Peer Community Journal

Section: Paleontology

RESEARCH ARTICLE

Published
2023-11-23

Cite as
Ioannis Patramanis, Jazmín Ramos-Madrigal, Enrico Cappellini and Fernando Racimo (2023) *PaleoProPhyler: a reproducible pipeline for phylogenetic inference using ancient proteins*, Peer Community Journal, 3: e112.

Correspondence
john.patraman@gmail.com

Peer-review
Peer reviewed and recommended by PCI Paleontology, https://doi.org/10.24072/pci.paleo.100220

PaleoProPhyler: a reproducible pipeline for phylogenetic inference using ancient proteins

Ioannis Patramanis [1], Jazmín Ramos-Madrigal [2], Enrico Cappellini [3], and Fernando Racimo [1,4]

Volume **3** (2023), article e112

https://doi.org/10.24072/pcjournal.344

Abstract

Ancient proteins from fossilized or semi-fossilized remains can yield phylogenetic information at broad temporal horizons, in some cases even millions of years into the past. In recent years, peptides extracted from archaic hominins and long-extinct mega-fauna have enabled unprecedented insights into their evolutionary history. In contrast to the field of ancient DNA - where several computational methods exist to process and analyze sequencing data - few tools exist for handling ancient protein sequence data. Instead, most studies rely on loosely combined custom scripts, which makes it difficult to reproduce results or share methodologies across research groups. Here, we present PaleoProPhyler: a new fully reproducible pipeline for aligning ancient peptide data and subsequently performing phylogenetic analyses. The pipeline can not only process various forms of proteomic data, but also easily harness genetic data in different formats (CRAM, BAM, VCF) and translate it, allowing the user to create reference panels for phyloproteomic analyses. We describe the various steps of the pipeline and its many functionalities, and provide some examples of how to use it. PaleoProPhyler allows researchers with little bioinformatics experience to efficiently analyze palaeoproteomic sequences, so as to derive insights from this valuable source of evolutionary data.

[1]Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen – Copenhagen, Denmark, [2]Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen – Copenhagen, Denmark, [3]Section for GeoGenetics, Globe Institute, University of Copenhagen – Copenhagen, Denmark, [4]Lundbeck GeoGenetics Centre, Globe Institute, University of Copenhagen – Copenhagen, Denmark

Contents

1. Introduction

Recent advances in protein extraction and mass spectrometry (Lanigan et al., 2020; Nogueira et al., 2021; Porto et al., 2011; Rüther et al., 2022) have made it possible to isolate ancient peptides from organisms that lived thousands or even millions of years ago. Certain ancient proteins have a lower degradation rate and can be preserved for longer than ancient DNA (Cappellini et al., 2014; Demarchi et al., 2016; Hendy, 2021; Warinner et al., 2022). These ancient proteins can be utilized by Peptide Mass Fingerprinting (PMF) methods (Ostrom et al., 2000), including ZooMS (Buckley et al., 2009), for genus or species identification (Buckley et al., 2010) and to single out fossil material of interest for further analyses including DNA sequencing (Brown et al., 2016, 2022), radiocarbon dating (Devièse et al., 2017) and shotgun proteomics (Brown et al., 2016; Welker et al., 2016). Shotgun proteomics in particular, utilizing tandem mass spectrometry, has enabled the reconstruction of the amino acid sequences of those proteins, which sometimes number in the hundreds (Cappellini et al., 2012; Warinner et al., 2014). These sequences contain evolutionary information and thus have the potential to resolve important scientific questions about the deep past, which are not approachable via other methods. Tooth enamel proteins and bone collagen in particular have been successfully extracted from multiple extinct species, in order to resolve their relationships to other species (Buckley, 2015; Buckley et al., 2019; Cappellini et al., 2019; Chen et al., 2019; Nielsen-Mars et al., 2009; Rybczynski et al., 2013; Welker et al., 2020, 2019, 2017).

Ancient proteomic studies typically use combinations of custom scripts and repurposed software, which require extensive in-house knowledge and phylogenetic expertise, and are not easily reproducible. Barriers to newcomers in the field include difficulties in properly aligning the fractured peptides with present-day sequences, translating available genomic data for comparison, and porting proteomic data into standard phylogenetic packages. The creation of automated pipelines like PALEOMIX (Schubert et al., 2014) and EAGER (Peltzer et al., 2016) have facilitated the streamlining and reproducibility of ancient DNA analyses, which has been particularly helpful for emerging research groups around the world. This has undoubtedly contributed to the growth of the field (Lan and Lindqvist, 2018). Yet, the field of palaeoproteomics still lacks a "democratizing" tool that is approachable to researchers of different backgrounds and expertises.

