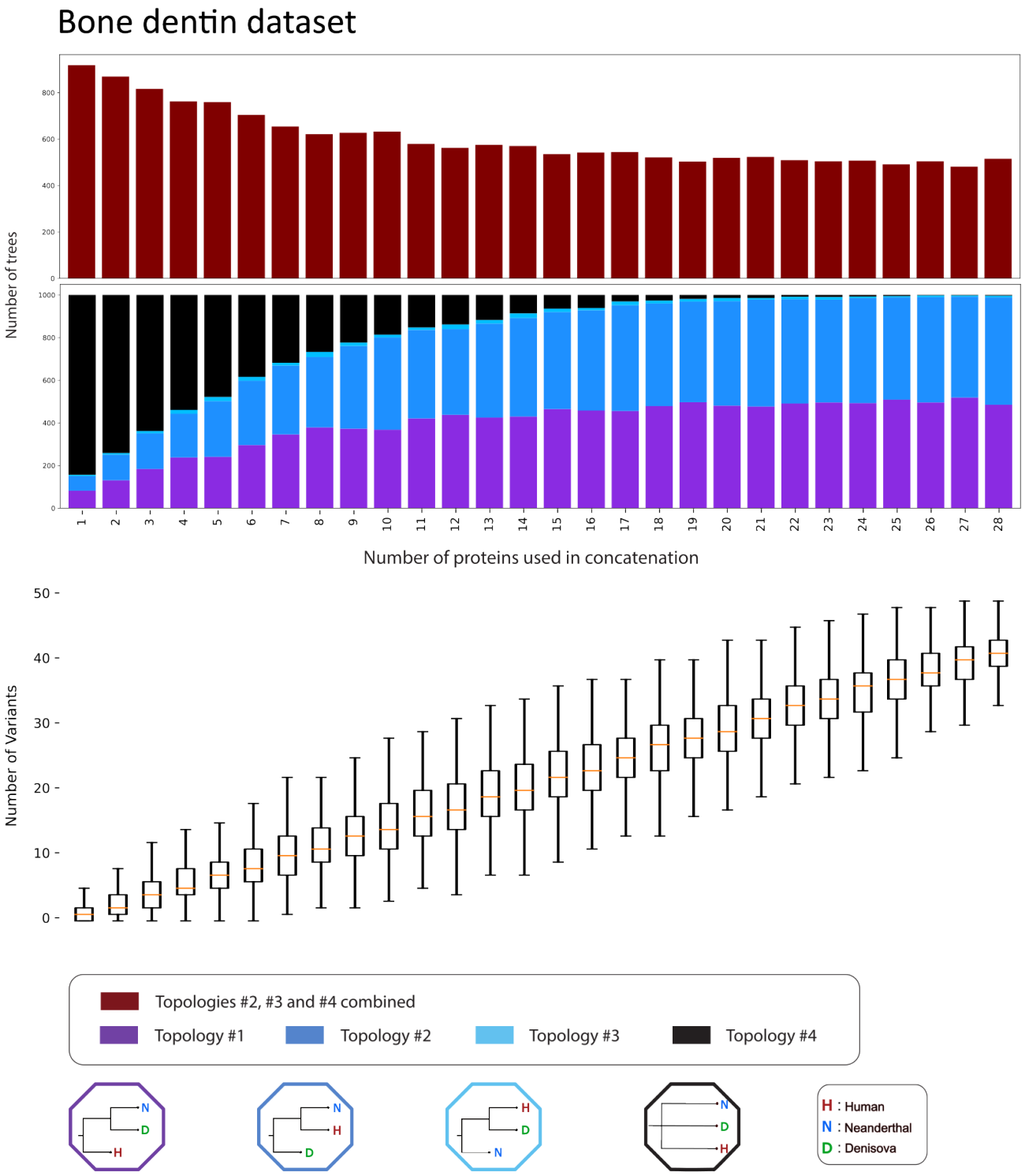

Figure 7: Iterative concatenation analysis for the dentin-bone dataset. The lower barplot showcases the percentage of trees supporting each of the 4 possible topologies, out of 1000, for an N number of proteins used in the concatenation. The 4 possible topologies and their corresponding colours are visible at the bottom of the plot. The number of variants present in the dataset creating each tree are visible below the bar plot as a box plot. For each box the orange line denotes the median of the variants present in the N number of proteins concatenation. Each box plot denotes the 25%, 50% (the median, the line in the middle of the box) and 75% quantiles of the distribution. The whiskers of each box denote extremely low values (25% quantile - 1.5 * interquantile range) and extremely high values (75% quantile + 1.5 * interquantile range) for that distribution. The table containing the exact mean, median, maximum and minimum variants for each N proteins is available in Supplementary (S3.3).

