



HAL
open science

Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0

Eric W Deutsch, Lydie Lane, Christopher M Overall, Nuno Bandeira, Mark D. Baker, Charles Pineau, Robert L Moritz, Fernando Corrales, Sandra Orchard, Jennifer E van Eyk, et al.

► **To cite this version:**

Eric W Deutsch, Lydie Lane, Christopher M Overall, Nuno Bandeira, Mark D. Baker, et al.. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. *Journal of Proteome Research*, 2019, 18 (12), pp.4108-4116. 10.1021/acs.jproteome.9b00542 . hal-02364721

HAL Id: hal-02364721

<https://univ-rennes.hal.science/hal-02364721>

Submitted on 27 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0

Eric W. Deutsch^{1,*}, Lydie Lane², Christopher M. Overall³, Nuno Bandeira⁴, Mark S. Baker⁵, Charles Pineau⁶, Robert L. Moritz¹, Fernando Corrales⁷, Sandra Orchard⁸, Jennifer E. Van Eyk⁹, Young-Ki Paik¹⁰, Susan T. Weintraub¹¹, Yves Vandenbrouck¹², and Gilbert S. Omenn^{1,13}

¹ Institute for Systems Biology, Seattle, WA, USA

² SIB Swiss Institute of Bioinformatics and Department of microbiology and molecular medicine, Faculty of medicine, University of Geneva, CMU, Michel Servet 1, 1211 Geneva 4, Switzerland

³ Centre for Blood Research, Departments of Oral Biological & Medical Sciences, and Biochemistry & Molecular Biology, Faculty of Dentistry, The University of British Columbia, Vancouver, Canada

⁴ Center for Computational Mass Spectrometry and Department of Computer Science and Engineering, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California 92093, United States

⁵ Department of Biomedical Sciences, Faculty of Medicine and Health Science, Macquarie University, NSW, 2109, Australia

⁶ Univ. Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, F-35042 Rennes cedex, France

⁷ Functional Proteomics Laboratory. Centro Nacional de Biotecnología. Spanish Research Council. ProteoRed-.ISCIII. Madrid, Spain

⁸ European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

⁹Advanced Clinical Biosystems Research Institute, The Smidt Heart Institute, Department of Medicine, Cedars Sinai Medical Center, Los Angeles, CA, USA

¹⁰ Yonsei Proteome Research Center, Yonsei University, 50 Yonsei-ro, Sudaemoon-ku, Seoul, 03720, Korea

¹¹ The University of Texas Health Science Center at San Antonio, San Antonio, Texas, USA

¹² Univ. Grenoble Alpes, CEA, INSERM, IRIG-BGE, U1038, F-38000, Grenoble, France

¹³ 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: Email: edeutsch@systemsbiology.org, Phone: 206-732-1200, Fax: 206-732-1299.

Abstract

The Human Proteome Organization's (HUPO) Human Proteome Project (HPP) developed Mass Spectrometry (MS) Data Interpretation Guidelines that have been applied since 2016. These guidelines have helped ensure that the emerging draft of the complete human proteome is highly accurate and with low numbers of false-positive protein identifications. Here, we describe an update to these guidelines based on consensus-reaching discussions with the wider HPP community over the past year. The revised 3.0 guidelines address several major and minor identified gaps. We have added guidelines for emerging data independent acquisition (DIA) MS workflows and for use of the new Universal Spectrum Identifier (USI) system being developed by the HUPO Proteomics Standards Initiative (PSI). In addition, we discuss updates to the standard HPP pipeline for collecting MS evidence for all proteins in the HPP, including refinements to minimum evidence. We present a new plan for incorporating MassIVE-KB into the HPP pipeline for the next (HPP 2020) cycle in order to obtain more comprehensive coverage of public MS data sets. The main checklist has been reorganized under headings and subitems and related guidelines have been grouped. In sum, Version 2.1 of the HPP MS Data Interpretation Guidelines has served well and this timely update to version 3.0 will aid the HPP as it approaches its goal of collecting and curating MS evidence of translation and expression for all predicted ~20,000 human proteins encoded by the human genome.

Keywords: guidelines, standards, Human Proteome Project, HPP, mass spectrometry, Universal Spectrum Identifier (USI), false-discovery rates, C-HPP, B/D-HPP, unicity checker.

Introduction

The Human Proteome Organization's¹ (HUPO) Human Proteome Project^{2,3} (HPP) was launched in 2010 as an international endeavor to build on the success of the Human Genome Project^{4,5} by characterizing the products of the ~20,000 human protein-coding genes. As of January 2019, 17,694 proteins demonstrated compelling mass spectrometry (MS) or non-MS protein-level evidence in neXtProt (i.e., PE1), leaving 2129 proteins without strong evidence (PE2,3,4) that were have been designated as the HPP's 'missing proteins'⁶. The PE2,3,4 missing proteins represented 10.7% of all neXtProt's PE2,3,4 proteins. The goals of the HPP are (1) to complete the protein 'parts' list, including isoforms, post-translational modifications (PTMs), and single amino acid variants, with characterization of their functions; and (2) to make proteomics an integral part of all multi-omics studies in life sciences. The Chromosome-centric HPP (C-HPP) consortium focused largely, but not exclusively, on the first two goals⁷, whereas the Biology and Disease HPP (B/D-HPP) focused largely on the latter goal, whilst recognizing that many studies will also uncover disease-specific or tissue-specific PE2,3,4 missing proteins. The progress in achieving these goals has been tracked yearly via a set of published metrics^{3,8-11} based on the major knowledge bases of the HPP, namely neXtProt¹², PeptideAtlas¹³⁻¹⁵, Human Protein Atlas¹⁶, and the ProteomeXchange^{17,18} consortium of proteomics data repositories.

In order to maintain a high standard of quality for the identifications in the compendium of human proteins, and to ensure that journal articles and data contributions are reporting

1
2
3 with equally high standards, a set of MS data guidelines was developed. The inaugural
4
5 version 1.0 guidelines were released in 2013 and mandated deposition of data to members
6
7 of the newly formed ProteomeXchange Consortium for proteomics/MS data and other
8
9 repositories for other kinds of biochemical data. Initial progress in protein detection was
10
11 rapid since there were many high abundance proteins present in common samples
12
13 available to catalog. However, it soon became apparent that, as the compendium of
14
15 proteins commonly seen in high abundance became complete, the control of false
16
17 positives during the hunt for missing proteins became imperative. Version 2.1 of the HPP
18
19 MS Data Interpretation Guidelines was developed and published in 2016¹⁹. These
20
21 guidelines went beyond data deposition requirements, setting out minimum standards for
22
23 the handling of false discovery rate (FDR) in the interpretation of MS data as well as
24
25 minimum standards to claim the detection of any missing protein or protein otherwise not
26
27 yet found in the HPP KB compendium of detected proteins.
28
29
30
31
32
33
34

35 Version 2.1 of the guidelines has been in force for the annual Journal of Proteome
36
37 Research HPP Special Issues beginning in 2016. For papers submitted outside the frame
38
39 of the Special Issue, the Editors of the Journal of Proteome Research and increasing
40
41 numbers of other proteomics journals now require these guidelines to be met for all
42
43 claims of missing protein identifications.
44
45
46
47
48

