Using the Human Plasma PeptideAtlas to study human plasma proteins

Running head: Using the Human Plasma PeptideAtlas

Methods in Molecular Biology

Serum/Plasma Proteomics

Terry Farrah[1*], Eric Deutsch[1] & Ruedi Aebersold[2]

[1] Institute for Systems Biology, 1441 N 34th St., Seattle, WA 98103, USA

[2] Institute of Molecular Systems Biology, ETH Zurich, Wolfgang-Pauli-Str. 16, 8093, Zurich, Switzerland

[*] To whom correspondence should be addressed. tfarrah@systemsbiology.org.

Abstract/Summary

PeptideAtlas is a web-accessible database of LC-MS/MS shotgun proteomics results from hundreds of experiments conducted in diverse laboratories. Ninety-one experiments on human plasma and serum are included in the subsection, or "build", named the Human Plasma PeptideAtlas. Using the PeptideAtlas web interface, users can browse and search identified peptides and identified proteins, view spectra, and select proteotypic peptides. Users can easily view auxiliary information such as chromosomal mapping, sequence alignments, and much more. Herein, the reader is instructed in the use of the Human Plasma PeptideAtlas through an illustrated example.

Keywords

proteomics, plasma, blood, serum, peptide, database, web server, computer application

1. Introduction

Shotgun proteomics using LC-MS/MS (liquid chromatography, tandem mass spectrometry) is currently the most powerful tool available for discovering proteins

present in human plasma (*1*). As the technique develops, more and more proteins can be identified in a single experiment. To compile the most comprehensive list of human plasma proteins, one approach is to collect all proteins identified in various individual proteomics experiments, as was done by HUPO (the Human Proteome Organization (*2*)) via the Human Plasma Proteome Project (HPPP) in 2003-2005 (*3*). In HPPP Phase I, protein identifications from 18 laboratories were combined, and all proteins identified by at least two laboratories were used to generate a list of 3020 proteins (*4*).

One deficiency of this approach is that each laboratory interprets its data in its own manner, resulting in protein lists that are not comparable. HPPP Phase II (*5*) aims to address this by collecting raw data, rather than protein identifications. The PeptideAtlas project (*6*), a key participant in this effort, collects raw data from experiments conducted in many different laboratories (including HPPP data) and processes them using a common computational pipeline. To date, PeptideAtlas has collected and interpreted 91 experiments from human plasma, yielding over 3 million identified spectra and 20,709 distinctly identified peptides, which provide evidence for at least 2170 different proteins (*7*). While HPPP Phase I resulted in more protein identifications, the PeptideAtlas protein identifications have a higher confidence, with a false discovery rate (FDR) of 1%.

PeptideAtlas is accessible to the public via a web interface. For human plasma (as well as a number of other species proteomes and subproteomes), it provides a large database of proteins, peptides and spectra, with supporting data such as probabilities, FDRs, genome mappings, sequence alignments, links to other databases, uniqueness of peptide-protein mappings, observability of peptides, predicted observable peptides, estimated protein abundances, and cross-references to other databases. The PeptideAtlas also provides many useful methods for accessing the data; the user may search by protein or peptide, or may construct a query to retrieve proteins or peptides with certain characteristics.

This chapter provides an introduction to using the Human Plasma PeptideAtlas by guiding the reader through an illustrated example.

2. PeptideAtlas Construction

First, we provide a description of how data is added to PeptideAtlas, illustrated in **Figure 1**. PeptideAtlas is organized into various *builds*, each encompassing data from a single proteome or subproteome. Each build begins with raw LC-MS/MS spectra contributed by the community. Data can be deposited into one of several repositories or sent directly to the PeptideAtlas project. We then search these spectra against a sequence database (*8*), a spectral library (*9*), or both. Each search assigns a peptide identification and score to each spectrum. Search results are mapped to a comprehensive reference protein database (for human builds, this is a combination of Swiss-Prot (*10*), Ensembl (*11*) and IPI (*12*)), and post-processed using the Trans-Proteomic Pipeline (*13*), a suite of software tools developed at the Institute for Systems Biology for assigning a probability of being correct to each peptide-spectrum match (PSM), distinct peptide identification and protein identification. The Trans-Proteomic Pipeline includes the tools PeptideProphet (*14*), InterProphet (*15*), and ProteinProphet (*16*). Finally, other software processing tools are applied to store PSM, peptide and protein identifications in PeptideAtlas, and to generate and store supporting data such as genome mappings and predicted proteotypic peptides. The atlas build is then made available to the community.
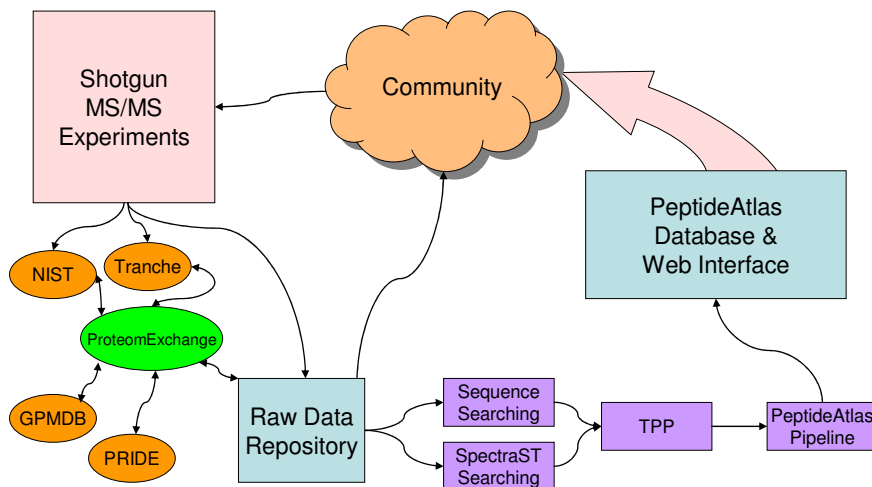

PeptideAtlas
Workflow

Figure 1: PeptideAtlas is built from data provided by the community, and is itself a community resource. Data can be submitted to one of several repositories, or sent directly to the PeptideAtlas project. The data are then processed via a uniform pipeline that includes searching, validating by the Trans-Proteomic Pipeline and post-processing. The results, as well as the raw data, are stored in a database and made accessible via a web interface (this figure first appeared in (*17*)).

PeptideAtlas and its web interface are continually under development, and data are constantly being added. Thus, what the reader sees when using PeptideAtlas may not always exactly match what is described in this tutorial.

3. Using PeptideAtlas

3.1 PeptideAtlas web interface

Go to *www.peptideatlas.org*. The front page provides basic search functionality and displays PeptideAtlas news. To access the full functionality of PeptideAtlas, click GO, without typing anything in the search box. You will see the page shown in **Figure 2**.
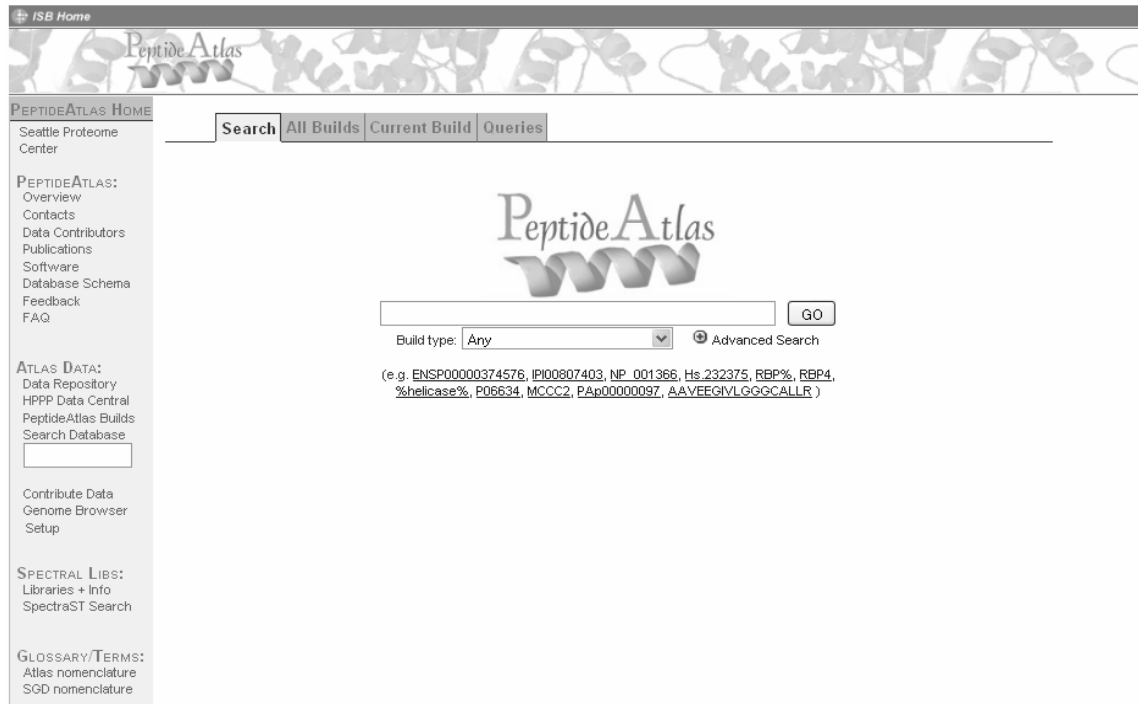
Figure 2: PeptideAtlas web interface. Near the top of the display are four tabs representing four categories of functionality. Search functionality is shown. Users may enter a protein accession, gene name, protein description, peptide accession, or peptide string to quickly access all matching peptides in all of the latest PeptideAtlas builds or in a particular build.

