Benchmarking the identification of a single degraded protein to explore optimal search strategies for ancient proteins

Ismael Rodriguez Palomo[1], Bharath Nair[1,2], Yun Chiang[2,3], Joannes Dekker[2,4], Benjamin Dartigues[5], Meaghan Mackie[2,6,7,8], Miranda Evans[1,4], Ruairidh Macleod[1], Jesper V. Olsen[6], Matthew J. Collins[1,2]

[1] McDonald Institute for Archaeological Research, University of Cambridge, United Kingdom
[2] Globe Institute, University of Copenhagen, Denmark
[3] The Nice Institute of Chemistry, Université Côte d'Azur, Nice, France
[4] Department of Archaeology, University of York, United Kingdom
[5] Department of Science and Technology, University of Bordeaux, France
[6] Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark
[7] School of Archaeology, University College Dublin, Ireland
[8] Archaeobiomics, Department of Life Sciences and Systems Biology, University of Turin, Italy

Keywords

Palaeoproteomics; beta-lactoglobulin; False Discovery Rate; benchmarking; *de novo*; open search

Abstract

Palaeoproteomics is a rapidly evolving discipline, and practitioners are constantly developing novel strategies for the analyses and interpretations of complex, degraded protein mixtures. The community has also established standards of good practice to interrogate our data. However, there is a lack of a systematic exploration of how these affect the identification of peptides, post-translational modifications (PTMs), proteins and their significance (through the False Discovery Rate) and correctness. We systematically investigated the performance of a wide range of sequencing tools and search engines in a controlled system: the experimental degradation of the single purified bovine β-lactoglobulin (BLG), heated at 95 ºC and pH 7 for 0, 4 and 128 days. We target BLG since it is one of the most robust and ubiquitous proteins in the archaeological record. We tested different reference database choices, a targeted dairy protein one, and the whole bovine proteome and the three digestion options (tryptic-, semi-tryptic- and non-specific searches), in order to evaluate the effects of search space and the identification of peptides. We also explored alternative strategies, including open search that allows for the global identification of PTMs based upon wide precursor mass tolerance and *de novo* sequencing to boost sequence coverage. We analysed the samples using Mascot, MaxQuant, Metamorpheus, pFind, Fragpipe and DeNovoGUI (pepNovo+, DirecTag, Novor), benchmarked these tools and discuss the optimal strategy for the characterisation of ancient proteins. We also studied physicochemical properties of the BLG that correlate with bias in the identification coverage.

Introduction

Proteomics software has been developed to identify and quantify proteins for different purposes and in various contexts, by processing and analysing mass spectrometry data (Aebersold & Mann, 2003). Many of these tools have been developed to analyse proteins in complex mixtures and map for post-translational modifications (PTMs). Most studies involve querying the mass spectra against a database of possible proteins (database search*, DBs*). Others aim at reconstructing the peptide sequences *de novo* from the spectra, without using any external data (*de novo* sequencing, *dnS*). Increasingly these methods are being applied beyond 'typical' tissues for which these software are optimised.

DBs approaches work effectively on modern tissues from well-studied organisms with complete sequence databases, typically matching up to 50% of MS/MS queries to peptides (peptide-spectrum matches, PSMs) in a typical tandem mass spectrometry experiment (Cox & Mann, 2008). Larger studies looking at PRIDE (Perez-Riverol et al., 2021) and MassIVE databases have determined that around 75% of submitted spectra remain unidentified (Griss et al., 2016; Bittremieux et al., 2022). Metaproteomics studies attempt to map proteins derived from complex samples such as microbiomes, in which it may be difficult to identify all the targets present without a proteogenomics strategy. The same is true for highly variable sequences, such as those found in immunoglobulins or novel sequences as found in cancers (Alfaro et al., 2017). *dnS* tools like PEAKS (Zhang et al., 2012) were initially developed to characterise these sequences. In addition to novel proteins introduced from the microbial community, the application of proteomics to food and forensics is complicated by an increase in the number of PTMs due to both enzymatic and chemical modifications or introduced by bacterial effectors (Liu et al., 2023).

Palaeoproteomics, the study of ancient proteins that are often fragmented and heavily modified, combines all of the above challenges. From studies of historical artworks to wider studies of material culture or skeletal remains in archaeology and palaeontology, chemical degradation adds further to the challenges of identification. In the field, most published ancient spectra remain unidentified. *De novo* sequencing has been decisive in studies looking at the phylogeny and evolutionary history of extinct organisms (Welker et al., 2015, 2019; Chen et al., 2019; Presslee et al., 2019; Cappellini et al., 2019). However, noise and missing ions might produce errors that can never be identified without matching to a database sequence. Muth and Renard (2018) evaluated the accuracy of different *dnS* software. They showed that only around 35% of the complete peptide sequences were correct on real HCD (higher-energy collisional dissociation) data and 85% on simulated data. On a different front, there are increasing efforts to minimise and identify modern protein contaminants in ancient samples. Deamidation rates have been proposed and applied to authenticate ancient proteins (Ramsøe et al., 2020, 2021). However, a clear correlation between

deamidation rates and time is yet to be established (Schroeter & Cleland, 2016). A deeper understanding of deamidation and other degradation patterns is key for ancient protein authentication.

At present, the field of palaeoproteomics has tended to use the same software tools as those originally designed for identification on targeted tissues. And due to the awareness of the specific challenges, the community sets and demands stringent approaches to interrogate the data (Hendy, Welker, et al., 2018; Hendy, Warinner, et al., 2018; Warinner et al., 2022). This includes the consideration of an increasingly wide array of post-translational modifications (PTMs), well-represented databases which encompass every potential constituent of samples, robust rules on the identification of peptides and proteins and statistically significant false discovery rates (FDRs) (Benjamini & Hochberg, 1995; Choi & Nesvizhskii, 2008; Käll et al., 2008). PSMs leading to extraordinary claims or novel peptides are manually checked, and at least 2 uniquely identified peptides are required to claim the presence of a specific protein. However, there is a lack of a systematic assessment of the effectiveness of these tools, approaches and standards in identifying ancient degraded proteins. The number of new strategies, software and packages being continuously developed to address problems with *dnS* and *DBs* already highlights issues with the different strategies. Because of all this, there is an increasing demand for all raw data to be made public in proteomics repositories for reanalysis. Along with it, the use of free and open-source software, standards and formats facilitates reproducibility and the adoption of common practices.

When using *DBs* software to analyse mass spectrometry data obtained from ancient protein samples, researchers often face decisions concerning parameter selection tailored to the specific sample types. These choices require balancing compromises and relying on a clear research question and hypothesis, and expert, multidisciplinary knowledge about the context of the samples. Broadly, one needs to consider the size of the search space, significance of results, computational resources and time. If the search space is too large, the memory usage and computation time increase and the significance of the identifications is compromised (Jeong et al., 2012; Noble, 2015). On the other hand, if the search space is too small, there is the risk of missing proteins present in the samples. In *de novo* sequencing, the peptide sequence is derived directly from the spectra, and therefore, it frees researchers from these decisions and compromises. However, care must be taken with the accuracy of the sequences (Muth & Renard, 2018; Beslic et al., 2023), especially in phylogenetic and evolutionary studies dealing with novel peptides of extinct, unsequenced species. In both *DBs* and *dnS*, peptides of interest are usually validated using BLASTp or tBLASTn (Altschul et al., 1990) against a reference database.

The search parameters that most impact the search space size and need more tuning are the type of enzymatic cleavage, the protein database, and the PTMs targeted. Evaluating the degree of sample degradation prompts the researcher to determine the suitable enzymatic cleavage method

(specific, semi-specific or non-specific) and the post-translational modifications (PTMs) to be targeted during the search. The specific archaeological and palaeontological questions that the researcher is trying to answer, along with the type of sample, determine the scope of the database. The size of the search space in turn affects the significance of the results. To control for the significance of multiple PSMs, the false discovery proportion (FDP) of a search is estimated using a Target-Decoy Competition (TDC) (Moore et al., 2002; Elias & Gygi, 2007), from which a False Discovery Rate (FDR) is obtained. A decoy database is built by reversing the sequences in the target database and it competes together for PSMs. It therefore contains known incorrect sequences that share statistical properties with the target database like amino acid distribution or peptide length. In the end, it will result in known incorrect PSMs that would have otherwise been considered positive. This allows sorting PSMs by a matching score and setting filters so that an expected given fraction (the FDR) of PSMs are false (the decoy ones). Thus the FDR is just an expectation of the actual proportion of false discoveries (the FDP). In fact, several studies have pointed out the divergence between the FDP and FDR (He et al., 2015; Madej & Lam, 2023; Ebadi et al., 2023), and the “instability” of the TDC approach that can lead to an underestimation of the FDP in high-resolution mass spectrometry (Couté et al., 2020). Moreover, FDRs are challenging in the study of complex or uncharacterised proteomes, as “treating each search result equally is methodically incorrect when peptides and proteins are not equally likely to be measured by LC-MS/MS and identified by search engines” (Wang et al., 2022).

In this study, we wish to explore which solution is optimal for degraded proteins given these constraints. For this, we create a deliberately simple experimental dataset which focuses not on the diversity of a complex mixture, but on the changes caused by mild and extreme degradation. Our model is a suspension of a small and well-characterised protein, two isoforms of a 156 amino acid long lectin, bovine β-Lactoglobulin (BLG), heated at neutral pH conditions for 0, 4 and 128 days, respectively. The three time points are designed to test the ability of these datasets to identify a mild (trypsin cleaved) and extreme example of protein degradation in which most surviving peptides are amenable to mass spectrometry and enzymatic digestion is not necessary. We explore a range of different software solutions and strategies to identify MS2 queries generated by this experiment and discuss the results in connection with the common practices and challenges in palaeoproteomics.

We chose BLG, as it is one of the most commonly preserved and identified proteins in dental calculus and ceramics and thus has become the main protein used in archaeology to explore the spread of dairying in the past (Warinner et al., 2014; Hendy, Warinner, et al., 2018; Hendy, Colonese, et al., 2018; Jeong et al., 2018; Charlton et al., 2019; Wilkin et al., 2020, 2021; Bleasdale et al., 2021; Tanasi et al., 2021). It is suggested that mineral surfaces play an important role in the survival of this protein (Fonseca et al., 2022), as it has been also found for other eggshell proteins (Demarchi et al., 2016). Evans et al. (2024) explored a range of protein

properties that might enable their survival in fresh ingredients and in cooked foodcrusts before and after burial. We explored as well several physico-chemical properties of the BLG: structural flexibility, solvent exposure, amyloidogenicity and isoelectric point; and linked them to the degradation, detection and identification of different regions of the BLG.

Materials and Methods

The target protein

BLG was purchased from Sigma Sigma Aldrich (CAS No.: 9045-23-2). It is a single purified protein with purity greater than 90% and traces of other bovine milk proteins. No further purification was undertaken.

Experimental setup

1.75 mg of BLG was dissolved in 7 mL of Molecular Grade Millipore water at pH ~7 (to a final concentration of 250 µg/mL) in borosilicate glass vials closed with screw caps with polytetrafluoroethylene (PTFE) lining on the inside.

Samples and blanks were placed on a VWR® dry block heater with a heated lid maintained at a constant temperature of 95 ^{o}C for 128 days. 200 µL was sampled from each glass vial at designated time lengths (0, 4 d, and 128 days) to 1 mL Eppendorf Safe-Lock® tubes. Samples were then subjected to vacuum centrifuge at 60 ^{o}C to dry the thermal degradation products and were stored at -20 ^{o}C. A USB data logger (ebro EBI 310-T1) connected to an external sensor was used for monitoring temperature throughout the experiment. A schematic representation of the experimental setup is shown in Figure 1a.

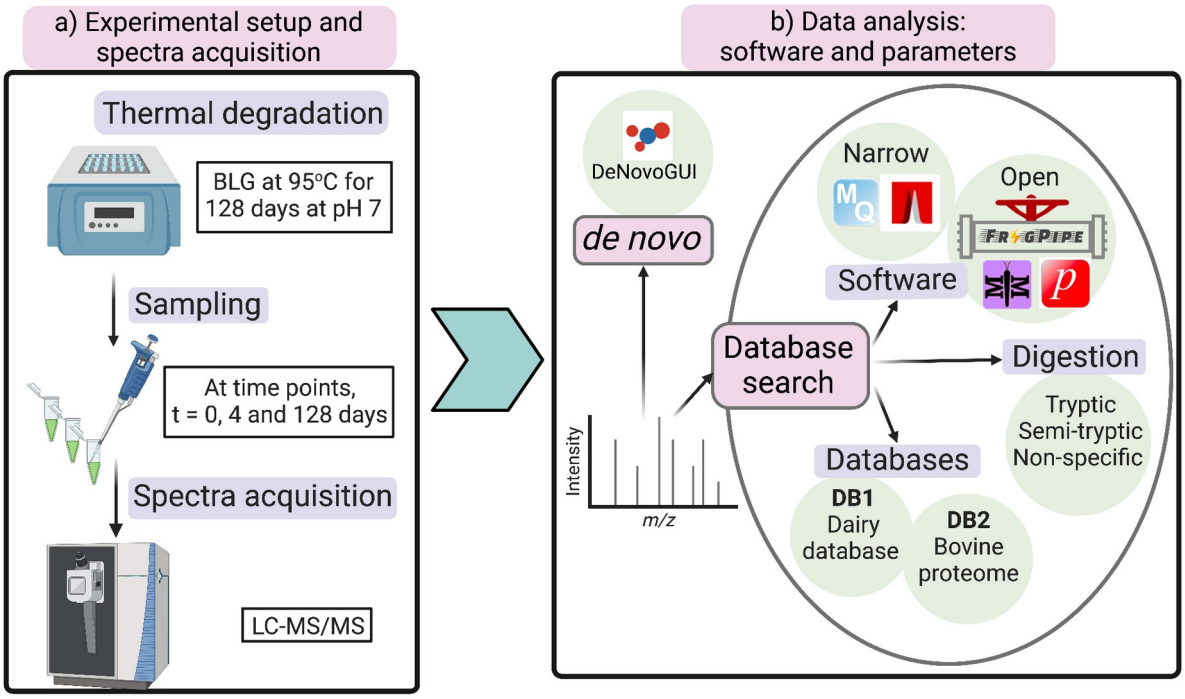


Figure 1. Schematic representation of the experimental setup (a) and the peptide identification strategies tested, including the different software, databases, and digestion (b).

LC-MS/MS

The peptides from the 0 and 4-day experiments were extracted following a previously published protocol (Cappellini et al., 2019) for degraded samples that consist of denaturation, reduction, alkylation, enzyme digestion, and desalting as described below.. Samples from each glass vial were resuspended to 300 µL solution consisting of 2 M guanidine hydrochloride (GuHCl, Thermo Scientific), 100 mMTris buffer (pH ~8), 20 mM 2-chloroacetamide (CAA) and 10 mM Tris (2-carboxyethyl)phosphine (TCEP) in ultrapure water. 0.2 µg of mass-spectrometry grade rLysC (Promega) enzyme was added prior to incubation with shaking for 2 h at 37 ºC. Samples were subsequently diluted to 0.6 M GuHCl and 0.4 µg of mass-spectrometry grade trypsin was added and subjected to overnight incubation with shaking at 37 ºC. The peptides at 128 days were alkylated but not digested by trypsin.