Another important issue in phyloproteomics is the relative scarcity of proteomic datasets (Brandt et al., 2022; Müller et al., 2020). There are currently tens of thousands of publicly available whole genome sequences, covering hundreds of species (Byrska-Bishop et al., 2021; Koepfli et al., 2015; Lewin et al., 2018; Prado-Martinez et al., 2013; Zhang et al., 2014). The amount of publicly available proteome sequences is much smaller. NCBI's list of sequenced genomes National Center for Biotechnology Information (NCBI), 2004 includes 78,420 species, out of which 30,530 are eukaryotes and 11,345 labeled as 'Animal'. For comparison, Uniprot's reference proteomes list (Uniprot consortium, 2021) contains a total of 23,805 entries of which 2,400 are

eukaryotes and around 950 are labeled as 'Metazoa'. For most vertebrate species, lab-generated protein data does not exist and phyloproteomic research is reliant on sequences translated *in silico* from genomic data. These, more often than not, are not sufficiently validated or curated (Bagheri et al., 2020). Ensembl's curated database of fully annotated genomes , and thus available proteomes, numbers only around 270 species (Martin, 2023). As a result, assembling a proper reference dataset for phyloproteomics can be challenging. Given how important rigorous taxon sampling is in performing proper phylogenetic reconstruction (Heath et al., 2008; Rosenberg and Kumar, 2003), having a complete and reliable reference dataset is crucial. In the case of proteins, the typically short sequence length and the low amounts of sequence diversity - due to the strong influence of purifying selection - means that absence of knowledge about a single amino acid polymorphism (SAP) can strongly affect downstream inferences (Chen et al., 2019; Demarchi et al., 2022; Opperdoes, 2003; Presslee et al., 2019).

2. Methods

To address all of the above issues, we present "PaleoProPhyler": a fully reproducible and easily deployable pipeline for assisting researchers in phyloproteomic analyses of ancient peptides. "PaleoProPhyler" is based on the workflows developed in earlier ancient protein studies (Cappellini et al., 2019; Welker et al., 2020, 2019), with some additional functionalities. It allows for the search and access of available reference proteomes, bulk translation of CRAM, BAM or VCF files into amino acid sequences in FASTA format, and various forms of phylogenetic tree reconstruction.

To maximize reproducibility, accessibility and scalability, we have built our pipeline using Snakemake (Mölder et al., 2021) and Conda (Inc., 2020). The Snakemake format provides the workflow with tools for automation and computational optimization, while Conda enables the pipeline to operate on different platforms, granting it ease of access and portability. The pipeline is divided into three distinct but interacting modules (Modules 1,2 and 3), each of which is composed of a Snakemake script and a Conda environment (Figure 1).

Module 1 is designed to provide the user with a baseline (curated) reference dataset as well as the resources required to perform the in silico translation of proteins from mapped whole genomes. The input of module 1 is a user-provided list of proteins and a list of organisms. The user also has the option of choosing a particular reference build. Utilizing the Ensembl API (Yates et al., 2015), the module will return 3 different resources for each requested protein and for each requested organism. These are : a) the reference protein sequence of that organism in FASTA format (Lipman and Pearson, 1985), b) the location (position and strand) of the gene that corresponds to that protein and c) the start and end of each exon and intron of that gene / isoform. The downloaded FASTA sequences are available individually but are also assembled into species- and protein-specific datasets. They can be immediately used as a reference dataset for either downstream phylogenetic analyses or as an input database for mass spectrometry software, like MaxQuant (Cox and Mann, 2008), Pfind (Chi et al., 2018), PEAKS (Ma et al., 2003) and others (Demichev et al., 2018; Kong et al., 2017; Perkins et al., 1999; Solntsev et al., 2018). The gene location information and the exon / intron tables can be utilized automatically by Module 2. For each requested protein, the module will select the Ensembl canonical isoform by default. Should the user desire a specific isoform or all protein coding isoforms of a protein, they have the ability to specify that as an option in the provided protein list.

Module 2 is designed to utilize the resources generated by Module 1 and to extract, splice and translate genes from whole genome data, into the proteins of interest. This module can handle some of the most commonly used genomic data file formats, including the BAM (Li et al., 2009), CRAM (Bonfield, 2022) and VCF (Danecek et al., 2011) formats. The easiest way to run Module 2 is to first run Module 1 for a set of proteins and a selected organism. This will generate all the necessary files and resources required for the protein translation. The selected organism will be used as a reference for the translation process. All genomic data to be translated must be mapped onto the same reference organism. The user can then run Module 2 simply by providing the organism's name (and optionally a reference version), as well as a list of the

samples to be translated. The user can also translate samples from a VCF file, but they will need to provide a reference genome in FASTA format, to complement the variation-only information of the VCF file. The translated protein sequences are available individually but are also assembled into individual- and protein-specific datasets.

Module 3 is designed to perform a phylogenetic analysis, with some modifications needed when working with palaeo-proteomic data. The input of this module is a FASTA file, containing all of the protein sequences from both the reference dataset and the ancient sample(s) to be analyzed. The dataset is automatically split into protein specific sub-datasets, each of which will be aligned and checked for Single Amino acid Polymorphisms (SAPs). The alignment is a two step process which includes first isolating and aligning the modern/reference dataset and then aligning the ancient samples onto the modern ones using Mafft (Katoh and Standley, 2013). Isobaric amino acids that cannot be distinguished from each other by some mass spectrometers are corrected to ensure the downstream phylogenetic analysis can proceed without issues. Specifically, any time an Isoleucine (I) or a Leucine (L) is identified in the alignment, all of the modern sequences are checked for that position. If all of them share one of the 2 amino acids, then the ancient samples are also switched to that amino acid. If both I and L appear on some present-day samples, both present-day and ancient samples are switched to an L. The user also has the option to provide an additional file named 'MASKED'. Using this optional file, the user can mask a present-day sample such that it has the same missing sites as an ancient sample. Finally a small report is generated for each ancient sample in the dataset, and a maximum likelihood phylogenetic tree is generated for each protein sub-dataset through PhyML (Guindon et al., 2010). All protein alignments are then also merged together into a concatenated dataset. The concatenated dataset is used to generate a maximum-likelihood species tree (Felsenstein, 1981) through PhyML and a Bayesian species tree (Mau and Newton, 1997; Rannala and Yang, 1996) through MrBayes (Huelsenbeck and Ronquist, 2001) or RevBayes (Höhna et al., 2016). The tree generation is parallelized using Mpirun (Message Passing Interface Forum, 2021).

