Journal of proteome research

Article

pubs.acs.org/jpr

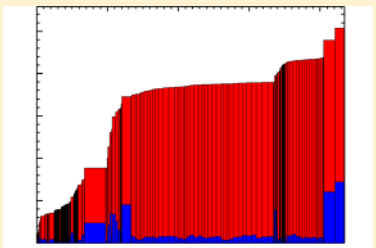
# The State of the Human Proteome in 2012 as Viewed through PeptideAtlas

Terry Farrah,* Eric W. Deutsch, Michael R. Hoopmann, Janice L. Hallows, Zhi Sun, Chung-Ying Huang, and Robert L. Moritz*

Institute for Systems Biology, 401 Terry Avenue North, Seattle, Washington 98109, United States

Supporting Information

**ABSTRACT:** The Human Proteome Project was launched in September 2010 with the goal of characterizing at least one protein product from each protein-coding gene. Here we assess how much of the proteome has been detected to date via tandem mass spectrometry by analyzing PeptideAtlas, a compendium of human derived LC−MS/MS proteomics data from many laboratories around the world. All data sets are processed with a consistent set of parameters using the Trans-Proteomic Pipeline and subjected to a 1% protein FDR filter before inclusion in PeptideAtlas. Therefore, PeptideAtlas contains only high confidence protein identifications. To increase proteome coverage, we explored new comprehensive public data sources for data likely to add new proteins to the Human PeptideAtlas. We then folded these data into a Human PeptideAtlas 2012 build and mapped it to Swiss-Prot, a protein sequence database curated to contain one entry per human protein coding gene. We find that this latest PeptideAtlas build includes at least one peptide for each of ∼12500 Swiss-Prot entries, leaving ∼7500 gene products yet to be confidently cataloged. We characterize these "PA-unseen" proteins in terms of tissue localization, transcript abundance, and Gene Ontology enrichment, and propose reasons for their absence from PeptideAtlas and strategies for detecting them in the future.



Growth of the Human PeptideAtlas: 470 datasets added from 2002 to present

**KEYWORDS:** *Human Proteome Project, PeptideAtlas, LC−MS/MS, database, protein inference*

## INTRODUCTION

A key goal in the field of proteomics is to characterize the entire complement of protein forms for a given species. This goal includes describing not only the most common protein form produced by each gene, but also its variants, including splice variants, post-translational modifications, products of enzymic processing, and different forms resulting from genetic variation. Equally important are the efforts to characterize which proteins are abundant in the various tissues and cell types, which are abundant during certain developmental stages, and which are abundant under various biological or environmental conditions.

The Chromosome-centric Human Proteome Project (C-HPP),[1] launched in 2010,[2] has taken as a manageable first step to characterize a single protein form corresponding to each currently identified protein-coding gene. This international effort has divided this ambitious goal among over two dozen countries comprised of multiple groups that will work collectively on identifying all of the proteins from their respective individual chosen chromosome. A variety of survey and targeted approaches will be used by the C-HPP groups, but all participants will find it useful to understand the baseline proteome coverage from extant MS data sets.

There are currently two primary proteomics methods: the identification of peptides or proteins via mass spectrometry, primarily LC−MS/MS, and the identification of proteins via immunoactivity such as immunohistochemistry using protein specific antibodies. In order to assess our progress toward identifying one protein form per coding gene, we must find out which proteins have been identified using each of these techniques.

The Human PeptideAtlas is uniquely suited to answer this question for LC−MS/MS. The broader PeptideAtlas project was begun nearly a decade ago to accurately reinterpret LC−MS/MS data from diverse sources and then map the resulting peptide identifications to genomes, in particular to the human genome.[3] In our laboratory's initial PeptideAtlas publication we identified peptides mapping to 27% of the human genes in Ensembl,[4] albeit with a gene/protein false discovery rate (FDR) likely 10% or greater. Over the years we have continually added publicly available human data to PeptideAtlas, and learned how to better control false identifications. In early 2012, the 1% protein-level FDR Human PeptideAtlas surpassed 50% genome coverage with the addition of several high-protein-coverage cell line data sets.[5−9]

The Human PeptideAtlas is a compendium of many different experiments—470 to date. Importantly, PeptideAtlas reinterprets all data using a uniform computational pipeline, the Trans-Proteomic Pipeline,[10] to a stringent protein FDR. PRIDE[11,12] is also a comprehensive proteomics data repository, but, rather than reprocess the raw data, it presents the peptide and protein identifications submitted by each investigator

**Special Issue:** Chromosome-centric Human Proteome Project

**Received:** October 25, 2012
**Published:** December 5, 2012

 © 2012 American Chemical Society  dx.doi.org/10.1021/pr301012j | J. Proteome Res. 2013, 12, 162−171

without further validation. Because protein inference methods vary considerably in the FDRs of their final protein lists, the FDR of all the protein identifications in PRIDE combined cannot be easily assessed. Further, when protein identifications from multiple sources are combined, the resulting FDR is always higher than the average FDR of the contributing experiments. This elevated FDR is due to the fact that, while the sets of true positives (correct identifications) of the various experiments tend to overlap, the false identifications tend to be randomly distributed.[13]

The Global Proteome Machine Database[14,15] (GPMDB), like PeptideAtlas, reprocesses raw data via a uniform pipeline. Further, it implements an automated process to scour the literature for publicly available data, and thus contains many more data sets than PeptideAtlas. The GPMDB Guide to the Human Proteome lists, by chromosome, the number of peptide observations for each Ensembl identifier along with the lowest (best) expectation value for that protein. One can apply any criteria one wishes to filter these lists and yield a final list of identified proteins. However, it is difficult to calculate the FDR for such a list. The expectation values are calculated with respect to each individual experiment. They have little meaning when calculating an FDR for results from multiple experiments combined. Indeed, PeptideAtlas is the only large compendium of shotgun proteomics results with a well-defined and stringent protein FDR for the entire compendium combined.

In the current work, we first increase the proteome coverage of PeptideAtlas by incorporating additional publicly available data sets, emphasizing experiments that identify classes of proteins missing from PeptideAtlas. The resulting Human PeptideAtlas build covers 62.4% of the human proteome as defined by Swiss-Prot with a protein-level FDR of 1%. We then analyze the remaining 37.6% to understand why these proteins are not yet included and to suggest strategies for observing them in the future efforts of the C-HPP.

## EXPERIMENTAL PROCEDURES

We judiciously added publicly available data to PeptideAtlas for the specific purpose of increasing the number of protein identifications. Our aim was to obtain a large number of new identifications by adding a moderate amount of data. First we added two large plasma data sets and one large cell line data set that had recently been contributed. We then looked for promising data in the GPMDB using two strategies: (a) reviewing Data sets of the Week, which tend to be high quality data sets, and selecting those which were very high quality, used new MS technology, had low-complexity samples due to a filtering method, or used cell types or tissue types not yet in PeptideAtlas, and (b) using an automated process to select GPMDB data sets containing many higher-confidence identifications for proteins not yet in PeptideAtlas. We also considered all articles published in *Molecular and Cellular Proteomics* that referenced the Tranche data repository[16] in the main text, and selected from these data sets using the same criteria we used for GPM Data sets of the Week.

