

Community benchmarking and evaluation of human unannotated microprotein detection by mass spectrometry based proteomics

Corresponding Author: Dr Anne-Ruxandra Carvunis

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This manuscript by Aaron Wacholder is an interesting observation that sheds some light on the difficulty of defining the unannotated proteome, which is deduced by Ribo-Seq analysis of translated RNAs recovered from purified ribosomes but is not included in the well-annotated and characterized proteome. A few manuscripts have described attempts to validate the existence of proteins translated from this unannotated part of the genome using LC-MS/MS analysis of tryptic digests of protein extracts or from immunopeptidomes, affinity purified from cells. It was previously claimed that immunopeptides are significantly derived from short-lived proteins, many of which are produced from unannotated transcripts aberrantly produced in cells. The source of microproteins among the HLA peptidome can be from DRiPs since they are rapidly degraded, and the resulting peptides are enriched during the affinity purification with the rest of the immunopeptidome.

Wacholder et al. noticed that only a fraction of the peptides discovered in each of these studies were detected in more than one of them. This raised the concern that either the identifications of the peptides/proteins translated from the unannotated transcripts are erroneously identified due to the high error rate in such analyses or that these proteins are very unstable and, therefore, are present in minute amounts among the tryptic peptides.

The manuscript seems to be sound, and the data is solid.

In a previous manuscript by some of these authors (Prensner et al. 2023, ref. 15, Figure 4A), they describe the same phenomena of small overlap in the sequences of identified peptides between the studies and discuss the possible reasons for this small overlap. A discussion of the differences and the additional data presented here is warranted.

The validation of the peptides can also be supported by the mass difference between the observed and calculated, not just by their retention times and MS/MS spectra. Is this data available?

The HLA peptide identifications can be supported by their fitness to the HLA binding motifs of the cells/tissues. Is there an indication for that?

Identification is not only supported by the identity in the retention times of synthetic and natural peptides but also by the similarity in their MS/MS fragmentation pattern.

Table S5 does not seem to contain the translation levels, as indicated in line 283. Please clarify.

Is there a correlation between the score of the sORFs identification and the peptides identification score by the MS/MS?

Please consider describing such a correlation if it exists.

Which software tools were used to identify the peptides or the non-canonical ORFs in the different studies? Maybe the differences are caused by the search engine used?

What is the meaning of (ORF in Tier 1A, 26 in Tier 1B, and 10 in Tier 2B (line 270)? Please explain.

Please consider in Figure 1B and Figure 1D to present the bars as numbers rather than percentages or also include the numbers, as these are derived from largely varying nu, numbers.

Line 329: Please provide a reference for the claim that many sORFs code for membrane proteins if there is such an indication.

How about studies that use tissues rather than cell lines? What about normal tissues and cells rather than cultured cell lines? Is it possible that cultured cancer cells produce aberrant microproteins due to defects in translation control, but this phenomenon is not prevalent in healthy tissues? It can also be that some cultured cancer cells are more defective in their control of the translational machinery. Is there an indication for this?

(Remarks to the Author)

In their manuscript "Detection of human unannotated microproteins by mass spectrometry-based proteomics: a community assessment", Wacholder and co-authors describe the analysis of twelve selected proteomics studies that all aimed to uncover evidence for as of yet unannotated human small proteins. While this is a very important area of research, I feel this contribution from a large group of experts in the field would benefit from providing additional detail and context to other model systems where such proteogenomics approaches have been done for more than two decades. Most importantly, I was missing a set of recommendations on how to carry out such studies to overcome the key problems in the more challenging case of the human genome, the underestimation of false positives when using too lenient false discovery rate (FDR) thresholds, the need to use local FDR estimates for novel proteins and the issue of how to deal with a reference genome sequence.

In the data analysis exercise presented, some aspects should likely be explained in a bit more detail. A rather interesting distinction between the quality of peptide spectrum matches (PSMs) from immunopeptidomics studies versus other studies aiming to identify novel human small ORFs was uncovered. Yet, potential underlying reasons are not further explored and remain unclear.

The 1% FDR threshold used at the peptide/protein level (sometimes even PSM level!) in the twelve selected studies (one even used 10%FDR at the peptide level) is clearly not stringent enough when aiming to discover novel human microproteins; this would almost ask for a re-analysis at a more stringent PSM level and considering local FDR rate for the class of novel CDS to arrive at smaller sets of novel human microproteins, containing substantially fewer false positives. Overall, it appears that the field of human proteogenomics could largely benefit from a set of guidelines/best practices to assist researchers in this highly important endeavor.

Comments

1. The manuscript should provide some definitions, more context and explanation of the key challenges in human proteogenomics compared to simpler model organisms.

It would be helpful to explain to the readers what threshold the authors rely on when they refer to short open reading frames (sORFs). In the literature, different thresholds are used at the level of the encoded proteins (100aa, 70aa, 50aa). Unless I missed it, this point is not explicitly mentioned but should be. Also, when aiming to identify small proteins, the two peptide rule mentioned needs to be taken with more caution, as often, few or even only one peptide are within the visible range of the mass spectrometer. I was not sure if data for all novel ORFs are provided, i.e., how many distinct peptides and PSMs imply these novel ORFs (overall 9414 peptides, line 111)?

Looking at table 1, it becomes apparent that different genome sequences and annotation releases have been used. These complicate tracking the genome coordinates of novel sORFs (exemplified by the fact that for 3-4 studies, the authors could not generate them), which are highly relevant. Did the authors use one genome sequence as a common reference for their re-analysis or did they rely on the references used by each study?

How would results of different proteogenomics studies be used in the future to further improve a (and which) human reference genome sequence? This is a key difference to proteogenomics e.g. in prokaryotes or other eukaryotic model systems, where the genome sequence of the bacterial strain in question can be readily sequenced and de novo assembled, eliminating this level of ambiguity. Even in plants like *A. thaliana*, having one reference genome simplifies the analysis substantially, compared to e.g. integrating results of studies from human cancer samples.

2. When attempting to identify novel human microproteins, using a 1% FDR rate at peptide/protein level is by no means stringent enough and not adequate; as the authors state, these thresholds are used for identification of proteins present in an annotated reference database (e.g. UniProt or NCBI RefSeq). In reference 16 the authors cite (Nesvizhskii AI, Nature Methods 2014), who has elaborated on the needs of more stringent FDRs, as have early proteogenomics studies in prokaryotes, which advocated for using more stringent cut-offs (PMID: 22114679), at times going even down to 0.01% FDR at the PSM level (PMID 29141959).

I was thus rather surprised to find that these insights have apparently not translated into human proteogenomics (yet), and that only few of the 12 studies worked with local FDRs and several still used a 1% PSM level FDR threshold. This would certainly make a set of guidelines desirable if not mandatory. In my view, the Douka study should be removed from the set; a 10% peptide level FDR (proteome-wide) should not qualify to be included in such a comparison.

3. Lots of work in the proteogenomics field has been done on prokaryotic organisms, where proteomics data has been searched against six frame translations of the genome sequence or more sophisticated custom search databases. While readily feasible for prokaryotes, this is more complicated for the substantially more complex eukaryotic genomes. A major advantage though is the availability of genomic coordinates of all novel sORFs. The lack of this data (in some cases) seems a complication of proteogenomics in humans as the different studies seem to have used different reference genome sequence data (please correct me if wrong here) and for sure different annotation releases (e.g. Human UniProtKB 2017-2021), etc.

4. Ribo-seq is indeed extremely useful as a basis to create an extended custom search database to find evidence for so far unannotated human microproteins by mass spectrometry. Mass spectrometry can help in identifying stable protein products; I believe the aspect of pervasive translation may even be stressed a bit more.

For the Ribo-seq approach, one must keep in mind that due to the higher sensitivity, a good fraction of the signals will represent proteins that get rapidly degraded and may not be found to be stably expressed. At least for *M. tuberculosis* H37Rv, a key bacterial model organism, pervasive translation was observed with Ribo-seq for over 2000 unannotated ORFs

(compared to roughly 4000 annotated CDS), only a minor fraction of which (90) showed evidence of purifying selection which was interpreted as evidence that they represent bona fide missed proteins (PMID: 35343439). Currently, several studies have employed both Ribo-seq and MS-based proteomics and show complementary sets of identifications.

