Proteomics Standards Initiative Extended FASTA Format (PEFF)

Pierre-Alain Binz[1], Jim Shofstahl[2], Juan Antonio Vizcaíno[3], Harald Barsnes[4,14], Robert J. Chalkley[5], Gerben Menschaert[6], Emanuele Alpi[3], Karl Clauser[7], Jimmy K. Eng[8], Lydie Lane[9,16], Sean L. Seymour[10], Luis Francisco Hernández Sánchez[11,15], Gerhard Mayer[17], Martin Eisenacher[17], Yasset Perez-Riverol[3], Eugene A. Kapp[12], Luis Mendoza[13], Peter R. Baker[5], Andrew Collins[18], Tim Van Den Bossche[19], and Eric W. Deutsch[13,*]

[1] CHUV Centre Hospitalier Universitaire Vaudois, CH-1011 Lausanne 14, Switzerland
[2] Thermo Fisher Scientific, 355 River Oaks Parkway, San Jose, CA, 95134, USA
[3] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
[4] Proteomics Unit, Department of Biomedicine, University of Bergen, N-5009 Bergen, Norway
[5] University California San Francisco, San Francisco, CA 94143, USA
[6] Biobix, Department of Data Analysis and Mathematical Modelling, Ghent University, 9000 Ghent, Belgium
[7] Broad Institute, Cambridge MA, USA
[8] University of Washington, Seattle, WA, 98195, USA
[9] SIB Swiss Institute of Bioinformatics, CH-1211 Geneva 4, Switzerland
[10] Seymour Data Science, LLC, San Francisco, CA, USA
[11] K.G. Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, 5021 Bergen, Norway
[12] Walter & Eliza Hall Institute of Medical Research and the University of Melbourne, VIC, 3052, Australia
[13] Institute for Systems Biology, Seattle, WA, 98109, USA
[14] Computational Biology Unit, Department of Informatics, University of Bergen, N-5008 Bergen, Norway
[15] Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway
[16] Department of Microbiology and Molecular Medicine, Faculty of Medicine, University of Geneva, CH-1211 Geneva 4, Switzerland
[17] Ruhr University Bochum, Medical Faculty, Medizinisches Proteom-Center, D-44801 Bochum, Germany
[18] Department of Functional and Comparative Genomics, Institute of Integrated Biology, University of Liverpool, Liverpool, L69 7ZB, United Kingdom
[19] VIB-UGent Center for Medical Biotechnology, Ghent University, 9000 Ghent, Belgium

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

Abstract

Mass spectrometry-based proteomics enables the high-throughput identification and quantification of proteins, including sequence variants and post-translational modifications (PTMs), in biological samples. However, most workflows require that such variations be included in the search space used to analyze the data, and doing so remains challenging with most analysis tools. In order to facilitate the search for known sequence variants and PTMs, the Proteomics Standards Initiative (PSI) has designed and implemented the PSI Extended FASTA Format (PEFF). PEFF is based on the very popular FASTA format but adds a uniform mechanism for encoding substantially more metadata about the sequence collection as well as individual entries, including support for encoding known sequence variants, PTMs, and proteoforms. The format is very nearly backwards compatible, and as such, existing FASTA parsers will require little or no changes to be able to read PEFF files as FASTA files, although without supporting any of the extra capabilities of PEFF. PEFF is defined by a full specification document, controlled vocabulary terms, a set of example files, software libraries, and a file validator. Popular software and resources are starting to support PEFF, including the sequence search engine Comet and the knowledge bases neXtProt and UniProtKB. Widespread implementation of PEFF is expected to further enable proteogenomics and top-down proteomics applications by providing a standardized mechanism for encoding protein sequences and their known variations. All the related documentation, including the detailed file format specification and example files, are available at http://www.psidev.info/peff.

Keywords: PEFF, Proteomics Standards Initiative, PSI, file formats, standards, mass spectrometry, FASTA, proteomics, proteogenomics.

Introduction

Mass spectrometry (MS) based proteomics has become the most commonly used technique for detecting the presence of and measuring the abundance of proteins in biological samples[1]. Although there are many variations, in the most common analysis workflows, proteins extracted from a sample are digested into peptides using a protease, and the resulting peptide mixture is separated by liquid chromatography in a manner that gradually introduces charged peptide ions into a mass spectrometer. As the ions stream in, the instrument measures the *m/z* of these precursor peptide ions, fragments them into many smaller ions, and acquires mass spectra of the ensemble of fragment ions, thereby creating a digital record of the content of each injected sample[2].

The interpretation of the mass spectra thus produced from each sample requires advanced software to determine putative peptide and protein identifications, confidence metrics for those identifications, and abundance measurements based on the signal intensities[3]. The software available for such processing includes free and open-source packages written by researchers in the community, commercial offerings from the instrument vendors themselves, as well as software tools from independent companies[4]. In typical analysis strategies the spectra are analyzed by matching their peak patterns to a search space of peptide ions that may be present in the sample, either in the form of a database of possibly present protein sequences or a library of previously identified spectra. In both cases, if the exact combination of peptide sequence, amino acid modifications, and charge state is not present in the search space, then the spectrum cannot be correctly identified. Several groups have demonstrated the ability to open the search space to consider unpredicted modifications[5–9], but these strategies generally lead to an overall decrease in identifications at a given FDR threshold, so are not widely adopted in bottom-up proteomics.