We selected a total of 27 data sets and were able to obtain 17 in full or almost in full (four from the authors, two from PRIDE, and 11 from Tranche) and four in large part (from Tranche). The remaining six data sets had been deposited in Tranche but could not be retrieved after multiple attempts, emphasizing the need for a stably funded publicly accessible repository for raw mass spectrometry data. One of the 17 full data sets was available only in Scaffold (Proteome Software) format and was not usable in our pipeline. Of the 20 full or partially downloaded data sets, 17 could be processed fully or partially using X!Tandem[17] + K-score.[18] These, along with the two large plasma data sets and the large cell line data set, were added to the Human PeptideAtlas. All 20 are listed in Table S1, Supporting Information.

Among the added data sets were several that were expected to provide coverage of protein categories shown to be underrepresented in PeptideAtlas by Gene Ontology analysis (data not shown), including samples of vitreous humor to increase coverage of proteins of sensory perception, seminal plasma to increase coverage of proteins of the reproductive system, a data set identifying new integral membrane proteins, and experiments targeting signaling proteins. Other data sets were selected to cover additional sample types not previously included in PeptideAtlas (e.g., mitotic spindle, nucleosome, and colorectal tissue).

These data sets, along with all the data sets we had included in the previous build, were processed through the latest PeptideAtlas build pipeline[19] to produce a final protein set with an FDR close to 1%. Briefly, all data sets were searched against a target-decoy sequence database consisting of the International Protein Index database[20] (IPI) and cRAP common contaminants (www.thegpm.org/crap), plus one decoy sequence for each target entry. Results were processed using the Trans-Proteomic Pipeline.[10] Identified peptides were mapped to a protein sequence database that included IPI v3.71,[20] Ensembl v67.37,[11] and the 2012_05 release of Swiss-Prot,[21,22] including splice variants and representing 20244 protein-coding genes. A PSM (peptide-spectrum match) FDR threshold of 0.0002 was applied to each data set to yield a list of 218 799 distinct identified peptides and a protein-level FDR of 0.8% as computed by Mayu.[13] See Table 1 for comparison with previous build.

**Table 1. Human PeptideAtlas, Before and After Recent Addition of Publicly Available Data**

| | Human PeptideAtlas June 2012 | Human PeptideAtlas October 2012 | percent increase |
|---|---|---|---|
| Experiments | 353 | 470 | 33% |
| PSMs (peptide-spectrum matches) | 14273527 | 43428145 | 204% |
| Distinct peptides | 189620 | 253690 | 34% |
| Swiss-Prot entries with at least one identified peptide | 11524 | 12629 | 9.6% |
| Swiss-Prot coverage | 56.9% | 62.4% | |

Over 62% (12629) of the Swiss-Prot entries were found to contain at least one identified peptide in either its canonical form or one of its variant forms. (Thirty-six entries identified only by semitryptic or nontryptic peptides are not included in this tally.) These entries formed the list referred to herein as *PA-seen* and the remaining 7614 entries formed the list *PA-unseen*. Note that in some cases two or more proteins in the PA-seen list have identical or overlapping sets of identified peptides. The PA-seen list is not intended to be a parsimonious (minimal-redundancy) protein list but to contain all Swiss-Prot entries with any peptide evidence in this atlas build.

About 1% (2397) of the distinct peptides mapped only to a sequence in either IPI or Ensembl and not to any Swiss-Prot


dx.doi.org/10.1021/pr301012j | J. Proteome Res. 2013, 12, 162−171

sequence. A parsimonious mapping of these peptides covers a total of 1291 IPI or Ensembl identifiers.

## RESULTS AND DISCUSSION

By way of our effort to fill the protein gaps in the Human PeptideAtlas, we supplemented our proteomics efforts with publicly available data sets, as described in the Experimental Procedures. Table 1 shows the current Human PeptideAtlas build details and compares it to the previous build. The Human PeptideAtlas now contains 470 experiments, a marked 32% increase from a year ago. The distribution of experiments across different tissues, cell types, and body fluids is provided in Table 2. This addition of experiments doubles the number of peptide spectrum matches, increases by 33% the number of distinct peptide identifications, and increases the coverage of Swiss-Prot entries from 56.9 to 62.4%.

Because the human section of Swiss-Prot is intended to represent the scientific community's current best estimate of the complete set of human protein coding genes, we consider only Swiss-Prot protein sequences in the current analysis. Henceforth, we will refer to the covered Swiss-Prot entries as *PA-seen* and those Swiss-Prot entries unobserved in PeptideAtlas as *PA-unseen*.

Many of the PA-unseen proteins have been reported in the results of various shotgun proteomics experiments. In fact, most have been reported in the PRIDE database. We calculated the Swiss-Prot coverage of PRIDE by first downloading all protein identifications for all 3694 human experiments (in PRIDE an experiment refers to a single LC−MS/MS run, whereas in PeptideAtlas, an experiment refers to a collection of LC−MS/MS runs performed on a single sample or closely related set of samples.) using PRIDE Biomart, 70% of which contained a UniProt accession (54182 distinct). We mapped these to Swiss-Prot using PICR,[23] obtaining 24435 distinct accessions. We removed accessions that were not valid in Swiss-Prot to obtain a list of 18935 Swiss-Prot accessions, representing 93.5% Swiss-Prot coverage. However, many experiments reported in PRIDE and elsewhere employ far less rigorous standards for protein identification than we do at PeptideAtlas. Further, simply taking a union of a large number of experiments causes the false positives to pile up. A simple computational simulation demonstrates that if the protein FDR for each of the 3684 human experiments in PRIDE is assumed to be a random value between 0.5% and 5%, and if the number of proteins identified by each experiment is a random value between 100 and 1000, then the false identifications alone will cover 93% of the proteome (see Figure S1, Supporting Information, for computer code). Many of the PA-unseen proteins that are identified in PRIDE are correct identifications, but we cannot easily discern which ones. These results stress the importance of a uniform analysis pipeline and further illustrate that there are many gene products yet to be detected by mass spectrometry.

To better understand how the goal of the C-HPP can be achieved through mass spectrometry, we explored in detail the proteins listed as PA-unseen. We examined their Swiss-Prot annotations, the tissues in which they had been detected via immunohistochemistry, their transcript abundances, and the properties they have in common via Gene Ontology analysis. We also calculated their basicity, hydrophobicity, and number of extramembrane tryptic peptides. The results of all these analyses are presented in a master table of all Swiss-Prot entries, Table S2 (Supporting Information). We developed several reasons why a protein may be PA-unseen, and we discuss them below, beginning with the more obvious and proceeding to the more subtle.