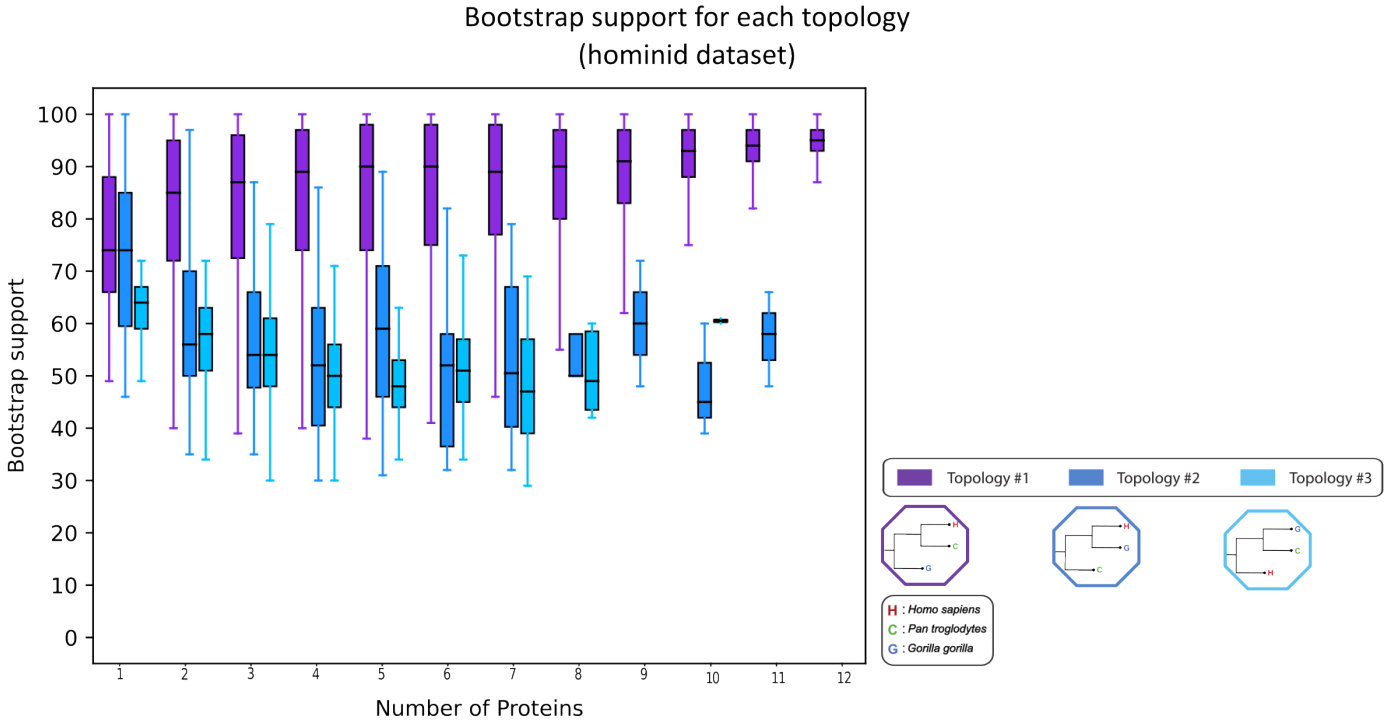

Figure 8: Box plot of the bootstrap values of the hominid dataset's generated trees. A bootstrap value is a metric generated along with a phylogenetic tree, representing the "confidence" of that tree. In our results, the bootstrap values of a tree range between 0 (lowest confidence) to 100 (highest confidence). Each box in this plot represents the distribution of these bootstrap values, grouped together by the number of proteins used to generate their corresponding tree (position on the x axis) and the topology that tree supported (colour). The three topologies corresponding to each colour are shown in the bottom right of the plot. Each box plot denotes the 25%, 50% (the median, the line in the middle of the box) and 75% quantiles of the distribution. The whiskers of each box denote extremely low values (25% quantile - 1.5 * interquantile range) and extremely high values (75% quantile + 1.5 * interquantile range) for that distribution.

Bibliography

[1] Charles A Lockwood, William H Kimbel, and John M Lynch. "Morphometrics and hominoid phylogeny: support for a chimpanzee–human clade and differentiation among great ape subspecies". In: *Proceedings of the National Academy of Sciences* 101.13 (2004), pp. 4356–4360. DOI: https://doi.org/10.1073/pnas.0306235101.

[2] Brian Villmoare. "Early Homo and the role of the genus in paleoanthropology". In: *American Journal of Physical Anthropology* 165 (2018), pp. 72–89. DOI: 10.1002/ajpa.23387.

[3] Ian Tattersall. "Evolutionary theory, systematics, and the study of human origins". In: *Journal of Anthropological Sciences* 100 (2022), pp. 1–26. DOI: 10.4436/JASS.10007.