Sequence database searching is still the most commonly used workflow, in which a search engine, such as Comet or X!Tandem, iterates through a list of input spectra, selects from a list of protein sequences a set of peptides that have the same precursor *m/z* within a selected tolerance, and scores each spectrum against a theoretical prediction of the fragments produced from each candidate peptide[10,11]. The most common format for this protein sequence database is the venerable FASTA format[12], a simple format that encodes an identifier, a free-text description, and the sequence for each protein. The format is very simple, used by most search engines and downstream processing tools, and is exported by nearly every purveyor of protein sequence lists. In cases where a sequence search engine does not use FASTA, there is a pre-indexing or pre-processing program to transform FASTA files into the needed format.

However, the FASTA format has several widely-recognized shortcomings. First, FASTA files cannot contain metadata about the collection itself: its origin, its production date, key assumptions and parameters used in its production, etc. Second, the description line for each entry is unstructured free text into which different file producers insert entry level metadata in a variety of ways that resists consistent interpretation by reading software packages; even the identifier of a single protein is subject to variations of parsing, making the mapping of proteins across different versions of a FASTA file difficult. Third, there is no mechanism for annotating the locations and nature of known post-translational modifications (PTMs) and sequence variants, which are becoming increasingly important in comprehensive analyses of datasets and to describe actual proteoforms. The UniProtKB/Swiss-Prot .DAT format does allow for encoding of variants and PTMs, but is not standardized or commonly used to inform database searching. A few software packages have custom mechanisms for searching for variants in knowledge bases (e.g. a second, refined search in X!Tandem[13]), but none of the implemented mechanisms are broadly accepted, much less ratified as a standard.

The Human Proteome Organization[14] (HUPO) Proteomics Standards Initiative[15,16] (PSI) has been developing and ratifying community-based standards for over 15 years[17]. The standards developed by the PSI range from formats[18] for MS input[19], mass spectrometer output[20], and output from downstream processing tools[21–25]. As proteogenomics studies become more widespread, interest in PTMs grows, and the available computational capacity expands, the deficiencies in the FASTA format have become an acute problem that would be well remedied with a community-developed enhanced standard from the PSI. All proposed standards are first subjected to the PSI Document Process[26], a three-level process of review that must be completed before any proposal is declared a ratified standard.

Here we present a new format from the PSI to address the need for an improved FASTA format, the PSI Extended FASTA Format (PEFF). In this article we first present an overview of the format, a brief description of its most salient features, and some example applications. We then describe the available PEFF resources, including the full specification, example files, format validators, software libraries, viewer applications, search engines that implement it, and data providers that already produce it. We finish with a discussion of important applications and considerations for this new format.

Format Description

The PEFF schema has two main sections as depicted in Figure 1. First is the file header section, which provides metadata about the collection itself, including support for independently describing several source databases that may be merged within one file. This section is absent from FASTA files. In PEFF files, each header line is prefixed with a "#" character (ASCII 35) so that FASTA readers -that are able to ignore comment lines beginning with "#"- can read PEFF files without software changes. In terms of readability, a space following the "#" is preferred, but not mandatory. Second is the individual sequence entries section, which appears in a similar pattern as FASTA files, albeit with more extensive and explicitly constrained annotation.

A crucial component of the PEFF schema is that a controlled vocabulary is used to specify the permitted keys in the key-value pairs encoded in a PEFF document[27]. This ensures that all values for the same concept are stored under the same key across all PEFF documents, quite unlike FASTA. There is a mechanism for formally defining custom keys to support cases where custom pipelines may wish to implement some non-standard key-value pairs.

Custom keys may be tied to concepts in other controlled vocabularies by providing a CURIE (compact URI) to that term. This is generally discouraged for publicly released files, but is available for judicious use. The PEFF controlled vocabulary keywords are stored in a special branch of the main PSI-MS controlled vocabulary[28] (https://www.ebi.ac.uk/ols/ontologies/ms), which is already widely available and extensively used in extant software and PSI formats. PTMs are encoded in PEFF with entries from the Unimod[29] or PSI-MOD[30] controlled vocabularies.

