Article

Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1

DEUTSCH, Eric W, *et al*.

Abstract

Every data-rich community research effort requires a clear plan for ensuring the quality of the data interpretation and comparability of analyses. To address this need within the Human Proteome Project (HPP) of the Human Proteome Organization (HUPO), we have developed through broad consultation a set of mass spectrometry data interpretation guidelines that should be applied to all HPP data contributions. For submission of manuscripts reporting HPP protein identification results, the guidelines are presented as a one-page checklist containing fifteen essential points followed by two pages of expanded description of each. Here, we present an overview of the guidelines and provide an in-depth description of each of the fifteen elements to facilitate understanding of the intentions and rationale behind the guidelines, both for authors and for reviewers. Broadly, these guidelines provide specific directions regarding how HPP data are to be submitted to mass spectrometry data repositories, how error analysis should be presented, and how detection of novel proteins should be supported with additional confirmatory evidence. These [...]

Reference

DEUTSCH, Eric W, *et al*. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *Journal of proteome research*, 2016, vol. 15, no. 11, p. 3961-3970

DOI : 10.1021/acs.jproteome.6b00392
PMID : 27490519

Perspective

Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1

Eric W. Deutsch, Christopher M Overall, Jennifer E. Van Eyk, Mark S. Baker, Young-Ki Paik, Susan T. Weintraub, Lydie Lane, Lennart Martens, Yves Vandenbrouck, Ulrike Kusebauch, William S. Hancock, Henning Hermjakob, Ruedi Aebersold, Robert L. Moritz, and Gilbert S Omenn

J. Proteome Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.jproteome.6b00392 • Publication Date (Web): 04 Aug 2016

Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1

Eric W. Deutsch[1,*], Christopher M. Overall[2], Jennifer E. Van Eyk[3], Mark S. Baker[4], Young-Ki Paik[5], Susan T. Weintraub[6], Lydie Lane[7], Lennart Martens[8,9], Yves Vandenbrouck[10], Ulrike Kusebauch[1], William S. Hancock[11], Henning Hermjakob[12,13], Ruedi Aebersold[14,15], Robert L. Moritz[1], and Gilbert S. Omenn[1,16]

[1] Institute for Systems Biology, Seattle, WA, USA
[2] Centre for Blood Research, Departments of Oral Biological & Medical Sciences, and Biochemistry & Molecular Biology, Faculty of Dentistry, University of British Columbia, Vancouver, Canada
[3]Advanced Clinical Biosystems Research Institute, Department of Medicine, Cedars Sinai Medical Center, Los Angeles, CA, USA
[4] Department of Biomedical Sciences, Faculty of Medicine and Health Science, Macquarie University, NSW, Australia
[5] Yonsei Proteome Research Center and Department of Biochemistry, Yonsei University, 50 Yonsei-ro, Sudaemoon-ku, Seoul, Korea
[6] The University of Texas Health Science Center at San Antonio, San Antonio, Texas, USA
[7] SIB Swiss Institute of Bioinformatics and Department of Human Protein Science, Faculty of medicine, University of Geneva, CMU, Michel Servet 1, 1211 Geneva 4, Switzerland
[8] Department of Medical Protein Research, VIB, Ghent, Belgium
[9] Department of Biochemistry, Ghent University, Ghent, Belgium
[10] French Proteomics Infrastructure, Biosciences and Biotechnology Institute of Grenoble (BIG), Université Grenoble Alpes, CEA, INSERM, U1038, Grenoble, France.
[11] Department of Chemical Biology, Northeastern University, Boston, Massachusetts, USA
[12] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
[13] National Center for Protein Sciences, Beijing, China
[14] Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
[15] Faculty of Science, University of Zurich, 8006 Zurich, Switzerland
[16] Departments of Computational Medicine & Bioinformatics, Internal Medicine, and Human Genetics and School of Public Health, University of Michigan, Ann Arbor, MI, 48109-2218, USA

*Address correspondence to: Eric W. Deutsch, Institute for Systems Biology, 401 Terry Ave N, Seattle, WA 98109, USA, Email: edeutsch@systemsbiology.org, Phone: 206-732-1200, Fax: 206-732-1299

Keywords: Guidelines, standards, Human Proteome Project, mass spectrometry, false-discovery rates, alternative protein matches

Abstract

Every data-rich community research effort requires a clear plan for ensuring the quality of the data interpretation and comparability of analyses. To address this need within the Human Proteome Project (HPP) of the Human Proteome Organization (HUPO), we have developed through broad consultation a set of mass spectrometry data interpretation guidelines that should be applied to all HPP data contributions. For submission of manuscripts reporting HPP protein identification results, the guidelines are presented as a one-page checklist containing fifteen essential points followed by two pages of expanded description of each. Here, we present an overview of the guidelines and provide an in-depth description of each of the fifteen elements to facilitate understanding of the intentions and rationale behind the guidelines, both for authors and for reviewers. Broadly, these guidelines provide specific directions regarding how HPP data are to be submitted to mass spectrometry data repositories, how error analysis should be presented, and how detection of novel proteins should be supported with additional confirmatory evidence. These guidelines, developed by the HPP community, are presented to the broader scientific community for further discussion.

Introduction

The flagship scientific project of the Human Proteome Organization (HUPO), known as the Human Proteome Project (HPP), is composed of 50 teams of scientists organized as the Chromosome-Centric HPP (C-HPP), the Biology and Disease-driven HPP (B/D-HPP), and the three resource pillars for Antibodies, Mass Spectrometry, and Knowledge Bases. The HPP is an international effort to advance the understanding of all aspects of the human proteome. Its initial primary aim is to develop a full "parts list" of proteins that are present in human cells, organs and biofluids. Beyond, the HPP aims to advance our understanding of protein interactions and functions in health and disease, and enable the widespread use of proteomics technologies through enhanced techniques and resources by the broader scientific community[1]. One of the major goals for the C-HPP in establishing the full parts list is to obtain conclusive mass spectrometry (MS) evidence for what are termed "missing proteins"—the set of polypeptide sequences predicted to be translated from the genome and transcriptome, but for which there is not yet sufficient high-stringency evidence that such translation takes place[2–4]. The conclusive detection of these missing proteins, which are specified as having a PE (protein existence) designation of 2, 3, or 4 in the neXtProt[5] knowledge base, as well as reported translation products from novel coding elements, requires compelling evidence. This includes an interpretation that clearly takes into account the inherent uncertainties currently found in high-throughput MS data acquisition techniques and sequence matching to still-evolving protein reference databases.

MS proteomics is a powerful technology that has enabled routine high-throughput identification and quantification of proteins in complex samples. There are several different MS techniques, including shotgun proteomics via data-dependent acquisition (DDA)[6,7], data-independent acquisition (DIA) (e.g., SWATH-MS[8]), and targeted proteomics via selected reaction monitoring (SRM; sometimes called multiple reaction monitoring, MRM). Each has different capabilities and strengths that can be brought to bear, depending on the goals of the analysis. Although many variations exist, a typical workflow involves

extracting and fractionating proteins from a sample, cleaving proteins into peptides using a protease such as trypsin, fractionating the obtained peptides to reduce complexity through methods such as liquid chromatography, and then introducing these fractionated peptides as charged ions into a mass spectrometer, typically by coupling chromatography to an electrospray device. The resulting peptide ions are subsequently fragmented in the instrument and spectral data of these fragments are recorded.