49 In the past year it has become apparent that, despite the advances in proteomics, the
50
51 increased difficulty of detecting the remaining ~10% of the human proteome requires an
52
53 update to the guidelines, as discussed on-line
54
55
56
57
58
59
60

1
2
3 <https://docs.google.com/document/d/167wLMYshQ3jUPJonxyk6TcOvT8GrZcqYOZAt>
4 [UGOK29w](#)), in the Bioinformatics Hub at the 2018 17th HUPO World Congress in
5
6 Orlando, USA, at the 2019 21st C-HPP Symposium in Saint-Malo, France, and at the
7
8 2019 18th HUPO World Congress in Adelaide, Australia. In these venues, the leadership
9
10 of the HPP, along with other interested contributors, debated 25 aspects of the existing
11
12 guidelines for journal articles as well as current practices of the pipelines that maintain
13
14 and refine the resources that comprise the HPP KB ecosystem.
15
16
17
18
19
20

21
22 Here, we describe the outcomes of these discussions, which are reflected in a refined set
23
24 of guidelines to take the HPP forward. First, we present a revised version 3.0 of the HPP
25
26 MS Data Interpretation Guidelines in the form of a brief one-page checklist and more
27
28 extensive three-page checklist documentation. Next, we discuss the reasoning behind the
29
30 changes to each guideline, often providing the set of options debated. Finally, we discuss
31
32 the reasoning behind the changes to overall HPP policy used by the HPP KB pipeline that
33
34 tracks and disseminates the best-available gathered understanding of the human proteome
35
36 as the HPP and the global community gear up to tackle the most difficult refractory
37
38 proteins of the human proteome.
39
40
41
42
43
44
45

46 **Changes to the guidelines**

47

48
49 Whilst version 3.0 of the checklist (Supplementary Material S1) looks similar to the
50
51 previous version 2.1 (<https://www.hupo.org/HPP-Data-Interpretation-Guidelines>), it
52
53 contains major differences. First, in addition to the requirement of a checkmark to
54
55 indicate that each requirement is fulfilled (or NA for not applicable or NC for not
56
57
58
59
60

1
2
3 completed, both of which require an explanation as to why this is not applicable or
4 complete at the bottom of the checklist), a new column requests a location where the
5 pertinent information may be found. This will typically be a reference to a page number
6 or a supplementary document. This additional information makes it much easier for the
7 journal editors, reviewers, and readers to find the section in the submission that fulfills
8 each guideline, which can sometimes be difficult and slows reviewing. Such a
9 requirement for page numbers is already common in submission checklists for many
10 bioinformatics journals.
11
12
13
14
15
16
17
18
19
20
21
22
23

24 A second substantial difference is the reorganization of the guidelines into numbered
25 major items and lettered subitems. Whereas the previous version had 15 major items,
26 some of which were highly related and needed to be considered together, the latest
27 version has only 9 major items, but some of those contain subitems that should be
28 considered as a group. We hope that this provides a better overall organizational
29 framework and is more user-friendly ensuring contributor completion. A few guidelines
30 have been deemphasized by being merged with important related guidelines. Two new
31 guidelines have been added, as discussed in detail below. In the following paragraphs
32 each guideline will be discussed briefly with emphasis on changes since version 2.1.0.
33
34 There are two parts of the guidelines: reports of well-established proteins and reports of
35 claims of novel detection of predicted proteins.
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Section 1

1
2
3 Guideline 1 remains essentially unchanged as a formal requirement that each manuscript
4 be submitted with a filled-in checklist describing the compliance of the manuscript with
5 the guidelines. If any manuscript is received for publication without a checklist, the
6 handling editors immediately request this be completed before sending any manuscript
7 for review. The extended description of Guideline 1 has been augmented to describe the
8 new requirement in column two for a page number with line number or paragraph
9 number, or other indication of the location (such as a specific supplemental table) of the
10 requested information.
11
12
13
14
15
16
17
18
19
20
21
22
23

24 Guideline 2 requiring deposition of all data sets into a ProteomeXchange repository has
25 been expanded into four subparts because, in the version 2.1 guidelines, these subparts
26 were concatenated into one sentence, where compliance suffered. There was a strong
27 tendency for authors to fulfill the first part of the requirement and move on without
28 addressing other components. In order to avoid this, the four main aspects of the previous
29 guideline 2 have been separated into four subitems 2a, b, c, d, which require complete
30 data deposition, deposition of analysis reference files, PXD identifier in abstract, and
31 reviewer credentials, respectively.
32
33
34
35
36
37
38
39
40
41
42
43
44

45 Guideline 3, requiring the use of the most recent neXtProt release, rather than older
46 versions, remains unchanged. Our understanding of the human proteome continues to
47 evolve rapidly and the use of older versions may lead to confusion and outdated claims.
48
49

50 Generally, neXtProt curators update regularly, with their prior January release relied
51
52
53
54
55
56
57
58
59
60

1
2
3 upon for HPP Journal of Proteome Research Special Issue data analysis/reanalysis and
4
5 which effectively is reflected as the annual HPP metrics^{3,6,9-11,20}.
6
7
8
9

10 Guideline 4 merges all previous FDR-related guidelines (4 – 9 in version 2.1) into a
11
12 single top-level entry with four subitems designed to streamline this section. Previous
13
14 top-level guidelines 7 and 8 request that authors consider that FDR calculations should be
15
16 reported with an appropriate number of significant digits (usually one or two), because
17
18 they are based on several imperfect assumptions, and that the required FDR calculations
19
20 and implied number of wrong identifications should be carefully considered in later
21
22 analysis of any resulting protein list. These points have been merged into part b of
23
24 guideline 4, which also addresses reporting of FDR values. The HPP community seems
25
26 to understand these aspects well and separate items no longer seem necessary.
27
28
29
30
31
32

33 **Section 2**

34
35 Whereas guidelines 1 to 4 apply to all manuscripts presenting MS data, the following
36
37 guidelines 5 to 9 apply only to manuscripts presenting evidence that could qualify the
38
39 newly-identified proteins for consideration as PE1 in neXtProt or to provide MS evidence
40
41 for PE1 proteins lacking MS data, so classified based on other types of data.
42
43
44
45
46

47 In the previous version of the guidelines, missing protein MS evidence was referred to as
48
49 “extraordinary detection claims”, reminiscent of the aphorism that “extraordinary claims
50
51 require extraordinary evidence”, often credited to Carl Sagan
52
53 (https://en.wikipedia.org/wiki/Sagan_standard) or Amos Bairoch
54
55
56
57
58
59
60

1
2
3 (https://en.wikipedia.org/wiki/Amos_Bairoch). The phrase “extraordinary detection
4
5 claims” was confusing to many, so this phrase has been replaced by “claims of new PE1
6
7 protein detection”. Such claims may apply to one of the “missing proteins” currently in
8
9 neXtProt with protein existence status of PE2,3,4. They may apply to a current PE5
10
11 protein, although most of these entries are annotated as pseudogenes in UniProtKB and
12
13 additional care should be applied to justify that their detection is not merely a variation of
14
15 the common PE1 protein that the predicted PE5 protein sequences closely resemble.
16
17 Finally, this assignation may apply to a protein not yet listed in neXtProt. These might
18
19 include: (i) an entry not yet manually reviewed in UniProtKB, (ii) a protein currently
20
21 annotated as a lncRNA, (iii) a smORF, or (iv) some other novel protein-coding element.
22
23
24
25
26 There are many new protein entries, including immunoglobulins, in annual releases from
27
28 neXtProt. The first three guidelines are specific to each of three different acquisition
29
30 technologies, whereas the two guidelines that follow apply to all three technologies—
31
32 DDA, SRM, and DIA-MS.
33
34
35
36
37