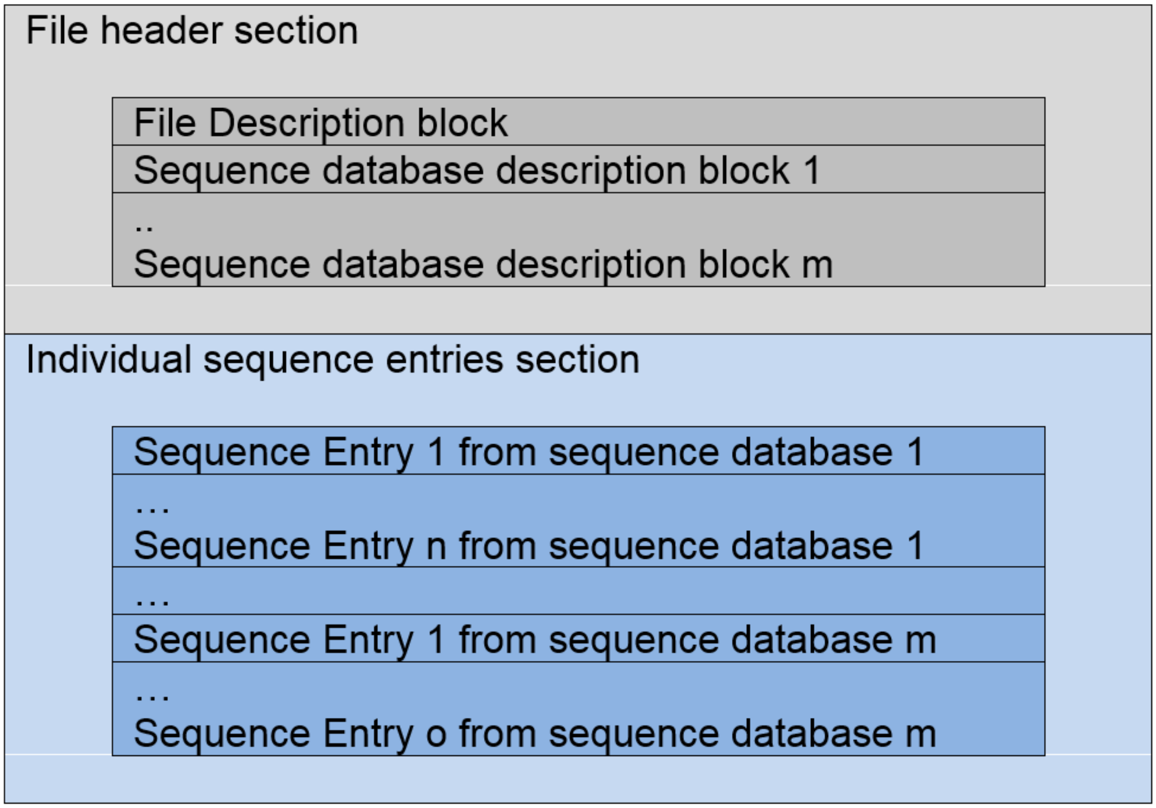

File header section
File Description block
Sequence database description block 1
..
Sequence database description block m
Individual sequence entries section
Sequence Entry 1 from sequence database 1
…
Sequence Entry n from sequence database 1
…
Sequence Entry 1 from sequence database m
…
Sequence Entry o from sequence database m

Figure 1. Overview of the PEFF schema. The file header section encodes metadata about the file itself and about the one or more sequence databases contained in the file. The individual sequence entries section encodes each of the individual sequences and the metadata associated with each entry.

The file header section has three main components. First, the preamble indicates the PEFF format version number. Second, a series of key-value pairs encodes metadata about the origin of the file. Third is a series of one or more key-value pair groups that describes each of the one or more constituent databases in the file. For example, a PEFF file may contain both neXtProt[31], RefSeq[32] sequences, and an explicit decoy sequence database in the same file and describe their origins individually.

The individual sequence entries section is essentially the same as in a FASTA file with the two main exceptions that all sequence identifiers must contain a source database prefix as defined in the file header section, and the rest of each description line is constrained to be a series of key-value pairs, where the keys are defined in the controlled vocabulary. This ensures consistent parsing by all readers that properly implement the PEFF specification. Table 1 lists an example (non-exhaustive) set of key-value pairs and their interpretation.

Example key-value	Interpretation
\PName=Tyrosine-protein kinase receptor	The full name of the protein is "Tyrosine-protein kinase receptor"
\GName=TYRO3	The source gene name for this entry is TYRO3
\TaxName=Homo sapiens \NcbiTaxId=9606	The taxonomy name associated with this entry is *Homo sapiens* and the NCBI Taxon ID is 9606
\PE=1	The UniProtKB protein evidence code for this entry is 1
\Length=890	The length of the protein entry is 890 amino acids
\Processed=(1|40|PEFF:0001021|signal peptide) (41|890|PEFF:0001020|mature protein)	The full length protein undergoes post-translational processing and amino acids 1-40 (counting 1 as the first) are separated from the rest as a signal peptide, while from 41-890 is the mature protein component
\ModResPsi=(681|MOD:00048|O4'-phospho-L-tyrosine)	Residue 681 has an annotated PTM in the source database as a phosphotyrosine, as

	fully defined as MOD:00048 within the PSI-MOD controlled vocabulary
\VariantSimple=(21|L)(68|R)	The source database annotates potential single amino-acid variants (SAAVs) at position 21 to leucine and at position 68 to arginine

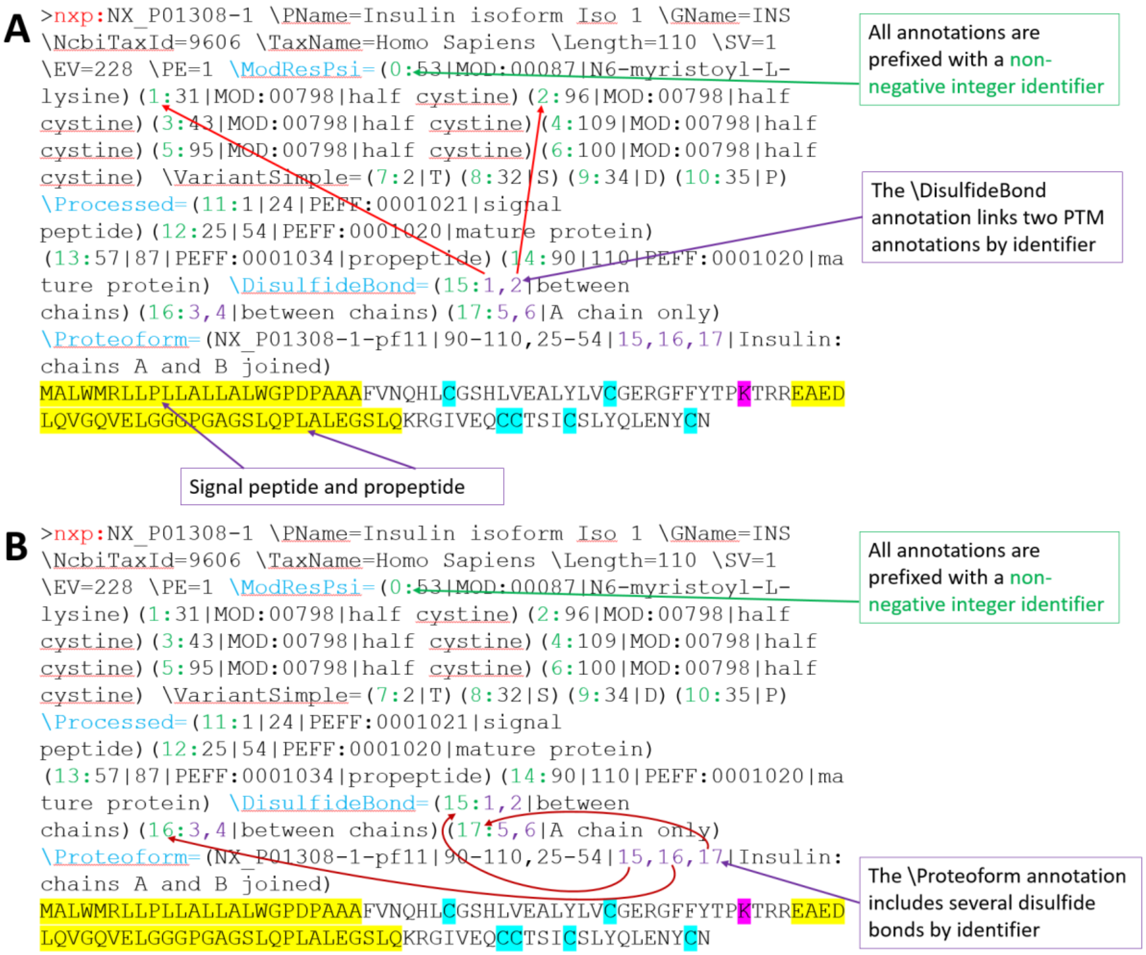
Table 1. A set of illustrative example key-value pairs that could appear in the description line of a PEFF file. All keys are defined in the PSI-MS controlled vocabulary (https://www.ebi.ac.uk/ols/ontologies/ms).

