Published in final edited form as:

J Proteome Res. 2015 September 4; 14(9): 3461–3473. doi:10.1021/acs.jproteome.5b00500.

The State of the Human Proteome in 2014/2015 as viewed through PeptideAtlas: enhancing accuracy and coverage through the AtlasProphet

Eric W. Deutsch^{1,*}, Zhi Sun¹, David Campbell¹, Ulrike Kusebauch¹, Caroline S. Chu¹, Luis Mendoza¹, David Shteynberg¹, Gilbert S. Omenn^{1,2}, and Robert L. Moritz¹

¹Institute for Systems Biology, Seattle, WA, USA

²Departments of Computational Medicine & Bioinformatics, Internal Medicine, Human Genetics and School of Public Health, University of Michigan, Ann Arbor, MI, USA

Abstract

The Human PeptideAtlas is a compendium of the highest quality peptide identifications from over 1000 shotgun mass spectrometry proteomics experiments collected from many different labs, all reanalyzed through a uniform processing pipeline. The latest 2015-03 build contains substantially more input data than past releases, is mapped to a recent version of our merged reference proteome, and uses improved informatics processing and the development of the AtlasProphet to provide the highest quality results. Within the set of ~20,000 neXtProt primary entries, 14,070 (70%) are confidently detected in the latest build, 5% are ambiguous, 9% are redundant, leaving the total percentage of proteins for which there are no mapping detections at just 16% (3166), all derived from over 133 million peptide-spectrum matches identifying more than 1 million distinct peptides using AtlasProphet to characterize and classify the protein matches. Improved handling for detection and presentation of single amino-acid variants (SAAVs) reveals the detection of 5,326 uniquely mapping SAAVs across 2,794 proteins. With such a large amount of data, the control of false positives is a challenge. We present the methodology and results for maintaining rigorous quality, along with a discussion of the implications of the remaining sources of errors in the build. We check our uncertainty estimates against a set of olfactory receptor proteins not expected to be present in the set. We show how the use of synthetic reference spectra can provide confirmatory evidence for claims of detection of proteins with weak evidence.

Keywords

shotgun proteomics; tandem mass spectrometry; repositories; PeptideAtlas; Human Proteome Project; observed proteome

*Address correspondence to: Institute for Systems Biology, 401 Terry Ave N, Seattle, WA 98109, USA, eric.deutsch@systemsbiology.org, Phone: 206-732-1200, Fax: 206-732-1299.

The authors have no conflicts of interest to declare.

Supporting Information

Supporting Information Available: This material is available free of charge via the Internet at <http://pubs.acs.org>

Introduction

Shotgun mass spectrometry (MS) proteomics is still the most widely used workflow for identifying and quantifying large numbers of proteins in complex samples. In this workflow, proteins are digested into peptides, which are then separated via liquid chromatography and injected as charged ions into a mass spectrometer¹. The instrument sequentially isolates these precursor ions, fragments them, and collects mass spectra of the fragment ions for each. These fragment ion spectra must then be subjected to complex informatics analysis to reconstruct the peptides that ultimately produced these spectra, and then map the peptides to a reference proteome².

Since the first Human PeptideAtlas build in 2004³, the number of high confidence spectra in the resource has increased over 500-fold, and the number of distinct peptides nearly 40-fold. In previous articles we have described this steady increase in coverage as well as new functionality for exploring the data⁴⁻⁷. As the number of available datasets has grown, we have created sub-builds for specific tissue or sample types. Last year's PeptideAtlas update for the Journal of Proteome Research 2nd C-HPP special issue focused on a comparison between three sample subtypes, kidney, urine, and blood plasma⁷. In addition to human builds, there are PeptideAtlas builds for many other important species, including recent new builds for cow⁸, horse⁹, and *C. albicans*¹⁰. A key focus of the PeptideAtlas is maintaining a well-understood and carefully controlled false discovery rate (FDR) for the identifications contained therein. To achieve this objective, all datasets obtained by PeptideAtlas are re-researched and processed by the components of the Trans-Proteomic Pipeline (TPP)¹¹, a suite of tools for the analysis and validation of shotgun proteomics data.

The Human Proteome Project (HPP) is an international effort to advance our understanding of the human proteome in a multi-pronged approach that includes characterizing its individual components, creating capabilities to assay all those components, and understanding how the system driven by the proteome changes in states of wellness and disease¹². PeptideAtlas has been a key component of the mass spectrometry pillar of the HPP, providing the primary reference for peptides and proteins reliably detected by experiments produced by HPP participants and others throughout the community. It is the primary mass spectrometry data source for neXtProt¹³, which is serving as the primary HPP knowledgebase. An overview of what is contained in neXtProt and its sources has been provided yearly^{13, 14}; the current status is described elsewhere in this issue (Omenn et al., *submitted for this issue*).

Here we present the state of the Human PeptideAtlas for the period 2014/2015. Substantial increases in the numbers of spectra and distinct peptides were achieved, and the increase in the number of confidently-identified proteins was 1044 distinct entries. In the sections below we describe the latest enhancements to the PeptideAtlas build methodology, present the latest results from the August 2014 build and the March 2015 build, and discuss issues surrounding build quality and our analysis of false positives. In the JPR Call for Papers for this 3rd C-HPP special issue, authors were instructed to use PeptideAtlas 2014-08 and neXtProt 2014-09-19, to facilitate comparisons of results.

Methods

The creation of the August 2014 and March 2015 Human PeptideAtlas builds follows the same workflow that has been previously described⁴⁻⁶, with new and additional bioinformatic enhancements that are described below. Briefly, all input experiments are searched with one or more sequence search engines, typically X!Tandem¹⁵ with the k-score plugin¹⁶ and Comet¹⁷ using a comprehensive sequence database that includes all of the UniProt^{18, 19} Complete Proteome set with all single amino-acid variants (SAAVs) in UniProt expanded into sequence snippets appended to each protein, 14 additional IPI²⁰ proteins with uniquely mapping peptides that have passed threshold in previous IPI-based builds and map nowhere else, and the full set of cRAP common contaminant proteins from GPM²¹. The database is appended with a like-sized set of decoy sequences that were generated by keeping all tryptic cleavage sites fixed and shuffling the order of amino acids between the cleavage sites with the addition of a lookup hash used so that every distinct input peptide is always shuffled the same way in order to match redundancy in the target sequences.

The results are then post-processed with PeptideProphet²² for each experiment separately, followed by additional modeling by iProphet²³ in order to combine the results from different search engines, further refine the probabilities, and assign robust probabilities to each PSM and distinct peptide sequence. These tools are designed to maximize the number of true positives and minimize the number of false positives that pass any applied threshold by considering all corroborating evidence for each PSM. The probability threshold is set individually for each dataset to achieve a constant peptide FDR for each dataset, and all PSMs passing the probability threshold are assembled into a master list, including all the decoy hits. This master list is evaluated by MAYU²⁴ to calculate final PSM-level, peptide-level, and protein-level FDRs. The PSM FDR threshold is then set in order to achieve an approximately 1% FDR at the protein level.

