PaleoProPhyler: a reproducible pipeline for phylogenetic reconstruction using ancient proteins

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

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

November 10, 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 respective position in the tree of life. 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 genomic and genetic data in different formats (CRAM, BAM, VCF) and translate it, allowing the user to create reference panels for phylogenetic 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 robustly analyze palaeoproteomic sequences, enabling them to derive powerful insights from this valuable source of evolutionary data.

PaleoProPhyler is released under the XXX license and available from Github

Statement of Need

Recent advances in protein extraction and mass spectrometry [52, 31, 46] have made it possible to isolate ancient proteins from organisms that lived thousands, or even millions, of years ago. When compared to ancient DNA, certain ancient proteins have a lower degradation rate and can be preserved for longer [11, 18, 24, 56]. The sequences of these proteins contain phylogenetic information and thus have the potential to resolve important scientific questions about the deep past, which are not approachable via other methods, like ancient genomics. Tooth enamel proteins in particular have been successfully extracted from multiple extinct species, in order to resolve their relationship to other species [59, 12, 58, 57, 7, 8].

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. In paleogenomics, the creation of automated pipelines like PALEOMIX [54] and EAGER [44] have facilitated the streamlining and expansion of data analysis, particularly in new and emerging research groups around the world. This has undoubtedly contributed to the growth of the field [30]. Yet, the field of paleoproteomics still lacks a "democratizing" tool that is approachable to researchers of different backgrounds and

expertises.

Another important issue in phylo-proteomics is the relative scarcity of proteomic datasets [40, 6]. There are currently tens of thousands of sequenced whole genomes, covering hundreds of species and subspecies, that are publicly available [32, 9, 47, 61, 28]. The amount of in-vitro generated proteomes is merely a fraction of that. For most vertebrate species, lab generated protein data doesn't even exist and researchers are reliant on sequences translated in silico from genomic data, in order to perform their analysis. These, more often than not, are not sufficiently validated or curated, may correspond to alternative isoforms or may simply be outright erroneous data [4] . As a result, assembling a proper reference dataset for a protein based tree inference can be challenging. Given how important rigorous taxon sampling is in performing proper phylogenetic reconstruction [51, 23], having a complete and reliable reference dataset is crucial. Reference database issues are even more important when one wants to conduct a phylo-proteomic analysis. Here the lack of sequence diversity, due to forces of selection, means that the presence or lack of a single amino acid polymorphism (SAP) in the dataset can affect evolutionary history estimations [42, 48, 19, 14].

To address all of the above issues, we present "Paleoprophyler": a fully reproducible and easily deployable pipeline with the aim to assist researchers in phylo-proteomic analyses of ancient peptides. "Paleoprophyler" is based on the workflows performed in [12, 58, 57]. It allows for the search and access of available reference proteomes, bulk translation of CRAM, BAM or VCF files into FASTA format amino acid sequences and extensive phylogenetic tree reconstruction. Below we provide a description of the pipeline, including its most important features and functionalities.

Description of the Pipeline

To maximize reproducibility, accessibility and scalability, we have built our pipeline using Snakemake [39] and Conda [1]. 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 easy access and portability. The pipeline is divided into three distinct but interacting modules, each of which is composed of a Snakemake script and a Conda environment. The modules are intended to synergize with each other, but can also be used independently.

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 this module is a user-provided list of proteins and a list of organisms. Both input lists should be in a simple TXT format. The user also has the option of choosing a particular reference build. Utilizing the Ensembl API [60], the module will return 3 different resources for each requested protein and for each requested organism / reference build. The 3 key resources are: a) The reference protein sequence in FASTA format [34]. b) The location (position and strand) of the gene that corresponds to the protein. c) The start and end of each exon and intron of that gene/isoform.

The downloaded FASTA sequences are available individually but will also be assembled into species- and protein-specific datasets and can be immediately used as a reference dataset for either phylogenetic purposes or as an input database for mass spectrometry software, like MaxQuant [16], Pfind [15], PEAKS [35] and others [20, 29, 55, 45]. The gene location information and the exon/intron tables are stored and can be utilized automatically by Module 2. For the requested proteins, 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 protein list TXT file.