The modules are intended to work with each other, but can also be used independently. An in depth explanation of each step of each module, as well as the code being run in the background, is provided on the software's Github page (Patramanis et al., 2023) as well as in the supplementary material.

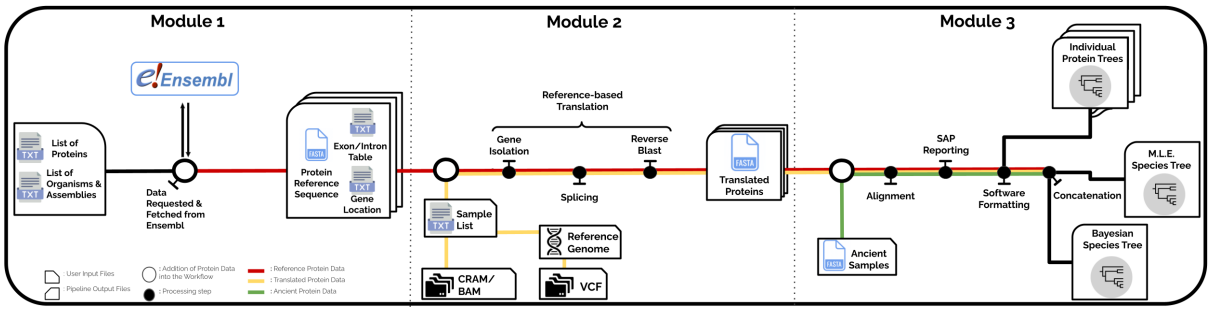

Figure 1 – Overview of the Pipeline

3. Applications

As proof of principle, we deploy this pipeline in the reconstruction of ancient hominid history using the publicly available enamel proteomes of *Homo antecessor* and *Gigantopithecus blacki*, in combination with translated genomes from hundreds of present-day and ancient hominid samples. In the process, we have generated the most complete and up to date, molecular hominid phyloproteomic tree (Figure 2). The process of generating the reference dataset and its phyloproteomic tree using PaleoProPhyler is covered in detail in the step-by-step Github Tutorial. The dataset used as input for the creation of the phylogenetic tree is available at Zenodo (Patramanis et al., 2022)