The data generated from the mass spectrometer are then subjected to extensive computational analysis to determine which peptide ions likely yielded the observed fragment ions, along with confidence metrics for identification and abundance measures[9]. There is a wide variety of informatics tools available for these data analysis tasks, both commercial and free and/or open source[10]. However, most of these tools are specific to only one type of MS technique. Confidence metrics reported by these tools are a crucial component of the data analysis because different approaches, instruments and analysis parameters result in different inherent uncertainties in data interpretation. These confidence metrics should be calculated at the peptide-spectrum-match (PSM) level, the aggregated peptide level, and the aggregated protein level, both at a global experiment level and individually. These confidence metrics must then be carefully considered when performing downstream interpretation of the results and functional validation of missing proteins.

Every data-rich community research effort requires a clear plan to ensure that data are of high quality and comparable between analyses. Over the years, several sets of guidelines have been developed in the field of proteomics, including those from within HUPO. Each set of guidelines has been distinct in its focus and goals; no single set of guidelines is applicable to all goals. The Minimum Information About a Proteomics Experiment (MIAPE) guidelines[11] developed by the HUPO Proteomics Standards Initiative (PSI)[12] focus specifically on the metadata annotation of experimental MS results. These metadata must describe what was done to execute the experiment with sufficient detail that the results may be properly interpreted or reproduced; MIAPE explicitly does not stipulate how an analysis is to be performed.

Several journals have developed their own guidelines, notably the *Journal of Proteome Research* (JPR) (http://pubs.acs.org/paragonplus/submission/jprobs/jprobs_proteomics_guidelines.pdf), *Molecular and Cellular Proteomics* (MCP)[13,14], and *Proteomics Clinical Applications* (http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291862-8354/homepage/ForAuthors.html#exp). These guidelines specify what information must be included in a submitted manuscript, as well as some basic expectations about how the acquired MS spectral data are interpreted. Three tiers of guidelines for targeted quantitative workflows, depending on the purpose of the assay, have been proposed by an NIH-NCI working group;[15] another group has benchmarked and proposed a set of guidelines specifically for proteogenomics efforts[16,17]. These latter two sets provide significant guidance on how an analysis should be performed, quite unlike the MIAPE approach, which requires extensive disclosure on whatever analysis was performed to be fully compliant.

In 2012, the HPP posted an initial version of guidelines (available at http://www.thehpp.org/guidelines) that focused primarily on ensuring that all data generated and published as part of the consortium effort were deposited to one of the ProteomeXchange Consortium[18] repositories for proteomics data, or another suitable repository for other data types. A further requirement of the HPP version 1.0 guidelines required an analysis threshold of no more than 1% false discovery rate (FDR) at the protein level. Despite having served the HPP well for the past three years, it has been recognized that the pursuit of confident identification of missing proteins, in particular, and also claims of novel translation products from long-non-coding RNAs or pseudogenes, required an updated set of guidelines with more stringent criteria[19].

A new set of guidelines, the HPP Mass Spectrometry Data Interpretation Guidelines Version 2.1, has, therefore, been developed and discussed among the HPP community members to address the stringency of data required to identify missing proteins or novel coding elements. These guidelines are intended to be applied to identification results rather than quantitative results. The guidelines are presented as a one-page checklist followed by two pages of expanded descriptions for each of the fifteen items in the checklist

(See Supplementary Material for this document). In this article we describe the development of the guidelines, and we provide a deeper discussion on the reasoning behind these guidelines. We also provide examples of common missteps seen in submitted manuscripts that prompted the development of the guidelines. These guidelines have been adopted as a requirement for articles that will be published as part of the HPP, and now are offered to the community for discussion and potential adoption elsewhere either in whole or by incorporation into other guidelines.

Development of the Guidelines

A set of preliminary guidelines and discussion points was brought to the HPP Bioinformatics Workshop at the 14th HUPO World Congress held in Vancouver, Canada, in September 2015. Each of the items was discussed and additional input collected. The proposed guidelines were also extensively further debated at the Bioinformatics Hub (Mohammed et al., in preparation) (http://www.psidev.info/hupo2015-bioinformatics-hub) at the same Congress. This informal venue and the post-Congress HPP Workshop enabled additional hours of discussion and refinement of the individual points. Following HUPO-2015, the proposed elements were written into a draft guidelines document consisting of a one-page checklist followed by two pages of additional detail about each of the checklist items. The document was circulated among the HUPO and HPP leadership, and further edited and refined. The document was then approved by the HPP and HUPO Executive Committees.

Version 2.0 of the guidelines was released on November 12, 2015 at http://thehpp.org/guidelines and at www.c-hpp.org. After this release, minor clarifications to the wording were applied in versions 2.0.1 and 2.0.2. Further minor clarifications in the wording were applied during the preparation of this article, resulting in version 2.0.3. These guidelines are in effect for all HPP papers submitted on or after November 12, 2015, and are specifically applicable for all manuscripts for the JPR C-HPP 2016 Special Issue. In response to the review of this article, guideline 2 was changed in a substantial manner, as

described below, and the guidelines document was updated to version 2.1.0 on July 6, 2016. Future changes to the guidelines will be described at the same URLs above. Wording changes or clarifications that do not change the intended interpretation of the guidelines will only invoke a version change in the third digit (i.e. version x.x.1 to x.x.2). A substantive change to one or more guidelines will increment the minor version digit (i.e. version x.0.x to x.1.0). A major rewrite would increment the major version digit (i.e., version 2.x.x to 3.0.0). The process for making additional changes to these guidelines is as follows: proposed changes are presented to the HPP Executive Committee for discussion and approval or rejection. Approved changes trigger an increment in the version number as described above, and the revised checklist and extended description is posted on the HPP web site, accompanied by announcements to HPP participants and HUPO membership.

Elaboration on the Guidelines

As discussed above, page one of the version 2.1.0 guidelines (available at http://thehpp.org/guidelines) is a checklist of fifteen items, with a check-box provided to signify compliance, and empty space for explanation by authors of any elements of non-adherence to the guidelines. Each box should be checked, marked as N/A (not applicable) for cases where the guideline is simply not relevant to a manuscript, or marked with an asterisk (*) for non-compliance (NC), which must be explained and justified in the space provided, extending to additional pages if necessary. In most instances, there should not be any NC markings. However, if any specific non-compliance is well justified, or if scenarios that were unforeseen by the drafters of these guidelines that prevent adherence do arise, reviewers or editors may make exceptions.

The second part is an expanded description of each of the fifteen items. This section provides additional points of clarification for each item and should be consulted by those not yet familiar with these guidelines.

The final and third part is this article, which provides a full description and discussion of the reasoning behind each guideline. The article should be read by all authors submitting an HPP manuscript, and by those who still have questions about the guidelines that were not answered in the expanded description. It is hoped that this three-tiered approach makes it as easy as possible to comply with the guidelines. Table 1 provides a list of the fifteen guidelines, each described in one or two sentences as listed in version 2.1.0 of the guidelines. Table 1 is provided here for ease of reading, but is not a substitute for the primary checklist available in the Supplementary Material and periodically updated at http://thehpp.org/guidelines.