5. The authors may want to provide a rough estimation of the number (or percentage) of sORFs that can be expected to be found in the human genome. In prokaryotes, studies in *E. coli* uncovered that 140 sORF encoded proteins (SEPs) below 50aa could be identified from 2010-2020 (see PMID: 32385980), i.e. roughly 3% when assuming ~4300 annotated CDS. With a cut-off of 100aa, this number would undoubtedly be substantially higher and amount to several percent of annotated CDS.

Here, based on the subset of PSMs analyzed (the authors should add what %age of all PSMs implying novel human proteins these represent), the rough extrapolations shown in Figure 2G and the fact that the 2021 Ouspenskaia dataset with the largest number of novel human microproteins predicted (4900) fared quite well in terms of the spectral quality assessment (but still suffers from a too lenient 1%PSM level FDR threshold), it appears that this could include hundreds to above 1000-2000 thousand novel human microproteins.

6. I was missing a take home message concerning manual evaluation of MS/MS spectra versus the ability to rely on in silico predicted spectra. This will majorly impact the ability to carry out larger scale analyses and become independent from expert evaluators.

7. Most helpful for the community would be a set of recommendations for an approach to identify novel human microproteins following a set of guidelines/best practices, such that obvious and avoidable errors can be minimized/eliminated (selecting too lenient FDRs at the spectrum, peptide or protein level, not considering local FDR among novel proteins, etc.).

What would the authors recommend in terms of reference genome sequence and latest annotation release (Table 1 lists Human UniPROT releases from 17 to 21; do their protein sequences also carry genome coordinates to locate them on the reference genome)? How can studies from cancer samples be integrated or should they rely on separate genome assemblies? What kind of tools can be employed to minimize the need for manual evaluation of MS/MS spectra. What kind of analyses can be carried out to prioritize among the list of hundreds/thousands of potentially novel human microproteins to focus on the most relevant ones?

Reviewer #3

(Remarks to the Author)

More and more sORFs and their encoded peptides were identified by Ribo-Seq and Mass spectrometry. The MS-based peptidomics is a reliable approach to finding novel microproteins. It's very important to figure out the issue with the detection quality and help find more functional microproteins. Overall, the manuscript is well written, with only a few minor issues that need to be addressed:

1. What are the criteria for collecting the human unannotated protein dataset? Only publish between 2019 and 2022, or any other conditions?
2. The authors emphasize the unannotated microproteins and focus on the short open reading frame-encoded peptide. The dataset includes both immunopeptides and non-HLA peptides. The immunopeptides differ from sORF-encoded peptides and may come from known large proteins. The authors should explain more about "unannotated microproteins" and tell the difference between peptides from various sources.
3. "A key motivation for initiating this community effort was the large variation in the number of validated unannotated proteins reported between studies, ranging from 6 to 4,903". The variation may be due to the sample size; for example, some papers only use one cell line, while others analyze plenty of tissue or cell samples. The author may do further statistical analysis to see how many sORF-encoded peptides can be identified in a single run or a single cell/tissue sample for every research. By the way, for reference 28, the sORFs number is 4903 in Table 1. But the original paper mentioned "We constructed a high-confidence database of translated nuORFs across tissues (nuORFdb) and used it to detect 3,555 translated nuORFs from MHC-I immunopeptidome mass spectrometry analysis" and "retaining 6,501 high confidence (FDR<1%) peptides from 3,261 nuORFs, across various nuORF types". So, could the authors explain where the number 4903 comes from?
4. For the dataset that 96% matched to annotated protein, most of which are tryptic peptides. Can not only say it's because they use a custom database of specific samples. Have any of those annotated proteins been proven not to be expressed in those samples? Can they show one example? Otherwise, if the peptide is not a unique peptide, then the identified peptide can not be used as existing evidence of a sORF-encoded peptide.
5. What is the main reason that makes HLA peptides different from non-HLA peptides, sample preparation, database search, or the characteristics of the peptides? Please have a more in-depth discussion on this based on publications. Is there any suggestion for improving the identification of non-HLA peptides?
6. Supplemental Table 3 is hard to search for information in the TXT form, maybe switching it to CSV form and putting the publications of each peptide in different columns.
7. Page 9 said "A total of 406 PSMs from 12 studies were evaluated.....Additionally, a common set of 10 negative control PSMs was included in each sample...". However, there are 439 mzspec and 15 control spec in Supplemental Table 4. Please double-check the number.
8. The information in Supplemental Table 5 is hard to follow. Could the author provide the expression data of each sORF related to Figure E?

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I have no further comment on the manuscript

Reviewer #2

(Remarks to the Author)

The authors have done a nice job and addressed almost all of the suggestions/questions raised in the review of their original manuscript entitled "Detection of human unannotated microproteins by mass spectrometry-based proteomics: a community assessment".

The revised version includes recommendations/guidelines for the community, provides possible explanations why immunopectidomics-derived peptides may have higher spectral quality, points out the value of custom search databases based on a specific genome sequence (if available), the value of strictly controlling the local FDR for novel proteins, and referenced two important papers covering the aspects of pervasively translated ORFs identified by RiboSeq and short proteins added over time to the genome of E.coli.

I understand the motivation to stick with the 12 datasets (instead of removing one case which had used too lenient FDR thresholds), and to opt to not provide specific FDR thresholds.

I support publication of the revised manuscript and believe it will become a very useful resource for researchers interested in identifying novel human microproteins.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

We greatly appreciate the time and effort by the three reviewers in carefully evaluating our initial submission and offering helpful feedback. We have made many changes to the text to improve clarity and add relevant context. Additionally, in response to reviewer concerns, we have added two new analyses to the revised manuscript presented in Supplementary Figures 1-2. Both analyses provide insight into how the manual evaluator ratings of peptide-spectra matches relate to attributes of the evaluated peptide or spectra. Finally, in response to multiple reviewer requests to add advice to the manuscript based on our findings, we have included advice both as a concise list of bullet points and as a detailed checklist in Appendix 2, which was modified from the Human Proteome Project Data Interpretation Guidelines. We thank all three reviewers for their many excellent suggestions, and we are confident that following these suggestions has substantially improved the manuscript. Please find below a point-by-point response to the reviewer comments.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This manuscript by Aaron Wacholder is an interesting observation that sheds some light on the difficulty of defining the unannotated proteome, which is deduced by Ribo-Seq analysis of translated RNAs recovered from purified ribosomes but is not included in the well-annotated and characterized proteome. A few manuscripts have described attempts to validate the existence of proteins translated from this unannotated part of the genome using LC-MS/MS analysis of tryptic digests of protein extracts or from immunopeptidomes, affinity purified from cells. It was previously claimed that immunopeptides are significantly derived from short-lived proteins, many of which are produced from unannotated transcripts aberrantly produced in cells. The source of microproteins among the HLA peptidome can be from DRiPs since they are rapidly degraded, and the resulting peptides are enriched during the affinity purification with the rest of the immunopeptidome.

Wacholder et al. noticed that only a fraction of the peptides discovered in each of these studies were detected in more than one of them. This raised the concern that either the identifications of the peptides/proteins translated from the unannotated transcripts are erroneously identified due to the high error rate in such analyses or that these proteins are very unstable and, therefore, are present in minute amounts among the tryptic peptides.

The manuscript seems to be sound, and the data is solid.

We thank the reviewer for their kind words and helpful feedback.

In a previous manuscript by some of these authors (Prensner et al. 2023, ref. 15, Figure 4A), they describe the same phenomena of small overlap in the sequences of identified peptides between the studies and discuss the possible reasons for this small overlap. A discussion of the differences and the additional data presented here is warranted.

Thank you for this suggestion. The main difference between our study and the data shown in Figure 4A in the Prensner et al., 2023 study in MCP is that our study reports on the poor

reproduction of MS-based identifications while the Prensner et al. 2023 Figure 4A highlights differences among ribo-seq databases. Nevertheless, these issues are related. In the revised manuscript, we now cite Prensner et al., 2023 to emphasize how the low overlap at the ORF level may partly explain the low overlap in detected unannotated peptides by mass spectrometry:

We do not interpret the high variability between studies as indicating that most reported detections are false: this high variability among reported detected peptides likely reflects in part the high variability in the size and composition of the sORF databases tested (Prensner et al. 2023, Table 1) and the quantity of proteomic data analyzed, as well as the diversity of cell types examined, MS techniques used, HLA allotypes among the immunopeptidomics studies, and search algorithms.

The validation of the peptides can also be supported by the mass difference between the observed and calculated, not just by their retention times and MS/MS spectra. Is this data available?