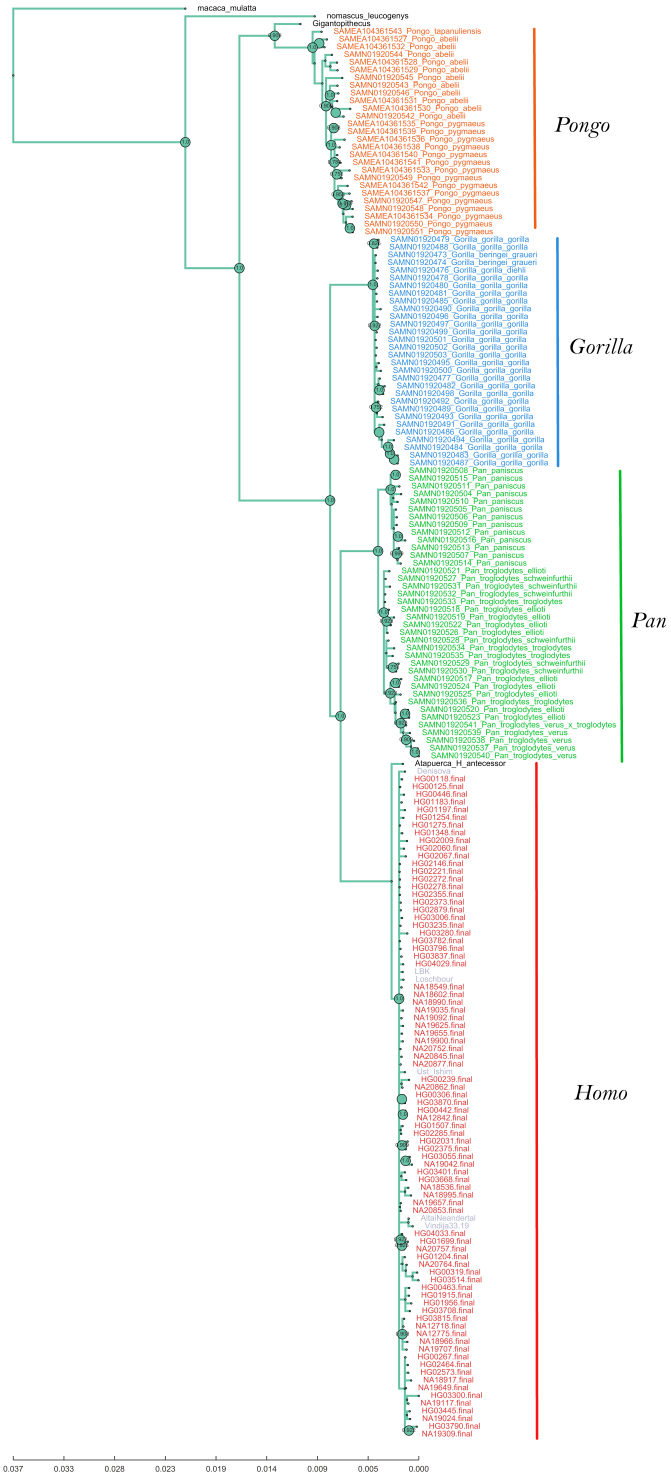

Figure 2 – Phyloproteomic tree generated using PaleoProPhyler's Module 3. The tree was constructed using 9 protein sequences obtained from enamel and includes more than 100 hominid individuals translated from genomic data, two individuals from published palaeoproteomic datasets as well as sequence data from a *Macaca* and a *Hylobates* individual, which are used to root the tree.

4. Protein Reference Dataset

In order to facilitate future analyses of ancient protein data, we also generated a publicly-available palaeoproteomic hominid reference dataset, using Modules 1 and 2. We translated 176 publicly available whole genomes from all 4 extant Hominid genera (Byrska-Bishop et al., 2021; Nater et al., 2017; Prado-Martinez et al., 2013). Details on the preparation of the translated samples can be found in the supplementary materials. We also translated multiple ancient genomes from VCF files, including those of several Neanderthals and one Denisovan (Mafessoni et al., 2020; Prüfer et al., 2017). Since the dataset is tailored for palaeoproteomic tree sequence reconstruction, we chose to translate proteins that have previously been reported as present in either teeth or bone tissue. We compiled a list of 1.696 proteins from previous studies (Acil et al., 2005; Alves et al., 2011; Castiblanco et al., 2015; Jágr et al., 2012; Park et al., 2009; Salmon et al., 2016) and successfully translated 1.543 of them. For each protein, we translated the canonical isoform as well as all alternative isoforms, leading to a total of 10.058 protein sequences for each individual in the dataset. Details on the creation of the protein list can be found in the supplementary materials. The palaeoproteomic hominid reference dataset is publicly available online at Zenodo, under the name 'Hominid Palaeoproteomic Reference Dataset' (Patramanis et al., 2022)

5. Closing remarks

The workflows presented here aim to facilitate phylogenetic reconstruction using ancient protein data to a wider audience, as well as to streamline these processes and enable greater reproducibility in the field. Although we highly encourage the use of the tools and methods utilized by our workflows, we still caution against the over interpretation of palaeoproteomic results. Deriving species relationships from ancient proteins is still a relatively new endeavor and as a result, our understanding of this data, their quantity and quality requirements, robustness and accuracy are all largely unexplored. We believe that palaeoproteomic data should therefore be used in combination with other sources of information in order to make accurate evolutionary inferences.

6. Author Contributions

- **Ioannis Patramanis**: Conceptualization, manuscript writing, code writing for the Snakemake scripts, compilation of the Conda environments and application of the pipelines to produce the results described in the 'Application' and 'Protein Reference Dataset' section.
- **Jazmin Ramos Madrigal**: Manuscript review, conceptualization and code for multiple R and bash scripts utilised by the Snakemake script as steps of the pipeline.
- **Enrico Cappellini** : Manuscript review and editing
- **Fernando Racimo** : Conceptualization, manuscript writing, review and editing

7. Acknowledgements

We thank Ryan Sinclair Paterson, Graham Gower, Alberto Taurozzi, Martin Petr, Evan Irving-Pease and other members of the Racimo and Cappellini groups, who provided valuable help and feedback throughout the project. We also thank Helen Fewlass for testing and identifying errors in the workflow. Preprint version 3 of this article has been peer-reviewed and recommended by Peer Community In Paleontology (https://doi.org/10.24072/pci.paleo.100220; Hlusko, 2023).

8. Conflicts of Interest

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. The authors declare the following non-financial conflict of interest: Fernando Racimo is a recommender for PCI Evolutionary Biology.

9. Data Availability

The Protein Reference Dataset is available on Zenodo:
10.5281/zenodo.7728060 (Patramanis et al., 2022)

PaleoProPhyler is publicly available on Github and Zenodo:
10.5281/zenodo.10122365 (Patramanis et al., 2023)

The tool requires a Linux OS (Operating System) and the installation of Conda. The github repository contains a tutorial for using the workflow presented here, with the proteins recovered from the *Homo antecessor* and *Gigantopithecus blacki* as examples. We welcome code contributions, feature requests, and bug reports via Github. The software is released under a CC-BY license.

10. 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. FR was additionally supported by a Villum Young Investigator Grant (project no. 00025300), a COREX ERC Synergy grant (ID 951385) and a Novo Nordisk Fonden Data Science Ascending Investigator Award (NNF22OC0076816). 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).

References

Acil Y, Mobasseri AE, Warnke PH, Terheyden H, Wiltfang J, Springer I (2005). *Detection of Mature Collagen in Human Dental Enamel*. *Calcified tissue international* **76**, 121–126. https://doi.org/10.1007/s00223-004-0122-0.