38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

Guideline 5 has become a guideline containing three subitems that merge several previous top-level guidelines into a single one focused on requirements for data dependent acquisition (DDA) MS workflows, commonly referred to as “shotgun proteomics”. Part 5a is essentially the same as previous guideline 10, which affirms that evidence spectra for new PE1 protein detection claims must be high mass-accuracy, high signal-to-noise ratio, and clearly annotated with peak interpretations. The previous guideline 11 enjoining authors to examine the spectra carefully for telltale signs of misidentification has been appended to the extended description of the new subitem “a”,

1
2
3 since, although laudable, it was difficult for many authors to perform effectively and is
4
5 less important in the presence of the guideline requiring comparison with a synthetic
6
7 spectrum.
8
9

10
11
12 Guideline 5b is similar to the previous guideline 12, seeking clear presentations of
13
14 synthetic peptide spectra that match endogenous peptide spectra. The guideline has been
15
16 augmented to include a recommendation by the guidelines revision team group that
17
18 spectra derived from digested recombinant proteins are suitable substitutes for those MS
19
20 spectra derived from peptides created with peptide synthesizing technologies. The
21
22 guideline has also been amended so that a retention time match between the target and
23
24 the synthetic peptide are no longer required, but rather suggested only if the target and
25
26 reference are both run on the same instrument. The use of public reference spectra from
27
28 synthetic peptides such as from SRMATlas²¹ is now specifically allowed.
29
30
31
32
33
34

35
36 Guideline 5c is completely new. A persistent problem with discussions about the merits
37
38 of certain spectra as evidence for new PE1 protein detection claims is the general
39
40 inability to identify specific spectra and access them easily in the data repositories for
41
42 close examination. PDF representations of MS spectra found in supplementary materials
43
44 are useful but resist close examination and the application of KB tools that reviewers or
45
46 readers might like to use for inspection of presented MS evidence. Furthermore, if
47
48 reprocessing of the data set does not yield the same result, it is very difficult to assess
49
50 what became of the key spectrum and why it does not reveal the same PSM in
51
52 reprocessing. In order to solve this problem, the HUPO Proteomics Standards Initiative²²⁻
53
54
55
56
57
58
59
60

1
2
3 ²⁴ (PSI) has developed the Universal Spectrum Identifier (USI) concept as a multi-part
4 key that can universally identify any acquired spectrum in a manner that any repository
5 containing the data set would be able to display or furnish the same spectrum via this
6 identifier. Guideline 5c now introduces a requirement for the provision of USIs for all
7 spectra that provide evidence for new PE1 protein detection claims, natural sample
8 observations and synthetic peptide spectra alike. See <http://psidev.info/USI> for more
9 information on how to create and use USIs.
10
11
12
13
14
15
16
17
18
19
20

21
22 Guideline 6 is the same as guideline 13 in version 2.1. It applies to selected/multiple
23 reaction monitoring (SRM/MRM) workflows²⁵, requiring that chromatogram traces of the
24 detected peptides be provided along with the matching chromatograms of heavy-labeled
25 reference synthetic peptides. It is important in SRM that both the intensity patterns and
26 the retention times match, since there are typically far fewer ions monitored than peaks
27 available in full spectra. We have added a request that the heavy-labeled reference
28 peptides should be spiked in at an abundance similar to the target peptides so that minor
29 impurities in the reference do not contribute to the target signal. If the heavy-labeled
30 spike-in has a 1% impurity in the form of light peptide, then, if the reference is spiked in
31 at 100 times the target peptide abundance, the impurity will contribute as much signal as
32 the target peptide, leading to an incorrect abundance or even a spurious detection. The
33 extended description reaffirms that guidelines 8 and 9 also apply to SRM as there has
34 been some confusion previously. This same guideline can also be applied to parallel
35 reaction monitoring²⁶ (PRM) data, although since PRM acquisition creates full MS/MS
36 spectra, full compliance with guideline 5 is also acceptable.
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Guideline 7 is a new guideline that addresses untargeted data independent acquisition (DIA) workflows such as SWATH-MS²⁷ and similar techniques²⁸. The version 2.1 guidelines were silent on DIA data sets as we felt that the technology was too new to write useful guidelines at that time. In the meantime, DIA has become a broadly-adopted technology. Although DIA has not yet been used to claim detection of new PE1 proteins, this will surely come. Guideline 7 is simple. It applies guidelines 5 and 6 depending on the mode of bioinformatics analysis of the DIA data. If the data are analyzed via extracted ion chromatograms (XICs) (sometimes called peptide-centric analysis) such as with OpenSWATH²⁹, Spectronaut³⁰, and SWATH 2.0, then the SRM guideline 6 applies. If the data are analyzed via extracted demultiplexed spectra (sometimes called spectrum-centric analysis) such as with DIA-Umpire³¹ and DISCO, then the DDA guidelines 5a-c apply. The next few years will show whether DIA can be used reliably for new PE1 detection claims and if this simple approach to a DIA guideline is sufficient. Of interest is the observation that the journal *Molecular and Cellular Proteomics* has recently developed a comprehensive set of guidelines for handling DIA data³². Authors are advised to examine these and use these where applicable to further support claims, although as yet they are not required as part of the HPP and/or *Journal of Proteome Research* guidelines.

Guideline 8 remains the same as the previous guideline 14, encouraging authors to consider alternate explanations for novel spectral matches. In many cases a single amino acid variant (SAAV) or a post-translational modification (PTM) creates an isobaric or

1
2
3 near-isobaric change that can mean the difference between mapping to a protein never
4 before detected with MS and a common protein observed by millions of PSMs. Despite
5 some useful tools available for authors to address this guideline (e.g., neXtProt peptide
6 uniqueness checker³³ and PeptideAtlas ProteoMapper³⁴), it remains one of the most
7 difficult to fulfill, since exact mappings are clear enough, but near mappings are difficult
8 and time consuming to assess. Nevertheless, the authors consider that this remains an
9 important guideline that researchers and reviewers should continue to consider when
10 presenting new PE1 protein detection claims.
11
12
13
14
15
16
17
18
19
20
21
22
23

24 Guideline 9 is a derivative of the previous guideline 15, although many aspects were
25 discussed extensively and several small modifications made. This guideline provides the
26 minimum MS requirements for the number and attributes of peptides that support the
27 claim of any new PE1 protein detection. The group reaffirmed that two uniquely-
28 mapping, non-nested peptides of nine or more amino acids should be the minimum
29 required for such a claim. However, various aspects of this requirement were discussed
30 extensively and clarifications made. First, the definition of non-nesting was clarified.
31 Strictly, the meaning of non-nested means that one peptide may not be completely
32 contained within another. The reasoning is that, although the observation of two nested
33 peptides increases the confidence that the peptides have been correctly identified, it does
34 not provide any additional evidence that the peptide has been correctly mapped to the
35 protein in question; i.e., if the longer peptide is mismapped and should instead map to
36 some part of the proteome that we do not yet fully understand (such an immunoglobulin
37 or some other variation), then the nested peptide will have exactly the same problem, and
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 provides no new information. An extension of one peptide beyond the other provides
4
5 some additional mapping confidence. However, the previous guidelines as written
6
7 permitted even a single amino acid extension. For example, a tryptic peptide
8
9 PEPTIDESR and a LysargiNase³⁵ (that cleaves before K/R instead of after K/R) peptide
10
11 KPEPTIDES would qualify as non-nested under the previous guidelines. We recommend
12
13 amendment of the guidelines to require that the total extent of the coverage of the two
14
15 nested peptides combined be at minimum 18 amino acids (2×9). This strategy had
16
17 already been implemented at neXtProt, and thus there is no change there, but does reflect
18
19 a change for PeptideAtlas and other interpretations of the guidelines.
20
21
22
23
24
25