**Table 2. Fifty-two Sample Types Included in the Human PeptideAtlas**

| sample type | experiments | total PSMs (1000s) |
|---|---|---|
| blood plasma | 143 | 28238 |
| cell line Jurkat T-cell | 79 | 655 |
| cell line HEK293 | 30 | 4181 |
| T-cell | 27 | 117 |
| brain | 22 | 163 |
| lung | 20 | 152 |
| B-cell | 16 | 72 |
| cell line ES iPS | 12 | 908 |
| blood PBMC | 11 | 299 |
| monocyte | 10 | 102 |
| neutrophil | 10 | 141 |
| nucleosome | 10 | 82 |
| cell line HeLa | 9 | 1575 |
| cell line LNCaP prostate cancer | 5 | 18 |
| cell line K562 erythroleukemia | 4 | 118 |
| urine | 4 | 3 |
| blood platelets | 3 | 3124 |
| breast | 3 | 45 |
| cell line HFH primary human fetal hepatocytes | 3 | 12 |
| cell line Huh7 (hepatocarcinoma) | 3 | 7 |
| cell line U2OS | 3 | 166 |
| hair | 3 | 21 |
| semen plasma | 3 | 109 |
| cell line JCaM (LCK-deficient T-cell line) | 2 | 1.2 |
| cell line THP-1 | 2 | 3 |
| cell line unspecified | 2 | 1324 |
| mitotic spindle | 2 | 116 |
| placenta | 2 | 353 |
| prostate | 2 | 125 |
| purified adducin protein | 2 | 0.3 |
| saliva | 2 | 17 |
| adipocyte | 1 | 431 |
| blood red cell | 1 | 2 |
| bone | 1 | 14 |
| cell line Caco-2 epithelial | 1 | 9 |
| cell line Colo-205 mitochondrial | 1 | 1.2 |
| cell line HCV-hh5 | 1 | 7 |
| cell line HH4 immortalized hepatocyte | 1 | 7 |
| cell line SiHa | 1 | 0.2 |
| cell line SqCC | 1 | 0.6 |
| colorectal | 1 | 5 |
| cerebral spinal fluid (CSF) | 1 | 25 |
| eye lens | 1 | 8 |
| eye vitreous humor | 1 | 205 |
| heart | 1 | 330 |
| in vitro protein expression | 1 | 0.2 |
| nail plate | 1 | 91 |
| nuclear envelope | 1 | 1.0 |
| ovary | 1 | 0.4 |
| ovary cancer | 1 | 3 |
| pancreas | 1 | 32 |
| prostate cancer | 1 | 701 |


dx.doi.org/10.1021/pr301012j | J. Proteome Res. 2013, 12, 162−171

## Protein Existence

Although Swiss-Prot is curated to include only sequences for which there is some indication that the corresponding protein exists, it is likely that some of these inferred genes are never transcribed as mRNA or never translated into protein. These, of course, will never be seen in any proteomics experiment. In particular, of the 7614 PA-unseen proteins, only 6814, or 89%, are annotated in neXtProt (version 3.0, release 2012−05−07, corresponding to the same Swiss-Prot release with a 1:1 identifier mapping) as having evidence at the protein or transcript level. The remaining 800 are predicted to exist on the basis of homology or a gene model, or are simply annotated as of "dubious" existence. Interestingly, 150, or 1.2%, of the PA-seen proteins also bear these "no evidence" annotations. We are investigating this set carefully for possible protein-level evidence for neXtProt .

Conversely, some protein-coding genes are not included in Swiss-Prot. In fact, 2397 peptides in the Human PeptideAtlas map only to entries in the IPI or Ensembl databases. These peptides and the corresponding 1291 putative protein sequences are not included in the analysis presented herein, even though it is likely that most of these peptides are correct identifications representing real proteins. Other protein-coding genes have not yet been included in any comprehensive protein sequence database, making them impossible to identify using software based on database searching. *De novo* identification algorithms,[24] as they become faster and more accurate, will help identify these, but currently present a formidable computational task.

## MS Workflow Limitations

Proteins known to exist are sometimes difficult to detect using common shotgun proteomics techniques. We analyzed each Swiss-Prot canonical form for several features which contribute to LC−MS/MS detectability. Hydrophobicity, the tendency of a molecule to repel water, was calculated as the fraction of total residues that are highly hydrophobic (leucine, isoleucine, valine, tryptophan, tyrosine, or phenylalanine); entries with a value of >0.35 were labeled hydrophobic. Basicity, the acid-neutralizing capability of a molecule, was calculated as the fraction of total residues that are basic (histidine, lysine, or arginine), minus the fraction that are acids (aspartic or glutamic acid); entries with a value of >0.15 were labeled basic. Swiss-Prot transmembrane region boundaries were used to determine which entries represented integral membrane proteins and which residues were extra-membrane residues. The total number of observable peptides was calculated as the number of extramembrane fully tryptic peptides of length 7−30.

Six PA-unseen proteins do not contain any tryptic peptides between 7 and 30 residues, which is the effective peptide length capability for most shotgun MS workflows. A prime example is the 60S ribosomal protein L41, which has a 24-residue sequence of MRAKWRKKRMRRLKRKRRKMRQRSK. By our analysis, this is the PA-unseen protein with the second highest transcript abundance (see Table S3, Supporting Information); however, it does not contain any tryptic peptides greater than two residues in length. For some other proteins, all the tryptic peptides are fully or partially embedded in a known or predicted transmembrane region; this is the case for 34 PA-unseen proteins. Surprisingly, it is also the case for four PA-seen proteins. One (P24311) has only a single PSM and thus, according to Mayu analysis, has an 18-fold greater likelihood than multiply observed proteins of being an incorrect identification (Mayu estimates the protein level FDR for single PSM hits at 8% vs 0.45% for multiple PSM hits). For two of the proteins (Q9P0S9, P58511), the transmembrane domains are predicted theoretically; the observed peptides can be considered evidence that the predictions may be incorrect. The fourth protein (P52511), predicted by similarity to another protein to span the membrane five times, has two splice isoforms (P52511−3 and P52511−5) that are missing large chunks near the N-terminus. Since these deletions disrupt some of the predicted membrane spanning regions, it is possible that these isoforms are not membrane bound and that the observed peptides came from these isoforms.

Other physiochemical protein properties, such as hydrophobicity, play a role in protein detection. Hydrophobic proteins are often insoluble in trypsin digestion protocols, and thus few, if any, peptides result from these proteins. Some of this insolubility can be overcome with the use of detergents. Unfortunately, many detergents are not compatible with mass spectrometry analysis, and those that are show varying degrees of effectiveness[25,26] Very basic proteins are also difficult to detect using the most common fragmentation technology, CID (collision-induced dissociation). The basic residues provide an abundance of protons, leading to high fragment charge states. CID is most effective with charge states +2 and +3. To date, all experiments in the Human PeptideAtlas employ CID. ETD (electron transfer dissociation), a newer technology, allows detection of higher charge density peptides and allows better detection of basic proteins.[27] High proportions of PA-unseen proteins are very hydrophobic or very basic (Table 3).

**Table 3. Proportion of PA-seen, PA-unseen Proteins with Properties Contributing to Poor Detectability[a]**

| | PA-seen | PA-unseen |
|---|---|---|
| Very hydrophobic | 10% | 24% |
| Very basic | 1.3% | 2.1% |
| Membrane protein | 20% | 35% |

[a]Very hydrophobic = LIVWYF > 35%; very basic = (HKR)-(DE) > 15%; membrane protein = has Swiss-Prot feature TRANSMEM.