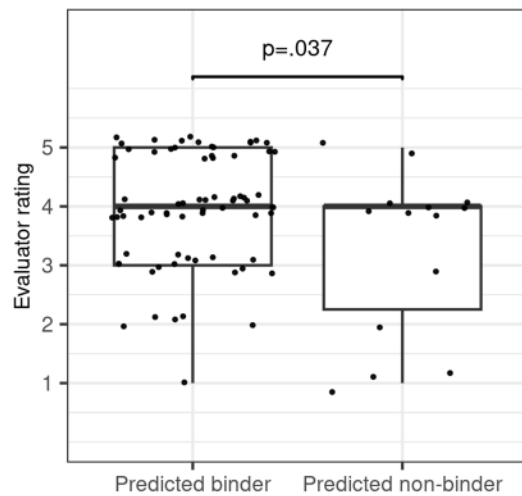
Yes, the observed and calculated precursor m/z values are available via USIs for each of the putative identifications. We evaluated PSMs primarily on the basis of the fragmentation spectra rather than the precursor mass differences because these deltas are generally best interpreted in the context of other identifications in the same MS run, since there can often be calibration drift, which is typically modeled and compensated for during full analysis but not visible via the USI. Furthermore, the precursor m/z values reported by the instrument are frequently off by one or more isotope deltas (e.g. 1.003355 m/z), which also adds complication. The mass difference data are available for anyone who wishes to further examine and analyze them.

The HLA peptide identifications can be supported by their fitness to the HLA binding motifs of the cells/tissues. Is there an indication for that?

We thank the reviewer for suggesting this analysis. We used NetMHC to predict HLA binding affinity for all HLA-derived peptides from our list of manually evaluated PSMs. We were able to obtain predictions for 85 peptides for which the original papers reported the HLA allele expressed in the cell and found that the HLA peptides with high fitness for HLA binding motifs had scored higher in our manual evaluations. We have included this analysis in the revised manuscript:

Agreement among evaluators was generally high. For the PSMs rated by two evaluators, ratings were well correlated ($r = 0.82$, $p < 10^{-10}$) (Figure 2A). Only 14 of 155 (9%) PSM scores differed by more than one point. The negative controls scored consistently poorly (average score of 1.5), as expected. Evaluator ratings were also well correlated ($r = 0.74$, $p < 10^{-10}$) with the dot product between the observed spectra and the spectra predicted by MS2PIP (Supplementary Figure 1).⁴⁸ Among immunopeptidomics studies, PSMs with peptides that were predicted to bind to MHC molecules by NetMHC⁴⁹ were rated more highly ($n = 71$, mean rating 3.94) than those with peptides not predicted to bind ($n = 14$, mean rating 3.29, $p = 0.037$ by permutation test, Supplementary Figure 2,

Supplementary Table 5), consistent with manual evaluation discriminating between true and false discoveries.

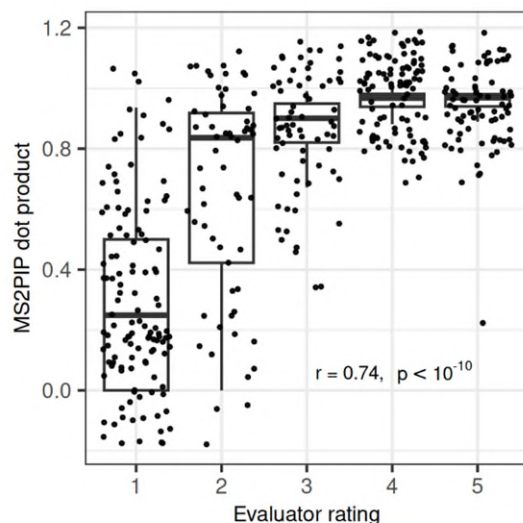


Supplementary Figure 2: Peptides that are predicted to bind HLAs are more highly rated by evaluators. For each evaluated peptide from Ouspenskaia et al. 2021, Martinez et al. 2020, or Chong et al. 2020, HLA binding was predicted using NetMHC. Any peptide meeting the NetMHC criteria for weak or strong binder to an HLA allele present in the cell type was considered a predicted binder. The distribution of evaluator ratings among predicted binders and peptides not predicted to bind HLAs is shown

Identification is not only supported by the identity in the retention times of synthetic and natural peptides but also by the similarity in their MS/MS fragmentation pattern.

We agree with the reviewer. For the revised manuscript, we have added an additional analysis addressing this point. We predicted the MS/MS spectra using MS2PIP and calculated the dot product between the predicted and observed spectra:

*Agreement among evaluators was generally high. For the PSMs rated by two evaluators, ratings were well correlated ($r = 0.82$, $p < 10^{-10}$) (Figure 2A). Only 14 of 155 (9%) PSM scores differed by more than one point. The negative controls scored consistently poorly (average score of 1.5), as expected. **Evaluator ratings were also well correlated ($r = 0.74$, $p < 10^{-10}$) with the dot product between the observed spectra and the spectra predicted by MS2PIP (Supplementary Figure 1).***



Supplementary Figure 1: Evaluator rating is strongly correlated with dot product between observed spectra and spectra predicted by MS2PIP. MS2PIP was used to generate predicted spectra for each evaluated PSM. The dot product between the predicted and observed spectra is shown for each PSM, with PSMs grouped by manual evaluator rating. The correlation between dot products and ratings is given.

Table S5 does not seem to contain the translation levels, as indicated in line 283. Please clarify.

We thank the reviewer for pointing out that this data was missing. We added a Table S7 to the revised manuscript, which we now reference together with Table S6. Table S7 includes translation levels for every ORF we analyzed in the section titled “Higher rated PSMs are derived from more highly expressed sORFs”, as well as the ORF coordinates.

Is there a correlation between the score of the sORFs identification and the peptides identification score by the MS/MS? Please consider describing such a correlation if it exists.

We investigated this question using the p-values generated by the iRibo program, which indicate confidence that a given ORF is translated based on accumulated ribo-seq data. There is no significant correlation between log iRibo p-value and spectra rating ($r = 0.098, p = 0.18$). We now note this in the results of the revised manuscript:

There was no significant correlation between log iRibo p-value, indicating level of confidence that the ORF is translated, and PSM rating ($r = 0.098, p = 0.18$).

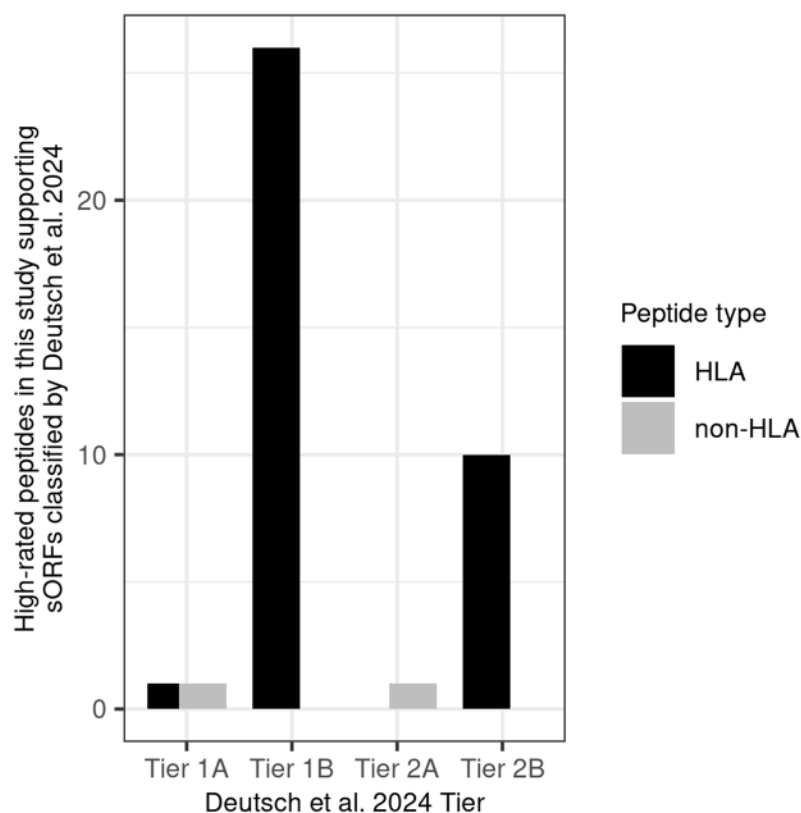
Many of us investigated this general question more deeply in a related preprinted article, Deutsch et al. 2024, where we also observed no correlation between ribo-seq based sORF scores and PSM match quality (<https://www.biorxiv.org/content/10.1101/2024.09.09.612016v1>).

