# PaleoProPhyler: a reproducible pipeline for phylogenetic inference using ancient proteins

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

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

December 12, 2022

## Summary

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[1, 2, 3, 4, 5, 6]. 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.

## Statement of Need

Recent advances in protein extraction and mass spectrometry [7, 8, 9] 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 [10, 11, 12, 13]. The sequences of these proteins 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 in particular have been successfully extracted from multiple extinct species, in order to resolve their relationships to other species [1, 2, 4, 6, 14, 5].

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 [15] and EAGER [16] 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 [17]. 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 rel-

ative scarcity of proteomic datasets [18, 19]. There are currently tens of thousands of publicly available whole genome sequences, covering hundreds of species [20, 21, 22, 23, 24]. The amount of publicly available proteome sequences is much smaller in comparison. For most vertebrate species, lab-generated protein data does not even 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 [25]. 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 [26, 27], 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 even a single amino acid polymorphism (SAP) can strongly affect downstream inferences [3, 28, 29, 30].

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 [2, 4, 6], 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 seuqences in FASTA format, and various forms of phylogenetic tree reconstruction.

## Description of the Pipeline

To maximize reproducibility, accessibility and scalability, we have built our pipeline using Snakemake [31] and Conda [32]. 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 1. The modules are intended to synergize 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 as well as in the supplementary material.

## Application

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 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'

## 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 204 publicly available whole genomes from all 4 extant Hominid genera [21, 22, 33]. 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 [34, 35]. 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 1696 proteins from previous studies [36, 37, 38, 39, 40, 41] 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'

## Availability and Community Guidelines

PaleoProPhyler is publicly available on github: The software requires the prior 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

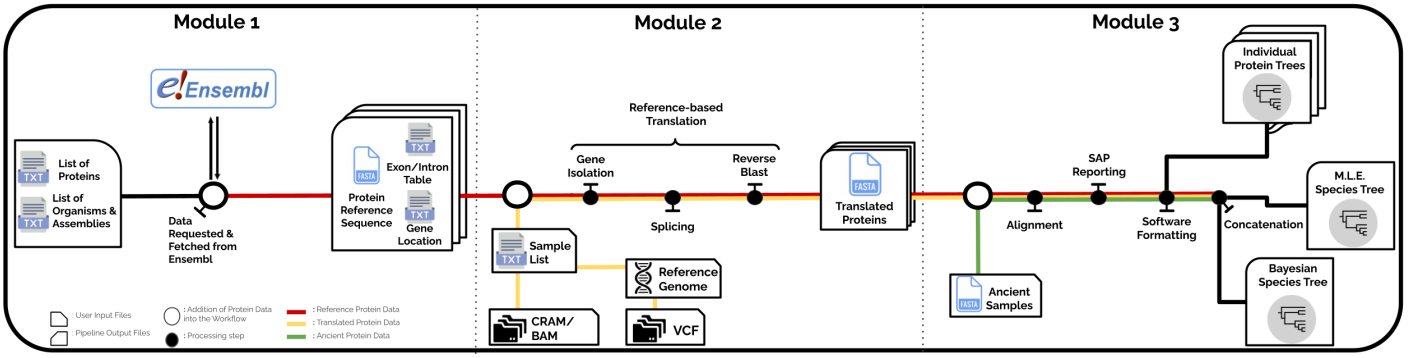


Figure 1: Overview of the Pipeline

reports via Github. The software is released under a CC-BY license.

## 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

## 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.

## 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 Eu- ropean Union's Horizon 2020 research and innovation program (grant agreement No. 101021361).