Discussion of individual Guidelines

1. Complete this HPP Data Interpretation Guidelines checklist and submit with your manuscript.

For ease of use for completion and compliance checking, the checklist is presented as a one-page table. This allows those familiar with the checklist to quickly assess compliance with each item and mark each element appropriately. Each item in the checklist must be checked, marked as N/A, or marked with an asterisk indicating non-compliance that must be further justified. Explanations for N/A entries or any other variances marked with an asterisk must be provided in the Author Comments section. Submission of a completed checklist was a requirement for initiation of peer review for the JPR 2016 HPP Special Issue.

Please note that it will be common for manuscripts to have at least one N/A entry. For example, for an SRM-only dataset, element 12 will likely be N/A, while for a dataset that does not include SRM data element 13 will be N/A. Element 9 will be N/A if there is only a single dataset analyzed. Although full compliance for all applicable elements is generally expected for manuscripts, in rare cases it may be appropriate to allow particular non-compliance. If authors feel that compliance for a particular element is

applicable but not achievable, the element may be asterisked and explained. Reviewers and editors may then consider whether the particular exception request is reasonable or should not be accepted. For example, a reasonable exception for element 2 would be a meta-analysis of 1,000 datasets that are all already publicly accessible in some form, although potentially not found in ProteomeXchange repositories. Element 15 already has a potential exception described; some proteins are so short or their sequences such that only one unique (i.e., proteotypic) peptide may be possible, even when considering multiple enzymatic digests.

2. Deposit all MS proteomics data (DDA, DIA, SRM), including analysis reference files (search database, spectral library), to a ProteomeXchange repository as a complete submission. Provide the PXD identifier(s) in the manuscript abstract and reviewer login credentials.

The 2012 HPP Guidelines were the first to require submission of data through one of the ProteomeXchange consortium repositories[18]. At that time, this was only PRIDE[20–22] for shotgun data, and PASSEL[23], a part of PeptideAtlas[24–26], for SRM data. Compliance was not universally enforced, but data were deposited to ProteomeXchange for most HPP Special Issue articles through 2015. Since the initial guidelines were put into place, the MassIVE and jPOST repositories have joined ProteomeXchange; as a result, there are now four repositories in the consortium. The new iProX repository has expressed interest in joining the ProteomeXchange Consortium. Access to the raw data is essential for the standardized reanalyses by the field. Indeed, reanalysis of MS data by PeptideAtlas and by GPMDB has greatly advanced data quality and comparability of analysis in the field of proteomics and provided insights into the metrics underlying these guidelines.

There are broadly two kinds of submissions supported by ProteomeXchange repositories: “partial” (also called “unsupported”) and “complete” (also called “supported”). While both require the same amount of information (metadata, raw data, and identification results), the key difference is that for a partial

submission, the receiving repository was not able to parse and fully load all of the data. In a complete submission, the identification results and identified spectra are fully loaded and searchable via the repository interface. The reason for this distinction is that there are many available software pipelines, many of which do not use standardized or otherwise common output formats. The repositories may not be able to support the parsing and loading of all possible formats on account of the limited resources available to the repositories. Although complete submissions are most desirable, the partial submission mechanism is supported by the repositories so that everyone can submit their data and results to ProteomeXchange, even if the formats were not fully supported.

With these latest 2015/2016 guidelines, past requirements have been upgraded to that of mandatory complete submission. This means that results must be submitted in a format that can be parsed by the receiving repository, and some software tools may have to be excluded because their output cannot (yet) be written or converted to a supported format. Although such a requirement was considered too demanding in 2012 because the PSI mzIdentML format[27] was not universally supported by the repositories at that time, it is now the case that mzIdentML is well supported and widely used. Although not unanimous, the consensus opinion was that complete submission has been widely achievable for some time, and workflows or tools that do not yet produce a suitable output need incentive to support complete submission to ProteomeXchange. For such software that still does not permit complete submissions, we hope that this raising of the bar will accelerate progress in this area. Complete submission is now clearly presented as our long term goal. The HPP decided that a requirement for complete submission would be an important component of such an effort, which has had broad support at the HUPO2015 Congress in Vancouver and throughout the HPP community. If only a minority of submitters are unable to submit data in a complete form, this should put pressure on developers of the software they use to develop solutions to this problem. The HPP special issue editors are willing to be generous in allowing exceptions to this guideline in the near term as we seek solutions for full compliance.

There was considerable debate about the details of the guideline about mandatory complete data submission relative to timing—that is, whether complete data deposition should be required prior to manuscript submission or not. It was generally agreed that submission after acceptance of a manuscript was too late, as this does not give reviewers the opportunity to verify that the submission is appropriate and matches all claims and descriptions found in the submitted manuscript. However, some felt that deposition of datasets prior to initial manuscript submission would place undue burden on repositories, since they are already operating with limited (human) resources, in that they would be forced to handle unnecessary data submissions for any manuscripts destined to be returned without peer review or rejected, leading to pollution of the databases. In versions 2.0.0 through 2.0.5, the current guideline was written so that authors had the option of first submitting their manuscript and waiting to deposit their data until the editors have signaled that the manuscript will be sent for review upon data deposition. In response to the reviewer comments and additional debate by the HPP leadership, this guideline was changed as of version 2.1.0 to reflect the requirement that all data must be deposited prior to submission of the manuscript. This policy is deemed simpler to implement and explain and desirable to expedite review.

3. Use the most recent version of the neXtProt reference proteome for all informatics analyses, particularly with respect to potential missing proteins.

The official reference knowledge base for the HPP is neXtProt[5], and it is crucial that claims of detection of missing proteins and possible translation products from other novel coding elements be compared with the most current neXtProt release, rather than any earlier version. In some cases, this will require re-processing data with an updated reference database prior to final submission of a manuscript. Manuscripts that specially claim detection of missing proteins in their abstract, followed by comments in the discussion section that some are no longer missing in the latest neXtProt release, which was permitted in 2015, are no longer acceptable. For submission to one of the HPP special issues, the call for papers will denote which version of neXtProt, PeptideAtlas and other resources are to be used.

4. Describe in detail the calculation of FDRs at the PSM, peptide, and protein levels.

Each manuscript must describe the methods that were used to calculate the false discovery rates at the PSM, distinct peptide, and protein levels. Importantly, no specific method is prescribed. Some methods or tools can calculate all three levels at once, while in some cases multiple tools must be used. Use and citation of existing tools is encouraged. It is not sufficient to state only "FDRs were calculated using tool X." Software versions, input parameters, apparent anomalies, input file formats, and output formats must all be specified. Any variances or modifications to a previously published methodology should be described. If custom, novel, or unpublished methods are used, they should be described in detail. If such novel or unpublished methods are used, then the results should be compared in some way with results from a more conventional analysis that has been previously published. Note that any assumptions should be clearly stated. Be specific about the distinction between a global FDR (the fraction of incorrect entities among all entities that pass the threshold) and a local FDR (the fraction of incorrect entities within a subset of entities that share the same score, usually expressed for each entity or for the threshold score in a list). The calculation at the peptide level may differentiate between different mass modifications, or aggregate over multiple modifications, at the discretion of the authors.

5. Report the PSM-, peptide-, and protein-level FDR values along with the total number of expected true positives and false positives at each level.