Which software tools were used to identify the peptides or the non-canonical ORFs in the different studies? Maybe the differences are caused by the search engine used?

Thank you for asking this question. In our revised manuscript, we included the software used by each study for mass spectrometry analysis to Supplementary Table 2. We did not observe any significant associations or evident patterns between the software and spectra ratings.

What is the meaning of (ORF in Tier 1A, 26 in Tier 1B, and 10 in Tier 2B (line 270)? Please explain.

Thank you for pointing out the need to explain the meaning of this tier system. To this aim, we included in our revised manuscript a supplementary figure comparing our results to the ORFs tiers in Deutsch et al. 2024 and detailed the tier definitions in the legends of this figure.



Supplementary Figure 4: Peptides highly rated in this study that also supported ORFs classified in Deutsch et al. 2024. The number of high-rated peptides (ratings of 4 or 5) analyzed in this study that were also validated in Deutsch et al. 2024 and used as support for ORFs of various tiers classified in that paper. Peptides are divided by whether they were found in HLA immunopeptidomics experiments or non-HLA conventional proteomics experiments. The ORF tier definitions, taken from Deutsch et al. 2024, are as follows. Tier 1A: ORF supported by at least two non-nested peptides from conventional proteomics experiments as well as Ribo-Seq data. Tier 1B: ORF supported by at least two non-nested peptides from HLA immunopeptidomics experiments as well as Ribo-Seq data. Tier 2A: ORF supported by at least one peptide from conventional proteomics experiments as well as Ribo-Seq data. Tier 2B: ORF supported by at least one peptide from HLA immunopeptidomics experiments as well as Ribo-Seq data.

Please consider in Figure 1B and Figure 1D to present the bars as numbers rather than percentages or also include the numbers, as these are derived from largely varying nu, numbers.

To address this concern, we revised Figure 1B and Figure 1D to show the numerator and denominator above the bars, thus providing the number of peptides identified in other studies in addition to the percentage.

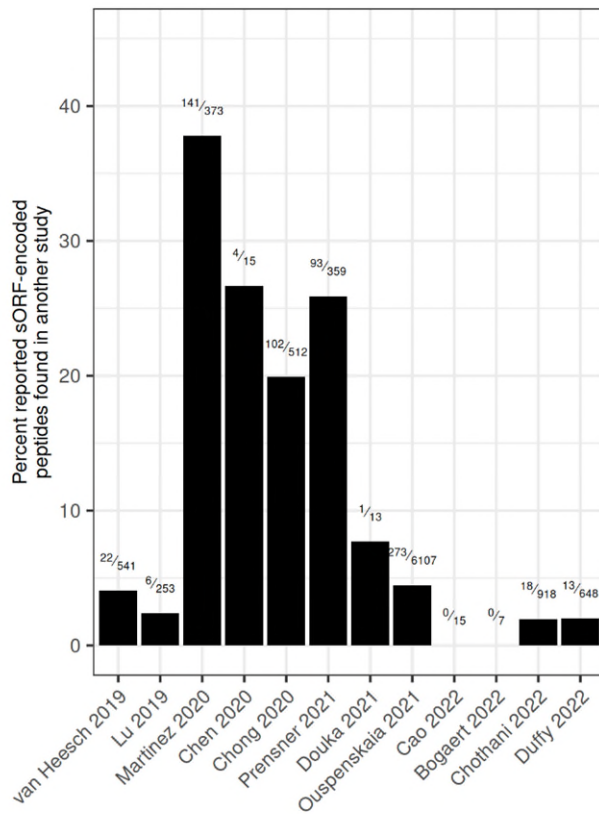


Figure 1B

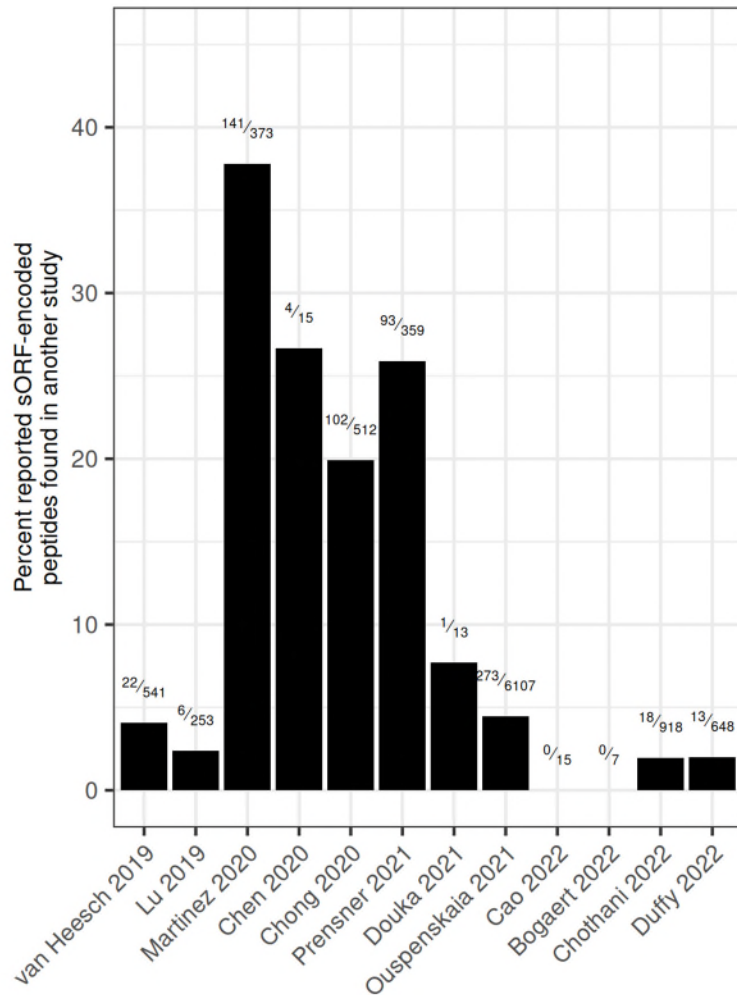


Figure 2D

Line 329: Please provide a reference for the claim that many sORFs code for membrane proteins if there is such an indication.

Thank you for pointing out this omission. In the revised manuscript, we provide a reference to Makarewich 2020, a review of membrane microproteins that supports this point.

How about studies that use tissues rather than cell lines? What about normal tissues and cells rather than cultured cell lines? Is it possible that cultured cancer cells produce aberrant microproteins due to defects in translation control, but this phenomenon is not prevalent in healthy tissues?

It can also be that some cultured cancer cells are more defective in their control of the translational machinery. Is there an indication for this?

The reviewer points out a fundamental question in microprotein biology: it is definitely plausible that cancer cell lines express some microproteins that are not expressed in healthy cells, for example because cancer cells could be defective in translation control. Based on the data we collected for this manuscript, however, we see no clear evidence of this. As shown in Table 1, some studies in our meta-analysis used cell lines, others used healthy tissues, and others used both. We see no significant differences between studies that used cell lines and studies that used healthy tissues in PSM quality or the number of high-quality PSMs. That being said, this observation does not imply that no important differences between the microproteomes of cell lines and tissues exist. The questions raised by the reviewer are of great interest and importance to the microprotein field and we hope to pursue them further in future studies designed more specifically to address them.

Reviewer #2 (Remarks to the Author):

In their manuscript “Detection of human unannotated microproteins by mass spectrometry-based proteomics: a community assessment”, Wacholder and co-authors describe the analysis of twelve selected proteomics studies that all aimed to uncover evidence for as of yet unannotated human small proteins. While this is a very important area of research, I feel this contribution from a large group of experts in the field would benefit from providing additional detail and context to other model systems where such proteogenomics approaches have been done for more than two decades. Most importantly, I was missing a set of recommendations on how to carry out such studies to overcome the key problems in the more challenging case of the human genome, the underestimation of false positives when using too lenient false discovery rate (FDR) thresholds, the need to use local FDR estimates for novel proteins and the issue of how to deal with a reference genome sequence.

We thank the reviewer for their kind words and helpful feedback. We have revised the paper to address these concerns, as explained further below.

