End-to-End Integration of Known Variants and Modifications from PEFF into the Trans-Proteomic Pipeline for Enriched MS/MS Sequence Determination

Luis Mendoza[1], Eric W. Deutsch[1], Jimmy K. Eng[2], and Robert L. Moritz[1]

[1] Institute for Systems Biology, Seattle, WA; [2] University of Washington, Seattle, WA

Institute for Systems Biology

Introduction

The recently released Proteomics Standards Initiative (PSI) Extended File Format (**PEFF**) allows for the specification of known mutations, post-translational modifications (PTMs), and other processing events of a given proteome in a unified, consistent format.

While a handful of database search engines have started to support PEFF, there is no end-to-end informatics pipeline for the validation, mapping, and visualization of those results.

The Trans-Proteomic Pipeline (**TPP**) is a widely used and well-validated **free and open source** suite of software tools that facilitates and standardizes proteomics analysis. **We describe recent enhancements and additions to TPP that enable full analysis, from raw file to the export of validated results and visualization, taking advantage of the rich information contained in PEFF**.

What is PEFF?

PEFF is a "unified format for protein sequence databases to be used by sequence search engines and other associated tools, to enable consistent extraction, display and processing of information such as post-translational modifications, mutations and other processing events," among others.

The format is plain text, largely FASTA-like for backwards compatibility.

"\" Keywords

```
>nxp:NX_Q5EE01-1 \DbUniqueId=NX_Q5EE01-1 \PName=Centromere protein
W isoform Iso 1 \GName=CENPW \NcbiTaxId=9606 \TaxName=Homo Sapiens
\Length=88 \SV=61 \EV=265 \PE=1 \VariantSimple=(4|L)(6|M)(6|V)
(8|P)(8|F)(11|R)(19|H)(19|C)(20|D)(24|Q)(28|L)(28|P)(31|R)(32|*)
(40|N)(41|F)(45|V)(47|F)(52|R)(53|*)(53|Q)(57|D)(59|G)(63|F)(64|V)
(12|H)(26|C)(62|T)(63|S)(74|R)(78|T)(80|M)(86|I)(86|G)
\Processed=(1|88|mature protein)
MALSTIVSQRKQIKRKAPRGFLKRVFKRKKPQLRLEKSGDLLVH
LNCLLFVHRLAEESRTNACASKCRVINKEHVLAAAKVILKKSRG
>nxp:NX_Q5EE01-1 \DbUniqueId=NX_Q5EE01-1 ...
...
```

Sequence

PEFF entry

For more information, please see **Poster MP 438.**

References

TPP	http://www.tppms.org
PEFF	http://www.psidev.info/peff
Comet	http://comet-ms.sourceforge.net
ProteoMapper	http://www.tppms.org/tools/pm

This work was funded in part by the National Institutes of Health grants R01GM087221, R24GM127667, U54EB020406, and P50GM076547.

Methods

Search Engine

The TPP includes the latest version of the **Comet** search engine, which includes PEFF support:

- Both PSI (*ModResPsi*) and Unimod (*ModResUnimod*) encoded **modifications** are supported; the choice of which to use is controlled by a search parameter. Comet currently supports **one** PEFF modification per peptide.
- "*VariantSimple*" **substitutions** are supported. Comet currently supports **one** PEFF amino acid substitution per peptide.
- A peptide can contain either a PEFF modification or an amino acid substitution but not both.
- Standard Comet variable modifications can be analyzed in addition to PEFF modifications and substitutions.
- Standard Comet static modifications are always applied; variable and PEFF modification masses are added on top of static modifications.

PepXML Extensions

In order to report and track the origin of the potential sequence variants and amino acid modifications detected in an experiment, the following extensions have been made to the *pepXML* format:

- A new ***source*** attribute was added to the <mod_aminoacid_mass> tag, with possible values of "*params*" and "*peff*", to represent whether the modification was found as a result of user-specified search parameters, or automatically due to being specified in the PEFF header of the protein in question.
- A **new** <aminoacid_substitution> **tag** to represent a deviation of the matched peptide sequence from the canonical protein sequence.