PEFF is primarily designed to encode a set of reference protein sequences and the associated collection of annotations on each protein, most commonly in the form of potential PTMs and sequence variants. However, any of the constituent databases can be defined in a PEFF header as being a database comprising proteoforms. A proteoform is defined as any one of the multitude of protein forms that can result from a single gene, including sequence variations, PTMs, and processing results[33]. There have been several other efforts to define nomenclatures, ontologies and notations for proteoforms[34,35], including the recent ProForma[36], although the latter focuses more on capturing the results of experimental analysis than being a mechanism for encoding the contents of a protein knowledge base.

There are two methods in which proteoforms can be defined in a PEFF file: the long method, wherein each entry is a different proteoform, and the compact method, wherein each entry defines a basic template and set of interchangeable annotations that may be assembled in different combinations to create multiple proteoforms per entry.

In the long method (denoted in each database header via the *isProteoformDb=true* flag), each sequence entry is required to be a single proteoform, where all key-value annotations that describe variation must apply to that sequence. For example, if five PTMs are listed, all are applicable to that specific proteoform entry. Sequence variation-defining key-value pairs are discouraged for proteoforms; however, if supplied, they must be applied. By using this extension, top-down proteomics and other similar applications can create and use a PEFF file of known proteoforms for analysis.

A more compact form is also available via the use of the *hasAnnotationIdentifiers=true* flag in the database header (*isProteoformDb=true* and *hasAnnotationIdentifiers=true* are mutually exclusive in the same database). In this form, as depicted in Figure 2, each sequence entry is a basic template with a set of potential variations, plus a special *\Proteoform* keyword that specifies which of the optional PTMs, sequence variants, disulfide bonds, and processing events should be applied to the template in combination to create individual proteoforms. In this form, the database may be used by ordinary bottom-up applications by ignoring the *\Proteoform* keyword, and also used by top-down applications by automatically expanding the proteoforms based on the listed annotation combinations.



```
>nxp:NX_P01308-1 \PName=Insulin isoform Iso 1 \GName=INS
\NcbiTaxId=9606 \TaxName=Homo Sapiens \Length=110 \SV=1
\EV=228 \PE=1 \ModResPsi=(0:53|MOD:00087|N6-myristoyl-L-
lysine)(1:31|MOD:00798|half cystine)(2:96|MOD:00798|half
cystine)(3:43|MOD:00798|half cystine)(4:109|MOD:00798|half
cystine)(5:95|MOD:00798|half cystine)(6:100|MOD:00798|half
cystine) \VariantSimple=(7:2|T)(8:32|S)(9:34|D)(10:35|P)
\Processed=(11:1|24|PEFF:0001021|signal
peptide)(12:25|54|PEFF:0001020|mature protein)
(13:57|87|PEFF:0001034|propeptide)(14:90|110|PEFF:0001020|ma
ture protein) \DisulfideBond=(15:1,2|between
chains)(16:3,4|between chains)(17:5,6|A chain only)
\Proteoform=(NX_P01308-1-pf11|90-110,25-54|15,16,17|Insulin:
chains A and B joined)
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

All annotations are prefixed with a non-negative integer identifier

The \DisulfideBond annotation links two PTM annotations by identifier

Signal peptide and propeptide