26
27 Guideline 9 now also contains a clarification for how to handle identical proteins. There
28
29 are 118 entries in the current January 2019 neXtProt release that have the same protein
30
31 sequence for at least one other entry. These entries reflect different gene loci that may
32
33 have synonymous-coding nucleotide variation but yield the exact same protein sequence.
34
35 This is an extreme case of highly homologous proteins. These 118 entries can be
36
37 retrieved by applying the SPARQL query NXQ_00231
38
39 (https://www.nextprot.org/proteins/search?mode=advanced&queryId=NXQ_00231) in
40
41 the neXtProt advanced search tool. They represent a total of 51 distinct protein
42
43 sequences. It was decided that for the purposes of PE status assignment, if two or more
44
45 qualifying peptides map uniquely to multiple identical proteins, then all such proteins
46
47 will be switched to PE1 as a group since they are indistinguishable from each other.
48
49
50 Nonetheless, it was noted that since their gene promoter regions are likely to differ, these
51
52 proteins may be expressed in different tissues, or under different spatiotemporal
53
54
55
56
57
58
59
60

1
2
3 circumstances or under different physiological or pathological conditions. As is the case
4
5 now, each will be counted individually as PE1.
6
7
8
9

10 The group further clarified that, while the two peptides presented as evidence do not need
11
12 to originate from the same sample or instrument, they do need to be presented together in
13
14 the paper. The practice of offering a single new suitable peptide to complement a pre-
15
16 existing different suitable peptide already in PeptideAtlas and neXtProt is permitted, but
17
18 the PeptideAtlas peptide spectrum must also be scrutinized and compared with a
19
20 synthetic peptide spectrum in accordance with the above guidelines with all evidence
21
22 presented in the paper.
23
24
25
26
27
28
29

30 **Changes to the HPP PE2,3,4 missing protein strategy**

31

32
33 The current basic process by which the HPP investigators manage the process of reducing
34
35 the number of missing proteins of the human proteome, herein called the “HPP pipeline”,
36
37 begins with the collection of MS data sets from the global community and deposition in
38
39 one of the ProteomeXchange repositories. The vast majority of data sets are deposited
40
41 into PRIDE^{36,37}, with some routed through MassIVE³⁸ and jPOST³⁹. These data sets may
42
43 come from experiments presented in HPP special issues such as this issue, or from
44
45 experiments performed by other members of the community in pursuit of their own
46
47 research objectives. After ProteomeXchange deposition, PeptideAtlas collects raw MS
48
49 data files and reprocesses those data using the tools of the Trans-Proteomic Pipeline
50
51 (TPP)⁴⁰⁻⁴². Thresholds are set extremely high in PeptideAtlas in order to obtain a 1%
52
53 protein-level FDR across the ensemble of all data sets. In November each year,
54
55
56
57
58
59
60

PeptideAtlas stops processing new data sets and creates an annual build reflecting the current state of the human proteome from MS evidence. In December the final peptide list is transferred to neXtProt for integration into neXtProt's next build/release based on their import of the most recent version of the human proteome from UniProtKB/Swiss-Prot. While all peptides that pass thresholds are visible in PeptideAtlas and neXtProt, only the proteins with two uniquely-mapping non-nested peptides with length 9 amino acid (AA) or greater, as called by neXtProt, are deemed to have sufficient evidence to be labeled as confidently detected PE1 proteins by MS methods. Therefore, neXtProt is the final arbiter to decide if a PE2,3,4 protein in UniProtKB is deemed PE1 in neXtProt and released as such for HUPO's HPP. Figure 1 provides a graphical summary of the current HPP pipeline.

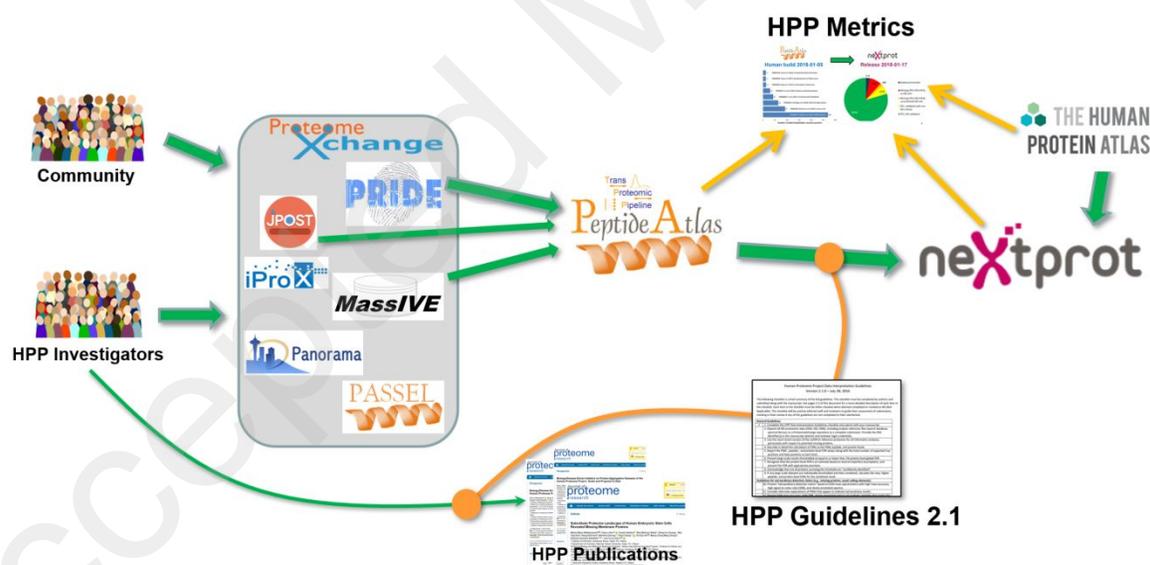


Figure 1. Overview of the 2019 HPP pipeline for data integration. HPP investigators publish their results constrained by the HPP guidelines. The data sets from these publications as well as other data sets from the community flow into the ProteomeXchange repositories. Currently a subset of the data sets from PRIDE, MassIVE, and JPOST are reprocessed by PeptideAtlas, the results of which are transferred to neXtProt constrained by the HPP guidelines. Information from PeptideAtlas, neXtProt, and Human Protein Atlas is summarized yearly in the HPP Metrics summary (this issue). Data from the Human Protein Atlas is also

1
2
3 transferred to and reprocessed by neXtProt as part of the HPP data cycle, although they are
4 not yet used to change PE status.
5