Near the top are four gray tabs representing four categories of functionality:

Search: Keyword search within a single build or across all the most current builds.

All Builds: Allows the user to view all available builds, to select a particular build for functions under the Current Build tab, and to see a peptide's presence across all builds.

Current Build: Allows the user to obtain information on a specific peptide or protein in the currently selected build.

Queries: Allows the user to retrieve information on a set of peptides, proteins, or transitions that satisfy user-specified criteria..

To see the functions that are available for each tab, place the cursor over the tab. Many of these options will be reviewed in detail later in the tutorial.

In the left sidebar are links that will take you to auxiliary pages within PeptideAtlas. Here are descriptions of some key pages:

Overview: Description of how PeptideAtlas is constructed.

Publications: What to cite if you use PeptideAtlas in your research.

Data Repository: Links for downloading much of the raw data used to construct PeptideAtlas.

HPPP Data Central: Background and links for the HUPO Human Plasma Proteome Project.

PeptideAtlas Builds: Lists and download links for all available PeptideAtlas builds.

Search Database: A conveniently accessible search bar with the same functionality as the Search tab (described below).

Contribute Data: How to contribute your own data to the PeptideAtlas project.

Libraries + Info: Access to spectral libraries created from data in PeptideAtlas builds as well as libraries from NIST (*18*) (National Institute of Standards and Technology) and links to other spectral library search resources. You can download the libraries available here and use them to perform spectral library searches of your own data.

SpectraST Search: A web server allowing you to perform a spectral library search using the SpectraST software (*9*).

After exploring these links, click *Search Database* in the left navigation bar, to return to the Search functionality.

3.2 Search for a protein by keyword

Using the Search functionality, you can enter a protein accession, a peptide sequence, a gene name, or a keyword or phrase, and retrieve links to all matching proteins. You can search one build or all current builds.

To illustrate the use of PeptideAtlas, we will investigate the presence of cytokine receptors in plasma. Cytokines are signaling molecules that are secreted by specific cells of the immune system and interact with receptors on the surfaces of other cells, thereby initiating various responses. Cytokine receptors perform their primary functions while embedded in the surfaces of cells. However, under some conditions, cytokine receptor extracellular portions are shed and may then perform a variety of secondary functions. More generally, it has been hypothesized that nearly every human protein will appear in plasma under some conditions.

Click the Build type dropdown menu. You will see a list of all current builds, including two human plasma builds. The more encompassing of these is named "Human Plasma". Click Human Plasma, and then, to search for cytokine receptor names that contain the word "interleukin", type %interleukin% in the search box (the % symbol is a wildcard character) and click GO.

Note that not all of the proteins listed in the results are actually observed in this atlas, but rather, the list includes all proteins within the database that match your search criteria. The rightmost column, N Peptide obs, lists the number of observations of each protein. To focus on observed proteins, perform a descending sort on the last column by clicking the downward-pointing gray triangle.

3.3 Protein View

The second most frequently observed interleukin receptor protein in the Human Plasma PeptideAtlas is Interleukin-6 receptor subunit beta (*IL6ST*). Click the identifier link (ENSP00000314481) for that result. You will be taken to the PeptideAtlas Protein View (**Figure 3**). Note that a different primary tab, Current Build, is now highlighted, and that the secondary tab, Protein, is selected.

For each protein in the reference proteome for a given build, a dynamic Protein View page summarizes the information available for that protein. The page is segmented into several collapsible sections that can be easily minimized by clicking the small icon in the orange section header.

Figure 3: Protein View, top section. For any protein included in a particular PeptideAtlas build, users may retrieve a page such as this displaying alternative names (most of which are clickable hyperlinks to external databases), the full name of the protein, and the number of distinct peptides and spectra (observations) that map to this protein within this build. The Protein View provides additional information further down on the page, as shown in Figures 4 and 5.

The top section provides basic information about the protein, including alternative names (most of which are hyperlinked to external databases), as well as the total number of spectra (observations) and distinct peptides that map to the protein. To learn more about interleukin-6 receptor subunit beta, click the UniProt link (P40189-2). Here, we see that this protein is also known by several other names, including the commonly used name gp130. For the purpose of this tutorial, we will refer to it as *IL-6R-beta*. The UniProt page also describes the structure and function of this molecule and provides literature references. We see that Isoform 1 is a membrane protein, whereas the smaller Isoform 2 (P40189-2) is secreted. Because the protein we are examining lists P40189-2 as its synonym, we know it is the smaller, secreted isoform. According to this page, IL-6R-beta is a signal-transducing molecule participating in the receptor systems for IL6 and several other cytokines.

Going back to the top section of the Protein View page (**Figure 3**), we see that two distinct peptides have been identified as IL-6R-beta from a total of 10 observations (spectra). Just below is an External Links section that takes you to other peptide and protein atlases: the Human Protein Atlas, which lists antibodies available for the protein, and the Global Proteome Machine, which is another database that collects and displays MS/MS search results to the community.
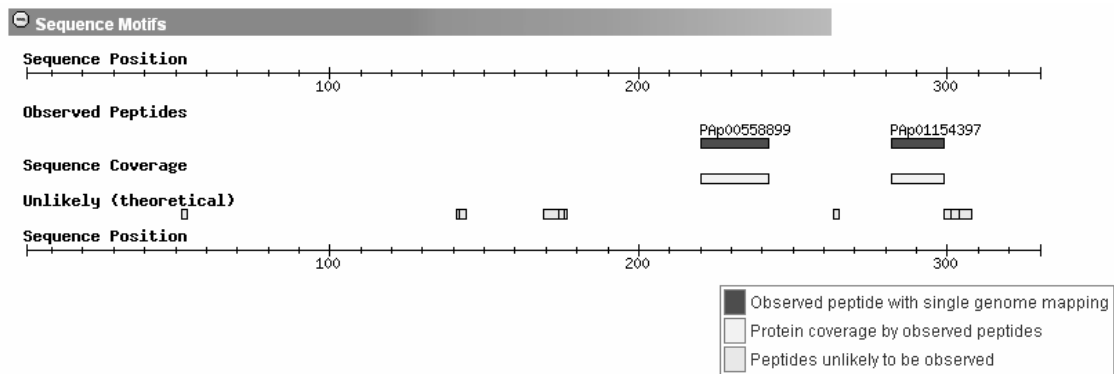
Figure 4: More details from the Protein View page. The Sequence Motifs section displays a linear representation of the protein, with differently shaded bars depicting observed peptides and peptides unlikely to ever be observed due to small size.