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. Module 2 can handle some of the most commonly used genomic data file formats, including the BAM [33], CRAM [5] and VCF [17] 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 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 reference version), as well as a list of the samples to be translated, both in TXT format. Should the user want to translate samples

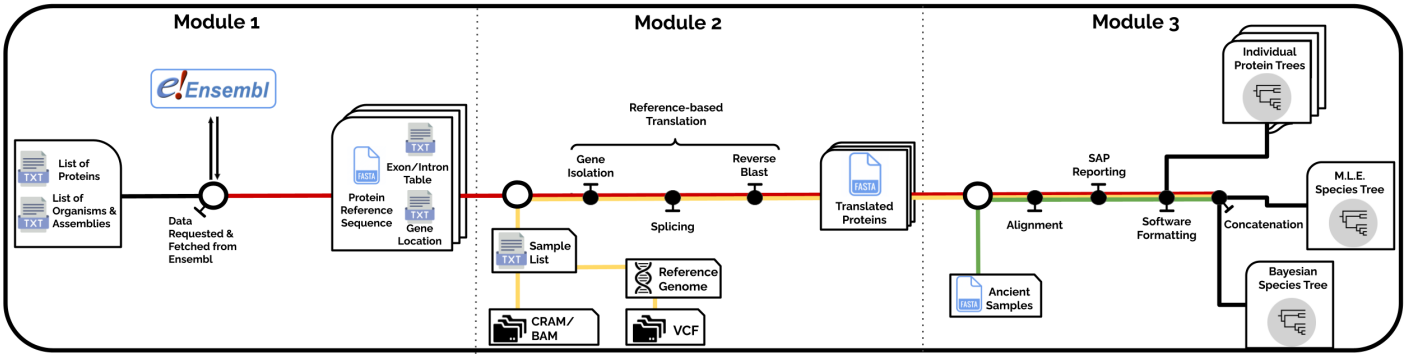

Figure 1: Pipeline Overview

from a VCF file, they will also need to provide a reference genome in a FASTA format, to complement the variation information of the VCF file. After executing the module, an initial 'normalization' step is performed, where all input files are formatted and indexed. Once this is complete, the locations of the genes are used to extract their sequence and the exon/intron information is used to splice them. These isolated and spliced genomic sequences are then BLASTed [10] onto the reference protein sequence, and the matching translated amino acids are stitched together into the final translated protein sequences. In the last step, the translated sequences are organized into 3 alternative databases: The 'Per protein' database, a folder containing one FASTA file for each translated protein. Each protein FASTA file contains the sequences of that protein for all samples The 'Per individual' database, a folder containing one FASTA file for each translated sample / individual. Each sample FASTA file contains all of the translated proteins for that sample. The 'All Protein Reference' database, a single FASTA file containing all translated proteins for all samples. Any of these fasta files can be instantly merged with an ancient protein dataset and be used in the third module.

Module 3 is designed to perform a standard phylogenetic analysis, with some modifications specifically designed for paleoproteomic data. The input of the Module is a FASTA file, containing all of the protein sequences from both the reference dataset and the ancient sample(s) to be analyzed. Accompanying this FASTA file should be a TXT file that contains the name of the dataset-FASTA file as well as the names of all of the ancient samples included in that dataset. The dataset will automatically be split into protein specific sub-datasets, each of which will be aligned and checked for 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 [27]. Isobaric amino acids that cannot be distinguished from each other by the Mass Spectrometer are corrected to ensure the downstream phylogenetic analysis can proceed without problems. Specifically, any time an Isoleucine (L) 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 modern samples, both modern and ancient samples are switched to an L. The user also has the option to provide an additional TXT file named 'MASKED'. Using this optional file, the user can 'mask' a modern sample with the missingness of 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 [22]. All protein sub-dataset alignments are then also merged together into a concatenated dataset. The concatenated dataset is used to generate a maximum-likelihood species tree [21] through PhyML and a Bayesian species tree [50, 37] through MrBayes [25]. The tree generation is parallelized using Mpirun [38].

