Check for updates

BRIEF REPORT

Spectra without stories: reporting 94% dark and unidentified ancient proteomes

[version 1; peer review: 2 approved, 1 approved with reservations]

Yun Chiang [1,2], Frido Welker [1], Matthew James Collins [1,3]

[1]Globe Institute, University of Copenhagen, Copenhagen, Denmark
[2]The Nice Institute of Chemistry, Universite Cote d'Azur, Nice, France
[3]McDonald Institute for Archaeological Research, University of Cambridge, Cambridge, England, UK

v1 **First published:** 15 Apr 2024, **4**:71
https://doi.org/10.12688/openreseurope.17225.1
Latest published: 15 Apr 2024, **4**:71
https://doi.org/10.12688/openreseurope.17225.1

Abstract

Background

Data-dependent, bottom-up proteomics is widely used for identifying proteins and peptides. However, one key challenge is that 70% of fragment ion spectra consistently fail to be assigned by conventional database searching. This 'dark matter' of bottom-up proteomics seems to affect fields where non-model organisms, low-abundance proteins, non-tryptic peptides, and complex modifications may be present. While palaeoproteomics may appear as a niche field, understanding and reporting unidentified ancient spectra require collaborative innovation in bioinformatics strategies. This may advance the analysis of complex datasets.

Methods

14.97 million high-impact ancient spectra published in Nature and Science portfolios were mined from public repositories. Identification rates, defined as the proportion of assigned fragment ion spectra, were collected as part of deposited database search outputs or parsed using open-source python packages.

Results and Conclusions

We report that typically 94% of the published ancient spectra remain unidentified. This phenomenon may be caused by multiple factors, notably the limitations of database searching and the selection of user-defined reference data with advanced modification patterns. These 'spectra without stories' highlight the need for widespread data

Open Peer Review

Approval Status ✓ ? ✓

	1	2	3
version 1 15 Apr 2024	✓ view	? view	✓ view

1. **Timothy Cleland**, Smithsonian Museum Conservation Institute, Suitland-Silver Hill, USA
2. **Glendon Parker**, University of California, California, USA
3. **Virginia Harvey**, University of Chester, Chester, UK

Any reports and responses or comments on the article can be found at the end of the article.

sharing to facilitate methodological development and minimise the loss of often irreplaceable ancient materials. Testing and validating alternative search strategies, such as open searching and de novo sequencing, may also improve overall identification rates. Hence, lessons learnt in palaeoproteomics may benefit other fields grappling with challenging data.

Keywords
palaeoproteomics, DDA, shotgun proteomics, database searching, bioinformatics challenges

This article is included in the Marie-Sklodowska-Curie Actions (MSCA) gateway.

This article is included in the European Research Council (ERC) gateway.

This article is included in the Horizon 2020 gateway.

Corresponding author: Yun Chiang (yun@palaeome.org)

Author roles: Chiang Y: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Welker F**: Conceptualization, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Collins MJ**: Conceptualization, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement Nos. 787282 and 948365). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 956351. FW is supported by Villum Fonden (no. 40747). MJC acknowledges support from Danmarks Grundforskningsfond (DNRF128) and Carlsbergfondet (CF18-1110).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2024 Chiang Y *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License , which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Chiang Y, Welker F and Collins MJ. **Spectra without stories: reporting 94% dark and unidentified ancient proteomes [version 1; peer review: 2 approved, 1 approved with reservations]** Open Research Europe 2024, **4**:71 https://doi.org/10.12688/openreseurope.17225.1

First published: 15 Apr 2024, **4**:71 https://doi.org/10.12688/openreseurope.17225.1

Introduction

Tandem mass spectrometry (LC-MS/MS) has been widely used for untargeted or shotgun proteomics. A standard method for automated fragment ion spectra (MS2) generation is data-dependent acquisition (DDA). In a typical DDA experiment, precursor ions are selected for fragmentation using their intensities observed in full MS1 scans and user-defined parameters (the top *N* method)[1]. Three common fragmentation techniques include electron-transfer dissociation (ETD), collision-induced dissociation (CID), and higher-energy collisional dissociation (HCD); the latter two methods typically generate b- and y- ions due to the breakages of amide bonds during collision[2]. The resulting MS2 data are clean and ready for peptide-spectrum matching and database searching, as fragments are generally derived from a single precursor within a small isolation window below 2 mass-to-charge (*m/z*) ratios[3]. Overall, data-driven shotgun proteomics has been robust and provides high–resolution data.