The following two sections, Sequence Motifs and Sequence, summarize the peptide coverage of the protein (**Figure 4**). A graphical diagram, similar to a genome browser view, summarizes all the peptides that map either uniquely or redundantly to the protein, plus information on segments unlikely to be observed with mass spectrometers, as well as signal peptides and transmembrane domains, where available. The observed peptides are highlighted in red in the actual protein sequence. We see that for IL-6R-beta, the two observed peptides are in the C-terminal half of the chain, covering only 12.7% of the likely observable sequence. It is not surprising that the protein coverage is so low, given that we expect cytokine receptor proteins to be present at very low abundance in plasma, and only under certain conditions.

show column descriptions

Distinct Observed Peptides (2)

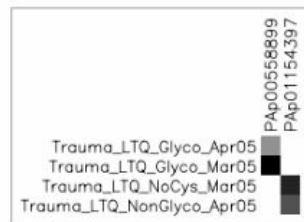| Peptide Accession	Pre AA	Peptide Sequence	Fol AA	Suitability Score	Best Prob	N Obs	EOS	RHS	N Protein Mappings	N Genome Locations	sample_ids	is_subpeptide
PAp01154397	R	SSFTVQDLKPFTEYVFR	I	1.00	1.000	5	1.00	41.85	5	1	2547,2550	
PAp00558899	K	VKPNPPHNLSVINSEELSSILK	L	0.83	1.000	5	0.50	38.60	4	0	2541,2540	

[View peptides and related proteins as a network with Cytoscape]

Predicted Highly Observable Peptides

Peptide Accession	Pre AA	Peptide Sequence	Fol AA	Suitability Score	Detectability Predictor Score	PeptideSieve Score
	K	GYWSDWSEEASGITYEDTIFSF	-	0.83	0.74	0.91
	K	DASTWSQIPPEDTASTR	S	0.80	0.76	0.84
	R	DTPTSCTVDYSTVYFVNIEVWVEAENALGK	V	0.64	0.62	0.65
	K	VTSDHINFDPVYK	V	0.56	0.47	0.64
PAp00025044	R	ETHLETNFTLK	S	0.53	0.60	0.46
PAp00029274	K	CMDYFHVNANYIVWK	T	0.36	0.59	0.13
PAp00558899	K	VKPNPPHNLSVINSEELSSILK	L	0.36	0.57	0.15

Sample peptide map:

Shows per-experiment expression for 40 most highly observed peptides

PAp00558899
PAp01154397

Trauma_LTQ_Glyco_Apr05
Trauma_LTQ_Glyco_Mar05
Trauma_LTQ_NoCys_Mar05
Trauma_LTQ_NonGlyco_Apr05

Observed in Samples:

2540 Trauma_LTQ_Glyco_Apr05
2541 Trauma_LTQ_Glyco_Mar05
2547 Trauma_LTQ_NonGlyco_Apr05
2550 Trauma_LTQ_NoCys_Mar05

Figure 5: The last four sections in the Protein View page. First, observed peptides are listed, with links to further details for each and to a Cytoscape representation of their protein mappings. Second, predicted highly observable peptides are listed, which are of use in the design of targeted proteomics experiments. Third, a graphical map shows which peptides are observed in which samples, with darker shading denoting more observations. Finally, links to the samples are provided.

Next is the Distinct Observed Peptides section, which lists all the observed peptides and maps them to the protein (**Figure 5**). Each peptide has a PeptideAtlas accession of the form PApxxxxxxxx; a peptide with a given sequence has the same accession in any

PeptideAtlas build. The table displays several attributes of the peptides, including the number of times they were observed in the selected build (N Obs), highest PeptideProphet probability among all observations (Best Prob), theoretically calculated hydrophobicity (RHS), and the samples in which the peptides were observed. The Empirical Observability Score (EOS) and Suitability Score metrics are listed as well. The EOS reflects the likelihood that if the protein is detectable in the sample, it will be detected via that peptide. The Suitability Score represents a ranking of how suitable the peptide is as a reference or proteotypic peptide. The score includes information about the total number of observations, the EOS, the best probability of identification. It also includes penalties if the peptides are not fully tryptic or contain missed cleavages or undesirable residues that might impact the suitability of the peptide for targeting (such as methionine, which is variably oxidized).

Immediately below is a Cytoscape (*19*) link, allowing users to see which observed peptides are also observed in related proteins. For IL-6R-beta, this network (**Figure 6**) shows the two peptides listed in the Protein View page, PAp01154397 and PAp00558899, connected to the protein we are examining (ENSP00000314481) and eleven additional proteins. The network also includes two peptides not seen in the Protein View page, PAp00415855 and PAp01420893, that each also map to some of these eleven additional proteins.
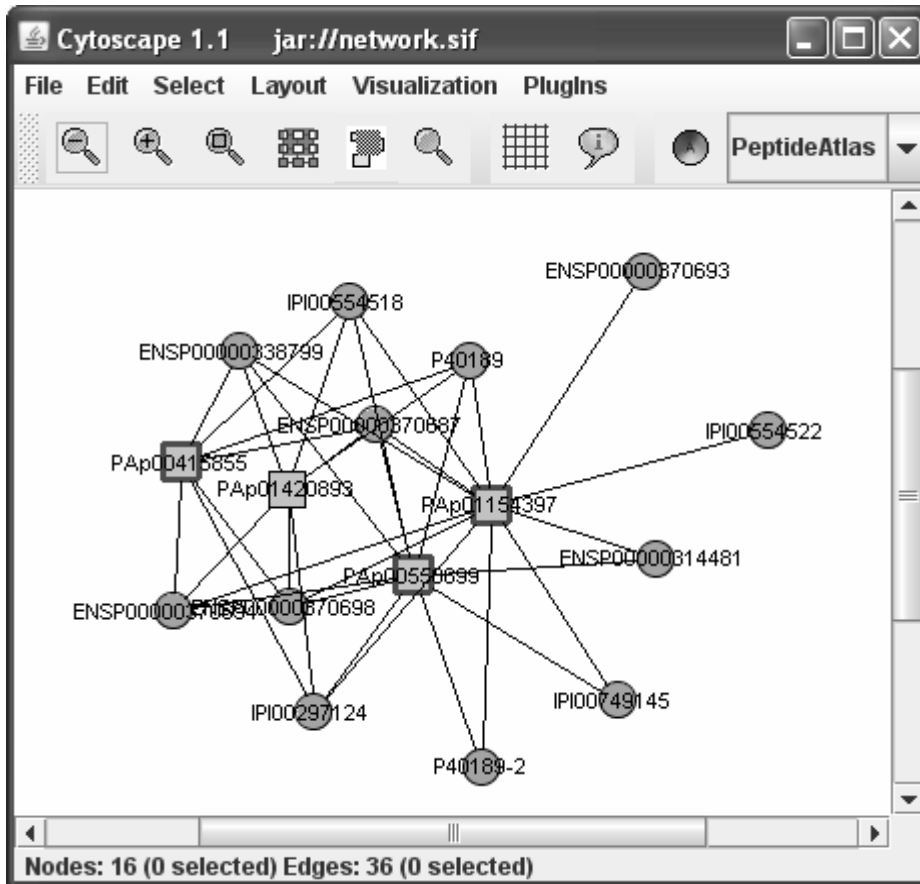
Figure 6: Cytoscape (*19*) depiction of peptide-protein mapping. Squares near the center denote peptides, and circles near the periphery denote proteins. Here, one peptide maps to all twelve proteins shown, while the other peptides map to fewer proteins. The user can grab and move the nodes. The Cytoscape graphic is hyperlinked from the Protein View page.

Below the Cytoscape link is the Predicted Highly Observable Peptides section, listing theoretical peptides for the protein. Each protein is digested *in silico* and both the PeptideSieve (*20*) and DetectabilityPredictor (*21*) software tools are used to predict which peptides might be easily detectable in a electrospray mass spectrometry platform. For low abundance or otherwise hard-to-detect proteins, these theoretical predictions are useful. For IL-6R-beta, we see in the left column that three theoretical peptides have

peptide accessions, indicating that they are actually observed somewhere in PeptideAtlas, although not necessarily in this build.

The process of reliably predicting observable peptides is currently quite weak, and this is reflected in the results seen here. Only one of the empirically observed peptides, PAp00558899, is listed as being highly observable, and it appears at the bottom of the list, with a Suitability Score of only 0.36. The results in this section, therefore, should not be given great weight.

Next, under the heading Sample peptide map, is a graphical depiction of the abundance of each peptide in each sample. Darker shades denote more observations.

Finally, the Observed in Samples section provides links to the samples in which the protein was observed. Note that IL-6R-beta was observed only in samples from trauma patients. IL-6R-beta is known to regulate cell growth and differentiation and to play an important role in immune response. Observation of IL-6R-beta in plasma after trauma may lead an expert to hypothesize further about IL-6R-beta's role in the body's response to trauma. To learn more about these samples, click the links.

