

Human Plasma PeptideAtlas

Eric Deutsch¹; Nichole King¹; Jimmy Eng²; Alexey Nesvizhskii³; Olga Vittek¹; Ruedi Aebersold⁴

¹Institute for Systems Biology, Seattle, WA; ²Fred Hutchinson Cancer Research Center, Seattle, WA; ³University of Michigan Medical School, Ann Arbor, MI; ⁴Institute for Molecular Systems Biology, ETH Zurich, Zurich, Switzerland

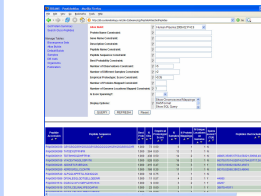


Empirical Proteotypic Scores

It has been frequently noted that when a protein is observed in a sample that is analyzed with LC-MS/MS techniques, some of the protein's component peptides are observed many times, while other component peptides are not observed at all, despite being in the observable mass range and otherwise having attributes appropriate for MS analysis. Several algorithms that attempt to predict observability based on sequence attributes have been put forward. These algorithms are often heavily influenced by the data with which they are trained. Peptides that are often observed and map uniquely to only one protein have been called "proteotypic". We now routinely calculate an empirical proteotypic score, defined below, for all peptides in a PeptideAtlas build. These scores do not rely on prediction algorithms, but merely reflect the frequency with which peptides are observed when the parent protein is observed.

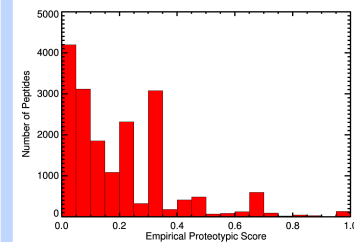
$$\text{Empirical Proteotypic Score (EPS)} = \frac{N_{\text{samples}}(\text{peptide})}{N_{\text{samples}}(\text{parent protein})}$$

We define the Empirical Proteotypic Score (EPS) as the number of samples within which the given peptide is observed divided by the number of samples for which the parent protein was observed. For example, if Protein X is observed in 10 samples within a PeptideAtlas build and its component peptide A is observed in 5 of those, the EPS is 0.5. Note that the number of times a peptide is observed within any given sample is not a factor.



PeptideAtlas web interface query results for proteotypic peptides suitable for targeted followup; selected peptides have elevated EPS values, unique protein mapping, and several observations.

Since shotgun-style experiments of complex samples will sometimes miss some proteins due to the large number of peptides present, a targeted experiment in which only peptides contained within specific proteins of interest are selected by the mass spectrometer will be more successful and time efficient. Using the PeptideAtlas web interface one can select a list of peptides based on the EPS score (defined) below and other attributes as an aid in the design of targeted experiments. For example, one can query the PeptideAtlas for the peptides matching constraints: contained within the desired list of target proteins, having EPS > 0.35, mapping to exactly one protein within the proteome, and having an observation count greater than 5. This list can then be used as an inclusion list of the mass spectrometer.



Histogram of Empirical Proteotypic Scores within the Plasma PeptideAtlas build. While most peptides in the build have a relatively low EPS, indicating that they are only observed rarely, a substantial number of peptides have EPS > 0.35, suggesting that they have greater than a one in three chance of being observed in an additional experiment of similar type as the input experiments, perhaps more for targeted experiments.

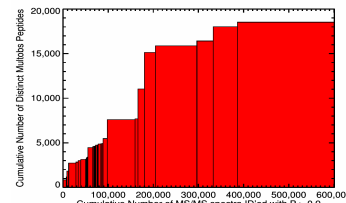
We have excluded peptides whose parent protein is observed in fewer than 3 samples, since the resulting EPS score is less meaningful.

Overview

Starting with nearly 10 million MS/MS spectra from 40 different experiments analyzing human plasma (and serum) samples, contributed from researchers around the world, we have assembled a PeptideAtlas compendium of peptides observed in all these data.

- All datasets searched with SEQUEST and processed through Trans-Proteomic Pipeline in a uniform way
- 600k spectra identified, yielding 18,529 distinct peptide sequences with at least two spectra.
- Estimated a false discovery rate of 1.4% for all 600k spectral identifications, and 2.7% for the list of distinct peptide sequences with at least two spectra
- Over 3000 proteins observed with high confidence
- Proteins identified in these samples exhibit a small but significant enrichment for secreted proteins and a depletion of transmembrane proteins as compared with all Ensembl proteins.
- Resulting distinct peptides and attributes are freely available for download or can be browsed via the PeptideAtlas web interface, including peptide and protein views as well as in tabular formats
- Empirical Proteotypic Score (EPS) is calculated for all peptides to enable selection of peptides that are known to be frequently observable in plasma samples

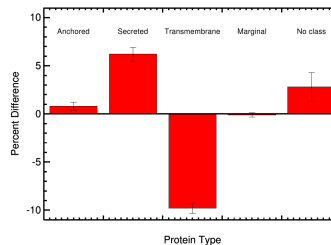
Results



Cumulative number of distinct peptides observed more than once as additional spectrum identifications are added to the atlas. Recently added, very large experiments have greatly increased the number of IDs in the atlas, but the number of new distinct peptides has only increased modestly.