The combination of DDA and database searching has been a powerful tool for the deep and global discoveries of proteomes. However, DDA is susceptible to the co-fragmentation of co-eluting, near-isobaric precursor ions, which may generate complex chimeric spectra[4]. Moreover, due to the stochastic nature of precursor sampling in DDA, peptide-level repeatability in technical replicates is reported to be below 60%, with low-abundance precursors often being under-sampled[5]. While features like dynamic exclusion (a mass spectrometer parameter that temporarily excludes most-abundant precursors), and match-between-runs (MBR, a bioinformatics algorithm that realigns unidentified peaks using retention time and other MS1 features) have been designed to mitigate the missing value problem[6,7], it has been estimated that only 30% of DDA MS2 spectra may be identified using database searching[8]. This value is corroborated by the experimental observation of two large-scale analyses that just over 25% of MS2 spectra submitted to PRIDE and MassIVE are identified[9,10].

Additionally, the efficiency of database searching may exacerbate the problem of unassigned spectra, leading to even lower rates of successful identification. This is primarily due to the limitations of the search space of a peptide matching algorithm, which is determined by a user-provided protein sequence database, digestion patterns, missed cleavages, a range of allowed peptide length, a predefined set of post-translational modifications (PTMs), and the mass tolerances of both precursors and fragment ions[11]. Since acquired MS2 data are searched against theoretical spectra generated *in silico*, sequences not covered by a database, miscleaved peptides, and large mass shifts due to unexpected PTMs will be overlooked during conventional database searching. The search space also affects the filtering of peptide-spectrum matches (PSMs), as target-decoy competition is widely used to calculate false discovery rates (FDRs), with decoys generated based on the same parameters as the target database[12].

These limitations are pronounced in fields where managing the optimal search space is challenging due to the lack of *a priori* knowledge, non-model organisms, non-canonical variants, and the heterogeneity of PTMs. For example, in metaproteomics, it is difficult to anticipate which species may be present in samples, and not all microbial sequences are fully documented[13]. The search space is also substantial in immunopeptidomics due to variants and the non-tryptic nature of human leukocyte antigen (HLA) molecules, particularly class I peptides[14,15]. As a result, the average identification rates of HLA data are around 10%. Similarly, it has been highlighted that less than 5 % of the phosphoproteome is identified, since phosphorylation sites are complex and vary in abundance[16].

These data analysis challenges are often summarised as the "dark matter of shotgun proteomics"[9,10,17]. While similar issues are anticipated in palaeoproteomics, given that ancient proteins may be fragmented, non-tryptic, heavily modified, low in abundance, and high in non-model organisms that are not covered by standard reference databases, it remains unclear how many ancient spectra are dark and unassigned. This research question is significant for three reasons. Firstly, the study of ancient proteins has provided invaluable information for constructing narratives about diets, culinary practices, social interactions, and evolutionary histories in the *longue durée*. Yet, these unassigned ancient mass spectrometry data are, figuratively, spectra without stories. They are untapped reservoirs of amino acid substitutions, novel peptides, and complex PTMs that almost certainly conceal insights into the human past and protein preservation pathways.

Secondly, archaeological materials are unique and often irreplaceable artefacts. Since ancient proteins need to be extracted, destructive sampling is inevitable without the development and optimisation of non-destructive extraction protocols. It is crucial to evaluate the efficiency of palaeoproteomics analysis considering the associated trade-offs between potential scientific gains and irreversible loss of valuable ancient materials. Lastly, the challenges of working with complex, damaged, and low-abundance ancient proteomes may require innovation in data acquisition, database management, scoring algorithms, and search strategies. These advancements may increase current analytical capabilities and may be applied to other fields dealing with similarly challenging datasets that require a large search space.

Methods

To address the question, we mined 15 high-impact datasets recently reviewed[18] and published in *Science* and *Nature Portfolio* journals[19–33], resulting in the collection of 14.97 million deposited, publicly available ancient MS2 spectra. These datasets represent the best practice of the field and the challenges of analysing a diverse range of unique ancient materials, including bones, teeth, dental calculus, paintings, and ceramic pots. Additional criteria were that only complete datasets that cover original peak lists (.mgf files) and search results (summary.txt or .mzid files) were included.

Specifically, Orbitrap™ hybrid mass analysers (Q Exactive™ or Exploris™ instruments), DDA, and HCD were employed

by all 15 datasets. 73.33% (11/15) of the search results were obtained from MaxQuant[34] while the remaining datasets (4/15) were analysed by Mascot[35]. Overall identification rates, defined as the ratios of total assigned MS2 spectra to submitted MS2 spectra, were collected as part of submitted MaxQuant outputs. For Mascot results, open-source Pyteomics[36] was used to parse deposited .mzid outputs and determine the ratios of identified spectra. To contrast these identification rates of ancient proteome datasets with the general trend of identified MS2 spectra in public repositories, we obtained this metric from two recent meta-clustering analyses; these are the average identification rate of the PRIDE MS2 data (n= 256 million)[9] and that of the MassIVE depository (n= 669 million)[10]. All figures were produced by non- proprietary python packages (Matplotlib[37]/Seaborn[38]) and annotated using Affinity Designer 2[39]. A free alternative for annotation is Inkscape.