[4] Carles Lalueza-Fox and M Thomas P Gilbert. "Paleogenomics of archaic hominins". In: *Current Biology* 21.24 (2011), R1002–R1009. DOI: 10.1016/j.cub.2011.11.021.

[5] Christina Warinner, Kristine Korzow Richter, and Matthew J Collins. "Paleoproteomics". In: *Chemical Reviews* (2022). DOI: 10.1021/acs.chemrev.1c00703.

[6] Christina M Nielsen-Marsh et al. "Extraction and sequencing of human and Neanderthal mature enamel proteins using MALDI-TOF/TOF MS". In: *Journal of Archaeological Science* 36.8 (2009), pp. 1758–1763. DOI: https://doi.org/10.1016/j.jas.2009.04.004.

[7] Beatrice Demarchi et al. "Protein sequences bound to mineral surfaces persist into deep time". In: *elife* 5 (2016), e17092. DOI: https://doi.org/10.7554/eLife.17092.

[8] Ryan Sinclair S Paterson et al. "A 20+ Ma old enamel proteome from Canada's High Arctic reveals diversification of Rhinocerotidae in the middle Eocene-Oligocene". In: *bioRxiv* (2024), pp. 2024–06. DOI: https://doi.org/10.1101/2024.06.07.597871.

[9] Love Dalén et al. "Deep-time paleogenomics and the limits of DNA survival". In: *Science* 382.6666 (2023), pp. 48–53. DOI: 10.1126/science.adh7943.

[10] Frido Welker et al. "The dental proteome of Homo antecessor". In: *Nature* 580.7802 (2020), pp. 235–238. DOI: https://doi.org/10.1038/s41586-020-2153-8.

[11] Frido Welker et al. "Enamel proteome shows that Gigantopithecus was an early diverging pongine". In: *Nature* 576.7786 (2019), pp. 262–265. DOI: https://doi.org/10.1038/s41586-019-1728-8.

[12] Jülide Kubat et al. "Geometric morphometrics and paleoproteomics enlighten the paleodiversity of Pongo". In: *Plos one* 18.12 (2023), e0291308. DOI: https://doi.org/10.1371/journal.pone.0291308.

[13] Palesa P Madupe et al. "Enamel proteins reveal biological sex and genetic variability within southern African Paranthropus". In: *bioRxiv* (2023), pp. 2023–07. DOI: https://doi.org/10.1101/2023.07.03.547326.

[14] Enrico Cappellini et al. "Proteomic analysis of a pleistocene mammoth femur reveals more than one hundred ancient bone proteins". In: *Journal of proteome research* 11.2 (2012), pp. 917–926. DOI: 10.1021/pr200721u.

[15] Ludovic Orlando et al. "Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse". In: *Nature* 499.7456 (2013), pp. 74–78. DOI: https://doi.org/10.1038/nature12323.

[16] Christina Warinner et al. “Pathogens and host immunity in the ancient human oral cavity”. In: *Nature genetics* 46.4 (2014), pp. 336–344. DOI: https://doi.org/10.1038/ng.2906.

[17] Ryan C Hill et al. “Preserved proteins from extinct Bison latifrons identified by tandem mass spectrometry; hydroxylysine glycosides are a common feature of ancient collagen”. In: *Molecular & Cellular Proteomics* 14.7 (2015), pp. 1946–1958. DOI: 10.1074/mcp.M114.047787.

[18] Rikai Sawafuji et al. “Proteomic profiling of archaeological human bone”. In: *Royal Society open science* 4.6 (2017), p. 161004. DOI: https://doi.org/10.1098/rsos.161004.

[19] Rosa R Jersie-Christensen et al. “Quantitative metaproteomics of medieval dental calculus reveals individual oral health status”. In: *Nature communications* 9.1 (2018), p. 4744. DOI: https://doi.org/10.1038/s41467-018-07148-3.

[20] Liam T Lanigan et al. “Multi-protease analysis of Pleistocene bone proteomes”. In: *Journal of proteomics* 228 (2020), p. 103889. DOI: 10.1016/j.jprot.2020.103889.

[21] Natalia Rybczynski et al. “Mid-Pliocene warm-period deposits in the High Arctic yield insight into camel evolution”. In: *Nature communications* 4.1 (2013), p. 1550. DOI: https://doi.org/10.1038/ncomms2516.

[22] Michael Buckley, Craig Lawless, and Natalia Rybczynski. “Collagen sequence analysis of fossil camels, Camelops and cf Paracamelus, from the Arctic and sub-Arctic of Plio-Pleistocene North America”. In: *Journal of proteomics* 194 (2019), pp. 218–225. DOI: 10.1016/j.jprot.2018.11.014.

