Review of PSI proBed 1.0.0 specification and supporting files

<anon>

This document reviews the Proteomics Standard Initiative specification and supporting files for a standard proteogenomics data format (proBed 1.0.0) based on the UCSC-originated BED file format. The review is in response to a request from the PSI Editor (Martin Eisenacher), and was based on materials posted at http://www.psidev.info/probed.

Summary

The overall assessment is that the document is on the whole clear, complete, and sufficiently detailed for the purposes of both data providers and tool developers. The supporting files are well-designed and example datasets appear to be well-formatted to the standard. It was straightforward to load the example datasets as custom data into the UCSC Genome Browser for visualization and data mining, and the display presents well (screenshots below). There were some issues uncovered that should be addressed, described below.

Significant Issues

1) The specification should include (perhaps begin with) a 1-page summary of the format -- the schema file (.as) provides good starting material for the summary.

2) The example schema file is inconsistently formatted and contains unnecessary filler words and punctuation that impair readability. Also, the file lacks datatypes for the new numeric fields (if a string-only representation is necessary, this should be an auxiliary file). As the schema is included in the bigBed formatted files, and is displayed in browser visualizations, its accuracy and readability is important. Note that a single standard schema file may be used with any proBed file adhering to the base standard (i.e. no optional fields); there is no need (at least at <anon>) for this file to be customized for each dataset (unnecessary labor and opportunity to introduce error) – I recommend removing this requirement (as in Section 10.2) from the specification. (Even if customizing line 1 is desired for some reason, note that line 2 of the schema is intended as the file format description – i.e. BED12+13 PSI proBed 1.0.0 would be appropriate – not the dataset identifier). A revised schema file (proBed.as) addressing these issue is included below.

3) Release of the specification should be accompanied by a file format validation tool that assures adherence to standard and provides helpful error messages for the most commonly expected errors.

Other Issues

Document

1) Section 1.1 – Description of BED format is confusing (12 mandatory fields with 3 required?). Suggest: " 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, so extending the format often adds fields beyond the first 12".

2) Section 5.5.5: *psmScore* field is missing numeric type (double)

3) Section 5.5.6: *fdr* field is missing numeric type (double)

4) Section 10.4: Missing final argument to bedToBigBed – the destination file. (PXD001524_reprocessed.bb). Mention that bedToBigBed with no arguments will print a usage message describing the arguments and options.

5) Footnote 1: Please use most recent UCSC Genome Browser NAR paper:

The UCSC Genome Browser database: 2017 update (doi: 10.1093/nar/gkw1134)

Files

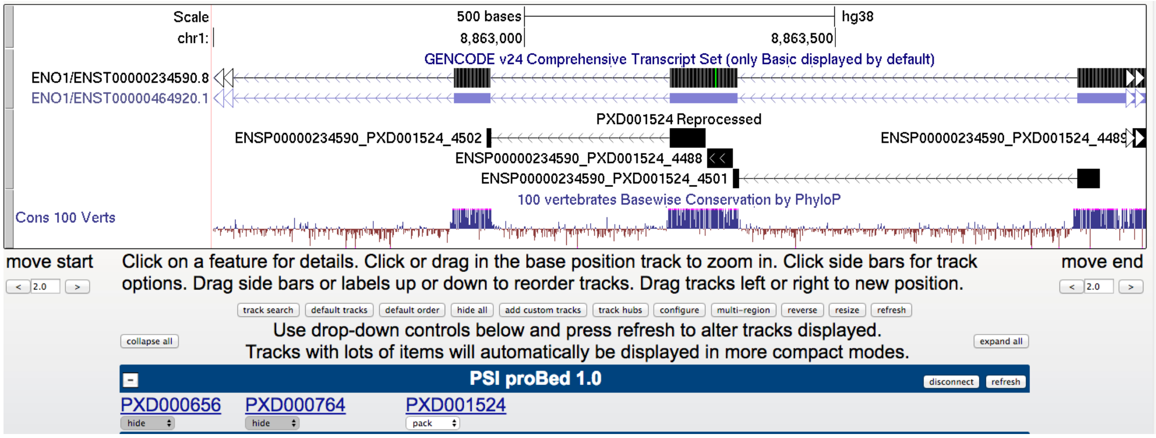
1) Check .bb files for non-printing character in .as file portion. Suggest regenerating with revised proBed.as.

Other Comments

1) Section 5.8: Consider adding a *psmCount* field to the specification for peptides reported based on multiple PSM's.