```
>nxp:NX_P01308-1 \PName=Insulin isoform Iso 1 \GName=INS
\NcbiTaxId=9606 \TaxName=Homo Sapiens \Length=110 \SV=1
\EV=228 \PE=1 \ModResPsi=(0:53|MOD:00087|N6-myristoyl-L-
lysine)(1:31|MOD:00798|half cystine)(2:96|MOD:00798|half
cystine)(3:43|MOD:00798|half cystine)(4:109|MOD:00798|half
cystine)(5:95|MOD:00798|half cystine)(6:100|MOD:00798|half
cystine) \VariantSimple=(7:2|T)(8:32|S)(9:34|D)(10:35|P)
\Processed=(11:1|24|PEFF:0001021|signal
peptide)(12:25|54|PEFF:0001020|mature protein)
(13:57|87|PEFF:0001034|propeptide)(14:90|110|PEFF:0001020|ma
ture protein) \DisulfideBond=(15:1,2|between
chains)(16:3,4|between chains)(17:5,6|A chain only)
\Proteoform=(NX_P01308-1-pf11|90-110,25-54|15,16,17|Insulin:
chains A and B joined)
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

All annotations are prefixed with a non-negative integer identifier

The \Proteoform annotation includes several disulfide bonds by identifier

Figure 2. Simplified depiction of how annotation identifiers can be referenced by other annotations to link them, such as for disulfide bonds and for proteoform definitions. Each annotation has a non-negative integer identifier, and other annotations may link to them. This example (somewhat simplified for clarity of presentation) for human insulin encodes: A) PTMs and disulfide bonds that link two PTMs; and B) a final proteoform that include two separate processed chains that are linked together via disulfide bonds.

Resources and Implementations

There are many components that help define PEFF in addition to this article, which merely provides a brief overview. Further details about PEFF can be obtained at the PSI web page for PEFF (http://www.psidev.info/peff) as well as at the GitHub repository page (https://github.com/HUPO-PSI/PEFF), where version-controlled files are managed.

The primary document is the official PEFF Format Specification (https://github.com/HUPO-PSI/PEFF/tree/master/Specification). This document has been jointly developed by the PEFF designers and subjected to the PSI Document Process in conjunction with many of the additional resources described below, prior to final ratification. The specification document presents all the details needed to implement a PEFF reader or writer successfully.

Accompanying the PEFF Format Specification is a series of example files, including a smallest possible valid PEFF file, a series of increasingly complex but human digestable examples, and a set of invalid files that can be used to test PEFF reading implementations. An important component of the PEFF Format Specification is the PEFF validator, which is able to read a PEFF file and report any warnings or errors on its adherence to the specification. The validator is available as a web application or can be downloaded at http://www.psidev.info/peff.

There is also a Perl library available for download for reading, writing, and modifying PEFF files. The Proteomics::PEFF Perl library comes with a tool that enables easy editing of PEFF files programmatically. For example, it can convert a FASTA file to a PEFF file, and it can add a series of additional PTMs or variants to individual proteins to an existing PEFF file, based on a simple tab-separated list of changes to make. The phpMs[37] toolkit (http://pgb.liv.ac.uk/phpMs) also supports the viewing and creation of PEFF files. Pyteomics 4.0[38], a proteomics software library for the Python language, supports PEFF reading. Implementations in other languages are underway. An up-to-date summary of implementations is available at http://psidev.info/peff.

The neXtProt knowledge base has been exporting PEFF files of its builds since 2015. However, it should be noted that the exports prior to February 2019 did not conform to the final PEFF Format Specification, but rather to earlier draft versions, which are subtly different. This is a natural outcome of the standards development process wherein neXtProt exported their data according to the active draft of the PEFF Format Specification to enable software testing of the format. UniProt[39] has implemented an export of its variation data using PEFF via the Proteins API[40] (https://www.ebi.ac.uk/proteins/api/doc).

The ultimate utility of PEFF will be in its implementation in proteomics search engines and downstream analysis and visualization software. As of this writing, the Comet search engine[41] has been adapted to read PEFF files (in addition to FASTA files) and process input MS data using the encoded variants and PTMs. The Trans-Proteomic Pipeline[42–44] (TPP) will soon implement PEFF in its downstream validation and visualization of data searched with Comet using PEFF input. The ProteoMapper[45] tool (http://www.peptideatlas.org/map) can search a PEFF file for a set of input peptide sequences, taking into account the protein variations encoded in PEFF. Submission of datasets to ProteomeXchange[46,47] supports the inclusion of the reference database used. Currently this usually means FASTA files; going forward, PEFF files should be similarly submitted or cited when they are used as a reference. A complete summary of supporting software and resources is available and will be maintained as tables of producers and consumers of PEFF at http://www.psidev.info/peff.

Discussion and Conclusions

The choice to expand on the basic structure of the FASTA format has not been made without dissenting opinions during the design of PEFF. Porting an existing FASTA parser to a PEFF parser will be quite easy for the most basic features. However, as the more advanced features of PEFF are parsed, the job of parsing a complex free-text format becomes considerably more difficult. Alternative encoding strategies such as a single XML (Extensible Markup Language) file and a side-car annotations file that is separate from a FASTA file were seriously considered. Parsing of complex sequence annotations from a PEFF-like XML format in general would be easier via the use of existing XML-parsing frameworks, but this requires completely new parsers and additional software dependencies. The PSI philosophy over the years has generally been to avoid side-car implementations since these types of files have a tendency to become separated from their siblings, thus causing information loss. In the end, the predominating opinion that PEFF should retain the FASTA format’s basic structure and thereby should enable a modest upgrade path for existing FASTA parsers rather than require completely new parsers prevailed.

Standard file formats are only as effective at the software that implements them. However, this precept can often be a chicken-and-egg problem in that it is often difficult to finalize a standard until it has been well tested by several implementations, and yet it is difficult to convince software developers to implement a format that has not yet been finalized. PEFF has finally achieved critical mass with one major search engine implementation (Comet) several major exporters (neXtProt and UniProt) supporting PEFF, and emerging research citing the use of PEFF in the workflow[48]. As a key point, several software libraries now support PEFF. Additionally, the Protein Prospector[6] search engine is currently in the process of implementing PEFF support (after previously supporting similar functionality with *ad hoc* formats). Therefore, we expect the number of implementing resources to expand rapidly once PEFF has been ratified by the PSI.