[23] ULF Lagerkvist. “”Two out of three”: an alternative method for codon reading.” In: *Proceedings of the National Academy of Sciences* 75.4 (1978), pp. 1759–1762. DOI: 10.1073/pnas.75.4.1759.

[24] Adam Eyre-Walker and Peter D Keightley. “The distribution of fitness effects of new mutations”. In: *Nature Reviews Genetics* 8.8 (2007), pp. 610–618. DOI: https://doi.org/10.1038/nrg2146.

[25] Zhengting Zou and Jianzhi Zhang. “Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations?” In: *Molecular biology and evolution* 32.8 (2015), pp. 2085–2096. DOI: 10.1093/molbev/msv091.

[26] Fábio K Mendes, Yoonsoo Hahn, and Matthew W Hahn. “Gene tree discordance can generate patterns of diminishing convergence over time”. In: *Molecular biology and evolution* 33.12 (2016), pp. 3299–3307. DOI: https://doi.org/10.1093/molbev/msw197.

[27] Georgii A Bazykin et al. “Extensive parallelism in protein evolution”. In: *Biology direct* 2.1 (2007), pp. 1–13. DOI: 10.1186/1745-6150-2-20.

[28] Antonis Rokas and Sean B Carroll. “Frequent and widespread parallel evolution of protein sequences”. In: *Molecular biology and evolution* 25.9 (2008), pp. 1943–1953. DOI: 10.1093/molbev/msn143.

[29] Joe Parker et al. “Genome-wide signatures of convergent evolution in echolocating mammals”. In: *Nature* 502.7470 (2013), pp. 228–231. DOI: 10.1038/nature12511.

[30] Andrew D Foote et al. “Convergent evolution of the genomes of marine mammals”. In: *Nature genetics* 47.3 (2015), pp. 272–275. DOI: https://doi.org/10.1038/ng.3198.

[31] Richard A Goldstein et al. “Nonadaptive amino acid convergence rates decrease over time”. In: *Molecular biology and evolution* 32.6 (2015), pp. 1373–1381. DOI: 10.1093/molbev/msv041.

[32] Robert C Griffiths and Paul Marjoram. “Ancestral inference from samples of DNA sequences with recombination”. In: *Journal of computational biology* 3.4 (1996), pp. 479–502. DOI: https://doi.org/10.1089/cmb.1996.3.4.

[33] Alexander L Lewanski, Michael C Grundler, and Gideon S Bradburd. “The era of the ARG: an empiricist’s guide to ancestral recombination graphs”. In: *ArXiv* (2023).

[34] Wayne P Maddison. “Gene trees in species trees”. In: *Systematic biology* 46.3 (1997), pp. 523–536. DOI: https://doi.org/10.1093/sysbio/46.3.523.

[35] Asger Hobolth et al. "Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection". In: *Genome research* 21.3 (2011), pp. 349–356. DOI: 10.1101/gr.114751.110.

[36] Kirk E Lohmueller, Carlos D Bustamante, and Andrew G Clark. "The effect of recent admixture on inference of ancient human population history". In: *Genetics* 185.2 (2010), pp. 611–622. DOI: https://doi.org/10.1534/genetics.109.113761.

[37] Daniel A Pollard et al. "Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting". In: *PLoS genetics* 2.10 (2006), e173. DOI: https://doi.org/10.1371/journal.pgen.0020173.

[38] Thomas Mailund, Kasper Munch, and Mikkel Heide Schierup. "Lineage sorting in apes". In: *Annual review of genetics* 48 (2014), pp. 519–535. DOI: 10.1146/annurev-genet-120213-092532.

[39] Vitor Sousa and Jody Hey. "Understanding the origin of species with genome-scale data: modelling gene flow". In: *Nature Reviews Genetics* 14.6 (2013), pp. 404–414. DOI: https://doi.org/10.1038/nrg3446.

[40] Brian Charlesworth, Carolina Bartolomé, and Véronique NoëL. "The detection of shared and ancestral polymorphisms". In: *Genetics Research* 86.2 (2005), pp. 149–157. DOI: 10.1017/S0016672305007743.

[41] Emilia Huerta-Sánchez et al. "Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA". In: *Nature* 512.7513 (2014), pp. 194–197. DOI: https://doi.org/10.1038/nature13408.

[42] Aylwyn Scally et al. "Insights into hominid evolution from the gorilla genome sequence". In: *Nature* 483.7388 (2012), pp. 169–175. DOI: https://doi.org/10.1038/nature10842.

[43] Zev N Kronenberg et al. "High-resolution comparative analysis of great ape genomes". In: *Science* 360.6393 (2018), eaar6343. DOI: 10.1126/science.aar634.

[44] Iker Rivas-González et al. "Pervasive incomplete lineage sorting illuminates speciation and selection in primates". In: *Science* 380.6648 (2023), eabn4409. DOI: 10.1126/science.abn44.