Results and discussion

We report that only 0.88 out of 14.97 million total submitted MS2 spectra are identified, indicating that 94.12% of the queries remain uncharacterised. The identification rates of palaeoproteomics datasets range from 0.47% to 12.61% (Figure 1). All of them consistently fall below the average identification rate of MS2 datasets deposited in PRIDE (25.78%) and MassIVE (26.28%). It should be noted that these identified MS2 spectra in the ancient datasets include peptides from both putative ancient proteins and contaminants, such as trypsin and human skin keratins. The validation of ancient proteins and how to effectively separate them from modern contaminants are ongoing debates[27]. However, it is assumed that the actual proportion of genuinely assigned ancient spectra is below the average identification rate of 5.88%.

Given that only 6% of the ancient MS2 spectra were successfully assigned, the 94% of dark, unidentified ancient proteomes highlights the need for widespread data sharing to facilitate methodological development and the re-evaluation of bioinformatics challenges inherent in the analyses of archaeological, palaeoanthropological, and palaeontological materials. Sharing data enables the re-analysis of often irreplaceable ancient materials, and fosters collaborative efforts in addressing the analytical challenges of complex archaeological data as seen in other related fields, especially ancient DNA[40,41]. While we acknowledge that the analysis of ancient MS2 data is often context-specific, depending on protein chemistry, preservation pathways, research questions, and the design of search space, we call for ongoing efforts to build comprehensive databases, refine search strategies for PTMs and ensure widespread data dissemination, to optimise search strategies and increase overall identification rates (summarised in Figure 2).

How to build a relevant database?

Indeed, creating a comprehensive database for ancient proteins is significantly challenging without *a priori* information. Generally, ancient proteins may be derived from proteinaceous tissues, microbial communities, and past human activities, including cooking, crafting, parchment making, and leather tanning. An ideal database for searching includes all potential candidates without additional protein sequences. However, given the diversity and complexity of ancient proteins, often an entire UniProt knowledge database (Swiss-Prot and TrEMBL), up to about 251 million proteins and/or curated databases, typically consisting of hundreds of protein targets, are used in the database searching of ancient MS2 data[18].

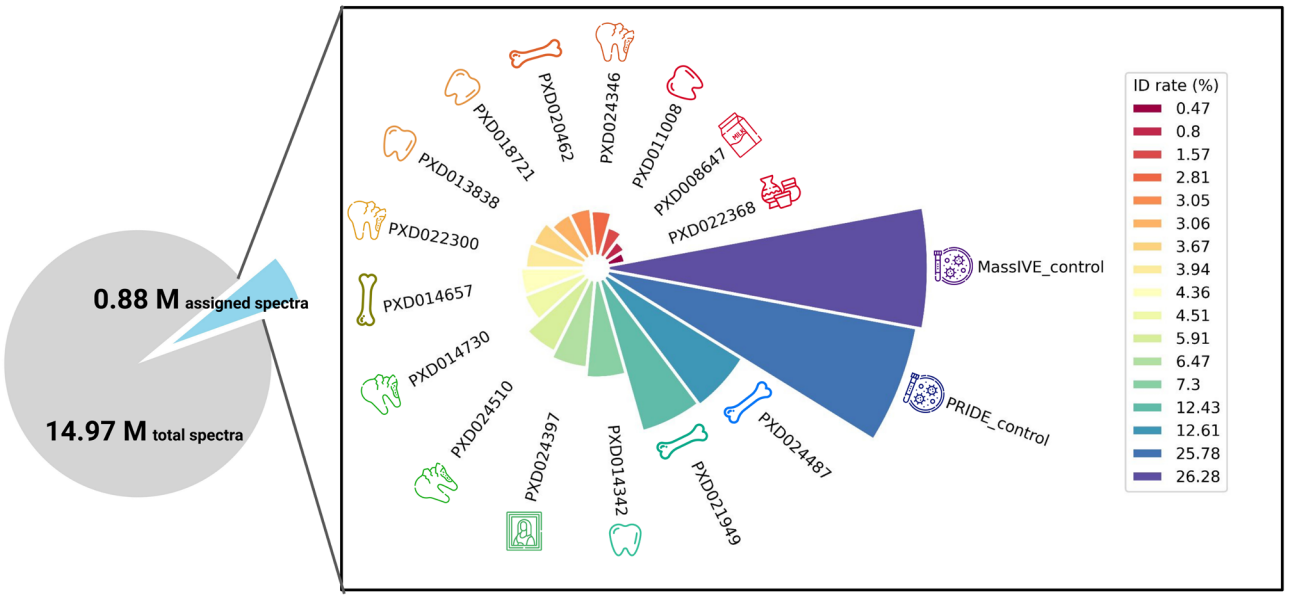


Figure 1. A pie chart of the total collected (14.97 million) and assigned MS2 spectra (0.88 million). Across all the 15 palaeoproteomics datasets containing skeletal, painting, and pottery samples, the overall identification rates (ID rate %) are low in comparison to two large-scale proteomics analyses (over 25%).