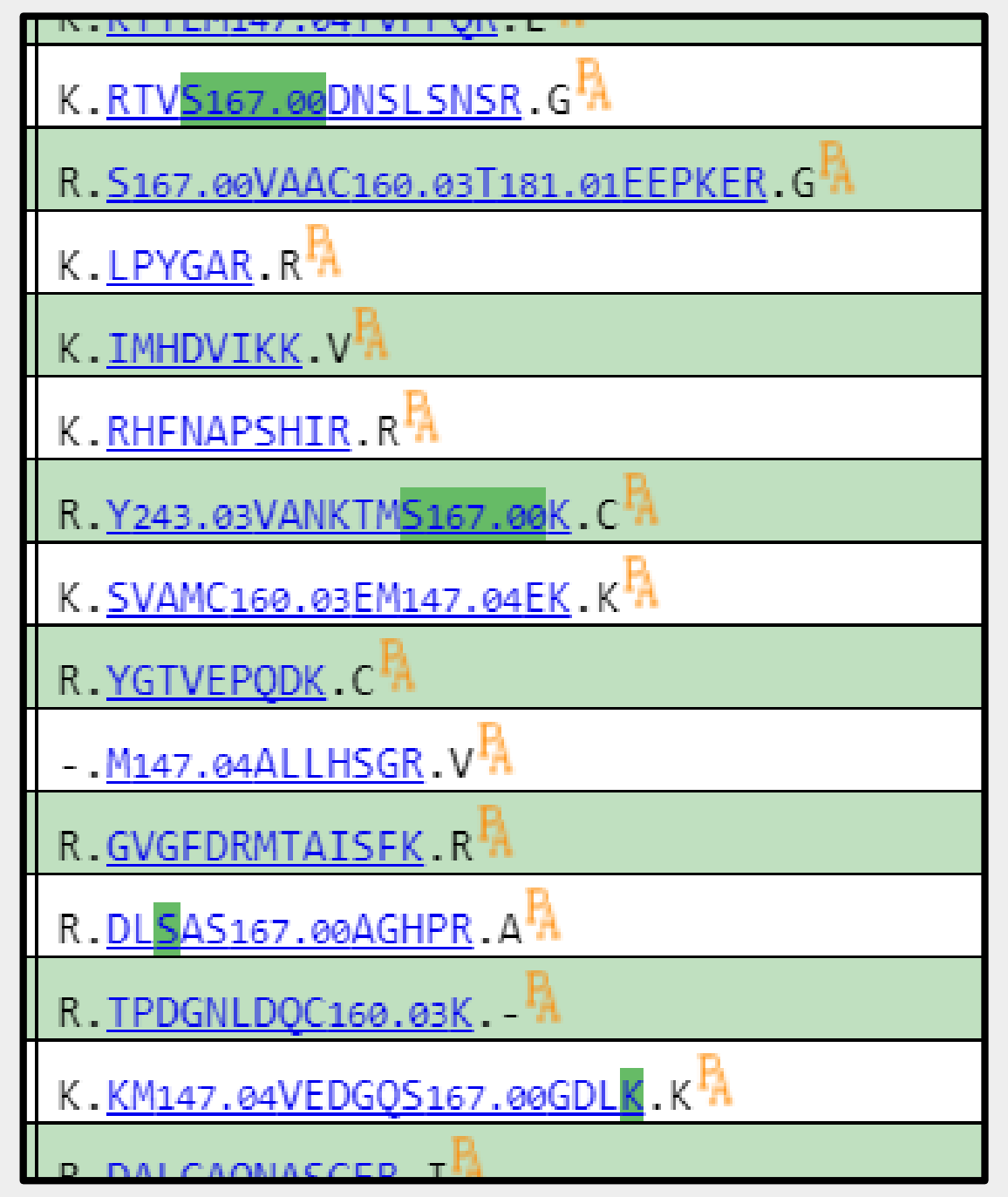
```
<search_hit hit_rank="1" peptide="RPAGPDLSPDGPR" peptide_prev_aa="R“
peptide_next_aa="S“ protein="nxp:NX_P22455-1“ num_tot_proteins="3“
num_matched_ions="4"tot_num_ions="24“ calc_neutral_pep_mass="1413.640193“
massdiff="0.979948" num_tol_term="2“ num_missed_cleavages="0“
num_matched_peptides="66272">
  <modification_info modified_peptide="RPAGPDLS[167]PDGPR">

    <mod_aminoacid_mass position="8" mass="166.998359" variable="79.966331“
     source="peff" id="MOD:00046"/>

    <aminoacid_substitution position="3" orig_aa="P"/>

  </modification_info>
  <search_score name="xcorr" value="1.002"/>
  <search_score name="deltacn" value="0.182"/>
  <search_score name="deltacnstar" value="0.000"/>
  <search_score name="spscore" value="24.2"/>
  <search_score name="sprank" value="6"/>
  <search_score name="expect" value="8.86E+00"/>

</search_hit>
```