Alves RD, Demmers JA, Bezstarosti K, van der Eerden BC, Verhaar JA, Eijken M, van Leeuwen JP (2011). *Unraveling the Human Bone Microenvironment beyond the Classical Extracellular Matrix Proteins: A Human Bone Protein Library*. *Journal of proteome research* **10**, 4725–4733. https://doi.org/10.1021/pr200522n.

Bagheri H, Severin AJ, Rajan H (2020). *Detecting and Correcting Misclassified Sequences in the Large-Scale Public Databases*. *Bioinformatics (Oxford, England)* **36**, 4699–4705. https://doi.org/10.1093/bioinformatics/btaa586.

Bonfield JK (2022). *CRAM 3.1: Advances in the CRAM File Format*. *Bioinformatics (Oxford, England)* **38**, 1497–1503. https://doi.org/10.1093/bioinformatics/btac010.

Brandt LØ, Taurozzi AJ, Mackie M, Sinding MHS, Vieira FG, Schmidt AL, Rimstad C, Collins MJ, Mannering U (2022). *Palaeoproteomics Identifies Beaver Fur in Danish High-Status Viking Age Burials-Direct Evidence of Fur Trade*. *Plos one* **17**, e0270040. https://doi.org/10.1371/journal.pone.0270040.

Brown S, Higham T, Slon V, Pääbo S, Meyer M, Douka K, Brock F, Comeskey D, Procopio N, Shunkov M, et al. (2016). *Identification of a New Hominin Bone from Denisova Cave, Siberia Using Collagen Fingerprinting and Mitochondrial DNA Analysis*. *Scientific reports* **6**, 23559. https://doi.org/10.1038/srep23559.

Brown S, Massilani D, Kozlikin MB, Shunkov MV, Derevianko AP, Stoessel A, Jope-Street B, Meyer M, Kelso J, Pääbo S, et al. (2022). *The Earliest Denisovans and Their Cultural Adaptation*. *Nature ecology & evolution* **6**, 28–35. https://doi.org/10.1038/s41559-021-01581-2.

Buckley M (2015). *Ancient Collagen Reveals Evolutionary History of the Endemic South American 'Ungulates'*. *Proceedings of the Royal Society B: Biological Sciences* **282**, 20142671. https://doi.org/10.1098/rspb.2014.2671.

Buckley M, Collins M, Thomas-Oates J, Wilson JC (2009). *Species Identification by Analysis of Bone Collagen Using Matrix-Assisted Laser Desorption/Ionisation Time-of-Flight Mass Spectrometry*. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry* **23**, 3843–3854. https://doi.org/10.1002/rcm.4316.

Buckley M, Lawless C, Rybczynski N (2019). *Collagen Sequence Analysis of Fossil Camels, Camelops and Cf Paracamelus, from the Arctic and Sub-Arctic of Plio-Pleistocene North America*. *Journal of proteomics* **194**, 218–225. https://doi.org/10.1016/j.jprot.2018.11.014.

Buckley M, Kansa SW, Howard S, Campbell S, Thomas-Oates J, Collins M (2010). *Distinguishing between Archaeological Sheep and Goat Bones Using a Single Collagen Peptide*. *Journal of Archaeological Science* **37**, 13–20. https://doi.org/10.1016/j.jas.2009.08.020.

Byrska-Bishop M, Evani ¸Zhao X, Basile ¸Abel ¸Regier ¸Corvelo A, Clarke ¸Musunuri R, Nagulapalli K, et al. (2021). *High Coverage Whole Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios. bioRxiv. 2021*. *Publisher Full Text*. https://doi.org/10.1101/2021.02.06.430068.

Cappellini E, Collins MJ, Gilbert MTP (2014). *Unlocking Ancient Protein Palimpsests*. *Science (New York, N.Y.)* **343**, 1320–1322. https://doi.org/10.1126/science.1249274.

Cappellini E, Jensen LJ, Szklarczyk D, Ginolhac A, da Fonseca RA, Stafford Jr TW, Holen SR, Collins MJ, Orlando L, Willerslev E, et al. (2012). *Proteomic Analysis of a Pleistocene Mammoth Femur Reveals More than One Hundred Ancient Bone Proteins*. *Journal of proteome research* **11**, 917–926. https://doi.org/10.1021/pr200721u.

Cappellini E, Welker F, Pandolfi L, Ramos-Madrigal J, Samodova D, Rüther PL, Fotakis AK, Lyon D, Moreno-Mayar JV, Bukhsianidze M, et al. (2019). *Early Pleistocene Enamel Proteome from Dmanisi Resolves Stephanorhinus Phylogeny*. *Nature* **574**, 103–107. https://doi.org/10.1038/s41586-019-1555-y.

Castiblanco GA, Rutishauser D, Ilag LL, Martignon S, Castellanos JE, Mejía W (2015). *Identification of Proteins from Human Permanent Erupted Enamel*. *European journal of oral sciences* **123**, 390–395. https://doi.org/10.1111/eos.12214.

Chen F, Welker F, Shen CC, Bailey SE, Bergmann I, Davis S, Xia H, Wang H, Fischer R, Freidline SE, et al. (2019). *A Late Middle Pleistocene Denisovan Mandible from the Tibetan Plateau*. *nature* **569**, 409–412. https://doi.org/10.1038/s41586-019-1139-x.