In the data analysis exercise presented, some aspects should likely be explained in a bit more detail. A rather interesting distinction between the quality of peptide spectrum matches (PSMs) from immunopeptidomics studies versus other studies aiming to identify novel human small ORFs was uncovered. Yet, potential underlying reasons are not further explored and remain unclear.

We are very interested in understanding why immunopeptidomics studies report higher quality PSMs for novel human small ORFs than those using conventional proteomics. Our results, together with those from other studies such as Deutsch et al. 2024 (<https://www.biorxiv.org/content/10.1101/2024.09.09.612016v1>), strongly suggest that the spectra generated in immunopeptidomics experiments have more true matches to sORFs than the spectra generated by shotgun proteomics. This could be for biological reasons: the sORFs are translated into proteins but quickly degraded and as such are not detectable in cell extracts. Or it could be for technical reasons: the conditions of immunopeptidomics are more conducive for identifying these peptides. Unfortunately, the examination of mass spectra in our manuscript is not sufficient to distinguish these possibilities because either scenario would produce the same observation: that immunopeptidomics experiments generate more high-quality sORF PSMs. So, we fully agree with the reviewer that the underlying reasons for this difference remain unclear and that this is a problem that must be explored further. But, unfortunately, our data is insufficient to do so. We believe that resolving this question may require the development of more sensitive proteomic technologies that will allow us to better detect and quantify low abundance microproteins in the cell.

To address the reviewer's concern, in our revised manuscript we have made some changes to the relevant section of discussion that make these points clearer:

Why do immunopeptidomics studies identify many high-quality PSMs supporting unannotated protein detections while studies using conventional enzymatic digests

identify only few? Many unannotated sequences found to be translated by Ribo-Seq lack signatures of evolutionary conservation and may not encode proteins that provide any benefit to the organism.^{5,15,57} It is plausible that many of these poorly conserved proteins are expressed but quickly degraded, and so can be found only as peptides bound to HLAs.^{14,58} However, there are also technical explanations for why HLA-bound peptides derived from unannotated microproteins may be easier to detect. Immunopeptidomics concentrates peptides bound to HLAs, which decreases sample complexity and may thereby enrich for low abundance microproteins. HLA peptides also have physico-chemical properties different from tryptic peptides that may affect detectability. Most immunopeptidomics datasets are from cancer samples, and some proteins may be expressed in some cancers but not in normal physiological conditions. Furthermore, microproteins may preferentially reside in cellular compartments that are hard to sample through non-HLA MS, such as membranes.²⁶ Moreover, the laboratories that perform immunopeptidomics are often distinct from those that analyze non-HLA data and may differ in their sample preparation techniques, experimental setup, and analytical choices. Understanding which factors are most important to explaining the difference between immunopeptidomics and conventional shotgun proteomics may require the development of more sensitive proteomic techniques for identifying low-abundance and short-lived microproteins in the cell.

The 1% FDR threshold used at the peptide/protein level (sometimes even PSM level!) in the twelve selected studies (one even used 10%FDR at the peptide level) is clearly not stringent enough when aiming to discover novel human microproteins; this would almost ask for a re-analysis at a more stringent PSM level and considering local FDR rate for the class of novel CDS to arrive at smaller sets of novel human microproteins, containing substantially fewer false positives.

We agree that our results indicate that most analyzed studies were not stringent enough. We hope our manuscript, by assessing previous claimed detections made in the literature, serves to warn against common practices (such as >1% FDR) that seem to generate many false positive claims.

Overall, it appears that the field of human proteogenomics could largely benefit from a set of guidelines/best practices to assist researchers in this highly important endeavor.

We agree with the reviewer that this would be useful, and have added a list of guidelines at the end of the revised manuscript:

Box 1: Advice for detection of novel microproteins using mass spectrometry-based proteomics

- *Ensure peptides appearing to support a novel protein detection uniquely support that protein:*
 - a. *Conduct a search using tools such as ProteoMapper⁴⁵ or PepQuery⁶⁰ to exclude peptides with possible matches to canonical proteins, including post- and co-translational modifications and common genetic variants. When possible, construct a sample-specific protein database that accounts for genotype. Do not assume a*

canonical protein is absent from the sample solely on the basis of gene transcription or translation evidence.

- b. Consider whether the peptide may come from a previously unannotated isoform of a known protein-coding gene, as gene annotation databases do not comprehensively capture all transcript diversity. Ideally, integrate short- or long-read transcriptomics data to determine whether the evidence supports an unannotated alternative transcript or splicing event that could explain the observed translation.*
- c. Pseudogene annotations can significantly impact microprotein discovery. Always check whether the peptide overlaps with a known pseudogene locus from either the Ensembl-GENCODE or RefSeq catalog.*
- *Ensure that the PSMs used to support a novel protein detection are high quality:*
 - a. Among PSMs that score highly in a search engine, spectra match quality can be further supported by comparison to experimental spectra generated from synthesized peptides, comparison to in silico fragmentation spectra generated by methods such as ProSit⁶¹ or MS2PIP,⁴⁸ and machine learning rescoring using approaches such as Oktoberfest⁵⁰ or MS2Rescore.⁶²*
 - b. Manual evaluation of a representative subset of PSMs is important to ensure reported detections are supported by high quality evidence.*
 - c. To accurately convey confidence in the list of unannotated protein detections, report local FDRs or FDRs specific to the list of unannotated proteins instead of or in addition to proteome-wide global FDR. The less stringent the FDR threshold used, the more it is necessary to examine candidates further to ensure they are correct.*
- *Make the MS data available in a public data repository. Report universal spectrum identifiers (USIs)⁶³ for all spectra supporting discovery of a novel protein.*

We also added more detailed guidelines, based on the existing Human Proteome Project guidelines, to appendix 2. We believe these additions will benefit researchers who endeavor to discover novel peptides.

Comments

1. The manuscript should provide some definitions, more context and explanation of the key challenges in human proteogenomics compared to simpler model organisms.

It would be helpful to explain to the readers what threshold the authors rely on when they refer to short open reading frames (sORFs). In the literature, different thresholds are used at the level of the encoded proteins (100aa, 70aa, 50aa). Unless I missed it, this point is not explicitly mentioned but should be.

We did not specify an explicit length because the starting point of our analyses was lists of proteins declared by previously published studies to be the unannotated detected proteins, and we did not impose any additional filters. We clarify this in the first paragraph of results in our revised manuscript:

From each study, we obtained a list of the unannotated proteins reported to be detected (of any length), together with the PSMs supporting these detections (Supplementary Tables 1-2).

Also, when aiming to identify small proteins, the two peptide rule mentioned needs to be taken with more caution, as often, few or even only one peptide are within the visible range of the mass spectrometer.

We fully agree with the reviewer and added a note on this to the introduction of our revised manuscript:

For example, it can be impossible to observe multiple unique supporting peptides for microproteins whose sequence is too short to hold multiple cleavage sites, or if only one peptide is within the mass-over-charge range of the spectrometer.

I was not sure if data for all novel ORFs are provided, i.e., how many distinct peptides and PSMs imply these novel ORFs (overall 9414 peptides, line 111)?

For each novel ORF, the peptides and PSMs supporting the ORF according to the original studies was taken from the supplementary data of each study. This information can be tracked back to the original study using Supplementary Table 2, which describes where we obtained data for each study. The reevaluation we performed for this manuscript caused us to reassess the level of support provided by each evaluated PSM. Supplementary Table 4 lists the evaluated PSMs, the rating assigned to the PSM by the evaluator, and the ORFs the PSM was claimed to support in the original study. Using these two tables, the number of distinct peptides and PSMs supporting each ORF according to both the original studies and our reanalysis can be found.

Looking at table 1, it becomes apparent that different genome sequences and annotation releases have been used. These complicate tracking the genome coordinates of novel sORFs (exemplified by the fact that for 3-4 studies, the authors could not generate them), which are highly relevant. Did the authors use one genome sequence as a common reference for their reanalysis or did they rely on the references used by each study?

We thank the reviewer for pointing out that this information was missing. We did use one genome sequence as a common reference, as we now note in the methods section of the revised manuscript:

The coordinates of each ORF with an evaluated peptide were taken from the supplementary data of each study and the ORF length determined. All ORF coordinates were converted to hg38 coordinates using LiftOver.

How would results of different proteogenomics studies be used in the future to further improve a (and which) human reference genome sequence? This is a key difference to proteogenomics e.g. in prokaryotes or other eukaryotic model systems, where the genome sequence of the bacterial strain in question can be readily sequenced and de novo assembled, eliminating this

level of ambiguity. Even in plants like *A. thaliana*, having one reference genome simplifies the analysis substantially, compared to e.g. integrating results of studies from human cancer samples.