A more in-depth explanation of each step of each module, as well as the code being run in the background, is provided on the Github page.

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, generating the most complete and up to date, molecular hominid tree. The process of generating the reference dataset and its phylogenetic 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 XXXXX (REF)

Protein Reference Dataset

As further proof of principle and in order to facilitate the future analysis of ancient protein data, we applied the first and second modules of the pipeline and generated a palaeoproteomic - hominid reference dataset. Using the first two modules, we translated publicly available whole genomes from all 4 extant Hominid genera [9, 47, 41]. Details on the preparation of the samples to be translated can be found in the supplementary materials. We also translated multiple ancient genomes from VCF files, including those of several Neanderthals and one Denisovan [49, 36]. Since the dataset is focused on palaeoproteomics, we chose to translate proteins that have previously been reported as present in either teeth or bone tissue. We compiled a list of XXXX(REF) proteins from previous works [13, 3, 2, 53, 26, 43] and successfully transalted XXXX of them. For each protein, both the canonical and all alternative protein coding isoforms were translated, leading to a total of XXXX (REF) number of protein sequences. Details on the creation of the protein list can be found in the supplementary materials. The palaeoproteomic hominid reference dataset is publicly available at —— (REF)

Availability and Community Guidelines

PaleoProPhyler is publicly available on github: The software is released under the XXX(REF) license, and requires the prior installation of Conda. The github