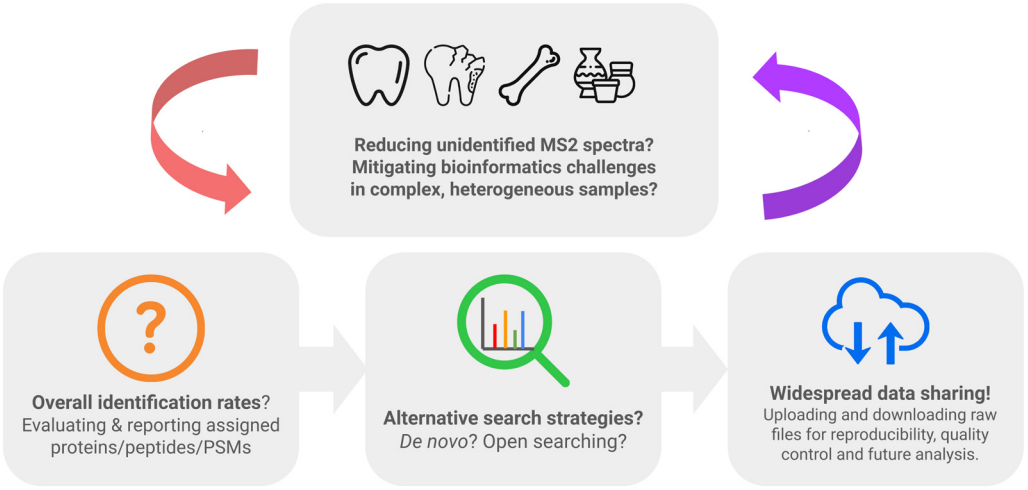


Figure 2. A flowchart that summarises key bottlenecks and suggested best practices to deal with the proteomics analysis of complex and heterogeneous samples.

Using an extensive database like the entire UniProt generates a substantial pool of candidates and increases the search space exponentially. The issues with a large search space are well documented in metaproteomics communities where NCBI RefSeq (over 113 million non-redundant proteins) have been used for the annotation of complicated protein systems[42]. It has been reported that the utilisation of large databases (over 1 million sequences) results in decreased PSMs, increased risks of false positives, and difficulties in using FDRs to control for confident identification[43]. In contrast, a focused and targeted approach may inadvertently raise false positive rates and may not capture the full spectrum of ancient proteins present in samples. Hence, a paradox is prevalent in palaeoproteomics: the need to balance the comprehensiveness of a database to maximise identification against the risk of increasing false positives and computational complexity.

A multi-step approach to database management may be beneficial; in metaproteomics, this generally includes an initial survey of a broad database to identify targets, followed by a subsequent run using a narrowed-down database[44]. Machine learning and *de novo* sequencing without databases are other alternatives. Since novel and extinct sequences may be present in ancient samples, a *de novo*-centric approach may discover amino acid substitutions relevant to evolutionary studies and improve overall proteome composition[45,46]. Validation and optimisation may be required for these alternative approaches, and a substantial collection of quality ancient spectra is needed to train relevant machine learning models. Despite these hurdles, such alternative strategies may mitigate the limitations of database management.

How to manage protein degradation & PTMs?

Apart from selecting a suitable database, the preservation states of ancient proteins may enlarge search space due to the increasing number of allowed missed cleavages, semi-tryptic/non-specific searching, and a wide range of associated PTMs. While in living organisms, PTMs mediate protein functions, structures, and interactions[47], in the context of palaeoproteomics, PTMs provide insights on the incorporation of proteins into archaeological materials, protein preservation pathways, and protein validation[18]. Indeed, ancient proteins are complex mixtures of multiple protein groups, minerals, and other organic molecules like carbohydrates and lipids. These complex mixtures are also subject to a wide range of environmental factors over time, including the pH of a burial environment, groundwater, soil chemistry, and temperature. The resulting complex and heterogeneous PTMs may complicate the analysis of MS2 data. For example, Maillard reactions and advanced glycation end products may alter cleavage efficiency during enzymatic digestion[48]. Glycation and lipidation are also labile PTMs, and they could be broken down during MS2 fragmentation, further affecting the assignment of PTM sites and PSMs[49].

Given these analytical challenges, one strategy is to leverage open searching provided by software such as Fragpipe[50]. An open search is characterised by a wide precursor tolerance window, generally extending up to +500 Daltons, to account for mass shifts between experimental data and theoretical values. MSFragger manages the enlarged search space by using a fragment ion index, mass binning, and precursor mass ordering to optimise the retrieval and scoring during peptide-spectrum matching; then open searching outputs are annotated by PTM-Shepherd that utilises signal-to-noise ratio filtering, retention time shifts, MS2 spectra resemblance and the entirety of the Unimod modification database[51]. Fragpipe also offers a workflow optimised for labile PTMs[49]. While FDR controls are still challenging for open searching[52], it could be advantageous for the global discoveries of PTMs and damage patterns.

