PSI Recommendation
PSI Proteomics Informatics Workgroup
Version rc-1.0.0, DRAFT

Tobias Ternent, European Bioinformatics Institute
Fawaz Ghali, University of Liverpool
David Fenyo, New York University Medical School
Andrew R Jones, University of Liverpool
Juan Antonio Vizcaíno, European Bioinformatics Institute

February 15, 2017

proBed: extension of the BED format for mapping peptides identified by mass spectrometry to a genome

Status of This Document

This document presents the version rc-1.0.0 specification of the proBed data format developed by members of the Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI) Proteomics Informatics (PI) Working Group. Distribution is unlimited.

Version of This Document
The current version of this document is rc-1.0.0, February 2017.

Abstract

The Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI) defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification. The Proteomics Informatics Working Group is developing standards for describing the results of identification and quantification processes for proteins, peptides, small molecules and protein modifications from mass spectrometry. This document defines a tab delimited text file format to report “proteogenomics” results i.e. the identification and mapping of peptide/protein sequences back against a genome, to assist in annotation efforts.

Contents

Abstract 1

1. Short Summary 4

2. Introduction 5

2.1 Background 5

2.2 Document Structure 5

3. Use Cases for proBed 5

4. Notational Conventions 6

5. Relationship to Other Specifications 6

5.1 The PSI Mass Spectrometry Controlled Vocabulary (CV) 7

6. Format specification 8

6.1 General details of the format specification 8

6.2 Null values and Data types 9

6.3 Header line 9

6.4 BED format standard fields 10

6.4.1 chrom 10

6.4.2 chromStart 10

6.4.3 chromEnd 11

6.4.4 name 11

6.4.5 score 12

6.4.6 strand 12

6.4.7 thickStart 12

6.4.8 thickEnd 13

6.4.9 reserved 13

6.4.10 blockCount 14

6.4.11 blockSizes 14

6.4.12 chromStarts 14

6.5 proBed specific fields 15

6.5.1 proteinAccession 15

6.5.2 peptideSequence 15

6.5.3 uniqueness 16

6.5.4 genomeReferenceVersion 17

6.5.5 psmScore 18

6.5.6 fdr 18
6.5.7 modifications 19
6.5.8 charge 20
6.5.9 expMassToCharge 20
6.5.10 calcMassToCharge 20
6.5.11 psmRank 21
6.5.12 datasetID 21
6.5.13 uri 21
6.6 Additional optional fields in proBed 22
6.7 How to represent intron and exon regions 22
6.8 Representation of peptides (as groups of PSMs) in proBed 23
6.9 Merging of proBed files 24
6.10 Other supporting materials 25
7. Conclusions 26
8. Authors 27
9. Contributors 28
10. References 28
11. Appendix I: proBed to bigBed conversion 29
11.1 Sorted proBed file 29
11.2 Supporting autoSQL file 29
11.3 Chromosome names file 31
11.4 Running the bigBed conversion tool 31
12. Intellectual Property Statement 32
TradeMark Section 32
Copyright Notice 32

1. Short Summary

The proBed specification describes a file format based upon the original BED format (1). Data are represented in lines of the original 12 BED columns, plus another 13 columns to report proteogenomics results: the identification and mapping of peptide/protein sequences back against a given genome. See the following table below for a quick summary of the proBed format fields in the order they should be used.

Datatype	Field name	Description	Origin
string	chrom	Reference sequence chromosome	BED
uint	chromStart	Start position of the first DNA base	BED
uint	chromEnd	End position of the last DNA base	BED
string	name	Unique name	BED
uint	score	Score	BED
char[1]	strand	+ or - for strand	BED
uint	thickStart	Coding region start	BED
uint	thickEnd	Coding region end	BED
uint	reserved	Always 0	BED
int	blockCount	Number of blocks	BED
int[blockCount]	blockSizes	Block sizes	BED
int[blockCount]	chromStarts	Block starts	BED
string	*proteinAccession*	*Protein accession number*	*proBed*
string	*peptideSequence*	*Peptide sequence*	*proBed*
string	*uniqueness*	*Peptide uniqueness*	*proBed*
string	*genomeReferenceVersion*	*Genome reference version number*	*proBed*
double	*psmScore*	*PSM score*	*proBed*
double	*fdr*	*Estimated global false discovery rate*	*proBed*
string	*modifications*	*Post-translational modifications*	*proBed*
int	*charge*	*Charge value*	*proBed*
double	*expMassToCharge*	*Experimental mass to charge value*	*proBed*
double	*calcMassToCharge*	*Calculated mass to charge value*	*proBed*
int	*psmRank*	*Peptide-Spectrum Match rank.*	*proBed*
string	*datasetID*	*Dataset Identifier*	*proBed*
string	*uri*	*Uniform Resource Identifier*	*proBed*

2. Introduction

2.1 Background

This document addresses the systematic description of peptide identification data retrieved from mass spectrometry (MS)-based experiments mapped to the genome. The original BED format (Browser Extensive Data, https://genome.ucsc.edu/FAQ/FAQformat.html - format1), developed by the UCSC (University of California, Santa Cruz) team, is used to describe genome coordinate data across lines, for use on annotation tracks (1).

In BED, data lines are formatted in plain text with white-space separated fields. Each data line represents one item mapped to the genome. The first three fields (genomic coordinates) are mandatory, and an additional 9 fields are standardized and commonly interpreted by genome browsers and other tools, totalling 12 “BED” fields, re-used here. The proBed format includes a further 13 fields to describe information primarily on peptide-spectrum matches (PSMs). The format can also accommodate peptides (as groups of PSMs), but in that case, some assumptions need to be taken in some of the fields (see Section 6).

Other variants of the BED format exist such as BigBed (2), a binary format based on BED, which represents a feasible way to store the same information present in BED in compressed binary files.