We agree with the reviewer that relying on a reference, as is common in human proteome studies, poses complications to analysis. In cancer in particular, structural rearrangements may lead to translation of RNAs whose existence is not obvious when looking at the reference genome. Better accounting for variation at the genome level would improve human proteogenomics studies by eliminating the ambiguity noted by the reviewer.

These complications can be addressed in several ways, including 1) using a pan-genome variant graph instead of a single reference genome, or (similar to the prokaryotic case) 2) sequencing all samples and creating custom protein databases, as was done in one of the previously published studies that we analyzed.

We briefly discuss this issue in the manuscript in the context of correctly assigning peptides that map to variants of canonical proteins:

Excluding potential hits to annotated proteins can be done with tools such as ProteoMapper⁴⁵ or the neXtProt peptide uniqueness checker⁴⁷, as suggested by the HUPO-HPP MS data interpretation guidelines²⁷, or, ideally, using sample-specific customized protein sequence databases based on sequenced genotypes.

Reciprocally, one can imagine proteogenomics could guide genome assemblies e.g. in cancer in the future. In summary, we believe that human proteome studies should move closer to the prokaryotic model and rely less on references.

We make a related point in a list of advice we have added to the revised manuscript:

- *Ensure peptides appearing to support a novel protein detection uniquely support that protein:*
 - a. *Conduct a search using tools such as ProteoMapper⁴⁵ or PepQuery⁶⁰ to exclude peptides with possible matches to canonical proteins, including post- and co-translational modifications and common genetic variants. When possible, construct a sample-specific protein database that accounts for genotype. Do not assume a canonical protein is absent from the sample solely on the basis of gene transcription or translation evidence.*

2. When attempting to identify novel human microproteins, using a 1% FDR rate at peptide/protein level is by no means stringent enough and not adequate; as the authors state, these thresholds are used for identification of proteins present in an annotated reference database (e.g. UniProt or NCBI RefSeq). In reference 16 the authors cite (Nesvizhskii AI, Nature Methods 2014), who has elaborated on the needs of more stringent FDRs, as have early proteogenomics studies in prokaryotes, which advocated for using more stringent cut-offs (PMID: 22114679), at times going even down to 0.01% FDR at the PSM level (PMID 29141959).

I was thus rather surprised to find that these insights have apparently not translated into human proteogenomics (yet), and that only few of the 12 studies worked with local FDRs and several still used a 1% PSM level FDR threshold. This would certainly make a set of guidelines desirable if not mandatory.

We appreciate the point made by the reviewer that FDR threshold practices have been different in the prokaryotic MS community than for human proteomics, and this may indeed explain in part why some studies in human proteomics report low quality PSMs.

We included a point about FDR on a list of advice we offer at the end of the manuscript:

- *To accurately convey confidence in the list of unannotated protein detections, report local FDRs or FDR specific to the list unannotated proteins instead of or in addition to proteome-wide global FDR. The less stringent the FDR threshold used, the more it is necessary to examine candidates further to ensure they are correct.*

We are reluctant to simply state that a specific FDR is needed, because different FDR thresholds might be appropriate for different purposes.

In my view, the Douka study should be removed from the set; a 10% peptide level FDR (proteome-wide) should not qualify to be included in such a comparison.

We agree that a 10% peptide level FDR is too high. However, we believe it is important to include all the studies meeting our predetermined criteria in the meta-analysis, and it is informative to see the full range of strategies employed, even if some of these strategies have problems.

3. Lots of work in the proteogenomics field has been done on prokaryotic organisms, where proteomics data has been searched against six frame translations of the genome sequence or more sophisticated custom search databases. While readily feasible for prokaryotes, this is more complicated for the substantially more complex eukaryotic genomes. A major advantage though is the availability of genomic coordinates of all novel sORFs. The lack of this data (in some cases) seems a complication of proteogenomics in humans as the different studies seem to have used different reference genome sequence data (please correct me if wrong here) and for sure different annotation releases (e.g. Human UniProtKB 2017-2021), etc.

The reviewer is correct that the different studies have relied on different genome annotations. The lack of sORF coordinates is not a fundamental technical limitation. Rather, published papers sometimes simply do not include this data. If the coordinates are given, we can use them even across different assemblies by converting to a common reference using LiftOver. To address the issue of studies using different proteome releases, we checked whether every peptide claimed to support a novel protein detection mapped to a protein that was present in UniProtKB/Swiss-Prot. in 2016: :

We determined whether each peptide mapped to a human annotated protein according to the 2023 build of the PeptideAtlas database⁶⁵ and whether each peptide mapped to a protein present in the 2016 version of UniProtKB/Swiss-Prot.¹⁸

In practice, this turns out not to be a major issue for this particular set of studies, because few of the novel proteins claimed to be detected in these studies have become annotated between 2016 and 2023, as we reported in both the initial submission and revised manuscript:

Only eight distinct annotated proteins matching reported unannotated peptides in 2023 were absent from UniProtKB/Swiss-Prot in 2016, indicating that annotation updates are not a major explanation for peptides reported to support unannotated proteins mapping to annotated proteins.

We agree that the use of different genome and protein databases pose complications to proteogenomic analysis, and that these issues are more challenging in humans than in prokaryotes. We believe we have addressed these complications in this work.

4. Ribo-seq is indeed extremely useful as a basis to create an extended custom search database to find evidence for so far unannotated human microproteins by mass spectrometry. Mass spectrometry can help in identifying stable protein products; I believe the aspect of pervasive translation may even be stressed a bit more.

For the Ribo-seq approach, one must keep in mind that due to the higher sensitivity, a good fraction of the signals will represent proteins that get rapidly degraded and may not be found to be stably expressed. At least for *M. tuberculosis* H37Rv, a key bacterial model organism, pervasive translation was observed with Ribo-seq for over 2000 unannotated ORFs (compared to roughly 4000 annotated CDS), only a minor fraction of which (90) showed evidence of purifying selection which was interpreted as evidence that they represent bona fide missed proteins (PMID: 35343439). Currently, several studies have employed both Ribo-seq and MS-based proteomics and show complementary sets of identifications.

We strongly agree about the value of mass spectrometry in identifying stable protein products in the face of pervasive translation. This was a major motivation for this manuscript.

We also agree that a good fraction of ribo-seq identified translated sequences may be poorly conserved, lack biological benefit, and get quickly degraded, and emphasize these points further in the discussion section of our revised manuscript:

Why do immunopeptidomics studies identify many high-quality PSMs supporting unannotated protein detections while studies using conventional enzymatic digests identify only few? Many unannotated sequences found to be translated by Ribo-Seq lack signatures of evolutionary conservation and may not encode proteins that provide any benefit to the organism.^{5,15,57} It is plausible that many of these poorly conserved proteins

are expressed but quickly degraded, and so can be found only as peptides bound to HLAs.^{14,58}

We now cite PMID 35343439 in this paragraph (reference 57), as well as similar findings in eukaryotes, to emphasize that many unannotated coding sequences lack signatures of selection.

5. The authors may want to provide a rough estimation of the number (or percentage) of sORFs that can be expected to be found in the human genome. In prokaryotes, studies in *E. coli* uncovered that 140 sORF encoded proteins (SEPs) below 50aa could be identified from 2010-2020 (see PMID: 32385980), i.e. roughly 3% when assuming ~4300 annotated CDS. With a cut-off of 100aa, this number would undoubtedly be substantially higher and amount to several percent of annotated CDS.

Here, based on the subset of PSMs analyzed (the authors should add what %age of all PSMs implying novel human proteins these represent), the rough extrapolations shown in Figure 2G and the fact that the 2021 Ouspenskaia dataset with the largest number of novel human microproteins predicted (4900) fared quite well in terms of the spectral quality assessment (but still suffers from a too lenient 1%PSM level FDR threshold), it appears that this could include hundreds to above 1000-2000 thousand novel human microproteins.

We thank the reviewer for pointing out that the reanalyzed percentage of all PSMs implying novel human proteins was missing for our submitted manuscript. 406 PSMs were selected out of 31,179 reported across the studies considered (1.3%). In the revised manuscript we now note this percentage:

A total of 406 PSMs from 12 studies were evaluated (1.3% of total), corresponding to 307 peptides from 204 unannotated proteins.