Membrane proteins present several challenges, particularly integral membrane proteins (IMPs). IMPs are proteins with one or more domains that span the phospholipid bilayer of membranes.[28] Because IMPs are bound to the insoluble membrane fraction of a cellular protein preparation, they are often discarded. Forgoing removal of the insoluble material prior to digestion will produce some peptides from loops outside the lipid bilayer in a technique known as membrane shaving.[25] However, the effectiveness of membrane shaving will depend highly on the choice of protease and the amount of protein exposed. Some proteins with multiple transmembrane domains, such as G protein-coupled receptors, expose few residues to the solvent. Use of multiple proteases with different specificities will allow release of peptides from more loops than use of trypsin alone and will thereby increase the number of proteins detected in a given experiment.[29] Additional complications with detecting IMPs result from their low abundance.[30] Often, detecting IMPs requires membrane enrichment followed by solubilization in detergent.[2530,31] These IMP-enriching steps are usually excluded from shotgun proteomic sample preparation unless specifically targeting these proteins. A high proportion of PA-unseen proteins are IMPs (35% vs 20%, see Table 3). Table 4 compares PA-seen and PA-


dx.doi.org/10.1021/pr301012j | J. Proteome Res. 2013, 12, 162−171

**Table 4. Integral Membrane Proteins (IMPs) in Swiss-Prot[a]**

| | integral membrane proteins | |
|---|---|---|
| | PA-seen | PA-unseen |
| total Swiss-Prot IMPs | 2695 | 2474 |
| average total tryptic peptides[b] | 30.1 | 21.8 |
| average extramembrane tryptic peptides[b] | 26.0 | 16.6 |
| average protein sequence length | 629 | 497 |
| average number membrane spanning segments | 3.3 | 4.8 |
| percent secreted | 5.5% | 2.4% |

[a]Included are all proteins with one or more known or putative transmembrane domains as reported by Swiss-Prot. [b]Only tryptic peptides of 7−30 residues were counted.

unseen IMPs with regard to several characteristics. On average, PA-unseen IMPs have fewer extramembrane tryptic peptides and more transmembrane domains than PA-seen IMPs.

## Sample Specificity

Some proteins may not be detected because they are present only in sample types (tissues, cell types, body fluids) that have not yet been analyzed. For a list of the sample types included in PeptideAtlas, see Table 2.

**Comparison with Human Protein Atlas.** To discover sample types that might yield many as-yet unobserved proteins, we looked in the Human Protein Atlas[32] (HPA, version 9.0, 2011−11−11) to gather tissue localization information for each Swiss-Prot entry. The HPA project utilizes antibody based proteomics to profile protein expression for about 12,000 genes in many different human tissues, cancer types, and cell lines. Forty-six different normal human tissues are covered, with multiple cell types analyzed in many cases, for a total of 66 different tissue/cell type combinations. We analyzed these to see if any were enriched for PA-unseen proteins (see Table S4, Supporting Information). We first used PICR to map the HPA's Ensembl identifiers to Swiss-Prot. 12,073 Swiss-Prot identifiers (8678 PA-seen, 3395 PA-unseen) were covered by the mapping. For 2493 of these there was a positive antibody reaction (level other than "Negative" or "None") with good reliability ("Supportive" or "High"), 331 of them in PA-unseen. (When these 331 are added to the 12629 proteins in PA-seen, we get a total of 12960 proteins with reliable LC−MS/MS or antibody evidence, giving 64.0% total confident proteome

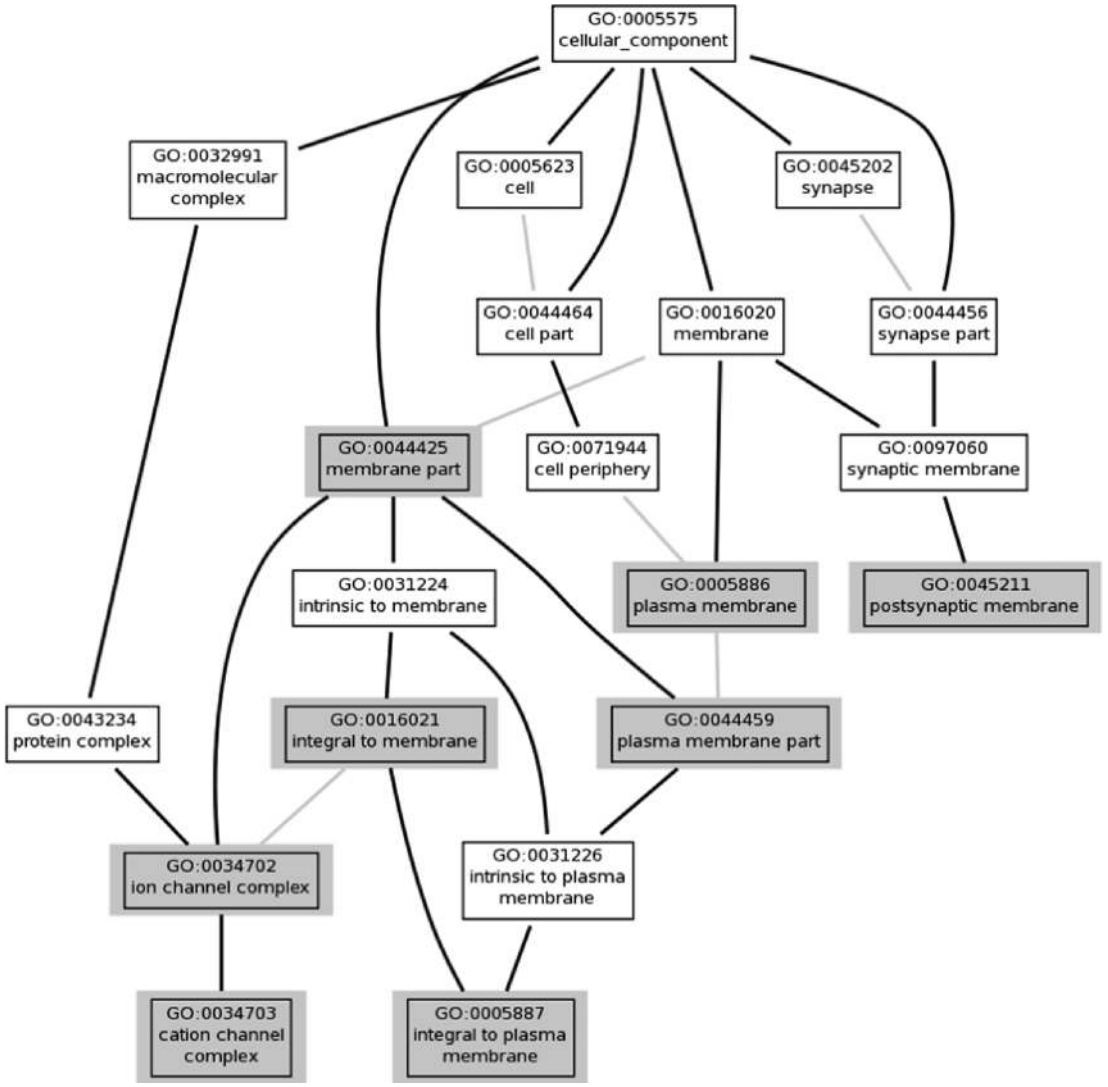



**Figure 1.** GO Cellular Component terms highly enriched among PA-unseen proteins. Terms with enrichment at *P*-value $\leq 10^{-10}$ are shaded, and only the nodes and edges which connect each of these terms with the root of the tree are depicted.


dx.doi.org/10.1021/pr301012j | *J. Proteome Res.* 2013, 12, 162−171

coverage.) For each identifier, we listed and counted the number of tissue/cell types for which there was an antibody reaction with good reliability (Table S2, Supporting Information, columns *n_reliable_staining_HuProtA* and *tissues_HuProtA*). For each tissue/cell type we also counted the number of proteins for which the HPA contained any reliable ("Supportive" or "High") reactivity at level "strong" (for staining) or "high" (for annotated protein expression) (Table S4, Supporting Information). Skeletal muscle myocytes, liver hepatocytes, and kidney cells in glomeruli showed the highest percentage of PA-unseen proteins (>10%). Samples from skeletal muscle, liver, and kidney cells have not yet been included in PeptideAtlas. Tissues from the digestive system had the highest number of PA-unseen proteins, including gall bladder (99), duodenum (87), upper stomach (83), lower stomach (74), colon (79), small intestine (78), and rectum (74). These tissues also have not yet been included in PeptideAtlas. Thus, increasing the diversity of tissue types analyzed in PeptideAtlas may help identify many of the PA-unseen proteins.