This document presents a specification, not a tutorial. As such, the presentation of technical details is deliberately direct. The role of the text is to describe the model and justify design decisions made. The document does not discuss how the models should be used in practice, consider tool support for data capture or storage, or provide comprehensive examples of the models in use. It is anticipated that tutorial material will be developed independently of this specification.

2.2 Document Structure

The remainder of this document is structured as follows. Section 3 lists use cases for the format. Section 4 is devoted to Notational Conventions throughout the document. Section 5 outlines the relationships between proBed and other file format specifications. Section 6 includes all the details of the format specification, listing all the required and optional fields. Section 6 is a brief summary of the conclusions. Sections 8, 9 and 10 are devoted to the list of Authors, contributors and references, respectively. There is one Appendix (section 11) devoted to illustrate how proBed files can be converted to the binary format bidBed, the most frequently used in annotation tracks. The last section of this document (section 12) contains the Intellectual Property Statement.

3. Use Cases for proBed

The following cases of usage have driven the development of the proBed data model, and are used to define the scope of the format in version 1.0.

1. proBed files should report genome coordinates for peptides identified from LC-MS/MS workflows. This will enable proteomics data to be more accessible in genome annotation and visualisation approaches.
2. It should be possible to export proteogenomic data from PSI standard formats, for example, mzIdentML/mzTab files into proBed.
3. It should be possible to open proBed files with “standard” software such as Microsoft Excel® or Open Office Spreadsheet. This should facilitate the usability of the format to people outside the fields of proteomics.
4. It should be possible for other reading software designed for regular BED files to consume proBed formatted files.
5. It should be possible to convert proBed files into the bigBed (2) format using the bedToBigBed tool (see Section 6.7 Other supporting materials, the version considered for this specification document is 2.87), as regular BED files can be. The binary bigBed files are the ones most commonly used in annotation tracks. This would enable other readers to consume an indexed compressed format for faster access, in order to read or display the data.

4. Notational Conventions

The key words “MUST,” “MUST NOT,” “REQUIRED,” “SHALL,” “SHALL NOT,” “SHOULD,” “SHOULD NOT,” “RECOMMENDED,” “MAY,” and “OPTIONAL” are to be interpreted as described in RFC-2119 (3).

5. Relationship to Other Specifications

The specification described in this document is not being developed in isolation; indeed, it is designed to be complementary to, and thus used in conjunction with, several existing and emerging models. Related specifications include the following:

Differences between proBed and proBAM. Two different file formats have been drafted by the PSI to enable interoperability between MS-based proteomics data and genome-centric data: proBed and proBAM.

Both file formats are based on existing, well-established genomics formats: BED and SAM (and its binary version BAM). Similar to the original BED format, the main purpose of the proBed format is to report genome coordinates for PSMs or peptides identified from MS workflows to be used as annotation tracks in genome-centric browsers (e.g. Ensembl, UCSC, IGV). Although the proBAM format can also be used to create annotation tracks at the PSM or peptide level, this format holds more information than proBed and represents alignment information, similar to the original SAM/BAM format.

The novel formats are designed in a way that both are as consistent as possible in presenting information that they share. The main differences are that in proBAM (1) all alignment information is available (underlying genomic sequence, CIGAR string); (2) extra peptide/PSM annotation can be presented (normal, variant, indel, etc); (3) analogously to BED and BAM, proBed is zero-based for genome coordinates whereas proBAM is one-based; and (4) more complete MS study result data can be exported in proBAM format, including both decoy information and low scoring and lower ranked PSMs. The latter point enables to use the

proBAM format to reanalyze the data using different thresholds or perform a gene level inference procedure. Both novel formats (proBed and proBAM) should be fully compatible with existing tools designed for the original BED and BAM files and proteogenomic data from MS PSI standard formats, mzIdentML/mzTab can be exported into these novel proteogenomic file formats. A proBAM to proBed conversion should be possible, but a reverse mapping (from proBed to proBAM) may not be possible in most circumstances.

Other related specifications include the following:

1. *mzIdentML* (http://www.psidev.info/mzidentml). mzIdentML is the PSI standard for capturing of peptide and protein identification data (4).
2. *mzTab (http://www.psidev.info/mztab).* mzTab is a light-weight, tab-delimited file format and PSI standard for capturing of peptide and protein identification data. mzTab files MAY reference mzIdentML files that then contain the detailed evidence of the reported identifications (5).
3. *BED (https://genome.ucsc.edu/FAQ/FAQformat.html#format1).* BED is a plain text format to define data in lines for annotation tracks, used as the basis for proBed.
4. *bigBed (https://genome.ucsc.edu/FAQ/FAQformat.html#format1.5).* bigBed is an indexed compressed binary format to represent the same data as the BED format (2).

5.1 The PSI Mass Spectrometry Controlled Vocabulary (CV)

The PSI-MS controlled vocabulary (6) is intended to provide terms for annotation of proBed files. The CV has been generated with a collection of terms from software vendors and academic groups working in the area of mass spectrometry and proteome informatics. Some terms describe attributes that must be coupled with a numerical value attribute in the CvParam element (e.g. MS:1002072 “p-value”) and optionally a unit for that value (e.g. MS:1001117, “theoretical mass”, units = “dalton”). The terms that require a value are denoted by having a “datatype” key-value pair in the CV itself: MS:1001172 "mascot:expectation value" value-type:xsd:double. Terms that need to be qualified with units are denoted with a “has_units” key in the CV itself (relationship: has_units: UO:0000221 ! dalton).

As recommended by the PSI CV guidelines, psi-ms.obo should be dynamically maintained via the psidev-ms-vocab@lists.sourceforge.net mailing list that allows any user to request new terms, in agreement with the community involved. Once a consensus is reached among the community the new terms are added within a few business days.