One of the driving applications for PEFF is proteogenomics[49], in which the variations unique to each sample from each distinct individual are important to the data analysis. In such scenarios, genomic sequencing, RNA-seq, or ribosome profiling (e.g., using PROTEOFORMER[50,51]; https://github.com/Biobix/proteoformer) will determine the variations unique to the sample, and that information will be used to create a custom sequence database specifically for that sample. PEFF provides an ideal format for this workflow. PEFF provides support for analysis workflows where nucleotide sequences are used as the primary sequence information. Each database within a PEFF file can be defined as an amino acid database or a nucleotide database. Molecule type can be mixed within a file, but not within one database. It is similarly intended that PEFF will enable top-down analysis workflows, as we better understand the full complement of proteoforms detectable in biological samples. In this context, the previously mentioned notation ProForma[36] has been recently developed by the Top Down Proteomics Consortium. Proforma uses a different style of notation that embeds the annotations into the sequence. We have not incorporated this format into the PEFF sequences component, since the proteoforms can equally be described in the PEFF format, and it is preferable not to offer several ways to encode the same information, since this increases the complexity for parsers.

The PSI is an open consortium of interested parties, and we encourage participation and critical feedback, suggestions and contributions to PEFF and other PSI formats *via* participation at PSI annual workshops, conference calls, the GitHub collaboration platform, and PSI mailing lists (see http://www.psidev.info).

Acknowledgements

We thank the many contributors to early versions of this format, including David Creasy, Matt Chambers, Phil Andrews, members of the UniProt Consortium, particularly Nicole Redaschi, Maria Jesus Martín, Claire O'Donovan, Peter McGarvey, and Amos Bairoch for mapping the proposal with UniProtKB concepts. We would like to thank Frédéric Nikitin, Pierre-André Michel, Monique Zahn and the whole neXtProt team at SIB for the PEFF implementation in neXtProt. We also want to thank Edward Turner, Leonardo Gonzales, Guoying Qi, Andrew Nightingale and Jie Luo from the UniProt development team at EMBL-EBI for the PEFF implementation in the Proteins API. EWD is supported in part by the National Institutes of Health (NIH) grants R24GM127667, R01GM087221, U19AG023122, and U54EB020406. LFHS is supported by the European Research Council and the Research Council of Norway. HB is supported by the Research Council of Norway and the Bergen Research Foundation. LL is supported by the SIB Swiss Institute of

Bioinformatics. EA is supported by the NIH under the UniProt grant U24HG007822 and by the European Molecular Biology Laboratory (EMBL). GM is funded by the BMBF grant de.NBI - German Network for Bioinformatics Infrastructure (FKZ 031 A 534A). The funding of ME is related to PURE and VALIBIO, projects of Northrhine-Westphalia. JAV wants to acknowledge funding from EMBL, the Wellcome Trust (grants WT101477MA and 208391/Z/17/Z) and NIH (grant R24GM127667). RJC and PRB are supported by the Adelson Medical Research Foundation.

Supporting Information

None

References

(1) 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. https://doi.org/10.1038/nmeth0910-681.

(2) Aebersold, R.; Mann, M. Mass Spectrometry-Based Proteomics. *Nature* **2003**, *422* (6928), 198–207. https://doi.org/10.1038/nature01511.

(3) 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. https://doi.org/10.1152/physiolgenomics.00298.2007.

(4) 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. https://doi.org/10.1016/j.jprot.2010.08.009.

(5) Tsur, D.; Tanner, S.; Zandi, E.; Bafna, V.; Pevzner, P. A. Identification of Post-Translational Modifications by Blind Search of Mass Spectra. *Nat. Biotechnol.* **2005**, *23* (12), 1562–1567. https://doi.org/10.1038/nbt1168.

(6) Chalkley, R. J.; Baker, P. R.; Medzihradszky, K. F.; Lynn, A. J.; Burlingame, A. L. In-Depth Analysis of Tandem Mass Spectrometry Data from Disparate Instrument Types. *Mol. Cell. Proteomics MCP* **2008**, *7* (12), 2386–2398. https://doi.org/10.1074/mcp.M800021-MCP200.

(7) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun Proteomics as Modified Peptides. *Nat. Biotechnol.* **2015**, *33* (7), 743–749. https://doi.org/10.1038/nbt.3267.

(8) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry-Based Proteomics. *Nat. Methods* **2017**, *14* (5), 513–520. https://doi.org/10.1038/nmeth.4256.

(9) Bittremieux, W.; Meysman, P.; Noble, W. S.; Laukens, K. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. *J. Proteome Res.* **2018**, *17* (10), 3463–3474. https://doi.org/10.1021/acs.jproteome.8b00359.

(10) Eng, J. K.; Searle, B. C.; Clauser, K. R.; Tabb, D. L. A Face in the Crowd: Recognizing Peptides through Database Search. *Mol. Cell. Proteomics MCP* **2011**, *10* (11), R111.009522. https://doi.org/10.1074/mcp.R111.009522.

(11) Verheggen, K.; Raeder, H.; Berven, F. S.; Martens, L.; Barsnes, H.; Vaudel, M. Anatomy and Evolution of Database Search Engines-a Central Component of Mass Spectrometry Based Proteomic Workflows. *Mass Spectrom. Rev.* **2017**. https://doi.org/10.1002/mas.21543.