6
7 The group discussed several ambiguities and refinements of this process and made
8
9 recommendations/decisions on how the HPP pipeline will be defined for the next few
10
11 years. The group also sought to clarify some terminology pertaining to peptides relevant
12
13 to the HPP Pipeline process, most notably the term “stranded peptide”, which has been
14
15 used in several different (possibly confusing) contexts in the past^{43,44}. After considerable
16
17 discussion, it was resolved that the term “stranded peptide” shall specifically refer to “a
18
19 peptide that meets the minimum length and mapping uniqueness requirements and has
20
21 publicly available evidence for its detection via MS, but the evidence is not within the
22
23 HPP Pipeline”. In order to become unstranded, this publicly available information must
24
25 be captured and validated by the HPP Pipeline. In addition, the term “singleton peptide”
26
27 shall refer to a peptide that meets the minimum length and mapping uniqueness
28
29 requirements but does not have the needed additional partner peptide to achieve the full
30
31 requirements for two non-nested peptides. Stranded peptides may be singletons or not;
32
33 singleton peptides may be stranded or not. This terminology is used further below.
34
35
36
37
38
39
40

41 The first refinement is for how SAAVs are handled with respect to mapping uniqueness.
42
43 The fundamental question is what degree of mutation should be considered when
44
45 mapping potentially uniquely-mapping peptides to the proteome. Should all SAAVs in
46
47 neXtProt be considered when mapping peptides, and in all permutations (e.g., if there are
48
49 three annotated mutation sites in a single peptide, should mapping all three residues that
50
51 are mutated be considered, or should just one at a time be considered)? Despite
52
53
54
55 substantial diversity in opinion, the consensus was that co-occurrence of nearby SAAVs
56
57
58
59
60

1
2
3 was very low, and therefore simply considering one mutation per peptide was sufficient.
4
5 All mutations in neXtProt, except for the somatic mutations from COSMIC⁴⁵, will be
6
7 considered during mapping of peptides to proteins.
8
9

10
11
12 The group discussed whether there should be some formal adjustment to the lower limit
13
14 requirement of two peptides of nine amino acids or longer. These requirements are
15
16 designed to ensure a certain level of confidence in the MS detection of missing proteins,
17
18 but this level was never really quantified in a way to justify that 2×9 should be
19
20 sufficient, but (2×8) or (3×8) or $(1 \times 9) + (1 \times 8) + (1 \times 7)$ should not. An example of
21
22 the latter comes in the form of the current state of the protein Q8N688 β -defensin 123,
23
24 which has multiple detections of 7 AA, 8 AA, and 9 AA peptides as shown in Figure 2
25
26 and is claimed by Wang et al⁴⁶. Because this protein is only 67 amino acids long, and the
27
28 mature form is only 47 AAs long after cleaving off the 20 AA-long signal peptide, these
29
30 are the only three tryptic peptides that can be expected. The obvious question is: should
31
32 this complete MS evidence be sufficient for PE1 status assignment? After substantial
33
34 discussion, it was decided that there would be no change to the 2×9 policy for now,
35
36 because building in a more intricate limit without a mathematical/statistical foundation
37
38 for doing so was inadvisable. It was deemed that the 2×9 policy was simple and clear
39
40 and worth retaining in the absence of a more compelling lower limit. However, two
41
42 future courses of action were recommended.
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Q8N688 Beta-defensin 123

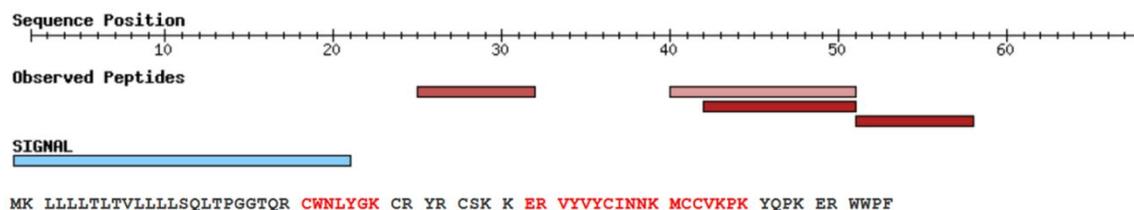


Figure 2. Depiction of the current status of Q8N688 Beta-defensin 123. The protein is only 47 AAs long after cleavage of the 20 AA signal peptide. Three distinct peptide sequences are detected (plus a fully nested peptide), but only one of the three meets guideline length requirements. Yet, all of the expected tryptic peptides (plus one missed cleavage product) are detected with excellent spectra. Should this be sufficient?

First, a sounder justification for the lower limit should be sought, perhaps one where a single probability formed the lower bound, and there might be multiple combinations that can achieve this probability. This would likely yield a per-protein metric since it is far easier and far more likely to obtain peptides that map to a very long protein than a very short one. In many cases, the use of multiple proteases might be needed to overcome the limitation of reliance on tryptic peptides, as might use of semi-tryptic and N or C-terminal peptides (see below and the 2018 HPP metrics³).

Second, guidelines v2.1 contained an “exceptions clause” for consideration of special cases. However, no mechanism was defined or implemented to deal with these special cases until now. The group recommended that a dedicated expert panel be formed by the HPP Knowledge Base Pillar Committee to judge whether particular proteins (including short proteins) that do not meet the guidelines precisely as written may indeed have sufficient evidence to meet the HPP’s desired level of confidence for PE1 status assignment. For each of the proteins recommended as candidates for elevation to PE1 without the minimum MS evidence, the panel would review the available spectra and prospects for obtaining additional MS evidence. In some cases, useful confirmatory non-

1
2
3 MS evidence may exist. If the obtainable evidence is excellent despite not meeting the
4
5 guidelines and further MS evidence is deemed unlikely, such proteins could be proposed
6
7 by the panel to neXtProt for assessment as PE1. β -defensin 123 (gene name DEFB123)
8
9 shown above was a prime initial exemplar candidate for the expert panel to consider.
10
11
12
13
14

15 The group discussed whether there are any proteins that should be declared too difficult
16
17 and unachievable, and should therefore be simply removed from the denominator of the
18
19 ratio describing the fraction of detectable proteins in the human proteome identified as
20
21 PE1 proteins. As an example, there are 15 proteins which **cannot** generate two uniquely
22
23 mapping 2×9 peptides even when using a series of five different common proteolytic
24
25 enzymes (trypsin, LysargiNase, GluC, AspN and chymotrypsin). See Supplemental Table
26
27 1 for a list of these proteins. Should such proteins be declared unattainable with MS
28
29 technologies? Remarkably, of these 15, nine are already designated as PE1, one of which
30
31 (C9JFL3) has remarkably good non-protease-specific peptides from N and C termini as
32
33 depicted in Figure 3. This is common when a protein is highly abundant in a sample.
34
35
36 Thus, the group decided that no proteins would be declared too difficult now since, if
37
38 enriched or purified to sufficient abundance, many might be accessed from the termini
39
40 with the aid of non-specific cleavages (e.g., through carboxypeptidases). Enrichment of
41
42 PTM-containing proteins, such as shown with SUMOylation^{3,47} may also be especially
43
44 effective here.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