In general, modifications SHOULD be sourced from Unimod (http://www.unimod.org/obo/unimod.obo) where possible.

6. Format specification

This section describes the structure of a proBed file.

6.1 General details of the format specification

A single proBed file SHOULD contain peptide or PSM-level identifications that have been mapped against one or more sets of gene models. A valid file MAY contain results mapped against only the “official gene models” or MAY contain results mapped to multiple sets of different gene models, so long as they have consistent chromosomal locations i.e. different gene model sets were produced from the same genome sequence release. There are no restrictions on the scope of one proBed file as containing one or more “experimental units” i.e. in one proBed file, an exporter MAY include a single LC-MS/MS run; merged pre-fractionation results from the same sample; merged results from different replicates; merged results from different samples and so on (discussion on merging proBed files specifically is covered in Section 6.9). It is generally encouraged that only results passing stringent statistical thresholds, such as <1% False Discovery Rate at the PSM/peptide-level and the protein-level have been applied before export to proBed, but this is not enforced in the format specifications. It is also encouraged that a process of protein inference is performed before export to enable the uniqueness of peptide-locus maps to be assessed in a non-trivial manner, but again this is not enforced by the specifications.

- **Field separator**
  The column delimiter is the Unicode Horizontal Tab character (Unicode codepoint 0009). In the original documentation (https://genome.ucsc.edu/FAQ/FAQformat.html#format1), it is also stated that blank spaces are allowed as field separator for BED files. However, in proBed blank spaces MUST NOT be used as a separators.
- **File encoding**
  The UTF-8 encoding of the Unicode character set is the preferred encoding for proBed files. However, parsers should be able to recognize commonly used encodings.
- **Case sensitivity**
  All column labels and field names are case-sensitive.
- **Dates**
  Dates and times MUST be supplied in the ISO 8601 format (“YYYY-MM-DD”, “YYYY-MM-DDTHH:MMZ” respectively).
- **Decimal separator**
  In proBAM files the dot (“.”) MUST be used as decimal separator. Thousand separators MUST NOT be used.
- **Params**
  proBed makes use of params. All parameters SHOULD be reported as CV parameters. If no suitable CV parameters exist, we encourage users to add them to the suitable CV or ontology, e.g. PSI-MS CV. User parameters SHOULD NOT be used by default. Parameters are always reported as:

[CV label, accession, name, value]. User params only contain a name and a value. Any field that is not available MUST be left empty.

```
[MS, MS:1001207, Mascot,]
[,,A user parameter, The value]
```

In case the name of the param contains commas, quotes MUST be added to avoid problems with parsing:
[label, accession, “first part of the param name, second part of the name”, value].

```
[MOD, MOD:00648, "N,O-diacetylated L-serine",]
```

6.2 Null values and Data types

There MUST NOT be any empty fields, and some MAY be nullable by use of the full stop “.” character. In general, null “.” values SHOULD NOT be used within any field if the information is available.

Data types
Only BED data types are permitted in proBed. These include:

- “uint” – an unsigned integer.
- “int” – a signed integer.
- “string” – a sequence of characters.
- “char[<N>]” – exactly <N> number of characters. The value used MUST be a positive integer.
- “int[<COLUMN_NAME>]” – a comma separated list of integers, with a number of elements equal to the value used for the <COLUMN_NAME> field for the data line.

All fields are ordered and so MUST NOT be changed. The data type of a nullable field MUST be “string”. In the case of the proBed specific fields, the intended proBed-specific data types are indicated for some fields (int, double), which can be implemented for software validation and mapping purposes. However, for compatibility purposes with all the BED related tools, in other contexts the additional data types may be treated as more general “string” data types.

6.3 Header line

The first line of the proBed file MAY be an optional header line. If present then it MUST start with a ‘#’ character followed by *proBed-version*. It is RECOMMENDED to include the proBed version number for clarity.

Example:	*# proBed-version 1.0*

6.4 BED format standard fields

The first 12 fields are directly taken from the BED format and are unchanged. In the original BED format specification, the only fully mandatory fields are the first three (“chrom”, “chromStart” and “chromEnd”). However, optional fields can only be added to BED files after the 12 original BED fields. This is why all 12 of the core BED columns MUST be present in proBed, even if it is acknowledged that there is some redundancy in the encoding. See https://genome.ucsc.edu/FAQ/FAQformat.html - format1, for more details. The Descriptions have been expanded here to describe how proteomics results specifically should be encoded.

In the Example section for each field, an example value for the data type is given in **bold** within the context of a complete and valid line of proBed data.

6.4.1 chrom

Description:	The reference sequence chromosome, using the nomenclature in standard practice from which protein sequences were obtained e.g. “1”, “Chr1” or “I” etc.
Type:	string
Use:	REQUIRED.
Null:	MUST NOT be “.”.
Example:	**1** 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

6.4.2 chromStart

Description:	The position of the first DNA base that the peptide has been mapped to. For example, if the N-terminal residue of the peptide is Methionine (from an “ATG” codon), the chromStart MUST contain the position of the “A”. Coordinates MUST be 0-based.
Type:	int *
Use:	REQUIRED.
Null:	MUST NOT be “.”.
Example:	1 **1043559** 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2

	604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

* It is considered as an unsigned integer 'unit' for compatibility with version 2.87 of the bigBed conversion tool.

6.4.3 chromEnd

Description:	The position of the last DNA base that the peptide has been mapped to. For example, if the C-terminal amino acid is lysine, and the last codon is AAG, chromEnd MUST contain the position of the "G". Coordinates MUST be 0-based.
Type:	int *
Use:	REQUIRED.
Null:	MUST NOT be ".".
Example:	1 1043559 **1043592** ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

* It is considered as an unsigned integer 'unit' for compatibility with version 2.87 of the bigBed conversion tool.