All peptides which are included in the build are then mapped to a more comprehensive sequence database that includes all of TrEMBL²⁷ and Ensembl²⁷ as well as all SAAVs catalogued in neXtProt, in addition to the sources mentioned above. This allows easier exploration of the PeptideAtlas results using the Ensembl or TrEMBL accession number domains (in addition to neXtProt and UniProt accessions), and also enables us to map all peptides onto the human genome. Therefore, all peptides (those that map to Ensembl proteins) will have full genome coordinates in the database. In many cases a peptide will span one or more introns and thus two or more sets of coordinates are associated with different parts of the peptide. It is important to note that the August 2014 build is mapped to the GRCh37 genome assembly, while the March 2015 build is mapped to the GRCh38 genome assembly, which means there is a small difference between the coordinates of most peptides in the two builds.

A significant change in the 2015-03 build is how the protein inference is performed. As previously described, in prior builds all the resulting iProphet-written pepXML files for each experiment were analyzed with ProteinProphet²⁸ for the final protein inference step. This had the advantage that the ProteinProphet algorithm is widely used and well regarded as successful in applying parsimony rules to develop the shortest list of proteins that can

explain the peptide evidence. However, it was primarily designed as a generic tool for the processing of individual datasets, and has a few shortcomings in the context of building a consensus proteome from many hundreds of experiments.

Therefore, we have developed a custom inference tool, which we will refer to as AtlasProphet here, and which we used for this release but remains a work in progress at this time. After additional refinement, these new ideas will be released in a new publicly-available version of ProteinProphet. The first innovation is how the AtlasProphet understands a reference proteome (i.e. the full list of proteins that the mass spectra are processed against). ProteinProphet treats all proteins in the reference proteome as equals, which results in cases where, when two or more proteins share several peptides, the one with the most peptides will become the main high probability protein and the others become subsumed. This leads to cases where a primary Swiss-Prot protein is labeled as subsumed to an isoform, TrEMBL, IPI or Ensembl entry, or cases where a Swiss-Prot protein existence (PE) = 1 protein is listed as subsumed to a PE = 5 protein on account of one additional peptide that might be explainable in other ways. In contrast, AtlasProphet is fully aware of a rank of importance of the different reference protein sources in the database as listed in Table 1, and will preferentially categorize proteins with a better rank as canonical, and categorize other proteins in a worse-ranked source relative to the canonical. This scheme is a major revision of the Cedar scheme we published previously⁵, which does already have some elements of Swiss-Prot entries outranking entries from other databases. Briefly, the process of the AtlasProphet is to load the reference proteome, assign rankings to proteins according to Table 1, load all peptides and their protein mappings, group together all proteins that share peptides, and then assign each of the proteins within each group a category based on its relationship to other proteins in the group, taking into account the unicity of all peptides in the group, along with the protein category assignments.

PeptideAtlas uses the PE values as reported by neXtProt and Swiss-Prot, but does not alter them. Briefly, PE = 1 means that the protein entry has been deemed by a set of rules or a curator of one of the two knowledgebases as being reliably observed in its translated form. PE = 2 means that a transcript from the gene has been observed in human samples, and there is every expectation that the protein is translated. PE = 3 means that the protein entry has a homolog in another species where it has been reliably detected. PE = 4 means that credible bioinformatic predictions suggest that the protein is translated. PE = 5 means that curators have deemed that the entry likely corresponds to a pseudogene that is not translated, but weak historical evidence has thus far prevented the entry from being completely deleted. While the majority of PE classifications between neXtProt and UniProtKB are the same, there are some differences because the two knowledgebases have some differing datasets as evidence.

In Table 1, the first column is the priority rank of each source of proteins. These ranks are used for resolving ties for protein categorization as described below. A lower number indicates a better rank. Fractional ranks are used to differentiate ranking that is already present in some of the sources. For example, neXtProt/UniProt already has a PE ranking system, as described above, where a lower number represents a better rank as well. We have accommodated PE ranks within our system by dividing the PE value by 10 and adding the

result to our rank numbers. Therefore a neXtProt 20k entry with a PE=2 will have a rank of 1.2. This means that higher PE values will give a protein entry a worse rank. Note that varsplic (isoform) entries do not have PE values assigned to them. We also note that all Swiss-Prot entries that are specifically excluded from neXtProt (primarily very old entries corresponding to immunoglobulin variable regions) are ranked worse than sequences included in neXtProt, irrespective of PE value. These entries are slated to be replaced in the future. The second column is the name of the source of proteins. The third column indicates the number of proteins from that source that were used as our reference for the 2015-03 build. These numbers will change for future builds as our use of annotations within these sources improves. The final column provides a specific description of each source. It is important to note that while our search database does not include Table 1 ranks 5 and 6 (TrEMBL not Complete Proteome, and Ensembl), we do add them to the mapping database. There are very few detectable peptides in these two sources that are not already contained in the other groups, but we include them by community requests so that all passing peptides can be seen mapping directly to these namespaces. In response to other community requests, we will also begin mapping to the RefSeq source of proteins with the next release, and will assess if mapping directly to GENCODE sequences would also be beneficial.

The term isoform is used to refer to one of several possible protein sequence variations derived from the same gene as listed in neXtProt and Swiss-Prot. These are usually different splice isoforms, but may also be extensions or shortened forms. Single amino acid changes, PTMs, or other post-translational processing such as signal peptide cleavage are not included as isoforms. Previous builds have mapped to all IPI proteins directly. Since the IPI set has now been long deprecated, we no longer map to this set; however, there is a small number of peptides with high probability that map to 14 IPI accessions and nowhere else. These have been retained until these discrepancies can be resolved, either as false positives or *bona fide* annotations that should be included in the reference knowledgebases. For example, IPI01022236 appears to be a splice isoform of P07437, which currently has no varsplic isoform entries, and whose alternate splicing junctions are well supported by multiple peptides. This evidence has been sent to neXtProt for inclusion in future releases. We anticipate that once these discrepancies are resolved, no more IPI entries will remain in future PeptideAtlas builds.