C9JFL3 Proline, histidine and glycine-rich protein 1

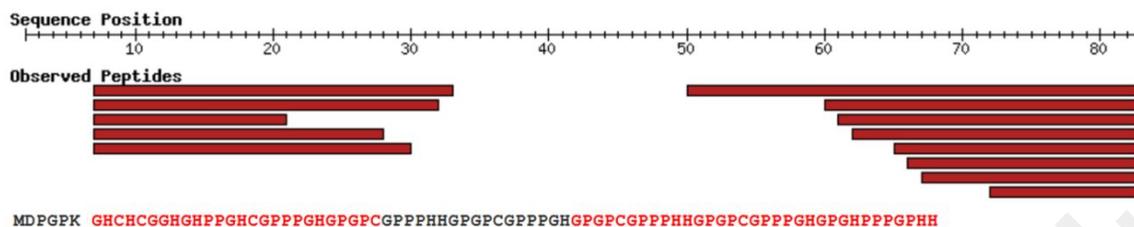


Figure 3. Depiction of the current status of C9JFL3, currently annotated as "Proline, histidine and glycine-rich protein 1". The protein is 83 amino acids long, but produces no useful fully tryptic peptides—only one that is too short and one that is too long. Yet, due to its high abundance in some samples, many miscleaved peptides are detected, easily providing the minimum evidence. The red bars indicate well detected peptides in PeptideAtlas. Multiple semi-tryptic peptides originate from the only cleavage site after the sixth amino acid. Multiple non-tryptic peptides originate from the C terminus.

A related and difficult class of proteins is the olfactory receptors. There are four PE1 entries based on non-MS evidence, and 401 PE2,3,4 entries that are annotated as being olfactory receptors in neXtProt. None of these has the requisite 2×9 uniquely mapping peptides found in PeptideAtlas. Of the 401 entries, 15 do have a single peptide mapping to them. However, a manual inspection of the available spectral evidence indicates that none of these can be called a solid detection. In most cases spectra are questionable or too short to be confident about the mapping. In all, the meager evidence for olfactory receptor proteins is far more consistent with false positives than real MS detections. This is perhaps moot since the evidence as is does not meet the guidelines but serves as an important reminder that PeptideAtlas does contain some false positives, and additional stringency of multiple detections and expert review of spectra are required for high confidence. The null hypothesis therefore remains that, among all 1500+ samples collated in the PeptideAtlas build, there are zero credible detections of olfactory receptors despite some previous hints to the contrary⁴⁸. Interestingly, if it is accepted that there have been zero credible detections of olfactory receptors via MS, one can use the putative matches

1
2
3 to olfactory receptors in any data set to provide an independent estimate of the true FDR
4 of the data set. In any case, after substantial discussion, the group felt that, although
5
6 detection of olfactory receptors by MS has proven to be extremely refractory⁴⁹, it should
7
8 not be insurmountable, and efforts should continue. The successful detection will likely
9
10 require isolation of the most appropriate olfactory cilia membrane samples, high levels of
11
12 detergent to free these proteins from the membrane, enrichment with affinity reagents,
13
14 and finally detection via MS of the enriched protein sample — a difficult challenge
15
16 indeed. If transcript levels are extremely low and expression of any single of the ~400
17
18 olfactory receptors is tightly limited to only one receptor in any cell at any time, detection
19
20 may be not feasible due to limit of detection of current MS and possibly antibody-based
21
22 methods.
23
24
25
26
27
28
29

30
31 The final major proposed change to the HPP pipeline is the addition of MassIVE-KB³⁸ to
32
33 the workflow. Whereas the current HPP pipeline (as described above and shown in
34
35 Figure 1) includes only PeptideAtlas as the data set reprocessing engine, it was agreed
36
37 that adding MassIVE-KB as a second reprocessing engine may have substantial benefits.
38
39 While data sets reprocessed in both PeptideAtlas and MassIVE-KB have substantial
40
41 overlap, this is not 100% and since MassIVE-KB has similar stringency criteria as
42
43 PeptideAtlas, HPP output quality levels would be expected to be similar. Yet, it is known
44
45 that there are protein detections in MassIVE-KB that meet the same criteria used by
46
47 PeptideAtlas and neXtProt that should be captured by the HPP Pipeline³⁸. To guard
48
49 against the possible doubling of FDR by combining these resources, the HPP Pipeline
50
51 will require that minimal evidence for a PE1 protein (i.e. two uniquely mapping non-
52
53
54
55
56
57
58
59
60

1
2
3 nested peptides of length 9AA or more) must come from either PeptideAtlas or Massive-
4
5 KB, but not a mixture of peptides from each. In other words, combining a singleton
6
7 peptide from one resource with a singleton from the other resource will not be deemed
8
9 sufficient until all evidence is reprocessed and validated by a single resource within the
10
11 HPP Pipeline.
12
13
14
15
16
17

18 **Conclusion**

19
20
21 As the HPP approaches one of its major initial goals (achieving credible detection of all
22
23 proteins coded by the human genome), the HPP MS Data Interpretation Guidelines that
24
25 served the project well since 2016 have now been clarified and enhanced with broad
26
27 consensus of the HPP leadership. These revisions address some previous ambiguities that
28
29 have emerged and address issues that seemed insignificant when the goal was distant.
30
31 The new guidelines provide an enhanced framework for ensuring that the evidence used
32
33 to substantiate future protein detection claims remains of very high quality. As such we
34
35 trust they will help guide the global proteomics community on the path to missing protein
36
37 discovery and functional understanding of proteins in the full biological detail of their
38
39 spatiotemporal networks, pathways, molecular complexes, transport, and localization.
40
41
42
43
44
45
46
47

48 **Supporting Information**

49
50 Supplementary Material 1: HPP Mass Spectrometry Data Interpretation Guidelines
51
52 Version 3.0 checklist and documentation.

53
54 Supplementary Table 1. neXtProt protein entries with only 0 or 1 uniquely mapping
55
56 peptides of length 9 AA or greater using any of 5 proteases.
57
58
59
60

Notes

The authors declare no competing financial interest.

Acknowledgements

This work was funded in part by the National Institutes of Health grants R01GM087221 (EWD/RLM), R24GM127667 (EWD), U54EB020406 (EWD), R01HL133135 (RLM), U19AG02312 (RLM), U54ES017885 (GSO), U24CA210967-01 (GSO), R01LM013115 (NB) and P41GM103484 (NB); National Science Foundation grants ABI-1759980 (NB), DBI-1933311 (EWD), and IOS-1922871 (EWD); Canadian Institutes of Health Research 148408 (CMO); Korean Ministry of Health and Welfare HI13C2098 (YKP); French Ministry of Higher Education, Research and Innovation, ProFI project, ANR-10-INBS-08 (YV); also in part by the National Eye Institute (NEI), National Human Genome Research Institute (NHGRI), National Heart, Lung, and Blood Institute (NHLBI), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of General Medical Sciences (NIGMS), and National Institute of Mental Health (NIMH) of the National Institutes of Health under Award Number U24HG007822 (SO) (the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health).