Based on the methodology described, report global FDR values at each of the three levels. Unless unusual methodology is employed, the PSM-level FDR should be lower than the peptide-level FDR, which should be lower than the protein-level FDR. The larger the dataset, the more extreme these differences become. In addition to the FDRs, report the total number of entities passing threshold at each level, and then also state the expected or estimated number of incorrect entities passing threshold at each level. This means that, in addition to stating that proteins are thresholded at a 1% global FDR, state that, for example, 5,000 proteins pass the threshold and, therefore, there are an estimated 50 incorrect identifications in the list. Some software packages do not report all this information, or even FDRs at each of the three levels.

However, even a simple strategy of counting all the PSMs, distinct peptides, and proteins that pass threshold, counting the corresponding decoys at each level that pass threshold, and entering those values into a spreadsheet to calculate the decoy rates and presumed corresponding false discovery rates would be sufficient.

6. Present large-scale results thresholded at equal to or lower than 1% protein-level global FDR.

Although there is not universal agreement on what the best threshold is, and it may vary based on the intent of the final protein list, the HPP has concluded that the baseline acceptable global FDR for a dataset should be at most 1% at the protein level. Lower than 1% is strongly encouraged. As described above, this will usually mean that the peptide-level and PSM-level FDRs will be far lower than 1% for large datasets. We note that, for some datasets, the local FDR should be the factor that should be used to set the threshold. Consider the extreme case where all identifications can be perfectly discriminated into correct and incorrect populations; in order to achieve a 1% global FDR, one is forced to add known incorrect identifications (with local FDR of 100%), which is clearly not an acceptable strategy. For some very high quality datasets where discrimination is excellent, it may be best to apply a local FDR threshold of 10% (where 1 in every 10 identifications near the threshold are incorrect), even though this may yield a global FDR far lower than 1%.

7. Recognize that the protein-level FDR is an estimate based on several imperfect assumptions, and present the FDR with appropriate precision.

There are many different approaches to estimating FDRs. The most common is the target-decoy approach, followed by a population modeling approach[28–33]. Both approaches make imperfect assumptions that affect the accuracy of the results. Decoys are not representative of all kinds of false positives. For example, identifications may be very nearly correct, but incorrect in one or two residues[34], and there tend to be rather few decoys at very stringent thresholds, leading to problems with small-number statistics.

Consider a hypothetical dataset with 1,010 proteins that pass threshold, ten of which are decoys; one might discard these ten decoys and presume there are another ten incorrect identifications among the remaining 1,000, leading to a 1% FDR. However, the exact scores and occurrence of decoys depends on many details of the exact decoy database used, and there could easily have been 9 or 11 decoys at the same effective threshold. Such a change in a single decoy would then yield a calculated FDR of 0.9% or 1.1% for 9 and 11, respectively. Clearly the precision with which the true uncertainty is known when such few decoys are present cannot be high. Model-based approaches may often fit well to the main part of the population, but may fit less well at the very tail of the distribution where the stringent threshold lies, leading to similar uncertainties. In addition, the model high-confidence tail can vary substantially depending on the mathematical function used for the model. In summary, FDR values in a manuscript should be quoted with appropriate precision; unjustified precision, *i.e.* more than two digits of precision, should be avoided.

8. Acknowledge that not all proteins surviving the threshold are "confidently identified."

It is important that careful FDR estimation is not left behind during subsequent analysis of the protein results. It is inappropriate to proceed with an analysis that treats all remarkable entries (e.g., missing proteins) in the resulting list as "confidently identified" when errors are known to exist in the list. In fact, the total number of remarkable identifications should be compared to the reported number of false positives (guideline 5). For example, if one expects 30 incorrect identifications in a result (such as 1% of 3000 proteins), then a claim of the detection of 10 missing proteins should be treated with great caution. The default hypothesis should be that these never-before-detected proteins (in mass spectrometry) are 10 of the expected 30 false positives. Orthogonal convincing evidence must be presented to rule out (or at least significantly constrain) this default hypothesis. See guidelines 10, 11, and 14.

9. If any large-scale datasets are individually thresholded and then combined, calculate the new, higher peptide- and protein-level FDRs for the combined result.

When several different proteomic (or MS) datasets are compared or combined in a manuscript, it is important to be mindful that the combined results will have a different, usually higher FDR. Consider the three cases in Figure 1. For example, in A, where there is no overlap in the correct proteins and no overlap in the incorrect identifications, the combined FDR is truly the same as in the original datasets. In case B, all of the correct identifications overlap, but the incorrect ones do not (because incorrect identifications usually scatter over the proteome). The combined FDR is twice as high as the original. The third case is a more real-world example where 50% of the correct identifications overlap, and none of the incorrect ones does. The resulting FDR is ~1.5%, which is much larger than the original FDRs. Caution is required with compendia of many experiments that have all been individually processed, thresholded, and then combined, as the false discovery rates will inflate considerably. If all of the data discussed in a manuscript are processed together with a single threshold, then this guideline will not be applicable.

10. Present "extraordinary detection claims" based on DDA mass spectrometry with high mass-accuracy, high signal-to-noise ratio (SNR), and clearly annotated spectra.

The concept of an "extraordinary detection claim" is purposely left somewhat vague in the guidelines. Two obvious examples in this category are missing proteins (predicted proteins lacking PE=1 neXtProt status) and novel coding elements (e.g., lncRNAs, novel exons, pseudogenes, or other sequences not listed in neXtProt as entries with protein existence level 1 through 4). However, authors and reviewers may consider other claims as extraordinary, such as a report of detection of a protein in a sample where the protein would not likely be present and the transcript cannot be detected, such as an olfactory receptor protein in liver.

Several journal guidelines already require annotated tandem mass spectra as supplementary material for single-hit proteins. We have extended this requirement to all extraordinary detection claims, even when supported by multiple peptides. Furthermore, the spectra must be of high signal-to-noise ratio, with a recommendation that the highest 5% intensity peaks should have a signal at least 20 times those of the

lowest 5% intensity peaks, which are presumed to be mostly noise. Although low mass-accuracy (i.e., ion trap) MS/MS spectra are still useful for many applications, MS/MS spectra supporting extraordinary detection claims should be acquired in higher mass-accuracy (Fourier-transform, Orbitrap, TOF, Q-Exactive, etc.) instruments.

11. Consider alternative explanations of PSMs that appear to indicate extraordinary results.

In cases where a peptide identification corresponding to an extraordinary claim appears to have a well annotated, high signal-to-noise ratio spectrum, consider whether a slightly different amino acid sequence that can map to a different, common protein also could be a credible explanation. An example is the case presented in Figure 5 of Deutsch *et al*. [26] of a spectrum in PeptideAtlas that appears to have excellent coverage and, thus, a very high score for olfactory receptor 5A2 (Q8NGI9), with just a few missing and unexplained peaks. However, careful scrutiny reveals a slightly different peptide sequence with an unconsidered mass modification that yields an even better match, and also maps to a very commonly seen protein (and peptide sequence), lactotransferrin (P02788). Such cases may be quite rare, but, among millions of mass spectra, some of the ones that appear to implicate extraordinary results will be cases such as this. This guideline does not require manual inspection of all spectra; rather it applies only to the exceptional case of an extraordinary claim for a previously unreported protein match.