6.4.4 name

Description:	Unique name for the BED line. For proBed, to ensure that this field is unique throughout the file, it is RECOMMENDED to use the convention: PROTEINACCESSION_DATASETID_UNIQUENUMBER or PROTEINACCESSION _UNIQUENUMBER Another option is to use the peptide sequence (plus modification) as the name. These are only recommendations to facilitate the work of reader software. The only formal requirement is that this field MUST be unique throughout the file.
Type:	string
Use:	REQUIRED.
Null:	MUST NOT be ".".
Example:	1 1043559 1043592 **ENSP00000368678_PXD001524_1462** 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

6.4.5 score

Description:	A score used for shading by visualisation software, MUST be between 0 (transparent) – 1000 (opaque). By default, it is recommended to set this to 1000. In proBed this field MAY be used optionally to represent counts of PSMs for a given "peptide" (representing a group of PSMs), or any quantification value that can be converted on a pseudo-absolute scale for the quantification of the peptide in the given sample. See Section 6.8 for details.
Type:	int *
Use	REQUIRED.
Null:	MUST NOT be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 **1000** + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

* It is considered as an unsigned integer 'unit' for compatibility with version 2.87 of the bigBed conversion tool.

6.4.6 strand

Description:	The strand, MUST either be "+" or "-".
Type:	char[1]
Use:	REQUIRED.
Null:	MUST NOT be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 **+** 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

6.4.7 thickStart

Description:	The thick shading start position of the peptide. This is the position when the coding region starts, which is displayed as a thick shaded area in a genome browser. This MUST be the same number as reported in the chromStart field for a particular data line. Coordinates MUST be 0-based.
Type:	int *
Use:	REQUIRED.

Null:	MUST NOT be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + **1043559** 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

* It is considered as an unsigned integer 'unit' for compatibility with version 2.87 of the bigBed conversion tool.

6.4.8 thickEnd

Description:	The thick shading end position of the peptide. This is the position when the coding region ends, which is displayed as a thick shaded area in a genome browser. This MUST be the same number as reported in the chromEnd field for a particular data line. Coordinates MUST be 0-based.
Type:	int *
Use:	REQUIRED.
Null:	MUST NOT be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 **1043592** 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

* It is considered as an unsigned integer 'unit' for compatibility with version 2.87 of the bigBed conversion tool.

6.4.9 reserved

Description:	This is a reserved field and MUST be always "0" to work with the BED tools. This is required for bigBed conversion tools.
Type:	uint *
Use:	REQUIRED.
Null:	MUST NOT be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 **0** 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

* In the original BED format documentation, this field is called "itemRgb" and it is used to specify the colour scheme of the blocks for visualisation purposes. However, this naming convention is not supported by the BED tools (version 2.87). We are complying here with the conventions supported at present.

6.4.10 blockCount

Description:	The number of blocks (exons) in the BED line i.e. the number of exons to which the peptide has been mapped. If the peptide maps to a single exon, then "1" must be reported. If the peptide crosses a splice junction i.e. mapping to >1 exon, then a value >1 MUST be reported. A value of 0 MUST NOT be reported.
Type:	int
Use:	REQUIRED.
Null:	MUST NOT be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 **1** 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

6.4.11 blockSizes

Description:	A list of the block sizes, and each MUST be a positive integer. The values MUST be separated by commas. The addition of all block sizes MUST be a value divisible by three.
Type:	int[blockCount]
Use:	REQUIRED.
Null:	MUST NOT be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 **33** 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

6.4.12 chromStarts

Description:	A list of block start positions, relative to the chromStart field value. The first value MUST be 0 – indicating the first DNA base to which the peptide has been mapped. For peptides mapped to multiple exons, subsequent values MUST be the start position of the subsequent exons. Values MUST be separated by commas. Coordinates MUST be 0-based. Note: Block end positions do not need to be specified because both the start position and block sizes information is present in the data line.
Type:	int[chromStarts]

Use:	REQUIRED.
Null:	MUST NOT be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 **0** ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

6.5 proBed specific fields

The remainder 13 fields extend the BED format and are specific for the proBed format. These fields are introduced as optional fields in the BED format. In the Example section, an example value for the data type is given in **bold** within the context of a complete and valid line of proBed data.

6.5.1 proteinAccession

Description:	The accession number or unique identifier of the protein in the database searched. This field does not need to be unique. Decoy peptides i.e. peptides that are mapped to decoy database in a target-decoy analysis approach, MUST NOT be included in proBed files. For peptides that can be mapped to both target and decoy peptides, the software MAY include or exclude such peptides based on local preference.
Type:	string
Use:	REQUIRED.
Null:	MUST NOT be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 **ENSP00000368678** FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

6.5.2 peptideSequence

Description:	The raw peptide sequence (without modifications). This field does not need to be unique.
Type:	string
Use:	REQUIRED.
Null:	MUST NOT be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 **FGALCEAETGR** unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2

	604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

6.5.3 uniqueness

Description:	The value provides information about the number of different genetic loci (unique combination of chromosome, chromStart and chromEnd) that the peptide has been mapped. This does not guarantee uniqueness genome-wide but only based on the gene models (genome build) used for the generation of the proBed file, since the same peptide could be classified as unique or not based on the use of different sets of gene models. If the peptide can only be mapped to one locus, it SHOULD have the value of “unique”. If the peptide can be mapped to more than one locus it SHOULD have the value of “not unique”. In the case of “not unique” peptides, more information MUST be provided optionally in brackets: 1. not-unique[super-set] 2. not-unique[same-set] 3. not-unique[subset] 4. not-unique[conflict] 5. not-unique[unknown] The following definitions apply. The application of cases 1-3 requires information from a “protein inference” process, relating peptides to a particular set of loci. One locus in this context is a protein sequence from a gene model at a given chromosomal location. 1. Category 1) SHOULD be applied if the peptide can be mapped to at least one other locus, but where other loci have less evidence than the current reported locus, indicating that this is most likely to be the correct and valid assignment for the peptide. 2. Category 2) SHOULD be applied if the peptide can be mapped to at least one other locus, which has the same (or almost identical – see below) level as evidence as the current locus. This situation usually occurs when the two or more proteins have been identified based on the same-set of peptides or spectra (depending on the protein inference approach taken). Some protein inference software may choose to classify two proteins together as same-set, even if one protein has a single peptide/spectrum with a very weak score, not carried by the other protein(s). In these cases, all peptides would be flagged as not-unique[same-set]. 3. Category 3) SHOULD be applied if the peptide can be mapped to at least one other locus, which has significantly more evidence for identification than the current mapping. This situation typically arises when the current locus (database protein) has a subset of peptides/spectra as one or more