[45] Richard E Green et al. "A draft sequence of the Neandertal genome". In: *science* 328.5979 (2010), pp. 710–722. DOI: 10.1126/science.1188021.

[46] Martin Kuhlwilm et al. "Ancient admixture from an extinct ape lineage into bonobos". In: *Nature ecology & evolution* 3.6 (2019), pp. 957–965. DOI: https://doi.org/10.1038/s41559-019-0881-7.

[47] Harvinder Pawar et al. "Ghost admixture in eastern gorillas". In: *Nature ecology & evolution* 7.9 (2023), pp. 1503–1514. DOI: https://doi.org/10.1038/s41559-023-02145-2.

[48] Nicolas Galtier. "An approximate likelihood method reveals ancient gene flow between human, chimpanzee and gorilla". In: *Peer Community Journal* 4 (2024). DOI: 10.24072/pcjournal.359.

[49] B Demarchi et al. *Ancient proteins resolve controversy over the identity of Genyornis eggshell.* 2021. DOI: https://doi.org/10.1073/pnas.2109326119.

[50] Alberto J Taurozzi et al. "Deep-time phylogenetic inference by paleoproteomic analysis of dental enamel". In: *Nature Protocols* (2024), pp. 1–32. DOI: https://doi.org/10.1038/s41596-024-00975-3.

[51] Mike Buckley and Caroline Wadsworth. "Proteome degradation in ancient bone: diagenesis and phylogenetic potential". In: *Palaeogeography, Palaeoclimatology, Palaeoecology* 416 (2014), pp. 69–79. DOI: https://doi.org/10.1016/j.palaeo.2014.06.026.

[52] Carine Froment et al. "Protein sequence comparison of human and non-human primate tooth proteomes". In: *Journal of Proteomics* 231 (2021), p. 104045. DOI: 10.1016/j.jprot.2020.104045.

[53] Ricardo Fong-Zazueta et al. "Phylogenetic signal in primate tooth enamel proteins and its relevance for paleoproteomics". In: *Genome Biology and Evolution* (2025), evaf007. DOI: https://doi.org/10.1093/gbe/evaf007.

[54] Maria Codlin et al. "New avian protein sequences improve resolution of palaeoproteomic approaches to taxonomic identification & reveal widespread intraspecies variability". In: *Research Square* (2025). DOI: https://doi.org/10.21203/rs.3.rs-5880462/v1.

[55] Javier Prado-Martinez et al. "Great ape genetic diversity and population history". In: *Nature* 499.7459 (2013), pp. 471–475. DOI: https://doi.org/10.1038/nature12228.

[56] Jeffrey D Wall. "Great ape genomics". In: *ILAR journal* 54.2 (2013), pp. 82–90. DOI: 10.1093/ilar/ilt048.

[57] Aisha Yousaf et al. "Current progress in evolutionary comparative genomics of great apes". In: *Frontiers in Genetics* 12 (2021), p. 657468. DOI: 10.3389/fgene.2021.657468.

[58] David Castellano and Kasper Munch. "Population genomics in the great apes". In: *Statistical Population Genomics* (2020), pp. 453–463. DOI: 10.1007/978-1-0716-0199-0_19.

[59] Matthias Meyer et al. "A high-coverage genome sequence from an archaic Denisovan individual". In: *Science* 338.6104 (2012), pp. 222–226. DOI: 10.1126/science.1224344.

[60] Kay Prüfer et al. "The complete genome sequence of a Neanderthal from the Altai Mountains". In: *Nature* 505.7481 (2014), pp. 43–49. DOI: https://doi.org/10.1038/nature12886.

[61] Kay Prüfer et al. "A high-coverage Neandertal genome from Vindija Cave in Croatia". In: *Science* 358.6363 (2017), pp. 655–658. DOI: 10.1126/science.aao1887.

[62] Patrick Leopold Rüther et al. "SPIN enables high throughput species identification of archaeological bone by proteomics". In: *Nature communications* 13.1 (2022), p. 2458. DOI: 10.1038/s41467-022-30097-x.

[63] Zandra Fagernäs et al. "Cleaning the Dead: Optimized decontamination enhances palaeoproteomic analyses of Pleistocene skeletal material". In: *bioRxiv* (2024), pp. 2024–06. DOI: https://doi.org/10.1101/2024.06.13.598810.

[64] Huan Xia et al. "Middle and Late Pleistocene Denisovan subsistence at Baishiya Karst Cave". In: *Nature* 632.8023 (2024), pp. 108–113. DOI: https://doi.org/10.1038/s41586-024-07612-9.

[65] Fergal J Martin et al. "Ensembl 2023". In: *Nucleic acids research* 51.D1 (2023), pp. D933–D941. DOI: https://doi.org/10.1093/nar/gkac958.