PepXML example

PepXMLViewer

PepXMLViewer enables the visual exploration of search results, either directly from the search engine, or validated via statistical tools such as *PeptideProphet* and *iProphet.* PEFF-specific enhanced includes:

- Highlighting mass modifications that originate from PEFF (in contrast to those that are the result of user-specified search parameters)
- Highlighting amino acid substitutions, with the original sequence viewable via mouse-over
- Adding the ability to sort and filter results based on source of modification, and/or presence of aa substitution
- Allowing the export of these results for further analysis.



```
K.RTVS167.00DNSLSNSR.G
R.S167.00VAAC160.03T181.01EEPKER.G
K.LPYGAR.R
K.IMHDVIKK.V
K.RHFNAPSHIR.R
R.Y243.03VANKTMS167.00K.C
K.SVAMC160.03EM147.04EK.K
R.YGTVEPQDK.C
-.M147.04ALLHSGR.V
R.GVGFDRMTAISFK.R
R.DLSAS167.00AGHPR.A
R.TPDGNLDQC160.03K.-
K.KM147.04VEDGQS167.00GDLK.K
```

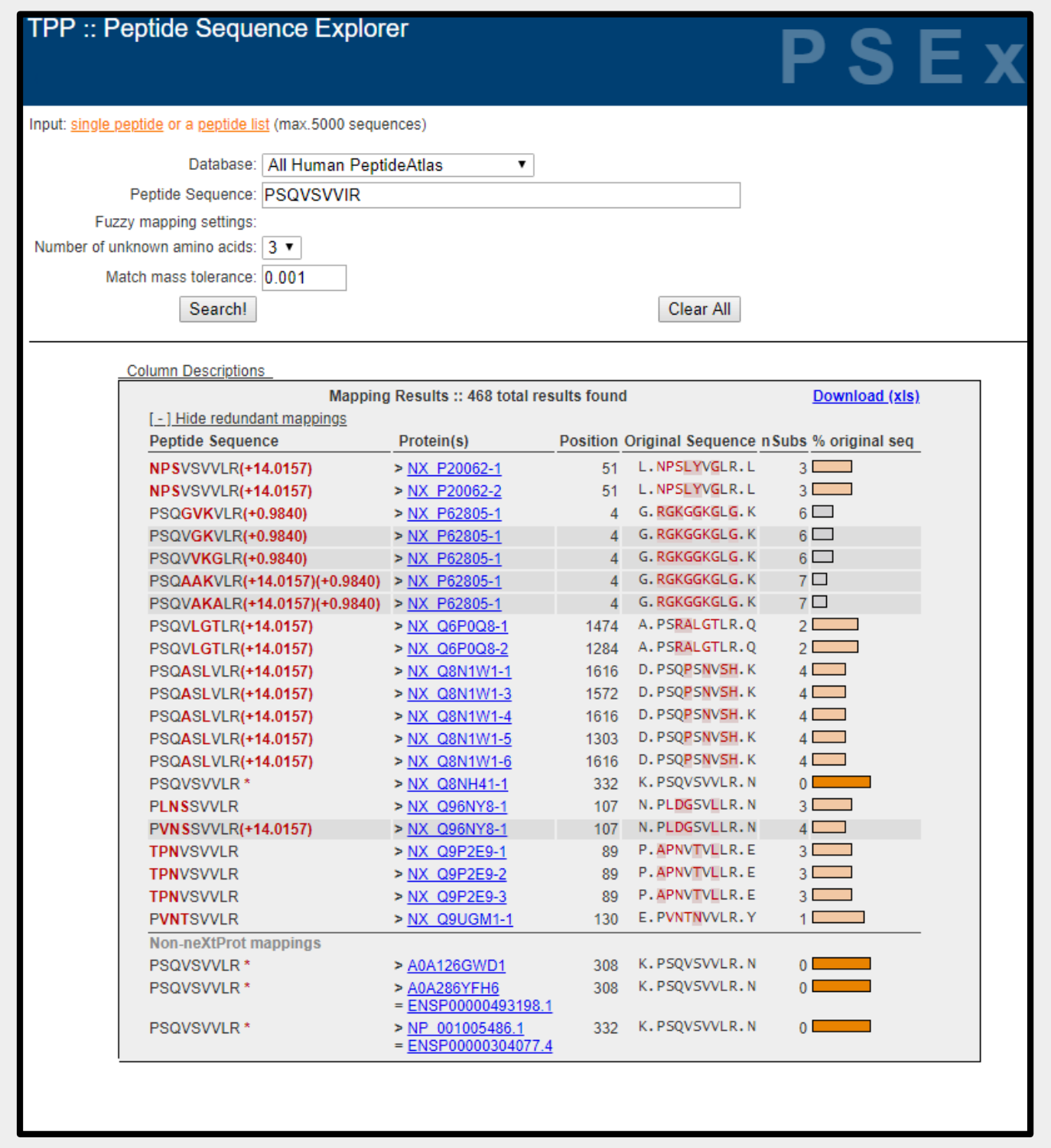
Sequence Mapping

The **ProteoMapper** tools, which consist of a novel peptide-sequence-to-protein mapping mechanism, were added to exhaustively map detected peptide sequences to all possible protein variations, replacing the role of *RefreshParser*.

These tools have been extended to enable to compute and record enzyme-specific values, such as Number of Missed Cleavages (**NMC**) and Number of Enzymatic Termini (**NTT**), based on modified sequences.

A stand-alone peptide-to-protein **mapping interface** has been added to enable users to explore potential sequence mappings, including testing for proteotypicity, with links to external resources (PeptideAtlas, neXtProt, etc).

Since the ProteoMapper tools make use of pre-indexed files of protein segments, an interface was added to the Petunia GUI to allow users to generate them.



TPP :: Peptide Sequence Explorer

Input: single peptide or a peptide list (max.5000 sequences)

Database: All Human PeptideAtlas

Peptide Sequence: PSQVSVVIR

Fuzzy mapping settings:

Number of unknown amino acids: 3

Match mass tolerance: 0.001

Mapping Results :: 468 total results found

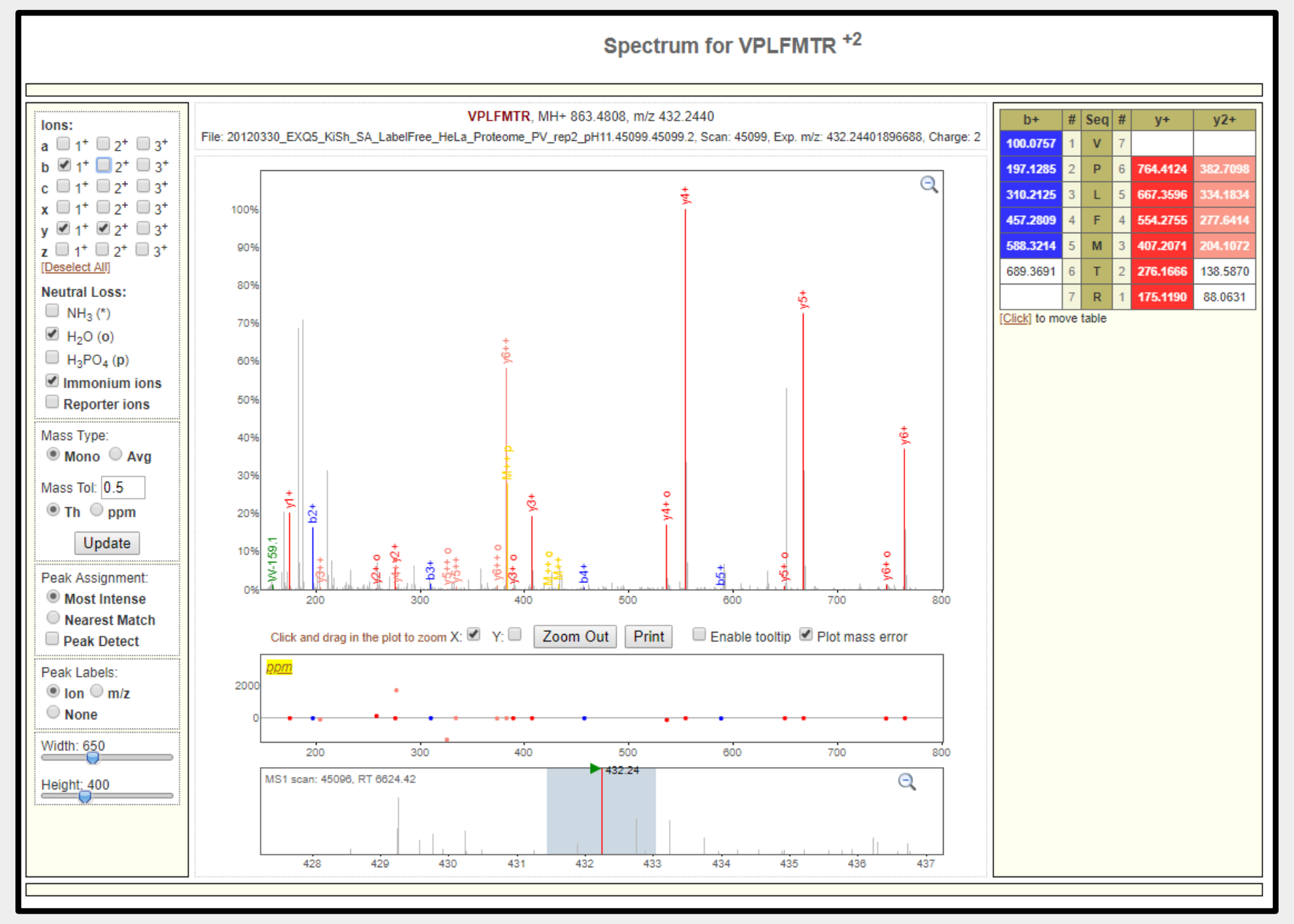
Peptide Sequence	Protein(s)	Position	Original Sequence	nSubs	% original seq
NPSVSVVLR(+14.0157)	> NX_P20062-1	51	L.NPSLYVGLR.L	3	
NPSVSVVLR(+14.0157)	> NX_P20062-2	51	L.NPSLYVGLR.L	3	
PSQGVKVLR(+0.9840)	> NX_P62805-1	4	G.RGKGGKGLG.K	6	
PSQVGKVLR(+0.9840)	> NX_P62805-1	4	G.RGKGGKGLG.K	6	
PSQVVKGLR(+0.9840)	> NX_P62805-1	4	G.RGKGGKGLG.K	6	
PSQAAKVLR(+14.0157)(+0.9840)	> NX_P62805-1	4	G.RGKGGKGLG.K	7	
PSQVAKALR(+14.0157)(+0.9840)	> NX_P62805-1	4	G.RGKGGKGLG.K	7	
PSQVLGTLR(+14.0157)	> NX_Q6P0Q8-1	1474	A.PSRALGTLR.Q	2	
PSQVLGTLR(+14.0157)	> NX_Q6P0Q8-2	1284	A.PSRALGTLR.Q	2	
PSQASLVLR(+14.0157)	> NX_Q8N1W1-1	1616	D.PSQPSNVSH.K	4	