On the next day, samples were acidified using 10 % *(v/v)* trifluoroacetic acid (TFA) and purified using in-house prepared C18 solid-phase extraction StageTips™ (Rappsilber et al., 2007) by following a previously published protocol (Cappellini et al., 2019) described as follows. Stage-Tips were prepared in-house wherein two disks, pre-pierced with a 14 gauge blunt-end needle, were pushed out of an Empore C18 membrane (3M) into the end of each P200 micropipette tip. The Stage-Tips were sequentially conditioned with 150 µL of methanol, 150 µL of 80% acetonitrile solution (containing 80% acetonitrile, 0.1% trifluoroacetic acid (TFA), and ultrapure molecular biology-grade H2O -v/v/v-), and 150 µL of 0.1% TFA in H2O (v/v). Acidified peptides were then

loaded onto the Stage-Tips and were immobilized on the C18 filter by centrifugation at 1300 ×g until all the solvent had passed through. The filter was washed with 150 µL of 0.1% TFA in H2O (v/v), was centrifuged at 1300 ×g until dry. Immobilized peptides within Stage-Tips were eluted using 150 µL of 50 % acetonitrile solution (consisting of 50 % acetonitrile, 0.1 % TFA, and ultrapure molecular grade water) and was stored at -20˚C.20 to 25 µL of eluted samples were transferred to the MS analysis facility at the Novo Nordisk Foundation Center for Protein Research (CPR), University of Copenhagen. Samples were subjected to nanoflow liquid chromatography-tandem mass spectrometry (nano-LC MS/MS) using an in-house packed column on an EASY-nLC ™ 1200 system connected to an Orbitrap Exploris 480 (Thermo Scientific, Germany) mass spectrometer. A column temperature of 40 ^{o}C was maintained with an integrated column oven. For each run, 3 uL of sample was separated using a linear gradient from 5% to 30% buffer B in 25 minutes followed by a 2-minute step to 45% buffer B. Following the linear gradient, the column was washed by increasing the concentration of buffer B to 80% in 2 minutes, and remaining there for 2 minutes. Finally, it was re-equilibrated back to 5% in 2 minutes and held for 2 minutes, resulting in a final acquisition time of 35 minutes.

The Orbitrap Exploris 480 was operated in data-dependent acquisition mode using a Top 10 method. The spray voltage was at 2 kV, the S-lens RF level was at 40%, and the ion transfer tube was kept at 275 °C. Full scan MS were acquired at a resolution of 120,000 in a mass range of 350-1400 with an Automatic Gain Control (AGC) target of 300 and a maximum ion injection time of 25 ms. Fragment MS/MS spectra were recorded using Higher-energy Collisional Dissociation (HCD) with a maximum ion injection time set to 118 ms with an AGC target value of 200 at a resolution of 60,000 with a fixed first mass of 100 *m/z*. Normalised collision energy (NCE) was 30 %. The isolation window was set to 1.2 *m/z*, and the dynamic exclusion was 20 seconds.

Search strategies

Open and narrow database search software

We chose various software based on different principles and algorithms (Table 1 and Figure 1b). We chose a range of free, open-source or closed-source software, but included Mascot, which is licensed, since it is frequently used in palaeoproteomics.

Table 1. Software used for benchmarking			
Software	Type of search	Platform	Citation
Fragpipe v19.1 with MSFragger v3.7	*DBs* open	University of Cambridge HPC[1]. 56 CPUs in Intel Cascade Lake node and 3.5GB per CPU	Kong et al., 2017; Chang et al., 2020; da Veiga Leprevost et al., 2020; Geiszler et al., 2021
	DBs narrow		
Open-pFind 3 v3.2.0	*DBs* open	MiniMax[2]. 24 AMD Ryzen Threadripper 3960X CPUs and 64 GB RAM	Chi, Liu, Yang, Zeng, Wu, Zhou, Niu, et al., 2018
Metamorpheus v0.0.320	*DBs* Multi-notch	Mjolnir[3]. 20 CPUs in AMD EPYC 4773 node and 20GB CPU	Solntsev et al., 2018
MaxQuant v2.1.0.0	*DBs* narrow	MiniMax[2]. 24 AMD Ryzen Threadripper 3960X CPUs and 64GB RAM.	Tyanova et al., 2016
		University of Cambridge HPC[1]. 56 CPUs in Intel Cascade Lake node and 3.5GB per CPU[†]	
Mascot v2.8.2	*DBs* narrow	Mascot Server[4]	Perkins et al., 1999
Novor v1.05.0573	*dnS*	MiniMax[2] with DeNovoGUI (Muth et al., 2014). 24 AMD Ryzen Threadripper 3960X CPUs and 64GB RAM	Ma, 2015
DirecTag v1.4.66	*dnS* hybrid		Tabb et al., 2008
PepNovo+ v3.1	*dnS*		Frank & Pevzner, 2005; Frank, 2009a; b

Table 1. Software used for benchmarking, the type of search and the computing platform or system in which they were run. [1]University of Cambridge High Performance Computing, Cambridge Service for Data-Driven Discovery (CSD3) https://www.hpc.cam.ac.uk/high-performance-computing. [2]MiniMax Computer Asus TRX40 Stone Magma High Performance Workstation, [3]Computing cluster at Globe Institute, University of Copenhagen https://mjolnir-ucph.readthedocs.io, [4]Mascot Server https://www.matrixscience.com/server.html. [†]Only tryptic and semi-tryptic runs were run on the University of Cambridge HPC, for time comparison with Fragpipe on the same system.

The LC-MS/MS files were run in Thermo RAW file format on pFind3, Metamorpheus, MaxQuant and Fragpipe. For Mascot and DeNovoGUI the files were transformed to MGF (Mascot Generic Format) using msconvert (Chambers et al., 2012).

DBs software can be categorised based on the approach for the identification of PTMs and the precursor mass tolerance. In narrow-search, the targeted PTMs are defined *a priori,* and predicted peptide mass including PTMs must match the measured precursor mass with a narrow error. In contrast, open search allows wide precursor mass tolerances in order to identify peptides with unaccounted PTMs "*on-the-fly*". Multi-notch sets a specific set of precursor mass shifts instead of a range. We note that most narrow-search software (e.g. Mascot, MaxQuant) can in theory also perform open search, by just establishing a large precursor mass tolerance. Unlike specifically

authored tools (e.g. pFind, Fragpipe or Metamorpheus), such software lacks other key components for open search. In Fragpipe, MSFragger implements a localization-aware *open search* and the fragment index allows a fast search (Kong et al., 2017; Yu et al., 2020), while Crystal-C (Chang et al., 2020) discards artefacts, and PTMshepherd (Geiszler et al., 2021) helps summarising and analysing open search results. pFind is also optimised for fast searches and is also able to localise PTM mass shifts (Chi, Liu, Yang, Zeng, Wu, Zhou, Niu, et al., 2018). Metamorpheus uses the G-PTM-D workflow, which runs a multi-notch search and an augmented narrow database search to speed up the process (Solntsev et al., 2018). We will refer to Metamorpheus as an open or multi-notch search software.

Conversely, open search software also runs in narrow or closed search mode. Fragpipe was run both in narrow and open mode, referred to as narrow-Fragpipe and open-Fragpipe. For both the open and narrow searches PeptideProphet (Keller et al., 2002; da Veiga Leprevost et al., 2020)) was run for the validation step. Percolator is not compatible with open search. pFind was only run in Open search mode, and it will be referred to as Open-pFind or pFind.

Each software was run using 6 different combinations of parameters (Table 2) that broadly change the search space. This can be appreciated in the number of peptides to search in the spectra that each one generates (Table 2). The selection of parameters is explained in the sections below.

Table 2. Database Search Parameters

Run #	Database	Enzymatic cleavage	# peptides
1	DB1: dairy database 227 proteins and 62,185 residues	Tryptic	9,302
2		Semi-tryptic	102,127
3		Non-specific	1,092,998
4	DB2: bovine proteome 37,606 proteins and 22,906,446 residues	Tryptic	3,640,393
5		Semi-tryptic	38,385,707
6		Non-specific	415,906,748

Table 2. The combination of search parameters used in each of the database search software. The number of proteins, residues and peptides from each database and enzymatic cleavage combination is shown. The number of peptides is calculated allowing up to 2 missed cleavages and peptide length range of 7 to and 25.

Dairy database and bovine proteome

Despite the fact that we know that the samples were almost purely composed of BLG and perhaps contaminants, we used two of the widely used database strategies, in order to provide robust and comparable statistics with other studies. Broadly, we define a wide database as one that covers a limited set of proteins from a wide range of species. Many studies use a wide, targeted database, for example (Cappellini et al., 2019; Evans et al., 2023) as they look for just specific proteins (dairy or enamel) from different species. On the opposite side, a deep database is one that represents the whole known proteome or proteomes from a specific species or taxa. These are used, for example, in medical studies that focus on a specific organism, like a human, mouse or zebrafish.

But in the study of ancient proteins usually a wide and deep strategy is adopted. A taxonomically broader database is used to explore the wider metaproteome, or dietary proteins with the entire SwissProt and TrEMBL (c. 251 million protein entries) (Bairoch & Apweiler, 2000) being the most typical target (Warinner et al., 2014; Wilkin et al., 2020, 2021; Bleasdale et al., 2021) although even larger ones are used (Jersie-Christensen et al., 2018).

Here we tested the two opposite types of databases: a wide one from (Evans et al., 2023) that includes common and reviewed dairy proteins (DB1) and as a deep one, the Bovine proteome (DB2) (Table 2 and Figure 1b). Both include common contaminants usually reported in palaeoproteomics (Evans et al., 2023) and 4 different bovine BLG variants. DB1 comprises 227 protein entries, while DB2 contains 37,606 proteins.

All software were run with both databases, except for narrow-Fragpipe, which was run only on the DB1.

Digestion: tryptic, semi-tryptic and non-specific

Most proteomics protocols include a digestion step, usually with trypsin. However, in palaeoproteomics the most degraded samples require no digestion step, removing the challenge of untangling diagenesis from the proteolytic background (Demarchi et al., 2016). This is the case with the 128-days sample. To test the differences in performance between using a tryptic, semi-tryptic and non-specific search we applied all three approaches on the three samples, each with an increasing degree of degradation (Table 2 and Figure 1). This is then reflected in the database search, which generates tryptic peptides (the two ends are tryptic, ie. result from a tryptic cleavage) out of the selected database to match the sample. However, in palaeoproteomics, samples are usually degraded, and hydrolysis generates a non-trivial amount of non-tryptic peptides. This means that the database search will also need to run on semi-tryptic mode, which generates peptides with just one tryptic end. The non-specific search produces all possible peptides within a certain length range. Semi-specific and especially non-specific searches increase the search space dramatically and therefore the running time and memory requirements. If combined with a large database, it might not be feasible to run for some software and systems.

Other parameters

We set the rest of the parameters as default or to values commonly used in Palaeoproteomics, to emulate what a standard run would be in a real scenario. We include glutamine and asparagine deamidation, oxidation of methionine, and N-terminal acetylation as variable PTMs, and cysteine carbamidomethylation as fixed. Peptide length range is 7 to 25 amino acids and the peptide mass range is 500 to 5000 Da. We allow up to 2 missed cleavages and turn off the match between runs feature in the software that implements it, so they can be compared and to keep identifications

independent between the 3 samples. Supplementary File S1 contains more detailed lists of parameters, compiled or adapted from the parameter files used in the set up of the searches.

Collation of Database Search (DBs) results

The different software reports results using different formats. These include pepXML, mzIdentML or a tab or comma-delimited table. We used Python 3.9 with lxml, pyteomics (Levitsky et al., 2019) and pandas packages to collate the results into a single PSM data frame. Each row is a single PSM from each sample and run. We processed all the results and extracted the following information from each PSM: Scan #, retention time, peptide sequence, protein accession, start and end position, whether it is a decoy match, calculated mass, delta mass, and q-value. q-value is the minimum FDR threshold at which the identification can be considered statistically significant. We calculated PSM-level q-values using pyteomics *q_value* function. We only include the best PSM per scan. When a peptide is matched against BLG from different species, if the bovine BLG is one of them, the start and end positions of the peptide are reported with respect to that one.

In instances where the same spectra is identified by different software and assigned to BLG, we calculated the Levenshtein distance between the two peptide identifications. This is the number of insertions, deletions and substitutions that transform one sequence into the other. We used the Python package Levenshtein (Bachmann, 2024).

After removing decoy matches and non BLG proteins from the PSM data frame, we pivoted it into a long format, taking each peptide's amino acid position into a single row. For example, a PSM row with peptide sequence TPEVDDEALEK, which starts at position 141 in the protein, is turned into 11 rows, one per amino acid, and with positions 141, 142, 143, … and 152. When all PSMs are considered, this allows for calculating each position's coverage along the protein sequence.

In this and subsequent analysis, the BLG amino acid positions will be given with respect to the beginning of the immature sequence, ie. counting the signal peptide (as they are given also in figures 7, 8 and 9). We do this to explicitly acknowledge the existence of the signal peptide, which at the same time is not present in the mature molecule in milk whey, and therefore there should not be any peptide starting before position 17. Moreover, signal peptides are always included in the protein database sequences used in database search engines.

De novo sequencing

These methods use machine learning techniques and employ graphs as data structures and algorithms applied to them. We run Novor, DirecTag and PepNovo+ using DeNovoGui (Muth et al., 2014) interface (Table 1 and Figure 1b). Novor uses two decision trees trained from peptide spectral libraries for its scoring PSM function. DirecTag, follows a hybrid approach and uses a

database to extend sequence tags derived from the spectra (Tabb et al., 2008; Zhang et al., 2012). PepNovo+ uses probabilistic network modelling and fragment ion intensity ranks prediction and scoring (Frank & Pevzner, 2005; Frank, 2009a; b) and is followed by a MS-BLAST (Shevchenko et al., 2001) query of the peptides against a database. We ran DeNovoGui with both sequence databases, DB1 and DB2, to understand how they can also affect these hybrid methods. pNovo+, which is included in DeNovoGUI did not produce enough meaningful results and was discarded from further analysis.

Mapping of *de novo* peptides

We mapped the *de novo* peptides back to the bovine BLG protein sequence allowing mismatches, that in this case account for sequencing errors. Since leucine (L) and isoleucine (I), aspartic (D) and glutamic acid (E) and deamidated asparagine (N) and glutamine (Q) have the same mass respectively, we 'blind' all the sequences, so that mismatches between these pairs are not counted as errors. I and L are encoded as a B, D and N as a X and E and Q as Z. Having all sequences blinded, we first create a 3, 4 and 5-mer index from the bovine BLG variants (BLG k-mer index). Then for each peptide, we generate a list of C and N-terminal 3, 4 and 5-mers and matched against the BLG k-mer index. These were used as seed matches that we extended by running a pairwise alignment using the Needleman-Wunsch (Needleman & Wunsch, 1970) algorithm. When a peptide matches to k-mers from different positions, the alignment with the highest identity is kept. In case of a tie, we keep the one originating from a longer k-mer, giving preference to C-terminal k-mers. We allow gaps in the alignment, accounting for errors in the *de novo* sequencing in which an amino acid is skipped or extra ones are added, due to noise in the spectra or missing fragment ions. Finally, alignments with overall identity below 70% are discarded. Figure 2 shows a diagram of the procedure. Once the peptides are mapped we pivot the data into long format as before, so we can calculate the coverage and accuracy of each position along the BLG sequence.