[66] Kazutaka Katoh and Daron M Standley. "MAFFT multiple sequence alignment software version 7: improvements in performance and usability". In: *Molecular biology and evolution* 30.4 (2013), pp. 772–780. DOI: 10.1093/molbev/mst010.

[67] Stéphane Guindon et al. "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0". In: *Systematic biology* 59.3 (2010), pp. 307–321. DOI: https://doi.org/10.1093/sysbio/syq010.

[68] Claude Elwood Shannon. "A mathematical theory of communication". In: *ACM SIGMOBILE mobile computing and communications review* 5.1 (2001), pp. 3–55. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[69] Tal Pupko et al. "Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues". In: *Bioinformatics* 18.suppl_1 (2002), S71–S77. DOI: 10.1093/bioinformatics/18.suppl_1.s71.

[70] Grant B.J. et al. "Bio3D: An R package for the comparative analysis of protein structures." In: *Bioinformatics* 22 (Nov. 2006), pp. 2695–2696. DOI: https://doi.org/10.1093/bioinformatics/btl461.

[71] Ioannis Patramanis et al. "PaleoProPhyler: a reproducible pipeline for phylogenetic inference using ancient proteins". In: *bioRxiv* (2022), pp. 2022–12. DOI: 10.24072/pcjournal.344.

[72] Masatoshi Nei. *Molecular evolutionary genetics*. Columbia university press, 1987. DOI: https://doi.org/10.7312/nei-92038.

[73] Andreas D Baxevanis and David Landsman. "Histone Sequence Database: a compilation of highly-conserved nucleoprotein sequences". In: *Nucleic acids research* 24.1 (1996), pp. 245–247. DOI: https://doi.org/10.1093/nar/24.1.245.

[74] David H Schlesinger and Gideon Goldstein. "Molecular conservation of 74 amino acid sequence of ubiquitin between cattle and man". In: *Nature* 255.5507 (1975), pp. 423–424. DOI: https://doi.org/10.1038/255423a0.

[75] Ioannis Patramanis et al. *Hominid Palaeoproteomic Reference Dataset*. Version 1.01. 2022. DOI: 10.5281/zenodo.7333226. URL: https://zenodo.org/records/7728060.

[76] Todd R Disotell. "Archaic human genomics". In: *American journal of physical anthropology* 149.S55 (2012), pp. 24–39. DOI: 10.1002/ajpa.22159.

[77] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[78] Samantha Brown et al. "Identification of a new hominin bone from Denisova Cave, Siberia using collagen fingerprinting and mitochondrial DNA analysis". In: *Scientific reports* 6.1 (2016), p. 23559. DOI: 10.1038/srep23559.

[79] Fahu Chen et al. "A late middle Pleistocene Denisovan mandible from the Tibetan Plateau". In: *nature* 569.7756 (2019), pp. 409–412. DOI: https://doi.org/10.1038/s41586-019-1139-x.

[80] Samantha Brown et al. "Zooarchaeology through the lens of collagen fingerprinting at Denisova Cave". In: *Scientific Reports* 11.1 (2021), p. 15457. DOI: https://doi.org/10.1038/s41598-021-94731-2.

[81] Fabrice Demeter et al. "A Middle Pleistocene Denisovan molar from the Annamite chain of northern Laos". In: *Nature communications* 13.1 (2022), p. 2557. DOI: https://doi.org/10.1038/s41467-022-29923-z.

[82] Ting-Fung Chan et al. "Natural variation in four human collagen genes across an ethnically diverse population". In: *Genomics* 91.4 (2008), pp. 307–314. DOI: https://doi.org/10.1016/j.ygeno.2007.12.008.

[83] David M Hudson et al. "Distinct post-translational features of type I collagen are conserved in mouse and human periodontal ligament". In: *Journal of periodontal research* 52.6 (2017), pp. 1042–1049. DOI: https://doi.org/10.1111/jre.12475.

[84] SJ Brookes et al. "Biochemistry and molecular biology of amelogenin proteins of developing dental enamel". In: *Archives of oral biology* 40.1 (1995), pp. 1–14. DOI: https://doi.org/10.1016/0003-9969(94)00135-X.

[85] Doris Bachtrog and Brian Charlesworth. "Reduced adaptation of a non-recombining neo-Y chromosome". In: *Nature* 416.6878 (2002), pp. 323–326. DOI: https://doi.org/10.1038/416323a.

[86] Jennifer F Hughes et al. "Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content". In: *Nature* 463.7280 (2010), pp. 536–539. DOI: https://doi.org/10.1038/nature08700.

[87] Sheng Sun and Joseph Heitman. "Should Y stay or should Y go: The evolution of non-recombining sex chromosomes". In: *BioEssays* 34.11 (2012), pp. 938–942. DOI: 10.1002/bies.201200064.