PSQASLVLR(+14.0157)	> NX_Q8N1W1-3	1572	D.PSQPSNVSH.K	4	
PSQASLVLR(+14.0157)	> NX_Q8N1W1-4	1616	D.PSQPSNVSH.K	4	
PSQASLVLR(+14.0157)	> NX_Q8N1W1-5	1303	D.PSQPSNVSH.K	4	
PSQASLVLR(+14.0157)	> NX_Q8N1W1-6	1616	D.PSQPSNVSH.K	4	
PSQVSVVLR *	> NX_Q8NH41-1	332	K.PSQVSVVLR.N	0	
PLNSSVVLR	> NX_Q96NY8-1	107	N.PLDGSVLLR.N	3	
PVNSSVVLR(+14.0157)	> NX_Q96NY8-1	107	N.PLDGSVLLR.N	4	
TPNVSVVLR	> NX_Q9P2E9-1	89	P.APNVTVLLR.E	3	
TPNVSVVLR	> NX_Q9P2E9-2	89	P.APNVTVLLR.E	3	
TPNVSVVLR	> NX_Q9P2E9-3	89	P.APNVTVLLR.E	3	
PVNTSVVLR	> NX_Q9UGM1-1	130	E.PVNTNVVLR.Y	1	
Non-neXtProt mappings					
PSQVSVVLR *	> A0A126GWD1	308	K.PSQVSVVLR.N	0	
PSQVSVVLR *	> A0A286YFH6 = ENSP00000493198.1	308	K.PSQVSVVLR.N	0	
PSQVSVVLR *	> NP_001005486.1 = ENSP00000304077.4	332	K.PSQVSVVLR.N	0	