(12) Pearson, W. R.; Lipman, D. J. Improved Tools for Biological Sequence Comparison. *Proc. Natl. Acad. Sci. U. S. A.* **1988**, *85* (8), 2444–2448.

(13) Craig, R.; Beavis, R. C. TANDEM: Matching Proteins with Tandem Mass Spectra. *Bioinforma. Oxf. Engl.* **2004**, *20* (9), 1466–1467. https://doi.org/10.1093/bioinformatics/bth092.

(14) Hanash, S.; Celis, J. E. The Human Proteome Organization: A Mission to Advance Proteome Knowledge. *Mol. Cell. Proteomics MCP* **2002**, *1* (6), 413–414.

(15) Orchard, S.; Hermjakob, H.; Apweiler, R. The Proteomics Standards Initiative. *Proteomics* **2003**, *3* (7), 1374–1376. https://doi.org/10.1002/pmic.200300496.

(16) 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.

(17) 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**. https://doi.org/10.1021/acs.jproteome.7b00370.

(18) Deutsch, E. W. File Formats Commonly Used in Mass Spectrometry Proteomics. *Mol. Cell. Proteomics MCP* **2012**, *11* (12), 1612–1621. https://doi.org/10.1074/mcp.R112.019695.

(19) Deutsch, E. W.; Chambers, M.; Neumann, S.; Levander, F.; Binz, P.-A.; Shofstahl, J.; Campbell, D. S.; Mendoza, L.; Ovelleiro, D.; Helsens, K.; et al. TraML--a Standard Format for Exchange of Selected Reaction Monitoring Transition Lists. *Mol. Cell. Proteomics MCP* **2012**, *11* (4), R111.015040. https://doi.org/10.1074/mcp.R111.015040.

(20) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; et al. MzML--a Community Standard for Mass Spectrometry Data. *Mol. Cell. Proteomics MCP* **2011**, *10* (1), R110.000133. https://doi.org/10.1074/mcp.R110.000133.

(21) 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 MCP* **2012**, *11* (7), M111.014381. https://doi.org/10.1074/mcp.M111.014381.

(22) Vizcaíno, J. A.; Mayer, G.; Perkins, S.; Barsnes, H.; Vaudel, M.; Perez-Riverol, Y.; Ternent, T.; Uszkoreit, J.; Eisenacher, M.; Fischer, L.; et al. The MzIdentML Data Standard Version 1.2, Supporting Advances in Proteome Informatics. *Mol. Cell. Proteomics MCP* **2017**, *16* (7), 1275–1285. https://doi.org/10.1074/mcp.M117.068429.

(23) Griss, J.; Jones, A. R.; Sachsenberg, T.; Walzer, M.; Gatto, L.; Hartler, J.; Thallinger, G. G.; Salek, R. M.; Steinbeck, C.; Neuhauser, N.; et al. The MzTab Data Exchange Format: Communicating Mass-Spectrometry-Based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Mol. Cell. Proteomics MCP* **2014**, *13* (10), 2765–2775. https://doi.org/10.1074/mcp.O113.036681.

(24) Walzer, M.; Qi, D.; Mayer, G.; Uszkoreit, J.; Eisenacher, M.; Sachsenberg, T.; Gonzalez-Galarza, F. F.; Fan, J.; Bessant, C.; Deutsch, E. W.; et al. The MzQuantML Data Standard for Mass Spectrometry-Based Quantitative Studies in Proteomics. *Mol. Cell. Proteomics MCP* **2013**, *12* (8), 2332–2340. https://doi.org/10.1074/mcp.O113.028506.

(25) Menschaert, G.; Wang, X.; Jones, A. R.; Ghali, F.; Fenyö, D.; Olexiouk, V.; Zhang, B.; Deutsch, E. W.; Ternent, T.; Vizcaíno, J. A. The ProBAM and ProBed Standard

Formats: Enabling a Seamless Integration of Genomics and Proteomics Data. *Genome Biol.* **2018**, *19* (1), 12. https://doi.org/10.1186/s13059-017-1377-x.

(26) Vizcaíno, J. A.; Martens, L.; Hermjakob, H.; Julian, R. K.; Paton, N. W. The PSI Formal Document Process and Its Implementation on the PSI Website. *Proteomics* **2007**, *7* (14), 2355–2357. https://doi.org/10.1002/pmic.200700064.

(27) Mayer, G.; Jones, A. R.; Binz, P.-A.; Deutsch, E. W.; Orchard, S.; Montecchi-Palazzi, L.; Vizcaíno, J. A.; Hermjakob, H.; Oveillero, D.; Julian, R.; et al. Controlled Vocabularies and Ontologies in Proteomics: Overview, Principles and Practice. *Biochim. Biophys. Acta* **2014**, *1844* (1 Pt A), 98–107. https://doi.org/10.1016/j.bbapap.2013.02.017.

(28) Mayer, G.; Montecchi-Palazzi, L.; Ovelleiro, D.; Jones, A. R.; Binz, P.-A.; Deutsch, E. W.; Chambers, M.; Kallhardt, M.; Levander, F.; Shofstahl, J.; et al. The HUPO Proteomics Standards Initiative- Mass Spectrometry Controlled Vocabulary. *Database J. Biol. Databases Curation* **2013**, *2013*, bat009. https://doi.org/10.1093/database/bat009.

(29) Creasy, D. M.; Cottrell, J. S. Unimod: Protein Modifications for Mass Spectrometry. *Proteomics* **2004**, *4* (6), 1534–1536. https://doi.org/10.1002/pmic.200300744.