	Mar 2006 Build
# Samples	40
# MS runs	38,000
# MS/MS acquired	9.9 million
# MS/MS ID'ed P>0.9	600,920
# Multobs Peptides P>0.9	18,529
~ # Proteins identified	3,300
% of Genes mapped to	15%

From nearly 10 million spectra, we obtain 600 thousand spectrum IDs, which coalesce into a list of 18,529 peptides seen more than once. We exclude the singletons due to the high error rate in this population. The final error rate in the 18,529 peptides is estimated at 2.7%. These peptides map to approximately 3300 distinct proteins; this number is approximate due to assumptions made when peptides map to multiple proteins.



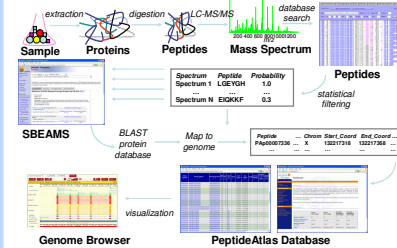
We compare the difference in protein class between the proteins identified in plasma as compared with the whole population of Ensembl proteins. Protein class is derived from the combined results of TMHMM and SignalP.

We find a small but significant enrichment for secreted proteins and a depletion of transmembrane proteins identified in the plasma samples as compared with all Ensembl proteins.

Introduction

PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of LC-MS/MS proteomics experiments and interfaces to access the datasets. Both previously published and unpublished raw experimental data are contributed from researchers around the world. All results of SEQUEST sequence searching have subsequently been processed through PeptideProphet to derive a probability of correct identification in a uniform manner to insure a high quality database.

Pipeline Overview:



- All 40 plasma/serum experiments searched with SEQUEST
- Processed through PeptideProphet to assign probabilities that assignment is correct
- Loaded into SBEAMS database
- All peptides mapped to Ensembl human genome
- All results loaded into PeptideAtlas database
- Available for download, queries, browsing within Ensembl genome browser, etc.

Selected uses for Human Plasma PeptideAtlas:

- Comparison of new peptide/protein identifications with previous work
- Experiment planning using lists of peptides known to be observable in plasma samples
- Contributing to the definition and annotation of the proteome
- Higher throughput searching against libraries of previously identified spectra
- Browse the human genome in Ensembl with tracks showing observed peptide sequences

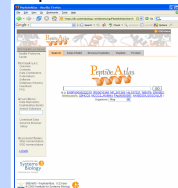
Major data contributions:

- 0.7 million from Seattle Proteome Center
- 2.7 million MS/MS spectra from HUPPO PPP datasets
- 1.8 million from NCI (National Cancer Institute)
- 1.9 million from PNLL (Pacific Northwest National Laboratory)
- 2.7 million from Novartis / GeneProt

Accessing the Data

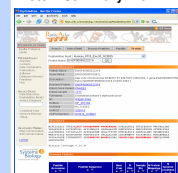
<http://www.peptideatlas.org/>

Simple Search Interface



- Search by protein names, protein descriptions, peptide sequences
- Raw MS output available for published or released experiments

Protein Summary View

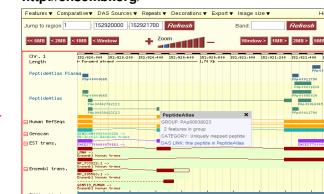


- Protein and peptide views summarizing atlas information
- Browse the human genome at Ensembl with PeptideAtlas tracks

Raw Data Repository



<http://ensembl.org/>



Conclusions

We have created and released a new build of the Human Plasma PeptideAtlas

- Available for download and browsing at <http://www.peptideatlas.org/>
- Now derived from nearly 10 million spectra in 40 experiments
- A useful tool for comparison with new data and planning of new experiments
- Future builds will include more contributed experiments. Greater diversity in experimental protocols and instrumental setups will allow the atlas to continue to expand.
- An additional 1 million spectra will be added soon
- Contribute your human plasma MS/MS experiments to the project to improve the atlas

Most recent general PeptideAtlas paper:

Frank Desiere, Eric W. Deutsch, Nichole L. King, Alexey I. Nesvizhskii, Parag Mallick, Jimmy Eng, Sharon Chen, James Eddes, Sandra N. Loevenich, and Ruedi Aebersold: "The PeptideAtlas Project", *Nucleic Acids Research*, 2006, 34, D655-D658

Previous Human Plasma PeptideAtlas paper derived from 2 million input spectra:

Eric W. Deutsch, Jimmy K. Eng, Hui Zhang, Nichole L. King, Alexey I. Nesvizhskii, Biayang Lin, Hookeun Lee, Eugene C. Yi, Reto Osola, and Ruedi Aebersold: "Human Plasma PeptideAtlas", *Proteomics*, 2005 Aug 5(13):3497-500

Acknowledgements:

This work has been funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28175. We gratefully acknowledge the participating HUPPO PPP labs, H. Zhang (ISB), B. Lin (ISB), T. Corradi (NCI), W. Qian (PNLL), T. Liu (PNLL), M. Pietsch (Novartis) for allowing us to use these data in the Plasma PeptideAtlas