Is palaeoproteomics data FAIR?

Lastly, like any other scientific datasets, palaeoproteomics data is progressing towards the FAIR principles[53]. Archaeological data is generally findable, accessible, and reusable thanks to the establishment and adherence to data-sharing practices. Acquired raw files from mass spectrometers, processed peak lists, and search results are often readily available to download from public repositories such as PRIDE and MassIVE. This fosters transparency and collaboration within the research community. We would urge the community to continue prioritising data dissemination, as mass spectrometry pipelines are rapidly advancing. It is crucial to ensure reproducibility and create possibilities for future research to build upon current findings, considering archaeological materials are finite resources and often cannot be resampled.

However, achieving true interoperability can be challenging due to the inherent complexity and diversity of palaeoproteomics experiments. We would highlight the usefulness of routinely reporting bulk identification rates for complex, heterogeneous materials. This simple quality control matrix is integral for the comprehensive evaluation of search engines and the performance of identification pipelines. In addition, it is advantageous to report the number of assigned PSMs under a predefined FDR, total identified peptides, and the total assignments of protein groups. These figures could be included as supplementary information and uploaded to public repositories. They may also be useful in evaluating the effectiveness of database searching strategies, and determining whether it is worth exploring open and *de novo* sequencing methods.

Overall, we demonstrate that low identification rates are consistently observed in the 15 million ancient spectra from the 15 high-profile palaeoproteomics studies. These spectra without stories potentially result from search space management, the absence of appropriate databases, and suboptimal search parameters due to the degraded and modified nature of ancient proteins. While the findability, accessibility, and reusability of archaeological data appear to align with the FAIR principles, simple and standardised quality measures are still needed to navigate the complexity of palaeoproteomics data and enhance interoperability. While open searching and *de novo* sequencing are promising approaches to tackle these analytical challenges, they will require benchmarking, validation, and optimisation. Overall, these bioinformatics advances offer a pathway towards uncovering the hidden stories within the 94% of unidentified ancient spectra, and lessons learnt in palaeoproteomics may be valuable in other fields facing similar challenges with complex datasets.

Ethics and consent

Ethical approval and consent were not required.

Data availability

All ancient datasets used in the paper are publicly available as follows:

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD008647[19].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD011008[20].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD013838[22].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD014342[23].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD014657[21].

PRIDE: palaeoproteomics data. Accession number PXD014730[24].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD018721[28].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD020462[30].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD021949[25].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD022300[26].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD022368[27].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD024346[32].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD024397[29].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD024487[31].

ProteomeXchange Consortium: palaeoproteomics data. Accession number PXD024510[33].

Acknowledgments

We would like to extend our gratitude to our colleagues for their regular discussions on bioinformatics challenges and how to improve the analysis of palaeoproteomics data. These discussions were helpful during the early stages of the paper design. We especially thank Ismael Rodriguez Palomo (University of Cambridge), Bharath Anila Bhuvanendran Nair (University of Cambridge), Joannes Adrianus Antonius Dekker (University of Copenhagen/ University of York), Max Emil Ermter Ramsøe (University of Copenhagen), and Benjamin Dartigues (University of Bordeaux).

References

1. Bateman NW, Goulding SP, Shulman NJ, *et al.*: **Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (DDA).** *Mol Cell Proteomics.* 2014; **13**(1): 329–338.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Lau KW, Hart SR, Lynch JA, *et al.*: **Observations on the detection of b- and y-type ions in the collisionally activated decomposition spectra of protonated peptides.** *Rapid Commun Mass Spectrom.* 2009; **23**(10): 1508–1514.
PubMed Abstract | Publisher Full Text

3. Davies V, Wandy J, Weidt S, *et al.*: **Rapid Development of Improved Data-Dependent Acquisition Strategies.** *Anal Chem.* 2021; **93**(14): 5676–5683.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Houel S, Abernathy R, Renganathan K, *et al.*: **Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies.** *J Proteome Res.* 2010; **9**(8): 4152–4160.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Tabb DL, Vega-Montoto L, Rudnick PA, *et al.*: **Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry.** *J Proteome Res.* 2010; **9**(2): 761–776.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Zhang Y, Wen Z, Washburn MP, *et al.*: **Effect of dynamic exclusion duration on spectral count based quantitative proteomics.** *Anal Chem.* 2009; **81**(15): 6317–6326.
PubMed Abstract | Publisher Full Text