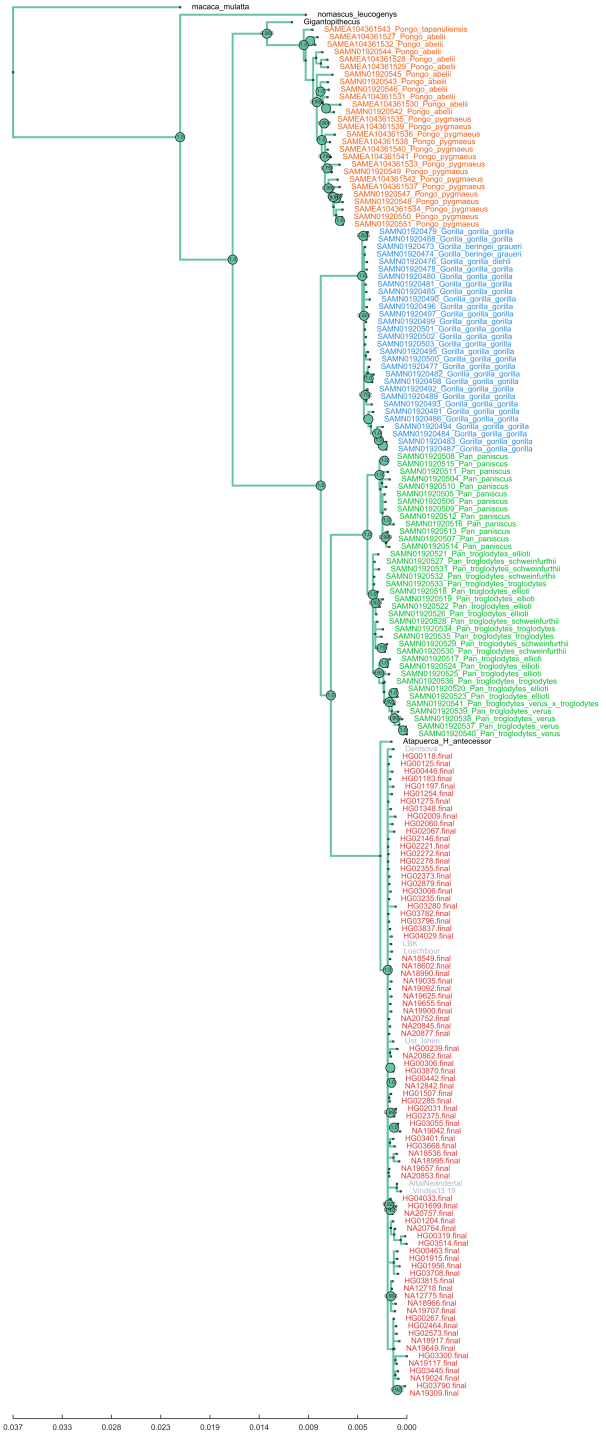

Figure 2: Plotted Phylogenetic Tree

repository contains a tutorial for utilizing the workflow presented here, using the proteins recovered from the Homo antecessor and Gigantopithecus blacki as examples. Users and contributors are welcome to contribute, request features, and report bugs via Github.

Author Contributions

Ioannis Patamanis: Manuscript writing, code for the Snakemake scripts, compiling 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 : Manuscript writing, review and editing

Acknowledgements

We thank Ryan Sinclair Paterson, Graham Gower, Alberto Taurozzi who provided help and feedback during the writing process.

Funding

The project received funding from the European Union's EU Framework Programme for Research and Innovation Horizon 2020, under Grant Agreement No. 861389 - PUSHH. FR was supported by a Villum Young Investigator Grant (project no. 00025300). E.C. has received funding from 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). (REF)

Bibliography

[1] Anaconda software distribution, 2020.

[2] Yahya Acil, Ali E Mobasseri, Patrick H Warnke, Hendrik Terheyden, Jörg Wiltfang, and Ingo Springer. Detection of mature collagen in human dental enamel. *Calcified tissue international*, 76(2):121–126, 2005.

[3] Rodrigo DAM Alves, Jeroen AA Demmers, Karel Bezstarosti, Bram CJ van der Eerden, Jan AN Verhaar, Marco Eijken, and Johannes PTM van Leeuwen. Unraveling the human bone microenvironment beyond the classical extracellular matrix proteins: a human bone protein library. *Journal of proteome research*, 10(10):4725–4733, 2011.

[4] Hamid Bagheri, Andrew J Severin, and Hridesh Rajan. Detecting and correcting misclassified sequences in the large-scale public databases. *Bioinformatics*, 36(18):4699–4705, 2020.

[5] James K Bonfield. Cram 3.1: advances in the cram file format. *Bioinformatics*, 38(6):1497–1503, 2022.

[6] Luise Ørsted Brandt, Alberto J Taurozzi, Meaghan Mackie, Mikkel-Holger S Sinding, Filipe Garrett Vieira, Anne Lisbeth Schmidt, Charlotte Rimstad, Matthew J Collins, and Ulla Mannering. Palaeoproteomics identifies beaver fur in danish high-status viking age burials-direct evidence of fur trade. *Plos one*, 17(7):e0270040, 2022.

[7] Michael Buckley. Ancient collagen reveals evolutionary history of the endemic south american 'ungulates'. *Proceedings of the Royal Society B: Biological Sciences*, 282(1806):20142671, 2015.

[8] 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. *Journal of proteomics*, 194:218–225, 2019.

[9] M Byrska-Bishop, US Evani, X Zhao, AO Basile, HJ Abel, AA Regier, A Corvelo, WE Clarke, R Musunuri, K Nagulapalli, et al. High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios. biorxiv. 2021. *Publisher Full Text*.

[10] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1):1–9, 2009.

[11] Enrico Cappellini, Matthew J Collins, and M Thomas P Gilbert. Unlocking ancient protein palimpsests. *Science*, 343(6177):1320–1322, 2014.

[12] Enrico Cappellini, Frido Welker, Luca Pandolfi, Jazmín Ramos-Madrigal, Diana Samodova, Patrick L Rüther, Anna K Fotakis, David Lyon, J Víctor Moreno-Mayar, Maia Bukhsianidze, et al. Early pleistocene enamel proteome from dmanisi resolves stephanorhinus phylogeny. *Nature*, 574(7776):103–107, 2019.