12. Present high mass-accuracy, high-SNR, clearly annotated spectra of synthetic peptides that match the spectra supporting the extraordinary detection claims.

One method for increasing the confidence in the correctness of an identified peptide is to compare the identification with a synthetic version of the peptide (i.e. same charge, same mass modifications, same instrument fragmentation). The synthetic peptide fragment spectra should be shown alongside the naturally-derived peptides, both with high spectrum quality and with similar peak intensity patterns between the natural peptide and the synthetic peptide. A match in chromatographic elution time also is a

strong confirmation, but not sufficient, that the peptide is correctly identified. As in the JPR C-HPP 2015 special issue, the editors may allow stepwise presentation of "candidate missing protein identifications", followed by an explanation of how the candidate fared upon application of these more stringent requirements. Such information may be a guide for others to seek more convincing evidence in the same type of specimen or in another specimen guided by transcript expression data.

13. If SRM verification for extraordinary detection claims is performed, present target traces alongside synthetic heavy-labeled peptide traces, demonstrating co-elution and very closely matching fragment mass intensity patterns.

SRM can be a useful technology to confirm the unambiguous identification of peptides that appear to support the extraordinary claim. Although its sensitivity can be better than conventional shotgun technologies, it is not vastly better and, since fewer ions are often used as evidence, it is imperative that SRM confirmation is performed with the use of spiked-in stable isotope-labeled synthetic peptides. Maximal corresponding fragments (transitions) must be monitored for both heavy and light ions and of predominantly higher mass transitions for better discrimination. The peak intensity order of those ions as well as elution pattern must match with high similarity. Traces down at the detection limit are usually not suitable, as the chance of spurious interferences is high at the detection limit. Furthermore, it is crucial to exclude the possibility of light-peptide contamination in heavy-labeled spike-ins providing spurious signal. For example, if a spike-in reference sample of heavy-labeled peptides contains 1% light peptide contamination, then all samples analyzed with that reference will exhibit a false detection if the heavy-labeled peptide signal is more than 100 times the level of detection. This should be prevented by spiking in heavy-labeled peptides at a comparable abundance as the target peptide, or demonstrating that the heavy-labeled reference has contamination much lower than a level at which putative target signals are detected.

14. Even when very high confidence peptide identifications are demonstrated, consider alternate mappings of the peptides to proteins other than the claimed extraordinary result. Consider isobaric sequence/mass modification variants, all known SAAVs, and unreported SAAVs.

Most of the earlier proteomic guidelines have been concerned with ensuring that peptide identifications are of high quality. But even with nearly irrefutable evidence that a peptide identification is correct, the peptide to protein mapping must also be considered very carefully. Clearly peptides that also map to a common, well-observed protein cannot be held up as evidence in support of an extraordinary detection claim, as the most likely explanation is that the peptide is derived from the common protein. Common laboratory contaminant protein sequences should always be considered (e.g., the GPM distributes the very comprehensive "cRAP" or "common Repository of Adventitious Proteins" set at http://www.thegpm.org/crap/). Direct mapping is easy to determine, but it is also necessary to consider alternative splice isoforms and single amino acid variants (SAAVs) in the mapping, as well. Substitutions of I/L must be accounted for as these are isobaric and cannot be distinguished by current MS/MS techniques used in mass spectrometers unless additional fragmentation routines are used[35,36]. There are other isobaric substitutions when one considers mass modifications. For example, deamidated N is equivalent to D, and deamidated Q is equivalent to E. Note that there are many more substitutions that are close but not exact, such as Q/K, that must be considered when analyzing low mass-accuracy spectra. Low mass-accuracy data cannot easily distinguish between Q and K, F and oxidized M, and similar pairs, which is another reason that guideline 11 excludes the use of low mass-accuracy ion trap spectra for confirming evidence of extraordinary detection claims. As well, there is always the possibility that a known or unknown PTM not taken into account during the search could lead alone or in combination with misidentification to an incorrect match. A tool to assist with this analysis is available at neXtProt at https://search.nextprot.org/view/unicity-checker and can be used to aid in compliance with this guideline. For example, in PeptideAtlas peptide SITDVLSADDIAAALQECQDPDTFEPQK

appears to uniquely map to PE=5 protein Putative oncomodulin-2 (P0CE71) amongst the core 20,000 neXtProt predicted proteins. However, when SAAVs are considered, one finds that it also maps to a known variant (dbSNP rs202012112) of PE=1 protein Oncomodulin-1 (P0CE72). This peptide is therefore no longer uniquely mapping, and cannot be held up as protein evidence for the existence of PE=5 protein P0CE71.

15. Support extraordinary detection claims by two or more distinct uniquely-mapping, non-nested peptide sequences of length ≥ 9 amino acids. When weaker evidence is offered for detection of a previously unreported protein or a coding element proposed translation product, justify that other peptides cannot be expected.

As outlined above, it is clear that an apparently very high quality uniquely-mapping peptide identification can still be incorrect as a protein match. In fact, in very large datasets using the thresholds advocated in these guidelines, there will surely still be a few such cases. Therefore, in order to engender additional confidence in extraordinary detection claims, we require the evidence of two distinct peptides of length 9 amino acids or more. Further, one of the peptides may not be fully nested within the other. Nested peptides are not counted, because, while they increase the confidence of the sequence being accurately identified, especially in the case of ragged peptides from termini, it does not generate additional confidence in the uniqueness of the peptide-to-protein mapping.

Very short peptides usually map to many different proteins, and there are abundant examples in PeptideAtlas where apparently "uniquely mapping" peptides can be better explained by mappings to variants or nearly identical isobaric peptides for other proteins. This problem is so rampant with peptides of length 6 or less, that they have long been completely discarded from PeptideAtlas and never shown. In PeptideAtlas peptides of length 7 are retained and shown, but there are many cases where one cannot feel confident that such short peptides are truly indicative of a protein detection alone. As a cautionary note,

there are also such cases for peptides of length 8, and we have therefore conservatively set a lower limit of 9 amino acids for peptides that are needed to confer the canonical designation in PeptideAtlas and the protein existence level 1 in neXtProt. It is useful to extend this same requirement for evidence of extraordinary detections. If it is desirable to present evidence that does not meet these criteria (covered in next paragraph), the implicated proteins may be offered as "candidate detections" to enable capture of this information by other researchers and use in potential future experiments.

In some rare cases there are proteins that simply do not contain enough uniquely mapping peptides of sufficient length to call a protein detected. For example, proteins with very few or excessive basic residues produce only a few extremely long peptides, if any, on the one hand, and produce many excessively short peptides on the other hand when trypsin is used as the cleavage reagent. The use of other enzymes, e.g., GluC or chymotrypsin, or chemical cleavage reagents may provide additional opportunities to detect a protein by generating different repertoires of peptides. It is still permissible to present evidence that does not fully meet this guideline if there is a strong justification that additional peptide evidence will be extraordinarily difficult to achieve. For example, if a single peptide or short peptides are all that can be reasonably expected for a missing protein, even with the use of multiple proteases, based on its sequence, and these are precisely the peptides that are observed, the community and the neXtProt curators may be convinced to relax this guideline in such special cases.