	other locations (including direct sub-sets of one or more proteins, and cases of the protein being multiply subsumed by two or more proteins. 4. Category 4) SHOULD be applied if the peptide can be mapped to at least one other locus, and the assignment of peptides between different loci does not resolve into same-set and subset relationships. In this case, it is assumed there may be independent evidence that protein evidence from more than one locus was observed in the sample, but the assignment of peptides cannot be performed conclusively. 5. Category 5) SHOULD be applied if protein inference/grouping has not been performed, but where the database peptide could have been derived from more than one locus. Note: if the peptide has been mapped to different predicted gene models at the same locus, but not to gene models at different genetic loci, a value of “unique” SHOULD be given. It is expected that the values and qualifiers applied in this data type, may be used in visualisation software e.g. to display different types of peptides in different ways.
Type:	string
Use:	REQUIRED.
Null:	MUST NOT be “.”.
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR **unique** Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

6.5.4 genomeReferenceVersion

Description:	The version of the genome build used as reference. This could represent a given gene set such as a particular GRC version, GENCODE, or a given Ensembl release.
Type:	string
Use:	Required.
Null:	MUST NOT be “.”.
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique **Homo_sapiens.GRCh38.77** 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

6.5.5 psmScore

Description:	One representative PSM score that needs to be used consistently in the file i.e. only one type of score is allowed across one entire file. They SHOULD be reported using PSI-MS CV terms. If a line in proBed represents one "peptide" (as a group of PSMs), the export software SHOULD report either the best score for all the PSMs that are part of that group or a consensus score derived across all PSMs for approaches that perform such analyses.
Type:	double *
Use:	Required.
Null:	MAY be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 **42.87803604921013** 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

* It is considered as a string for compatibility with version 2.87 of the bigBed conversion tool.

6.5.6 fdr

Description:	A cross-platform measure of the likelihood of the identification being incorrect. If a line in proBed represents a single PSM, this value SHOULD normally be the PSM q-value, local FDR, PEP or equivalent value. However, if a single line represents a "peptide" (as a group of PSMs), equivalent values at the peptide level SHOULD be reported instead. The value MUST be given using PSI-MS CV terms, as shown in the example.
Type:	double *
Use:	Required.
Null:	MAY be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 **1.0938989483608807E-5** 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

* It is considered as a string for compatibility with version 2.87 of the bigBed conversion tool.

6.5.7 modifications

Description:	Semicolon-separated list of modifications identified on the peptide, with the following format: **{position}-{modification identifier}** The position gives the peptide position starting from 1. Modifications on the N-terminus of the peptide MUST be reported as 0, and C-terminal modifications MUST be reported as length of the peptide+1. Valid modification identifiers are either PSI-MOD or UNIMOD accession (including the "MOD:" / "UNIMOD:" prefix). Single amino acid polymorphisms (amino acid substitutions) can be reported as modifications. Example: UNIMOD:676 means Trp->Gly substitution (http://www.unimod.org/modifications_view.php?editid1=676).
Type:	string
Use:	REQUIRED.
Null:	MAY be "." (if no modifications are reported).
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 **5-UNIMOD:4** 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524 No modifications in the peptides are reported as ".": 1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 **.** 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

In the case that a modification cannot be reported with an accurate term in UNIMOD then the term MS:1001460 "unknown modification" MUST be used instead, with the delta mass in Daltons as the value.

All (identified) variable modifications as well as fixed modifications MUST be reported for every identification. If no modifications are present, then "." MUST be reported.

The reference system for the location of the protein modification MUST be done at the amino acid level.

6.5.8 charge

Description:	The value of the charge of the peptide identified. This field SHOULD be "." if a line in proBed represents a peptide (as a group of PSMs).
Type:	int *
Use:	REQUIRED.
Null:	MAY be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 **2** 604.77 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

* It is considered as a String for compatibility with version 2.87 of the bigBed conversion tool.

6.5.9 expMassToCharge

Description:	The value of the experimental mass to charge for the precursor ion. This field SHOULD be "." if a line in proBed represents a peptide (as a group of PSMs).
Type:	double *
Use:	REQUIRED.
Null:	MAY be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 **604.77** 604.772 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

* It is considered as a string for compatibility with version 2.87 of the bigBed conversion tool.

6.5.10 calcMassToCharge

Description:	The value of the calculated mass to charge for the precursor ion. This field MAY be "." if a line in proBed represents a peptide both with and without modifications (as a group of PSMs) but in all other cases this value SHOULD be reported.
Type:	double *
Use:	REQUIRED.
Null:	MAY be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 **604.772** 1 PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

* It is considered as a string for compatibility with version 2.87 of the bigBed conversion tool.

6.5.11 psmRank

Description:	The rank of the score of the reported PSM. This field SHOULD be "." if a line in proBed represents a "peptide" (as a group of PSMs).
Type:	int *
Use:	REQUIRED.
Null:	MAY be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 **1** PXD001524_proteoannotator_reprocessed http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

* It is considered as a string for compatibility with version 2.87 of the bigBed conversion tool.