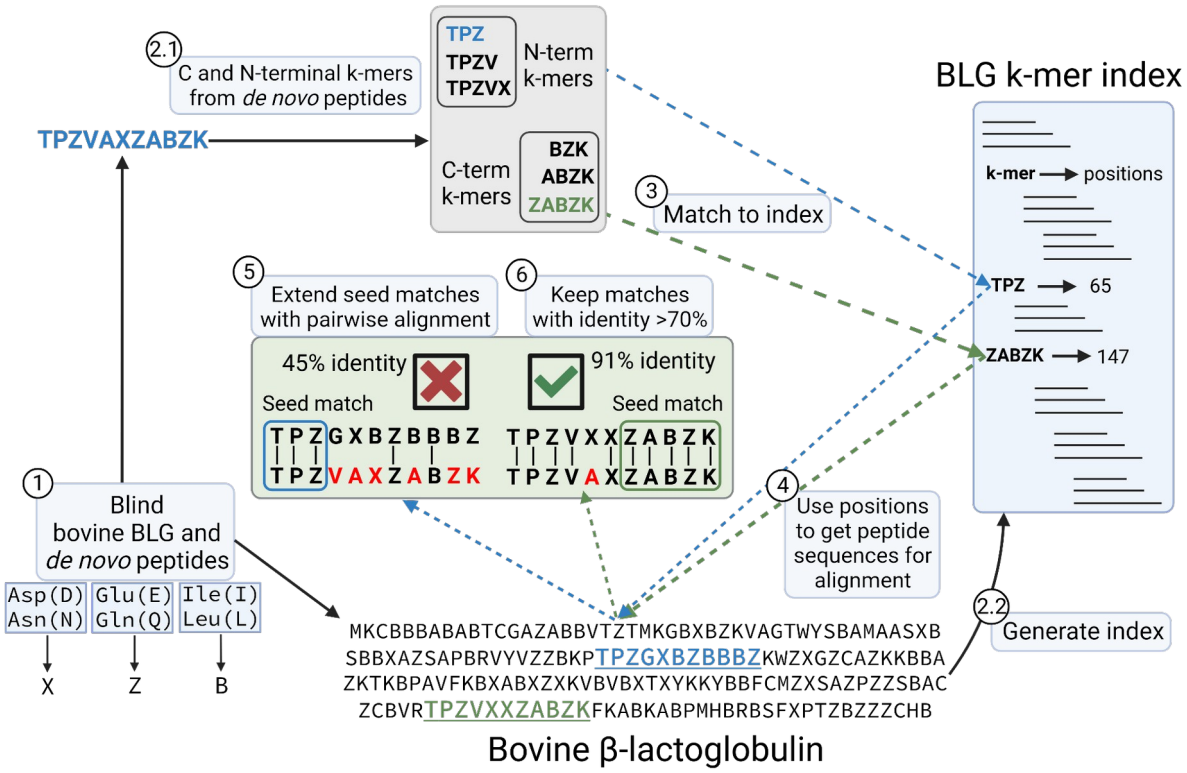


Figure 2. Schematic of the procedure to map the *de novo* peptides to the bovine β-lactoglobulin sequence, while allowing sequencing errors. 1) *de novo* peptides and BLG are 'blinded'. 2.1) C and N-terminal k-mers are extracted from the peptides. 2.2) BLG index is generated, with k-mers as keys pointing at the positions in the protein. Positions are with respect to the immature BLG containing the signal peptide 3) In this example, the peptide TPZVAZABZK matches 2 different k-mers from the BLG k-mer index, which come from positions 65 and 147 of the BLG product sequence. 4), 5) This results in two possible alignments of the query peptide against BLG. 6) The alignment with the highest accuracy is kept and overall, only alignments with identity above 70% are considered. In case of ties, the alignment with the longest seed match and coming from a C-terminal k-mer would be kept.

Peptide physico-chemical properties

We extracted or calculated several physico-chemical properties of the BLG: B-factors, relative solvent accessibility (RSA), amyloid propensity and isoelectric point.

We extracted the B-factors from the crystallographic structure of the BLG from PDB accession 7KOT. When determining the three-dimensional structure of a protein using X-ray crystallography, the B-factors are calculated from the attenuation of X-ray scattering due to thermal displacement of the atoms (Sun et al., 2019). B-factors have been extensively used to measure the flexibility of the backbone and side chains of proteins (Sun et al., 2019). The solvent accessible surface area (ASA), in $Å^2$, was extracted from the 7KOT structure from PDB using DSSP (Kabsch & Sander, 1983; Joosten et al., 2011). The RSA was calculated by dividing the ASA by the MaxASA obtained from (Tien et al., 2013). We calculated the isoelectric point of all identified peptides using the *pI* function in Pyteomics. For reference, we also calculated the isoelectric point along the BLG sequence using a sliding window of size 12. To estimate the probability of amyloid formation, we

used 3 different software, APPNN (Família et al., 2015), AmyloGram (Burdukiewicz et al., 2017) and ANuPP (Prabakaran et al., 2021).

Results and discussion

The LC-MSMS produced a total of 8,731 MS2 spectra for the 0 days sample and 9,108 and 10,141 for the 4 and 128 days respectively. The increase in the total number of spectra is consistent with the higher non-enzymatic hydrolysis expected in the more degraded samples, which would yield a wider variety of peptides.

Is FDR enough?

We calculated q-values for each PSMs in each database search software run and sample using pyteomics *q_value* function. Then we calculated the fraction of PSMs above a range of q-values from $1 \cdot 10^{-5}$ to $5 \cdot 10^{-2}$ and plotted the fraction of identified spectra (Figure 3). Supplementary Table S1 summarises the absolute number of PSMs for each run and software at FDR 0.01 and 0.05, which are highlighted as a blue dashed line in Figure 3, while Supplementary Figure S1 shows the same information only for FDR 0.01. We observe several trends:

1) Looking at each software separately, the fraction of identifications above the same FDR threshold, first slightly increases in the 4-days sample but then decreases in the 128-days.
2) Within the same software, on the 0 days sample, tryptic or semi-tryptic searches give the highest number of identifications, and only as degradation increases, in the 4 and 128 days, semi-tryptic, and then non-specific searches identify more spectra.
3) In each software and sample, the sets of parameters producing more identifications are always the ones using the smaller, dairy-specific DB1.
4) At the least stringent FDR, 0.05, and in the best case, less than 60% of the spectra is identified for the 0 and 4 days samples. For the 128 days sample, this is reduced to less than 30%.

The amount of identified PSMs at a given FDR are widely affected by the search space (Figure 3), controlled by the choice of the database and enzymatic search. And in general, it is remarkable that for such a simple system explored in this study, vast amounts of spectra remain unidentified. This is more pronounced for the most degraded and therefore complex sample (Figure 3). MaxQuant and Mascot identify less than 20% of spectra for the 0 days and 10% for the 128 days. This is significantly lower than the other software, almost for any parameter set. This makes them particularly affected by the decrease in the amount of PSMs when using the larger database.

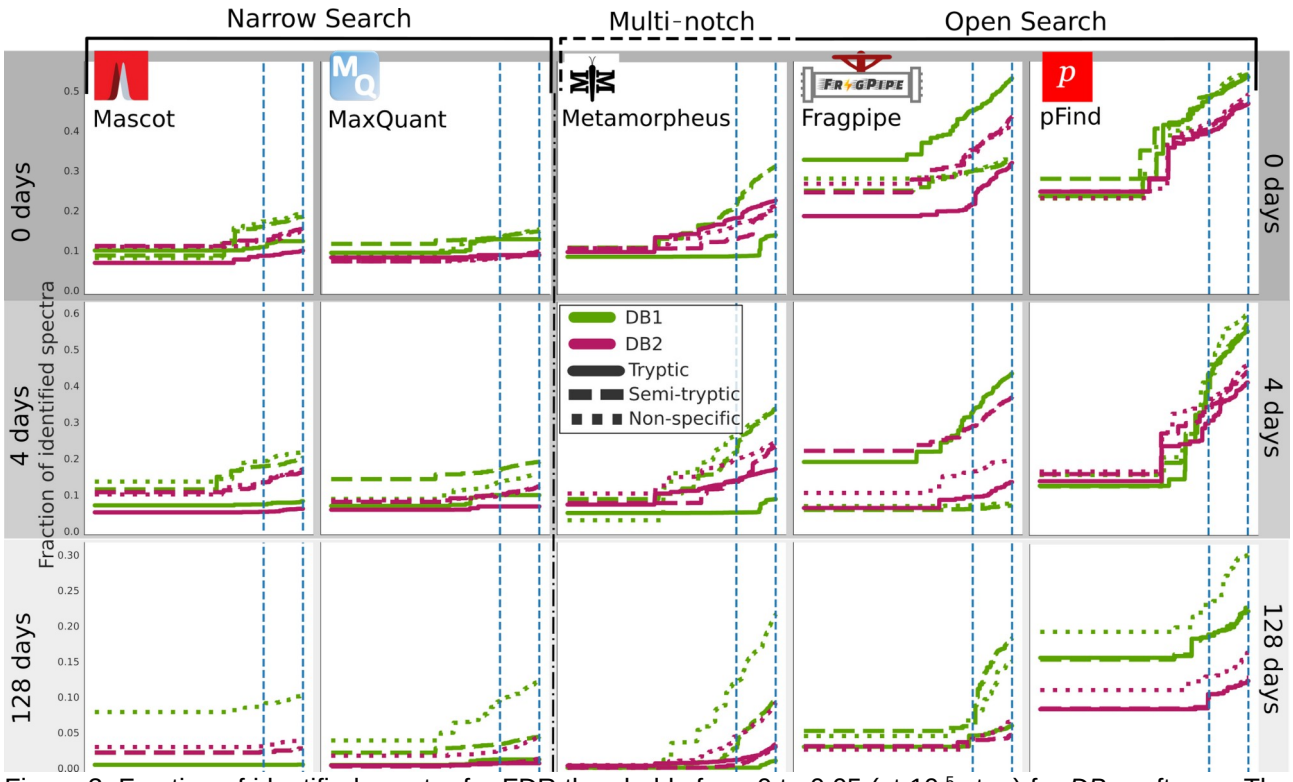


Figure 3. Fraction of identified spectra for FDR thresholds from 0 to 0.05 (at 10^{-5} step) for *DBs* software. The rows represent the increasing heating time (0, 4, 128 days), with each software represented in a column: MaxQuant and Mascot are narrow search software, Metamorpheus multi-notch and Fragpipe and pFind open search. Dairy database searches (DB1) searches are represented in green, and bovine proteome searches (DB2) in magenta. The solid lines are for tryptic searches, dashed for semi-tryptic and dotted for non-specific. The blue dashed lines mark the 0.05 and 0.01 FDR thresholds that are usually applied.

The main reason behind the reduction on the number of PSMs when we expand the search space is already known in proteomics and in statistics general: if by Increasing the search space, the proportion of peptides in the database that are not present in the samples also increases, there will be more chances of random and decoy matches with high scores (Jeong et al., 2012; Noble, 2015). Therefore, to achieve the same desired FDR, the score threshold to accept positive matches becomes more stringent, rendering fewer accepted PSMs (Jeong et al., 2012; Noble, 2015). This means that large databases containing a large number of proteins that are not present in the samples, like DB2, inadvertently boost the FDR and apparently reduce the number of identifications. Something similar applies to the use of non-specific searches for samples that are well-preserved and less degraded. This adds to the inherent divergence between FDR and FDP of the TDC (He et al., 2015; Couté et al., 2020; Madej & Lam, 2023; Ebadi et al., 2023). But the answer to this problem cannot just be to use small, targeted databases, even though it would reduce the computing time and resources and might produce more PSMs. We can speculate what would happen in the opposite scenario, using a very small database against a complex sample, for example dental calculus. To start, we might miss some of the contents of the sample. Another more inadvertent effect is that wrong, low-quality matches between a spectrum and a peptide might surface with a low enough FDR (since the search space is small). This is further pronounced if the correct peptide, which could have competed and achieved a better PSM, is absent in the

database. This is why common contaminants are included in the searches, even though we are not interested in them.

In general, we want the set of proteins and peptides that are in a sample and the set of what we are looking for (search space) to match as closely as possible (Jeong et al., 2012; Noble, 2015). While the content of a sample is not known *a priori*, it is clear in our simple system that using the whole bovine proteome to interrogate our data is utterly unnecessary and counter-productive. Going into a more realistic scenario, with real archaeological samples, having a clear research question and hypotheses and a good understanding of the historical, archaeological, diagenetic and taphonomic context will always help to delineate the optimal database and parameters given the compromises exposed. Studies like those of Wilkin et al. (2020, 2021), Bleasdale et al. (2021) and Hendy et al. (2018; 2018) are examples of different strategies and challenges regarding the complexity of samples and scope of the databases. They highlight the difficulty of building an unbiased and targeted database that encompasses all the contents of a sample. Targeted databases for different types of tissues like bone or enamel, microbiomes or other mixtures like foodcrusts need to be further developed.

Finally, another mechanism that reduces the amount of PSMs is the exclusion of correct matches because of a poor separation of the score distributions between the target and decoy matches. The low amount of identifications at any FDR on MaxQuant and Mascot, and the dramatic change when switching from 0.01 to 0.05 on Fragpipe, Metamorpheus, and pFind to a lesser extent (Figure 3 and Supplementary Table S1), indicates this is likely happening here. The choice of a score metric for the PSMs is of great importance, because a good scoring function will be able to assign low scores to wrong matches, and vice versa. Therefore, wrong matches can be excluded without eliminating correct matches. Furthermore, a subsequent classifier can be trained to separate the score distributions of the target and decoy hits, integrating different metrics such as the primary search score, peptide length, misscleavages, or PTMs. The use of post-processing classifiers is advised to set up accordingly and is now routinely integrated in most software pipelines: Fragpipe uses PeptideProphet (Keller et al., 2002; da Veiga Leprevost et al., 2020) or Percolator (The et al., 2016) (not used in this study, as it is not compatible with open searches), pFind integrates a linear classifier (Chi, Liu, Yang, Zeng, Wu, Zhou, Wang, et al., 2018). MaxQuant uses Posterior Error Probabilities (PEP) in a similar way (Tyanova et al., 2016) and Mascot also integrates Percolator in its pipeline.

Open search vs Narrow search

In the previous section, we commented on the implications of the database and enzyme digestion choices in the general TDC approach and FDR. However, the presence of a wide variety of PTMs in ancient, degraded proteins, offers further challenges to identification.

The largest differences in the number of PSMs are between Narrow and Open search software (Figure 3 and Supplementary Figure S1 Table S). Open search furnishes more identifications in all 3 samples, provided that DB1 is used with the appropriate enzymatic digestion (or lack). MaxQuant, Mascot and Fragpipe-Narrow tend to have a similar number of PSMs across the samples and enzymatic cleavage and behave similarly in terms of which settings produce more PSMs for each sample. On the other hand, the open search software has more variability, especially open-Fragpipe (Figure 3, Supplementary Figure S1 and Supplementary Table S1).

Open search is capable of identifying spectra from peptides bearing PTMs that are not accounted for and therefore remain unidentified (Chick et al., 2015). This is so because, in practice, the search space in this software is more unconstrained, while at the same time implementing algorithms that allow a fast search. FragPipe-open finds mass shifts not corresponding to the variable or fixed PTMs set, in 46.6%, 31.8% and 17% of the PSMs below FDR 0.05 for the 0, 4 and 128 days, respectively. The decrease in the identification of PTMs with the more degraded samples might seem counterintuitive. Nonetheless, the backbone cleavage and accumulation of PTMs in the 128 days samples might be so extensive that the Open search cannot even identify those spectra, and only a smaller fraction of the modified peptides is identified. In addition to the problems with FDR already commented, Kong et al. (2017) reported inherent problems with the TDC approach when using narrow search and an increase in the statistical power of open searches. They also reported that 68% of spectra reassigned by open search MSFragger to a different peptide from narrow search, had greater support in the former. They suggest there are potential false positives in spectra matched to unmodified peptides. However, Freeston et al. (2024) explored the trade-offs and discrepancies between open and narrow searches and dispute these claims. They attribute this to the presence of chimeric peptides and flaws in FDR control in PeptideProphet and Percolator. They found that narrow searches can perform better at identifying unmodified peptides. Fragpipe's performance with respect to its narrow search counterpart is lower in non-specific search of the 128 days sample and non-specific and semi-tryptic for the 4 days sample. This, as well as Fragpipe's more unstable performance, might be a consequence of the findings by Freeston et al. (2024). Like Fragpipe, pFind also includes a rescoring or reranking step, while Metamorpheus only relies on the primary search score to discriminate between true and false matches (Wenger & Coon, 2013).

