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

Reviews round 1

Recommender

The submitted manuscript is an important contribution to paleoproteomics. However, the reviewers have identified several issues with the clarity of the manuscript text and figures. I recommend that the authors carefully address these on a point-by-point basis, particularly as this manuscript should serve as a guide for best practices in paleoproteomics.

We understand the importance of this work as a guide and its implications in palaeoproteomics. In line with this, the reviewer's comments have been helpful and insightful and we have taken into consideration and addressed them all. The assessments are more nuanced and detailed now, but at the same clearer on what the best practices would involve, according to the results.

We have also addressed the comments regarding the clarity of figures, captions or concepts.

Reviewer 1

Line 223 - Dairy database Bleasdale 2021 and Wilkin 2021 did not look for only dairy sequences, the goal was to identify any dietary proteins from consumption. The samples were searched against Swissprot entirely and also included a custom dairy database as well.
A: We corrected this in the manuscript. These studies are indeed examples of a wide and deep approach, ie targeting many different protein families, or entire proteomes, from a variety of organisms.

Line 341 - "less accepted PSMs" should be "fewer accepted PSMs"
A: Corrected it in the manuscript.

Lines 318-321 Can you explain what the q_value is telling us? Is this like the expect value of the peptide ID? Or is this something related to FDR?

A: q-value is the minimum FDR at which an identification is deemed statistically significant. Lower q-values indicate higher confidence in the identification. We have included its definition in the manuscript.

Lines 334-336. If there is no examination of the quality of the IDs then is there a point of comparing them? Maybe I'm missing something, but if FragPipe is making a ton of IDs but they are unreliable or not found through other searches, do they really count? Are the IDs too good to be true? I'm honestly asking, as this has been a large concern of mine when deciding which program to use.
A: Thank you for raising this concern. The comparison of IDs generated by proteomic softwares is indeed a key aspect of evaluating their performance. However, as you pointed out, the quantity of IDs alone does not necessarily fully reflect their reliability. And this is why we say "FDR is not enough" and we aim to highlight that by comparing the number of identifications when enzymatic type of search and the database change, and between software.
To ensure a meaningful comparison, one must consider several factors such as 1) quality control measures, 2) cross validation with other tools, and 3) context specific evaluation of IDs. These are briefly explained below:

- Quality control measures: Like any other proteomic software, FragPipe does incorporate various quality control measures like FDR control, decoy database searches, and very strict filtering criteria to minimize false positives. Therefore, the IDs reported are not just numerous, but also subject to rigorous validation processes.
- Cross validation using other tools: To address the reliability of high quantity of IDs, one could compare its results with those obtained from other tools. Such a cross-validation will help identifying "consensus" IDs - those that are consistently found across different tools - that are most likely to be true positives.
- Context specific evaluation of IDs: While FragPipe may generate a high number of IDs, one must also consider the contextual (archaeological or palaeontological) relevance of these IDs, their consistency across samples, and their overlap with known proteomes. This helps in determining whether the additional IDs contribute to meaningful biological insights or are likely false positives.

We have tried to address this concern by comparing the identifications between software, in instances in which, eg. MaxQuant and pFind provide an identification for the same spectra.

Line 380 - Does "Figure 2" mean Figure 3? Also, I would reiterate in the legend to Figure 3 which are narrow and which are open again. Otherwise readers may have to flip around the paper to remind themselves which is which.
We have added it in the figure and in the caption

Line 397 - Referring to "Figure 3" is actually Figure 4 (with Venn diagrams). Check all figure mentions throughout, as many are incorrect.
Corrected throughout the text

Figure 4 - can you add a total of all PSMs recovered from each entire Venn diagram? This would help rather than readers adding them up.

We have added a table and figure with total PSMs. We have changed the Venn Diagrams for UpSet plots, so we can compare more sets together. In addition, the UpSet plots do not represent distinct peptides anymore but number of PSMs. This clarifies the comparisons.

Figure 5. What are the shaded regions underneath on the amino acid counts? This is mentioned in the main text at different points, but should also be referred to in the figure legend.
In order to simplify the figures, we have removed the shaded areas and no longer refer to them, but to specific positions and ranges in the sequences.

Overall - I am missing any discussion of the narrow searches. It would be great to have some assessment of how MQ and Mascot perform as these are the most commonly used programs in ancient proteomics. Are these less reliable than the open searches? They do seem to ID fewer peptides, but are the open search IDs of good quality? If not within the scope of this paper, this would be a great next step. In conclusion, I would make a stronger stance on how to best go about searches. Maybe some sort of table or figure that outlines the suggestions of using an open search first and then narrowing the space. Especially that there is little to no discussion of MQ and Mascot, the most frequently used programs in ancient proteomics. Also, I would add more detail and a figure to show what an ideal data analysis pipeline would be.
We have added further and more detailed assessments on MaxQuant and Mascot, especially in comparison with the open search software, which we also hope will address the previous comment about the reliability of the IDs. We also compare and discuss their identifications with the Open searches.