The extrapolation in Figure 4G is our best estimate of the number of presently unannotated microproteins supported by high quality PSMs within the specific set of studies that were published within our study window. We predict 3,706 proteins are supported by high quality immunopeptidomics PSMs in the studies we analyzed and 137 by high-quality PSMs in conventional, shotgun proteomics experiments. Dividing by around 20,000 human canonical proteins, this would imply that hypothetical microproteins supported by at least one high-quality immunopeptide PSM amount to around 19% of annotated proteins, and microproteins supported by at least one high quality conventional proteomics PSM is around 0.7% of annotated proteins.

We are reluctant to say that this means there are thousands of unannotated microproteins because, in our view, confidence in a protein requires more than a single high-quality PSM from a shotgun MS experiment. We view the work presented in our manuscript as one part of the picture towards discovery of novel small proteins, which would also include evolutionary analysis, genetics experiments, Ribo-seq, and more.

We note this in the discussion section of the initial submission and revised manuscript:

It is important to note that false positives can occur across the full range of PSM quality; a low-quality spectrum does not prove that a claimed detection is a false positive; nor is a high-quality spectrum conclusive evidence of detection. The gold standard for rigorous MS-based proteomics data validation requires demonstration that a synthetic peptide generates the observed spectrum and is retained on the liquid chromatography column to the same extent as the originally detected peptide, and that the endogenous spectrum is eliminated when the ORF is disabled genetically. Supporting evidence for the biological significance of a protein with inconclusive MS support can also come from outside proteomics, such as by demonstrating the evolutionary conservation of its amino acid sequence or reporting phenotypic impacts upon genetic perturbations.^{23,53}

For the revised manuscript, we now also cite PMID 32385980 in the introduction.

6. I was missing a take home message concerning manual evaluation of MS/MS spectra versus the ability to rely on in silico predicted spectra. This will majorly impact the ability to carry out larger scale analyses and become independent from expert evaluators.

In our opinion, unfortunately the current state of the field still requires manual evaluation for a sample of the discoveries. We added a list of advice which includes this point in our revised manuscript:

- *Manual evaluation of a representative subset of PSMs is important to ensure reported detections are supported by high quality evidence.*

7. Most helpful for the community would be a set of recommendations for an approach to identify novel human microproteins following a set of guidelines/best practices, such that obvious and avoidable errors can be minimized/eliminated (selecting too lenient FDRs at the spectrum, peptide or protein level, not considering local FDR among novel proteins, etc.).

What would the authors recommend in terms of reference genome sequence and latest annotation release (Table1 lists Human UniPROT releases from 17 to 21; do their protein sequences also carry genome coordinates to locate them on the reference genome)? How can studies from cancer samples be integrated or should they rely on separate genome assemblies? What kind of tools can be employed to minimize the need for manual evaluation of MS/MS spectra. What kind of analyses can be carried out to prioritize among the list of hundreds/thousands of potentially novel human microproteins to focus on the most relevant ones?

We agree with the reviewer that this would improve the manuscript, and have added a list of recommendations at the end of the manuscript that cover the issues raised in this point:

Box 1: Advice for detection of novel microproteins using mass spectrometry-based proteomics

- *Ensure peptides appearing to support a novel protein detection uniquely support that protein:*
 - a. *Conduct a search using tools such as ProteoMapper⁴⁵ or PepQuery⁶⁰ to exclude peptides with possible matches to canonical proteins, including post- and co-translational modifications and common genetic variants. When possible, construct a sample-specific protein database that accounts for genotype. Do not assume a canonical protein is absent from the sample solely on the basis of gene transcription or translation evidence.*
 - b. *Consider whether the peptide may come from a previously unannotated isoform of a known protein-coding gene, as gene annotation databases do not comprehensively capture all transcript diversity. Ideally, integrate short- or long-read transcriptomics data to determine whether the evidence supports an unannotated alternative transcript or splicing event that could explain the observed translation.*
 - c. *Pseudogene annotations can significantly impact microprotein discovery. Always check whether the peptide overlaps with a known pseudogene locus from either the Ensembl-GENCODE or RefSeq catalog.*
- *Ensure that the PSMs used to support a novel protein detection are high quality:*
 - a. *Among PSMs that score highly in a search engine, spectra match quality can be further supported by comparison to experimental spectra generated from synthesized peptides, comparison to in silico fragmentation spectra generated by methods such as Prosit⁶¹ or MS2PIP,⁴⁸ and machine learning rescoring using approaches such as Oktoberfest⁵⁰ or MS2Rescore.⁶²*
 - b. *Manual evaluation of a representative subset of PSMs is important to ensure reported detections are supported by high quality evidence.*
 - c. *To accurately convey confidence in the list of unannotated protein detections, report local FDRs or FDRs specific to the list of unannotated proteins instead of or in addition to proteome-wide global FDR. The less stringent the FDR threshold used, the more it is necessary to examine candidates further to ensure they are correct.*
- *Make the MS data available in a public data repository. Report universal spectrum identifiers (USIs)⁶³ for all spectra supporting discovery of a novel protein.*

We additionally added more detailed guidelines to the revised manuscript, based on the Human Proteome Project guidelines, in Appendix 2.

Reviewer #3 (Remarks to the Author):

More and more sORFs and their encoded peptides were identified by Ribo-Seq and Mass spectrometry. The MS-based peptidomics is a reliable approach to finding novel microproteins. It's very important to figure out the issue with the detection quality and help find more functional microproteins. Overall, the manuscript is well written, with only a few minor issues that need to be addressed:

We thank the reviewer for their kind words and helpful feedback.

1. What are the criteria for collecting the human unannotated protein dataset? Only publish between 2019 and 2022, or any other conditions?

Those are the only conditions. We see how the wording at the beginning of results was potentially misleading in the initial submission, so we changed wording slightly for the revised manuscript:

To evaluate the extent to which unannotated proteins can be detected in proteomics data, our group of microprotein researchers assembled in 2023 to conduct a literature search for all papers reporting human unannotated protein detections published between 2019 and 2022. We identified 12 such studies published in this time window (Table 1).

2. The authors emphasize the unannotated microproteins and focus on the short open reading frame-encoded peptide. The dataset includes both immunopeptides and non-HLA peptides. The immunopeptides differ from sORF-encoded peptides and may come from known large proteins. The authors should explain more about "unannotated microproteins" and tell the difference between peptides from various sources.

We thank the reviewer for highlighting the need to clarify these questions in our revised manuscript. Both the immunopeptides and non-HLA peptides were reported by the studies in our meta-analysis to derive from unannotated microproteins. For both, these claims could be correct, or the peptides could be derived from known proteins. To clarify any ambiguity, we have added a sentence at the beginning of results in the revised manuscript:

To evaluate the extent to which unannotated proteins can be detected in proteomics data, our group of microprotein researchers assembled in 2023 to conduct a literature search for all papers reporting human unannotated protein detections published between 2019 and 2022. We identified 12 such studies published in this time window (Table 1). Seven studies searched for unannotated proteins in conventional proteomics data, while two studies searched for peptides derived from unannotated proteins in immunopeptidomics data, and three studies searched both classes of proteomics data.

We have also elaborated in introduction that the spectra produced in a proteomics experiment are derived from both annotated and unannotated protein sources:

In both conventional proteomics experiments and immunopeptidomics experiments, the collected spectra will generated from peptides derived from both annotated and unannotated proteins in the sample. Confident inference of an unannotated protein detection requires that the peptide uniquely supports an unannotated protein; i.e., that we can exclude the possibility that it derives from a protein in a curated protein sequence database.

3. "A key motivation for initiating this community effort was the large variation in the number of validated unannotated proteins reported between studies, ranging from 6 to 4,903". The variation may be due to the sample size; for example, some papers only use one cell line, while others analyze plenty of tissue or cell samples.

We fully agree with the reviewer that the sample size is a factor that affects the number of unannotated protein detections. In our revised manuscript, we have edited this section to point this out:

We do not interpret the high variability between studies as indicating that most reported detections are false: this high variability among reported detected peptides likely reflects in part the high variability in the size and composition of the sORF databases tested (Table 1)¹⁶ and the quantity of proteomic data analyzed, as well as the diversity of cell types examined, MS techniques used, HLA allotypes among the immunopeptidomics studies, and search algorithms. Nevertheless, in the absence of robust replicability to establish confidence, a closer assessment of the strength of evidence provided in each study for their reported detected unannotated proteins is needed.