[13] Gina A Castiblanco, Dorothea Rutishauser, Leopold L Ilag, Stefania Martignon, Jaime E Castellanos, and Wilson Mejía. Identification of proteins from human permanent erupted enamel. *European journal of oral sciences*, 123(6):390–395, 2015.

[14] Fahu Chen, Frido Welker, Chuan-Chou Shen, Shara E Bailey, Inga Bergmann, Simon Davis, Huan Xia, Hui Wang, Roman Fischer, Sarah E Freidline, et al. A late middle pleistocene denisovan mandible from the tibetan plateau. *nature*, 569(7756):409–412, 2019.

[15] Hao Chi, Chao Liu, Hao Yang, Wen-Feng Zeng, Long Wu, Wen-Jing Zhou, Xiu-Nan Niu, Yue-He Ding, Yao Zhang, Rui-Min Wang, et al. Open-pfind enables precise, comprehensive and rapid peptide

identification in shotgun proteomics. *BioRxiv*, page 285395, 2018.

[16] Jürgen Cox and Matthias Mann. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367–1372, 2008.

[17] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[18] Beatrice Demarchi, Shaun Hall, Teresa Roncal-Herrero, Colin L Freeman, Jos Woolley, Molly K Crisp, Julie Wilson, Anna Fotakis, Roman Fischer, Benedikt M Kessler, et al. Protein sequences bound to mineral surfaces persist into deep time. *elife*, 5, 2016.

[19] Beatrice Demarchi, Josefin Stiller, Alicia Grealy, Meaghan Mackie, Yuan Deng, Tom Gilbert, Julia Clarke, Lucas J Legendre, Rosa Boano, Thomas Sicheritz-Pontén, et al. Ancient proteins resolve controversy over the identity of genyornis eggshell. *Proceedings of the National Academy of Sciences*, page e2109326119, 2022.

[20] Vadim Demichev, Christoph B Messner, Spyros I Vernardis, Kathryn S Lilley, and Markus Ralser. Dia-nn: Neural networks and interference correction enable deep coverage in high-throughput proteomics. *bioRxiv*, page 282699, 2018.

[21] Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

[22] Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Systematic biology*, 59(3):307–321, 2010.

[23] Tracy A Heath, Shannon M Hedtke, and David M Hillis. Taxon sampling and the accuracy of phylogenetic analyses. *Journal of systematics and evolution*, 46(3):239–257, 2008.

[24] Jessica Hendy. Ancient protein analysis in archaeology. *Science Advances*, 7(3):eabb9314, 2021.

[25] John P Huelsenbeck and Fredrik Ronquist. Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

[26] Michal Jágr, Adam Eckhardt, Statis Pataridis, and Ivan Mikšík. Comprehensive proteomic analysis of human dentin. *European journal of oral sciences*, 120(4):259–268, 2012.

[27] Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[28] Klaus-Peter Koepfli, Benedict Paten, Genome 10K Community of Scientists, and Stephen J O'Brien. The genome 10k project: a way forward. *Annu. Rev. Anim. Biosci.*, 3(1):57–111, 2015.

[29] Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, and Alexey I Nesvizhskii. Msfragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature methods*, 14(5):513–520, 2017.

[30] Tianying Lan and Charlotte Lindqvist. Technical advances and challenges in genome-scale analysis of ancient dna. *Paleogenomics*, pages 3–29, 2018.

[31] Liam T Lanigan, Meaghan Mackie, Susanne Feine, Jean-Jacques Hublin, Ralf W Schmitz, Arndt Wilcke, Matthew J Collins, Enrico Cappellini, Jesper V Olsen, Alberto J Taurozzi, et al. Multi-protease analysis of pleistocene bone proteomes. *Journal of proteomics*, 228:103889, 2020.