In order to highlight the differences between algorithms within Open search, we created UpSet diagrams in Figure 4A with the common spectra identified in the different open search software and enzyme cleavage runs and for each of the samples. For FragPipe and Metamorpheus, on the 128 days sample, most spectra are identified only on the non-specific search, and the difference between the 3 types of digestion increases with sample degradation (Figure 4A). In contrast, in pFind most identified spectra are common in the 3 searches throughout the 3 samples, although the totals decrease (Figure 4A). Moreover, most of the spectra identified by Metamorpheus and

Fragpipe are also identified by pFind, while few spectra are exclusively identified by Fragpipe and/or Metamorpheus. Open-pFind takes a different approach than MSFragger or Metamorpheus; it finds k-mer tags in the spectra that are extended until there is a match to the peptide database, and given a mass shift (Chi, Liu, Yang, Zeng, Wu, Zhou, Niu, et al., 2018). Open-pFind is capable of finding non-tryptic peptides even in tryptic search in the 128 days sample (Figure 4A). The mass shift between the precursor ion and the peptide candidate is treated as a potential modification, and PTMs are proposed by testing all valid positions according to Unimod (Creasy & Cottrell, 2004).

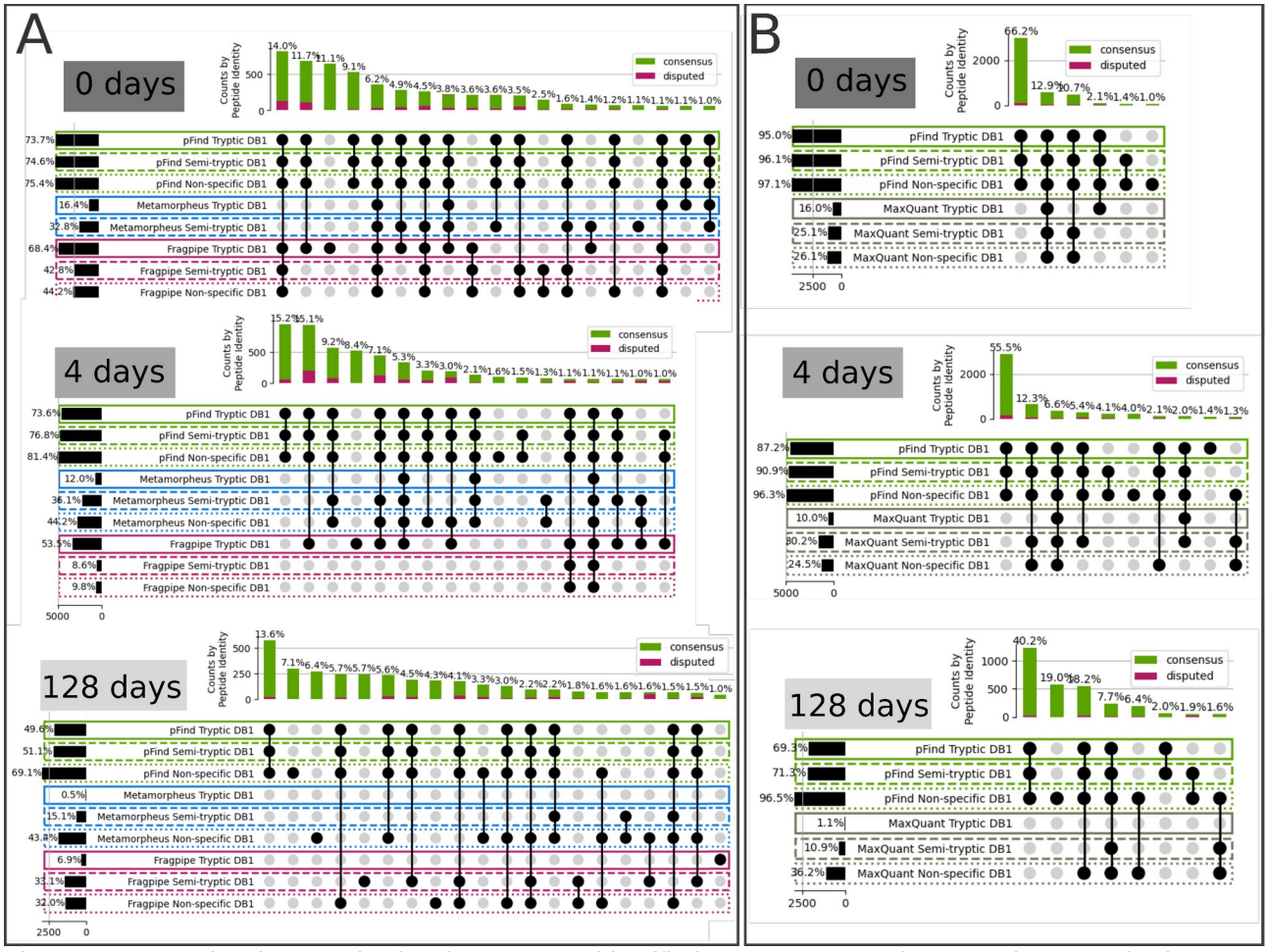


Figure 4. UpSet plots by sample showing common identified PSMs across software and enzymatic cleavage. PSMs below FDR 0.05 are considered. A) UpSet plots for pFind (green), Metamorpheus (blue) and Fragpipe (pink) and tryptic (solid line), semi-tryptic (dashed) and non-specific (dotted) searches. B) MaxQuant and pFind UpSet plots. All cases are with the dairy database (DB1). The vertical bars indicate the sizes of each set and the percentages are with respect to the size of the union of all sets. The degree of agreement between the peptide sequences of each set is represented in red (average pairwise Levenshtein distance less or equal than 1) and green (average pairwise Levenshtein distance greater than 1).

In Figure 4B we compared pFind and MaxQuant using UpSet diagrams, again considering PSMs below 0.05 FDR. The sets of spectra identified by MaxQuant are always contained in the set identified by at least one pFind run. In the most extreme case, for the 128 days sample, only 8% of the spectra are identified by the semi-tryptic or non-specific runs in MaxQuant, and not by the tryptic run on pFind (Figure 4B).

Open-pFind, even in the tryptic search on the most degraded sample, “unlocks” a spectra space that the other software does not identify. And in general, the Open search software provides identifications for spectra for which narrow search does not. We found that those PSMs exclusively identified by pFind tend to have a slightly lower score than those also identified by MaxQuant (Wilcoxon rank-sum test *p-value=0*). But even though significant, the magnitude of the difference remains small and many pFind-exclusive PSMs score high (Figure 5).

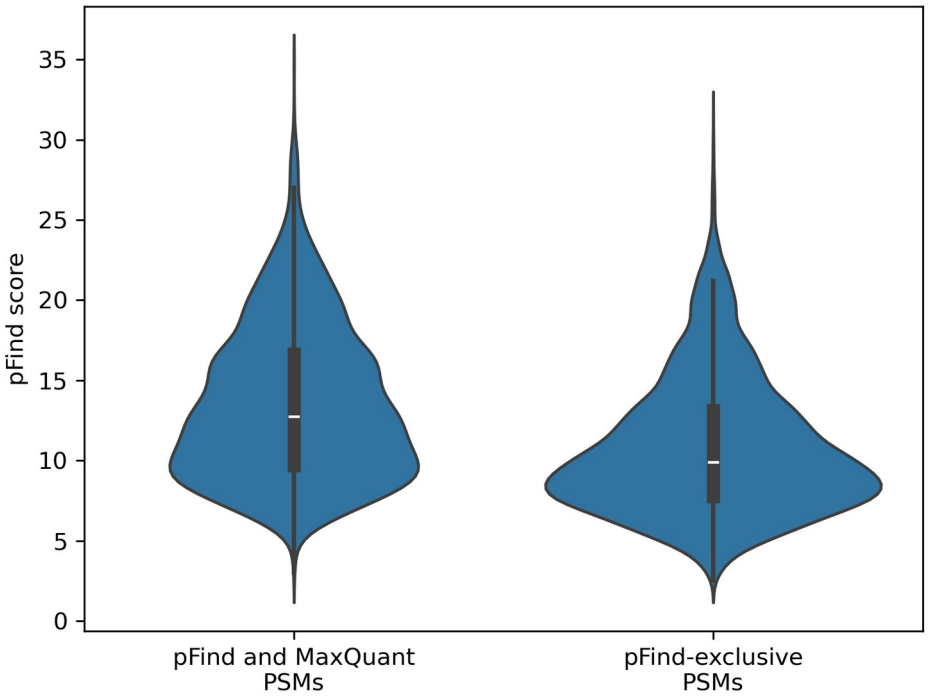


Figure 5. Distribution of the pFind scores for PSMs below 0.05 FDR identified both by pFind and MaxQuant and only identified by pFind.

For the cases, when they both provide a peptide identification for the same spectra, we explored to what extent the pFind and MaxQuant agree. Overall, the level of agreement decreases on the 128-days sample when running semi-tryptic or non-specific searches (Figure 6, left column bar plots). This is, with the more complex sample and unconstrained search space. We also compared the Andromeda and pFind scores for each PSM and the Levenshtein distances (the extent of disparity) between the identified sequences (Figure 6, right column scatter plots). There is an overall correlation between the scores, with those having the highest Levenshtein distances showing low degree of correlation, or scoring low in one or both software (Figure 6). However, disputed peptide sequences still overlap broadly with those where there is a consensus.

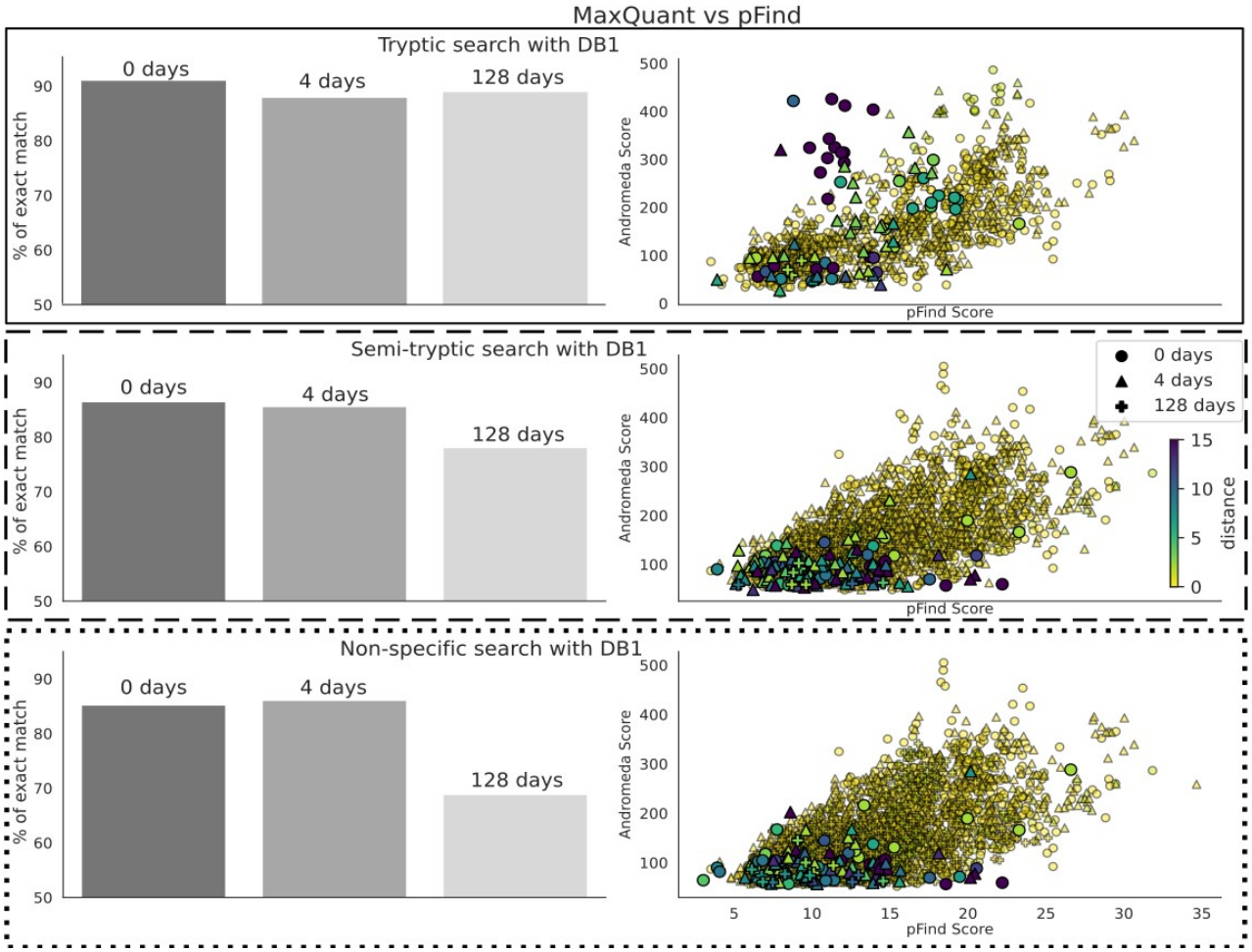


Figure 5. Comparison of the identifications made by Fragpipe and MaxQuant using the dairy database and tryptic search (top panels, with a solid outline), semi-tryptic (middle, with dashed outline) and non-specific (bottom with dotted outline). When both software identify the same spectra, we calculate the distance between the peptide sequences, which is expected to be 0. Panels on the left are a histogram of the Levenshtein distances, split by sample (0, 4 and 128 days). The text is the percentage of common spectra for which both sequences are identical. Panels on the right are the scores given by the software to that same PSM. The color of each point is again the Levenshtein distance.

Lastly, we compare the running times between pFind, FragPipe and MaxQuant. It has already been reported that the 2 open search software run much faster than other narrow search ones (Kong et al., 2017; Chi, Liu, Yang, Zeng, Wu, Zhou, Niu, et al., 2018). Here, we also run a subset of the MaxQuant runs on the Cambridge HPC (Table 1) with the same amount of CPUs in order to compare with Fragpipe in the same system. MaxQuant and pFind were executed on the same system (Table 1). We found that pFind runs about 40 times faster than MaxQuant on MiniMax and FragPipe about 30 times on Cambridge HPC.

The abundance and range of PTMs is likely to be limited in this study, given the low complexity of the sample. For example, there is no lactose that would generate lactosylations (Czerwenka et al., 2006; Bosman et al., 2021). However, on a real, complex, archaeological or palaeontological ancient sample, we can expect a plethora of post-translational modifications, from the ones occurring *in vivo,* to those induced by human activities, bacteria and time. A simple narrow search,

as seen here and as frequently reported in the literature and by researchers, leaves massive, and potentially very informative, amounts of data unexplored. As we will comment in further detail, we recommend including open search software in the analysis workflow in Palaeoproteomics; at least for PTM discovery to augment narrow searches, similarly to what Metamorpheus G-PTM-D does (Solntsev et al., 2018).