3.4 Peptide View

To learn more about the peptide PAp01154397 (SSFTVQDLKPFTEYVFR), click the link on the Protein View page in the Distinct Observed Peptides section, which will take you to the Peptide View page (**Figure 7**).

Human Plasma PeptideAtlas 2009-11 ⌄ Search Peptide Name ⌄ for: PA

⊖ PAp01154397

Peptide Accession	PAp01154397
Peptide Sequence	SSFTVQDLKPFTEYVFR
Best Probability	1.00
Times Observed:	5
Avg Molecular Weight	2063.04
pI (approx)	5.8
SSRCalc relative hydrophobicity	41.85
# Samples	2
# Protein Samples	2
Proteotypic score	1.0
Number of builds in which Peptide found	3
Organisms in which Peptide found	1

⊖ Genome Mappings: 1

Chr	Protein	Residues on Exon	Exon Range	Strand
5	ENSP00000314481 [5 more]	SSFTVQDLKPFTEYVFR	55292066 - 55292116	-
n/a	IPI00297124 [5 more]	SSFTVQDLKPFTEYVFR	n/a	n/a

Compare Proteins

⊖ Modified Peptides

Modified Sequence	Charge	Mono Parent m/z	Best Prob	# Obs	# Siblings	Sample IDs	Consensus Spectra
SSFTVQDLKPFTEYVFR	2	1032.5260	1	2	0.6	2550	▦ ▦
SSFTVQDLKPFTEYVFR	3	688.6866	1	3	0.6	2547,2550	▦ ▦

⊖ Individual Spectra

Modified Sequence	Chg	Smpl	Instr	Prob	Spectrum Name	Spectrum
SSFTVQDLKPFTEYVFR	2	2550	LTQ	0.999	Tao-PlasM12-GNC-14_28Mar05_Doc_0105-07.14256.14256.2	▦
SSFTVQDLKPFTEYVFR	2	2550	LTQ	1	Tao-PlasM12-GNC-15_28Mar05_Doc_0105-06.12705.12705.2	▦
SSFTVQDLKPFTEYVFR	3	2550	LTQ	1	Tao-PlasM12-GNC-14_28Mar05_Doc_0105-07.14232.14232.3	▦
SSFTVQDLKPFTEYVFR	3	2547	LTQ	1	Tao-PlasM12-NNG-14_31Mar05_Doc_0105-07.13708.13708.3	▦
SSFTVQDLKPFTEYVFR	3	2547	LTQ	1	Tao-PlasM12-NNG-15_31Mar05_Doc_0105-06.12619.12619.3	▦

⊖ Observed in Samples:

2547	Trauma_LTQ_NonGlyco_Apr05
2550	Trauma_LTQ_NoCys_Mar05

Figure 7: Peptide View page. Information on any peptide within a build can be retrieved by either the PeptideAtlas peptide accession or peptide sequence. The top section of the page shows general information about the peptide and its presence in the selected PeptideAtlas build. Below that, each distinct genome mapping is detailed, with the last line listing mappings to sequences from protein databases that do not provide genomic information. Third is a listing of all the different modifications and charge states that were observed, with graphical depictions of the consensus spectra for each. Fourth is a listing of each individual spectrum that was identified to this peptide, with links to graphical depictions of the actual spectra. Finally, links are provided to the samples in which this peptide was observed.

For each peptide observed in the data for a given build, this dynamic Peptide View page summarizes the available information. Like the Protein View page, it is segmented into several collapsible sections. The first section provides a number of attributes for the peptide, including predicted hydrophobicity and pI, as well as the number of spectra supporting the identifications. We can see that this peptide is found in other builds (a total of 3, as of this writing); click on the numeral to see more information in a new browser tab using the peptide summary view (**Figure 8**).

As of this writing, this peptide is found in the current (2009-11) human plasma build, the previous build (2009-05), and in a special human plasma build containing only non-glycocapture experiments (Human Plasma Non-Glyco PeptideAtlas 2009-07). In all builds, it is seen less than once per million spectra.
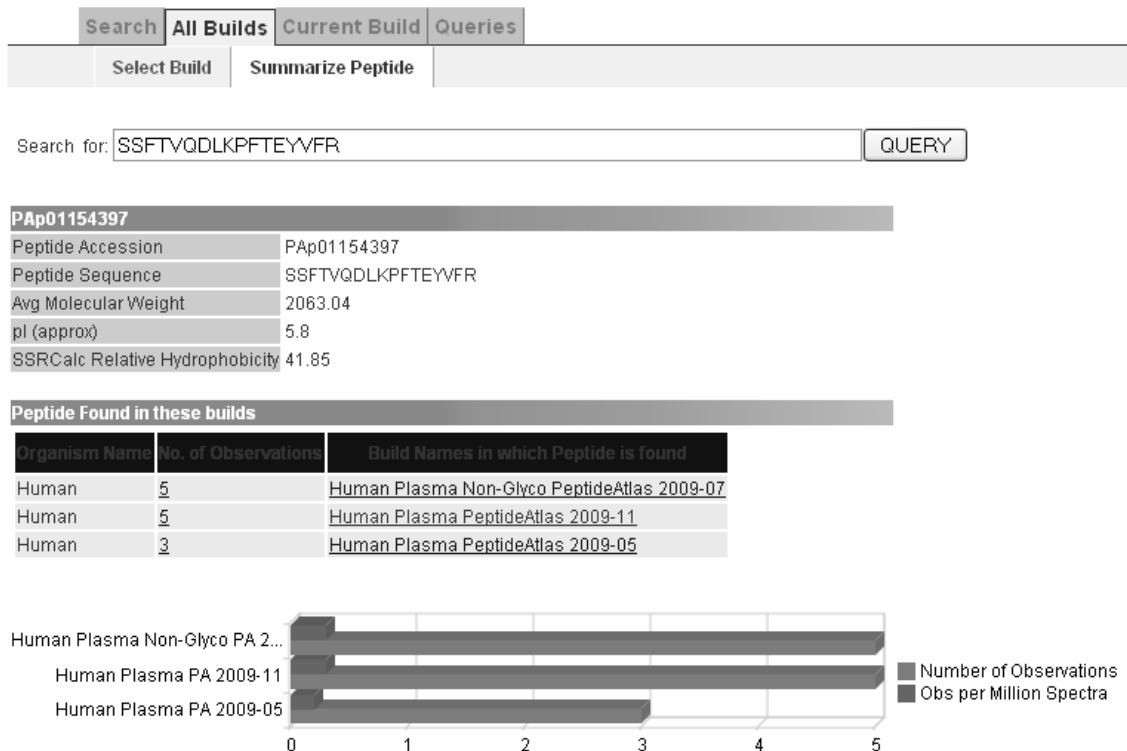
Figure 8: Peptide summary view. Whereas the Peptide View shows information on a peptide within a specific PeptideAtlas build, the peptide summary view shows all the builds in which this peptide was observed. The bar graph shows the abundance of the peptide in each build, both in absolute terms (blue/lighter) and in observations per million spectra (red/darker).

Return to the browser tab displaying the Peptide View page (**Figure 7**) for further examination.

The second section, Genome Mappings, displays the peptide-to-protein and chromosomal mapping information. Since the peptide-to-protein mapping can be multiplex and confusing, this section tries to simplify the mapping information. For IL6-RB, there are two rows in the Genome Mappings table, but only the row for the Ensembl sequence

database contains mapping information. We see that this protein maps to six different Ensembl entries, all of which correspond to a single genome location on chromosome 5. If you click the chromosomal coordinates in the Exon Range column (55292066 – 55292116), you will be taken to the Ensembl page for those coordinates in a new browser tab.

The Compare Proteins link displays an alignment of all the proteins to which a peptide maps, highlighting the peptides observed for each isoform or different protein (**Figure 9**).

Figure 9: A sequence alignment of all proteins to which a peptide maps, accessible by clicking Compare Proteins from the Peptide View page (Figure 7). Lettered superscripts denote sets of identical proteins. For example, all sequences marked "A" are identical. In this illustration, all the sequences are identical in the displayed region (the N-terminal region). The user can scroll to the left to see the rest of the alignment (see Figure 10).