We have made a stronger instance on the use of Open Search in general and have added a short section summarising an ideal pipeline in Palaeoproteomics combining the strengths of each method.

Reviewer 2

Palomo et al. apply various proteomic searching tools to experimentally degraded beta-lactoglobulin. This is a paper that has important implications in palaeoproteomics. A couple of things that I would like to see included is a closed MSFragger search to compare with the other closed searches and a detailed supplementary list of parameters for each algorithm. I realize they are included on Zenodo, but having them with the paper will be beneficial.
We have included a separate spreadsheet with the parameters used in each software. This is derived directly from the parameters files used in each software.

With Metamorpheus, what search type was used? Specifically was classic search used for tryptic, non-tryptic, and nonspecific or the modern indexed search versions?

A: We used the latest Enhanced Global Post-Translational Modification Discovery (G-PTM-D) algorithm. It creates a fragment index, if this is what is referred to.

For denovo/any search, especially on the 128 day samples, is there a correlayion to the level of aminoacid identification to the isoelectric point of the peptides? It seems like there is a bias toward certain sections of coverage for the tryptic peptides compared to the non-specific searches. Can one of these approaches help find different types of peptides/protein fragments based on the composition of the peptides?
We have calculated the isoelectric point along the BLG and it does seem to explain certain coverage biases. To complement this, we also calculated other phisico-chemical properties that can help understand these patterns along with the isoelectric point.

Can you include a comparison of the number of PSMs detected per algorithm as well? Also, I'd be interested to see Figure 4 with PSM identifications instead/in addition to the unique peptide counts, so the unit of comparison between the algorithms is the same.
We have added in Supplementary Figure 1 and Supplementary Table 1 the total PSMs. We have also changed the Venn Diagrams for UpSet plots and these now represent PSMs rather than unique peptides

Line 126: Should 156 long be 156 amino acid long?
Corrected in the text

Line 148: What aqueous solution was used?
A: We used ultrapure molecular grade water. We have corrected this in the manuscript.

Figure 1: For the schematic, change the LC-MS/MS picture to an Exploris.
A: Thanks for pointing out the instrument swap. We have updated the figure with Exploris 480.

Line 160-161: Include a short form of the Cappellini et al. 2019 extracϴon protocol beyond the various basic summary included here.
A: Added detailed protocol in the manuscript.

Table 1. List the versions of pFind 3, Metamorpheus, Mascot, Novor, DirecTag, and PepNovo+.
A: We have added the information in table 1

Line 194-196: Why was Fragpipe only run on a cluster instead of also on the MiniMax workstation? Fragpipe can natively run Thermo RAW file format as well. Conversely, why wasn't the same peak picked mzML file used with pFind3, Metamorpheus, and Maxquant.
We run Fragpipe on a cluster because the peptide fragment index (which is the reason behind the high speed) takes up enough memory to make the large proteome tricky to run on that workstation, as it would exceed the memory limits. At the same time it is easier to set up a batch of runs on a cluster on Fragpipe. On the other hand, MaxQuant could not be run on the cluster as it has a hardwired time wall that it would exceed in most runs. Although it was possible, we also note that it tends to be less stable on Linux systems.

We do not believe this would have changed the results. But note we run a subset of the runs on MaxQuant on the same cluster in order to compare running times with Fragpipe. This way we can compare running times of MaxQuant vs pFind on MaxQuant and MaxQuant vs Fragpipe on the cluster.

We note mzML vs RAW discrepancy. We have rerun Fragpipe using the Thermo RAW file.

Line 321: Figure 2 should be Figure 3.
Corrected in the text

Figure 3: What do the 2 dashed blue lines represent? Additionally, make sure that the colour choices are colorblind accessible. With a simulator Tryptic DB2 is very similar to Semi DB2 and Trypttc DB1 is very similar to NS DB1.

The blue lines mark the 0.01 and 0.05 FDR thresholds. We have mentioned this in the caption of the figure.

We have changed the colors here, which now only represent the database, while the type of tryptic search is represented in either solid, dashed or dotted lines. The figure is now clearer and it is easier to get the general trends.

Figure 4: A series of Upset plots may be easier to understand/compare than these 3 circle Venn diagrams.
We have now added UpSet plots. They indeed make the comparison between multiple sets easier.

Figure S2 was not included in this preprint.
Corrected in the text.

Figure 5 A, B, D: Make sure to use a colorblind palette. The importance of the amino acid colors here is completely lost with this palette.
To decrease the amount of colours, now amino acids colours are represented by side-chain properties. And we use the muted colorblind-safe palette from: https://personal.sron.nl/~pault/#fig:scheme_vibrant