Files and Screenshots

UCSC Browser view of dataset/region in example Ensembl screenshot:

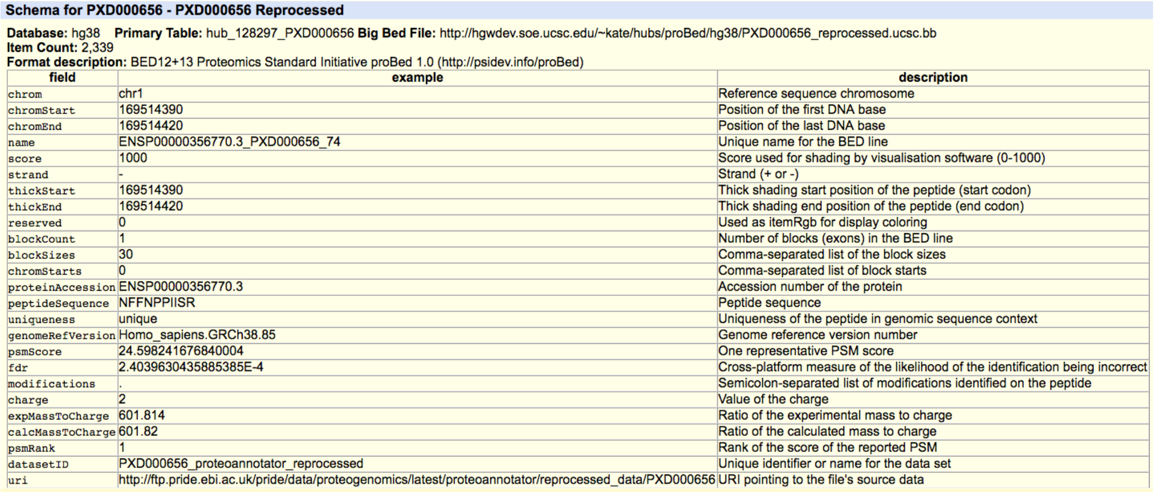


Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

PXD000656 hide
PXD000764 hide
PXD001524 pack

display of proBed schema:



Schema for PXD000656 - PXD000656 Reprocessed

Database: hg38 **Primary Table:** hub_128297_PXD000656 **Big Bed File:** http://hgwdev.soe.ucsc.edu/~kate/hubs/proBed/hg38/PXD000656_reprocessed.ucsc.bb
Item Count: 2,339
Format description: BED12+13 Proteomics Standard Initiative proBed 1.0 (http://psidev.info/proBed)

field	example	description
chrom	chr1	Reference sequence chromosome
chromStart	169514390	Position of the first DNA base
chromEnd	169514420	Position of the last DNA base
name	ENSP00000356770.3_PXD000656_74	Unique name for the BED line
score	1000	Score used for shading by visualisation software (0-1000)
strand	-	Strand (+ or -)
thickStart	169514390	Thick shading start position of the peptide (start codon)
thickEnd	169514420	Thick shading end position of the peptide (end codon)
reserved	0	Used as itemRgb for display coloring
blockCount	1	Number of blocks (exons) in the BED line
blockSizes	30	Comma-separated list of the block sizes
chromStarts	0	Comma-separated list of block starts
proteinAccession	ENSP00000356770.3	Accession number of the protein
peptideSequence	NFFNPPIISR	Peptide sequence
uniqueness	unique	Uniqueness of the peptide in genomic sequence context
genomeRefVersion	Homo_sapiens.GRCh38.85	Genome reference version number
psmScore	24.598241676840004	One representative PSM score
fdr	2.4039630435885385E-4	Cross-platform measure of the likelihood of the identification being incorrect
modifications	.	Semicolon-separated list of modifications identified on the peptide
charge	2	Value of the charge
expMassToCharge	601.814	Ratio of the experimental mass to charge
calcMassToCharge	601.82	Ratio of the calculated mass to charge
psmRank	1	Rank of the score of the reported PSM
datasetID	PXD000656_proteoannotator_reprocessed	Unique identifier or name for the data set
uri	http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/proteoannotator/reprocessed_data/PXD000656	URI pointing to the file's source data

proBed.as (for Section 10.2 of doc, and to include as supporting file in release package)

```
table proBed
"BED12+13 Proteomics Standard Initiative proBed 1.0 (http://psidev.info/proBed)"
(
string  chrom;          "Reference sequence chromosome"
uint    chromStart;     "Position of the first DNA base"
uint    chromEnd;       "Position of the last DNA base“
string  name;           "Unique name for the BED line"
uint    score;          "Score used for shading by visualisation software (0-1000)"
char[1]  strand;        "Strand (+ or -)"
uint    thickStart;     "Thick shading start position of the peptide (start codon)"
uint    thickEnd;       "Thick shading end position of the peptide (end codon)"
uint    reserved;       "Used as itemRgb for display coloring"
int  blockCount;        "Number of blocks (exons) in the BED line"
int[blockCount]  blockSizes; "Comma-separated list of the block sizes"
int[blockCount] chromStarts; "Comma-separated list of block starts"
string  proteinAccession;    "Accession number of the protein"
string  peptideSequence;     "Peptide sequence"
string  uniqueness;          "Uniqueness of the peptide in genomic sequence context"
string  genomeRefVersion;    "Genome reference version number"
double  psmScore;            "One representative PSM score"
double  fdr;                 "Cross-platform measure of the likelihood of the              identification being incorrect“
string  modifications;       "Semicolon-separated list of modifications identified on the                  peptide”
uint  charge;                "Value of the charge"
double  expMassToCharge;     "Value of the experimental mass to charge"
double  calcMassToCharge;    "Value of the calculated mass to charge"
uint    psmRank;             "Rank of the score of the reported PSM”
string  datasetID;           "Unique identifier or name for the data set”
string  uri;                 "URI pointing to the file's source data"
)
```