7. Cox J, Hein MY, Luber CA, *et al.*: **Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ.** *Mol Cell Proteomics.* 2014; **13**(9): 2513–2526.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Kreimer S, Belov ME, Danielson WF, *et al.*: **Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-Up Proteomic Profiling.** *J Proteome Res.* 2016; **15**(10): 3563–3573.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Griss J, Perez-Riverol Y, Lewis S, *et al.*: **Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets.** *Nat Methods.* 2016; **13**(8): 651–656.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Bittremieux W, May DH, Bilmes J: **A learned embedding for efficient joint analysis of millions of mass spectra.** *Nat Methods.* 2022; **19**(6): 675–678.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Kumar D, Yadav AK, Dash D: **Choosing an Optimal Database for Protein Identification from Tandem Mass Spectrometry Data.** *Methods Mol Biol.* 2017; **1549**: 17–29.
PubMed Abstract | Publisher Full Text

12. Choi H, Nesvizhskii AI: **False discovery rates and related statistical concepts in mass spectrometry-based proteomics.** *J Proteome Res.* 2008; **7**(1): 47–50.
PubMed Abstract | Publisher Full Text

13. Duan H, Cheng K, Ning Z, *et al.*: **Assessing the Dark Field of Metaproteome.** *Anal Chem.* 2022; **94**(45): 15648–15654.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Caron E, Kowalewski DJ, Koh CC, *et al.*: **Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry.** *Mol Cell Proteomics.* 2015; **14**(12): 3105–3117.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Vizcaíno JA, Kubiniok P, Kovalchik KA, *et al.*: **The Human Immunopeptidome Project: A Roadmap to Predict and Treat Immune Diseases.** *Mol Cell Proteomics.* 2020; **19**(1): 31–49.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Needham EJ, Parker BL, Burykin T, *et al.*: **Illuminating the dark phosphoproteome.** *Sci Signal.* 2019; **12**(565): eaau8645.
PubMed Abstract | Publisher Full Text

17. Chick JM, Kolippakkam D, Nusinow DP, *et al.*: **A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides.** *Nat Biotechnol.* 2015; **33**(7): 743–749.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Warinner C, Korzow Richter K, Collins MJ: **Paleoproteomics.** *Chem Rev.* 2022; **122**(16): 13401–13446.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Hendy J, Colonese AC, Franz I, *et al.*: **Ancient proteins from ceramic vessels at Çatalhöyük West reveal the hidden cuisine of early farmers.** *Nat Commun.* 2018; **9**(1): 4064.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Cappellini E, Welker F, Pandolfi L, *et al.*: **Early Pleistocene enamel proteome from Dmanisi resolves Stephanorhinus phylogeny.** *Nature.* 2019; **574**(7776): 103–107.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Tsutaya T, Mackie M, Koenig C, *et al.*: **Palaeoproteomic identification of breast milk protein residues from the archaeological skeletal remains of a neonatal dog.** *Sci Rep.* 2019; **9**(1): 12841.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Welker F, Ramos-Madrigal J, Kuhlwilm M, *et al.*: **Enamel proteome shows that *Gigantopithecus* was an early diverging pongine.** *Nature.* 2019; **576**(7786): 262–265.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Welker F, Ramos-Madrigal J, Gutenbrunner P, *et al.*: **The dental proteome of *Homo antecessor*.** *Nature.* 2020; **580**(7802): 235–238.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Wilkin S, Ventresca Miller A, Taylor WTT, *et al.*: **Dairy pastoralism sustained eastern Eurasian steppe populations for 5,000 years.** *Nat Ecol Evol.* 2020; **4**(3): 346–355.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Coutu AN, Taurozzi AJ, Mackie M, *et al.*: **Palaeoproteomics confirm earliest domesticated sheep in southern Africa ca. 2000 BP.** *Sci Rep.* 2021; **11**(1): 6631.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Wilkin S, Ventresca Miller A, Fernandes R, *et al.*: **Dairying enabled Early Bronze Age Yamnaya steppe expansions.** *Nature.* 2021; **598**(7882): 629–633.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Ramsøe A, Crispin M, Mackie M, *et al.*: **Assessing the degradation of ancient milk proteins through site-specific deamidation patterns.** *Sci Rep.* 2021; **11**(1): 7795.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Demeter F, Zanolli C, Westaway KE, *et al.*: **A Middle Pleistocene Denisovan molar from the Annamite Chain of northern Laos.** *Nat Commun.* 2022; **13**(1): 2557.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Di Gianvincenzo F, Peggie D, Mackie M, *et al.*: **Palaeoproteomics guidelines to identify proteinaceous binders in artworks following the study of a 15th-century painting by Sandro Botticelli's workshop.** *Sci Rep.* 2022; **12**(1): 10638.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Ntasi G, Palomo IR, Marino G, *et al.*: **Molecular signatures written in bone proteins of 79 AD victims from Herculaneum and Pompeii.** *Sci Rep.* 2022; **12**(1): 8401.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Rüther PL, Husic IM, Bangsgaard P, *et al.*: **SPIN enables high throughput species identification of archaeological bone by proteomics.** *Nat Commun.* 2022; **13**(1): 2458.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Scorrano G, Nielsen SH, Vetro DL, *et al.*: **Genomic ancestry, diet and microbiomes of Upper Palaeolithic hunter-gatherers from San Teodoro cave.** *Communications Biology.* 2022; **5**(1): 1262.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Ventresca Miller AR, Wilkin S, Bayarsaikhan J, *et al.*: **Permafrost preservation reveals proteomic evidence for yak milk consumption in the 13th century.** *Commun Biol.* 2023; **6**(1): 351.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Cox J, Mann M: **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.** *Nat Biotechnol.* 2008; **26**(12): 1367–1372.
PubMed Abstract | Publisher Full Text

