bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The proBAM and proBed standard formats: enabling a seamless integration of genomics and proteomics data

Gerben Menschaert[1,‡,§], Xiaojing Wang[2,3,‡,§], Andrew R. Jones[4], Fawaz Ghali[4,5], David Fenyö[6,7], Volodimir Olexiouk[1], Bing Zhang[2,3], Eric W. Deutsch[8], Tobias Ternent[9] and Juan Antonio Vizcaíno[9,§]

[1] Department of Mathematical Modeling, Statistics and Bioinformatics, Ghent University, Ghent, Belgium.

[2] Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, Texas, USA.

[3] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA.

[4] Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom.

[5] School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester, M1 5GD, United Kingdom.

[6] Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, New York, USA.

[7] Institute for Systems Genetics, New York University School of Medicine, New York, New York, USA.

[8] Institute for Systems Biology, Seattle, WA, USA.

[9] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom.

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

‡ equally contributing authors

§ corresponding authors

To whom correspondence may be addressed: Dr. Gerben Menschaert, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure links 653, 9000 Gent, Belgium. Tel +3292649922; E-mail: Gerben.Menschaert@ugent.be. Dr. Xiaojing Wang, Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, Texas, USA; and Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA. E-mail: Xiaojing.Wang@bcm.edu. Dr. Juan Antonio Vizcaíno, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom. E-mail: juan@ebi.ac.uk.

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Summary

On behalf of The Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI), we are here introducing two novel standard data formats, proBAM and proBed, that have been developed to address the current challenges of integrating mass spectrometry based proteomics data with genomics and transcriptomics information in proteogenomics studies. proBAM and proBed are adaptations from the well-defined, widely used file formats SAM/BAM and BED respectively, and both have been extended to meet specific requirements entailed by proteomics data. Therefore, existing popular genomics tools such as SAMtools and Bedtools, and several very popular genome browsers, can be used to manipulate and visualize these formats already out-of-the-box. We also highlight that a number of specific additional software tools, properly supporting the proteomics information available in these formats, are now available providing functionalities such as file generation, file conversion, and data analysis. All the related documentation to the formats, including the detailed file format specifications, and example files are accessible at http://www.psidev.info/probam and http://www.psidev.info/probed.

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Mass spectrometry (MS) based proteomics approaches have advanced enormously over the last decade, and are becoming increasingly prominent as an essential tool for post-genomic research. Proteomics approaches enable the identification, quantification and characterization of proteins, peptides, and post-translational protein modifications (PTMs) such as phosphorylation, providing information about protein expression and functional states [1]. Despite the instrumental role of the underlying genome in proteomics data analysis, it is only relatively recently when the field of proteogenomics started to gain prominence [2-4].

In proteogenomics, proteomics data is combined with genomics and/or transcriptomics information, typically by using sequence databases generated from DNA sequencing efforts, RNA-Seq experiments [5], Ribo-Seq approaches [6, 7], and long-non-coding RNAs [8], among others, in the MS-based identification process. Peptide sequences are mapped back to gene models *via* their genomic coordinates, demonstrating evidence of new translational events (e.g. novel splice junctions). Proteogenomics studies can be used to improve genome annotation and are increasingly utilized to understand the information flow from genotype to phenotype in complex diseases such as cancer [9-11] and to support personalized medicine studies [12].

Since 2002, the Proteomics Standards Initiative (PSI, http://www.psidev.info) of the Human Proteome Organization (HUPO) [13] has taken the role of developing open

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

80 community standard file formats for different aspects of MS based proteomics analysis
81 and data types. At present, well-established data standards are available for instance, for
82 representing raw MS data (the mzML data format [14]), peptide and protein
83 identifications (mzIdentML [15] and mzTab [16]) and quantitative information
84 (mzQuantML [17] and mzTab).
85
86 The existence of compatible and interoperable data formats is a way to facilitate and
87 advance "multi-omics" studies, and a clear need in proteogenomics, due to the growing
88 importance of the field [9, 10, 18, 19]. However, no standard file format had been
89 established so far for proteogenomics data exchange. To address this problem, we here
90 present two novel standard data formats called proBAM and proBed. As suggested by
91 their names, these two formats are adapted from their genomics counterparts BAM/SAM
92 [20, 21] and BED (Browser Extensible Data) [22], where proBAM stands for proteomics
93 BAM file (compressed binary version of the Sequence Alignment/Map (SAM) format)
94 and proBed stands for proteomics BED file. A key feature of these formats is that they
95 can seamlessly accommodate both regular genomic mapping information and specifics
96 related to proteomics data, i.e. peptide-to-spectrum matches (PSM) or peptide sequence
97 information. Existing popular genomics tools as SAMtools [20, 21] and Bedtools [23,
98 24], or the most widely used genome browsers such as Ensembl [25], the University of
99 California Santa Cruz (UCSC) Genome Browser [26], JBrowse [27] and the Integrative
100 Genomics Viewer (IGV) [28], can be used to manipulate and visualize proteomics data in
101 these formats already. We believe that both proBAM and proBed are essential to merge

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the growing amount of proteomics information with the available genomics/transcriptomics data.

Experimental Procedures

The development of these data formats has taken place since 2014 and it has been an open process *via* conference calls and discussions at the PSI annual meetings. Both format specifications have been submitted to the PSI document process [29] for review. The overall goal of this process, analogous to an iterative scientific manuscript review, is that all formalized standards are thoroughly assessed. This process is handled by the PSI Editor and external reviewers who can provide feedback on the format specifications. Additionally, there is a phase for public comments, ensuring the involvement of heterogeneous points of view from the community.

Both formats use Controlled Vocabulary (CV) terms and definitions as part of the PSI-MS CV [30], also used in other PSI data formats. All the related documentation, including the detailed file format specifications and example files, are available at http://www.psidev.info/probam and http://www.psidev.info/probed.

Overview of the proBAM and proBed formats

The proteogenomics formats proBAM and proBed are designed to store a genome-centric representation of proteomics data (Figure 1). As mentioned above, both formats are highly compatible with their originating genomics counterparts, thus benefiting already from a plethora of existing tools developed by the genomics community.

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

proBAM overview

The BAM format was originally designed to hold alignments of short DNA or RNA reads to a reference genome [20, 21]. A BAM file typically consists of a header section storing metadata and an alignment section storing mapping data (Figure 1, Figure 2 and Supplemental Table 1). The metadata can include information about the sample identity, technical parameters in data generation (such as library, platform, etc) and data processing (such as mapping tool used, duplicate marking, etc). Essential information includes where reads are aligned, how good the alignment is and the quality of the reads. Specific fields or tags are designed to represent or encode such information. The proBAM format inherits all these features. In this case, sequencing reads are replaced by PSMs (see proBAM specification document for full details, https://goo.gl/EW1cqB).

It should be noted that, since the tags used in BAM usually have recognized meanings, we did not attempt to repurpose any of them but rather created new ones to accommodate specific proteomics data types such as PSM scores, charge states, and protein PTMs (Figure 2 and proBAM specification document section 4.4.1 for full description on PSM specific tags). We also envisioned that additional fields and tags may be necessary to hold additional aspects of proteomics data. We thus designed a "Z?" tag as an extension anchor. Analogously to proBed, the format can also accommodate peptides (as groups of PSMs with the same peptide sequence). At the moment of writing, the proBAM format is under review as part of the PSI document process (http://www.psidev.info/probam). *As a note to editors and reviewers, we are aiming to conclude the PSI review at the same time*

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

147 *as the manuscript describing proBAM/proBed is deemed suitable for publication, at*
148 *which point we can announce a finalized standard.*
149
150 *proBed overview*
151 The original BED format (https://genome.ucsc.edu/FAQ/FAQformat.html#format1),
152 developed by the UCSC, provides a flexible way to define data lines that can be
153 displayed as annotation tracks. proBed is an extension to the original BED file format
154 [26]. In BED, data lines are formatted in plain text with white-space separated fields.
155 Each data line represents one item mapped to the genome. The first three fields
156 (corresponding to genomic coordinates) are mandatory, and an additional 9 fields are
157 standardized and commonly interpreted by genome browsers and other tools, totalling 12
158 BED fields, re-used here. The proBed format includes a further 13 fields to describe
159 information primarily on peptide-spectrum matches (PSMs) (Figure 1, Figure 2 and
160 Supplemental Table 1). The format can also accommodate peptides (as groups of PSMs
161 with the same peptide sequence), but in that case, some assumptions need to be taken in
162 some of the fields (see proBed specification document Section 6.8 for details,
163 https://goo.gl/FM2w66). At the moment of writing, the proBed format has completed the
164 PSI internal review process, so the first version of the standard has been formalized
165 (version 1.0, http://www.psidev.info/probed).
166
167 *Distinct features of proBAM and proBed and their use cases*
168 The proBAM and proBed formats differ in similar ways as their genomic counterparts do,
169 although representing analogous information. In fact, proBAM and proBed are

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

complementary and have different use cases. Figure 3 shows two examples of proBAM and proBed visualization tracks of the same datasets. An IGV and Ensembl visualization are presented including multiple splice-junction peptides (Figure 3.A) and a novel translation initiation event in the HDGF gene locus (Figure 3.B), respectively.

Similar to the designed purposes of SAM/BAM, the basic concepts behind the proBAM format are: i) to provide genome coordinates as well as detailed mapping information, including CIGAR, flag, nucleotide sequences, etc; ii) to hold richer proteomics related information; and iii) to serve as a well-defined interface between PSM identification and downstream analyses. Therefore, the proBAM format contains much more information about the peptide-gene mapping statuses as well as PSM related information, when compared to proBed. Peptide and nucleotide sequences are inherently embedded in proBAM, which can be useful for achieving improved visualization by tools such as IGV. This feature enables intuitive display of the coverage of a region of interest, peptides at splice junctions, single nucleotide/amino acid variation, and alternative spliced isoforms (Figure 3), among others. Therefore, proBAM can hold the full MS proteomics result set, whereupon further downstream analysis can be performed: gene-level inference [31], basic spectral count based quantitative analysis, reanalysis based on different scoring systems and/or FDR (False Discovery Rate) thresholds.

The proBed format, on the other hand, is more tailored for storing only the final results of a given proteogenomics analysis, without providing the full details. The BED format is commonly used to represent genomic features. Thus, proBed stores browser track information at the PSM and/or peptide level mainly for visualisation purposes. As a key point, proBed files can be converted to BigBed [32], a binary format based on BED,

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

193 which represents a feasible way to store the same information present in BED as
194 compressed binary files, and is the final routinely used format as annotation tracks. It
195 should be noted that a proBAM to proBed conversion should be possible, and vice versa.
196 However, "null" values for some of the Tags would be logically expected for the
197 mapping from proBed to proBAM.
198
199 *Software implementations*
200 Both proBAM and proBed are fully compatible out-of-the-box with existing tools
201 designed for the original SAM/BAM and BED files. Therefore, existing popular tools in
202 the genomics community can readily be applied to read, merge and visualize these
203 formats (Table 1). As mentioned already, several stand-alone and web genome browsers
204 are available to visualize these formats e.g. UCSC browser, Ensembl, Integrative
205 Genomics Viewer, and JBrowse.
206
207 Routinely used command line tools as SAMtools allow to manipulate (index, merge, sort)
208 alignments in proBAM. Bedtools, seen as the "Swiss-army knife" tools for a wide-range
209 of genomic analysis tasks, allows similar actions to both formats, including among others,
210 intersection, merging, count, shuffling and conversion functionality. With the UCSC
211 'bedToBigBed' converter tool (http://hgdownload.soe.ucsc.edu/admin/exe/), one can also
212 convert the proBed to bigBed. In this context, it is important to note that bedToBigBed
213 version 2.87 is highlighted in the proBed format specification as the reliable version that
214 can be used to create bigBed files coming from proBed (version 1.0) files.
215

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table1. Existing software implementations of the proBAM and proBed formats (by June 2017).

Name	Description	URL	purpose
ms-data-core-api *	Open-source Java library to handle different proteomics data standard formats	https://github.com/PRIDE-Utilities/ms-data-core-api	**write/ convert**
PGConverter *	Command-line tool to convert between the following formats: mzIdentML -> mzTab -> proBed -> bigBed	https://github.com/PRIDE-Toolsuite/PGConverter	
proBAMr *	Bioconductor package to convert MS-shotgun identification results into proBAM	http://bioconductor.org/packages/release/bioc/html/proBAMr.html	
proBAMconvert *	Command-line and GUI tool to create proBAM or proBed from mzIdentML, mzTab or pepXML.	http://probam.biobix.be/	
UCSC Genome Browser	Web-based genome browser	https://genome.ucsc.edu/	**visualize**
Ensembl	Web-based genome browser	http://www.ensembl.org/	
Integrative Genomics Viewer (IGV)	Stand-alone, high-performance visualization tool for interactive exploration of large, integrated genomic datasets	http://software.broadinstitute.org/software/igv/	
JBrowse	Embeddable genome browser built completely with JavaScript and	http://jbrowse.org/	

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

	HTML5		
SAMtools	Tool package that provides various utilities for manipulating alignments in the SAM format (including sorting, merging and indexing)	http://samtools.sourceforge.net/	**manipulate**
Bedtools	A "Swiss-army knife" of tools for a wide-range of genomics analysis tasks	http://bedtools.readthedocs.io/en/latest/	
proBAMtools *	R package to perform downstream analysis of proBAM files	http://proteogenomics.zhang-lab.org/	**analyse**
PGConverter *	It contains a proBed validation module	https://github.com/PRIDE-Toolsuite/PGConverter	**validate**
BamUtil	An original SAM/BAM format validation package	https://github.com/statgen/bamUtil	

* Software supports full features of the format (including proteomics information).

There is also software specifically written for proBAM and proBed, supporting all the proteomics related features. In fact, proteogenomics data encoded in the PSI standard formats mzIdentML and mzTab can be converted into proBAM and proBed, although it should be noted that the representation for proteogenomics data in mzIdentML has only been formalized recently [33]. In this context, first of all, the open-source Java library ms-data-core-api, created to handle different proteomics file formats using the same interface, can be used to write proBed [34]. A Java command line tool, PGConverter (https://github.com/PRIDE-Toolsuite/PGConverter), is also able to convert from

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mzIdentML and mzTab to proBed and bigBed. Analogously, several tools are available to write proBAM files, such as the Bioconductor proBAMr package. An additional R package, called proBAMtools, is also available to analyze fully exported MS-based proteomics results in proBAM [31]. proBAMtools was specifically designed to perform various analyses using proBAM files, including functions for genome-based proteomics data interpretation, protein and gene inference, count-based quantification, and data integration. It also provides a function to generate a peptide-based proBAM file coming from a PSM-based one.

ProBAMconvert is another intuitive tool that enables the conversion from mzIdentML, mzTab and pepXML (another popular proteomics open format) [35] to both peptide- or PSM-based proBAM and proBed (http://probam.biobix.be) [36]. It is available as a command line interface (CLI) and a graphical user interface (GUI for Mac OS X, Windows and Linux). As CLI it is also wrapped in a Bioconda package (https://bioconda.github.io/recipes/probamconvert/README.html) and in a Galaxy tool, available from the public test toolshed (https://testtoolshed.g2.bx.psu.edu/view/galaxyp/probamconvert). The PGConverter tool also allows the validation of proBed files. For proBAM files, a validator is available that checks the validity of the original SAM/BAM format (https://github.com/statgen/bamUtil), although additional proteogenomics data verification still needs to be implemented.

Discussion

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We strongly believe that having available these two novel data formats (proBAM and proBed) constitutes an essential milestone for the continuous development of the field of proteogenomics. Successful promotion of proBAM and proBed requires support from software vendors, individual investigators, publishers, and data repositories. We will promote them following the typical channels used by the PSI. Therefore, further efforts will be focused on implementing these formats, not only using newly generated proteomics data but also on datasets already available in the public domain. In this context it is important to highlight that MS-based proteomics datasets are now routinely deposited in public repositories such as PRIDE [37], PeptideAtlas [38], MassIVE (https://massive.ucsd.edu) and jPOST [39] gathered in the ProteomeXchange Consortium (http://www.proteomechange.org [40]). In fact, an enormous amount of MS data is available in the public domain that can be used for proteogenomics studies, something that it is increasingly happening [41, 42]. The PRIDE database, located in the European Bioinformatics Institute (EMBL-EBI), plans to fully implement proBed in the coming months, facilitating the integration and visualisation of public proteomics data in Ensembl. In this context, it is also important to note that proBAM files generated from several large proteomics datasets have been already preloaded in a JBrowse-based genome browser (http://proteogenomics.zhang-lab.org/), facilitating the access to this data to a broader audience, both within and outside the proteomics community.

Additionally, we have already been actively pushing the use of these formats in big Consortia such as Clinical Proteomic Tumor Analysis Consortium (CPTAC). We hope the data released by such projects will inspire new tools that support these two formats.

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We expect that their existence will facilitate integration, visualization and exchange throughout both the proteomics and genomics communities, and will help multiple proteogenomics endeavours in trying to interpret proteomics results and/or refine gene model annotation by means of protein level validation.

The formats will be fully maintained by the PSI group using the strategy applied for all existing standard formats. If changes in the formats were needed that would not make them compatible with existing software, the formats would change their version number, and they would re-enter a new round of review in the PSI document process. Some future possible expansions for both formats could consider extended mechanisms to encode quantitative proteomics data. There is a mechanism to report PSM counts in proBed, but it is limited at present. Additionally, PSM counts can be calculated, at both gene and protein levels, from proBAM files. In the future, quantification support could be extended to additional workflows (e.g. intensity-based approaches).

We also highly encourage proteogenomics data providers to report PSMs to these two formats as part of their data exports, so it can be visualized by genome browsers directly and it is possible to re-analyse it within a genome context. We expect that the release and usage of proBed and proBAM will increase data sharing and integration between both the genomics and proteomics communities. The PSI remains a free and open consortium of interested parties, and we encourage critical feedback, suggestions and contributions *via* attendance at a PSI annual meeting, conference calls or our mailing lists (see http://www.psidev.info/).

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

296

297 **Acknowledgements**

298

299 J.A.V., T.T., A.R.J. and F.G. want to acknowledge funding by the BBSRC grants
300 "ProteoGenomics" [grant number BB/L024225/1] and "PROCESS" [grant number
301 BB/K01997X/1], and A.R.J. wants to acknowledge BBSRC grant BB/L005239/1. G.M.
302 is a Fellow of the Research Foundation – Flanders (FWO-Vlaanderen)
303 [G.M.,12A7813N]. X.W. and B.Z. are supported by National Cancer Institute award
304 U24CA159988 and U24CA210954. E.W.D. acknowledges funding from NIGMS grant
305 number R01GM087221 and NIBIB grant number U54EB020406. D.F. is supported by
306 National Cancer Institute award U24CA210972 and by contract 13XS068 from Leidos
307 Biomedical Research, Inc. Finally, the colleagues in the Proteomics Standards Initiative,
308 including the reviewers of the proBAM and proBed format specifications in the PSI
309 document process, are acknowledged for helpful discussions and feedback. We also want
310 to thank Andy Yates (Ensembl team), for his useful comments.

311

312 The author(s) declare(s) that they have no competing interests.

313

314 **References**

315

316 1. Aebersold R, Mann M: **Mass-spectrometric exploration of proteome structure and**
317 **function.** *Nature* 2016, **537**:347-355.
318 2. Menschaert G, Fenyo D: **Proteogenomics from a bioinformatics angle: A growing field.**
319 *Mass Spectrom Rev* 2015.

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3. Nesvizhskii AI: **Proteogenomics: concepts, applications and computational strategies.** *Nat Methods* 2014, **11:**1114-1125.

4. Ruggles KV, Krug K, Wang X, Clauser KR, Wang J, Payne SH, Fenyo D, Zhang B, Mani DR: **Methods, tools and current perspectives in proteogenomics.** *Mol Cell Proteomics* 2017.

5. Wang X, Slebos RJ, Wang D, Halvey PJ, Tabb DL, Liebler DC, Zhang B: **Protein identification using customized protein sequence databases derived from RNA-Seq data.** *J Proteome Res* 2012, **11:**1009-1017.

6. Crappe J, Ndah E, Koch A, Steyaert S, Gawron D, De Keulenaer S, De Meester E, De Meyer T, Van Criekinge W, Van Damme P, Menschaert G: **PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration.** *Nucleic Acids Res* 2015, **43:**e29.

7. Olexiouk V, Crappe J, Verbruggen S, Verhegen K, Martens L, Menschaert G: **sORFs.org: a repository of small ORFs identified by ribosome profiling.** *Nucleic Acids Res* 2016, **44:**D324-329.

8. Volders PJ, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J, Mestdagh P: **An update on LNCipedia: a database for annotated human lncRNA sequences.** *Nucleic Acids Res* 2015, **43:**D174-180.

9. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F, et al: **Proteogenomics connects somatic mutations to signalling in breast cancer.** *Nature* 2016, **534:**55-62.

10. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, et al: **Proteogenomic characterization of human colon and rectal cancer.** *Nature* 2014, **513:**382-387.

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

11. Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou JY, Petyuk VA, Chen L, Ray D, et al: **Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer.** *Cell* 2016, **166:**755-765.

12. Barbieri R, Guryev V, Brandsma CA, Suits F, Bischoff R, Horvatovich P: **Proteogenomics: Key Driver for Clinical Discovery and Personalized Medicine.** *Adv Exp Med Biol* 2016, **926:**21-47.

13. Deutsch EW, Albar JP, Binz PA, Eisenacher M, Jones AR, Mayer G, Omenn GS, Orchard S, Vizcaino JA, Hermjakob H: **Development of data representation standards by the human proteome organization proteomics standards initiative.** *J Am Med Inform Assoc* 2015, **22:**495-506.

14. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Rompp A, Neumann S, Pizarro AD, et al: **mzML--a community standard for mass spectrometry data.** *Mol Cell Proteomics* 2011, **10:**R110 000133.

15. Jones AR, Eisenacher M, Mayer G, Kohlbacher O, Siepen J, Hubbard SJ, Selley JN, Searle BC, Shofstahl J, Seymour SL, et al: **The mzIdentML data standard for mass spectrometry-based proteomics results.** *Mol Cell Proteomics* 2012, **11:**M111 014381.

16. Griss J, Jones AR, Sachsenberg T, Walzer M, Gatto L, Hartler J, Thallinger GG, Salek RM, 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* 2014, **13:**2765-2775.

17. Walzer M, Qi D, Mayer G, Uszkoreit J, Eisenacher M, Sachsenberg T, Gonzalez-Galarza FF, Fan J, Bessant C, Deutsch EW, et al: **The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics.** *Mol Cell Proteomics* 2013, **12:**2332-2340.

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

18. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, et al: **A draft map of the human proteome.** *Nature* 2014, **509:**575-581.

19. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al: **Mass-spectrometry-based draft of the human proteome.** *Nature* 2014, **509:**582-587.

20. **The SAM/BAM Format Specification Working Group (2014) Sequence alignment/map format specification** [http://samtools.github.io/hts-specs/SAMv1.pdf]

21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

22. **BED format** [http://genome.ucsc.edu/FAQ/FAQformat.html - format1]

23. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Curr Protoc Bioinformatics* 2014, **47:**11 12 11-34.

24. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.

25. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-Silva D, Cummins C, Clapham P, et al: **Ensembl 2017.** *Nucleic Acids Res* 2017, **45:**D635-D642.

26. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D, Gonzalez JN, Guruvadoo L, et al: **The UCSC Genome Browser database: 2017 update.** *Nucleic Acids Res* 2017, **45:**D626-D634.

27. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH: **JBrowse: a next-generation genome browser.** *Genome Res* 2009, **19:**1630-1638.

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

28. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer.** *Nat Biotechnol* 2011, **29:**24-26.

29. Vizcaino JA, Martens L, Hermjakob H, Julian RK, Paton NW: **The PSI formal document process and its implementation on the PSI website.** *Proteomics* 2007, **7:**2355-2357.

30. Mayer G, Montecchi-Palazzi L, Ovelleiro D, Jones AR, Binz PA, Deutsch EW, Chambers M, Kallhardt M, Levander F, Shofstahl J, et al: **The HUPO proteomics standards initiative-mass spectrometry controlled vocabulary.** *Database (Oxford)* 2013, **2013:**bat009.

31. Wang X, Slebos RJ, Chambers MC, Tabb DL, Liebler DC, Zhang B: **proBAMsuite, a Bioinformatics Framework for Genome-Based Representation and Analysis of Proteomics Data.** *Mol Cell Proteomics* 2016, **15:**1164-1175.

32. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D: **BigWig and BigBed: enabling browsing of large distributed datasets.** *Bioinformatics* 2010, **26:**2204-2207.

33. Ghali F, Krishna R, Perkins S, Collins A, Xia D, Wastling J, Jones AR: **ProteoAnnotator--open source proteogenomics annotation software supporting PSI standards.** *Proteomics* 2014, **14:**2731-2741.

34. Perez-Riverol Y, Uszkoreit J, Sanchez A, Ternent T, Del Toro N, Hermjakob H, Vizcaino JA, Wang R: **ms-data-core-api: an open-source, metadata-oriented library for computational proteomics.** *Bioinformatics* 2015, **31:**2903-2905.

35. Deutsch EW, 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:**1150-1159.

36. Olexiouk V, Menschaert G: **proBAMconvert: a conversion tool for proBAM/proBed.** *J Proteome Res* 2017.

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

37. Vizcaino JA, Csordas A, Del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, et al: **2016 update of the PRIDE database and its related tools.** *Nucleic Acids Res* 2016, **44**:11033.

38. Deutsch EW, Lam H, Aebersold R: **PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows.** *EMBO Rep* 2008, **9**:429-434.

39. Okuda S, Watanabe Y, Moriya Y, Kawano S, Yamamoto T, Matsumoto M, Takami T, Kobayashi D, Araki N, Yoshizawa AC, et al: **jPOSTrepo: an international standard data repository for proteomes.** *Nucleic Acids Res* 2017, **45**:D1107-D1111.

40. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N, et al: **ProteomeXchange provides globally coordinated proteomics data submission and dissemination.** *Nat Biotechnol* 2014, **32**:223-226.

41. Martens L, Vizcaino JA: **A Golden Age for Working with Public Proteomics Data.** *Trends Biochem Sci* 2017, **42**:333-341.

42. Vaudel M, Verheggen K, Csordas A, Raeder H, Berven FS, Martens L, Vizcaino JA, Barsnes H: **Exploring the potential of public proteomics data.** *Proteomics* 2016, **16**:214-225.

Figure Titles and Legends:

Figure 1. Overview of the proBAM and proBed proteogenomics standard formats. Both proBAM and proBed can be created from well-established proteomics standard formats containing peptide and protein identification information (mzTab and mzIdentML, blue box), which are derived from their corresponding MS-data spectrum files (mzML, brown box). The proBAM and proBed formats (green box) contain similar PSM related and genomic mapping information, yet proBAM contains more details,

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

437 including enzymatic (protease) information, key in proteomics experiments (enzyme
438 type, mis-cleavages, enzymatic termini, etc) and mapping details (CIGAR, flag, etc).
439 Additionally, proBAM is able to hold a full MS-based proteomics identification result
440 set, enabling further downstream analysis in addition to genome-centric visualization, as
441 it is also the purpose for proBed (purple box).
442
443 **Figure 2. Fields of proBAM and proBed format.** A proBed file holds 12 original BED
444 columns (highlighted by a bold box) and 13 additional proBed columns. The proBAM
445 alignment record contains 11 original BAM columns (highlighted by a bold box) and 21
446 proBAM-specific columns, using the TAG:TYPE:VALUE format. Each row in the table
447 represents a column in proBAM and proBed. The rows are colored to reflect the
448 categories of information provided in the two formats (see color legend at the bottom, the
449 header section of proBAM format is not included here). The rows without any
450 background color in the proBAM table represent original BAM columns that are not used
451 in proBAM but that are retained for compatibility. The last row in grey indicates the
452 customized columns that could be potentially used.
453
454 **Figure 3. Visualization of proBAM and proBed files in genome browsers.** a) IGV
455 visualisation: proBAM (green box) and proBed (red box) files coming from the same
456 dataset (accession number PXD001524 in the PRIDE database). proBed files are usually
457 loaded as annotation tracks in IGV whereas proBAM files are loaded in the mapping
458 section. b) Ensembl visualization: proBAM (green box) and proBed (red box) files
459 derived from the same dataset (accession number PXD000124) illustrating a novel

bioRxiv preprint doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

translational event. The N-terminal proteomics identification result points to an alternative translation initiation site (TIS) for the gene HDGF at a near-cognate start-site located in the 5'-UTR of the transcript (blue box).

doi: https://doi.org/10.1101/152579; this version posted June 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

HUPO-PSI MS output standard format

mzML

.XML

mzTab

.TSV

mzIdentML

.XML

HUPO-PSI MS identifications standard formats

proBed

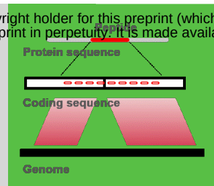
3 stricly mandatory fields from original BED format

chrom
chromStart
chromEnd

9 extra core BED fields

name
score
strand
thickStart
thickEnd
reserved
blockCount
blockSizes
chromStarts

original BED



13 extra proBed fields

proteinAccession
peptideSequence
uniqueness
genomeReferenceVersion
psmScore
fdr
modifications
charge
expMassToCharge
calcMassToCharge
psmRank
datasetID
uri

proBed specific .TSV

- PSM or peptide level
- encouraged to only export valid results

proBAM

HEADER section

@HD VN SO (header line)
@SQ SN LN AS SP (reference sequence lines)
@PG ID VN CL (program line)
@CO (comment lines)

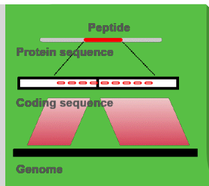
follows a TAG:VALUE format (except @CO), where TAG is two-letter abbrev.

ALIGNMENT section

11 mandatory fields from original SAM/BAM format

QNAME FLAG RNAME POS MAPQ CIGAR RNEXT PNEXT TLEN SEQ QUAL

original BAM



@GA AS VN AD (genome annotation line)

extra mandatory proteogenomics fields

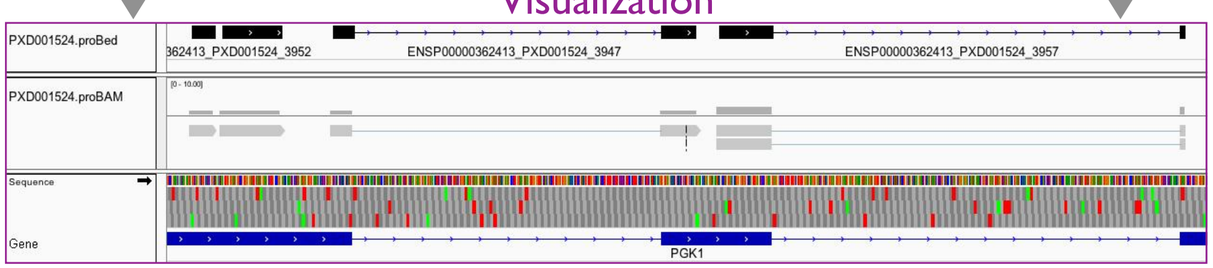
follow a TAG:TYPE:VALUE format (analogous to SAM)
List of tags:
NH XA XB XC XE
XF XG XI XL XM
XN XO XP XQ XR
XS XT XU YA YB
YP Z?

proBAM specific .TSV

- PSM or peptide level
- possible to export only valid or full results

HUPO-PSI proteogenomics standard formats

Visualization



- for annotation/visualization purposes as genome browser tracks

Downstream analysis

- **interpretation**: gene, transcript isoform inference
- **integration**: co-analyse with gen-, transcript-, translat-omics data

- annotation/visualization based on valid results
- downstream (re-)analysis based on full results

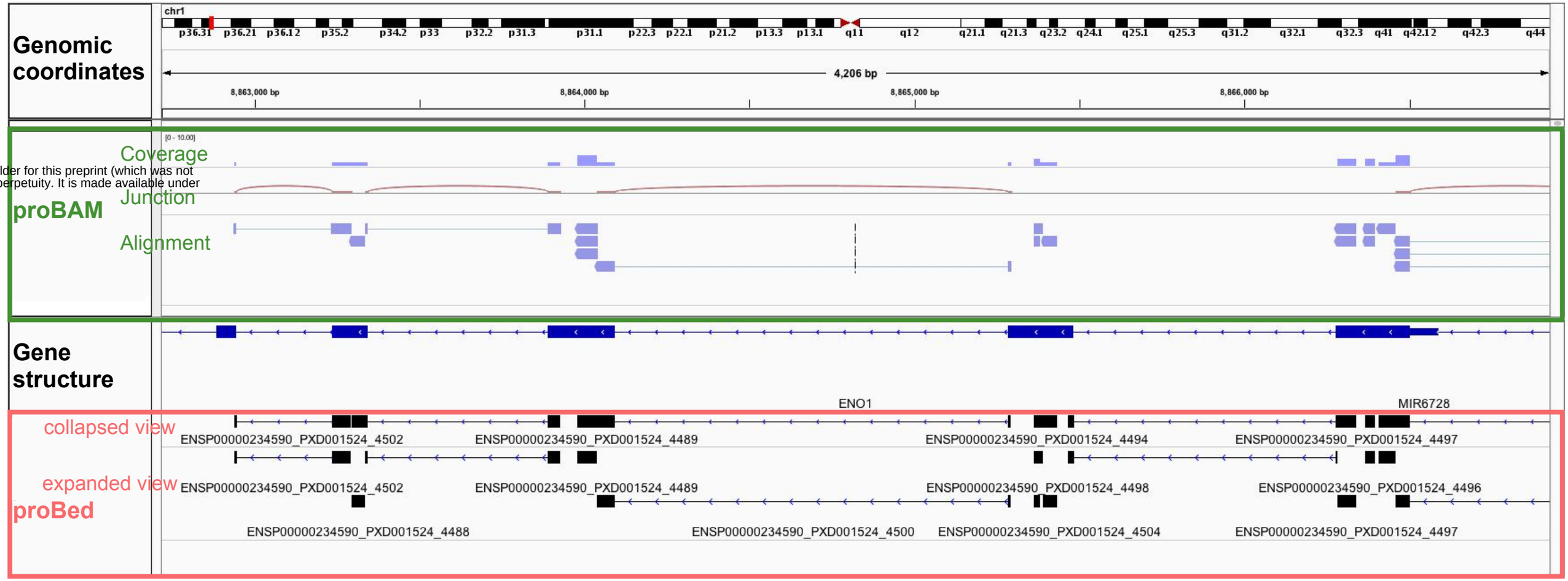
purpose

proBAM	Description	Example
QNAME	Spectrum name	index=7096_PXD001524
FLAG	Bitwise FLAG	16
RNAME	Reference sequence NAME	chr21
POS	1-based leftmost mapping POSition	33907431
MAPQ	-	255
CIGAR	CIGAR string	23M1628N28M
RNEXT	-	*
PNEXT	-	0
TLEN	-	0
SEQ	Coding sequence	TCGACCATTTTCAGCAAG CAAATTGATCAGATTGGT AGTGAGGGGAGAGAA
QUAL	-	*
XL	Number of peptides to which the spectrum maps	XL:i:1
XM	Modification(s): semicolon-separated list of modifications	XM:Z:*
XB	Mass error (experimental - calculated)	XB:f:0.0002109709
XQ	PSM FDR (i.e. q-value or 1-PEP)	XQ:f:1.06E-04
XS	PSM score	XS:f: 79.78288685
NH	Number of genomic locations to which the peptide sequence maps	NH:i:1
XO	Peptide uniqueness (1…5)	XO:Z:unique
XC	Peptide Charge	XC:i:2
XI	Peptide intensity	XI:f:*
XP	Peptide sequence from the original search result	XP:Z:FSPLTTNLINLLAENGR
XR	Reference peptide sequence	XR:Z:FSPLTTNLINLLAENGR
XF	Reading frame of the peptide (0, 1, 2)	XF:Z:0,1
XA	Whether the peptide is well annotated (0,1,2)	XA:i:0
XG	Peptide type (N, V, W, J, A, M, C, E, B, O, T, R, I, G, D, U, X)	XG:Z:N
YP	Protein accession ID from the original search	YP:Z:ENSP00000290299
XE	Enzyme used in the experiment	XE:i:1
XN	Number of missed cleavages in the peptide	XN:i:0
XT	Enzyme specificity (0, 1, 2, 3)	XT:i:3
YA	Following amino acids (2 AA)	YA:Z:LS
YB	Preceding amino acids (2 AA)	YB:Z:ER
XU	Uniform Resource Identifier	.
Z?	Custom fields	.

proBed	Description	Example
chrom	Reference sequence chromosome	chr21
chromStart	Start position of the first DNA base	33907430
chromEnd	End position of the last DNA base	33909107
name	Unique name	ENSP00000290299_3845
score	Score	276
strand	+ or - for strand	-
thickStart	Coding region start	33907430
thickEnd	Coding region end	33909107
reserved	Always 0	0
blockCount	Number of blocks	2
blockSizes	Block sizes	25,26
chromStarts	Block starts	0,1651
psmScore	PSM score	79.78288685
fdr	Estimated global false discovery rate	1.06E-04
modifications	Post-translational modifications	15-UNIMOD:7
expMassToCharge	Experimental mass to charge value	936.499
calcMassToCharge	Calculated mass to charge value	936.497
psmRank	Peptide-Spectrum Match rank.	1
charge	Charge value	2
peptideSequence	Peptide sequence	FSPLTTNLINLLAENGR
uniqueness	Peptide uniqueness	unique
proteinAccession	Protein accession number	ENSP00000290299
genomeReferenceVersion	Genome reference version number	Homo_sapiens.GRCh38.77
datasetID	Dataset Identifier	PXD001524_reprocessed
uri	Uniform Resource Identifier	.

Color legend
Genomic locations
Mapping details
Nucleotide sequence
PSM information
Peptide information
Protein information
Enzyme information
Data source

A.



older for this preprint (which was not
perpetuity. It is made available under

B.