Another innovation in the 2015-03 build is a refinement of the protein categories since previously published by Farrah et al.⁵ A few additional categories are now organized within four groups as shown in Table 2, in order to make their detection status more precise and more understandable. The four major groups are canonical, ambiguous, redundant, and not observed (column 1). Columns 2 lists the new categories as well as the groups into which the categories are sometimes aggregated. The canonical group is the set of proteins that are deemed high confidence detections, although they should not be considered without errors (see discussion of error rates below). The ambiguous group contains proteins of various more specific categories that denote that, while they contain one or more peptides that might be correct evidence of their detection, there are complications (beyond poor PSMs) that indicate that they cannot qualify for canonical yet. The redundant group includes various categories that indicate that a protein has no uniquely mapping peptides, and, therefore,

while the protein may truly have been detected, the evidence peptides map to multiple proteins, and therefore the protein does not belong in a parsimonious list. The table provides a detailed description of the meaning of each protein category within these groups. The difference between identical and indistinguishable categories is that identical proteins have exactly the same sequence and are therefore either reference duplicates or, if originating from different chromosomal loci, are impossible to differentiate based on sequence and would be discarded if not for the desire to view all accessions as entries in the atlas. Indistinguishable proteins cannot be distinguished with the available evidence, but since they do differ in predicted sequence, they could possibly be distinguished with additional evidence; the potential of suitable tryptic peptides for distinguishing purposes is not considered here. In cases where two or more proteins compete for identical rank, the alphanumerically lower accession wins over higher accessions, with the exception that for UniProt-style accessions, those that begin with P win over Q, which wins over all others. For example, following the order P12345 > P34567 > Q12345 > A12345 > B12345 > B34567, if P12345 and P34567 were identical in sequence, P34567 would always be categorized identical, and P12345 some higher category; if they were both different in sequence but indistinguishable, P34567 would be indistinguishable (redundant), and P12345 would be the indistinguishable representative (ambiguous) (or weak or insufficient evidence if appropriate).

In this new scheme, we have introduced crucial, though arbitrary, qualifiers including the number and length of uniquely mapping peptides. Whereas in previous builds even a single 7 AA peptide was sufficient to categorize a protein as canonical, there must now be at least 2 uniquely mapping peptides of length 9. The length limitation has been introduced to overcome the fact that very short peptides can often receive very high scores based on a few spurious peaks plus the fact that short peptides are more likely to be not truly uniquely mapping despite the fact that our imperfect understanding of the human reference proteome (including variants of all types) would imply that they are. PeptideAtlas has for a long time rejected peptides of length 6 and below for this problem. Now that the Human PeptideAtlas has topped 1 million peptides and 100 million PSMs, we have found that peptides of length 7 and 8 are exhibiting this problem as well. PeptideAtlas and neXtProt both share a commitment to ensuring the highest quality mass spectrometric evidence for annotations of proteins, but there is still some difference in the details of these acceptance thresholds between PeptideAtlas canonical and neXtProt PE=1 as discussed in Omenn et al. (*submitted, this issue*).

We also decided it is logical to require at least two distinct peptides before elevating a protein to canonical, since a single hit within such a vast amount of data brings a certain level of suspicion. In future builds we may consider making an exception for proteins where it seems likely that only one or two peptides are amenable for detection based on standard methods, and those exact expected peptides are detected. However, while two distinct peptides are required, these two peptides are allowed to be partially sequence overlapping. While having overlapping peptides generally brings confidence that neither identification is due to random noise, there is some danger that both identifications suffer from the same

homology-based misidentification (see below for further discussion about these types of errors).

While nearly all of the proteins in the PeptideAtlas are classified in an automated manner using the workflow described above, some (but not all) of the more unlikely identifications, such as those of $PE > 1$ and weak or insufficient evidence in the PeptideAtlas classification have been examined manually. In cases where the PSM is determined to be either incorrect or of insufficient quality to be confident in its accuracy, the PSM is marked as rejected. If all PSMs corresponding to a protein are marked as rejected, then the protein itself is automatically put in the “rejected” category. Whenever a spectrum is marked as not being credible evidence for a specific peptide sequence, it remains so labeled for all future builds as well, although additional evidence can alter the protein category. There are, of course, too many spectra for curators to examine all of them, but it is worthwhile and in fact important to examine manually the evidence for single-hit proteins and proteins that are previously annotated in neXtProt as not having definitive protein evidence in order to ensure that the automated statistical scoring is performing as expected, and at the same time discard obviously incorrect or insufficiently convincing PSMs. While PeptideAtlas curators have not yet examined most of the spectra supporting lower-evidence proteins, all spectra are easily accessible to everyone and can be viewed using the Lorikeet spectrum viewer or downloaded for additional examination.

Results

Overall Build Results

The results of the August 2013, August 2014 and March 2015 Human PeptideAtlas builds are summarized in Table 3. The 2014 and 2015 builds are derived from processing over 400 million MS/MS spectra from ~100,000 MS runs, and contain over 100 million high quality PSMs, 1.0 million distinct peptide sequences, and over 14,000 distinct proteins at the target threshold of 1% FDR at the protein level. To achieve this level of stringency, the PSM-level FDR is approximately 0.00009, and the peptide-level FDR is 0.0003. Between the 2014 and 2015 builds, the number of PSMs has increased by over 15 million. Yet, fewer than 4,000 new distinct peptides were added. Of the 1.0 million distinct peptides, only about 300 are expected to be false positives based on our calculated peptide-level FDR. The total number of canonical proteins went down because of the new strategy for classifying proteins as described above. There are three notable differences in the reference proteome: neXtProt has been added as the highest level of the reference; in addition to the UniProt “Complete Proteome” set (UniProtCP), all of the rest of TrEMBL has now been added; nearly all of the IPI database has been removed except for the 14 “orphan” sequences as described above.

The number of proteins claimed depends significantly on the reference proteome used and the criteria used to claim a confident detection. The strategy for doing this within PeptideAtlas has evolved over time. Although there were not a lot of new data added between the August 2014 and March 2015 builds, the method of counting proteins did change significantly, as described in Methods. While Table 3 lists the total number of canonical proteins derived from the merged reference proteome, the numbers in Table 4 are limited to the subset of proteins listed in neXtProt, which is a smaller list (not including,

e.g., immunoglobulins, orphan IPI proteins, etc.). Table 4 lists the number of proteins in the various categories for both the 2014-08 and 2015-03 builds. In contrast, there were a lot of data added between 2013 and 2014 as presented below. The overall number of canonical proteins is lower for the 2015-03 build because proteins with less certain evidence were moved to different categories (in the ambiguous group). The total number of proteins for which there are no detected peptides was reduced by 2%, reducing the total percentage of proteins for which there are no mapping detections at all to just 16% (3166) of the 20,061. These numbers differ somewhat from the HPP “missing proteins” in that the HPP “missing proteins” are a combination of not-observed proteins, some ambiguous proteins, and explicitly do not include the PE=5 proteins (see Omenn et al. this issue).

We compared the total PSM and peptide counts in the latest PeptideAtlas build with those from the latest National Institute of Standards and Technology (NIST) spectral libraries (<http://peptide.nist.gov>) and the summary of GPMDB (http://gpmdb.thegpm.org/statistics_species.html). The NIST libraries have ~208,000 distinct peptides in the ion trap library, and ~433,000 distinct peptides in the union of the high-resolution MS/MS libraries. There are ~69,000 distinct peptides contained in the NIST libraries that are not in PeptideAtlas (~12,000 of these are for peptides of length 6 or below, which are automatically discarded by PeptideAtlas), while ~606,000 peptides are not present in the NIST libraries. See Supplementary Figure 1 for a Venn diagram depicting the overlap between PeptideAtlas and the NIST libraries. The GPMDB ²¹ claims 1.5 billion peptide observations for human samples, which is 10 times larger than the number of PSMs in PeptideAtlas, but the stringency is lower, with low-quality PSMs matching to nearly every protein.