**Gene Ontology Analysis.** To determine which Gene Ontology (GO) terms are enriched among the PA-unseen proteins, we employed the GOstats package[33] (Bioconductor) running under the R statistical software. UniProt accessions (Swiss-Prot is a subset of UniProt) were mapped to Entrez gene IDs, and then the map was reversed and multiple mappings were resolved using the org.Hs.eg.db annotation package. 1612 (8%) of the Swiss-Prot IDs were missing from the map and thus were not included in this analysis. These tended to be putative or poorly characterized proteins and the majority (70%) were PA-unseen. The analysis hyperGTest was run on the PA-unseen protein list with a P-value cutoff of $10^{-10}$ and parameters conditional=TRUE and testDirection=rep for all three GO ontologies. See Figure 1 and Figures S2 and S3 (Supporting Information) for GO terms found to be over-represented in PA-unseen. For each PA-unseen protein, associated GO terms from these figures are listed in Table S2 (Supporting Information), column *enriched_GO_terms*.

Several of the experiments included in PeptideAtlas took steps to enrich for membrane proteins. Still, membrane proteins are poorly detected. All eight of the GO Cellular Component terms found to be highly enriched in PA-unseen are related to membranes, including two related specifically to ion channel complexes that span the membrane multiple times (Figure 1). Similarly, the GO biological process and molecular function analyses found many terms related to membrane proteins to be enriched in PA-unseen proteins (see Figures S2 and S3, Supporting Information).

GO biological process analysis shows that most olfactory receptors are missing from PeptideAtlas (see Supporting Information Figure S3). Not only are these receptors membrane proteins, but they are located in a tissue not represented in PeptideAtlas: the olfactory epithelium in the nasal cavity.[34]

The remaining GO terms enriched in PA-unseen have no clear explanation in terms of sample type, but other explanations are proposed in the sections below.

A final note regarding sample specificity: many of the PA-seen proteins were observed in only one type of sample in the Human PeptideAtlas (Table S5, Supporting Information). Because PeptideAtlas is far from comprehensive in its coverage of different sample types, these proteins cannot be said to be specific to these sample types. However, they may be considered candidates for tissue or cell type specificity. The particular proteins may be gleaned from Table S2 in Supporting Information, columns *n_sample_types_HuPA* and *sample_types_HuPA*.

## Low Abundance and Transient Expression

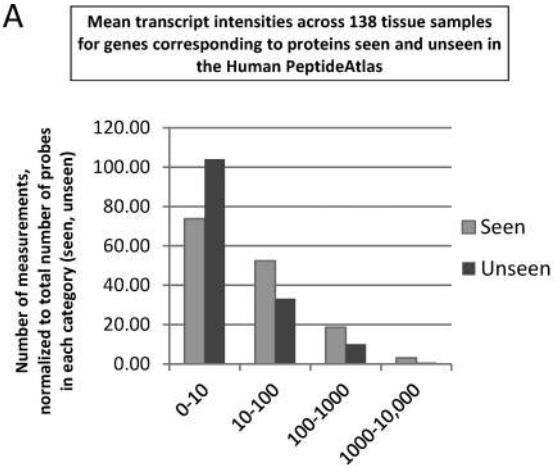
Secreted proteins such as cytokines and hormones are produced in very low abundance. They are among the least abundant proteins in blood plasma and are rarely detected in proteomics experiments.[35,36] These proteins were seen by GO analysis to be enriched in PA-unseen proteins in the previous Human PeptideAtlas build (data not shown), but enough of them were observed in the newly added data targeting signaling proteins (Table S1, Supporting Information) that most of the associated terms (*extracellular region* (CC); *growth factor activity*, *G-protein coupled receptor binding*, *cytokine receptor binding* (MF)) are no longer sufficiently enriched to appear in Figure 1. Still, slightly more (9.5% vs 8.2%) PA-unseen proteins bear the Swiss-Prot term "Secreted" than PA-seen (Table 4). The term *cytokine activity* remains enriched in PA-unseen, suggesting cytokines are still significantly under-represented, perhaps because they are among the lowest abundance secreted proteins. As additional work in secretome analysis is performed, additional identifications of these rare proteins can be expected. Use of multiple proteases will increase the detection of proteins near the limit of detectibility by multiplying the number of distinct peptides per protein. Because peptides from the same protein differ in their detectibility, for some fraction of these proteins, the additional peptides will prove to be more detectible than those created by trypsin alone and will allow detection of those proteins. Of course, for very low abundance proteins, the component peptides will all be very low abundance, and use of multiple proteases is not likely to help.

Transiently expressed proteins pose another challenge for detection by shotgun proteomics. Some proteins are only expressed during certain developmental or other biological processes. One such example is the cyclins, which are transiently expressed during, and regulate progression through, the cell cycle. Thus, cyclins are not expressed in non-proliferative tissues, while aberrant expression of cyclins is a common feature of many cancers.[37] Eight of the 25 cyclins in Swiss-Prot are in PA-unseen. Another example is proteins expressed only during fetal development. Data from fetal tissue is not included in PeptideAtlas; accordingly, two biological process GO terms related to development, *pattern specification* and *regionalization*, are enriched in PA-unseen. Other proteins, for example some neural and/or brain proteins (seen by GO analysis to be enriched in PA-unseen proteins, see Supporting Information Figure S2, *neuropeptide receptor activity*, and Figure S3, *neurological system process*), have a very short half-life. Transient expression and short half-life may present the same difficulties as low-abundance.