6.5.12 datasetID

Description:	A unique identifier or name for the data set
Type:	string
Use:	REQUIRED.
Null:	MAY be ".".
Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 **PXD001524_proteoannotator_reprocessed** http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524

6.5.13 uri

Description:	A URI (Uniform Resource Identifier) pointing to the file's source data e.g. the website or FTP location for a given dataset (i.e. a dataset in PRIDE Archive http://www.ebi.ac.uk/pride/archive/projects/PXD000764), or the original file that was converted to proBed. Another possibility would be to add the URI of the specific PSM reported in a given proBed line.
Type:	string
Use:	REQUIRED.
Null:	MAY be ".".

Example:	1 1043559 1043592 ENSP00000368678_PXD001524_1462 1000 + 1043559 1043592 0 1 33 0 ENSP00000368678 FGALCEAETGR unique Homo_sapiens.GRCh38.77 42.87803604921013 1.0938989483608807E-5 5-UNIMOD:4 2 604.77 604.772 1 PXD001524_proteoannotator_reprocessed **http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD001524**

6.6 Additional optional fields in proBed

Additional fields MAY be added to the end of rows. The optional fields SHOULD be consistent across all data lines. The information stored within an optional field is completely up to the resource that generates the file. The values SHOULD be of a consistent data type throughout the column, preferably using the standard cvParam encoding:

[PSI-MS, MS:1002356, PSM-level combined FDRScore, 5.0130232398289357E-5]

It is important to highlight that, if optional fields are present, the related supporting .as file would need to be up-to-date when covering proBed to the bigBed format (see details in Section 11.2, Supporting autoSQL file, and see Appendix I). This SHOULD include any new field definition name in CamelCase, a short description through a comment, and the data type. Names MUST be unique.

6.7 How to represent intron and exon regions

In all typical cases, a PSM or peptide is matched to one or more exons at a given locus on the genome. This may result in either one continuous uninterrupted block of coordinates (mapping of a peptide to one exon), or many blocks (for peptides that span across intron junctions.

To describe one continuous region on a proBed data line then the “blockCount” field would be “1”, the “blockSizes” field would be the difference between the “chromStart” and “chromEnd” fields, and the “chromStarts” field must be “0”.

Example:	A single block, of size 24 would have the values: “1” for “blockCount”, “24” for “blockSize”, and “0” for “chromStarts”, would be defined in a data line as: *1 8861404 8861428 ENSP00000234590_PXD001524_1255 1000 - 8861404 8861428 0* ***1*** ***24*** ***0*** *…*

If multiple blocks are to be described instead for a data line, i.e. the peptide has been mapped to more than one exon region (across one or more introns), then the total number of blocks MUST be listed for the “blockCount” field, the sizes of each block

listed for the “blockSizes” field as a comma separated list, and the start positions listed in the “chromStarts” field as a comma separated list with the first value must be “0”.

Example:	Two blocks: the first has a length of 8, the second has a length of 58.The starting relative positions are 0 and 295. This would mean the fields are as follows: “2” for the “blockCount”, “8,58” for “blockSize”, and “0,295” for “chromStarts”. The data line would be defined as: *1 8862939 8863292 ENSP00000234590_PXD001524_1266 1000 - 8862939 8863292 0* ***2*** ***8,58 0,295****…*

6.8 Representation of peptides (as groups of PSMs) in proBed

The proBed format is primarily designed to represent PSMs. However, it can also accommodate the reporting of “peptides” (representing groups of PSMs). Grouping of PSMs in peptides is flexible, depending on the needs of the file producer. Two main situations are envisioned:

- If the producer is only interested in reporting peptide sequences without considering modifications, then PSMs should be grouped by their raw peptide sequence. In this case, “modifications” SHOULD be stated as null.

- If the producer is interested in reporting modifications as well, then PSMs should be grouped by peptide sequence plus modifications i.e. multiple rows per modification status of a given peptide.

PSMs representing the same peptide sequence (with or without modifications being considered) but having different charge values should be always grouped together, for simplification purposes.

In any case, some conventions need to be followed as well for the fields psmScore (Section 6.5.4), fdr (Section 6.5.5), charge (Section 6.5.7), expMassToCharge (Section 6.5.8) and psmRank (Section 6.5.9).

If reporting peptide-level data, the field “score” (Section 6.4.5) SHOULD be used to represent PSM counts. It is a score used for shading by visualisation software, and it MUST be between 0 (transparent) – 1000 (opaque). It is RECOMMENDED to set this to 1000 when reporting PSM-level data. See https://genome.ucsc.edu/FAQ/FAQformat.html#format1 for documentation about the shading in the original BED format. For cases where peptide-level data is being reported, the score attribute should map the PSM counts (spectral counting) or any other quantification value that can be used to give an estimate for absolute quantitation of the protein onto a consistent scale, depending on the size of the input data set. As a general guidance, the following default mapping (**Table 1**) MAY be applied in case of spectral counting data.

Score	Number of PSMs grouped in a peptide sequence
<167	1
167-277	2-4
278-388	5-7
389-499	8-10
500-610	11-13
611-722	14-16
723-833	17-19
834-944	20-22
>944	>22

Table 1. Recommended values for the “score” field in case number of PSMs are represented using this field (for cases where peptide-level results are reported).

6.9 Merging of proBed files

proBed files can be merged, for example for meta-analyses or collating data from different samples. The field “name” (Section 6.4.4) must be unique throughout the file. Files SHOULD only be merged if they have been generated by the same software, following the same conventions for the naming and content of the different fields, and ideally the same reference system used for the genome coordinatesx although this is not mandatory (see Section 6.5.4). In addition, it is RECOMMENDED not to use the “score” field to represent PSM counts in merged BED files, unless PSM counts are updated during the merge process (see Section 6.8).