The raw data, either in mzML or raw instrument format, for the experiments used here are available in the PeptideAtlas raw data repository (<http://www.peptideatlas.org/repository/>) along with the search results and TPP tool output. For very large experiments where the raw data are already accessible in another repository, the raw data are not duplicated in the PeptideAtlas raw data repository interface. The results of the build process are available in the build download area (<http://www.peptideatlas.org/builds/>) in several different formats.

Contribution of new datasets

Since the 2013 Human PeptideAtlas build last described⁷, there has been considerable addition of new datasets. The four most notable large tranches of data that have been added come from the CPTAC Consortium²⁹, which extensively analyzed TCGA Project (<http://cancergenome.nih.gov>) tumors in several different facilities, the Pandey Lab³⁰, which analyzed 24 different adult and fetal tissue samples and 6 cell types (publicly available under accession PXD000561), the Küster Lab³¹, which analyzed 36 different adult normal tissues (publicly available under accession PXD000865, although their web site also includes results from numerous other datasets not produced by their lab under that accession), and the Mirzaei Lab³², which analyzed HeLa cells with combinations of seven different enzymes for the digestion step. Figure 1 depicts the impact that these 4 tranches of data have had on the PeptideAtlas between 2013 and 2015. The number of distinct peptide sequences in the Human PeptideAtlas build doubled from less than 0.5 million to ~1 million.

However, the contribution to the number of neXtProt 20k proteins categorized as canonical added was far more modest, increasing from 13,026 to just 14,070. Together these datasets have significantly increased the sequence coverage of proteins already included and added another ~8% proteins. However, this is a rather small increase considering the tremendous amount of data that was added and the much larger numbers of proteins claimed to have been identified in refs 30 and 31 due in large part to the rejection of application of an FDR threshold at the protein level and use of quite lax FDR filters for PSM and peptide levels (see Omenn et al, this issue).

Sample categories and sample-specific builds

In addition to the full proteome builds described above, there are 20 tissue/fluid specific builds, such as for brain, kidney, lung, plasma, and urine. Each of these builds based on human tissue contains only a subset of samples specific for the tissue or biofluid; data derived from cultured cell lines are not included in these sample-specific builds so as not to create artificial tissue proteomes. For these builds, the same process described above is used, but the input dataset list is limited to those of each sample type. There is also an “Other” build that includes all other samples not included in the 20 selected sample types. Each of these builds maintains a threshold so that each has a protein-level FDR of 1%. However, it is very important to note that aggregating all these individual builds will not precisely equal the main build because the PSM-level thresholds set in the build process are quite different. As there are fewer data in each of the individual sample builds, the probability threshold can be lower to achieve the same desired protein-level FDR. These sample-specific builds should be used when one wants to compare a set of proteins observed in one sample type versus another, or if a specific research question is best limited within one sample type. Users should be warned that aggregating several or all of these builds will yield a much higher FDR. If one wishes to limit or categorize entries in the main build by sample type, there is a query option that allows one to filter the result by sample type.

One important change in the most recent Human all-sample PeptideAtlas builds is that they are no longer vastly dominated by normal tissue/biofluid samples and cell line samples. These latest builds now contain a very large number of high-quality spectra from tumor-derived samples from the CPTAC consortium analyzed by the latest generation LTQ Orbitrap Velos and Q Exactive mass spectrometry instruments.

At the PeptideAtlas web site at <http://www.peptideatlas.org/hupo/c-hpp/> is a listing of the number of proteins by chromosome, by PE level, and by category. The tabular data are available for both the 2014-08 build and the 2015-03 build. These results are summarized by AtlasProphet protein category in Figure 2. Ultimately, these categories are intended to answer the question “Is a specific protein of interest detected in PeptideAtlas?” If the questioner wants a black and white answer, then “canonical” or not is the answer. If a more nuanced answer is desirable, then the four groups outlined in Figure 2 may be used as the answer. Approximately 70% of the neXtProt 20,061 protein entries are listed as canonical. Another 5% are categorized as ambiguous for a variety of reasons as discussed above, including 70 indistinguishable representatives. Another 9% have peptides that map to them, but only in a non-unique manner and therefore are not needed to explain any peptides and

are not needed in a parsimonious list of proteins. Finally, there are still 16% of the proteins amongst the neXtProt 20k that have no peptides at all that pass our stringent quality criteria. The most nuanced answer to the question of the presence of a protein in PeptideAtlas is one of the 10 categories in Table 2. Of course, at lower probability thresholds, some additional proteins would have correct identifications, but other protein entries would gather false identifications at an even faster rate.

Support for Single Amino Acid Variants

The existence of SAAV-containing peptides is often overlooked in proteomics data analyses, although it is gaining additional attention in the context of the analysis of proteomics data in conjunction with RNA-seq (see review³³). However, many SAAVs are already known and curated in UniProt and neXtProt, and PeptideAtlas now has partial support for SAAVs contained in those resources. UniProt currently contains over 70,000 human SAAVs, while neXtProt curates over 1.1 million SAAVs from various sources. Thus far, the 1,011 human experiments (for PeptideAtlas, an experiment is a single sample irrespective of possible pre-fractionation, and thus usually includes several MS runs) in PeptideAtlas have been sequence searched against a database containing the 70,000 UniProt SAAVs as described above in the Methods section. Therefore, most of the SAAVs listed in UniProt could have been discovered given sufficiently high quality PSMs. In some cases, a peptide can be discovered via its existence in the search database, and even appear to be uniquely mapping therein; however, when mapped to a much larger space of a great number of known SAAVs, it is sometimes the case that the peptide is no longer uniquely mapping. To guard against these cases, all identified peptides are mapped to an expansion of the full list of neXtProt SAAVs. Using this methodology, we have identified apparent evidence in the data contained in PeptideAtlas for a total of 10,039 distinct SAAV entries for neXtProt SAAVs. Of these, 5,326 sites have uniquely mapping peptides within 2,794 neXtProt proteins. For the rest, there is some ambiguity in the mapping of the peptides; in some cases a peptide can map to one protein with a SAAV and another protein in its reference form. Of note is a subset of ~400 SAAVs for which the curated variant is seen almost exclusively because the reference genome, GRCh 38, itself constructed from a very small number of individuals, is in fact the rare variant. This reveals the underappreciated circumstance that several hundred peptides in widely-observed proteins are not routinely detected in proteomics experiments unless SAAVs are properly handled, because the reference genome and proteome used are based on a rare variant.

In order to enable users to explore the SAAV information in PeptideAtlas, we have added two interface extensions. First, in the protein view page, when there are any known Swiss-Prot or neXtProt variants, a new section is displayed, listing each of these variants. These variants are primarily SAAVs, but also include signal peptides, propeptides, and other variations not explicitly curated as a separate entry (i.e. splice isoform). The SAAVs that are detected are displayed first before those not detected. An example is shown as Figure 3 for the human MMS19 nucleotide excision repair protein homolog (Q96T76). Of its 11 SAAVs listed in Swiss-Prot, four are detected and listed first. The display can be switched to the neXtProt list of SAAVs, which may yield a greater number. The dbSNP identifiers are listed for each entry when available. By clicking on the hyperlink in the second column, one can

see on a separate web page a full listing of all peptides that support the reference sequence or the SAAV sequence.

A queryable and filterable listing of all SAAVs detected thus far in PeptideAtlas is available in the web interface as [Show SAAVs] selection under the [Queries] tab. The user may view and sort all SAAVs or simply a subset of them that match certain criteria. An example screenshot is available as Supplementary Figure 3. Hyperlinks enable further examination of the evidence supporting the original sequence as well as the variations, both as tabular identification information and ultimately as individual annotated spectra. Frequency information for each of the variants may be explored via hyperlinks to the Kaviar database³⁴.

Discussion

PeptideAtlas quality metrics

The stated quality metric of all PeptideAtlas builds for the last five years has been a 1% FDR at the protein level using a target/decoy strategy; the HUPO HPP adopted this same threshold as a guideline from its initiation. To achieve this protein FDR, the PSM- and peptide-level FDR thresholds must be far lower. This has the unfortunate effect of discarding many correct identifications, but this is necessary in order to limit the number of incorrect identifications in the set. Such a tradeoff between false positives and false negatives is well-recognized in statistics. Further, as the size of a data compendium grows, the disparity between the PSM-level FDR and the protein level FDR grows wider. It is crucial to note that the 1% FDR is calculated for the entire build as a whole, not individual experiments. If each experiment were filtered at 1% FDR and then the results combined, nearly every protein in the proteome would have hits to it and the combined result would have a far higher FDR³⁵.

However, while a 1% FDR at the protein level sounds satisfyingly low when stated in that way, with ~15,000 non-redundant proteins in the build, this means there are ~150 incorrectly identified proteins in the build, which may seem less satisfyingly low. This same calculation of multiplying the FDR times the population can show why a 1% peptide-level FDR with 1 million peptides and a 1% PSM-level FDR with 133 million PSMs will yield an unacceptable number of false positives that would easily cover most of the proteome. Nearly every PSM going into the build atlas has a very high probability, and there seems little leverage left in individual PSM probabilities to push for even higher quality. It is very rare to see poor-looking PSMs in the latest human build. However, manual inspection of proteins that are unlikely to be present (never before reported, or “one-hit wonders”, especially with a single spectrum or short peptide) reveals that some PSMs with very high scores and that appear to be outstanding matches may nonetheless still be incorrect. It may be that manual inspection will be more effective in reducing the FDR rather than applying even more stringent probability thresholds at the PSM level.

Recently a new “picked” FDR algorithm was advanced for very large heterogeneous datasets²⁵. It was reported to outperform a “classic” target-decoy FDR strategy, but it is not clear whether it outperforms the MAYU algorithm or the R Factor approach²⁶, which both

provide corrections to the “classic” FDR approach. We have had good experience with the MAYU algorithm and have used it for these builds. However, we will assess whether the “picked” FDR approach yields a different or better result in the next human build.

Olfactory receptors as a quality test

It was recently reported by Ezkurdia et al.³⁶ that olfactory receptors (ORs) may make a useful test set of proteins for evaluating datasets if no tissue samples expected to have such proteins are included. Two recent compendia of MS/MS data^{30, 31} were examined manually by Ezkurdia et al. with the conclusion that no credible PSMs for ORs could be verified³⁶ out of 108 and 200 claimed matches, respectively. These two datasets (reprocessed by our pipeline) are included in the latest PeptideAtlas in addition to many other datasets. None of the sample types entering the PeptideAtlas seems to be likely to harbor ORs. We have therefore examined the OR entries in PeptideAtlas to assess the quality of the build and determine if there are any credible detections of ORs. The output of the pipeline for the 2015-03 build yielded zero OR proteins categorized as canonical, zero categorized as indistinguishable representative, and three categorized as weak from a total of 425 ORs listed as neXtProt entries.

These identifications were examined carefully to see if the PSM evidence seemed motivating for claiming detection of these proteins. I3L273 has a single phosphopeptide PSM of good quality (Supplementary Figure 4), but may well better match an unknown highly homologous peptide and does not seem strong enough evidence to claim detection of this protein. Q9H255 also has just one PSM and the spectrum is certainly not of sufficiently high quality for a convincing match. Q8NGI9 has two apparently excellent PSMs for the same semi-tryptic peptide, but upon very close inspection the spectrum can be better explained by a slightly altered sequence, which may be a variant of lactotransferrin (Supplementary Figures 5a and 5b). In the end all three of these non-redundant detections are not definitive and have been rejected. We therefore conclude that none of the datasets included in PeptideAtlas contains high quality spectra derived from OR proteins. That there were three stated detections (although in the “weak” category) is significantly lower than our stated 1% protein-level FDR. This lends support to the postulate that the quality of a compendium of MS/MS data derived from samples not expected to contain ORs may be evaluated by counting the number of putative OR detections. There are reports that ORs may be functional in non-neural cell types, and transcripts have been claimed for a few ORs in non-neural samples, suggesting that some proteins annotated as ORs may be annotated incorrectly, or the proteins may have other functions in other tissues. In any case, in order to claim that any OR has been detected in proteomic samples, more compelling evidence should be required. The August 2013 PeptideAtlas build did have two OR canonical entries (gene entries), both of which have now been reclassified and removed after similar manual scrutiny.

Largest sources of errors

Mis-identification to a highly homologous peptide ion—due to the stringent threshold used to build the latest atlas, poor PSMs are no longer the greatest source of errors. Rather, two new sources of errors have emerged. The first is PSMs that are of very high

quality and receive an outstanding score, but instead should have been matched to a very similar peptide but with an unconsidered mass modification or sequence variation. An example of this kind discovered while reviewing evidence for PE=5 proteins is the peptide EITALAPSIMK, which appears to be uniquely mapping to and therefore implicate the existence of protein POTEKP, a PE=5 putative beta-actin-like protein. A PSM for this peptide (Supplementary Figure 6) is outstanding, with only one y2 ion missed. However, careful inspection of the spectrum reveals that it is of such high quality that there is no reason to justify the missing y2, and there is a nearby unexplained peak. If one maps the peptide EITALAPSTMK, where the K has been dimethylated, one finds a perfect match with no missed peaks (Supplementary Figure 7). This dimethylation is likely to derive from sample handling, rather than inherent dimethylation of the intact protein in the original specimen. The peptide EITALAPSTMK maps directly to actin itself, and that peptide has been observed 63,000 times in PeptideAtlas without the dimethylation. This is a classic case of an almost-correct identification entering the analysis result because the completely correct answer was not in the search space. Although we describe this single example in a recent tumor tissue analysis set, further evaluation has identified ~1% of total peptide identifications were methylated to various levels raising important issues regarding confidence not only in identification of peptides but also in quantitation, especially for low abundance proteins which are being identified as candidate biomarkers.

Mis-identification of a protein despite a correct peptide identification—the second largest case of error now is where the PSM is correct, but is mapped to the incorrect protein, usually on account of an unconsidered SAAV or other variant. Peptides of length 6 AA or lower have always been discarded in PeptideAtlas because of this problem. Even if correctly identified, there is too much opportunity for the peptide to map to a different protein with unconsidered variants, including I/L substitutions as well as Q/K and N/D substitutions, which can be difficult to distinguish at low resolution or lax tolerances. Now that there are over 1 million distinct peptides, including over 45,000 distinct peptides of length 7, we find that most of these peptides map to several proteins, often via known SAAVs that we have now considered. Many of the ones that still do appear to map uniquely to an otherwise undetected protein are likely to map instead to a more common protein via an unconsidered variant. Investigators performing searches for “missing proteins” should specifically consider such SAAV explanations before claiming the missing proteins to have been detected.

An important issue that bears further investigation is the possibility that immunoglobulins, with their huge potential variability, may be an explanation of such single-hit proteins. Although there are 126 immunoglobulin variable region entries that are included in PeptideAtlas and our search databases (although specifically excluded from neXtProt), these Swiss-Prot entries are very old and do not capture the full variability potential present in immunoglobulins. Therefore, there may be a significant number of entries in PeptideAtlas where the peptide is surely correctly identified, but the peptide is not uniquely mapping for the protein as implied by the current reference proteome, because it can also be found in an immunoglobulin. This effect is almost surely present, but its scale is currently not known. We have begun working with the IMGT database ³⁷ to expand our reference list of

immunoglobulins, and hope to have a more complete set of immunoglobulins in future PeptideAtlas releases.

Confirmation of spectra using synthetic peptides

For proteins with few supporting peptides, a positive match against the spectrum of a corresponding synthetic peptide provides substantially increased confidence in the available evidence. SRMATlas^{38, 39} is the largest compendium of spectra for synthetic peptides for several species including human. Comparison of spectra in the Human PeptideAtlas vs. spectra from synthetic human tryptic peptides contained in the SRMATlas provides solid evidence to peptide identities, although care must be taken to ensure that compared spectra use similar fragmentation mechanisms.

As an example, protein Q9UJA2 Cardiolipin synthase is a PE=2 predicted protein with only 2 distinct peptides that map to it in the 2015 PeptideAtlas. One (AAAFYVR) is only 7 amino acids long, is semi-tryptic in this protein, and maps to another much better observed protein in a fully tryptic manner; thus there is no good reason to use this as evidence for the detection of Q9UJA2. The other peptide (YFNPCYATAR) maps uniquely to this protein and its one known isoform. A single uniquely mapping peptide is normally not very definitive evidence for a protein. However, a comparison between the HCD spectrum of the peptide in PeptideAtlas and a QTOF spectrum of the corresponding human synthetic peptide reveals a nearly perfect match (Figure 4), lending great confidence that the peptide is correctly identified. However, there is still a possibility that the peptide derives from a different protein (such as an immunoglobulin as discussed above) in a manner that is not currently understood. An examination of the protein sequence reveals all other potential tryptic peptides fall into regions of the protein not likely to be observed (such as the signal sequence or transmembrane regions) or are too long, too short, or too extreme in hydrophobicity. The peptide YFNPCYATAR is the only one that has no reasons not to be seen, thereby lending significant confidence that this single-peptide protein identification is genuine.

Conclusion

We have presented the current state of the PeptideAtlas in August 2014 and March 2015, which represents a substantial improvement over previous builds. It includes several additional large-scale publicly accessible datasets, which double the total number of distinct peptides in the Human PeptideAtlas, and increase the number of canonical neXtProt proteins by 1044 since 2013 (Figure 1). In addition to including more data, we have enhanced the scheme by which we categorize proteins based on the available evidence.

There still remain over 3000 proteins for which there is no peptide evidence that passes our threshold in PeptideAtlas, including all of the 425 olfactory receptor proteins and over 500 PE=5 entries, which are thought to be pseudogenes or other non-transcribed and non-translated genes. Of the rest, some may also not be translated, some are not suitable for detection with bottom-up proteomics using conventional digestion enzymes, and still others may be of such low abundance or only abundant under highly unusual conditions that detection is extremely difficult¹⁴.

PeptideAtlas will continue to expand and improve in the coming years as increases in depth of mass spectrometry experiments become available. More datasets will be added as they are made available in ProteomeXchange⁴⁰ repositories for shotgun data such as PRIDE and MassIVE.

Everyone in the community is encouraged to support this effort by submitting their raw spectra, metadata, and experimental results to ProteomeXchange. New and old datasets will be re-searched with an improved search strategy that considers more SAAVs, PTMs, splice variants, and predicted proteins in the search space. There will be more extensive manual curation to better understand the sources of errors and reduce the FDR even further. There will also be improved support for datasets that have been digested with an enzyme other than trypsin, which until very recently has been rare. With these enhancements PeptideAtlas will continue to advance as it has for the past 12 years as the most comprehensive compendium of the highest quality identifications produced from the shotgun proteomics experiments made public by the research community.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was funded in part by the American Recovery and Reinvestment Act (ARRA) funds through National Institutes of Health from the NHGRI grant RC2HG005805, the NIGMS grants R01GM087221 and 2P50GM076547 to the Center for Systems Biology, the National Institute of Biomedical Imaging and Bioengineering grant U54EB020406, the National Science Foundation MRI grant 0923536, the NIEHS grant U54ES017885 to the University of Michigan, and the EU FP7 'ProteomeXchange' grant 260558.

