Human Proteome Project Data Interpretation Guidelines
Version 2.1.0 – July 28, 2016

The following checklist is a brief summary of the full guidelines. This checklist must be completed by authors and submitted along with the manuscript. See pages 2-3 of this document for a more detailed description of each item in the checklist. Each item in the checklist must be either checked when deemed completed or marked as NA (Not Applicable). The checklist will be used by editorial staff and reviewers to guide their assessment of submissions, marking in their review if any of the guidelines are not completed to their satisfaction.

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.

Author comments (use this space and extra pages to explain any nonadherence in the above checklist):

(see extended description for each of the above items on page 2 and 3 below)

Extended Detail on Checklist items:

1. **Complete this HPP Data Interpretation Guidelines checklist and submit with your manuscript**. Page 1 of this document must be submitted as supplementary material for the editor/reviewers. The completed checklist is required before a manuscript will be sent to reviewers. Each item in the checklist must be either checked or marked as NA (Not Applicable). Please explain NA entries or any other variances in the Author Comments section. Manuscripts received without a checklist will be returned without review.
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**. All depositions are required to be "Complete" submissions instead of "Partial" submissions. ProteomeXchange deposition must be completed prior to submission of the manuscript to the journal. Reviewer login information at the repository must be provided in the manuscript submission if the dataset is not already publicly released.
3. **Use the most recent version of the neXtProt reference proteome for all informatics analyses, particularly with respect to potential missing proteins.** Informatics analysis should always be presented in comparison with the most recent proteome references, rather than older versions thereof. For the HPP special issues, the required version will be listed in the call for papers.
4. **Describe in detail the calculation of FDRs at the PSM, peptide, and protein levels**. Describe which tools are used to estimate the false discovery rate (FDR) at the peptide-spectrum-match (PSM) level, at the distinct peptide sequence level, and at the protein level. Briefly describe the approach and what assumptions are made or implied, and any corrections for the fraction of the proteome covered. If you use novel or uncommon tools and criteria, compare your results with results with tools that are widely used in the community.
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**. Report the actual numbers of true positives and false positives at each level based on the thresholds used.
6. **Present large-scale results thresholded at equal to or lower than 1% protein-level global FDR**. The 1% is somewhat arbitrary but well accepted and remains set as the upper limit. For many datasets from modern instrumentation, achieving a 1% global FDR may include very low quality results with a **local** FDR worse than 10%, which is undesirable. A global FDR lower than 1% is encouraged, but it should never be higher than 1%. Similarly, PSMs, peptides, and proteins with a local FDR worse than 10% should not be included.
7. **Recognize that the protein-level FDR is an estimate based on several imperfect assumptions, and present the FDR with appropriate precision**. For example if decoys are used to estimate the number of expected errors, realize that there other types of errors that are not modeled well by decoys, and therefore the calculated FDR may be considerably lower than the true FDR. Do not report the FDR with many significant digits.
8. **Acknowledge that not all proteins surviving the threshold are "confidently identified".** The common mistake of thresholding at 1% FDR and then assuming that all surviving results are correct, no matter how surprising, must be avoided. Sometimes the number of estimated false positives equals or exceeds the number of missing proteins claimed to be 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**. When datasets are combined, the true positives will mostly overlap, while the false positives will be scattered randomly across the proteome and thus overlap far less. This means that the FDR will be higher in the combined dataset.
10. **Present "extraordinary detection claims" based on DDA mass spectrometry with high mass-accuracy, high signal-to-noise ratio (SNR), and clearly annotated spectra**. Annotated spectra (i.e. spectra with the matched peaks clearly labeled) must be provided in the supplementary material for the manuscript. While low mass-accuracy and low SNR spectra can still be useful for many experiments, they are not acceptable for claims of

extraordinary detections. Time-of-flight, FT-ICR, and Orbitrap-type instruments are considered as having high mass accuracy (when properly calibrated) in these guidelines.

11. **Consider alternate explanations of PSMs that appear to indicate extraordinary results**. The spectra should be examined closely to determine if there are peaks missing that should be expected, if there are peaks present that are unexplained, and if a small alteration of the putative sequence would yield a much better match. This may indicate a false positive of a kind that is not modeled well by decoys.
12. **Present high mass accuracy, high-SNR, clearly annotated spectra of synthetic peptides that match the spectra supporting the extraordinary detection claims**. Synthetic peptides are powerful tools for determining the correct identification of spectra. For each PSM corresponding to an extraordinary detection claim, a synthetic peptide should be created and run through the same high mass-accuracy instrument to verify that the intensity patterns of the spectra and the retention times are a very close match.
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**. All SRM runs performed must have spiked-in heavy labeled peptides corresponding to the putative identifications. Annotated chromatograms must be provided in the supplementary material for the manuscript. Remember that solid peptide sequence evidence does not alter the uncertainties in matching that peptide uniquely to a protein (guideline 14).
14. **Even when very high confidence peptide identifications are demonstrated, consider alternate mappings of the peptide to proteins other than the claimed extraordinary result. Consider isobaric sequence/mass modification variants, all known SAAVs, and unreported SAAVs**. Even when a peptide identification is shown to be very highly confident, care should be taken when mapping it to a protein or novel coding element. Consider whether I=L, N[Deamidated]=D, Q[Deamidated]=E, GG=N, Q≈K, F≈M[Oxidation], or other isobaric or near isobaric substitutions could change the mapping of the peptide from an extraordinary result to a mapping to a commonly-observed protein. Consider if a known single amino-acid variation (SAAV) in neXtProt could turn an extraordinary result into an ordinary result. Consider if a single amino-acid change, not yet annotated in a well-known source, could turn an extraordinary result into a questionable result. Check more than one reference proteome (e.g., RefSeq may have entries that UniProt and Ensembl do not, and vice versa). A tool to assist with this analysis is available at neXtProt at https://search.nextprot.org/view/unicity-checker.
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**. Single-peptide detections simply have too high a chance of being some type of pernicious false positive to be sufficient for claiming an extraordinary result. Likewise, short peptides of length 8 or smaller have relatively few peaks and have an increased chance of mapping to immunoglobulins or other sequences not readily apparent in the reference proteome. Nested peptides (where one sequence is fully subsumed within another) do provide additional confidence that the peptide identification is correct, but provide no additional evidence that the peptide-to-protein mapping is unique. In rare cases only a single uniquely mapping peptide can be reasonably expected even when applying different proteases; this may then be sufficient if the case is well justified. Alternatively, if it is desirable to present evidence that does not meet these criteria for extraordinary claims, the implicated proteins may be offered as “candidate detections” to enable capture of this information by other researchers for follow up by further experiments.