references

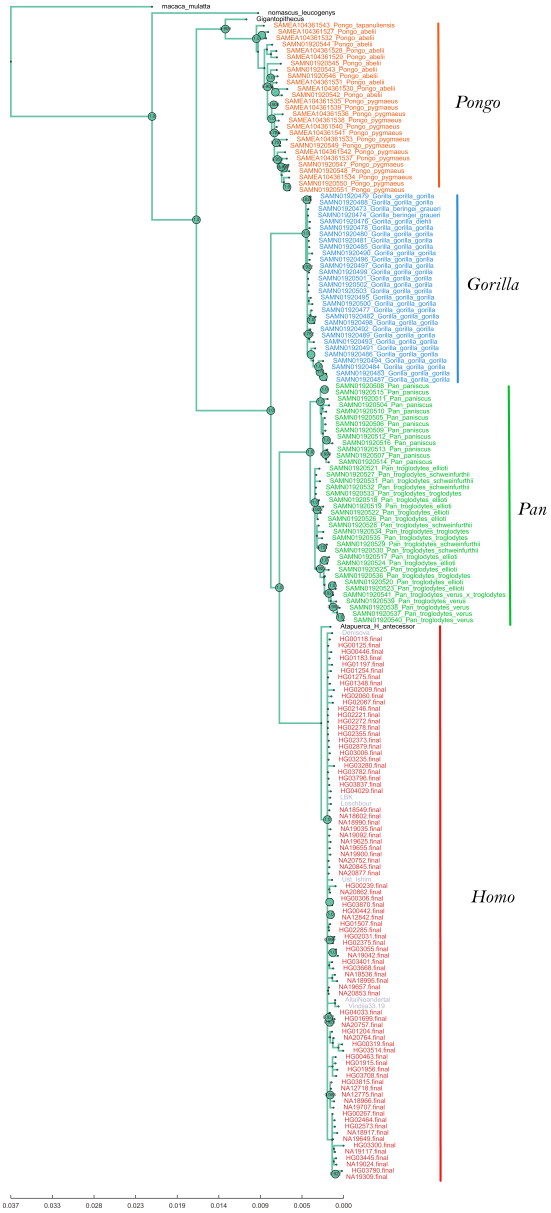


Figure 2: Phyloproteomic tree generated using PaleoProPhyler's Module 3. The tree was constructed using 9 protein sequences obtained from enamel and includes more than 100 hominid individuals translated from genomic data, two individuals from published palaeoproteomic datasets as well as sequence data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

# Bibliography


[1] Frido Welker et al. "Middle Pleistocene protein sequences from the rhinoceros genus Stephanorhinus and the phylogeny of extant and extinct Middle/Late Pleistocene Rhinocerotidae". In: *PeerJ* 5 (2017), e3033.

[2] Enrico Cappellini et al. "Early Pleistocene enamel proteome from Dmanisi resolves Stephanorhinus phylogeny". In: *Nature* 574.7776 (2019), pp. 103–107.

[3] Fahu Chen et al. "A late middle Pleistocene Denisovan mandible from the Tibetan Plateau". In: *nature* 569.7756 (2019), pp. 409–412.

[4] Frido Welker et al. "Enamel proteome shows that Gigantopithecus was an early diverging pongine". In: *Nature* 576.7786 (2019), pp. 262–265.

[5] 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.

[6] Frido Welker et al. "The dental proteome of Homo antecessor". In: *Nature* 580.7802 (2020), pp. 235–238.

[7] Patrick Leopold Rüther et al. "SPIN enables high throughput species identification of archaeological bone by proteomics". In: *Nature communications* 13.1 (2022), pp. 1–14.

[8] Liam T Lanigan et al. "Multi-protease analysis of Pleistocene bone proteomes". In: *Journal of proteomics* 228 (2020), p. 103889.

[9] Isabel Maria Porto et al. "New techniques for the recovery of small amounts of mature enamel proteins". In: *Journal of Archaeological Science* 38.12 (2011), pp. 3596–3604.

[10] Enrico Cappellini, Matthew J Collins, and M Thomas P Gilbert. "Unlocking ancient protein palimpsests". In: *Science* 343.6177 (2014), pp. 1320–1322.

[11] Beatrice Demarchi et al. "Protein sequences bound to mineral surfaces persist into deep time". In: *elife* 5 (2016).

[12] Jessica Hendy. "Ancient protein analysis in archaeology". In: *Science Advances* 7.3 (2021), eabb9314.

[13] Christina Warinner, Kristine Korzow Richter, and Matthew J Collins. "Paleoproteomics". In: *Chemical Reviews* (2022).

[14] Michael Buckley. "Ancient collagen reveals evolutionary history of the endemic South American 'ungulates'". In: *Proceedings of the Royal Society B: Biological Sciences* 282.1806 (2015), p. 20142671.

[15] Mikkel Schubert et al. "Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX". In: *Nature protocols* 9.5 (2014), pp. 1056–1082.

[16] Alexander Peltzer et al. "EAGER: efficient ancient genome reconstruction". In: *Genome biology* 17.1 (2016), pp. 1–14.