We performed a transcript abundance analysis to test whether protein nondetectability correlates with low transcript abundance, as shown elsewhere.[38] We used BioGPS[39] to gather transcript abundances for each Swiss-Prot entry, and compared these data to the PA-seen and PA-unseen proteins. The BioGPS database contains results for 138 samples covering 84 different tissue or cell line types. These were analyzed using an Affymetrix U133a chip and a custom designed GNF1H chip, together containing 18400 transcripts and variants covering "14500 well characterized human genes".[40,41] 14686 Swiss-Prot IDs could be mapped to Affymetrix probe IDs using the ID


dx.doi.org/10.1021/pr301012j | J. Proteome Res. 2013, 12, 162−171

A




B

BioGPS transcriptomics analysis

| | PA-seen | PA-unseen |
|---|---|---|
| Swiss-Prot identifiers | 12629 | 7614 |
| Swiss-Prot identifiers mappable to gene ID | 12053 | 6606 |
| gene IDs mappable to probeIDs | 10405 | 4281 |
| percent Swiss-Prot mapped to probeIDs | 82% | 56% |
| average mean intensity across all tissues | 120.9 | 47.6 |

**Figure 2.** Microarray transcript analysis for proteins seen and unseen in PeptideAtlas. PA-seen proteins tend to fall into the higher mean intensity bins (A) and have, on average, 2.5 times the mean intensity across all tissues (B, line 5). The microarray contained more probes for genes for PA-seen proteins than for PA-unseen because the array was biased toward proteins of high general interest (B, line 4). See text for details.

mapping tool at uniprot.org.[42] Intensities (average difference (AD) of probe hybridization intensities) for each protein in each of the 138 samples were obtained from the BioGPS cross-reference matrix, and an average intensity across all 138 samples was computed for each protein. The average intensity across all 138 tissue samples for the PA-seen proteins is much greater than the average intensity for PA-unseen proteins (Figure 2), in agreement with our expectation. We further examined those PA-unseen proteins with the highest transcript abundance. Some have properties already discussed here that hinder detection (hydrophobicity, basicity, IMPs), but for others it is unclear why they have not yet been observed (Table S3, Supporting Information).

## Informatics Limitations

To search all 470 experiments using X!Tandem consumed about 138 CPU-weeks on a modest 2010-era compute cluster. With each new build, only newly added data needs to be searched, so these searches were spread out over many months. However, if 100 compute nodes were available at once, it would take only 10 days to search all the data. Seven additional CPU-days were consumed processing and loading these results into PeptideAtlas. This compute time grows approximately linearly with the amount of data. Over the years, data set sizes have tended to grow. However, assuming continuing speed-ups in computing hardware, addition of more and ever-larger data sets to PeptideAtlas appears to be sustainable.

There are several biological process and molecular function GO terms enriched in the PA-unseen proteins (Figures S2, S3, Supporting Information) for reasons as yet unknown. These include terms related to DNA binding, transcription factor activity, G-protein coupled receptor signaling pathways and cyclic nucleotide metabolism. It is notable that all these processes are regulated by or involve post-translational modifications (PTMs). Binding of transcription factors and other proteins to DNA is regulated by a large variety of PTMs, including methylation, acetylation, phosphorylation and ubiquitylation.[43−47] G-protein coupled receptor signaling pathways utilize phosphorylation extensively,[48] and cyclic nucleotide metabolism is regulated by phosphorylation.[49] If these PTMs are included in the database searches, it is possible that more of these proteins could be detected, especially those with abundances near the limit of detection. While it might be argued that not all the peptides in such a protein will contain PTMs, so the proteins should be detectable without searching for the PTMs, searching for them will increase the number of confidently identified peptides and allow more of these proteins to pass the threshold for inclusion in PeptideAtlas. We expect the resulting increase in total proteome coverage to be modest: if 10% of the proteome is involved in processes regulated by PTMs, and 10% of those proteins have abundances near the limit of detection, then this approach will be of assistance in detecting only an additional 1% of the proteome.

Twenty-two experiments in the Human PeptideAtlas incorporated enrichment for phospho-peptides, and these data were searched for phosphorylation modifications. Using our standard search engine, X!Tandem, each potential modification on a specific amino acid approximately doubles the search time. Searching for potential phosphorylation on serine, threonine, and tyrosine thus multiplied the search time for these experiments about 8-fold. Searching for multiple modifications could multiply the search time by a factor of 100 or more. With the emerging availability of large-scale cloud computing resources, this can be computationally feasible. However, one must also pay the price of an increased error rate

due to the enlarged search space. Better results may be obtained by using recently developed software that addresses the speed and accuracy issues inherent in searching for PTMS.[50−54]

Finally, many of the PA-unseen proteins are likely in the data, but are not detected with high enough confidence to meet the very stringent thresholds implemented in PeptideAtlas. Future improvements in informatics will improve this, but excluding large numbers of false positives will always result in discarding valuable true positives.

### Estimating Observability

In order to assist researchers in prioritizing protein targets, we formulated some observability metrics for putative proteins in PA-unseen. First, we calculated an integer observability score for each putative protein, with possible scores ranging from 0 to 5, based upon the Swiss-Prot evidence code, transcript intensity, Human Protein Atlas evidence, and number of observable tryptic peptides. Second, we classified each putative protein as either *observable, observable with special handling* (for secreted, integral membrane, hydrophobic, or basic proteins), *likely unobservable* (for proteins with no transcript evidence), and *unobservable* (for proteins with no observable peptides). These metrics are detailed in Table S6 (Supporting Information), and values for each Swiss-Prot entry are listed in Table S2 (Supporting Information), columns *observability_score, observability,* and *observability_notes.* As expected, proteins in PA-seen have higher observability scores, and are more likely to be annotated as *observable*, than those in PA-unseen.

## CONCLUSION

We have presented the state of the observed MS/MS proteome using publicly available data in PeptideAtlas, uniformly processed to high stringency to October, 2012. In order to increase the number of observed proteins as much as possible, we sought out many publicly available data sets expected to increase proteome coverage. These new data sets did expand PeptideAtlas to contain nearly three times as many PSMs, 15% more peptide sequences, and 5% more proteins. This suggests that similar efforts to bring in additional high quality data sets will slowly increase the total protein coverage beyond the current 62.4%, but likely not by major increments.

Some proteins not seen in LC−MS/MS experiments have been detected in SRM experiments. PASSEL, the PeptideAtlas SRM Experiment Library,[55] is a recent addition to PeptideAtlas that provides an online catalog of publicly accessible SRM experimental results. Eight human experiments have been contributed to date. Of the 917 human Swiss-Prot entries represented so far in PASSEL, only 764, or 83%, are in PA-seen. SRM is clearly an effective technology for detecting proteins in PA-unseen. The related SRMAtlas resource,[56,57] a collection of SRM transitions covering all human entries in Swiss-Prot and validated on both synthetic peptides and synthetic proteins, will provide an excellent starting point for systematically attempting to detect as-yet undetected proteins using SRM technology.

We intend that our analysis here will assist others in the Human Proteome Project, especially those working to complete coverage for individual chromosomes, in detecting PA-unseen proteins. We suggest a three-pronged approach. First, efforts should be made to obtain unrepresented sample types such as tissues of the eye and nasal system. Second, we suggest the application of more vigorous and sophisticated experimental efforts to specifically capture membrane proteins, very basic proteins, and very hydrophobic proteins. Third, we recommend employing a systematic targeted approach using SRMAtlas to immediately deploy assays for the low abundance proteins. The effort for each protein should be focused on samples shown likely to contain that protein based on transcript and/or immunohistochemistry evidence. These strategies are feasible and should be employed by all chromosome-centric groups working in cooperation. Of course, this is not an exclusive list and other approaches may also be successful.

Finally, several studies on promising sample types were not included because their data are not yet publicly available. ProtomeXchange now provides a centralized pathway for submitting data to the proteomics data repositories, with LC−MS/MS data being stored in PRIDE, and empirical SRM data in PASSEL. Journals should require authors to make not only results, but also *raw* data, public and available via ProteomeXchange, to most efficiently advance the goals of the Human Proteome Project.

If any readers have data that detect with high confidence proteins not currently in PeptideAtlas, we encourage submission of the raw data along with basic metadata to ProteomeXchange or PeptideAtlas and they will be included in future builds.