In the sequence alignment for this peptide, we see the same twelve proteins that we saw in the Cytoscape display. The lettered superscripts indicate which sequences are in a set of identical sequences. To show just one of each set, de-select all checkboxes except the uppermost for each letter. Also, leave selected the sequence with no superscript. Then click "Align selected sequences" to align only those sequences (**Figure 10**).

```
ⁿINFDPVYKVKPNPPHNLSVINSEELSSILKLTVTNPSIKSVIILKYNIQYRTKDASTVSQIPPEDTASTRSSFTVQDLKPFTEYVFRIRCMKED
ⁿINFDPVYKVKPNPPHNLSVINSEELSSILKLTVTNPSIKSVIILKYNIQYRTKDASTVSQIPPEDTASTRSSFTVQDLKPFTEYVFRIRCMKED
ⁿINFDPVYKVKPNPPHNLSVINSEELSSILKLTVTNPSIKSVIILKYNIQYRTKDASTVSQIPPEDTASTRSSFTVQDLKPFTEYVFRIRCMKED
ⁿINFDPVYKVKPNPPHNLSVINSEELSSILKLTVTNPSIKSVIILKYNIQYRTKDASTVSQIPPEDTASTRSSFTVQDLKPFTEYVFRIRCMKED
---------LKPNPPHNLSVINSEELSSILKLTVTNPSIKSVIILKYNIQYRTKDASTVSQIPPEDTASTRSSFTVQDLKPFTEYVFRIRCMKED
        . .. .:*************************************************************************************
```

Figure 10: The portion of the sequence alignment shown in Figure 9 that contains the two observed peptides for these sequences. The alignment is restricted to show only the five distinct sequences. Green highlighting (right) indicates the peptide that was described in the Peptide View page from which this alignment was generated. It will be found in all sequences depicted. Blue highlighting (left) shows other observed peptides. In this alignment, the peptide highlighted on the left is seen in only four of the five sequences.

Scrolling to the right, find the first two highlighted peptides. The green one is the one we are currently investigating. The blue one is the other observed peptide for ENSP00000314481. Note that the bottom sequence does not contain the blue peptide. We see that there are a total of four variants (actual or predicted) that contain both peptides. Two of them, P40189 and P40189-2, are identified with Swiss-Prot accessions and correspond to the two splice variants we saw in the UniProt entry for IL-6R-beta. Swiss-Prot is a highly curated protein subset of UniProt; it includes only proteins with good evidence for existence.
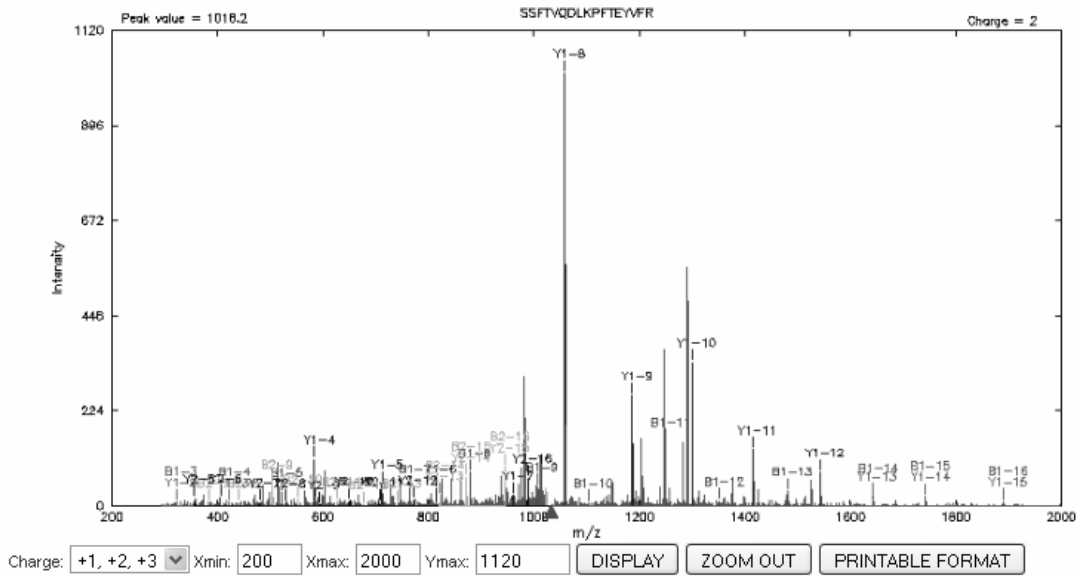
Scrolling further to the right, two peptides are highlighted in blue. Neither of these appears in ENSP00000314481; these are the two additional peptides we saw in the

Cytoscape display. Indeed, one of the sequences to which the peptides map is P40189, the longer splice variant described in the UniProt entry.

Going back to the Peptide View (**Figure 7**) using your browser's back button, the next section, Modified Peptides, lists all of the different observed peptide ions, i.e. the different charge states or mass modifications that were observed. For each peptide ion, the predicted monoisotopic precursor m/z is listed along with the number of observations, samples in which they were seen, and hyperlinks to visualize the consensus spectra for each peptide ion. We see that this peptide is observed in charge states 2 (two observations, one sample) and 3 (three observations, two samples).

The next section, Individual Spectra, lists every spectrum that supports the identification of this peptide, along with individual attributes of the match such as the probability of being correctly identified. View the first spectrum in a new browser tab by clicking its spectrum icon (**Figure 11**).

Finally, at the bottom is a listing of the experiments that included the peptide along with some simple charts depicting the relative number of spectral counts in each experiment.

Peak value = 1018.2 SSFTVQDLKPFTEYVFR Charge = 2

Charge: +1, +2, +3 Xmin: 200 Xmax: 2000 Ymax: 1120 DISPLAY ZOOM OUT PRINTABLE FORMAT

Residue	ion	B	Y	ion +1
S	B1-1	88.0	--	Y1-17
S	B1-2	175.1	1977.0	Y1-16
F	B1-3	322.1	1890.0	Y1-15
T	B1-4	423.2	1742.9	Y1-14
V	B1-5	522.2	1641.9	Y1-13
Q	B1-6	650.3	1542.8	Y1-12
D	B1-7	765.3	1414.7	Y1-11
L	B1-8	878.4	1299.7	Y1-10
K	B1-9	1006.5	1186.6	Y1-9
P	B1-10	1103.6	1058.5	Y1-8
F	B1-11	1250.6	961.5	Y1-7
T	B1-12	1351.7	814.4	Y1-6
E	B1-13	1480.7	713.4	Y1-5
Y	B1-14	1643.8	584.3	Y1-4
V	B1-15	1742.8	421.3	Y1-3
F	B1-16	1889.9	322.2	Y1-2
R	B1-17	--	175.1	Y1-1

Figure 11: Spectrum view for spectrum identified as SSFTVQDLKPFTEYVFR, charge 2. A graphical depiction shows peaks identified as Y-ions in blue, and as B-ions in red (colors not visible in illustration). Overlapping Y-ion and B-ion peaks are depicted in green. Because this identification is for charge 2, it comprises fragment ions in both charge state 1 and charge state 2. Identified peaks for charge state 1 are listed in a chart beneath the spectrum, with colored text allowing easy correlation with the peaks. A similar chart for charge state 2 can be viewed by scrolling down.

3.5 Select a build

You can construct queries to retrieve proteins or peptides fulfilling certain criteria from a specified build. Before doing so, you should select the build by going to All Builds > Select Build (**Figure 12**).
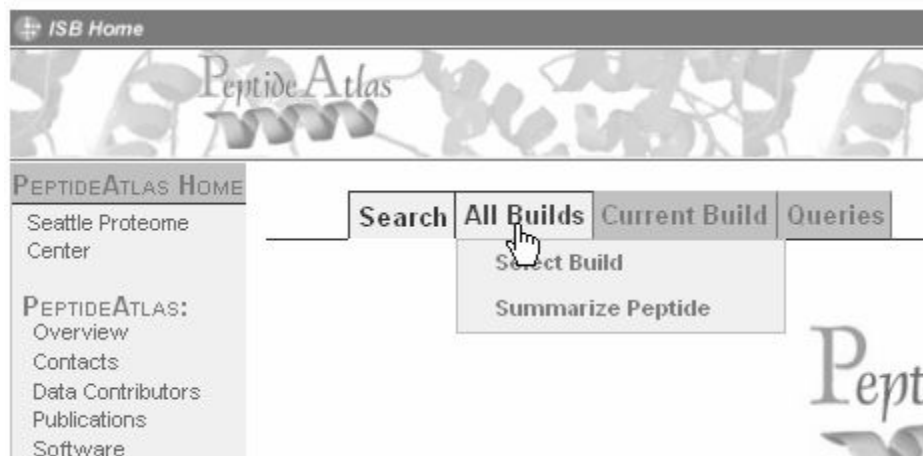

Figure 12: Four primary tabs are displayed on most PeptideAtlas pages, with each allowing access to a set of secondary tabs. The Select Build functionality can be accessed from the primary tab All Builds.

