TPB The Proteome Browser

Data types and mappings

January 2013

Contents

Overview

This document describes the data types currently implemented within TPB, as well as mappings from each data source to these types in a source dependant manner. For each data source (currently neXtProt, GPM and Human Protein Atlas), mappings are provided between the raw source data and TPB data types as well as the thresholds used to define the quality score (i.e. definitions of green, yellow, red and black traffic lights).

Current TPB data types

Protein Expression (PE)

Description: Evidence for the presence of protein expression. It is a summary of numerous underlying data types, currently from MS- or antibody-based methods, or curated annotation. Note that it is not a measure of expression level (quantitative).

Data level: 1

Parent data type: N/A

Child data types: PE MS, PE ANTI, PE OTH

Protein Expression by Mass Spectrometry (PE MS)

Description: Direct MS-based evidence for protein expression. It is a summary of several underlying data types.

Data level: 2

Parent data type: PE

Child data types: PE MS ANN, PE MS PROB, PE MS SAM

Annotation of protein expression by Mass Spectrometry (PE MS ANN)

Description: Annotated, indirect evidence for MS-based detection of protein expression.

Data level: 3

Parent data type: PE MS

Child data types: None

Direct data sources: neXtProt

Probability-based MS detection of protein expression (PE MS PROB)

Description: Evidence for protein expression by MS, based upon the highest probability in a single analysis.

Data level: 3

Parent data type: PE MS

Child data types: None

Direct data sources: GPM

Frequency of MS detection (PE MS SAM)

Description: Repeated detection of protein expression by MS.

Data level: 3

Parent data type: PE MS

Child data types: None

Direct data sources: GPM

Protein expression by antibody technologies (PE ANTI)

Description: Antibody-based evidence for protein expression. It is a summary of several underlying data types.

Data level: 2

Parent data type: PE

Child data types: PE ANTI ANN, PE ANTI IHC

Annotation of antibodies (PE ANTI ANN)

Description: Annotated availability of antibodies in Human Protein Atlas

Data level: 3

Parent data type: PE ANTI

Child data types: None

Direct data sources: neXtProt

Immunohistochemical detection of protein expression (PE ANTI IHC)

Description: Detection of protein expression using immunohistochemical methods.

Data level: 3

Parent data type: PE ANTI

Child data types: PE ANTI IHC NORM

Immunohistochemical detection in normal tissues (PE ANTI IHC NORM)

Description: Detection of protein expression in “normal” (non-diseased) tissue by immunohistochemical methods.

Data level: 4

Parent data type: PE ANTI IHC

Child data types: None

Direct data sources: Human Protein Atlas

Other evidence for protein expression (PE OTH)

Description: Any non MS- or antibody-based evidence for protein expression.

Data level: 2

Parent data type: PE

Child data types: PE OTH CUR

Curated annotation of protein expression (PE OTH CUR)

Description: Curated annotation of protein expression.

Data level: 3

Parent data type: PE OTH

Child data types: None

Direct data sources: neXtProt

Source to data type and quality score mappings

Introduction

For each source repository, a summary of the source files utilised is provided, along with mappings of each data type that is derived from the source and the colour mappings used.

neXtProt

Source file format: XML

Source repository:

Data file(s): ftp://ftp.nextprot.org/pub/current_release/xml/nextprot_all.xml.gz

Schema: ftp://ftp.nextprot.org/pub/current_release/xml/nextprotExport.xsd

Data mappings:

1. TPB data type: PE OTH CUR

   Source data: XPath: proteins/protein/proteinExistence@value

   Quality score: Based on direct mapping from source data to following colour levels.

4 (green):	"protein level"
3 (yellow):	N/A
2 (red):	"transcript level"
1 (black):	"homology", "predicted" or "uncertain"

2. TPB data type: PE ANTI ANN

   Source data: XPath: proteins/protein/xrefs/xref where @category="Antibody databases", @database="HPA" and @accession starts with "CAB" or "HPA"

   Quality score: Based on count of number of antibodies available. Note, only one entry per protein entry

4 (green):	N/A
3 (yellow):	count >1
2 (red):	count =1
1 (black):	count=0

3. TPB data type: PE MS ANN

   Source data: XPath: proteins/protein/xrefs/xref where @category="Proteomic databases" and @database="PeptideAtlas" or "PRIDE"

   Quality score: Based on the @database value

4 (green):	N/A
3 (yellow):	PeptideAtlas
2 (red):	PRIDE
1 (black):	N/A

GPM

Source file format: XML (customised by R.Beavis for TPB)

Source repository:

Data file(s): URL to the current version is available in an RSS feed at http://gpmdb.thegpm.org/tpb/current.xml

Schema: gpm2tpb_schema.xsd

Data mappings:

1. TPB data type: PE MS PROB

   Source data: XPath: gpmdbsummary/protein/identification@beste

   Quality score: Based on the highest log(e) score for each protein

   4 (green): less than or equal to -10

   3 (yellow): less than or equal to -3 and greater than -10

   2 (red): less than or equal to -1 and greater than -3

   1 (black): higher than -1

2. TPB data type: PE MS SAM

   Source data: XPath: gpmdbsummary/protein/identification@samples

   Quality score: Based on number of samples in which the protein was detected

   4 (green): 100 or more samples

   3 (yellow): 20 to 99 samples

   2 (red): 1 to 19 samples

   1 (black): not detected

Human Protein Atlas

Source file format: XML

Source repository:

Data file(s): http://www.proteinatlas.org/download/proteinatlas.xml.zip

Schema: http://www.proteinatlas.org/download/proteinatlas.xsd

Data mappings:

1a. TPB data type: PE ANTI IHC NORM

Source data: XPath: proteinAtlas/entry/tissueExpression@type="APE"&@technology="IH"

Quality score: Based on reliability scores generated by HPA (documentation available at http://www.proteinatlas.org/about/quality+scoring#re) though with more stringent colour mappings. Only proteins with positive expression are provided the following quality scores, any tissue that has no expression is given a quality score of 1 (black).

4 (green):	High
3 (yellow):	Medium
2 (red):	Low or Very low
1 (black):	No protein expression in tissue

1b. TPB data type: PE ANTI IHC NORM

Source data: XPath: proteinAtlas/entry/tissueExpression@type="staining"&@technology="IH"

Quality score: Based on validation scores generated by HPA (documentation available at http://www.proteinatlas.org/about/quality+scoring#va) though with more stringent colour mappings. Only proteins with positive staining are provided the following quality scores, any tissue that has no staining is given a quality score of 1 (black).

4 (green):	N/A
3 (yellow):	Supportive
2 (red):	Uncertain, Non-supportive
1 (black):	Negative for protein staining in tissue