[32] Harris A Lewin, Gene E Robinson, W John Kress, William J Baker, Jonathan Coddington, Keith A Crandall, Richard Durbin, Scott V Edwards, Félix Forest, M Thomas P Gilbert, et al. Earth biogenome project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17):4325–4333, 2018.

[33] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[34] David J Lipman and William R Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, 1985.

[35] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, 17(20):2337–2342, 2003.

[36] Fabrizio Mafessoni, Steffi Grote, Cesare de Filippo, Viviane Slon, Kseniya A Kolobova, Bence Viola, Sergey V Markin, Manjusha Chintalapati, Stephane Peyrégne, Laurits Skov, et al. A high-coverage neandertal genome from chagyrskaya cave. *Proceedings of the National Academy of Sciences*, 117(26):15132–15136, 2020.

[37] Bob Mau and Michael A Newton. Phylogenetic inference for binary data on dendograms using markov chain monte carlo. *Journal of Computational and Graphical Statistics*, 6(1):122–131, 1997.

[38] Message Passing Interface Forum. *MPI: A Message-Passing Interface Standard Version 4.0*, June 2021.

[39] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, et al. Sustainable data analysis with snakemake. *F1000Research*, 10, 2021.

[40] Johannes B Müller, Philipp E Geyer, Ana R Colaço, Peter V Treit, Maximilian T Strauss, Mario Oroshi, Sophia Doll, Sebastian Virreira Winter, Jakob M Bader, Niklas Köhler, et al. The proteome landscape of the kingdoms of life. *Nature*, 582(7813):592–596, 2020.

[41] Alexander Nater, Maja P Mattle-Greminger, Anton Nurcahyo, Matthew G Nowak, Marc De Manuel, Tariq Desai, Colin Groves, Marc Pybus, Tugce Bilgin Sonay, Christian Roos, et al. Morphometric, behavioral, and genomic evidence for a new orangutan species. *Current Biology*, 27(22):3487–3498, 2017.

[42] Fred R Opperdoes. Phylogenetic analysis using protein sequences. *The phylogenetics handbook a practical approach to DNA and protein phylogeny*, pages 207–235, 2003.

[43] Eun-Sung Park, Hye-Sim Cho, Tae-Geon Kwon, Sin-Nam Jang, Sang-Han Lee, Chang-Hyeon An, Hong-In Shin, Jae-Young Kim, and Je-Yoel Cho. Proteomics analysis of human dentin reveals distinct protein expression profiles. *Journal of proteome research*, 8(3):1338–1346, 2009.

[44] Alexander Peltzer, Günter Jäger, Alexander Herbig, Alexander Seitz, Christian Kniep, Johannes Krause, and Kay Nieselt. Eager: efficient ancient genome reconstruction. *Genome biology*, 17(1):1–14, 2016.

[45] David N Perkins, Darryl JC Pappin, David M Creasy, and John S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS: An International Journal*, 20(18):3551–3567, 1999.

[46] Isabel Maria Porto, Helen Julie Laure, Frederico Barbosa de Sousa, Jose Cesar Rosa, and Raquel Fernanda Gerlach. New techniques for the recovery of small amounts of mature enamel proteins. *Journal of Archaeological Science*, 38(12):3596–3604, 2011.

[47] Javier Prado-Martinez, Peter H Sudmant, Jeffrey M Kidd, Heng Li, Joanna L Kelley, Belen Lorente-Galdos, Krishna R Veeramah, August E Woerner, Timothy D O'connor, Gabriel Santpere, et al. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, 2013.

[48] Samantha Presslee, Graham J Slater, Francois Pujos, Analia M Forasiepi, Roman Fischer, Kelly Molloy, Meaghan Mackie, Jesper V Olsen, Alejandro Kramarz, Matias Taglioretti, et al. Data from: Palaeoproteomics resolves sloth phylogeny. 2019.