References

- (1) Hanash, S.; Celis, J. E. The Human Proteome Organization: A Mission to Advance Proteome Knowledge. *Mol. Cell. Proteomics MCP* **2002**, *1* (6), 413–414. <https://doi.org/10.1074/mcp.r200002-mcp200>.
- (2) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C.; Corthals, G. L.; Costello, C. E.; et al. The Human Proteome Project: Current State and Future Direction. *Mol. Cell. Proteomics MCP* **2011**. <https://doi.org/10.1074/mcp.O111.009993>.
- (3) Omenn, G. S.; Lane, L.; Overall, C. M.; Corrales, F. J.; Schwenk, J. M.; Paik, Y.-K.; Van Eyk, J. E.; Liu, S.; Snyder, M.; Baker, M. S.; et al. Progress on Identifying and Characterizing the Human Proteome: 2018 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* **2018**, *17* (12), 4031–4041. <https://doi.org/10.1021/acs.jproteome.8b00441>.
- (4) Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial Sequencing and Analysis of the Human Genome. *Nature* **2001**, *409* (6822), 860–921. <https://doi.org/10.1038/35057062>.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
- (5) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; et al. The Sequence of the Human Genome. *Science* **2001**, *291* (5507), 1304–1351. <https://doi.org/10.1126/science.1058040>.
 - (6) Omenn, G. S.; Lane, L.; Overall, C. M.; Corrales, F. J.; Schwenk, J. M.; Paik, Y.-K.; Van Eyk, J. E.; Liu, S.; Pennington, S.; Snyder, M. P.; et al. Progress on Identifying and Characterizing the Human Proteome: 2019 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* **2019**. <https://doi.org/10.1021/acs.jproteome.9b00434>.
 - (7) 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. <https://doi.org/10.1038/nbt.2152>.
 - (8) Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; Lundberg, E.; Omenn, G. S. Metrics for the Human Proteome Project 2013–2014 and Strategies for Finding Missing Proteins. *J. Proteome Res.* **2014**, *13* (1), 15–20. <https://doi.org/10.1021/pr401144x>.
 - (9) 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. <https://doi.org/10.1021/acs.jproteome.5b00499>.
 - (10) Omenn, G. S.; Lane, L.; Lundberg, E. K.; Beavis, R. C.; Overall, C. M.; Deutsch, E. W. Metrics for the Human Proteome Project 2016: Progress on Identifying and Characterizing the Human Proteome, Including Post-Translational Modifications. *J. Proteome Res.* **2016**, *15* (11), 3951–3960. <https://doi.org/10.1021/acs.jproteome.6b00511>.
 - (11) Omenn, G. S.; Lane, L.; Lundberg, E. K.; Overall, C. M.; Deutsch, E. W. Progress on the HUPO Draft Human Proteome: 2017 Metrics of the Human Proteome Project. *J. Proteome Res.* **2017**, *16* (12), 4281–4287. <https://doi.org/10.1021/acs.jproteome.7b00375>.
 - (12) 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), D764–770. <https://doi.org/10.1093/nar/gku1178>.
 - (13) 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. <https://doi.org/10.1186/gb-2004-6-1-r9>.
 - (14) 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), D655–658. <https://doi.org/10.1093/nar/gkj040>.
 - (15) 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. <https://doi.org/10.1021/acs.jproteome.5b00500>.
- (16) Uhlén, M.; Fagerberg, L.; Hallström, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A.; et al. Proteomics. Tissue-Based Map of the Human Proteome. *Science* **2015**, *347* (6220), 1260419. <https://doi.org/10.1126/science.1260419>.
- (17) 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. <https://doi.org/10.1038/nbt.2839>.
- (18) Deutsch, E. W.; Csordas, A.; Sun, Z.; Jarnuczak, A.; Perez-Riverol, Y.; Ternent, T.; Campbell, D. S.; Bernal-Llinares, M.; Okuda, S.; Kawano, S.; et al. The ProteomeXchange Consortium in 2017: Supporting the Cultural Change in Proteomics Public Data Deposition. *Nucleic Acids Res.* **2017**, *45* (D1), D1100–D1106. <https://doi.org/10.1093/nar/gkw936>.
- (19) Deutsch, E. W.; Overall, C. M.; Van Eyk, J. E.; Baker, M. S.; Paik, Y.-K.; Weintraub, S. T.; Lane, L.; Martens, L.; Vandenbrouck, Y.; Kusebauch, U.; et al. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J. Proteome Res.* **2016**, *15* (11), 3961–3970. <https://doi.org/10.1021/acs.jproteome.6b00392>.
- (20) Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; Lundberg, E.; Omenn, G. S. Metrics for the Human Proteome Project 2013–2014 and Strategies for Finding Missing Proteins. *J. Proteome Res.* **2014**, *13* (1), 15–20. <https://doi.org/10.1021/pr401144x>.
- (21) Kusebauch, U.; Campbell, D. S.; Deutsch, E. W.; Chu, C. S.; Spicer, D. A.; Brusniak, M.-Y.; Slagel, J.; Sun, Z.; Stevens, J.; Grimes, B.; et al. Human SRMATlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell* **2016**, *166* (3), 766–778. <https://doi.org/10.1016/j.cell.2016.06.041>.
- (22) Orchard, S.; Hermjakob, H.; Apweiler, R. The Proteomics Standards Initiative. *Proteomics* **2003**, *3* (7), 1374–1376. <https://doi.org/10.1002/pmic.200300496>.
- (23) 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. JAMIA* **2015**, *22* (3), 495–506. <https://doi.org/10.1093/jamia/ocv001>.
- (24) Deutsch, E. W.; Orchard, S.; Binz, P.-A.; Bittremieux, W.; Eisenacher, M.; Hermjakob, H.; Kawano, S.; Lam, H.; Mayer, G.; Menschaert, G.; et al. Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. *J. Proteome Res.* **2017**, *16* (12), 4288–4298. <https://doi.org/10.1021/acs.jproteome.7b00370>.
- (25) Picotti, P.; Aebersold, R. Selected Reaction Monitoring-Based Proteomics: Workflows, Potential, Pitfalls and Future Directions. *Nat. Methods* **2012**, *9* (6), 555–566. <https://doi.org/10.1038/nmeth.2015>.