References

1. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003; 422:198–207. [PubMed: 12634793]
2. Deutsch EW, Lam H, Aebersold R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol Genomics*. 2008; 33(1):18–25. [PubMed: 18212004]
3. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, Fausto N, Hafen E, Hood L, Katze MG, Kennedy KA, Kregenow F, Lee H, Lin B, Martin D, Ranish JA, Rawlings DJ, Samelson LE, Shiio Y, Watts JD, Wollscheid B, Wright ME, Yan W, Yang L, Yi EC, Zhang H, Aebersold R. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol*. 2004; 6(1):R9. [PubMed: 15642101]
4. Deutsch EW, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep*. 2008; 9(5):429–34. [PubMed: 18451766]
5. Farrah T, Deutsch EW, Omenn GS, Campbell DS, Sun Z, Bletz JA, Mallick P, Katz JE, Malmstrom J, Ossola R, Watts JD, Lin B, Zhang H, Moritz RL, Aebersold R. A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol Cell Proteomics*. 2011; 10(9) M110 006353.
6. Farrah T, Deutsch EW, Hoopmann MR, Hallows JL, Sun Z, Huang CY, Moritz RL. The state of the human proteome in 2012 as viewed through PeptideAtlas. *J Proteome Res*. 2013; 12(1):162–71. [PubMed: 23215161]
7. Farrah T, Deutsch EW, Omenn GS, Sun Z, Watts JD, Yamamoto T, Shteynberg D, Harris MM, Moritz RL. State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the

- kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. *J Proteome Res.* 2014; 13(1):60–75. [PubMed: 24261998]
8. Bislev SL, Deutsch EW, Sun Z, Farrah T, Aebersold R, Moritz RL, Bendixen E, Codrea MC. A Bovine PeptideAtlas of milk and mammary gland proteomes. *Proteomics.* 2012; 12(18):2895–9. [PubMed: 22837157]
9. Bundgaard L, Jacobsen S, Sorensen MA, Sun Z, Deutsch EW, Moritz RL, Bendixen E. The Equine PeptideAtlas: a resource for developing proteomics-based veterinary research. *Proteomics.* 2014; 14(6):763–73. [PubMed: 24436130]
10. Vialas V, Sun Z, Loureiro y Penha CV, Carrascal M, Abian J, Monteoliva L, Deutsch EW, Aebersold R, Moritz RL, Gil C. A *Candida albicans* PeptideAtlas. *J Proteomics.* 2014; 97:62–8. [PubMed: 23811049]
11. Keller A, Eng J, Zhang N, Li XJ, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol.* 2005; 1 2005 0017.
12. Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, Beretta L, Bergeron J, Borchers CH, Corthals GL, Costello CE, Deutsch EW, Domon B, Hancock W, He F, Hochstrasser D, Marko-Varga G, Salekdeh GH, Sechi S, Snyder M, Srivastava S, Uhlen M, Wu CH, Yamamoto T, Paik YK, Omenn GS. The human proteome project: current state and future direction. *Mol Cell Proteomics.* 2011; 10(7) M111 009993.
13. Gaudet P, Argoud-Puy G, Cusin I, Duek P, Evalet O, Gateau A, Gleizes A, Pereira M, Zahn-Zabal M, Zwahlen C, Bairoch A, Lane L. neXtProt: organizing protein knowledge in the context of human proteome projects. *J Proteome Res.* 2013; 12(1):293–8. [PubMed: 23205526]
14. Lane L, Bairoch A, Beavis RC, Deutsch EW, Gaudet P, Lundberg E, Omenn GS. Metrics for the Human Proteome Project 2013–2014 and strategies for finding missing proteins. *J Proteome Res.* 2014; 13(1):15–20. [PubMed: 24364385]
15. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics.* 2004; 20(9):1466–7. [PubMed: 14976030]
16. MacLean B, Eng JK, Beavis RC, McIntosh M. General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics.* 2006; 22(22): 2830–2. [PubMed: 16877754]
17. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. *Proteomics.* 2013; 13(1):22–4. [PubMed: 23148064]
18. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 2004; 32:D115–9. (Database issue). [PubMed: 14681372]
19. UniProt C. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2014; 42(Database issue):D191–8. [PubMed: 24253303]
20. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics.* 2004; 4(7):1985–8. [PubMed: 15221759]
21. Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res.* 2004; 3(6):1234–42. [PubMed: 15595733]
22. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem.* 2002; 74:5383–5392. [PubMed: 12403597]
23. Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, Mendoza L, Moritz RL, Aebersold R, Nesvizhskii AI. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics.* 2011; 10(12) M111 007690.
24. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, Hengartner MO, Aebersold R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics.* 2009; 8(11):2405–17. [PubMed: 19608599]

25. Savitski MM, M WI, Hahne H, Kuster B, Bantscheff M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol Cell Proteomics*. 2015
26. Shanmugam AK, Yocum AK, Nesvizhskii AI. Utility of RNA-seq and GPMDB protein observation frequency for improving the sensitivity of protein identification by tandem MS. *J Proteome Res*. 2014; 13(9):4113–9. [PubMed: 25026199]
27. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. 2003; 31(1):365–70. [PubMed: 12520024]
28. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*. 2003; 75:4646–4658. [PubMed: 14632076]
29. Ellis MJ, Gillette M, Carr SA, Paulovich AG, Smith RD, Rodland KK, Townsend RR, Kinsinger C, Mesri M, Rodriguez H, Liebler DC, Clinical Proteomic Tumor Analysis, C. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer discovery*. 2013; 3(10):1108–12. [PubMed: 24124232]
30. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabudde NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TS, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H, Pandey A. A draft map of the human proteome. *Nature*. 2014; 509(7502):575–81. [PubMed: 24870542]
31. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmaier A, Faerber F, Kuster B. Mass-spectrometry-based draft of the human proteome. *Nature*. 2014; 509(7502):582–7. [PubMed: 24870543]
32. Guo X, Trudgian DC, Lemoff A, Yadavalli S, Mirzaei H. Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics. *Mol Cell Proteomics*. 2014; 13(6):1573–84. [PubMed: 24696503]
33. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat Methods*. 2014; 11(11):1114–25. [PubMed: 25357241]
34. Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics*. 2011; 27(22):3216–7. [PubMed: 21965822]
35. Deutsch EW, Mendoza L, Shteynberg D, Slagel J, Sun Z, Moritz RL. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clinical Applications*. 2015 accepted.
36. Ezkurdia I, Vazquez J, Valencia A, Tress M. Analyzing the First Drafts of the Human Proteome. *J Proteome Res*. 2014; 13(8):3854–3855. [PubMed: 25014353]
37. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, Regnier L, Ehrenmann F, Lefranc G, Duroux P. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res*. 2009; 37:D1006–12. (Database issue). [PubMed: 18978023]
38. Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R. Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell*. 2009; 138(4):795–806. [PubMed: 19664813]
39. Kusebauch U, Deutsch EW, Campbell DS, Sun Z, Farrah T, Moritz RL. Using PeptideAtlas, SRMatlas, and PASSEL: Comprehensive Resources for Discovery and Targeted Proteomics. *Current protocols in bioinformatics / editorial board, Andreas D. Baxevanis ... [et al.]*. 2014; 46:13 25 1–13 25 28.

40. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dienes JA, Sun Z, Farrah T, Bandeira N, Binz PA, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus HJ, Albar JP, Martinez-Bartolome S, Apweiler R, Omenn GS, Martens L, Jones AR, Hermjakob H. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol.* 2014; 32(3):223–6. [PubMed: 24727771]

Figure 1.

Summary of the increase in content of the Human PeptideAtlas with the addition of five very large tranches of data (plus various smaller datasets) from 2013 to 2015. The left panel shows that the number of distinct peptide sequences doubled from less than 0.5 million to ~1 million. The right panel shows that the number of distinct canonical proteins (neXtProt 20k baseline) only increased by 8%. The 13,026 has been corrected from the previously published value of 13,377 to compensate for the use of neXtProt as the reference for this chart instead of the more comprehensive reference used in 2013. The 2014 numbers are different than reported at the HUPO 2014 Congress on account of applying the AtlasProphet algorithm to obtain more accurate counts for the 2014-08 build.

Figure 2.

Summary of the mapping of the Human 2015-03 PeptideAtlas peptides to the neXtProt core 20,061 proteins. The proteins are apportioned in 10 categories with 4 groups. The relative sizes of the groups are depicted in the pie chart.

Figure 3. Example of SAAV information in PeptideAtlas protein display for MMS19 (Q96T76). At the top is a partial listing from Swiss-Prot of PeptideAtlas-detected SAAVs and their detected frequencies. At the bottom is a sequence alignment view showing the reference sequence as well as three of the aligned, detected SAAVs. On the left is the region 50-100, showing two SAAVs, one (at position 68) observed exclusively (i.e. no peptides match the unmodified version), and one (at position 98) that has never been observed. On the right is the region 500 – 570, showing the alignment of two SAAV peptides where both the reference form and the changed forms have been observed, although the changed forms only infrequently. By looking in Kaviar³⁴ at the exclusively seen SAAV corresponding to dbSNP entry rs2275586, one finds that, at position chr10:97481001, the overall frequency of the reference sequence is only ~4%, while ~96% of genotypes analyzed for inclusion in the Kaviar database version 2.0 have the A → G forms (Supplementary Figure 2).

Figure 4.

Comparison of fragmentation spectra for 2+ ions of peptide YFNP CYATAR in protein Q9UJA2. Top: high resolution HCD spectrum using an Orbitrap Velos from a fetal testis sample³⁰ in PeptideAtlas. Bottom: high resolution CID spectrum from a synthetic peptide using a 6530 QTOF. Despite being from different instruments and sources, the spectra are very similar, lending great confidence that the peptide has been correctly identified. Most of the unlabeled low-mass ions are immonium ions and other annotatable fragments.

Table 1

Rank order of protein reference sources considered by the AtlasProphet algorithm when assigning protein categories in PeptideAtlas beginning with the 2015-03 build. See text for a detailed explanation and discussion of the table.

Rank	Protein reference source	2015 Count	Description
1.1 - 1.5	neXtProt 20k	20,061	Canonical proteins listed in neXtProt as the Swiss-Prot primary displayed isoforms. These make up the primary HPP target list
2.1 - 2.5	Swiss-Prot not neXtProt	129	Additional Swiss-Prot entries that are excluded from neXtProt, e.g. immunoglobulin variable region sequences
3	neXtProt Varsplic (isoforms)	21,841	Additional neXtProt and extra Swiss-Prot curated splice isoforms of the canonicals
4.1 - 4.5	UniProt Complete Proteome (CP)	47,672	UniProt TrEMBL entries that are categorized as being part of the "Complete Proteome" since they contain an entry in Ensembl, and are not in previous sources
5.1 - 5.5	TrEMBL not CP	74,269	Additional TrEMBL entries not part of the "Complete Proteome" group
6	Ensembl	99,436	All human Ensembl entries, which all can be mapped exactly to the human genome
7	IPI	14	A few remaining IPI entries retained because they appear to have valid peptides that do not map elsewhere
8	Contaminant	115	All non-human contaminants from the GPM cRAP database
9	Decoy	89,117	All decoy entries from the search database, which does not include ranks 5,6

Table 2

Summary of the new protein categories used for the 2015-03 build. See text for further discussion of the table.

Group	Category	Description
canonical	canonical	The top category for a protein that has 2 distinct peptides of length 9 that are uniquely mapping within the set of proteins of the same or better ranked protein reference source (Table 3).
ambiguous	indistinguishable representative	A protein that is part of a set of proteins within the same reference source that together share all their peptides within the set. This implies that at least one protein within the group has been definitively detected, but it is unclear which one. One or more are selected to be the representative(s) of the group from which the true detection cannot be discerned. If a protein that would be an indistinguishable representative qualifies for weak, insufficient evidence, or marginally distinguished, those categories take precedence over an indistinguishable representative.
	marginally distinguished	A protein that shares most of its peptides with a canonical protein of higher rank but also seems to have some independent evidence in the form of a few uniquely-mapping peptides within its source or higher as well. It may be an isoform of the canonical or a separate protein.
	weak	A protein with only one uniquely-mapping peptide with length of at least 9.
	insufficient evidence	A protein whose only uniquely mapping peptides are of length 7 or 8. There is too high a chance that the peptide, even if identified correctly, maps to another region of the proteome that is not represented in our reference.
redundant	indistinguishable	A protein that shares all the same peptides with one or more proteins in a higher or equal category and is therefore not needed to explain the peptides in a parsimonious list.
	subsumed	A protein that shares all of its peptides with one or more proteins in a higher or equal category that have additional peptides. This protein is not needed to explain peptides in a parsimonious list.
	identical	A protein whose sequence is identical to another protein in the same or a higher reference source. In a set of identical sequence protein entries, all but one are labeled as identical.
not observed	not detected	A protein that has no peptides above threshold mapping to it.
	rejected	A protein that has some peptides and PSMs above threshold mapping to it, but all have been rejected by human curators as being false identifications, or of insufficient quality to support the identification of a protein.

Table 3

Comparison of basic statistics for the 2013, 2014, and 2015 builds

	Human 2013-08	Human 2014-08	Human 2015-03
Reference Proteome			neXtProt 2014-12
	UniProtCP 2012-10	UniProtCP 2014-06	UniProtCP 2014-12
			UniProt all TrEMBL 2014-12
	Ensembl 67.37	Ensembl 75.37	Ensembl 78
	IPI 3.71	IPI 3.87	IPI orphan 14
	cRAP 2012	cRAP 2014	cRAP 2015
Reference Genome	GRCh 37	GRCh 37	GRCh 38
Number of Samples	515	971	1011
Number of MS runs	85,091	99,567	103,387
Input Spectra searched	252M	414M	470M
PSMs above threshold	61M	114M	133M
Distinct Peptides	338,013	1,021,823	1,025,698
Canonical Proteins	12,644	15,195	14,629

Table 4

Comparison of PeptideAtlas protein categories for the neXtProt 20k proteins only. Note that the number of canonical proteins is different than in Table 3 because the reference is different.

Group	Category	Human 2014-08	Human 2015-03
neXtProt 20k		20126	20061
canonical	canonical	14946	14070
ambiguous	indistinguishable representative	-	70
	marginally distinguished	612	471
	weak	-	538
	insufficient evidence	-	29
redundant	indistinguishable	189	74
	subsumed	683	1718
	NTT-subsumed	448	-
	identical	14	23
not observed	not detected	3226	3166
	rejected	0	5