By default, only *default builds* are displayed. These are the latest versions of all builds available to the public. Click the small plus-sign icon to see older builds.

As we saw earlier when using Search, two default human plasma builds are available. On this page, their longer, more descriptive names are displayed:

Human Plasma PeptideAtlas 2009-11

Human Plasma Non-Glyco PeptideAtlas 2009-07

The first build listed corresponds to the one named Human Plasma in the Search interface, and this is the more comprehensive build. Select it by clicking the radio button

to its left. Note that near the top of the page, it now says, "Your current build is: Human Plasma PeptideAtlas 2009-11".

Before leaving this page, click the hyperlink for this atlas to view its Build Summary Page (**Figure 13**).

Figure 13: Build Summary Page for the Human Plasma PeptideAtlas, showing information about the build as a whole.

The top section, Build Overview, gives general information about the build. Most entries are self-explanatory.

For most current builds, the *PSM FDR threshold* describes how PSMs (peptide-spectrum matches) were selected for inclusion in this build. This build of the Human Plasma PeptideAtlas is unusual in that it is a combination of two other builds, each created with a separate PSM FDR. Thus, this build shows N/A for the PSM FDR threshold. Older builds were created with a PeptideProphet probability threshold, usually of 0.9. This is a very generous threshold that results in more false identifications than found in the current build of the Human Plasma PeptideAtlas.

Total observations refers to the total number of PSMs, in other words, the total number of identified spectra.

Further down on the Build Summary page are sections describing the contribution of each sample and a detailed FDR analysis. Click the *show column descriptions* links to learn how to interpret these tables.[1] The Mayu Decoy-Based FDR Analysis table shows, in the 19th column, that this build's protein FDR is 0.00940853, or just under 1%. This is a very high confidence level; of the more than 2000 *canonical* proteins identified (see *Browse Proteins* section below for definition of canonical), we expect only about 20 false identifications.

3.6 Search Proteins

Let's now look up all cytokine receptors in the Human Plasma PeptideAtlas. We can get such a list from Amigo, the Gene Ontology web browser (*http://amigo.geneontology.org*), by typing *cytokine receptor* in the Amigo search bar, selecting *GO terms,* and clicking Submit Query. Three GO terms will be returned. Click the *286 gene products* link for *cytokine receptor activity* (the exact number may be different). On the resulting page, set two filters: click *protein* under Gene Product Type, and *Homo sapiens* under Species. Then click Set Filters. Next, just above the filters section, click *Download all associations information in gene association format.* Your results will be displayed right in the browser. From your browser, save this page in a file with a .tsv extension. Then edit the file (perhaps using Excel) to contain only the UniProt accessions (column 2) and save the file. This yields a list of about 68 accessions, all of which come from the curated Swiss-Prot section of UniProt. Some accessions will be duplicates; in this case there are about 54 distinct accessions.

---

[1] . As of this writing, the Sample Contribution table shows data reflecting the number of *multiply observed* peptides, a vestige of the time when the Trans-Proteomic Pipeline was less good at distinguishing between true and false hits; at that time we believed that peptides observed only once were not reliable enough.

Now, go to the Search Proteins function under the Queries tab, upload the file you just saved, and press QUERY (**Figure 14**). After a minute or two, you will get a table with one line for each input accession. If an accession does not appear in the reference database for the atlas build, the word UNKNOWN will appear in the *hit* column, and you will need to find another accession to get information on that protein. In this case, because all our input accessions were from Swiss-Prot, and the reference database for this PeptideAtlas build includes Swiss-Prot, UNKNOWN should not appear.

Most of the entries have no observations (n_obs = 0). To bring those that were observed to the top, click the downward-pointing triangle in the n_obs header.

Figure 14: Search Proteins functionality takes a PeptideAtlas build and a list of protein accessions as input, and returns one line per accession, showing how many times each protein is observed in that build. Accessions for identical sequences in the build's reference database are also shown. If an input accession is not found in the reference database, it is reported as UNKNOWN (not shown).

We see that nine distinct cytokine receptors are observed in the Human Plasma PeptideAtlas. The first column, *prot_id*, shows the accession we searched for. The second, *hit*, shows the accession used for that protein in the reference database. The fourth column, *equiv_ids*, shows other accessions from the reference database for that protein sequence. To learn about each protein, click links for each protein under *hits* and the Protein View will be displayed. Or, to learn about all nine at once, download those nine into a file and look them up using the Browse Proteins function. To do so, go to the bottom of the page. You will find several download formats available (**Figure 15**).

Q96F46	Q96F46	0	ENSP00000320936
Q99062	Q99062-1	0	ENSP00000355406;ENSP00000362198
Q99650	Q99650-1	0	ENSP00000274276
Q99665	Q99665-1	0	ENSP00000262345

Displayed rows 1 - 50 of 55
Result Page
[1] 2 of 2
Page Size: 50   Page Number: 1   VIEWRESULTSET
Download ResultSet in Format: Excel, XML, TSV, CSV
[Annotate this Resultset] Name: " (2009-11-09 16:5   6)

URL to recall this result set: https://db.systemsbiology.net/sbeams/cgi/shortURL?key=d05gpftj
URL to re-execute this query: https://db.systemsbiology.net/sbeams/cgi/shortURL?key=cqy5tynw

Figure 15: Results from Browse Peptides, Browse Proteins, and Search Proteins can be downloaded in any of four formats by clicking links at the bottom of the page.

Click *CSV,* choose a filename, and save. Then edit the file so that it contains just the top nine distinct accessions that are in the first column (delete the other rows and columns, including the header row and any duplicates). Be sure to save the final file in a text format (.txt, .csv, or .tsv), rather than an Excel format.

Now, go to Browse Proteins under the Queries tab (**Figure 16**).

3.7 Browse Proteins

This page allows you to view detailed information on a set of proteins. The set can either be uploaded as a list, as for the Search Proteins function, or be specified via constraints, or a combination of both.

Figure 16: Browse Proteins functionality. For a particular PeptideAtlas build, a user can retrieve proteins that fulfill a set of constraints. To retrieve all proteins, select no constraints. By default, constraints are specified to show a highly non-redundant list of proteins (shown). This query will return about 2500 proteins. To retrieve all protein sequences in the reference database that contain observed peptides, deselect *canonical, possibly distinguished,* **and** *no redundant relationships.* **This will return**

about 16,000 protein sequences, many of which are identical to one another. To learn more about any constraint, click its question mark icon.

Continuing with our example, enter or browse to the file containing the list of nine accessions you just created. Then, remove the default constraints under Presence Level and Redundancy by scrolling to the blank entry at the bottom of each menu and selecting that. Under Display Options, select *Show Estimated Abundances.* Finally, click QUERY.

A table containing about 66 proteins will be returned. This illustrates the redundancy of the reference database; all proteins from the reference database that are either indistinguishable from (same set of observed peptides) or identical to (sequence-identical) one of the nine input proteins are displayed. Only the top 50 are displayed. To see all 66 proteins, type 66 into the Page Size box at the end of the table and select VIEW_RESULT_SET.

Note that nine of the entries are labeled canonical, and that they are exactly the ones in our query list. This is not an accident: from among each set of indistinguishable and/or identical database entries, PeptideAtlas selects one to assign a *Presence Level,* preferring Swiss-Prot entries, and further preferring Swiss-Prot entries without splice-variant suffixes. We can apply a Presence Level constraint of *canonical* and request *no redundant relationships* to see only those nine proteins. Select those constraints, plus *Show Estimated Abundances* under Display Options. Then—important—browse to the file again, and press QUERY to see just the nine proteins of interest (**Figure 17**).