- 1
2
3 (26) Rauniyar, N. Parallel Reaction Monitoring: A Targeted Experiment Performed
4 Using High Resolution and High Mass Accuracy Mass Spectrometry. *Int. J. Mol.*
5 *Sci.* **2015**, *16* (12), 28566–28581. <https://doi.org/10.3390/ijms161226120>.
- 6
7 (27) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.;
8 Aebersold, R. Targeted Data Extraction of the MS/MS Spectra Generated by Data-
9 Independent Acquisition: A New Concept for Consistent and Accurate Proteome
10 Analysis. *Mol. Cell. Proteomics MCP* **2012**, *11* (6), O111.016717.
11 <https://doi.org/10.1074/mcp.O111.016717>.
- 12
13 (28) Distler, U.; Kuharev, J.; Tenzer, S. Biomedical Applications of Ion Mobility-
14 Enhanced Data-Independent Acquisition-Based Label-Free Quantitative
15 Proteomics. *Expert Rev. Proteomics* **2014**, *11* (6), 675–684.
16 <https://doi.org/10.1586/14789450.2014.971114>.
- 17
18 (29) Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinović, S. M.; Schubert,
19 O. T.; Wolski, W.; Collins, B. C.; Malmström, J.; Malmström, L.; et al.
20 OpenSWATH Enables Automated, Targeted Analysis of Data-Independent
21 Acquisition MS Data. *Nat. Biotechnol.* **2014**, *32* (3), 219–223.
22 <https://doi.org/10.1038/nbt.2841>.
- 23
24 (30) Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Miladinović, S. M.; Cheng, L.-Y.;
25 Messner, S.; Ehrenberger, T.; Zanotelli, V.; Butscheid, Y.; Escher, C.; et al.
26 Extending the Limits of Quantitative Proteome Profiling with Data-Independent
27 Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver
28 Microtissues. *Mol. Cell. Proteomics MCP* **2015**, *14* (5), 1400–1410.
29 <https://doi.org/10.1074/mcp.M114.044305>.
- 30
31 (31) Tsou, C.-C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A.-C.;
32 Nesvizhskii, A. I. DIA-Umpire: Comprehensive Computational Framework for
33 Data-Independent Acquisition Proteomics. *Nat. Methods* **2015**, *12* (3), 258–264, 7
34 p following 264. <https://doi.org/10.1038/nmeth.3255>.
- 35
36 (32) Chalkley, R. J.; MacCoss, M. J.; Jaffe, J. D.; Röst, H. L. Initial Guidelines for
37 Manuscripts Employing Data-Independent Acquisition Mass Spectrometry for
38 Proteomic Analysis. *Mol. Cell. Proteomics MCP* **2019**, *18* (1), 1–2.
39 <https://doi.org/10.1074/mcp.E118.001286>.
- 40
41 (33) Schaeffer, M.; Gateau, A.; Teixeira, D.; Michel, P.-A.; Zahn-Zabal, M.; Lane, L.
42 The NeXtProt Peptide Uniqueness Checker: A Tool for the Proteomics
43 Community. *Bioinforma. Oxf. Engl.* **2017**, *33* (21), 3471–3472.
44 <https://doi.org/10.1093/bioinformatics/btx318>.
- 45
46 (34) Mendoza, L.; Deutsch, E. W.; Sun, Z.; Campbell, D. S.; Shteynberg, D. D.; Moritz,
47 R. L. Flexible and Fast Mapping of Peptides to a Proteome with ProteoMapper. *J.*
48 *Proteome Res.* **2018**, *17* (12), 4337–4344.
49 <https://doi.org/10.1021/acs.jproteome.8b00544>.
- 50
51 (35) Huesgen, P. F.; Lange, P. F.; Rogers, L. D.; Solis, N.; Eckhard, U.; Kleifeld, O.;
52 Goulas, T.; Gomis-Rüth, F. X.; Overall, C. M. LysargiNase Mirrors Trypsin for
53 Protein C-Terminal and Methylation-Site Identification. *Nat. Methods* **2015**, *12*
54 (1), 55–58. <https://doi.org/10.1038/nmeth.3177>.
- 55
56 (36) Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.;
57 Gevaert, K.; Vandekerckhove, J.; Apweiler, R. PRIDE: The Proteomics
58
59
60

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
- Identifications Database. *Proteomics* **2005**, *5* (13), 3537–3545.
<https://doi.org/10.1002/pmic.200401303>.
- (37) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; et al. The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data. *Nucleic Acids Res.* **2019**, *47* (D1), D442–D450.
<https://doi.org/10.1093/nar/gky1106>.
- (38) Pullman, B. S.; Wertz, J.; Carver, J.; Bandeira, N. ProteinExplorer: A Repository-Scale Resource for Exploration of Protein Detection in Public Mass Spectrometry Data Sets. *J. Proteome Res.* **2018**, *17* (12), 4227–4234.
<https://doi.org/10.1021/acs.jproteome.8b00496>.
- (39) Moriya, Y.; Kawano, S.; Okuda, S.; Watanabe, Y.; Matsumoto, M.; Takami, T.; Kobayashi, D.; Yamanouchi, Y.; Araki, N.; Yoshizawa, A. C.; et al. The JPOST Environment: An Integrated Proteomics Data Repository and Database. *Nucleic Acids Res.* **2019**, *47* (D1), D1218–D1224. <https://doi.org/10.1093/nar/gky899>.
- (40) Keller, A.; Eng, J.; Zhang, N.; Li, X.; Aebersold, R. A Uniform Proteomics MS/MS Analysis Platform Utilizing Open XML File Formats. *Mol. Syst. Biol.* **2005**, *1*, 2005.0017. <https://doi.org/10.1038/msb4100024>.
- (41) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazan, B.; et al. A Guided Tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10* (6), 1150–1159.
<https://doi.org/10.1002/pmic.200900375>.
- (42) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L. Trans-Proteomic Pipeline, a Standardized Data Processing Pipeline for Large-Scale Reproducible Proteomics Informatics. *Proteomics Clin. Appl.* **2015**, *9* (7–8), 745–754. <https://doi.org/10.1002/prca.201400164>.
- (43) Elguoshy, A.; Hirao, Y.; Xu, B.; Saito, S.; Quadery, A. F.; Yamamoto, K.; Mitsui, T.; Yamamoto, T.; Chromosome X Project Team of JProS. Identification and Validation of Human Missing Proteins and Peptides in Public Proteome Databases: Data Mining Strategy. *J. Proteome Res.* **2017**, *16* (12), 4403–4414.
<https://doi.org/10.1021/acs.jproteome.7b00423>.
- (44) Macron, C.; Lane, L.; Núñez Galindo, A.; Dayon, L. Deep Dive on the Proteome of Human Cerebrospinal Fluid: A Valuable Data Resource for Biomarker Discovery and Missing Protein Identification. *J. Proteome Res.* **2018**.
<https://doi.org/10.1021/acs.jproteome.8b00300>.
- (45) Forbes, S. A.; Beare, D.; Boutselakis, H.; Bamford, S.; Bindal, N.; Tate, J.; Cole, C. G.; Ward, S.; Dawson, E.; Ponting, L.; et al. COSMIC: Somatic Cancer Genetics at High-Resolution. *Nucleic Acids Res.* **2017**, *45* (D1), D777–D783.
<https://doi.org/10.1093/nar/gkw1121>.
- (46) Wang, Y.; Chen, Y.; Zhang, Y.; Wei, W.; Li, Y.; Zhang, T.; He, F.; Gao, Y.; Xu, P. Multi-Protease Strategy Identifies Three PE2 Missing Proteins in Human Testis Tissue. *J. Proteome Res.* **2017**, *16* (12), 4352–4363.
<https://doi.org/10.1021/acs.jproteome.7b00340>.
- (47) Hendriks, I. A.; Lyon, D.; Young, C.; Jensen, L. J.; Vertegaal, A. C. O.; Nielsen, M. L. Site-Specific Mapping of the Human SUMO Proteome Reveals Co-

Modification with Phosphorylation. *Nat. Struct. Mol. Biol.* **2017**, *24* (3), 325–336.
<https://doi.org/10.1038/nsmb.3366>.

(48) Ezkurdia, I.; Vázquez, J.; Valencia, A.; Tress, M. Analyzing the First Drafts of the Human Proteome. *J. Proteome Res.* **2014**, *13* (8), 3854–3855.
<https://doi.org/10.1021/pr500572z>.

(49) Hwang, H.; Jeong, J. E.; Lee, H. K.; Yun, K. N.; An, H. J.; Lee, B.; Paik, Y.-K.; Jeong, T. S.; Yee, G. T.; Kim, J. Y.; et al. Identification of Missing Proteins in Human Olfactory Epithelial Tissue by Liquid Chromatography-Tandem Mass Spectrometry. *J. Proteome Res.* **2018**.
<https://doi.org/10.1021/acs.jproteome.8b00408>.

For TOC Only