Chi H, Liu C, Yang H, Zeng WF, Wu L, Zhou WJ, Niu XN, Ding YH, Zhang Y, Wang RM, et al. (2018). *Open-pFind Enables Precise, Comprehensive and Rapid Peptide Identification in Shotgun Proteomics*. *bioRxiv : the preprint server for biology*, 285395. https://doi.org/10.1101/285395.

Cox J, Mann M (2008). *MaxQuant Enables High Peptide Identification Rates, Individualized Ppb-Range Mass Accuracies and Proteome-Wide Protein Quantification*. *Nature biotechnology* **26**, 1367–1372. https://doi.org/10.1038/nbt.1511.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. (2011). *The Variant Call Format and VCFtools*. *Bioinformatics (Oxford, England)* **27**, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Demarchi B, Hall S, Roncal-Herrero T, Freeman CL, Woolley J, Crisp MK, Wilson J, Fotakis A, Fischer R, Kessler BM, et al. (2016). *Protein Sequences Bound to Mineral Surfaces Persist into Deep Time*. *elife* **5**. https://doi.org/10.7554/eLife.17092.

Demarchi B, Stiller J, Grealy A, Mackie M, Deng Y, Gilbert T, Clarke J, Legendre LJ, Boano R, Sicheritz-Pontén T, et al. (2022). *Ancient Proteins Resolve Controversy over the Identity of Genyornis Eggshell*. *Proceedings of the National Academy of Sciences*, e2109326119. https://doi.org/10.1073/pnas.2109326119.

Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M (2018). *DIA-NN: Neural Networks and Interference Correction Enable Deep Coverage in High-Throughput Proteomics*. *bioRxiv : the preprint server for biology*, 282699. https://doi.org/10.1038/s41592-019-0638-x.

Devièse T, Karavanić I, Comeskey D, Kubiak C, Korlević P, Hajdinjak M, Radović S, Procopio N, Buckley M, Pääbo S, et al. (2017). *Direct Dating of Neanderthal Remains from the Site of Vindija Cave and Implications for the Middle to Upper Paleolithic Transition*. *Proceedings of the National Academy of Sciences* **114**, 10606–10611. https://doi.org/10.1073/pnas.1709235114.

Felsenstein J (1981). *Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach*. *Journal of molecular evolution* **17**, 368–376. https://doi.org/10.1007/BF01734359.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010). *New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0*. *Systematic biology* **59**, 307–321. https://doi.org/10.1093/sysbio/syq010.

Heath TA, Hedtke SM, Hillis DM (2008). *Taxon Sampling and the Accuracy of Phylogenetic Analyses*. *Journal of systematics and evolution* **46**, 239–257. https://doi.org/10.3724/SP.J.1002.2008.08016.

Hendy J (2021). *Ancient Protein Analysis in Archaeology*. *Science Advances* **7**, eabb9314. https://doi.org/10.1126/sciadv.abb9314.

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F (2016). *RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language*. *Systematic biology* **65**, 726–736. https://doi.org/10.1080/10618600.1997.10474731.

Huelsenbeck JP, Ronquist F (2001). *MRBAYES: Bayesian Inference of Phylogenetic Trees*. *Bioinformatics (Oxford, England)* **17**, 754–755. https://doi.org/10.1093/bioinformatics/17.8.754.

Inc. A (2020). *Anaconda Software Distribution*. https://docs.anaconda.com/. Version Vers. 2-2.4.0.

Jágr M, Eckhardt A, Pataridis S, Mikšík I (2012). *Comprehensive Proteomic Analysis of Human Dentin*. *European journal of oral sciences* **120**, 259–268. https://doi.org/10.1111/j.1600-0722.2012.00977.x.

Katoh K, Standley DM (2013). *MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability*. *Molecular biology and evolution* **30**, 772–780. https://doi.org/10.1093/molbev/mst010.

Koepfli KP, Paten B, Scientists G1C, O'Brien SJ (2015). *The Genome 10K Project: A Way Forward*. *Annual Review of Animal Biosciences* **3**, 57–111. https://doi.org/10.1146/annurev-animal-090414-014900.

Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI (2017). *MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–Based Proteomics*. *Nature methods* **14**, 513–520. https://doi.org/10.1038/nmeth.4256.

Lan T, Lindqvist C (2018). *Technical Advances and Challenges in Genome-Scale Analysis of Ancient DNA*. *Paleogenomics*, 3–29. https://doi.org/10.1007/13836_2018_54.

Lanigan LT, Mackie M, Feine S, Hublin JJ, Schmitz RW, Wilcke A, Collins MJ, Cappellini E, Olsen JV, Taurozzi AJ, et al. (2020). *Multi-Protease Analysis of Pleistocene Bone Proteomes*. *Journal of proteomics* **228**, 103889. https://doi.org/10.1016/j.jprot.2020.103889.

Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, et al. (2018). *Earth BioGenome Project: Sequencing Life for the Future of Life*. *Proceedings of the National Academy of Sciences* **115**, 4325–4333. https://doi.org/10.1073/pnas.1720115115.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009). *The sequence alignment/map format and SAMtools*. *Bioinformatics* **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Lipman DJ, Pearson WR (1985). *Rapid and Sensitive Protein Similarity Searches*. *Science (New York, N.Y.)* **227**, 1435–1441. https://doi.org/10.1126/science.2983426.