[49] Kay Prüfer, Cesare De Filippo, Steffi Grote, Fabrizio Mafessoni, Petra Korlević, Mateja Hajdinjak, Benjamin Vernot, Laurits Skov, Pinghsun Hsieh, Stéphane Peyrégne, et al. A high-coverage neandertal genome from vindija cave in croatia. *Science*, 358(6363):655–658, 2017.

[50] Bruce Rannala and Ziheng Yang. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of molecular evolution*, 43(3):304–311, 1996.

[51] Michael S Rosenberg and Sudhir Kumar. Taxon sampling, bioinformatics, and phylogenomics. *Systematic Biology*, 52(1):119, 2003.

[52] Patrick Leopold Rüther, Immanuel Mirnes Husic, Pernille Bangsgaard, Kristian Murphy Gregersen, Pernille Pantmann, Milena Carvalho, Ricardo Miguel Godinho, Lukas Friedl, João Cascalheira, Alberto John Taurozzi, et al. Spin enables high throughput species identification of archaeological bone by proteomics. *Nature communications*, 13(1):1–14, 2022.

[53] Cristiane R Salmon, Ana Paula O Giorgetti, Adriana Franco Paes Leme, Romênia R Domingues, Enilson Antonio Sallum, Marcelo C Alves, Tamara N Kolli, Brian L Foster, and Francisco H Nociti Jr. Global proteome profiling of dental cementum under experimentally-induced apposition. *Journal of proteomics*, 141:12–23, 2016.

[54] Mikkel Schubert, Luca Ermini, Clio Der Sarkissian, Hákon Jónsson, Aurélien Ginolhac, Robert Schaefer, Michael D Martin, Ruth Fernández, Martin Kircher, Molly McCue, et al. Characterization of ancient and modern genomes by snp detection and phylogenomic and metagenomic analysis using paleomix. *Nature protocols*, 9(5):1056–1082, 2014.

[55] Stefan K Solntsev, Michael R Shortreed, Brian L Frey, and Lloyd M Smith. Enhanced global post-translational modification discovery with metamorpheus. *Journal of proteome research*, 17(5):1844–1851, 2018.

[56] Christina Warinner, Kristine Korzow Richter, and Matthew J Collins. Paleoproteomics. *Chemical Reviews*, 2022.

[57] Frido Welker, Jazmín Ramos-Madrigal, Petra Gutenbrunner, Meaghan Mackie, Shivani Tiwary, Rosa Rakownikow Jersie-Christensen, Cristina Chiva, Marc R Dickinson, Martin Kuhlwilm, Marc de Manuel, et al. The dental proteome of homo antecessor. *Nature*, 580(7802):235–238, 2020.

[58] Frido Welker, Jazmín Ramos-Madrigal, Martin Kuhlwilm, Wei Liao, Petra Gutenbrunner, Marc de Manuel, Diana Samodova, Meaghan Mackie, Morten E Allentoft, Anne-Marie Bacon, et al. Enamel proteome shows that gigantopithecus was an early diverging pongine. *Nature*, 576(7786):262–265, 2019.

[59] Frido Welker, Geoff M Smith, Jarod M Hutson, Lutz Kindler, Alejandro Garcia-Moreno, Aritza Villaluenga, Elaine Turner, and Sabine Gaudzinski-Windheuser. Middle pleistocene protein sequences from the rhinoceros genus stephanorhinus and the phylogeny of extant and extinct middle/late pleistocene rhinocerotidae. *PeerJ*, 5:e3033, 2017.

[60] Andrew Yates, Kathryn Beal, Stephen Keenan, William McLaren, Miguel Pignatelli, Graham RS Ritchie, Magali Ruffier, Kieron Taylor, Alessandro Vullo, and Paul Flicek. The ensembl rest api: Ensembl data for any language. *Bioinformatics*, 31(1):143–145, 2015.

[61] Guojie Zhang, Cai Li, Qiye Li, Bo Li, Denis M Larkin, Chul Lee, Jay F Storz, Agostinho Antunes, Matthew J Greenwold, Robert W Meredith, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, 346(6215):1311–1320, 2014.