Complementing the comprehensive information provided for all Swiss-Prot entries in Supporting Information Table S2, the PeptideAtlas web interface (www.peptideatlas.org) provides a powerful tool for exploring the data discussed here. A link to a tutorial,[58] along with a preformulated PeptideAtlas query to retrieve all PA-seen proteins and other relevant resources, may be found at www.peptideatlas.org/hupo/c-hpp. The raw data for many of the 470 experiments in the Human PeptideAtlas, including 17 of the 20 newly added data sets (see Table S1, Supporting Information, for details), is available for download at www.peptideatlas.org/repository.

## ASSOCIATED CONTENT

### Supporting Information

Figures S1−S3 and Tables S1−S6. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

### Corresponding Author

*E-mail: Terry.Farrah@systemsbiology.org. Phone: 206-732-1348. Fax: 206-732-1299. E-mail: Robert.Moritz@systemsbiology.org. Phone: 206-732-1200. Fax: 206-732-1299.

### Author Contributions

These people provided large high quality unpublished raw data sets: John R. Yates III, The Scripps Research Institute; Aaron Aslanian, The Scripps Research Institute and Salk Institute for Biological Studies; Xuemei Han, The Scripps Research Institute; Samir M. Hanash, MD Anderson Cancer Center; Chee-Hong Wong, DOE Joint Genome Institute; Akhilesh Pandey, Johns Hopkins; Krishna R. Murthy, Institute of Bioinformatics, International Technology Park, Bangalore and Amrita School of Biotechnology, Amrita Vishwa Vidyapeetham and Vittala International Institute Of Ophthalmology; Ingo Lindner, Roche Diagnostics; Angélique Augustin, F. Hoffmann-La Roche Ltd; Nikolaos Berntenis, F. Hoffmann-La Roche Ltd; Paul Cutler, F. Hoffmann-La Roche Ltd; Axel Ducret, F. Hoffmann-La Roche Ltd.

## Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

Many thanks to all whose data appears in the Human PeptideAtlas, especially Ihor Batruch, Boris Macek, Karsten Krug, Mike Snyder, George Mias, and Hogune Im, who provided their data directly when we were unable to retrieve it from Tranche. Thanks to David Campbell for PeptideAtlas support, to Joseph Slagel for help with large compute jobs, to the entire Moritz lab, especially Julian Watts, for technical assistance and insightful discussions, and to Burak Kutlu and Kerry Deutsch for R and GOStats assistance. This work has was supported in part by American Recovery and Reinvestment Act (ARRA) funds through grant number R01 HG005805 from the National Institutes of Health National Human Genome Research Institute, and the National Institute of General Medical Sciences, under grant no. R01 GM087221, 2P50 GM076547/Center for Systems Biology, the National Science Foundation MRI (grant no. 0923536), the EU FP7 grant 'ProteomeXchange' (grant no. 260558), and support from the Luxembourg Centre for Systems Biomedicine and the University of Luxembourg.

## ■ REFERENCES

(1) Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30* (3), 221−3.

(2) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.; Deutsch, E. W.; Domon, B.; Hancock, W.; He, F.; Hochstrasser, D.; Marko-Varga, G.; Salekdeh, G. H.; Sechi, S.; Snyder, M.; Srivastava, S.; Uhlen, M.; Wu, C. H.; Yamamoto, T.; Paik, Y. K.; Omenn, G. S. The human proteome project: current state and future direction. *Mol. Cell. Proteomics* **2011**, *10* (7), M111 009993.

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

(4) Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fairley, S.; Fitzgerald, S.; Gil, L.; Gordon, L.; Hendrix, M.; Hourlier, T.; Johnson, N.; Kahari, A. K.; Keefe, D.; Keenan, S.; Kinsella, R.; Komorowska, M.; Koscielny, G.; Kulesha, E.; Larsson, P.; Longden, I.; McLaren, W.; Muffato, M.; Overduin, B.; Pignatelli, M.; Pritchard, B.; Riat, H. S.; Ritchie, G. R.; Ruffier, M.; Schuster, M.; Sobral, D.; Tang, Y. A.; Taylor, K.; Trevanion, S.; Vandrovcova, J.; White, S.; Wilson, M.; Wilder, S. P.; Aken, B. L.; Birney, E.; Cunningham, F.; Dunham, I.; Durbin, R.; Fernandez-Suarez, X. M.; Harrow, J.; Herrero, J.; Hubbard, T. J.; Parker, A.; Proctor, G.; Spudich, G.; Vogel, J.; Yates, A.; Zadissa, A.; Searle, S. M. Ensembl 2012. *Nucleic Acids Res.* **2012**, *40*, D84−90.

(5) Munoz, J.; Low, T. Y.; Kok, Y. J.; Chin, A.; Frese, C. K.; Ding, V.; Choo, A.; Heck, A. J. The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Mol. Syst. Biol.* **2011**, *7*, 550.

(6) Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szymborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R. The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **2011**, *7*, 549.

(7) Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **2012**, *11* (3), M111 014050.

(8) Sharma, K.; Vabulas, R. M.; Macek, B.; Pinkert, S.; Cox, J.; Mann, M.; Hartl, F. U. Quantitative proteomics reveals that Hsp90 inhibition preferentially targets kinases and the DNA damage response. *Mol. Cell. Proteomics* **2012**, *11* (3), M111 014654.

(9) Nagaraj, N.; Wisniewski, J. R.; Geiger, T.; Cox, J.; Kircher, M.; Kelso, J.; Paabo, S.; Mann, M. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **2011**, *7*, 548.

(10) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.; Martin, D. B.; Nesvizhskii, A. I.; Aebersold, R. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10* (6), 1150−9.

(11) Jones, P.; Cote, R. G.; Cho, S. Y.; Klie, S.; Martens, L.; Quinn, A. F.; Thorneycroft, D.; Hermjakob, H. PRIDE: new developments and new datasets. *Nucleic Acids Res.* **2008**, *36*, D878−83.

(12) Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R. PRIDE: the proteomics identifications database. *Proteomics* **2005**, *5* (13), 3537−45.

(13) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; 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.

(14) Craig, R.; Cortens, J. P.; Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **2004**, *3* (6), 1234−42.

(15) Fenyo, D.; Eriksson, J.; Beavis, R. Mass spectrometric protein identification using the global proteome machine. *Methods Mol. Biol.* **2010**, *673*, 189−202.

(16) Falkner, J. A.; Andrews, P. C. Tranche: Secure Decentralized Data Storage for the Proteomics Community. *J. Biomol. Tech.* **2007**, *18* (1), 3.

(17) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466−7.

(18) MacLean, B.; Eng, J. K.; Beavis, R. C.; McIntosh, M. General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* **2006**, *22* (22), 2830−2.

(19) Farrah, T.; Deutsch, E. W.; Omenn, G. S.; Campbell, D. S.; Sun, Z.; Bletz, J. A.; Mallick, P.; Katz, J. E.; Malmstrom, J.; Ossola, R.; Watts, J. D.; Lin, B.; Zhang, H.; Moritz, R. L.; Aebersold, R. A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell. Proteomics* **2011**, *10* (9), M110 006353.

(20) Kersey, P. J.; 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.

(21) Boutet, E.; Lieberherr, D.; Tognolli, M.; Schneider, M.; Bairoch, A. UniProtKB/Swiss-Prot. *Methods Mol. Biol.* **2007**, *406*, 89−112.