Results

Naturally-occurring mutations are present in many, if not most, organisms, yet the vast majority of protein identification and validation is made against a single canonical database of sequences that contain little or no knowledge of such variants. A further complication arises when post-translational modifications that are not specified in the search parameters are present in the sample, which leads to incorrect search results for those spectra.

By automatically incorporating previously observed and validated sequence variants and PTMs, a larger share of high quality spectra are confidently identified, increasing sensitivity while potentially decreasing the false discovery error rate.

SAAV/SNP example: prior to using PEFF modifications and mapping, one of the highest-scoring spectra that was assigned to a (incorrect) decoy sequence in *PeptideAtlas* was found to map to a well characterized protein via a SAAV that has been observed in several experiments:



Spectrum for VPLFMTR +2

VPLFMTR, MH+ 863.4808, m/z 432.2440

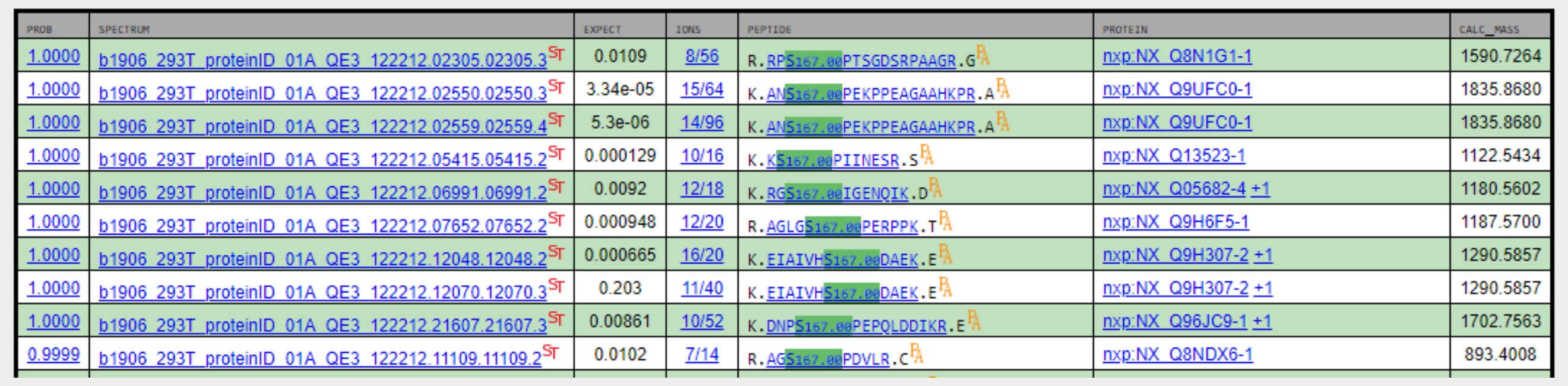
File: 20120330_EXQ5_KiSh_SA_LabelFree_HeLa_Proteome_PV_rep2_pH11.45099.45099.2, Scan: 45099, Exp. m/z: 432.24401896688, Charge: 2