6.10 Other supporting materials

The following example instance documents are available and between them cover all the use cases supported. All example files and the originating files that were converted to proBed can be downloaded from https://goo.gl/wojrR4.

Three sets of proBed example files have been generated (at https://goo.gl/JvUvbU).

a) PXD001524_reprocessed.pro.bed https://goo.gl/CtMjgQ – example proBed file converted from the mzTab example file indicated below. This file has been further processed to only report upon chromosomes that are mentioned in the chrom_sizes.txt file.

Additional related files to example one (available at https://goo.gl/XtcbeX):
b) PXD001524_reprocessed.mzid https://goo.gl/s6hi1z – example mzIdentML file with genome annotation.
c) PXD001524_reprocessed.mztab https://goo.gl/HdcaQv – example mztab file converted from the mzidentML example file.
d) UCSC bed to big bed converter tool – v2.87 http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64.v287/bedToBigBed.
e) chrom_sizes.txt https://goo.gl/1cZDMb – example chromosome sizes file.
f) proBed-1.0.0.as https://goo.gl/wvUP2I – proBed autoSQL file, supporting the file conversion to bigBed.
g) PXD001524_reprocessed.bb https://goo.gl/bHfGB7 – example bigBed file converted using the example proBed file, aSQL file, chromosome sizes file, and converter tool.

Second set of proBed example and related files (available at https://goo.gl/MpBqnt):
h) PXD000656_reprocessed.pro.bed https://goo.gl/EvgIfk – 2nd example proBed file converted from the 2nd mzTab example file.
i) PXD000656_reprocessed.mzid https://goo.gl/rMS9Mt – 2nd example mzIdentML file with genome annotation.
j) PXD000656_reprocessed.mztab https://goo.gl/zwJLXH – 2nd example mztab file converted from the 2nd mzidentML example file.
k) PXD000656_reprocessed.bb https://goo.gl/SUjA9O – 2nd example bigBed file converted using the 2nd example proBed file.

Third set of proBed example and related files (also available at https://goo.gl/MpBqnt):
l) PXD000764_reprocessed.pro.bed https://goo.gl/8Cw6iG – 3rd example proBed file converted from the 3rd mzTab example file.
m) PXD000764_reprocessed.mzid https://goo.gl/UOT7k5 – 3rd example mzIdentML file with genome annotation.
n) PXD000764_reprocessed.mztab https://goo.gl/6rm6fX – 3rd example mztab file converted from the 3rd mzidentML example file.
o) PXD000764_reprocessed.bb https://goo.gl/X4QLlp – 3rd example bigBed file converted using the 3rd example proBed file.

The ProteoAnnotator (7) software pipeline produces annotated mzIdentML data which with genome coordinates for identified peptides.

7. Conclusions

This document contains the specifications for using the proBed format to represent results from peptide and protein identification pipelines, in the context of a proteogenomics investigation. This specification constitutes a proposal for a standard from the Proteomics Standards Initiative. These artefacts are currently undergoing the PSI document process, which will result in a standard officially sanctioned by PSI.

8. Authors

Tobias Ternent
European Bioinformatics Institute (EMBL-EBI)
Hinxton, United Kingdom
tobias@ebi.ac.uk

Fawaz Ghali
Institute of Integrative Biology, University of Liverpool
Liverpool, United Kingdom
F.Ghali@liverpool.ac.uk

David Fenyo,
Center for Health Informatics and Bioinformatics, New York University Medical School
New York, USA
david.fenyo@gmail.com

Andrew R. Jones,
Institute of Integrative Biology, University of Liverpool
Liverpool, United Kingdom
Andrew.Jones@liverpool.ac.uk

Juan Antonio Vizcaíno
European Bioinformatics Institute (EMBL-EBI)
Hinxton, United Kingdom
juan@ebi.ac.uk

Correspondence – Tobias Ternent (tobias@ebi.ac.uk) and Juan Antonio Vizcaíno (juan@ebi.ac.uk).

9. Contributors

In addition to the authors, the following people contributed to the model development, gave feedback or tested proBed:

Andy Yates, European Bioinformatics Institute (EMBL-EBI), United Kingdom.
Gerben Menschaert, Ghent University, Belgium.

10. References

1. Tyner C.; Barber, G.P.; Casper J.; Clawson, H.; Diekhans, M.; Eisenhart, C.; Fischer, C.M.; Gibson, D.; Gonzalez, J.N.; Guruvadoo, L.; Haeussler, M.; Heitner, S.; Hinrichs, A.S.; Karolchik, D.; Lee, B.T.; Lee, C.M.; Nejad, P.; Raney, B.J.; Rosenbloom, K.R.; Speir, M.L.; Villarreal, C.; Vivian, J.; Zweig, A.S.; Haussler, D.; Kuhn, R.M.; Kent, W.J. The UCSC GenomeBrowser database: 2017 update. *Nucleic Acids Res.* **2017**, 45(D1), D626-34.
2. Kent, W. J.; Zweig, A. S.; Barber, G.; Hinrichs, A. S.; Karolchik, D., BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **2010,** 26, (17), 2204-7.
3. Bradner, S; Key words for use in RFCs to Indicate Requirement Levels. In Internet Engineering Task Force. RFC 2119: **1997**.
4. 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.; Julian, R.; Binz, P. A.; Deutsch, E. W.; Hermjakob, H.; Reisinger, F.; Griss, J.; Vizcaino, J. A.; Chambers, M.; Pizarro, A.; Creasy, D., The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics* **2012,** 11, (7), M111 014381.
5. Griss, J.; Jones, A. R.; Sachsenberg, T.; Walzer, M.; Gatto, L.; Hartler, J.; Thallinger, G. G.; Salek, R. M.; Steinbeck, C.; Neuhauser, N.; Cox, J.; Neumann, S.; Fan, J.; Reisinger, F.; Xu, Q. W.; Del Toro, N.; Perez-Riverol, Y.; Ghali, F.; Bandeira, N.; Xenarios, I.; Kohlbacher, O.; Vizcaino, J. A.; Hermjakob, H., The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol Cell Proteomics* **2014,** 13, (10), 2765-75.
6. Mayer, G.; Montecchi-Palazzi, L.; Ovelleiro, D.; Jones, A. R.; Binz, P.-A.; Deutsch, E. W.; Chambers, M.; Kallhardt, M.; Levander, F.; Shofstahl, J.; Orchard, S.; Antonio Vizcaíno, J.; Hermjakob, H.; Stephan, C.; Meyer, H. E.; Eisenacher, M., The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. *Database* **2013,** 2013.
7. Ghali, F.; Krishna, R.; Perkins, S.; Collins, A.; Xia, D.; Wastling, J.; Jones, A. R., ProteoAnnotator--open source proteogenomics annotation software supporting PSI standards. *Proteomics* **2014,** 14, (23-24), 2731-41.