Discussion

Although there are already several sets of guidelines relevant to proteomics, there is minimal content overlap. Most of the other guidelines focus on specific basic metadata that must be provided, while the HPP guidelines focus mostly on addressing the need to provide well accepted evidence for analysis quality and for new identifications whilst reducing the probability of false positive identifications that seem to implicate proteins that were never before seen and are thus highly sought after as "missing

proteins". In this sense, these guidelines complement other data inclusion guidelines that may also apply. For example, for manuscripts submitted to JPR, the journal data submission guidelines also apply. There is minimal overlap between these two sets of guidelines; where these do overlap, complying with the HPP guidelines should include compliance with the JPR guidelines.

The overarching theme for the HPP guidelines is that careful control of false positives is crucial for unambiguous protein identification when the goal of the work is to present claims of comprehensive datasets of increasingly nearly complete proteomes and for the inclusion of never-before-confidently-detected proteins. Unless the number of errors present in a final protein list is much smaller than the number of claimed novel discoveries, confirmatory orthogonal evidence must be presented to demonstrate that the novel claims are not merely one of the false positives.

There is one additional guideline that was considered and not included in the current release. There was consideration for a requirement for a confirming detection in a second sample. At present, two peptides from a single sample is all that is required. Majority consensus was that requiring detection of a protein from at least two separate samples (i.e. biological replicates rather than technical replicates) was raising the bar too high, and this guideline was not included. It may be considered for future guidelines upon consultation with the proteomics community.

Another situation may arise that provides evidence for a missing protein, but which does not meet the guidelines to its positive identification. For example, in PTM peptide enrichment studies, e.g., glyco and phospho proteomics, and N and C terminomics[38], a single peptide from a protein is often identified with high confidence. The point of the study may be to characterize the PTMs rather than to identify proteins, but such studies also provide an orthogonal approach to provide proteomic evidence for proteins, especially useful for missing proteins. For such studies with high quality spectra and peptides that meet the spectral assignment and peptide identification guidelines otherwise, these peptides can be designated

as potentially having come from the missing protein. In any case specific caveats need to be stated, e.g., that peptide evidence was found for a missing protein or that "candidate missing proteins" were detected by these high confidence peptides, but that further evidence is required for high confidence identification. The hope is that these identifications can stimulate other groups to specifically seek further evidence of the missing protein in that tissue, for example, or by using such approaches incorporated into broader studies to identify recalcitrant missing proteins.

Although there has been extensive discussion and refinement of the guidelines, the first real test of the guidelines has been this JPR 2016 HPP special issue. All submitted manuscripts were required to comply with these guidelines. Completed checklists were submitted with the manuscripts. The special issue editors agreed to perform a first pass of compliance checking before the manuscripts were sent out for review. Reviewers were then asked to consider the guidelines as they review the manuscripts. Authors generally complied with the guidelines, either upon submission or, in multiple cases, during revision. We anticipate that JPR will consider adopting these guidelines for papers claiming identification of Missing Proteins in regular journal issues. We encourage all journals, whether inside or outside the field of proteomics, to consider and adopt these guidelines.

There is potential opportunity for integration with other guidelines, but this task will need effort from the respective stakeholders. Many of the guidelines, including these, are directed to a specific purpose and may not apply well in other experimental designs. The reasonable desire to have a single set of guidelines might only result in a large and unwieldy document with many "if-then" sections for different strategies. Despite these considerations, these HPP guidelines break new ground regarding the somewhat narrow focus about claims of novel protein detection, and many of these individual guidelines may be suitable for inclusion in more general fit-for-purpose guidelines.

Conclusion

We have presented the latest version (2.1.0) of the HPP MS Data Interpretation Guidelines. These guidelines expand substantially on the version 1.0 guidelines, which only required any kind of ProteomeXchange deposition and a 1% protein-level FDR threshold. For manuscript submission to the Journal of Proteome Research the primary guidelines comprise a one-page checklist followed by two pages of extended information. This article provides an in-depth history, reasoning, and expanded discussion of each of the guidelines so that the community may fully understand their intent and consider whether broader application to other projects is appropriate. The previous 2012 guidelines served the HPP well for three years. These guidelines will be further refined and expanded by the HPP as the field advances.

Acknowledgements

This work was funded in part by the National Institutes of Health through NIGMS grant R01GM087221 and NIBIB grant U54EB020406 (EWD) and NIEHS grant U54ES017885 (GSO). The authors have no conflicts of interest to declare.

Supporting Information

Supporting Information: Checklist document with extended description of the guidelines. Supplementary Document S1.

References

(1) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.; et al. The human proteome project: current state and future direction. *Mol. Cell Proteomics* **2011**, *10* (7), M111.009993.

(2) Paik, Y.-K.; Jeong, S.-K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H.-J.; Na, K.; Choi, E.-Y.; Yan, F.; et al. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30* (3), 221–223.
(3) Paik, Y.-K.; Omenn, G. S.; Overall, C. M.; Deutsch, E. W.; Hancock, W. S. Recent Advances in the Chromosome-Centric Human Proteome Project: Missing Proteins in the Spot Light. *J. Proteome Res.* **2015**, *14* (9), 3409–3414.
(4) Horvatovich, P.; Lundberg, E. K.; Chen, Y.-J.; Sung, T.-Y.; He, F.; Nice, E. C.; Goode, R. J.; Yu, S.; Ranganathan, S.; Baker, M. S.; et al. Quest for Missing Proteins: Update 2015 on Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2015**, *14* (9), 3415–3431.
(5) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198–207.
(6) Nilsson, T.; Mann, M.; Aebersold, R.; Yates, J. R.; Bairoch, A.; Bergeron, J. J. M. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* **2010**, *7* (9), 681–685.
(7) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell Proteomics* **2012**, *11* (6), O111.016717.
(8) Picotti, P.; Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **2012**, *9* (6), 555–566.
(9) Deutsch, E. W.; Lam, H.; Aebersold, R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol. Genomics* **2008**, *33* (1), 18–25.
(10) Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* **2010**, *73* (11), 2092–2123.
(11) Taylor, C. F.; Paton, N. W.; Lilley, K. S.; Binz, P.-A.; Julian, R. K.; Jones, A. R.; Zhu, W.; Apweiler, R.; Aebersold, R.; Deutsch, E. W.; et al. The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* **2007**, *25* (8), 887–893.
(12) Deutsch, E. W.; Albar, J. P.; Binz, P.-A.; Eisenacher, M.; Jones, A. R.; Mayer, G.; Omenn, G. S.; Orchard, S.; Vizcaíno, J. A.; Hermjakob, H. Development of data representation standards by the human proteome organization proteomics standards initiative. *J Am Med Inform Assoc* **2015**, *22* (3), 495–506.
(13) Bradshaw, R. A.; Burlingame, A. L.; Carr, S.; Aebersold, R. Reporting protein identification data: the next generation of guidelines. *Mol. Cell Proteomics* **2006**, *5* (5), 787–788.
(14) Burlingame, A.; Carr, S. A.; Bradshaw, R. A.; Chalkley, R. J. On Credibility, Clarity, and Compliance. *Mol. Cell Proteomics* **2015**, *14* (7), 1731–1733.
(15) Carr, S. A.; Abbatiello, S. E.; Ackermann, B. L.; Borchers, C.; Domon, B.; Deutsch, E. W.; Grant, R. P.; Hoofnagle, A. N.; Hüttenhain, R.; Koomen, J. M.; et al. Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. *Mol. Cell Proteomics* **2014**, *13* (3), 907–917.
(16) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **2014**, *11* (11), 1114–1125.