b+	#	Seq	#	y+	y2+
100.0757	1	V	7		
197.1285	2	P	6	764.4124	382.7098
310.2125	3	L	5	667.3596	334.1834
457.2809	4	F	4	554.2755	277.6414
588.3214	5	M	3	407.2071	204.1072
689.3691	6	T	2	276.1666	138.5870
	7	R	1	175.1190	88.0631

USI :

```
Canonical protein sequence:  K.VPLFMSR.A
Observed peptide sequence :  K.VPLFMTR.A
Known variant dbSNP:rs1147990
```

PTM example: in a standard Comet search with **no** modifications specified in the user parameters, about **0.6%** of spectra (65 out of 10,223) with validated **probabilities** of **1.00** by PeptideProphet are matched to sequences with previously-observed modifications, including *Phosphorylation*:



PROB	SPECTRUM	EXPECT	IONS	PEPTIDE	PROTEIN	CALC_MASS
1.0000	b1906_293T_proteinID_01A_QE3_122212.02305.02305.3 ST	0.0109	8/56	R.RPS167.00PTSGDSRPAAGR.G	nxp:NX_Q8N1G1-1	1590.7264
1.0000	b1906_293T_proteinID_01A_QE3_122212.02550.02550.3 ST	3.34e-05	15/64	K.ANS167.00PEKPPEAGAAHKPR.A	nxp:NX_Q9UFC0-1	1835.8680
1.0000	b1906_293T_proteinID_01A_QE3_122212.02559.02559.4 ST	5.3e-06	14/96	K.ANS167.00PEKPPEAGAAHKPR.A	nxp:NX_Q9UFC0-1	1835.8680
1.0000	b1906_293T_proteinID_01A_QE3_122212.05415.05415.2 ST	0.000129	10/16	K.KS167.00PIINESR.S	nxp:NX_Q13523-1	1122.5434
1.0000	b1906_293T_proteinID_01A_QE3_122212.06991.06991.2 ST	0.0092	12/18	K.RGS167.00IGENQIK.D	nxp:NX_Q05682-4 +1	1180.5602
1.0000	b1906_293T_proteinID_01A_QE3_122212.07652.07652.2 ST	0.000948	12/20	R.AGLGS167.00PERPPK.T	nxp:NX_Q9H6F5-1	1187.5700
1.0000	b1906_293T_proteinID_01A_QE3_122212.12048.12048.2 ST	0.000665	16/20	K.EIAIVHS167.00DAEK.E	nxp:NX_Q9H307-2 +1	1290.5857
1.0000	b1906_293T_proteinID_01A_QE3_122212.12070.12070.3 ST	0.203	11/40	K.EIAIVHS167.00DAEK.E	nxp:NX_Q9H307-2 +1	1290.5857
1.0000	b1906_293T_proteinID_01A_QE3_122212.21607.21607.3 ST	0.00861	10/52	K.DNPS167.00PEPQLDDIKR.E	nxp:NX_Q96JC9-1 +1	1702.7563
0.9999	b1906_293T_proteinID_01A_QE3_122212.11109.11109.2 ST	0.0102	7/14	R.AGS167.00PDVLR.C	nxp:NX_Q8NDX6-1	893.4008

Conclusions and Availability

The various updates to TPP allow users to search, validate, visualize results from PEFF searches, and to link to relevant sources of knowledge for further verification.

These features will be available as of the next release of TPP, version **5.3.0**, planned for Summer 2019.