**­­**

**qcML (PSI)**

Status of This Document

This document presents a draft specification of the qcML data format developed by members of the Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI) Quality Control (QC) Working Group. Distribution is unlimited.

Version of This Document

The current version of this document is 0.0.9, April …, 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

# 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.

# Introduction

## 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](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.

## 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.

# Use Cases for qcML

The following cases of usage have driven the development of the qcML data model, and are used to define the scope of the format in version 0.0.9

1. qcML is for reporting quality metrics calculated by QC tools (such as Quameter, rawMeat, …).
2. qcML is for handover of quality reports of measured MS runs to researchers.
3. qcML is for storing / archiving QC metrics next to MS files / results in databases / LIMS.

In later versions specifying thresholds for QC metrics or flagging metrics / runs / sets of runs as “quality unfulfilled” is not modelled in the current version. Once that is modelled, a qcML file may be enriched in a second-pass with thresholds and flags together with some rationale, who or how that “quality unfulfilled” flag was set.

# 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).

# Relationship to Other Specifications

*bigBed (*[*https://genome.ucsc.edu/FAQ/FAQformat.html#format1.5*](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).

## 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](mailto: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.

# Format specification

…

<runquality> section: (1-to-many) quality metrics for one (raw) file

<setquality> section: (0-to-many) if it is present, all the quality metrics for one (raw) file should be in the same qcML file

# 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.

# Authors

Tobias Ternent

European Bioinformatics Institute (EMBL-EBI)

Hinxton, United Kingdom

[tobias@ebi.ac.uk](mailto:tobias@ebi.ac.uk)

Fawaz Ghali

Institute of Integrative Biology, University of Liverpool

Liverpool, United Kingdom

[F.Ghali@liverpool.ac.uk](mailto: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](mailto:david.fenyo@gmail.com)

Andrew R. Jones,

Institute of Integrative Biology, University of Liverpool

Liverpool, United Kingdom

[Andrew.Jones@liverpool.ac.uk](mailto:Andrew.Jones@liverpool.ac.uk)

Juan Antonio Vizcaíno

European Bioinformatics Institute (EMBL-EBI)

Hinxton, United Kingdom

[juan@ebi.ac.uk](mailto:juan@ebi.ac.uk)

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

# 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.

# 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.

# 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.

## 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 |

## 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"*  *)* |

## 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*  *…* |

## 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 |

# 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."