De novo sequencing coverage bias and accuracy

We have seen that when we use external knowledge and data to search the spectra, we are always required to make choices that will affect the number, quality and significance of the identifications. Open search software can search for unspecified PTMs. But *de novo* sequencing (*dnS*) software can in theory reconstruct the peptide sequence just from the spectra without the use of an external database. However, sequencing errors might arise (Muth & Renard, 2018; Beslic et al., 2023), which, without external validating data, can have a high impact in the discovery of potential novel peptides from extinct or unsequenced organisms and phylogenetic studies. In this section we look at the interplay between degradation, sequence coverage bias and accuracy of *dnS*. We also discuss how the choice of the database affects the accuracy of the results in this case, since DirecTag and PepNovo+ rely on it.

Since we have a very simple system, with a single and known protein, it is easy to map the *de novo* sequenced peptides back to the BLG to obtain sequence coverage and also identify sequence errors. We follow the procedure explained in Mapping of de novo peptides and Figure 2. Finally, peptides with less than 70% overall identity with the BLG sequence are discarded. We found that overall, alignments coming from C-terminal seed matches have a greater identity than those from N-terminal k-mers, when none of them are given priority (Figure S3). This is in line with previous studies reporting that long *b-ions* are unstable and therefore *y-ions* are preferentially generated in HCD (Shao et al., 2014; Medzihradszky & Chalkley, 2015). Therefore, HCD produces better ion fragment coverage on the C-terminal and in turn less sequencing errors.

As with *DBs*, we studied the effect of the size of the sequence database for *de novo* methods that use it at any point of the pipeline, like DirecTag and PepNovo+. When the larger DB2 is used in DeNovoGUI, DirecTag and PepNovo+ attempt at covering larger parts of the BLG sequence, but at the cost of decreasing the overall accuracy of the peptides (Figure 7A, B, C and Supplementary Figure S2). In contrast, Novor achieves almost identical results (Figure 7A and Supplementary Figure S2). Within DeNovoGUI, DirecTag and PepNovo+ rely on an external database, either for a tagging hybrid method or for a subsequent MS-BLAST, respectively. We can draw a parallel observation as when the FDR is inadvertently boosted in *DBs* if the database contains many irrelevant sequences. In this case, incorrectly *de novo* sequenced peptides or tags are queried against a plethora of sequences from DB2, resulting in spurious matches that end up validating or

wrongly extending them. Novor does not use a database at any point, but the overall amino acid accuracy is always lower (Figure 7A and Supplementary Figure S2). Beslic et al. (2023) also reported Novor performed worse than other *dnS* software in terms of amino acid and peptide recall.

To study sequence accuracy along the BLG we take every BLG-assigned sequenced peptide position and 1) count how many times it is observed and 2) identify if it is correct or not in order to calculate the accuracy at each position. Incorrect amino acids produce a negative count (Figure 7 B, C). For Novor, inaccuracies accumulate with the more degraded samples (Figure 7A). In contrast, DirecTag and PepNovo+ decrease their coverage, but inaccuracies only occur when using the DB2. These inaccuracies are not spread evenly across the BLG, but are more pronounced in specific areas (Figure 7A, B, C). Moreover, regions presenting good coverage, between positions 80 and 90 or around 160, also present the highest number of wrong observations (Figure 7 C)

In line with the database effect, DirecTag and PepNovo+ are more conservative and overall present less hits (150 for the best position), because sequencing errors are prevented and wrong peptides discarded upon matching against a database. Novor, in contrast, attempts at sequencing everything and has improved coverage but returns more erroneous peptides.

De novo sequencing is a critical tool in palaeoproteomic studies looking at the phylogenetic relations of extinct animals (Welker et al., 2015, 2019; Chen et al., 2019; Presslee et al., 2019; Cappellini et al., 2019). Crucially, these studies use PEAKS, MaxQuant and BLAST, in sequential workflows tailored to the identification and validation of single amino acid polymorphisms and novel peptides. The combination of hybrid methods and *dnS* and *DBs* reflects the compromise between the discovery of novel sequences and the minimization of sequencing errors. As seen in Figure 7, sequencing errors can accumulate in highly degraded proteins and regions. Software that does not rely on any external database is more affected by this. At the same time, a database can be a curse, rather than a blessing, if it contains too many irrelevant sequences, like DB2. Therefore, these studies use databases restricted to a specific species, or type of protein like collagen or tissue, like bone or enamel proteomes.

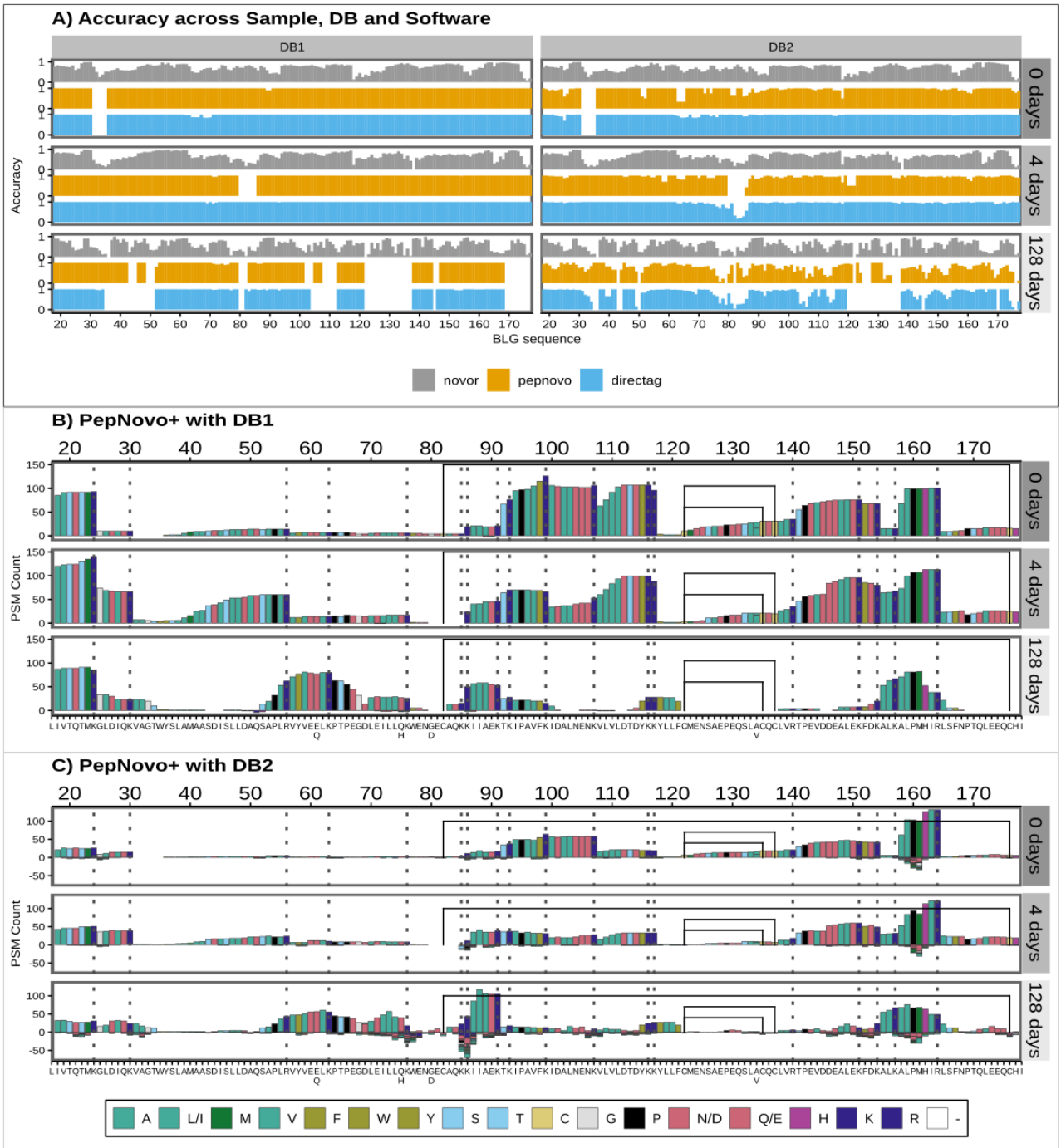


Figure 7. Positional sequence coverage and amino acid accuracy. Coverage is measured in the y-axes as the number of times an amino acid position is covered by an identified peptide, while accuracy is the proportion of those that are correctly identified. The BLG signal peptide is not shown, but positions are given counting it, and therefore start at 17, just to highlight that it would not be present in the mature BLG found in milk. Panel A) shows the amino acid accuracy for the 3 *de novo* tools using either the dairy database (DB1) or the bovine proteome (DB2). Panels B and C show the coverage for PepNovo+ using DB1 and DB2 databases. Values below 0 correspond to mismatches between the peptide sequence and the actual BLG sequence. In panels A and B, the grey dotted lines mark the tryptic sites and the black lines are the disulfide bonds between Cysteines in positions 82 and 176, 122 and 135 and the alternative 137. Amino acids are coloured by main properties: hydrophobic ones (A, L/I, M, V) are in teal green, with an olive shade for the aromatic ones (F, W; Y). Acidic amino acids (D and E) are in rose (or their amide group counterparts N and D, as they are not differentiated here), while basic ones are in dark blue (K and R) or purple (H). Other hydrophilic amino acids (S and T) are in cyan. Cysteines (C) are highlighted in yellow, glycine (G) in grey and proline (P) black. Colour scheme source: https://personal.sron.nl/~pault/

Peptide chemical properties and coverage bias

In the previous section, we reported a coverage bias for *dnS*, which also changes as degradation proceeds. Figure 8 shows the coverage for pFind tryptic search (A) and MaxQuant with tryptic (B) and semi-tryptic (C) search with the dairy specific DB1. There are sections of the sequence and peptides with a much greater support than others, meaning they were identified in more different PSMs. In the 128 days sample, this coverage bias dramatically changes and some regions become “invisible”, while others increase their relative coverage. In line with the findings of previous sections, pFind tryptic search and MaxQuant non-specific search coverage of the 128-days sample is similar (Figure 8A and C), while MaxQuant’s tryptic search fails to identify broad sections of the BLG (Figure 8B). We explored whether structural flexibility, relative solvent accessibility (RSA), amyloidogenicity and isoelectric point along BLG protein sequence and 3D structure correlate with these patterns. For this we extracted the B-factors and relative solvent accessibility from the 7KOT 3D structure from PDB. We estimated the amyloid formation propensity using different algorithms and predictors and calculated the isoelectric point (Figure 9).

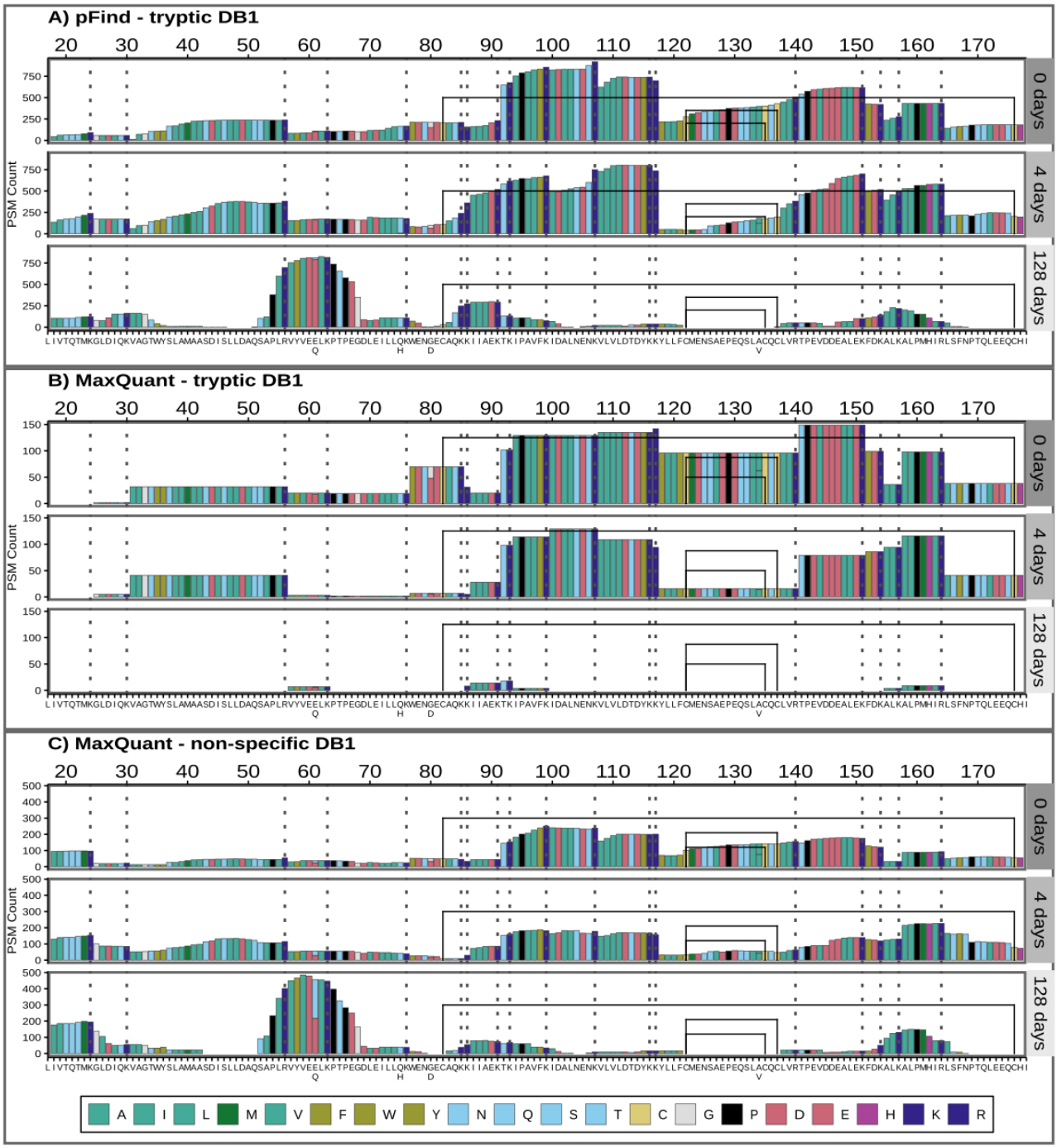


Figure 8. Sequence coverage for pFind tryptic search (A) and MaxQuant with tryptic and non-specific search (B and C). The BLG signal peptide is not shown, but positions are given counting it, and therefore start at 17, just to highlight that it would not be present in the mature BLG found in milk. PSMs below 0.05 FDR are included. Amino acids are coloured by main properties: in teal green hydrophobic ones (A, I, L, M, V), with an olive shade for the aromatic ones (F, W, Y). Acidic amino acids (D and E) are in rose, while basic ones are in dark blue (K and R) or purple (H). Other hydrophilic amino acids (N, Q, S and T) are in cyan. Cysteines (C) are highlighted in yellow, glycine (G) in grey and proline (P) black. Colour scheme source: https://personal.sron.nl/~pault/

The region between positions 85 and 117 has the greatest coverage in the non-degraded (0 days) and moderately degraded (4 days) samples. It coincides with a region with a high isoelectric point, several exposed tryptic sites and low probability of amyloid aggregate formation. However, after extensive degradation its coverage greatly decreases and some highly flexible sections (97 to 117)

are barely identified. In contrast, the relative coverage of the region between positions 53 and 67 increases only after extensive degradation and using non-specific search. The only 2 tryptic sites here are relatively not as exposed and the isoelectric point is low.

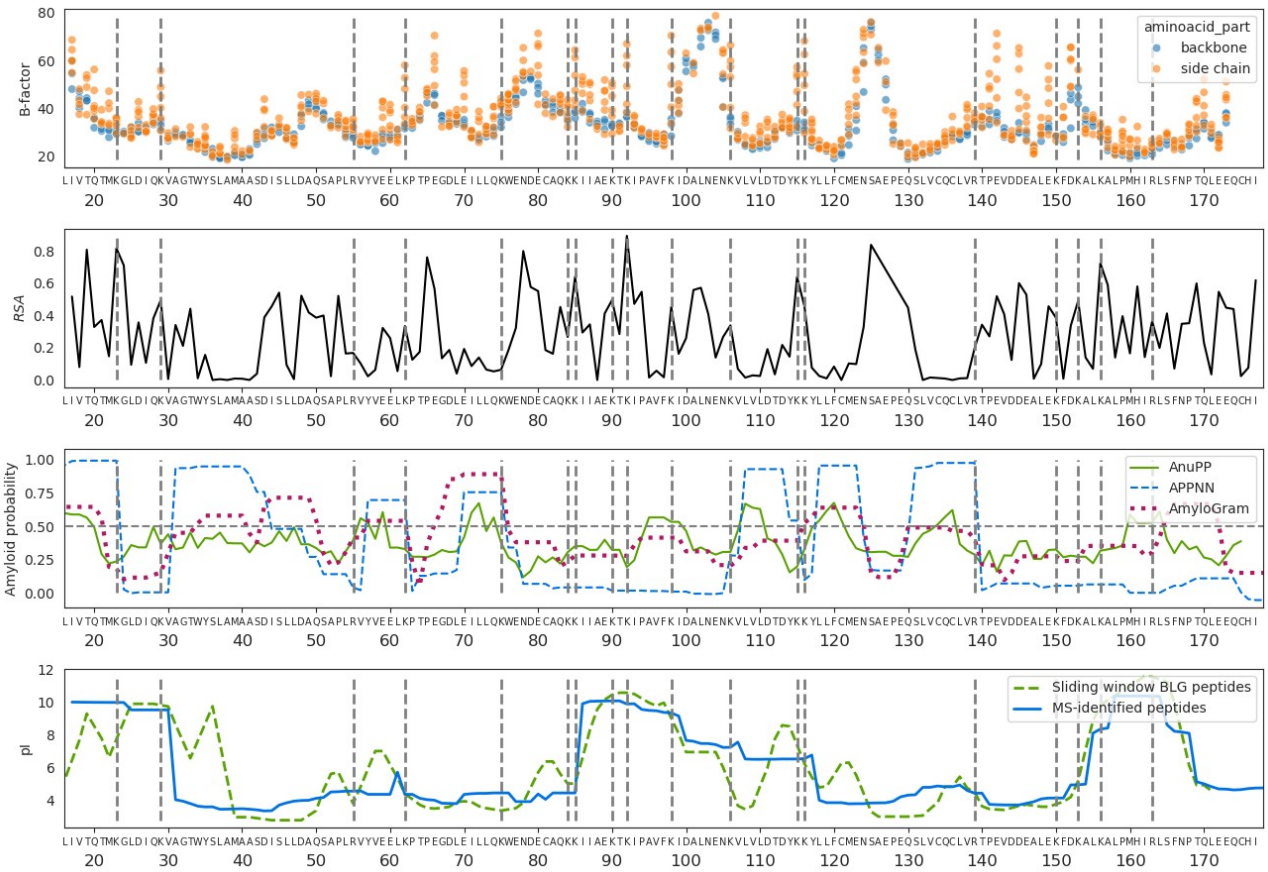


Figure 9. Physico-chemical properties along the beta-lactoglobulin. The top graph are the B-factors of all atoms except hydrogens. The carboxylic and alpha carbons and nitrogens from the amino group along the backbone are in blue, while other side chain atoms are in orange. The second graph shows the relative solvent accessibility of the BLG folded structure. The third panel is the positional amyloid formation propensity estimated by AnuPP (Prabakaran et al., 2021), ANNPP (Família et al., 2015) and AmyloGram (Burdukiewicz et al., 2017) software. The last graph is the isoelectric point. The tryptic sites (lysines and arginines) are marked as grey dashed horizontal lines.

The region between positions 105 and 140, features compact and rigid beta-sheets connected by flexible loops in the 3D structure, with high propensity of forming amyloid structures, it is negatively charged (low isoelectric point) and lacks tryptic sites (Figure 9). This region's coverage is overall reduced from the 0 to 4 days samples and then is not observed in the 128 days sample.

It is therefore a combination of processes and mechanisms that will determine whether a protein or peptide will be successfully extracted, detected in the mass spectrometer and identified (Evans et al., 2024). First there needs to be well exposed tryptic sites (if using trypsin and tryptic or semi-tryptic search), otherwise, those regions can only appear after denaturalisation and hydrolysis of the backbone (128-days sample) (Evans et al., 2024), and then only observed using semitryptic or non-specific search. However, peptides can start forming aggregates after they are enzymatically

cleaved or hydrolysed (4 days), and highly flexible, polar and charged regions are also more vulnerable to being hydrolyzed by water (Fonseca et al., 2024). At the same time, other properties might protect peptides from hydrolysis and degradation. The peptide T_{141}PEVDDEALEK_{151} is one of the most encountered BLG peptides in the literature (Wilkin et al., 2020, 2021); (Bleasdale et al., 2021). It presents highly flexible and negatively charged side-chains (Figure 9). Here, in contrast, it is hardly identified after extensive degradation. This suggests that the real conditions in which the peptide is found, mainly the presence of a mineral surface, contributes to its preservation, as it has been suggested using molecular simulations (Hagiwara et al., 2009; Fonseca et al., 2022). The mineral surface and other characteristics of a real sample also determines the extent to which proteins are successfully extracted in the laboratory, which is not discussed here.

The last main factor determining the detection of peptides is the ionisation efficiency by the electrospray. Peptides with high isoelectric point are already positively charged and therefore easier to be detected when running in the electrospray in positive ion mode. Negatively charged peptides, with low pI, will depend entirely on ionisation efficiencies at the electrospray (Liigand et al., 2017).

Towards an integrated Paleoproteomics pipeline

Results can vary greatly between MS/MS identification strategies, even for such a simple system as the one explored here. Moreover, there are challenges that each method cannot single-handedly solve. To tackle this, we devise an standardised approach that combines the powers of the 3 methods discussed (*de novo,* open and narrow search), suited for the study of ancient proteins. This has already been done in other more informal and *ad-hoc* approaches that combine *de novo* sequencing and database search (Welker et al., 2015, 2019; Chen et al., 2019; Presslee et al., 2019; Cappellini et al., 2019). Meta-search engines like DeNovoGUI (Muth et al., 2014), SearchGUI (Barsnes & Vaudel, 2018) or platforms like PeptideShaker (Vaudel et al., 2015) focus on aggregating results from different software. This is important to catch disputed sequences or give more reliability to the consensus ones.

The pipeline would integrate the sequence discovery of *de novo* sequencing with the PTM discovery of open search, and a process to narrow down sequence databases. These can be used to (1) refine a protein sequence database and (2) expand the PTMs to search at a final narrow search (Figure 10) that would be more reliable. At the end, peptide identifications by different software can be integrated to increase the confidence of the results.

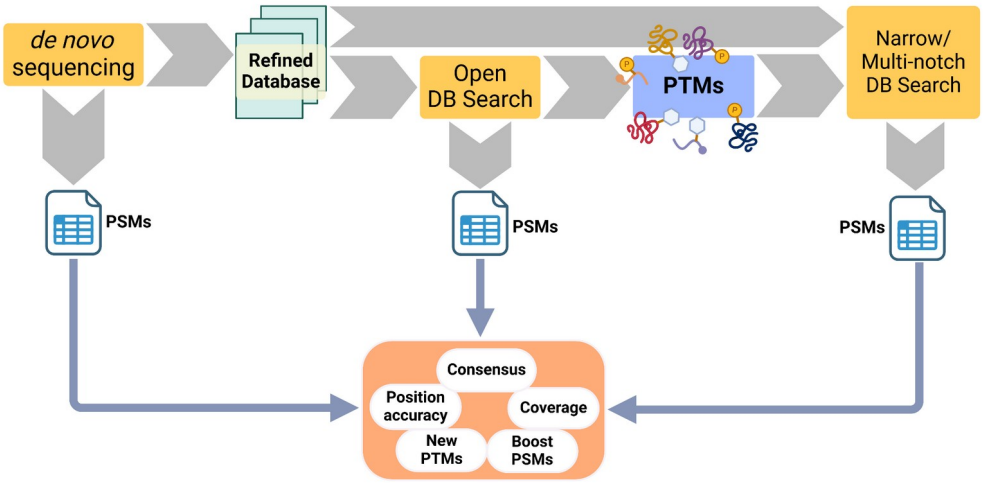


Figure 10. Proposed pipeline for the analysis of ancient proteins. It starts with a *de novo* sequencing step that refines or produces a targeted database. This database is then used for an open search and the novel, abundant or relevant PTMs are extracted. The refined database and the PTMs are then fed into a closed or a multi-notch database search. At the end the PSMs of the 3 software runs are combined, to check for consensus or disputed sequences, study position accuracy and coverage. This would increase the confidence of the results, while producing more PSMs than each of the steps alone, and identify PTMs reliably.

Moreover, we believe it is useful to study sequence coverage and positional information to identify commonly identified regions and inaccuracies. Finally, the identification of a wider array of PTMs can enrich the information that we can gather from ancient proteins in archaeological and palaeontological samples.

Conclusion

Palaeoproteomics is a maturing discipline that is developing new strategies to overcome its own challenges. Teams working on computational proteomics are constantly releasing new software and updates with new features. These implement a wide variety of underlying methods and algorithms that we broadly group in *de novo* sequencing or database search. Palaeoproteomics frequently borrows methodology and software from the broader field of proteomics. Yet there are unique challenges to applying these tools to palaeoproteomics, such as the difficulties associated with selecting a properly constrained database or accounting and identifying PTMs. It is therefore important to assess the pertinence and validity of the different strategies and parameters in the study of ancient proteins. Here we have done so by using a simple and progressively degraded protein model, β-lactoglobulin, heated at pH 7 for 0, 4 and 128 days.

We highlight the importance of constraining the search space in *DBs* to improve the significance and reduce computing times. Setting the enzymatic digestion according to the level of degradation and using targeted databases are the most important factors. At the same time, Open search software can identify more peptides with PTMs that would have been otherwise unidentified. Moreover, the study of new PTMs, like glycations, can open the door to answering new questions

in archaeology. *De novo* sequencing goes one step further in freeing researchers from database selection, which we have shown to hugely impact the results. However, the possibility of sequencing mistakes cannot be disregarded and in practice, external databases and complementary *DBs* are commonly used. Recently, new *dnS* programs like Casanovo (Yilmaz et al., 2022), InstaNovo (Eloff et al., 2023), or π-HelixNovo (Yang et al., 2023) are reporting promising performance. They use transformer neural network architecture, (as used in Large Language Models), which seem suited to protein sequences as they can be seen as another language. A language that is revealed in spectral data, in which the presence or absence of peaks and their intensities are intercorrelated.

As it has already been done in more informal and *ad-hoc* approaches, we propose a standardised approach that combines the powers of the different approaches discussed here. *De novo* sequencing software, especially those using transformers, and Open search can be used to discover novel sequences and PTMs and initially constrain the search space of possible protein contents and modifications in a complex sample. Using this information, a final, constrained *narrow database search* with targeted database and PTMs would boost identifications confidently.

Data and code availability

Python and R notebooks used to process the *DBs* and *dn*S data can be found in https://github.com/ismaRP/PPbenchmark.git and 10.5281/zenodo.13785336. RAW LC-MSMS and PSM data resulting from the analysis with Fragpipe, pFind, Metamorpheus, MaxQuant, Mascot and DeNovoGUI can be accessed in Zenodo 10.5281/zenodo.13785293. MaxQuant and Fragpipe parameters and workflow files are also included.

Acknowledgements

IRP is currently funded by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 956410. At the time of producing and writing this work BN was funded by European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No. 801199 and with MC by the European Union's EU Framework Programme for Research and Innovation Horizon 2020 under Grant Agreement No. 787282 (B2C). MC and MM were also supported by the Danish National Research Foundation (DNRF128). YC and JD are funded by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No. 956351.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

References

Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207. https://doi.org/10.1038/nature01511

Alfaro JA, Ignatchenko A, Ignatchenko V, Sinha A, Boutros PC, Kislinger T (2017) Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines. *Genome medicine*, **9**, 62. https://doi.org/10.1186/s13073-017-0454-9

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bachmann M (2024) *Levenshtein*.

Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research*, **28**, 45–48. https://doi.org/10.1093/nar/28.1.45

Barsnes H, Vaudel M (2018) SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *Journal of proteome research*, **17**, 2552–2555. https://doi.org/10.1021/acs.jproteome.8b00175

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, **57**, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Beslic D, Tscheuschner G, Renard BY, Weller MG, Muth T (2023) Comprehensive evaluation of peptide de novo sequencing tools for monoclonal antibody assembly. *Briefings in bioinformatics*, **24**, bbac542. https://doi.org/10.1093/bib/bbac542

Bittremieux W, May DH, Bilmes J, Noble WS (2022) A learned embedding for efficient joint analysis of millions of mass spectra. *Nature methods*, **19**, 675–678. https://doi.org/10.1038/s41592-022-01496-1

Bleasdale M, Richter KK, Janzen A, Brown S, Scott A, Zech J, Wilkin S, Wang K, Schiffels S, Desideri J, Besse M, Reinold J, Saad M, Babiker H, Power RC, Ndiema E, Ogola C, Manthi FK, Zahir M, Petraglia M, Trachsel C, Nanni P, Grossmann J, Hendy J, Crowther A, Roberts P, Goldstein ST, Boivin N (2021) Ancient proteins provide evidence of dairy consumption in eastern Africa. *Nature communications*, **12**, 632. https://doi.org/10.1038/s41467-020-20682-3

Bosman GP, Oliveira S, Simons PJ, Sastre Torano J, Somsen GW, Knippels LMJ, Haselberg R, Pieters RJ, Garssen J, Knipping K (2021) Limited lactosylation of beta-lactoglobulin from cow's milk exerts strong influence on antigenicity and degranulation of mast cells. *Nutrients*, **13**, 2041.

https://doi.org/10.3390/nu13062041

Burdukiewicz M, Sobczyk P, Rödiger S, Duda-Madej A, Mackiewicz P, Kotulska M (2017) Amyloidogenic motifs revealed by n-gram analysis. *Scientific reports*, **7**, 12961. https://doi.org/10.1038/s41598-017-13210-9

Cappellini E, Welker F, Pandolfi L, Ramos-Madrigal J, Samodova D, Rüther PL, Fotakis AK, Lyon D, Moreno-Mayar JV, Bukhsianidze M, Rakownikow Jersie-Christensen R, Mackie M, Ginolhac A, Ferring R, Tappen M, Palkopoulou E, Dickinson MR, Stafford TW Jr, Chan YL, Götherström A, Nathan SKSS, Heintzman PD, Kapp JD, Kirillova I, Moodley Y, Agusti J, Kahlke R-D, Kiladze G, Martínez-Navarro B, Liu S, Sandoval Velasco M, Sinding M-HS, Kelstrup CD, Allentoft ME, Orlando L, Penkman K, Shapiro B, Rook L, Dalén L, Gilbert MTP, Olsen JV, Lordkipanidze D, Willerslev E (2019) Early Pleistocene enamel proteome from Dmanisi resolves Stephanorhinus phylogeny. *Nature*. https://doi.org/10.1038/s41586-019-1555-y

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology*, **30**, 918–920. https://doi.org/10.1038/nbt.2377

Chang H-Y, Kong AT, da Veiga Leprevost F, Avtonomov DM, Haynes SE, Nesvizhskii AI (2020) Crystal-C: A Computational Tool for Refinement of Open Search Results. *Journal of proteome research*, **19**, 2511–2515. https://doi.org/10.1021/acs.jproteome.0c00119

Charlton S, Ramsøe A, Collins M, Craig OE, Fischer R, Alexander M, Speller CF (2019) New insights into Neolithic milk consumption through proteomic analysis of dental calculus. *Archaeological and anthropological sciences*. https://doi.org/10.1007/s12520-019-00911-7

Chen F, Welker F, Shen C-C, Bailey SE, Bergmann I, Davis S, Xia H, Wang H, Fischer R, Freidline SE, Yu T-L, Skinner MM, Stelzer S, Dong G, Fu Q, Dong G, Wang J, Zhang D, Hublin J-J (2019) A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature*, **569**, 409–412. https://doi.org/10.1038/s41586-019-1139-x

Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, Gygi SP (2015) A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified

peptides. *Nature biotechnology*, **33**, 743–749. https://doi.org/10.1038/nbt.3267

Chi H, Liu C, Yang H, Zeng W-F, Wu L, Zhou W-J, Niu X-N, Ding Y-H, Zhang Y, Wang R-M, Wang Z-W, Chen Z-L, Sun R-X, Liu T, Tan G-M, Dong M-Q, Xu P, Zhang P-H, He S-M (2018) Open-pFind enables precise, comprehensive and rapid peptide identification in shotgun proteomics. *bioRxiv*, 285395. https://doi.org/10.1101/285395

Chi H, Liu C, Yang H, Zeng W-F, Wu L, Zhou W-J, Wang R-M, Niu X-N, Ding Y-H, Zhang Y, Wang Z-W, Chen Z-L, Sun R-X, Liu T, Tan G-M, Dong M-Q, Xu P, Zhang P-H, He S-M (2018) Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nature biotechnology*, **36**, 1059–1061. https://doi.org/10.1038/nbt.4236

Choi H, Nesvizhskii AI (2008) False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *Journal of proteome research*, **7**, 47–50. https://doi.org/10.1021/pr700747q

Couté Y, Bruley C, Burger T (2020) Beyond Target-Decoy Competition: Stable Validation of Peptide and Protein Identifications in Mass Spectrometry-Based Discovery Proteomics. *Analytical chemistry*, **92**, 14898–14906. https://doi.org/10.1021/acs.analchem.0c00328

Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, **26**, 1367–1372. https://doi.org/10.1038/nbt.1511

Creasy DM, Cottrell JS (2004) Unimod: Protein modifications for mass spectrometry. *Proteomics*, **4**, 1534–1536. https://doi.org/10.1002/pmic.200300744

Czerwenka C, Maier I, Pittner F, Lindner W (2006) Investigation of the lactosylation of whey proteins by liquid chromatography-mass spectrometry. *Journal of agricultural and food chemistry*, **54**, 8874–8882. https://doi.org/10.1021/jf061646z

Demarchi B, Hall S, Roncal-Herrero T, Freeman CL, Woolley J, Crisp MK, Wilson J, Fotakis A, Fischer R, Kessler BM, Rakownikow Jersie-Christensen R, Olsen JV, Haile J, Thomas J, Marean CW, Parkington J, Presslee S, Lee-Thorp J, Ditchfield P, Hamilton JF, Ward MW, Wang CM, Shaw MD, Harrison T, Domínguez-Rodrigo M, MacPhee RDE, Kwekason A, Ecker M, Kolska Horwitz L, Chazan M, Kröger R, Thomas-Oates J, Harding JH, Cappellini E, Penkman K, Collins MJ (2016) Protein sequences bound to mineral surfaces persist into deep time. *eLife*, **5**. https://doi.org/10.7554/eLife.17092

Ebadi A, Freestone J, Noble WS, Keich U (2023) Bridging the False Discovery Gap. *Journal of proteome research*, **22**, 2172–2178. https://doi.org/10.1021/acs.jproteome.3c00176

Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein

identifications by mass spectrometry. *Nature methods*, **4**, 207–214. https://doi.org/10.1038/nmeth1019

Eloff K, Kalogeropoulos K, Morell O, Mabona A, Jespersen JB, Williams W, van Beljouw SPB, Skwark M, Laustsen AH, Brouns SJJ, Ljungars A, Schoof EM, Van Goey J, Keller U auf D, Beguir K, Carranza NL, Jenkins TP (2023) De novo peptide sequencing with InstaNovo: Accurate, database-free peptide identification for large scale proteomics experiments. *bioRxiv*, 2023.08.30.555055. https://doi.org/10.1101/2023.08.30.555055

Evans M, Hagan R, Boyd O, Bondetti M, Craig O, Collins M, Hendy J (2024) The impact of cooking and burial on proteins: a characterisation of experimental foodcrusts and ceramics. *bioRxiv*, 2024.04.03.587902. https://doi.org/10.1101/2024.04.03.587902

Evans M, Lundy J, Lucquin A, Hagan R, Kowalski Ł, Wilczyńki J, Bickle P, Adamczak K, Craig OE, Robson HK, Hendy J (2023) Detection of dairy products from multiple taxa in Late Neolithic pottery from Poland: an integrated biomolecular approach. *Royal Society open science*, **10**, 230124. https://doi.org/10.1098/rsos.230124

Família C, Dennison SR, Quintas A, Phoenix DA (2015) Prediction of peptide and protein propensity for amyloid formation. *PloS one*, **10**, e0134679. https://doi.org/10.1371/journal.pone.0134679

Fonseca B, Freeman CL, Collins MJ (2022) Conformational analysis and water dynamics: a molecular dynamics study on the survival of a β-lactoglobulin peptide in the archaeological record. *Chemical physics*, **561**, 111602. https://doi.org/10.1016/j.chemphys.2022.111602

Fonseca B, Freeman CL, Collins MJ (2024) The Survival of β-lactoglobulin Peptides in the Archaeological Record: Vulnerability vs. Sequence Variation. *bioRxiv*, 2024.09.13.612646. https://doi.org/10.1101/2024.09.13.612646

Frank AM (2009a) A ranking-based scoring function for peptide-spectrum matches. *Journal of proteome research*, **8**, 2241–2252. https://doi.org/10.1021/pr800678b

Frank AM (2009b) Predicting intensity ranks of peptide fragment ions. *Journal of proteome research*, **8**, 2226–2240. https://doi.org/10.1021/pr800677f

Frank A, Pevzner P (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry*, **77**, 964–973. https://doi.org/10.1021/ac048788h

Freestone J, Noble WS, Keich U (2024) Reinvestigating the correctness of decoy-based false discovery rate control in proteomics tandem mass spectrometry. *Journal of proteome research*, **23**, 1907–1914. https://doi.org/10.1021/acs.jproteome.3c00902

Geiszler DJ, Kong AT, Avtonomov DM, Yu F, Leprevost F da V, Nesvizhskii AI (2021) PTM-Shepherd:

Analysis and Summarization of Post-Translational and Chemical Modifications From Open Search Results. *Molecular & cellular proteomics: MCP*, **20**, 100018. https://doi.org/10.1074/mcp.TIR120.002216

Griss J, Perez-Riverol Y, Lewis S, Tabb DL, Dianes JA, Del-Toro N, Rurik M, Walzer MW, Kohlbacher O, Hermjakob H, Wang R, Vizcaíno JA (2016) Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nature methods*, **13**, 651–656. https://doi.org/10.1038/nmeth.3902

Hagiwara T, Sakiyama T, Watanabe H (2009) Molecular simulation of bovine beta-lactoglobulin adsorbed onto a positively charged solid surface. *Langmuir: the ACS journal of surfaces and colloids*, **25**, 226–234. https://doi.org/10.1021/la8024149

He K, Fu Y, Zeng W-F, Luo L, Chi H, Liu C, Qing L-Y, Sun R-X, He S-M (2015) A theoretical foundation of the target-decoy search strategy for false discovery rate control in proteomics. *arXiv [stat.AP]*.

Hendy J, Colonese AC, Franz I, Fernandes R, Fischer R, Orton D, Lucquin A, Spindler L, Anvari J, Stroud E, Biehl PF, Speller C, Boivin N, Mackie M, Jersie-Christensen RR, Olsen JV, Collins MJ, Craig OE, Rosenstock E (2018) Ancient proteins from ceramic vessels at Çatalhöyük West reveal the hidden cuisine of early farmers. *Nature communications*, **9**, 4064. https://doi.org/10.1038/s41467-018-06335-6

Hendy J, Warinner C, Bouwman A, Collins MJ, Fiddyment S, Fischer R, Hagan R, Hofman CA, Holst M, Chaves E, Klaus L, Larson G, Mackie M, McGrath K, Mundorff AZ, Radini A, Rao H, Trachsel C, Velsko IM, Speller CF (2018) Proteomic evidence of dietary sources in ancient dental calculus. *Proceedings. Biological sciences / The Royal Society*, **285**. https://doi.org/10.1098/rspb.2018.0977

Hendy J, Welker F, Demarchi B, Speller C, Warinner C, Collins MJ (2018) A guide to ancient protein studies. *Nature ecology & evolution*, **2**, 791–799. https://doi.org/10.1038/s41559-018-0510-x

Jeong K, Kim S, Bandeira N (2012) False discovery rates in spectral identification. *BMC bioinformatics*, **13 Suppl 16**, S2. https://doi.org/10.1186/1471-2105-13-S16-S2

Jeong C, Wilkin S, Amgalantugs T, Bouwman AS, Taylor WTT, Hagan RW, Bromage S, Tsolmon S, Trachsel C, Grossmann J, Littleton J, Makarewicz CA, Krigbaum J, Burri M, Scott A, Davaasambuu G, Wright J, Irmer F, Myagmar E, Boivin N, Robbeets M, Rühli FJ, Krause J, Frohlich B, Hendy J, Warinner C (2018) Bronze Age population dynamics and the rise of dairy pastoralism on the eastern Eurasian steppe. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, E11248–E11255. https://doi.org/10.1073/pnas.1813608115

Jersie-Christensen RR, Lanigan LT, Lyon D, Mackie M, Belstrøm D, Kelstrup CD, Fotakis AK, Willerslev E, Lynnerup N, Jensen LJ, Cappellini E, Olsen JV (2018) Quantitative metaproteomics of medieval dental

calculus reveals individual oral health status. *Nature communications*, **9**, 4744. https://doi.org/10.1038/s41467-018-07148-3

Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend G (2011) A series of PDB related databases for everyday needs. *Nucleic acids research*, **39**, D411–9. https://doi.org/10.1093/nar/gkq1105

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637. https://doi.org/10.1002/bip.360221211

Käll L, Storey JD, MacCoss MJ, Noble WS (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research*, **7**, 29–34. https://doi.org/10.1021/pr700600n

Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry*, **74**, 5383–5392. https://doi.org/10.1021/ac025747h

Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature methods*, **14**, 513–520. https://doi.org/10.1038/nmeth.4256

Levitsky LI, Klein JA, Ivanov MV, Gorshkov MV (2019) Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *Journal of proteome research*, **18**, 709–714. https://doi.org/10.1021/acs.jproteome.8b00717

Liigand J, Laaniste A, Kruve A (2017) PH effects on electrospray ionization efficiency. *Journal of the American Society for Mass Spectrometry*, **28**, 461–469. https://doi.org/10.1007/s13361-016-1563-1

Liu X, Jin J, Yuan Y, Xian W, Tang Z, Fu J (2023) The ever increasing necessity of mass spectrometry in dissecting protein post-translational modifications catalyzed by bacterial effectors.

Ma B (2015) Novor: real-time peptide de novo sequencing software. *Journal of the American Society for Mass Spectrometry*, **26**, 1885–1894. https://doi.org/10.1007/s13361-015-1204-0

Madej D, Lam H (2023) Modeling Lower-Order Statistics to Enable Decoy-Free FDR Estimation in Proteomics. *Journal of proteome research*, **22**, 1159–1171. https://doi.org/10.1021/acs.jproteome.2c00604

Medzihradszky KF, Chalkley RJ (2015) Lessons in de novo peptide sequencing by tandem mass spectrometry. *Mass spectrometry reviews*, **34**, 43–63. https://doi.org/10.1002/mas.21406

Moore RE, Young MK, Lee TD (2002) Qscore: an algorithm for evaluating SEQUEST database search

results. *Journal of the American Society for Mass Spectrometry*, **13**, 378–386. https://doi.org/10.1016/S1044-0305(02)00352-5

Muth T, Renard BY (2018) Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Briefings in bioinformatics*, **19**, 954–970. https://doi.org/10.1093/bib/bbx033

Muth T, Weilnböck L, Rapp E, Huber CG, Martens L, Vaudel M, Barsnes H (2014) DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. *Journal of proteome research*, **13**, 1143–1146. https://doi.org/10.1021/pr4008078

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**, 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Noble WS (2015) Mass spectrometrists should search only for peptides they care about. *Nature methods*, **12**, 605–608. https://doi.org/10.1038/nmeth.3450

Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu DJ, Prakash A, Frericks-Zipper A, Eisenacher M, Walzer M, Wang S, Brazma A, Vizcaíno JA (2021) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic acids research*, **50**, D543–D552. https://doi.org/10.1093/nar/gkab1038

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567. https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

Prabakaran R, Rawat P, Kumar S, Michael Gromiha M (2021) ANuPP: A versatile tool to predict aggregation nucleating regions in peptides and proteins. *Journal of molecular biology*, **433**, 166707. https://doi.org/10.1016/j.jmb.2020.11.006

Presslee S, Slater GJ, Pujos F, Forasiepi AM, Fischer R, Molloy K, Mackie M, Olsen JV, Kramarz A, Taglioretti M, Scaglia F, Lezcano M, Lanata JL, Southon J, Feranec R, Bloch J, Hajduk A, Martin FM, Salas Gismondi R, Reguero M, de Muizon C, Greenwood A, Chait BT, Penkman K, Collins M, MacPhee RDE (2019) Palaeoproteomics resolves sloth relationships. *Nature ecology & evolution*. https://doi.org/10.1038/s41559-019-0909-z

Ramsøe A, Crispin M, Mackie M, McGrath K, Fischer R, Demarchi B, Collins MJ, Hendy J, Speller C (2021) Assessing the degradation of ancient milk proteins through site-specific deamidation patterns. *Scientific reports*, **11**, 7795. https://doi.org/10.1038/s41598-021-87125-x

Ramsøe A, van Heekeren V, Ponce P, Fischer R, Barnes I, Speller C, Collins MJ (2020) DeamiDATE 1.0: Site-specific deamidation as a tool to assess authenticity of members of ancient proteomes. *Journal of archaeological science*, **115**, 105080. https://doi.org/10.1016/j.jas.2020.105080

Rappsilber J, Mann M, Ishihama Y (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature protocols*, **2**, 1896–1906. https://doi.org/10.1038/nprot.2007.261

Schroeter ER, Cleland TP (2016) Glutamine deamidation: an indicator of antiquity, or preservational quality? *Rapid communications in mass spectrometry: RCM*, **30**, 251–255. https://doi.org/10.1002/rcm.7445

Shao C, Zhang Y, Sun W (2014) Statistical characterization of HCD fragmentation patterns of tryptic peptides on an LTQ Orbitrap Velos mass spectrometer. *Journal of proteomics*, **109**, 26–37. https://doi.org/10.1016/j.jprot.2014.06.012

Shevchenko A, Sunyaev S, Loboda A, Shevchenko A, Bork P, Ens W, Standing KG (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Analytical chemistry*, **73**, 1917–1926. https://doi.org/10.1021/ac0013709

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

Sun Z, Liu Q, Qu G, Feng Y, Reetz MT (2019) Utility of B-factors in protein science: Interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chemical reviews*, **119**, 1626–1665. https://doi.org/10.1021/acs.chemrev.8b00290

Tabb DL, Ma Z-Q, Martin DB, Ham A-JL, Chambers MC (2008) DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *Journal of proteome research*, **7**, 3838–3846. https://doi.org/10.1021/pr800154p

Tanasi D, Cucina A, Cunsolo V, Saletti R, Di Francesco A, Greco E, Foti S (2021) Paleoproteomic profiling of organic residues on prehistoric pottery from Malta. *Amino acids*, **53**, 295–312. https://doi.org/10.1007/s00726-021-02946-4

The M, MacCoss MJ, Noble WS, Käll L (2016) Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *Journal of the American Society for Mass Spectrometry*, **27**, 1719–1727. https://doi.org/10.1007/s13361-016-1460-7

Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO (2013) Maximum allowed solvent accessibilites

of residues in proteins. *PloS one*, **8**, e80635. https://doi.org/10.1371/journal.pone.0080635

Tyanova S, Temu T, Cox J (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature protocols*, **11**, 2301–2319. https://doi.org/10.1038/nprot.2016.136

Vaudel M, Burkhart JM, Zahedi RP, Oveland E, Berven FS, Sickmann A, Martens L, Barsnes H (2015) PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature biotechnology*, **33**, 22–24. https://doi.org/10.1038/nbt.3109

da Veiga Leprevost F, Haynes SE, Avtonomov DM, Chang H-Y, Shanmugam AK, Mellacheruvu D, Kong AT, Nesvizhskii AI (2020) Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nature methods*, **17**, 869–870. https://doi.org/10.1038/s41592-020-0912-y

Wang S, Feng S, Pan C, Guo X (2022) FineFDR: Fine-grained Taxonomy-specific False Discovery Rates Control in Metaproteomics. In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* , pp. 287–292. IEEE. https://doi.org/10.1109/BIBM55620.2022.9995401

Warinner C, Korzow Richter K, Collins MJ (2022) Paleoproteomics. *Chemical reviews*, **122**, 13401–13446. https://doi.org/10.1021/acs.chemrev.1c00703

Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, Radini A, Hancock Y, Tito RY, Fiddyment S, Speller C, Hendy J, Charlton S, Luder HU, Salazar-García DC, Eppler E, Seiler R, Hansen LH, Castruita JAS, Barkow-Oesterreicher S, Teoh KY, Kelstrup CD, Olsen JV, Nanni P, Kawai T, Willerslev E, von Mering C, Lewis CM Jr, Collins MJ, Gilbert MTP, Rühli F, Cappellini E (2014) Pathogens and host immunity in the ancient human oral cavity. *Nature genetics*, **46**, 336–344. https://doi.org/10.1038/ng.2906

Welker F, Collins MJ, Thomas JA, Wadsley M, Brace S, Cappellini E, Turvey ST, Reguero M, Gelfo JN, Kramarz A, Burger J, Thomas-Oates J, Ashford DA, Ashton PD, Rowsell K, Porter DM, Kessler B, Fischer R, Baessmann C, Kaspar S, Olsen JV, Kiley P, Elliott JA, Kelstrup CD, Mullin V, Hofreiter M, Willerslev E, Hublin J-J, Orlando L, Barnes I, MacPhee RDE (2015) Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. *Nature*, **522**, 81–84. https://doi.org/10.1038/nature14249

Welker F, Ramos-Madrigal J, Kuhlwilm M, Liao W, Gutenbrunner P, de Manuel M, Samodova D, Mackie M, Allentoft ME, Bacon A-M, Collins MJ, Cox J, Lalueza-Fox C, Olsen JV, Demeter F, Wang W, Marques-Bonet T, Cappellini E (2019) Enamel proteome shows that Gigantopithecus was an early diverging pongine. *Nature*, **576**, 262–265. https://doi.org/10.1038/s41586-019-1728-8

Wenger CD, Coon JJ (2013) A proteomics search algorithm specifically designed for high-resolution tandem

mass spectra. *Journal of proteome research*, **12**, 1377–1386. https://doi.org/10.1021/pr301024c

Wilkin S, Miller AV, Taylor WTT, Miller BK, Hagan RW, Bleasdale M, Scott A, Gankhuyg S, Ramsøe A, Uliziibayar S, Trachsel C, Nanni P, Grossmann J, Orlando L, Horton M, Stockhammer PW, Myagmar E, Boivin N, Warinner C, Hendy J (2020) Dairy pastoralism sustained eastern Eurasian steppe populations for 5,000 years. *Nature Ecology & Evolution*, **4**, 346–355. https://doi.org/10.1038/s41559-020-1120-y

Wilkin S, Ventresca Miller A, Fernandes R, Spengler R, Taylor WT-T, Brown DR, Reich D, Kennett DJ, Culleton BJ, Kunz L, Fortes C, Kitova A, Kuznetsov P, Epimakhov A, Zaibert VF, Outram AK, Kitov E, Khokhlov A, Anthony D, Boivin N (2021) Dairying enabled Early Bronze Age Yamnaya steppe expansions. *Nature*, **598**, 629–633. https://doi.org/10.1038/s41586-021-03798-4

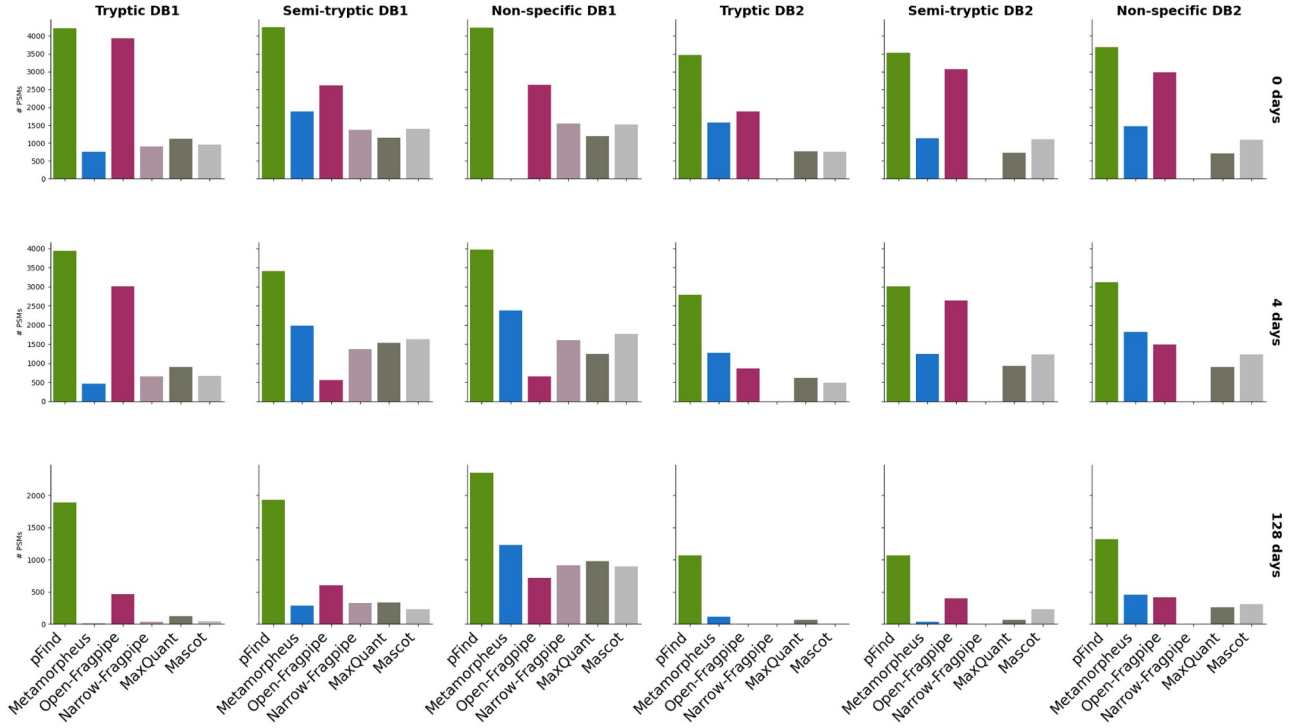
Yang T, Ling T, Sun B, Liang Z, Xu F, Huang X, Xie L, He Y, Li L, He F, Wang Y, Chang C (2023) Introducing π-HelixNovo for practical large-scale de novo peptide sequencing. *bioRxiv*, 2023.07.15.549133. https://doi.org/10.1101/2023.07.15.549133

Yilmaz M, Fondrie WE, Bittremieux W, Oh S, Noble WS (2022) *De novo* mass spectrometry peptide sequencing with a transformer model. *bioRxiv*, 2022.02.07.479481. https://doi.org/10.1101/2022.02.07.479481

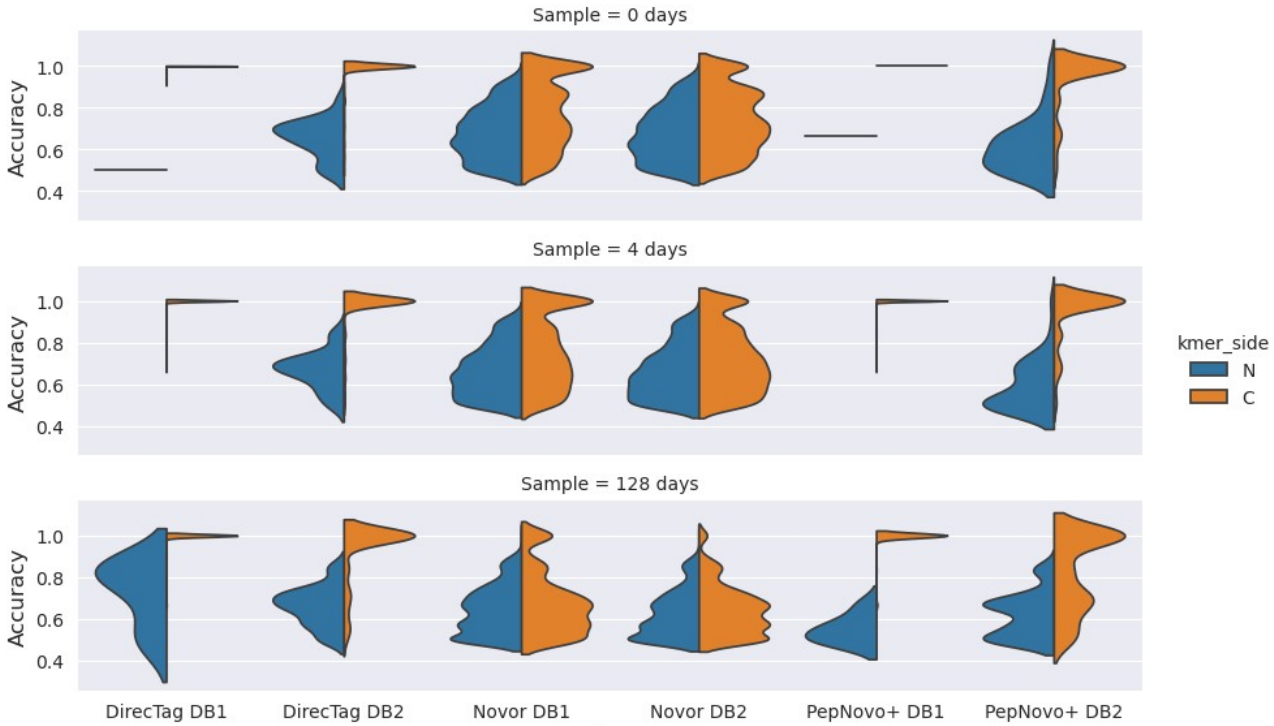
Yu F, Teo GC, Kong AT, Haynes SE, Avtonomov DM, Geiszler DJ, Nesvizhskii AI (2020) Identification of modified peptides using localization-aware open search. *Nature communications*, **11**, 4065. https://doi.org/10.1038/s41467-020-17921-y

Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B (2012) PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & cellular proteomics: MCP*, **11**, M111.010587. https://doi.org/10.1074/mcp.M111.010587

Supplementary Data



Supplementary Figure S1. Number of PSMs at FDR 0.01 for each database search software, enzyme cleavage and database.



Supplementary Figure S2. Accuracy distribution of *de novo* peptides aligned to the bovine β-lactoglobulin. In this plot, we chose the best alignment for each peptide, irrespective of whether it came from a N or C-terminal kmer seed alignment. We separate the accuracy of alignments coming from either a N-terminal (blue) or C-terminal (orange) seed k-mer.

Supplementary Table S1. Number of PSMs per run and software and at FDR 0.01 and 0.05								
Digestion	Database	Engine	FDR 0.01			FDR 0.05		
			0 days	4 days	128 days	0 days	4 days	128 days
Tryptic	DB1	Mascot	949	668	46	1076	753	49
		MaxQuant	1119	902	123	1119	902	123
		Fragpipe-narrow	891	648	31	917	725	40
		Metamorpheus	748	466	6	1209	802	104
		Fragpipe-open	3930	3004	464	4639	3949	605
		pFind-open	4206	3938	1887	4666	5018	2258
	DB2	Mascot	757	484	0	869	560	0
		MaxQuant	772	619	69	772	619	69
		Metamorpheus	1578	1273	115	1967	1558	304
		Fragpipe-open	1886	859	0	2790	1229	0
		pFind-open	3468	2785	1069	4088	3738	1257
Semi-tryptic	DB1	Mascot	1397	1625	231	1612	1836	280
		MaxQuant	1153	1528	338	1292	1738	447
		Fragpipe-narrow	1343	1096	299	1443	1252	381
		Metamorpheus	1877	1983	289	2712	3064	1030
		Fragpipe-open	2618	552	602	2874	644	1851
		pFind	4246	3403	1927	4726	5212	2328
	DB2	Mascot	1109	1225	231	1344	1470	280
		MaxQuant	719	931	66	853	1099	136
		Metamorpheus	1127	1237	38	1845	2107	345
		Fragpipe-open	3073	2636	401	3798	3358	472
		pFind	3520	3006	1066	4188	4007	1305
Non-specific	DB1	Mascot	1506	1757	893	1689	1991	1036
		MaxQuant	1183	1241	974	1336	1422	1233
		Fragpipe-narrow	1424	970	657	1530	1088	766
		Metamorpheus	0	2382	1231	0	3067	2192
		Fragpipe-open	2621	652	718	2863	687	1516
		pFind	4221	3967	2351	4796	5482	3067
	DB2	Mascot	1093	1235	311	1377	1543	413
		MaxQuant	710	900	264	852	1127	451
		Metamorpheus	1467	1817	455	1791	2255	934
		Fragpipe-open	2973	1493	417	3682	1839	666
		pFind	3682	3123	1321	4283	4200	1662

Supplementary Table S2. Number of PSMs for each run and software and at FDR thresholds 0.01 and 0.05.

Supplementary Files

S1. parameters_table.xlsx is a spreadsheet with all the search parameters for MaxQuant, Mascot, Metamorpheus, pFind and Fragpipe. These are derived directly from the parameter files used to set up each software and run. Variable parameters (enzyme or database, or related to open/narrow search in Fragpipe) are highlighted.