35. Perkins DN, Pappin DJ, Creasy DM, *et al.*: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis.* 1999; **20**(18): 3551–3567.
PubMed Abstract | Publisher Full Text

36. Levitsky LI, Klein JA, Ivanov MV, *et al.*: **Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework.** *J Proteome Res.* 2019; **18**(2): 709–714.
PubMed Abstract | Publisher Full Text

37. Hunter JD: **Matplotlib: A 2D Graphics Environment.** 2007; **9**(3): 90–95.
Publisher Full Text

38. Waskom M: **seaborn: statistical data visualization.** *J Open Source Softw.* 2021; **6**(60): 3021.
Publisher Full Text

39. Serif (Europe) Ltd: **Affinity Designer 2 (Version 2)[Computer software]**. 1987.
Reference Source

40. Anagnostou P, Capocasa M, Milia N, *et al.*: **When data sharing gets close to 100%: what human paleogenetics can teach the open science movement.** *PLoS One.* 2015; **10**(3): e0121409.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Der Sarkissian C, Velsko IM, Fotakis AK, *et al.*: **Ancient Metagenomic Studies: Considerations for the Wider Scientific Community.** *mSystems.* 2021; **6**(6): e0131521.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Park SKR, Jung T, Thuy-Boun PS, *et al.*: **ComPIL 2.0: An Updated Comprehensive Metaproteomics Database.** *J Proteome Res.* 2019; **18**(2): 616–622.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Jagtap P, Goslinga J, Kooren JA, *et al.*: **A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies.** *Proteomics.* 2013; **13**(8): 1352–1357.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Zhang X, Ning Z, Mayne J, *et al.*: **MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota.** *Microbiome.* 2016; **4**(1): 31.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

45. Gutenbrunner P, Kyriakidou P, Welker F, *et al.*: **Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra.** *bioRxiv.* 2021; 2021.09.04.458985.
Publisher Full Text

46. Welker F: **Elucidation of cross-species proteomic effects in human and hominin bone proteome identification through a bioinformatics experiment.** *BMC Evol Biol.* 2018; **18**(1)**: 23.**
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

47. Mann M, Jensen ON: **Proteomic analysis of post-translational modifications.** *Nat Biotechnol.* 2003; **21**(3): 255–261.
PubMed Abstract | **Publisher Full Text**

48. Cho, S, Duong VA, Mok JH, *et al.*: **Enrichment and analysis of glycated proteins.** *Rev Anal Chem.* 2022; **41**(1): 83–97.
Publisher Full Text

49. Polasky DA, Geiszler DJ, Yu F, *et al.*: **MSFragger-Labile: A Flexible Method to Improve Labile PTM Analysis in Proteomics.** *Mol Cell Proteomics.* 2023; **22**(5)**:** 100538.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

50. Kong AT, Leprevost FV, Avtonomov DM, *et al.*: **MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics.** *Nat Methods.* 2017; **14**(5): 513–520.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

51. Geiszler DJ, Kong AT, Avtonomov DM, *et al.*: **PTM-Shepherd: Analysis and Summarization of Post-Translational and Chemical Modifications From Open Search Results.** *Mol Cell Proteomics.* 2021; **20**: 100018.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

52. Freestone J, Keich U: **Re-investigating the correctness of decoy-based false discovery rate control in proteomics tandem mass spectrometry.** *bioRxiv.* 2023; 2023.06.21.546013.
Publisher Full Text

53. Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ? ✔

Version 1

Reviewer Report 19 June 2024

https://doi.org/10.21956/openreseurope.18615.r40925

© **2024 Harvey V.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Virginia Harvey**
University of Chester, Chester, England, UK

Chiang et al. set about analysing almost 15 million ancient fragment ion (MS2) spectra from Nature and Science portfolios and looked for identification rates across the dataset. They found that over 94% of the queries remained unmatched/uncharacterised, resulting in only 6% of MS2 spectra that had actually been genuinely assigned to peptide sequence database matches. Of these, some still could have been contaminant matches, making the actual assigned MS2 spectra even less than this! This paper is very well written, clear, and makes the important case for effective data sharing practices to be employed by everyone in the field of palaeoproteomics (and beyond) to best allow for methodological developments, and to get the most output out of often irreplaceable ancient samples. I recommend indexing without changes.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Palaeoproteomics, zoology, marine biology, ZooMS