[17] Tianying Lan and Charlotte Lindqvist. "Technical advances and challenges in genome-scale analysis of ancient DNA". In: *Paleogenomics* (2018), pp. 3–29.

[18] Johannes B Müller et al. "The proteome landscape of the kingdoms of life". In: *Nature* 582.7813 (2020), pp. 592–596.

[19] Luise Ørsted Brandt et al. "Palaeoproteomics identifies beaver fur in Danish high-status Viking Age burials-direct evidence of fur trade". In: *Plos one* 17.7 (2022), e0270040.

[20] Harris A Lewin et al. "Earth BioGenome Project: Sequencing life for the future of life". In: *Proceedings of the National Academy of Sciences* 115.17 (2018), pp. 4325–4333.

[21] M Byrska-Bishop et al. "High Coverage Whole Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios. bioRxiv. 2021". In: *Publisher Full Text* ().

[22] Javier Prado-Martinez et al. "Great ape genetic diversity and population history". In: *Nature* 499.7459 (2013), pp. 471–475.

[23] Guojie Zhang et al. "Comparative genomics reveals insights into avian genome evolution and adaptation". In: *Science* 346.6215 (2014), pp. 1311–1320.

[24] Klaus-Peter Koepfli et al. "The Genome 10K Project: a way forward". In: *Annu. Rev. Anim. Biosci.* 3.1 (2015), pp. 57–111.

[25] Hamid Bagheri, Andrew J Severin, and Hridesh Rajan. "Detecting and correcting misclassified sequences in the large-scale public databases". In: *Bioinformatics* 36.18 (2020), pp. 4699–4705.

[26] Michael S Rosenberg and Sudhir Kumar. "Taxon sampling, bioinformatics, and phylogenomics". In: *Systematic Biology* 52.1 (2003), p. 119.

[27] Tracy A Heath, Shannon M Hedtke, and David M Hillis. "Taxon sampling and the accuracy of phylogenetic analyses". In: *Journal of systematics and evolution* 46.3 (2008), pp. 239–257.

[28] Fred R Opperdoes. "Phylogenetic analysis using protein sequences". In: *The phylogenetics handbook a practical approach to DNA and protein phylogeny* (2003), pp. 207–235.

[29] Samantha Presslee et al. "Data from: Palaeoproteomics resolves sloth phylogeny". In: (2019).

[30] Beatrice Demarchi et al. "Ancient proteins resolve controversy over the identity of Genyornis eggshell". In: *Proceedings of the National Academy of Sciences* (2022), e2109326119.

[31] Felix Mölder et al. "Sustainable data analysis with Snakemake". In: *F1000Research* 10 (2021).

[32] *Anaconda Software Distribution*. Version Vers. 2-2.4.0. 2020. URL: https://docs.anaconda.com/.

[33] Alexander Nater et al. "Morphometric, behavioral, and genomic evidence for a new orangutan species". In: *Current Biology* 27.22 (2017), pp. 3487–3498.

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

[35] Fabrizio Mafessoni et al. "A high-coverage Neandertal genome from Chagyrskaya Cave". In: *Proceedings of the National Academy of Sciences* 117.26 (2020), pp. 15132–15136.

[36] Gina A Castiblanco et al. "Identification of proteins from human permanent erupted enamel". In: *European journal of oral sciences* 123.6 (2015), pp. 390–395.

[37] Rodrigo DAM Alves et al. "Unraveling the human bone microenvironment beyond the classical extracellular matrix proteins: a human bone protein library". In: *Journal of proteome research* 10.10 (2011), pp. 4725–4733.

[38] Yahya Acil et al. "Detection of mature collagen in human dental enamel". In: *Calcified tissue international* 76.2 (2005), pp. 121–126.

[39] Cristiane R Salmon et al. "Global proteome profiling of dental cementum under experimentally-induced apposition". In: *Journal of proteomics* 141 (2016), pp. 12–23.

[40] Michal Jágr et al. "Comprehensive proteomic analysis of human dentin". In: *European journal of oral sciences* 120.4 (2012), pp. 259–268.

[41] Eun-Sung Park et al. "Proteomics analysis of human dentin reveals distinct protein expression profiles". In: *Journal of proteome research* 8.3 (2009), pp. 1338–1346.