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G (2003). *PEAKS: Powerful Software for Peptide de Novo Sequencing by Tandem Mass Spectrometry*. *Rapid communications in mass spectrometry* **17**, 2337–2342. https://doi.org/10.1002/rcm.1196.

Mafessoni F, Grote S, de Filippo C, Slon V, Kolobova KA, Viola B, Markin SV, Chintalapati M, Peyrégne S, Skov L, et al. (2020). *A High-Coverage Neandertal Genome from Chagyrskaya Cave*. *Proceedings of the National Academy of Sciences* **117**, 15132–15136. https://doi.org/10.1073/pnas.2004944117.

Martin FJ (2023). *Ensembl 2023*. https://www.ensembl.org/info/about/species.html.

Mau B, Newton MA (1997). *Phylogenetic Inference for Binary Data on Dendograms Using Markov Chain Monte Carlo*. *Journal of Computational and Graphical Statistics* **6**, 122–131. https://doi.org/10.1080/10618600.1997.10474731.

Message Passing Interface Forum (2021). *MPI: A Message-Passing Interface Standard Version 4.0*. Manual.

Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Tward žiok SO, Kanitz A, et al. (2021). *Sustainable Data Analysis with Snakemake*. *F1000Research* **10**. https://doi.org/10.12688/f1000research.29032.2.

Müller JB, Geyer PE, Colaço AR, Treit PV, Strauss MT, Oroshi M, Doll S, Virreira Winter S, Bader JM, Köhler N, et al. (2020). *The Proteome Landscape of the Kingdoms of Life*. *Nature* **582**, 592–596. https://doi.org/10.1038/s41586-020-2402-x.

Nater A, Mattle-Greminger MP, Nurcahyo A, Nowak MG, De Manuel M, Desai T, Groves C, Pybus M, Sonay TB, Roos C, et al. (2017). *Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species*. *Current Biology* **27**, 3487–3498. https://doi.org/10.1016/j.cub.2017.09.047.

National Center for Biotechnology Information (NCBI) (2004). *Genome*. *Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information*.

Nielsen-Mars CM, Stegemann C, Hoffmann R, Smith T, Feeney R, Toussaint M, Harvati K, Panago þoulou E, Hublin JJ, Richards MP (2009). *Extraction and Sequencing of Human and Neanderthal Mature Enamel Proteins Using MALDI-TOF/TOF MS*. *Journal of Archaeological Science* **36**, 1758–1763. https://doi.org/10.1016/j.jas.2009.04.004.

Nogueira FC, Neves LX, Pessoa-Lima C, Langer MC, Domont GB, Line SRP, Leme AFP, Gerlach RF (2021). *Ancient Enamel Peptides Recovered from the South American Pleistocene Species Notiomastodon Platensis and Myocastor Cf. Coypus*. *Journal of Proteomics* **240**, 104187. https://doi.org/10.1016/j.jprot.2021.104187.

Opperdoes FR (2003). *Phylogenetic Analysis Using Protein Sequences*. *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Ed. by and Vandamme A-M Lemey P Salemi M, 207–235. https://doi.org/10.1017/CBO9780511819049.011.

Ostrom PH, Schall M, Gandhi H, Shen TL, Hauschka PV, Strahler JR, Gage DA (2000). *New Strategies for Characterizing Ancient Proteins Using Matrix-Assisted Laser Desorption Ionization Mass Spectrometry*. *Geochimica et Cosmochimica Acta* **64**, 1043–1050. https://doi.org/10.1016/S0016-7037(99)00381-6.

Park ES, Cho HS, Kwon TG, Jang SN, Lee SH, An CH, Shin HI, Kim JY, Cho JY (2009). *Proteomics Analysis of Human Dentin Reveals Distinct Protein Expression Profiles*. *Journal of proteome research* **8**, 1338–1346. https://doi.org/10.1021/pr801065s.

Patramanis I, Madrigal JR, Cappellini E, Racimo F (2022). *Hominid Palaeoproteomic Reference Dataset - v1.0.1*. https://doi.org/10.5281/zenodo.7728060.

Patramanis I, Madrigal JR, Cappellini E, Racimo F (2023). *Palaeoprophyler - Public Release v1.0.1*. https://doi.org/10.5281/zenodo.10122365.

Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, Nieselt K (2016). *EAGER: Efficient Ancient Genome Reconstruction*. *Genome biology* **17**, 1–14. https://doi.org/10.1186/s13059-016-0918-z.

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999). *Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data*. *ELECTROPHORESIS: An International Journal* **20**, 3551–3567. https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

Porto IM, Laure HJ, de Sousa FB, Rosa JC, Gerlach RF (2011). *New Techniques for the Recovery of Small Amounts of Mature Enamel Proteins*. *Journal of Archaeological Science* **38**, 3596–3604. https://doi.org/10.1016/j.jas.2011.08.030.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'connor TD, Santpere G, et al. (2013). *Great Ape Genetic Diversity and Population History*. *Nature* **499**, 471–475. https://doi.org/10.1038/nature12228.