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 12 June 2024

https://doi.org/10.21956/openreseurope.18615.r40929

© 2024 Parker G. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Glendon Parker
University of California, California, USA

Spectra without Stories: reporting 94% dark and unidentified ancient proteomes.
The author's conduct a meta-analysis of previously submitted paleoproteomic datasets that were published in high-quality / high-impact publications. These datasets for the most part originate from the network of highly productive collaborations centred around the Globe Institute at the University of Copenhagen. As such the datasets have consistent sample processing, data generation, and analytical workflows. The authors provide commentary about the degree of unidentified spectra in each sample and conclude that there are higher rates of unassigned spectra in paleoproteomics than more complex data dependent acquisition datasets.
In terms of a meta-analysis the manuscript is not controversial, but also does not really address any of the hypotheses that could be tested given the advantage of the larger combined datasets available or suggested by the authors. An obvious and necessary item to correct is the lack of a confidence interval in the estimate of the size of the 'dark paleoproteomes'. With no interval the number has no meaning.
The following is an incomplete list of suggestions in no particular order.

1. The language used is tentative. The words 'may' or 'could' show up several times at critical moments and undercut the impact of the manuscript. The study is a meta-analysis, and so definitive statements may be difficult to justify. Ideally meta-analyses test hypotheses that can only be examined in a larger context. Nevertheless, the authors should be definitive when justified, but also refrain from speculation when appropriate.
2. The first paragraph of the introduction is overly detailed, and a broader framing of the author's analysis is underdeveloped. We recommend the use of a background section to separate the necessary detail from the broader themes and hypotheses that are normally examined in an Introduction.
3. Is the use of high-quality PRIDE and MassIVE control datasets a relevant comparator for paleoproteomics? These are cell culture datasets. Fully differentiated tissues typically have lower yields and fewer detected proteins. Biological tissue proteomes provide less data but may be more relevant. Surely the authors have access to a modern skeletonized sample dataset.
4. The issue raised by the first reviewer is pertinent. The figure is attractive but not that

informative. Likewise, data that would help provide context, such the age of the sample, or degradation indices such as deamidation, or median peptide length are left out. A relationship between PSM efficiency and total spectra number would be more informative and easily accessible. Likewise, a ratio of PSMs with and without PTM modifications would be informative as well.

5. The datasets, while representing a range of targets, are selective. Some such as pottery are under-represented. Other groups are conducting paleoproteomics from a broader range of sample types, processing strategies, instruments, and analytical workflows. They may not be as good in quality or impact, but conclusions that are drawn from a larger, broader dataset will be more representative, universal, and applicable.
6. The authors could provide more detail on what is meant by heterogeneity of PTMs. PTMs come in two distinct types, biological and exogenous chemical modifications. These have fundamentally different etiologies and impacts. It was gratifying to see Maillard reactions mentioned but these likewise should be expanded on.
7. Finally, it is probably best to refrain from mentioning a strategy for a refined data search strategy without doing it. It is a good idea and leaving it out will protect the authors from losing priority.

Sample degradation in paleoproteomic samples is a challenge at many levels, in addition to low yields, the chemistries at play increase variation in peptide targets, distributing the remaining peptide backbone over an increasing number of peptide species. Spectra discarded by PSM algorithms may be a rich source of additional data and older datasets may continue to provide information with new bioinformatic workflows, well after the samples have been processed and fragmentation data generated. We look forward to seeing this work unfold.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
No

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Forensic Proteomics, Paleoproteomics, Biochemistry, Mass Spectrometry. Proteomic Genotyping, Proteomic Sex Estimation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 10 May 2024

https://doi.org/10.21956/openreseurope.18615.r39932

© 2024 Cleland T. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Timothy Cleland
Smithsonian Museum Conservation Institute, Suitland-Silver Hill, Maryland, USA

Chiang et al. evaluate the levels of spectral identification in paleoproteomic data that has been previously published in Science and Nature portfolio journals.

"To address the question... Science and Nature Portfolio journals" I understand limiting the choices of journal, but it seems unhelpful to use predominantly publications that use MaxQuant for database searching. Mascot also tends to be rather conservative in its identifications and the majority of the field does not use it. Can you explore other publications that use alternative search strategies as well (e.g., Proteome Discoverer/Sequest, MSFragger, MetaMorpheus)?

"The identification rates of palaeoproteomics datasets range from 0.47% to 12.61%" Can you give both the overall percentages and also split this by search algorithm because they aren't the same?

Figure 1. Please use a colorblind accessible palette. These colors are difficult to distinguish and PXD014657 becomes almost invisible in multiple colorblind viewer types. Also, I think this figure would be better as a regular bar chart making the comparison easy without having to have an index. Alternatively add the ID rate % to the end of the bars instead. Also, add the first author of each publication with its PXD number for ease of comparison because the data availability section uses numerical citations without easy connection to the authors.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Paleoproteomics, mass spectrometry, proteomics, bone

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.