Biosequence Name	Presence Level	Protein Prophet Prob	Mult Hyp Test Prob	N Obs	N Distinct Peptides	Estimated ng/ml	Uncertainty ng/ml	Redundancy Relationship	Redundant With Respect To	Protein Group	Seq Uniq Prots in Grp	Protein Group Seq Alignmt	Protein Desc
Q9NPH3	canonical	1.000	1.000	23	6	220.000	30x			Q9NPH3	5		Interleukin-1 receptor accessory protein OS=H
Q01638	canonical	1.000	1.000	14	5	210.000	30x			Q01638	4		Interleukin-1 receptor-like 1 OS=Homo sapiens
P40189	canonical	1.000	1.000	12	4	210.000	30x			P40189	5		Interleukin-6 receptor subunit beta OS=Homo s
P48357	canonical	0.999	1.000	3	2	60.000	30x			P48357	6		Leptin receptor OS=Homo sapiens GN=LEPR P
Q99650	canonical	0.999	1.000	8	2	64.000	30x			Q99650	3		Oncostatin-M specific receptor subunit beta O
P27930	canonical	0.964	1.000	10	1	130.000	30x			P27930	1		Interleukin-1 receptor type II OS=Homo sapiens
P17181	canonical	0.964	1.000	1	1	19.000	30x			P17181	4		Interferon-alpha/beta receptor alpha chain OS
P08887	canonical	0.964	1.000	1	1	26.000	30x			P08887	2		Interleukin-6 receptor subunit alpha OS=Homo
P26992	canonical	0.952	1.000	1	1	5.900	10x			P26992	2		Ciliary neurotrophic factor receptor alpha OS=

Displayed rows 1 - 9 of 9

Figure 17: Results of a Browse Proteins query requesting information on nine cytokine receptors. The accessions for these proteins were specified in an uploaded file. All of the proteins are labeled *canonical*, which means that each is considered to be the most representative protein (or one of the most representative proteins) in its protein group. A link is provided in the eleventh column to display all members of any group. Protein probabilities from PeptideProphet (*14*) and a multiple hypothesis-testing technique (*22*) are shown. For large PeptideAtlas builds such as this one, protein probabilities are much less reliable than peptide probabilities and should not be given great weight. The N Obs column shows the number of identified spectra for each protein. Abundance in ng/mL is estimated using a spectral counting method calibrated to abundances reported in the literature (*23*). A multiplicative uncertainty factor is provided for each estimated abundance.

We can see that the nine cytokine receptor proteins observed in the Human Plasma PeptideAtlas include IL-1 receptor accessory protein, IL-1 receptor like protein 1, IL-6R-beta, leptin receptor, oncostatin-M specific receptor subunit beta, IL-1 receptor type II, interferon-alpha/beta receptor alpha chain, IL-6 receptor subunit alpha, and CNTF receptor alpha, and all are observed at low abundances, as expected. The highest estimated abundance is 210 ng/mL with an uncertainty of 30 x. This means that the estimated abundance must be divided by 30 and multiplied by 30 to obtain a lower and upper bound. Therefore, the abundance for the first protein in the list, IL-1 receptor accessory protein, is estimated to be between 7.3 and 6900 ng/mL. Abundances in the Human Plasma PeptideAtlas are estimated based upon the total number of observations (N Obs), calibrated against a set of plasma protein abundances reported in the literature (*23*).

You may wonder how the abundances of these receptors compare to the abundances of other receptors found in this atlas. To retrieve all higher abundance receptors, enter %receptor% in the Description Constraint, >100 in the Number of Observations Constraint, and *no redundant relationships* in the Redundancy constraint. Also, select the

Query Option *Show Estimated Abundances.* Leave all the other constraints unspecified, and click QUERY. We see that four proteins are returned. Although all contain the term *receptor* in their descriptions, further investigation shows that only one, Q86VB7 (Scavenger receptor cysteine-rich type 1 protein M130), is actually a receptor, with an estimated abundance of 730 ng/mL (uncertainty 10 x). Plasma proteins can be present at over 10^6 ng/mL, so we see that even this receptor has a fairly low abundance. We can conclude that, assuming all receptor proteins contain the word *receptor* in their descriptions, no receptor protein is estimated to have an average abundance in human plasma greater than 730 ng/ml.

Proteins that share many observed peptides are partitioned into *protein groups*. To examine the protein group for IL-6R-beta, our original example, enter its Swiss-Prot accession, P40189, into the Protein Group Representative Constraint, select *Show Estimated Abundances,* leave all other constraints unspecified (be sure to erase the ones you entered for Description and Number of Observations), and click QUERY. You will see all proteins from the reference database that belong to the same protein group as IL-6R-beta (**Figure 18**).

Each protein has a Presence Level or a Redundancy Relationship, as follows:

Canonical: The single protein (or, for a few groups, several proteins) that represents the group. Usually canonical proteins have a Swiss-Prot accession. All canonical proteins in an atlas build are well-distinguished from one another; any pair of canonicals will share no more than 80% of their observed peptides. There are 2057 canonical proteins in this build. This number serves as a very conservative count of the number of distinct proteins observed.

Possibly Distinguished: (none in this group) A protein that contains some observed peptides that distinguish it from the canonical(s), but those distinguishing peptides are fewer than 20% of the total number of observed peptides in the protein. There are 506 possibly distinguished proteins in this build.

Subsumed: A protein for which the observed peptides are a subset of the observed peptides for a canonical or possibly distinguished protein. A subsumed protein is a protein that may have been observed, but there is no peptide to distinguish it from its subsuming protein(s).

NTT-subsumed: (none in this group) A protein that has an identical set of peptides to a canonical or possibly distinguished protein, but has fewer tryptic termini. Again, there is no peptide to distinguish it.

Indistinguishable: A protein with observed peptides that are identical to those of a canonical, possibly distinguished, or subsumed protein. Taken together, the canonical, possibly distinguished, subsumed, NTT-subsumed, and indistinguishable proteins of a build comprise an exhaustive list of sequence-unique proteins that may have been observed in the given experiments; for the Human Plasma PeptideAtlas, this list includes over 8000 proteins.

Identical: A protein from the reference database for the build that is identical in sequence to a protein with one of the above labels.

More detailed explanations of this terminology can be found by clicking *Protein ID terms* near the bottom of the left navigation bar.

It is usually not useful to see the proteins labeled *identical*. To remove them from the display, click *Indistinguishable* in the Redundancy constraint menu, and then click QUERY again. You will then see the five non-identical proteins in this group. To see the sequence alignment for the group, click the icon under Protein Group Seq Alignmt. This is the same sequence alignment you saw when clicking Compare Proteins from the Peptide View page.

Biosequence Name ▲▼	Presence Level ▲▼	Protein Prophet Prob ▲▼	Mult Hyp Test Prob ▲▼	N Obs ▲▼	N Distinct Peptides ▲▼	Redundancy Relationship ▲▼	Redundant With Respect To ▲▼	Protein Group ▲▼	Seq Uniq Prots in Grp ▲▼	Protein Group Seq Alignmt ▲▼	
P40189	canonical	1.000	1.000	12	4			P40189	5		Interleukin-6 receptor subunit beta OS=Ho
IPI00554518						identical	ENSP00000370694	P40189	5		IPI:IPI00554518.1|TREMBL:Q5FC04|ENSEMI
ENSP00000370694						indistinguishable	P40189	P40189	5		pep:known chromosome:NCBI36:5:552703
ENSP00000338799						identical	P40189	P40189	5		pep:known chromosome:NCBI36:5:552703
ENSP00000370687						identical	P40189	P40189	5		pep:known chromosome:NCBI36:5:552724
ENSP00000370698						identical	P40189	P40189	5		pep:known chromosome:NCBI36:5:552666
IPI00297124						identical	P40189	P40189	5		IPI:IPI00297124.2|SWISS-PROT:P40189-1|T
IPI00554522						identical	ENSP00000370693	P40189	5		IPI:IPI00554522.1|TREMBL:Q5FC05|ENSEMI
ENSP00000370693	subsumed	0.000	1.000	4	1		P40189	P40189	5		pep:known chromosome:NCBI36:5:552703
P40189-2	subsumed	0.000	1.000	7	2		P40189	P40189	5		Isoform 2 of Interleukin-6 receptor subunit
ENSP00000314481						indistinguishable	P40189-2	P40189	5		pep:known chromosome:NCBI36:5:552817
IPI00749145						identical	P40189-2	P40189	5		IPI:IPI00749145.1|SWISS-PROT:P40189-2|E

Displayed rows 1 - 12 of 12

Figure 18: Browse Proteins results for the protein group whose representative is Interleukin-6 receptor subunit beta (P40189). Two proteins, labeled *subsumed*, are mapped to by peptide sets that are proper subsets of the peptide sets mapping to P40189. ENSP000000370694, labeled *indistinguishable*, is mapped to by the exact same peptide set as P40189. Likewise, ENSP00000314481 is indistinguishable from P40189-2. All the other entries in this table are marked *identical;* each is identical in sequence to some other member of this group.