The author may do further statistical analysis to see how many sORF-encoded peptides can be identified in a single run or a single cell/tissue sample for every research.

We agree with the reviewer that this analysis would need to be done to fully understand why studies vary in the number of reported detections. Our primary interest in this manuscript, however, is in assessing PSM quality across studies for claimed unannotated protein detections. This explains our study design, where we evaluated only a random subset of the PSMs for each study. For the larger studies, we directly evaluated only a small proportion of the PSMs reported, therefore many individual samples will have few or no PSMs evaluated. As a result, our study design unfortunately does not allow us to systematically assess how many sORF-encoded peptides can be identified in a single sample.

By the way, for reference 28, the sORFs number is 4903 in Table 1. But the original paper mentioned "We constructed a high-confidence database of translated nuORFs across tissues (nuORFdb) and used it to detect 3,555 translated nuORFs from MHC-I immunopeptidome mass spectrometry analysis" and "retaining 6,501 high confidence (FDR<1%) peptides from 3,261 nuORFs, across various nuORF types". So, could the authors explain where the number 4903 comes from?

Thank you for pointing out this apparent inconsistency. As stated in Supplementary Table S3 of the submitted and revised manuscript, we took PSMs from Supplementary Tables S3, S6, S8, S9, S12 of Ouspenskaia et al. 2021 (ref 28). Each PSM is associated with an ORF identifier, and the total number of unique ORF identifiers for all PSMs across these supplementary tables in Ouspenskaia et al. 2021 is 4,903. This number is larger than the 3,261 number cited in the quote selected by the reviewer, which corresponds to Supplementary Table S3 of Ouspenskaia et al. 2021 only (to the exclusion of their tables S6, S8, S9, S12).

4. For the dataset that 96% matched to annotated protein, most of which are tryptic peptides. Can not only say it's because they use a custom database of specific samples. Have any of those annotated proteins been proven not to be expressed in those samples? Can they show one example? Otherwise, if the peptide is not a unique peptide, then the identified peptide can not be used as existing evidence of a sORF-encoded peptide.

We believe this point by the reviewer is exactly what we intended to convey in the initial submission. We fully agree that “if the peptide is not a unique peptide, then the identified peptide cannot be used as existing evidence of a sORF-encoded peptide.” The authors (Duffy et al. 2022) tried to avoid the problem by creating custom databases of all the proteins in the sample expressed using Ribo-Seq. But, as suggested by the reviewer’s questions, an inability to find that a CDS is translated by Ribo-Seq does not prove that the protein it encodes does not exist in the sample. We think this is likely a flaw in the study by Duffy et al. 2022. In the initial submission, we tried to convey these points concisely but we understand from the reviewer’s question that our original writing failed to be sufficiently clear. For the revised manuscript, we have made some edits that we hope improve clarity:

For Duffy et al. 2022³⁴, spectra searches were conducted against custom databases of both annotated and unannotated proteins inferred to be expressed in the specific type of brain tissue or cell based on Ribo-Seq data, while all other studies included the full set of human annotated proteins in their protein database. Likely, annotated proteins not detected by Ribo-Seq may still be present in the sample, leading to peptides from annotated proteins potentially being falsely assigned to unannotated proteins.

5. What is the main reason that makes HLA peptides different from non-HLA peptides, sample preparation, database search, or the characteristics of the peptides? Please have a more in-depth discussion on this based on publications. Is there any suggestion for improving the identification of non-HLA peptides?

We very much share and appreciate the interest of the reviewer in better understanding the distinction between HLA and non-HLA data. It is true, as suggested by the reviewer, that sample preparation and the characteristics of the peptides do differ between HLA studies and non-HLA studies, but are these the factors that matter, or others, such as the type of statistical analyses performed? We are trying our best to interpret this difference but we are confined within the limits of our post-hoc, observational study design. Unfortunately, too many factors differ systematically between the HLA and non-HLA data, so our post-hoc study design does not

allow for clear hypothesis testing. To avoid making any claim that we cannot back up with strong evidence, we regret to conclude that neither our study, nor any other study that we are aware of, gives a clear answer to this question.

Nevertheless, we provide plausible explanations in the discussion section. :

Why do immunopeptidomics studies identify many high-quality PSMs supporting unannotated protein detections while studies using conventional enzymatic digests identify only few? Many unannotated sequences found to be translated by Ribo-Seq lack signatures of evolutionary conservation and may not encode proteins that provide any benefit to the organism.^{5,15,57} It is plausible that many of these poorly conserved proteins are expressed but quickly degraded, and so can be found only as peptides bound to HLAs.^{14,58} However, there are also technical explanations for why HLA-bound peptides derived from unannotated microproteins may be easier to detect. Immunopeptidomics concentrates peptides bound to HLAs, which decreases sample complexity and may thereby enrich for low abundance microproteins. HLA peptides also have physico-chemical properties different from tryptic peptides that may affect detectability. Most immunopeptidomics datasets are from cancer samples, and some proteins may be expressed in some cancers but not in normal physiological conditions. Furthermore, microproteins may preferentially reside in cellular compartments that are hard to sample through non-HLA MS, such as membranes.²⁶ Moreover, the laboratories that perform immunopeptidomics are often distinct from those that analyze non-HLA data and may differ in their sample preparation techniques, experimental setup, and analytical choices. Understanding which factors are most important to explaining the difference between immunopeptidomics and conventional shotgun proteomics may require the development of more sensitive proteomic techniques for identifying low-abundance and short-lived microproteins in the cell.

In our view, the reason immunopeptidomics studies are able to find so many more peptides from unannotated proteins remains one of the major open questions in the human microprotein field.

6. Supplemental Table 3 is hard to search for information in the TXT form, maybe switching it to CSV form and putting the publications of each peptide in different columns.

We remade Supplementary Table 3 following this advice: it is now a CSV file where each column corresponds to a study, each row corresponds to a peptide, and the cell indicates whether the peptide was found in the study.

7. Page 9 said "A total of 406 PSMs from 12 studies were evaluated.....Additionally, a common set of 10 negative control PSMs was included in each sample...". However, there are 439 mzspec and 15 control spec in Supplemental Table 4. Please double-check the number.

We thank the reviewer for checking so carefully and pointing out this discrepancy. The cause is that some PSMs were given a rating of NA by evaluators when they were unable to evaluate the

PSM. In particular, for a few USIs there were issues preventing the PSM from being properly displayed on the website. A total of 424 distinct non-control PSMs were given for evaluation of which 406 were given an actual non-NA rating. If controls and NA ratings are excluded from Supplemental Table 4 in the initial submission, the correct number of 406 is obtained. In the revised manuscript, we have added an explanation of this process in the methods section:

PSMs for each study were evaluated by a group of six expert evaluators. Each evaluator rated a random sample of PSMs from each study. A total of 424 PSMs from 12 studies were given for evaluation, out of which 406 were given ratings, as a few PSMs could not be displayed from the input USI. Out of the 406 PSMs evaluated, 155 were evaluated by two evaluators each to enable determination of the overall consistency between evaluators. Evaluations were done by visual inspection of the PSM using the ProteomeCentral USI web application (<https://proteomecentral.proteomexchange.org/usii/>) in May to June 2023. The evaluators were told to use no other information except the PSM as displayed on the USI application. A common set of 10 negative control PSMs was given to each evaluator; the evaluators were not informed of the existence of these controls. These negative controls consisted of high-scoring decoy-spectrum matches manually selected from among the strongest 30 decoy-spectrum matches in Duffy et al. 2022.³⁴ Each PSM was rated on a scale of 1-5; the rating scale is given in Appendix 1.

As for the 15 negative control PSMs in the initial submission supplemental table, this was an error on our part in outputting the table. There were only 10 negative controls. For the revised manuscript, we have removed the incorrect negative control entries from Supplemental Table 4. The error did not go into any analysis or affect how the non-control PSMs were reported in the supplemental table. We again thank the reviewer for catching this error and have double-checked to ensure the supplemental tables associated with the revised manuscript are correct.

8. The information in Supplemental Table 5 is hard to follow. Could the author provide the expression data of each sORF related to Figure E?

Supplemental Table 5 is simply providing a list of accession identifiers of all Ribo-Seq experiments that went into the analysis. To provide the expression data of each sORF related to Figure E, we constructed a new supplemental table, Supplemental Table 6, which we reference alongside Supplemental Table 5 in our revised manuscript.