(30) Montecchi-Palazzi, L.; Beavis, R.; Binz, P.-A.; Chalkley, R. J.; Cottrell, J.; Creasy, D.; Shofstahl, J.; Seymour, S. L.; Garavelli, J. S. The PSI-MOD Community Standard for Representation of Protein Modification Data. *Nat. Biotechnol.* **2008**, *26* (8), 864–866. https://doi.org/10.1038/nbt0808-864.

(31) Gaudet, P.; Michel, P.-A.; Zahn-Zabal, M.; Britan, A.; Cusin, I.; Domagalski, M.; Duek, P. D.; Gateau, A.; Gleizes, A.; Hinard, V.; et al. The NeXtProt Knowledgebase on Human Proteins: 2017 Update. *Nucleic Acids Res.* **2017**, *45* (D1), D177–D182. https://doi.org/10.1093/nar/gkw1062.

(32) O’Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. *Nucleic Acids Res.* **2016**, *44* (D1), D733-745. https://doi.org/10.1093/nar/gkv1189.

(33) Smith, L. M.; Kelleher, N. L.; Consortium for Top Down Proteomics. Proteoform: A Single Term Describing Protein Complexity. *Nat. Methods* **2013**, *10* (3), 186–187. https://doi.org/10.1038/nmeth.2369.

(34) Schlüter, H.; Apweiler, R.; Holzhütter, H.-G.; Jungblut, P. R. Finding One’s Way in Proteomics: A Protein Species Nomenclature. *Chem. Cent. J.* **2009**, *3*, 11. https://doi.org/10.1186/1752-153X-3-11.

(35) Natale, D. A.; Arighi, C. N.; Blake, J. A.; Bona, J.; Chen, C.; Chen, S.-C.; Christie, K. R.; Cowart, J.; D’Eustachio, P.; Diehl, A. D.; et al. Protein Ontology (PRO): Enhancing and Scaling up the Representation of Protein Entities. *Nucleic Acids Res.* **2017**, *45* (D1), D339–D346. https://doi.org/10.1093/nar/gkw1075.

(36) LeDuc, R. D.; Schwämmle, V.; Shortreed, M. R.; Cesnik, A. J.; Solntsev, S. K.; Shaw, J. B.; Martin, M. J.; Vizcaino, J. A.; Alpi, E.; Danis, P.; et al. ProForma: A Standard Proteoform Notation. *J. Proteome Res.* **2018**, *17* (3), 1321–1325. https://doi.org/10.1021/acs.jproteome.7b00851.

(37) Collins, A.; Jones, A. R. PhpMs: A PHP-Based Mass Spectrometry Utilities Library. *J. Proteome Res.* **2018**, *17* (3), 1309–1313. https://doi.org/10.1021/acs.jproteome.7b00783.

(38) Levitsky, L. I.; Klein, J. A.; Ivanov, M. V.; Gorshkov, M. V. Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *J. Proteome Res.* **2019**, *18* (2), 709–714. https://doi.org/10.1021/acs.jproteome.8b00717.

(39) UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2018**. https://doi.org/10.1093/nar/gky1049.

(40) Nightingale, A.; Antunes, R.; Alpi, E.; Bursteinas, B.; Gonzales, L.; Liu, W.; Luo, J.; Qi, G.; Turner, E.; Martin, M. The Proteins API: Accessing Key Integrated Protein and Genome Information. *Nucleic Acids Res.* **2017**, *45* (W1), W539–W544. https://doi.org/10.1093/nar/gkx237.

(41) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *Proteomics* **2013**, *13* (1), 22–24. https://doi.org/10.1002/pmic.201200439.

(42) 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.

(43) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; et al. A Guided Tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10* (6), 1150–1159. https://doi.org/10.1002/pmic.200900375.

(44) 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.

(45) Mendoza, L.; Deutsch, E. W.; Sun, Z.; Campbell, D. S.; Shteynberg, D. D.; Moritz, R. L. Flexible and Fast Mapping of Peptides to a Proteome with ProteoMapper. *J. Proteome Res.* **2018**. https://doi.org/10.1021/acs.jproteome.8b00544.

(46) 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.

(47) 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.

(48) Wippel, H. H.; Santos, M. D. M.; Clasen, M. A.; Kurt, L. U.; Nogueira, F. C. S.; Carvalho, C. E.; McCormick, T. M.; Neto, G. P. B.; Alves, L. R.; da Gloria da Costa Carvalho, M.; et al. Comparing Intestinal versus Diffuse Gastric Cancer Using a PEFF-Oriented Proteomic Pipeline. *J. Proteomics* **2017**. https://doi.org/10.1016/j.jprot.2017.10.005.

(49) Nesvizhskii, A. I. Proteogenomics: Concepts, Applications and Computational Strategies. *Nat. Methods* **2014**, *11* (11), 1114–1125. https://doi.org/10.1038/nmeth.3144.

(50) Crappé, J.; Ndah, E.; Koch, A.; Steyaert, S.; Gawron, D.; De Keulenaer, S.; De Meester, E.; De Meyer, T.; Van Criekinge, W.; Van Damme, P.; et al. PROTEOFORMER: Deep Proteome Coverage through Ribosome Profiling and MS Integration. *Nucleic Acids Res.* **2015**, *43* (5), e29. https://doi.org/10.1093/nar/gku1283.

(51) Verbruggen, S.; Ndah, E.; Van Criekinge, W.; Gessulat, S.; Kuster, B.; Wilhelm, M.; Van Damme, P.; Menschaert, G. PROTEOFORMER 2.0: Further Developments in the Ribosome Profiling-Assisted Proteogenomic Hunt for New Proteoforms. *Mol. Cell. Proteomics MCP* **2019**. https://doi.org/10.1074/mcp.RA118.001218.