(17) Ruggles, K. V.; Tang, Z.; Wang, X.; Grover, H.; Askenazi, M.; Teubl, J.; Cao, S.; McLellan, M. D.; Clauser, K. R.; Tabb, D. L.; et al. An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. *Mol. Cell Proteomics* **2016**, *15* (3), 1060–1071.

(18) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223–226.

(19) Omenn, G. S.; Lane, L.; Lundberg, E. K.; Beavis, R. C.; Nesvizhskii, A. I.; Deutsch, E. W. Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *J. Proteome Res.* **2015**, *14* (9), 3452–3460.

(20) Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R. PRIDE: the proteomics identifications database. *Proteomics* **2005**, *5* (13), 3537–3545.

(21) Ternent, T.; Csordas, A.; Qi, D.; Gómez-Baena, G.; Beynon, R. J.; Jones, A. R.; Hermjakob, H.; Vizcaíno, J. A. How to submit MS proteomics data to ProteomeXchange via the PRIDE database. *Proteomics* **2014**, *14* (20), 2233–2241.

(22) Vizcaíno, J. A.; Csordas, A.; Del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **2016**, *44* (D1), D447–D456.

(23) Farrah, T.; Deutsch, E. W.; Kreisberg, R.; Sun, Z.; Campbell, D. S.; Mendoza, L.; Kusebauch, U.; Brusniak, M.-Y.; Hüttenhain, R.; Schiess, R.; et al. PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics* **2012**, *12* (8), 1170–1175.

(24) Desiere, F.; Deutsch, E. W.; Nesvizhskii, A. I.; Mallick, P.; King, N. L.; Eng, J. K.; Aderem, A.; Boyle, R.; Brunner, E.; Donohoe, S.; et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **2005**, *6* (1), R9.

(25) Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R. The PeptideAtlas project. *Nucleic Acids Res.* **2006**, *34* (Database issue), D655–D658.

(26) Deutsch, E. W.; Sun, Z.; Campbell, D.; Kusebauch, U.; Chu, C. S.; Mendoza, L.; Shteynberg, D.; Omenn, G. S.; Moritz, R. L. State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet. *J. Proteome Res.* **2015**, *14* (9), 3461–3473.

(27) Jones, A. R.; Eisenacher, M.; Mayer, G.; Kohlbacher, O.; Siepen, J.; Hubbard, S. J.; Selley, J. N.; Searle, B. C.; Shofstahl, J.; Seymour, S. L.; et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell Proteomics* **2012**, *11* (7), M111.014381.

(28) Gaudet, P.; Michel, P.-A.; Zahn-Zabal, M.; Cusin, I.; Duek, P. D.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Teixeira, D.; et al. The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Res.* **2015**, *43* (Database issue), D764–D770.

(29) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207–214.

(30) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* **2010**, *604*, 55–71.

(31) Choi, H.; Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7* (1), 47–50.
(32) Savitski, M. M.; Wilhelm, M.; Hahne, H.; Kuster, B.; Bantscheff, M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Mol. Cell Proteomics* **2015**, *14* (9), 2394–2404.
(33) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74* (20), 5383–5392.
(34) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75* (17), 4646–4658.
(35) Colaert, N.; Degroeve, S.; Helsens, K.; Martens, L. Analysis of the resolution limitations of peptide identification algorithms. *J. Proteome Res.* **2011**, *10* (12), 5555–5561.
(36) Armirotti, A.; Millo, E.; Damonte, G. How to discriminate between leucine and isoleucine by low energy ESI-TRAP MSn. *J. Am. Soc. Mass Spectrom.* **2007**, *18* (1), 57–63.
(37) Lebedev, A. T.; Damoc, E.; Makarov, A. A.; Samgina, T. Y. Discrimination of leucine and isoleucine in peptides sequencing with Orbitrap Fusion mass spectrometer. *Anal. Chem.* **2014**, *86* (14), 7017–7022.
(38) Huesgen, P. F.; Lange, P. F.; Rogers, L. D.; Solis, N.; Eckhard, U.; Kleifeld, O.; Goulas, T.; Gomis-Rüth, F. X.; Overall, C. M. LysargiNase mirrors trypsin for protein C-terminal and methylation-site identification. *Nat. Methods* **2015**, *12* (1), 55–58.
(39) Marino, G.; Eckhard, U.; Overall, C. M. Protein Termini and Their Modifications Revealed by Positional Proteomics. *ACS Chem. Biol.* **2015**, *10* (8), 1754–1764.