11. Appendix I: proBed to bigBed conversion

Conversion from a proBed file into the bigBed format (binary version, usually used in annotation tracks) (2) can be performed with the UCSC bedToBigBed converter tool v2.87). The converter tool requires the following as input:

- An input proBed file. The file extension of the file needs to be .bed. The file MUST be pre-sorted in ascending numerical order of chromosome number and start position, respectively.
- A supporting autoSQL file. This acts as a definition file, listing and describing the fields that are present (including the ones incorporated into proBed).
- A supporting text file listing the chromosome sizes.

For examples of these files, see Section 6.10, Other supporting materials.

11.1 Sorted proBed file

Sorting proBed lines in ascending order can be achieved using a UNIX sort command. The sorted file MUST be plain-text and have a .bed file extension.

Example:	*PXD001524_reprocessed.pro.bed https://goo.gl/CtMjgQ* *$ sort -k1,1 -k2,2n PXD001524_reprocessed.pro.bed > PXD001524_reprocessed.sorted.pro.bed*

11.2 Supporting autoSQL file

The conversion tool from proBed into bigBed requires, amongst other parameters, a supporting autoSQL file. This file MUST be plain text and have .as as the file extension. It describes the fields present in the proBed file. By default, the format MUST be as follows (see more details at http://www.linuxjournal.com/article/5949).

An autoSQL

Example:	proBed-1.0.0.as https://goo.gl/wvUP2I The proBed aSQL schema file has the contents:

	table proBed *"BED12+13 PSI proBed 1.0.0"* *(* *string chrom; "Reference sequence chromosome"* *uint chromStart; "Start position of the first DNA base"* *uint chromEnd; "End position of the last DNA base"* *string name; "Unique name"* *uint score; "Score"* *char[1] strand; "+ or - for strand"* *uint thickStart; "Coding region start"* *uint thickEnd; "Coding region end"* *uint reserved; "Always 0"* *int blockCount; "Number of blocks"* *int[blockCount] blockSizes; "Block sizes"* *int[blockCount] chromStarts; "Block starts"* *string proteinAccession; "Protein accession number"* *string peptideSequence; "Peptide sequence"* *string uniqueness; "Peptide uniqueness"* *string genomeReferenceVersion; "Genome reference version number"* *double psmScore; "PSM score"* *double fdr; "False-discovery rate"* *string modifications; "Post-translational modifications"* *int charge; "Charge value"* *double expMassToCharge; "Experimental mass to charge value"* *double calcMassToCharge; "Calculated mass to charge value"* *int psmRank; "Peptide-Spectrum Match rank."* *string datasetID; "Dataset Identifier"* *string uri; "Uniform Resource Identifier"* *)*

11.3 Chromosome names file

The bedToBigBed converter tool also requires a supporting file that reports the chromosome names and their maximum sizes. Such a file can be generated using the Ensembl REST Servers. This file must be tab-separated plain-text with information described in two fields (the chromosome name, and its maximum size), with an optional file extension.

Example:	The Ensembl Python script chromosome sizes generator is available at: https://gist.github.com/andrewyatz/a3687b573364f65904e2 chrom_sizes.txt https://goo.gl/E3FKW3 File contents: *1 248956422* *10 133797422* *11 135086622* *…*

11.4 Running the bigBed conversion tool

The bedToBigBed conversion tool is run using the following command structure:

$ ./bedToBigBed. -as=<ASQLFILE> -type=<TYPE> -tab <SORTEDBEDFILE> <CHROMSIZESFILE> <BIGBEDFILE>

- <ASQLFILE> relates to the supporting aSQL filename.
- <TYPE> relates to the number of BED standard and additional fields used in the format “bed<BED FIELDS NUMBER>+<OTHER FIELDS NUMBER>”
- <CHROMSIZESFILE> relates to the chromosomes sizes filename.
- <BIGBEDFILE> relates to the output bigBed file name. The output bigBed file MUST have a .bb file extension.

Example:	*$ ./bedToBigBed -as=proBed-1.0.0.as -type=bed12+13 -tab PXD001524_reprocessed.pro.bed chrom_sizes.txt PXD001524_reprocessed.bb*

12. Intellectual Property Statement

The PSI takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the PSI Chair.

The PSI invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to practice this recommendation. Please address the information to the PSI Chair (see contacts information at PSI website).

TradeMark Section

Microsoft Excel®

Copyright Notice

Copyright (C) Proteomics Standards Initiative (2017). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the PSI or other organizations, except as needed for the purpose of developing Proteomics Recommendations in which case the procedures for copyrights defined in the PSI Document process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the PSI or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE PROTEOMICS STANDARDS INITIATIVE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY

THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE."