Menschaert, G, Wang, X, Jones, AR, Ghali, F, Fenyo, D, Olexiouk, V, Zhang, B, Deutsch, EW, Ternent, T and Vizcaino, JA

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

http://researchonline.ljmu.ac.uk/id/eprint/11511/

Citation (please note it is advisable to refer to the publisher’s version if you intend to cite from this work)

Menschaert, G, Wang, X, Jones, AR, Ghali, F, Fenyo, D, Olexiouk, V, Zhang, B, Deutsch, EW, Ternent, T and Vizcaino, JA (2018) The proBAM and proBed standard formats: enabling a seamless integration of genomics and proteomics data. Genome Biology, 19. ISSN 1474-760X

LJMU has developed LJMU Research Online for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact researchonline@ljmu.ac.uk
The proBAM and proBed standard formats: enabling a seamless integration of genomics and proteomics data

Gerben Menschaert1*, Xiaojing Wang2,3*, Andrew R. Jones4, Fawaz Ghali4,5, David Fenyö6,7, Volodimir Olexiouk1, Bing Zhang8,9, Eric W. Deutsch10, Tobias Ternent11 and Juan Antonio Vizcaíno11*

Abstract

On behalf of The Human Proteome Organization (HUPO) Proteomics Standards Initiative, we introduce here 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 of the well-defined, widely used file formats SAM/BAM and BED, respectively, and both have been extended to meet the specific requirements entailed by proteomics data. Therefore, existing popular genomics tools such as SAMtools and Bedtools, and several widely used genome browsers, can already be used to manipulate and visualize these formats “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, including the detailed file format specifications and example files, are accessible at http://www.psidev.info/probam and at http://www.psidev.info/probed.

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 are combined with genomics and/or transcriptomics information, typically by using sequence databases generated from DNA-sequencing efforts, RNA-sequencing (RNA-seq) experiments [5], ribosome-profiling (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, 14] has taken the role of developing open community standard file formats for different aspects of MS-based proteomics analysis and data types. At present, well-established data standards are available, for instance, for representing raw MS data.
The existence of compatible and interoperable data formats is a way to facilitate and advance "multi-omics" studies, and a clear need in proteomics, due to the growing importance of the field [9, 10, 20, 21]. However, no standard file format had been established so far for proteogenomics data exchange. To address this problem, we present here two novel standard data formats called proBAM and proBed. As suggested by their names, these two formats are adapted from their genomics counterparts BAM/SAM [22, 23] and BED (Browser Extensible Data) [24], where proBAM stands for proteomics BAM file (compressed binary version of the Sequence Alignment/Map (SAM) format) and proBed stands for proteomics BED file. A key feature of these formats is that they can seamlessly accommodate both regular genomic mapping information and specifics related to proteomics data, i.e. peptide-to-spectrum matches (PSMs) or peptide sequence information. Existing popular genomics tools as SAMtools [22, 23] and Bedtools [25, 26], or the most widely used genome browsers such as Ensembl [27], the University of California Santa Cruz (UCSC) Genome Browser [28], JBrowse [29], and the Integrative Genomics Viewer (IGV) [30], can be used to manipulate and visualize proteomics data in these formats already. We believe that both proBAM and proBed are essential to merge the growing amount of proteomics information with the available genomics/transcriptomics data.

**Overview of the proBAM and proBed formats**

The proteogenomics formats proBAM and proBed are designed to store a genome-centric representation of proteomics data (Fig. 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.

**proBAM overview**

The BAM format was originally designed to hold alignments of short DNA or RNA reads to a reference genome [22, 23]. A BAM file typically consists of a header section storing metadata and an alignment section storing mapping data (Figs. 1 and 2; Additional file 1: Table S1A). 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, http://www.psidev.info/probam#proBAM_specs).

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 (Fig. 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).

**proBed overview**

The original BED format (https://genome.ucsc.edu/FAQ/FAQformat.html#format1), developed by the UCSC, provides a flexible way to define data lines that can be displayed as annotation tracks. proBed is an extension to the original BED file format [28]. 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 (corresponding to genomic coordinates) are mandatory and an additional nine fields are standardized and commonly interpreted by genome browsers and other tools, totaling 12 BED fields, re-used here. The proBed format includes a further 13 fields to describe information primarily on peptide-spectrum matches (PSMs) (Figs. 1 and 2;}

---

**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 [31] 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. At the moment of writing, the PSI review process has been finalized for both formats and version 1.0 of both of them is stable.

Both formats use controlled vocabulary (CV) terms and definitions as part of the PSI-MS CV [32], 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 at http://www.psidev.info/probed.
Additional file 1: Table S1B). The format can also accommodate peptides (as groups of PSMs with the same peptide sequence), but in that case, some assumptions need to be taken in some of the fields (see proBed specification document section 6.8 for details, http://www.psidev.info/probed#proBed_specs).

**Distinct features of proBAM and proBed and their use cases**

The proBAM and proBed formats differ in similar ways as their genomic counterparts do, although representing analogous information. In fact, proBAM and proBed are 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 (Fig. 3a) and a novel translation initiation event in the HDGF gene locus (Fig. 3b), respectively.

Similar to the designed purposes of SAM/BAM, the basic concepts behind the proBAM format are: (1) to provide genome coordinates as well as detailed mapping information, including enzymatic (protease) information, key in proteomics experiments (enzyme type, mis-cleavages, enzymatic termini, etc.) and mapping details (CIGAR, flag, etc.). Additionally, proBAM is able to hold a full MS-based proteomics identification result set, enabling further downstream analysis in addition to genome-centric visualization, as it is also the purpose for proBed.
region of interest, peptides at splice junctions, single nucleotide/amino acid variation, and alternative spliced isoforms (Fig. 3), among others. Therefore, proBAM can hold the full MS proteomics result set, wherein further downstream analysis can be performed: gene-level inference [33], basic spectral count based quantitative analysis, reanalysis based on different scoring systems, and/or false discovery rate (FDR) thresholds.

Fig. 2 Fields of proBAM and proBed format. A proBed file holds 12 original BED columns (highlighted by a bold box) and 13 additional proBed columns. The proBAM alignment record contains 11 original BAM columns (highlighted by a bold box) and 21 proBAM-specific columns, using the TAG:TYPE:VALUE format. Each row in the table represents a column in proBAM and proBed. The rows are colored to reflect the categories of information provided in the two formats (see color legend at the bottom, the header section of proBAM format is not included here). The rows without any background color in the proBAM table represent original BAM columns that are not used in proBAM but that are retained for compatibility. The last row in the proBAM table indicates the customized columns that could be potentially used.

| proBAM | Description | Example |
|--------|-------------|---------|
| QNAME | Spectrum name | index=7096_FXD001524 |
| FLAG  | Browse FLAG | 16      |
| RNNAME| Reference sequence NAME | chr21 |
| POS   | 1-based leftmost mapping POSITION | 33907431 |
| MAPQ  | - | 255 |
| CIGAR | CIGAR string | 23M16SN28M |
| RNEXT | - | * |
| PNEXT | - | 0 |
| TLEN  | - | 0 |
| STIQ  | Coding sequence | TCGACCATTTTACGCAAG CAAAATGACAGATGTTAAGTGAGGCGGAGAGAA |
| QUAL  | - | * |

| 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 | ENSP0000290299_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 |
| blockStarts | block starts | 0,1651 |
| psmScore | PSM score | 79.78288685 |
| file | 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 |
| pmRank | Peptide-Spectrum Match rank | 1 |
| charge | Charge value | 2 |
| peptideSequence | Peptide sequence | FSLPTTVNLLAENGK |
| uniqueness | Peptide uniqueness | unique |
| proteinAccession | Protein accession number | ENSP0000290299 |
| genomeReferenceVersion | Genome reference version number | Homo_sapiens.GRC38.77 |
| 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
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 visualization purposes. As a key point, proBed files can be converted to BigBed [34], a binary format based on BED, which represents a feasible way to store the same information present in BED as compressed binary files, and is the final routinely used format as annotation tracks. It should be noted that a proBAM to proBed conversion should be possible and vice versa. However, “null” values for some of the Tags would be logically expected for the mapping from proBed to proBAM.

Software implementations

Both proBAM and proBed are fully compatible out-of-the-box with existing tools designed for the original SAM/BAM and BED files. Therefore, existing popular tools in the genomics community can readily be applied to read, merge and visualize these formats (Table 1). As mentioned already, several stand-alone and web genome browsers are available to visualize these formats, e.g. UCSC browser, Ensembl, Integrative Genomics Viewer, and JBrowse. For visualizing MS/MS identification results, an integrated proteomics data visualization tool, PDV (Table 1), currently accepts proBAM and matched spectrum file as input.

Routinely used command line tools such as SAMtools allow to manipulate (index, merge, sort) alignments in proBAM. Bedtools, seen as the “Swiss-army knife” tools...
for a wide range of genomic analysis tasks, allows similar actions to both formats, including, among others, intersection, merging, count, shuffling, and conversion functionality. Conversion from proBAM to CRAM format is also enabled by tools as SAMtools, Scramble, or Picard. With the UCSC “bedToBigBed” converter tool (http://hgdownload.soe.ucsc.edu/admin/exe/), one can also convert the proBed to bigBed. In this context, it is important to note that bedToBigBed version 2.87 is highlighted in the proBed format specification as the reliable version that can be used to create bigBed files coming from proBed (version 1.0) files.

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 [35]. In this context, first of all, the open-source Java library ms-data-core-api, created to handle different proteomics data standard formats

| 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                     |                 |
| PGConverter *                 | Command-line tool to convert between the following formats: mzIdentML -> mzTab -> proBed -> bigBed | https://github.com/PRIDE-Toolsuite/PGConverter                      | write/convert   |
| 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 HTML5        | http://jbrowse.org/                                                 |                 |
| PDV                           | PDV accepts proBAM and proBed as the input files to visualize the original spectrum of a PSM. | http://pdv.zhang-lab.org/                                           |                 |
| SAMtools                      | Tool package that provides various utilities for manipulating alignments in the SAM format (including sorting, merging, indexing and conversion to CRAM) | 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).
additional proteogenomics data verification still needs to be implemented.

Discussion
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 [39], PeptideAtlas [40], MassIVE (https://massive.ucsd.edu), and jPOST [41] gathered in the ProteomeXchange Consortium (http://www.proteomexchange.org/) [42]. In fact, an enormous amount of MS data are available in the public domain that can be used for proteogenomics studies, something that it is increasingly happening [43, 44]. The PRIDE database, located in the European Bioinformatics Institute (EMBL-EBI), plans to fully implement proBed in the coming months, facilitating the integration and visualization 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 these 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 the Clinical Proteomic Tumor Analysis Consortium (CPTAC). We hope the data released by such projects will support new tools that support these two formats. We expect that their existence will facilitate integration, visualization, and exchange throughout both the proteomics and genomics communities, and will help multiple proteogenomics endeavors 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 they can be visualized by genome browsers directly and it is possible to re-analyze 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/).

Additional file

Additional file 1: Table S1. Detailed description on the two formats presented, proBAM (S1A) and proBed (S1B). (XLSX 46 kb)

Acknowledgements
JAV, TT, ARJ, and FG acknowledge funding by the BBSRC grants “ProteoGenomics” (grant no. BB/L024225/1) and “PROCESS” (grant no. BB/K01997X/1). ARJ acknowledges BBSRC grant BB/005239/1. GM is a Fellow of the Research Foundation – Flanders (FWO-Vlaanderen) (GM, 12A7813N). XW and BZ are supported by National Cancer Institute award U24CA159988 and U24CA210954. EWD acknowledges funding from NIGMS grant nos. R24GM127667 and R01GM087221 and NIH grant no. U54EB020406. DF is supported by National Cancer Institute award U24CA210972 and by contract 13X5068 from Leidos Biomedical Research, Inc. Finally, the colleagues in the Proteomics Standards Initiative, including the reviewers of the proBAM and proBed format specifications in the PSI document process, are acknowledged for helpful discussions and feedback. The authors also thank Andy Yates (Ensembl team) for his useful comments.

Availability of data and materials
All the related documentation, including the detailed file format specifications, example files, and links to available software tools, are accessible at http://www.psidev.info/probam and at http://www.psidev.info/probed.

Ethics approval and consent to participate
Nothing to declare.

Authors’ contributions
GM, XW, ARJ, VO, BZ, EWD, and JAV developed the proBAM format. TT, FG, DF, ARJ, and JAV developed proBed. GM, XW, and JAV drafted the manuscript. All authors read, revised, and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
1Department of Mathematical Modeling, Statistics and Bioinformatics, Ghent University, Coupure links 653, 9000 Gent, Belgium. 2Greehey Children’s Cancer Research Institute, The University of Texas Health Science Center at
San Antonio, Texas, USA. 3Department of Epidemiology and Biostatistics, The University of Texas Health Science Center at San Antonio, San Antonio, Texas, USA. 4Institute of Integrative Biology, University of Liverpool, Liverpool, UK. 5School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester M1 5GD, UK. 6Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, NY, USA. 7Institute for Systems Genetics, New York University School of Medicine, New York, NY, USA. 8Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX, USA. 9Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. 10Institute for Systems Biology, Seattle, WA, USA. 11European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

Received: 9 June 2017 Accepted: 7 December 2017 Published online: 31 January 2018

References

1. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature. 2016;537:347–55.
2. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods. 2014;11:1114–25.
3. Ruggles KV, Krug K, Wang X, Clauser KR, Wang J, Payne SH, et al. Methods, tools and current perspectives in proteogenomics. Mol Cell Proteomics. 2017;16:595–81.
4. Menschaert G, Fenyo D. Proteogenomics from a bioinformatics angle: A growing field. Mass Spectrom Rev. 2017;36:684–99.
5. Wang X, Slebos RJ, Wang D, Halvey PJ, Tabb DL, Liebler DC, et al. Protein identification using customized protein sequence databases derived from RNA-Seq data. J Proteome Res. 2012;11:1009–17.
6. Crappe J, Nidad E, Koch A, Steyaert S, Gawron D, De Keulenaer S, et al. PROTEOFINDER: deep proteome coverage through ribosome profiling and MS integration. Nucleic Acids Res. 2015;43:e25.
7. Olexiouk V, Van Crielkine W, Menschaert G. An update on sDRFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. 2017; https://doi.org/10.1093/nar/gkx1130.
8. Volders PJ, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J, et al. An update on LNCipedia: a database for annotated human InrRNA sequences. Nucleic Acids Res. 2015;43:174–180.
9. Martens P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016;534:65–62.
10. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014;513:382–7.
11. Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, et al. Integrated proteogenomic characterization of human high-grade serious ovarian cancer. Cell. 2016;166:755–65.
12. Barberi R, Guyven R, Brandsma CA, Suit S, Bischoff R, Horvatovich P. Proteogenomics: key driver for clinical discovery and personalized medicine. Adv Exp Med Biol. 2016;926:41–47.
13. Deutsch EW, Albar JP, Binz PA, Eisenacher M, Jones AR, Mayer G, et al. Development of data representation standards by the human proteome organization proteomics standards initiative. J Am Med Inform Assoc. 2015;22:495–506.
14. Deutsch EW, Orchard S, Binz PA, Bittremieux W, Eisenacher M, Hermjakob H, et al. Proteomics standards initiative: fifteen years of progress and future work. J Proteome Res. 2017;16:4288–98.
15. Martens L, Chambers M, Strum M, Kessner D, Levander F, Shofstahl J, et al. mzML—a community standard for mass spectrometry data. Mol Cell Proteomics. 2011;10:R1 10 00013.
16. Jones AR, Eisenacher M, Mayer G, Kohlhaier Q, Siepen J, Hubbard SJ, et al. The mzIdentML data standard for mass spectrometry-based proteomics results. Mol Cell Proteomics. 2012;11:111:014381.
17. Griss J, Jones AR, Sachsingen T, Walzer M, Gatto L, Hartler J, et al. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Mol Cell Proteomics. 2014;13:2765–75.
18. Walzer M, Qi D, Mayer G, Uzskoreit J, Eisenacher M, Sachsingen T, et al. The mzQuantiML data standard for mass spectrometry-based quantitative studies in proteomics. Mol Cell Proteomics. 2013;12:2332–40.
19. Vazcaino JA, Walzer M, Jimenez RC, Bittremieux W, Bouyssie D, Carapito C, et al. A community proposal to integrate proteomics activities in ELIXIR. F1000Res. 2017. https://doi.org/10.26888/f1000research.11751.
20. Kim MS, Pinto SM, Getnet D, Nirajog RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature. 2014;509:575–81.
21. Wilhelm M, Schlegel J, Hahn M, Ghali AM, Lieberenz M, Savitski MM, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509:582–7.
22. The SAM/BAM Format Specification Working Group. Sequence alignment/map format specification. 2014. http://samtools.github.io/bint-specs/SAMv1.pdf.
23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Hornet N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
24. BED format. http://genome.ucsc.edu/FAQ/FAQformat.html.
25. Quinlan AR. BEDTools: The Swiss-Army tool for genome feature analysis. Curr Protoc Bioinformatics. 2014;47:11.
26. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:641–2.
27. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. Nucleic Acids Res. 2017;45:D633–42.
28. Tyler C, Barber GP, Casper J, Claswson H, Diekman H, Eisenhart C, et al. The UCSC Genome Browser database: 2017 update. Nucleic Acids Res. 2017;45:D26–34.
29. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. Genome Res. 2009;19:1630–8.
30. Robinson JT, Thorvaldsdottir H, Winckler W, Gutmann M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.
31. Vizcaino JA, Martens L, Hermjakob H, Julian RK, Paton NW. The PSI formal document process and its implementation on the PSI website. Proteomics. 2007;7:2355–31.
32. Mayer G, Montecchi-Palazzi L, Ovelliello D, Jones AR, Binz PA, Deutsch EW, et al. The HuPPO proteomics standards initiative-mass spectrometry controlled vocabulary. Database (Oxford). 2013;2013:bat009.
33. 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;15:1164–75.
34. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karholik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010;26:2204–7.
35. Ghal H, Krishna R, Perkins S, Collins A, Xia D, Wastling J, et al. ProteoAnnotator—an open source proteogenomics annotation software supporting PSI standards. Proteomics. 2014;14:2731–41.
36. Perez-Riverol Y, Uszkoreit J, Sanchez A, Ternent T, Del Toro N, Hermjakob H, et al. ms-data-core-api: an open-source, metadata-oriented library for computational proteomics. Bioinformatics. 2015;31:2903–5.
37. Deutsch EW, Mendoza L, Sheteinberg D, Farah T, Lam H, Tasman N, et al. A guided tour of the Trans-Proteomic Pipeline. Proteomics. 2010;10:1150–9.
38. Olexiouk V, Menschaert G, proBAMsuite: a conversion tool for proBAM/proBed. J Proteome Res. 2017;16:2639–44.
39. Vizcaino JA, Csordas A, Del-Toro N, Dianes JA, Griss J, Lavidas L, et al. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 2016;44:1033.
40. Deutsch EW, Lam H, Aebersold R, PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep. 2008;9:429–34.
41. Okuda S, Watanabe Y, Moriya Y, Kawano S, Yamamoto T, Matsumoto M, et al. jPOSTrepo: an international standard data repository for proteomes. Nucleic Acids Res. 2017;45:D1107–11.
42. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Riis D, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol. 2014;32:223–6.
43. Martens L, Vizcaino JA. A golden age for working with public proteomics data. Trends Biochem Sci. 2017;42:333–41.
44. Vaudel M, Verheggen K, Csordas A, Raeder H, Berven FS, Martens L, et al. Exploring the potential of public proteomics data. Proteomics. 2016;16:214–25.