[88] Melissa A Wilson Sayres. "Genetic diversity on the sex chromosomes". In: *Genome biology and evolution* 10.4 (2018), p. 1064. DOI: 10.1093/gbe/evy039.

[89] Nicolas Andre Stewart et al. "Sex determination of human remains from peptides in tooth enamel". In: *Proceedings of the National Academy of Sciences* 114.52 (2017), pp. 13649–13654. DOI: https://doi.org/10.1073/pnas.1714926115.

[90] Glendon J Parker et al. "Sex estimation using sexually dimorphic amelogenin protein fragments in human enamel". In: *Journal of Archaeological Science* 101 (2019), pp. 169–180. DOI: https://doi.org/10.1016/j.jas.2018.08.011.

[91] Ana Maria Casas-Ferreira et al. "Fast methods based on mass spectrometry for peptide identification. Application to sex determination of human remains in tooth enamel". In: *Microchemical Journal* 181 (2022), p. 107645. DOI: https://doi.org/10.1016/j.microc.2022.107645.

[92] Timothy P Cleland et al. "SPEED-E: A modified version of the sample preparation by Easy extraction and Digestion (-free) protocol for enamel-based sex estimation in archaeological remains". In: *Journal of Archaeological Science* 168 (2024), p. 106006. DOI: https://doi.org/10.1016/j.jas.2024.106006.

[93] Julia A Gamble et al. "Advancing sex estimation from amelogenin: Applications to archaeological, deciduous, and fragmentary dental enamel". In: *Journal of Archaeological Science: Reports* 54 (2024), p. 104430. DOI: https://doi.org/10.1016/j.jasrep.2024.104430.

[94] Claire Koenig et al. "Automated High-Throughput Biological Sex Identification from Archeological Human Dental Enamel Using Targeted Proteomics". In: *Journal of Proteome Research* 23.11 (2024), pp. 5107–5121. DOI: https://doi.org/10.1021/acs.jproteome.4c00557.

[95] Lily R Adair et al. "LAP-MALDI MS analysis of amelogenin from teeth for biological sex estimation". In: *Journal of Pharmaceutical and Biomedical Analysis* (2024), p. 116599. DOI: https://doi.org/10.1016/j.jpba.2024.116599.

[96] Federico Lugli et al. "Enamel peptides reveal the sex of the Late Antique 'Lovers of Modena'". In: *Scientific Reports* 9.1 (2019), p. 13130. DOI: https://doi.org/10.1038/s41598-019-49562-7.

[97] Carine Froment et al. "Analysis of 5000 year-old human teeth using optimized large-scale and targeted proteomics approaches for detection of sex-specific peptides". In: *Journal of proteomics* 211 (2020), p. 103548. DOI: https://doi.org/10.1016/j.jprot.2019.103548.

[98] Marta Cintas-Peña et al. "Amelogenin peptide analyses reveal female leadership in Copper Age Iberia (c. 2900–2650 BC)". In: *Scientific Reports* 13.1 (2023), p. 9594. DOI: https://doi.org/10.1038/s41598-023-36368-x.

[99] Anne S Cole and John E Eastoe. *Biochemistry and oral biology*. Butterworth-Heinemann, 2014.

[100] *Ensemble Homo sapiens AMELY orthologs*. https://www.ensembl.org/Homo_sapiens/Gene/Compara_Ortholog?db=core;g=ENSG00000099721;r=Y:6865918-6911752. Accessed: 2023-10-29.

[101] Malcolm L Snead et al. "Of mice and men: anatomy of the amelogenin gene". In: *Connective Tissue Research* 22.1-4 (1989), pp. 727–735. DOI: https://doi.org/10.3109/03008208909114125.

[102] Jan E Janečka et al. "Horse Y chromosome assembly displays unique evolutionary features and putative stallion fertility genes". In: *Nature communications* 9.1 (2018), pp. 1–15. DOI: https://doi.org/10.1038/s41467-018-05290-6.

[103] Kazuhiko Kawasaki et al. "The evolution of unusually small amelogenin genes in cetaceans; pseudogenization, X–Y gene conversion, and feeding strategy". In: *Journal of molecular evolution* 88 (2020), pp. 122–135. DOI: https://doi.org/10.1007/s00239-019-09917-0.

[104] Michael Buckley. "Zooarchaeology by Mass Spectrometry (ZooMS)". In: *Handbook of Archaeological Sciences* 1 (2023), pp. 483–499. DOI: https://doi.org/10.1007/978-1-4419-0465-2_2418.

[105] Rebecca C Griffin et al. "A new approach to amino acid racemization in enamel: testing of a less destructive sampling methodology". In: *Journal of forensic sciences* 53.4 (2008), pp. 910–916. DOI: https://doi.org/10.1111/j.1556-4029.2008.00753.x.