3.8 Design MRM assays

Multiple reaction monitoring (MRM; also called selected reaction monitoring or SRM) is a proteomics technique that aims to detect the presence of a small number of specific proteins in a sample. For such assays, the mass spectrometer is configured to monitor unique ion signatures, called transitions, of predetermined peptides in order to achieve a detection or confident upper limit for desired peptides to the exclusion of all other peptides. A transition is the combination of precursor m/z value and a product ion m/z value. The design of MRM assays is currently a topic of great interest, and the methods used are under constant development.

PeptideAtlas can be used in conjunction with any number of other tools to efficiently select peptides and transitions for such assays. New features and builds are currently being added to PeptideAtlas to make it even more useful for this task (*24*). Because the

methods used in this field are in such flux, we will only provide a very simplified example here.

Let us design an assay for detecting our example protein, the shorter, secreted isoform of IL-6R-beta. First, we select the peptides we wish to detect. We want to choose *proteotypic* peptides, which are peptides that are easily observable using the available instrument *and* which map uniquely to the protein of interest. If we have the luxury of choosing as many as five peptides, it may make sense to choose the two that are observed, plus the three with the highest suitability scores (refer **Figure 5**). We check the N Protein Mappings column for the observed peptides to see if they are unique to our protein. We see that they have 4 and 5 mappings, all to splice isoforms of the same gene. So if we design our assay to use these peptides, the assay will be specific for products of this gene, but will not distinguish the shorter isoform from the others. Let us assume that this is acceptable.

We should also check the theoretical peptides to see if they are unique to IL-6R-beta, perhaps using the BLAST tool (http://www.ncbi.nlm.nih.gov/BLAST).

.

The next task is transition selection. We will probably be executing our MRM assay on a triple quadrupole instrument. However, the observed peptides were all likely observed on ion trap instruments, because these are the most commonly used instruments for the shotgun experiments contributed to PeptideAtlas. The relative intensities of the fragment ions in triple quad instruments can be quite different from those in ion traps, but are generally fairly similar (*25*), so we will ideally synthesize these five peptides, obtain triple quad spectra, and select transitions from those spectra. If this is not possible, we will manually examine the consensus spectra stored in PeptideAtlas for our selected peptides and choose a few higher intensity transitions for each peptide.

A special build of PeptideAtlas, called MRMAtlas, contains only MS/MS spectra from triple-quad instruments. Only a relatively small number of such spectra are currently available in the MRMAtlas; none are available for human plasma at the time of this

writing, but should become available very soon. However, the data contained therein provide the best available transitions for the proteins represented in these special builds. MRMAtlas is growing and will eventually contain spectra for most of the proteins in the Human Plasma PeptideAtlas.

4. Summary and Conclusion

PeptideAtlas provides a wealth of information and tools for exploring the human plasma proteome. Much can be learned with this resource that cannot be learned from any single experiment. Here, exploring the world of cytokine receptor proteins, we learned that nine have been observed with high confidence among 91 plasma experiments. We examined one receptor protein, the shorter isoform of Interleukin-6 receptor subunit beta, in-depth, and found that within these 91 experiments, it has been observed only in samples from trauma patients. Of the two observed peptides that map to this isoform, we learned that one of them appears in all five IL-6R-beta isoforms found in a comprehensive sequence database, while the other appears in only four isoforms. The tool allowed us to examine the spectra underlying these identifications, to confirm their reliability and/or to select transitions for targeted proteomics assays. We were able to easily access additional information and data displays such as predicted proteotypic peptides, genome mappings, and sequence alignments.

A primary purpose of the PeptideAtlas project is to support the discovery of biomarkers to diagnose and stage human disease, to select effective treatments for an individual, and to assess the progress of treatment. Success in this arena will advance and improve medicine by enabling earlier diagnosis of disease and providing more effective treatment options for patients. Plasma, an easily obtained specimen that contains proteins from many or perhaps all other body tissues, is of tremendous interest for biomarker discovery, and PeptideAtlas facilitates this discovery by integrating data from many diverse experiments and presenting it in a user-friendly and information-rich web resource. Such integrative resources are essential if we are to harness the power of all experimental work

done in plasma proteomics and make progress in this hugely challenging task of transforming the practice of medicine.

Acknowledgements

References

(1)   Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422**, 198-207.
(2)   Hanash, S., and Celis, J. E. (2002) The Human Proteome Organization: a mission to advance proteome knowledge. *Mol Cell Proteomics* **1**, 413-4.
(3)   Omenn, G. S. (2004) The Human Proteome Organization Plasma Proteome Project pilot phase: reference specimens, technology platform comparisons, and standardized data submissions and analyses. *Proteomics* **4**, 1235-40.
(4)   Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W., Menon, R., Hermjakob, H., Apweiler, R., Haab, B. B., Simpson, R. J., Eddes, J. S., Kapp, E. A., Moritz, R. L., Chan, D. W., Rai, A. J., Admon, A., Aebersold, R., Eng, J., Hancock, W. S., Hefta, S. A., Meyer, H., Paik, Y. K., Yoo, J. S., Ping, P., Pounds, J., Adkins, J., Qian, X., Wang, R., Wasinger, V., Wu, C. Y., Zhao, X., Zeng, R., Archakov, A., Tsugita, A., Beer, I., Pandey, A., Pisano, M., Andrews, P., Tammen, H., Speicher, D. W., and Hanash, S. M. (2005) Overview of the HUPO

Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **5**, 3226-45.

(5) Omenn, G. S., Aebersold, R., and Paik, Y. K. (2009) 7(th) HUPO World Congress of Proteomics: launching the second phase of the HUPOPlasma Proteome Project (PPP-2) 16-20 August 2008, Amsterdam, The Netherlands. *Proteomics* **9**, 4-6.

(6) Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R. (2006) The PeptideAtlas project. *Nucleic Acids Res* **34**, D655-8.

(7) Farrah, T., Deutsch, E., Campbell, D., Omenn, G., Aebersold, R. (in preparation) A high-confidence quantitative human plasma proteome in the PeptideAtlas.

(8) Eng, J., McCormack, A. L., and Yates, J. R. (1994) An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **5**, 976-989.

(9) Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., Stein, S. E., and Aebersold, R. (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655-67.

(10) Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-70.

(11) Hubbard, T. J., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A., and Birney, E. (2007) Ensembl 2007. *Nucleic Acids Res* **35**, D610-7.

(12) Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4**, 1985-8.

(13) Keller, A., Eng, J., Zhang, N., Li, X. J., and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* **1**, 2005 0017.

(14) Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74**, 5383-5392.

(15) Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., Mendoza, L., Aebersold, R., and Nesvizhskii, A. Postprocessing and validation of tandem mass spectrometry datasets improved by iProphet. *in preparation*.

(16)   Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75**, 4646-4658.

(17)   Deutsch, E. W. (2010) The PeptideAtlas Project. *Methods in Molecular Biology* **604**, 319-331.

(18)   National Institute of Standards and Technology.

(19)   Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504.

(20)   Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., Kuster, B., and Aebersold, R. (2007) Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* **25**, 125-31.

(21)   Tang, H., Arnold, R. J., Alves, P., Xun, Z., Clemmer, D. E., Novotny, M. V., Reilly, J. P., and Radivojac, P. (2006) A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* **22**, e481-8.

(22)   States, D. J., Omenn, G. S., Blackwell, T. W., Fermin, D., Eng, J., Speicher, D. W., and Hanash, S. M. (2006) Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat Biotechnol* **24**, 333-8.

(23)   Zhang, N., Deutsch, E. W., Farrah, T., Lam, H., Picotti, P., Mendoza, L., Mirzaei, H., Watts, J., and Aebersold, R. (2010) Absolute protein quantification estimated by spectral counting using large datasets in PeptideAtlas. *in preparation*.

(24)   Deutsch, E. W., Lam, H., and Aebersold, R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* **9**, 429-34.

(25)   Sherwood, C. A., Eastham, A., Lee, L. W., Peterson, A., Eng, J. K., Shteynberg, D., Mendoza, L., Deutsch, E. W., Risler, J., Tasman, N., Aebersold, R., Lam, H., and Martin, D. B. (2009) MaRiMba: a software application for spectral library-based MRM transition list assembly. *J Proteome Res* **8**, 4396-405.