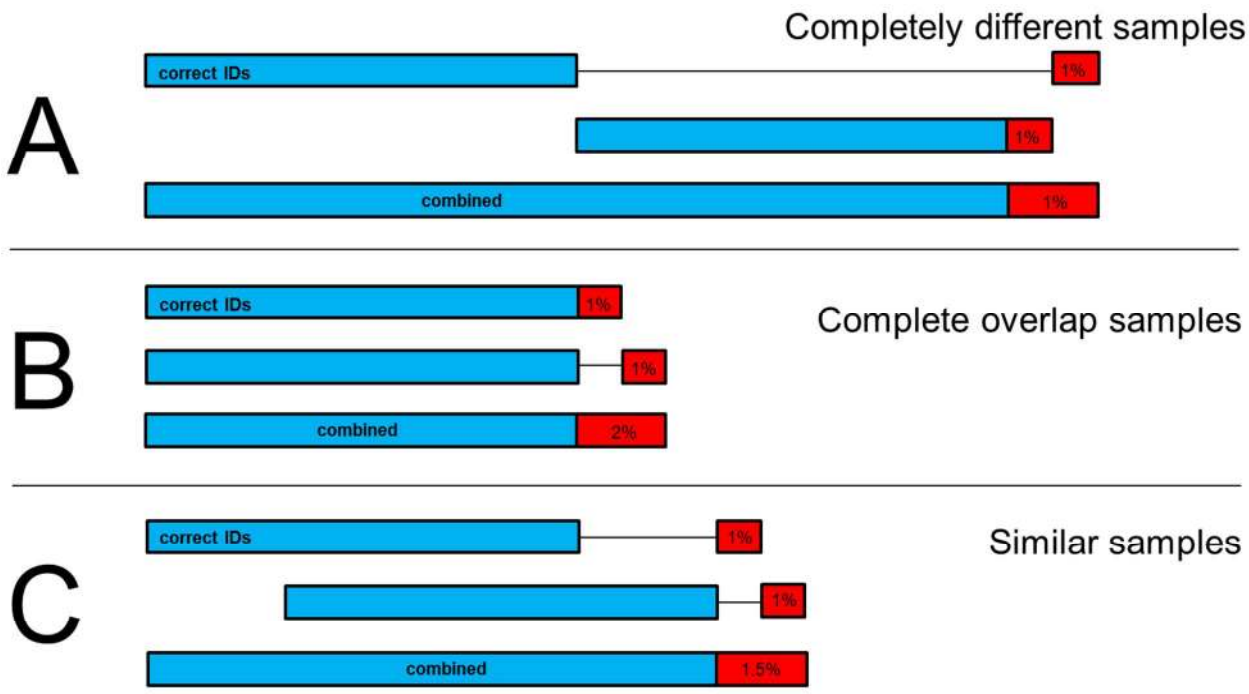
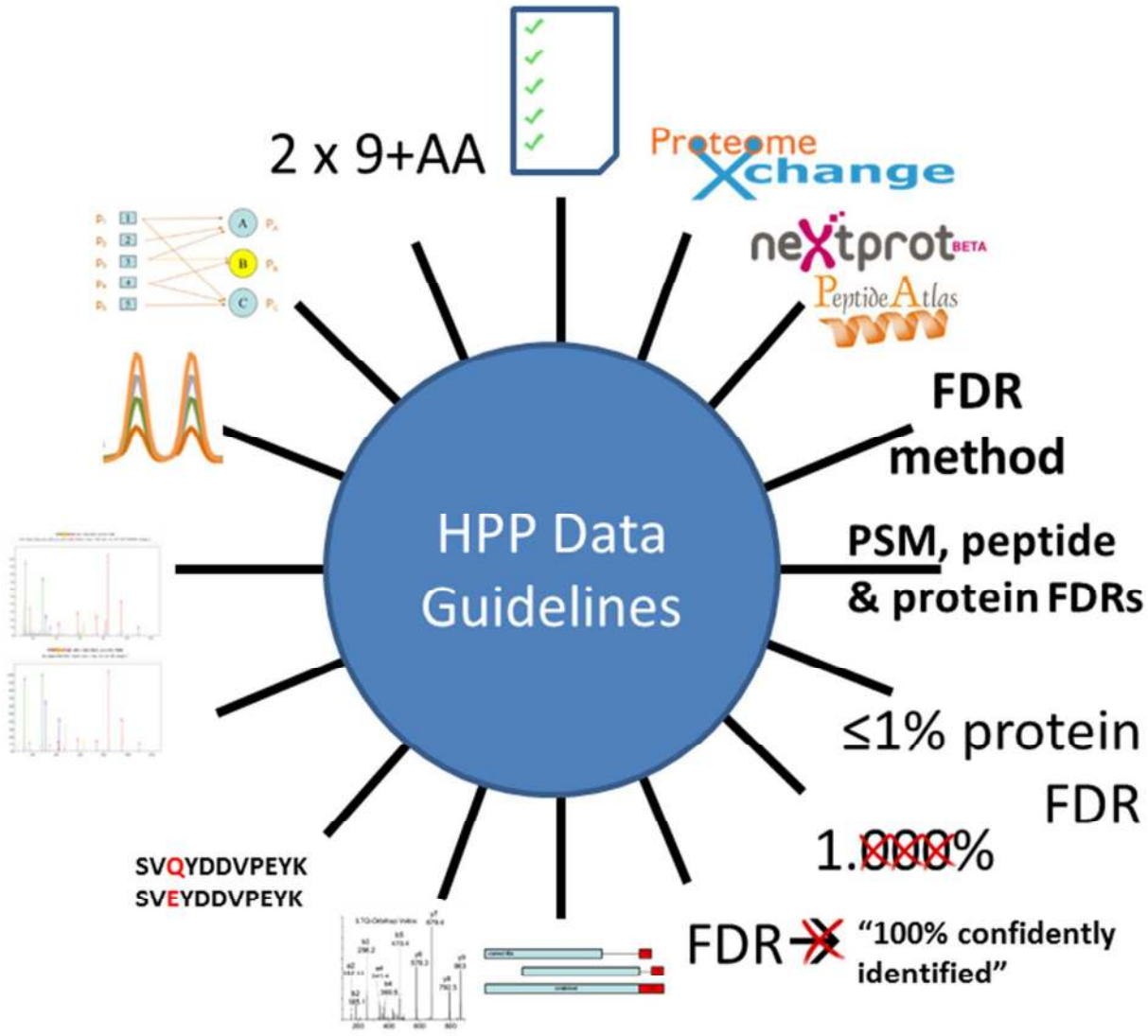


Figure 1. Three scenarios for how false discovery rates combine. True positives are shown in blue, and false positives in red. The red false positive boxes are depicted approximately 10 times larger than they should be for enhanced readability. A) If none of the true positives and 1% false positives overlap, then the final FDR does not expand. B) If all of the true positives overlap, but none of the false positives overlap (because they are false and random), then the final FDR is double the original rates. C) In a real-world scenario where the intersection of true positives overlaps by 50% and the false positives do not overlap, the combined FDR is 1.5%. This effect compounds as more datasets are merged.

Table 1. Checklist of the HUPO-2015 Human Proteome Project Data Interpretation Guidelines version 2.1.0.

General Guidelines:	
√	1. Complete this HPP Data Interpretation Guidelines checklist and submit with your manuscript.
	2. Deposit all MS proteomics data (DDA, DIA, SRM), including analysis reference files (search database, spectral library), to a ProteomeXchange repository as a complete submission. Provide the PXD identifier(s) in the manuscript abstract and reviewer login credentials.
	3. Use the most recent version of the neXtProt reference proteome for all informatics analyses, particularly with respect to potential missing proteins.
	4. Describe in detail the calculation of FDRs at the PSM, peptide, and protein levels.
	5. Report the PSM-, peptide-, and protein-level FDR values along with the total number of expected true positives and false positives at each level.
	6. Present large-scale results thresholded at equal to or lower than 1% protein-level global FDR.
	7. Recognize that the protein-level FDR is an estimate based on several imperfect assumptions, and present the FDR with appropriate precision.
	8. Acknowledge that not all proteins surviving the threshold are "confidently identified".
	9. If any large-scale datasets are individually thresholded and then combined, calculate the new, higher peptide- and protein-level FDRs for the combined result.
Guidelines for extraordinary detection claims (e.g., missing proteins, novel coding elements)	
	10. Present "extraordinary detection claims" based on DDA mass spectrometry with high mass-accuracy, high signal-to-noise ratio (SNR), and clearly annotated spectra.
	11. Consider alternate explanations of PSMs that appear to indicate extraordinary results.
	12. Present high mass-accuracy, high-SNR, clearly annotated spectra of synthetic peptides that match the spectra supporting the extraordinary detection claims.
	13. If SRM verification for extraordinary detection claims is performed, present target traces alongside synthetic heavy-labeled peptide traces, demonstrating co-elution and very closely matching fragment mass intensity patterns.
	14. Even when very high confidence peptide identifications are demonstrated, consider alternate mappings of the peptides to proteins other than the claimed extraordinary result. Consider isobaric sequence/mass modification variants, all known SAAVs, and unreported SAAVs.
	15. Support extraordinary detection claims by two or more distinct uniquely-mapping, non-nested peptide sequences of length ≥9 amino acids. When weaker evidence is offered for a previously unreported protein or a coding element proposed translation product, justify that other peptides cannot be expected.


2 x 9+AA
ProteomeXchange
neXtprot BETA
PeptideAtlas
FDR method
PSM, peptide & protein FDRs
≤1% protein FDR
1.000%
FDR
"100% confidently identified"
HPP Data Guidelines
SVQYDDVPEYK
SVEYDDVPEYK

TOC Figure