Presslee S, Slater GJ, Pujos F, Forasiepi AM, Fischer R, Molloy K, Mackie M, Olsen JV, Kramarz A, Taglioretti M, et al. (2019). *Data from: Palaeoproteomics Resolves Sloth Phylogeny*. https://doi.org/10.1038/s41559-019-0909-z.

Prüfer K, De Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, et al. (2017). *A High-Coverage Neandertal Genome from Vindija Cave in Croatia*. *Science (New York, N.Y.)* **358**, 655–658. https://doi.org/10.1126/science.aao1887.

Rannala B, Yang Z (1996). *Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference*. *Journal of molecular evolution* **43**, 304–311. https://doi.org/10.1007/BF02338839.

Rosenberg MS, Kumar S (2003). *Taxon Sampling, Bioinformatics, and Phylogenomics*. *Systematic Biology* **52**, 119. https://doi.org/10.1080/10635150390132894.

Rüther PL, Husic IM, Bangsgaard P, Gregersen KM, Pantmann P, Carvalho M, Godinho RM, Friedl L, Cascalheira J, Taurozzi AJ, et al. (2022). *SPIN Enables High Throughput Species Identification of Archaeological Bone by Proteomics*. *Nature communications* **13**, 1–14. https://doi.org/10.1038/s41467-022-30097-x.

Rybczynski N, Gosse JC, Richard Harington C, Wogelius RA, Hidy AJ, Buckley M (2013). *Mid-Pliocene Warm-Period Deposits in the High Arctic Yield Insight into Camel Evolution*. *Nature communications* **4**, 1550. https://doi.org/10.1038/ncomms2516.

Salmon CR, Giorgetti APO, Leme AFP, Domingues RR, Sallum EA, Alves MC, Kolli TN, Foster BL, Nociti Jr FH (2016). *Global Proteome Profiling of Dental Cementum under Experimentally-Induced Apposition*. *Journal of proteomics* **141**, 12–23. https://doi.org/10.1016/j.jprot.2016.03.036.

Schubert M, Ermini L, Sarkissian CD, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M, et al. (2014). *Characterization of Ancient and Modern Genomes by SNP Detection and Phylogenomic and Metagenomic Analysis Using PALEOMIX*. *Nature protocols* **9**, 1056–1082. https://doi.org/10.1038/nprot.2014.063.

Solntsev SK, Shortreed MR, Frey BL, Smith LM (2018). *Enhanced Global Post-Translational Modification Discovery with MetaMorpheus*. *Journal of proteome research* **17**, 1844–1851. https://doi.org/10.1021/acs.jproteome.7b00873.

Uniprot consortium (2021). *UniProt Readme file*. URL: ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/README (visited on 08/14/2023).

Warinner C, Korzow Richter K, Collins MJ (2022). *Paleoproteomics*. *Chemical Reviews*. https://doi.org/10.1021/acs.chemrev.1c00703.

Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, Radini A, Hancock Y, Tito RY, Fiddyment S, et al. (2014). *Pathogens and Host Immunity in the Ancient Human Oral Cavity*. *Nature genetics* **46**, 336–344. https://doi.org/10.1038/ng.2906.

Welker F, Hajdinjak M, Talamo S, Jaouen K, Dannemann M, David F, Julien M, Meyer M, Kelso J, Barnes I, et al. (2016). *Palaeoproteomic Evidence Identifies Archaic Hominins Associated with the Châtelperronian at the Grotte Du Renne*. *Proceedings of the National Academy of Sciences* **113**, 11162–11167. https://doi.org/10.1073/pnas.1605834113.

Welker F, Ramos-Madrigal J, Gutenbrunner P, Mackie M, Tiwary S, Rakownikow Jersie-Christensen R, Chiva C, Dickinson MR, Kuhlwilm M, de Manuel M, et al. (2020). *The Dental Proteome of Homo Antecessor*. *Nature* **580**, 235–238. https://doi.org/10.1038/s41586-020-2153-8.

Welker F, Ramos-Madrigal J, Kuhlwilm M, Liao W, Gutenbrunner P, de Manuel M, Samodova D, Mackie M, Allentoft ME, Bacon AM, et al. (2019). *Enamel Proteome Shows That Gigantopithecus Was an Early Diverging Pongine*. *Nature* **576**, 262–265. https://doi.org/10.1038/s41586-019-1728-8.

Welker F, Smith GM, Hutson JM, Kindler L, Garcia-Moreno A, Villaluenga A, Turner E, Gaudzinski-Windheuser S (2017). *Middle Pleistocene Protein Sequences from the Rhinoceros Genus Stephanorhinus and the Phylogeny of Extant and Extinct Middle/Late Pleistocene Rhinocerotidae*. *PeerJ* **5**, e3033. https://doi.org/10.7717/peerj.3033.

Yates A, Beal K, Keenan S, McLaren W, Pignatelli M, Ritchie GR, Ruffier M, Taylor K, Vullo A, Flicek P (2015). *The Ensembl REST API: Ensembl Data for Any Language*. *Bioinformatics (Oxford, England)* **31**, 143–145. https://doi.org/10.1093/bioinformatics/btu613.

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al. (2014). *Comparative Genomics Reveals Insights into Avian Genome Evolution and Adaptation*. *Science (New York, N.Y.)* **346**, 1311–1320. https://doi.org/10.1126/science.1251385.