[106] Nicolas Andre Stewart et al. "The identification of peptides by nanoLC-MS/MS from human surface tooth enamel following a simple acid etch extraction". In: *RSC advances* 6.66 (2016), pp. 61673–61679. DOI: 10.1039/c6ra05120k.

[107] Katharina Rebay-Salisbury et al. "Child murder in the Early Bronze Age: proteomic sex identification of a cold case from Schleinbach, Austria". In: *Archaeological and anthropological sciences* 12.11 (2020), p. 265. DOI: 10.1007/s12520-020-01199-8.

[108] Isabelle Fabrizi et al. "Low-Invasive Sampling Method with Tape-Disc Sampling for the Taxonomic Identification of Archeological and Paleontological Bones by Proteomics". In: *Journal of Proteome Research* 23.8 (2024), pp. 3404–3417.

[109] Theis Zetner Trolle Jensen et al. "Tryps-IN: A streamlined palaeoproteomics workflow enables ZooMS analysis of 10,000-year-old petrous bones from Jordan rift-valley". In: *Journal of Archaeological Science: Reports* 52 (2023), p. 104238. DOI: https://doi.org/10.1016/j.jasrep.2023.104238.

[110] Shevan Wilkin et al. "Sequential trypsin and ProAlanase digestions unearth immunological protein biomarkers shrouded by skeletal collagen". In: *Iscience* 27.5 (2024). DOI: 10.1016/j.isci.2024.109663.

[111] Zandra Fagernäs et al. "Digging deeper into ancient skeletal proteomes through consecutive digestion with multiple proteases". In: *Journal of Proteomics* 298 (2024), p. 105143. DOI: https://doi.org/10.1016/j.jprot.2024.105143.

[112] Sebastien Gallien et al. "Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer". In: *Molecular & cellular proteomics* 11.12 (2012), pp. 1709–1723. DOI: 10.1074/mcp.O112.019802.

[113] Ragnheiur Diljá Ásmundsdóttir et al. "Early Holocene preservation differences between cortical and trabecular bone proteomes". In: *Journal of Archaeological Science: Reports* 57 (2024), p. 104643. DOI: https://doi.org/10.1016/j.jasrep.2024.104643.

[114] Rodriguez-Palomo Ismael et al. "Benchmarking the identification of a single degraded protein to explore optimal search strategies for ancient proteins". In: *Peer Community Journal* 4 (2024). DOI: https://doi.org/10.24072/pcjournal.491.

[115] Yun Chiang, Frido Welker, and Matthew James Collins. "Spectra without stories: reporting 94% dark and unidentified ancient proteomes". In: *Open Research Europe* 4.71 (2024), p. 71. DOI: https://doi.org/10.12688/openreseurope.17225.1.

[116] Hayley C Lanier, Huateng Huang, and L Lacey Knowles. "How low can you go? The effects of mutation rate on the accuracy of species-tree estimation". In: *Molecular Phylogenetics and Evolution* 70 (2014), pp. 112–119. DOI: 10.1016/j.ympev.2013.09.006.

[117] Xiyun Jiao, Tomáš Flouri, and Ziheng Yang. "Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow". In: *National Science Review* 8.12 (2021), nwab127. DOI: 10.1093/nsr/nwab127.

[118] Jordan Douglas, Cinthy L Jiménez-Silva, and Remco Bouckaert. "StarBeast3: adaptive parallelized Bayesian inference under the multispecies coalescent". In: *Systematic Biology* 71.4 (2022), pp. 901–916. DOI: https://doi.org/10.1093/sysbio/syac010.

[119] Simon H Martin and Chris D Jiggins. "Interpreting the genomic landscape of introgression". In: *Current opinion in genetics & development* 47 (2017), pp. 69–74. DOI: https://doi.org/10.1016/j.gde.2017.08.007.

[120] Jeffrey D Wall and Michael F Hammer. "Archaic admixture in the human genome". In: *Current opinion in genetics & development* 16.6 (2006), pp. 606–610. DOI: 10.1016/j.gde.2006.09.006.

[121] Michael Dannemann and Fernando Racimo. "Something old, something borrowed: admixture and adaptation in human evolution". In: *Current opinion in genetics & development* 53 (2018), pp. 1–8. DOI: 10.1016/j.gde.2018.05.009.

[122] Rebecca R Ackermann et al. "Hybridization in human evolution: Insights from other organisms". In: *Evolutionary Anthropology: Issues, News, and Reviews* 28.4 (2019), pp. 189–209. DOI: 10.1002/evan.21787.

[123] Simon Ceder et al. "March, Tree, Stream—The Knowledge Production of Early Human Evolution—". In: (2018).