(22) 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.; Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003**, *31* (1), 365−70.

(23) Wein, S. P.; Cote, R. G.; Dumousseau, M.; Reisinger, F.; Hermjakob, H.; Vizcaino, J. A. Improvements in the protein identifier cross-reference service. *Nucleic Acids Res.* **2012**, *40*, W276−80.

(24) Hughes, C.; Ma, B.; Lajoie, G. A. De novo sequencing methods in proteomics. *Methods Mol. Biol.* **2010**, *604*, 105−21.

(25) Speers, A. E.; Wu, C. C. Proteomics of integral membrane proteins−theory and application. *Chem. Rev.* **2007**, *107* (8), 3687−714.

(26) Klammer, A. A.; MacCoss, M. J. Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J. Proteome Res.* **2006**, *5* (3), 695−700.

(27) Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (26), 9528−33.

(28) Lu, B.; McClatchy, D. B.; Kim, J. Y.; Yates, J. R., 3rd Strategies for shotgun identification of integral membrane proteins by tandem mass spectrometry. *Proteomics* **2008**, *8* (19), 3947−55.
(29) Glatter, T.; Ludwig, C.; Ahrne, E.; Aebersold, R.; Heck, A. J.; Schmidt, A. Large-Scale Quantitative Assessment of Different In-Solution Protein Digestion Protocols Reveals Superior Cleavage Efficiency of Tandem Lys-C/Trypsin Proteolysis over Trypsin Digestion. *J. Proteome Res.* **2012**, *11* (11), 5145−56.
(30) Savas, J. N.; Stein, B. D.; Wu, C. C.; Yates, J. R., 3rd Mass spectrometry accelerates membrane protein analysis. *Trends Biochem. Sci.* **2011**, *36* (7), 388−96.
(31) Gilmore, J. M.; Washburn, M. P. Advances in shotgun proteomics and the analysis of membrane proteomes. *J. Proteomics* **2010**, *73* (11), 2078−91.
(32) Uhlen, M.; Oksvold, P.; Fagerberg, L.; Lundberg, E.; Jonasson, K.; Forsberg, M.; Zwahlen, M.; Kampf, C.; Wester, K.; Hober, S.; Wernerus, H.; Bjorling, L.; Ponten, F. Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **2010**, *28* (12), 1248−50.
(33) Falcon, S.; Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **2007**, *23* (2), 257−8.
(34) DeMaria, S.; Ngai, J. The cell biology of smell. *J. Cell Biol.* **2010**, *191* (3), 443−52.
(35) Anderson, N. L.; Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **2002**, *1* (11), 845−67.
(36) Schiess, R.; Wollscheid, B.; Aebersold, R. Targeted proteomic strategy for clinical biomarker discovery. *Mol. Oncol.* **2009**, *3* (1), 33−44.
(37) Canavese, M.; Santo, L.; Raje, N. Cyclin dependent kinases in cancer: potential for therapeutic intervention. *Cancer Biol. Ther.* **2012**, *13* (7), 451−7.
(38) Van, P. T.; Schmid, A. K.; King, N. L.; Kaur, A.; Pan, M.; Whitehead, K.; Koide, T.; Facciotti, M. T.; Goo, Y. A.; Deutsch, E. W.; Reiss, D. J.; Mallick, P.; Baliga, N. S. Halobacterium salinarum NRC-1 PeptideAtlas: toward strategies for targeted proteomics and improved proteome coverage. *J. Proteome Res.* **2008**, *7* (9), 3755−64.
(39) Wu, C.; Orozco, C.; Boyer, J.; Leglise, M.; Goodale, J.; Batalov, S.; Hodge, C. L.; Haase, J.; Janes, J.; Huss, J. W., 3rd; Su, A. I. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* **2009**, *10* (11), R130.
(40) Su, A. I.; Wiltshire, T.; Batalov, S.; Lapp, H.; Ching, K. A.; Block, D.; Zhang, J.; Soden, R.; Hayakawa, M.; Kreiman, G.; Cooke, M. P.; Walker, J. R.; Hogenesch, J. B. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (16), 6062−7.
(41) Su, A. I.; Cooke, M. P.; Ching, K. A.; Hakak, Y.; Walker, J. R.; Wiltshire, T.; Orth, A. P.; Vega, R. G.; Sapinoso, L. M.; Moqrich, A.; Patapoutian, A.; Hampton, G. M.; Schultz, P. G.; Hogenesch, J. B. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (7), 4465−70.
(42) UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D71−5.
(43) Wang, C.; Powell, M. J.; Popov, V. M.; Pestell, R. G. Acetylation in nuclear receptor signaling and the role of sirtuins. *Mol. Endocrinol.* **2008**, *22* (3), 539−45.
(44) Erce, M. A.; Pang, C. N.; Hart-Smith, G.; Wilkins, M. R. The methylproteome and the intracellular methylation network. *Proteomics* **2012**, *12* (4−5), 564−86.
(45) Kouzarides, T. Acetylation: a regulatory modification to rival phosphorylation? *EMBO J.* **2000**, *19* (6), 1176−9.
(46) Cohen, P. The regulation of protein function by multisite phosphorylation–a 25 year update. *Trends Biochem. Sci.* **2000**, *25* (12), 596−601.
(47) Vertegaal, A. C. Uncovering ubiquitin and ubiquitin-like signaling networks. *Chem. Rev.* **2011**, *111* (12), 7923−40.
(48) Tobin, A. B.; Butcher, A. J.; Kong, K. C. Location, location, location...site-specific GPCR phosphorylation offers a mechanism for cell-type-specific signalling. *Trends Pharmacol. Sci.* **2008**, *29* (8), 413−20.
(49) Francis, S. H.; Blount, M. A.; Corbin, J. D. Mammalian cyclic nucleotide phosphodiesterases: molecular mechanisms and physiological functions. *Physiol. Rev.* **2011**, *91* (2), 651−90.
(50) Fu, Y.; Xiu, L. Y.; Jia, W.; Ye, D.; Sun, R. X.; Qian, X. H.; He, S. M. DeltAMT: a statistical algorithm for fast detection of protein modifications from LC-MS/MS data. *Mol. Cell. Proteomics* **2011**, *10* (5), M110 000455.
(51) Savitski, M. F.; Savitski, M. M. Unbiased detection of posttranslational modifications using mass spectrometry. *Methods Mol. Biol.* **2010**, *673*, 203−10.
(52) Shilov, I. V.; Seymour, S. L.; Patel, A. A.; Loboda, A.; Tang, W. H.; Keating, S. P.; Hunter, C. L.; Nuwaysir, L. M.; Schaeffer, D. A. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics* **2007**, *6* (9), 1638−55.
(53) Mazur, M. T.; Fyhr, R. An algorithm for identifying multiply modified endogenous proteins using both full-scan and high-resolution tandem mass spectrometric data. *Rapid Commun. Mass Spectrom.* **2011**, *25* (23), 3617−26.
(54) Na, S.; Bandeira, N.; Paek, E. Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell. Proteomics* **2012**, *11* (4), M111 010199.
(55) Farrah, T.; Deutsch, E. W.; Kreisberg, R.; Sun, Z.; Campbell, D. S.; Mendoza, L.; Kusebauch, U.; Brusniak, M. Y.; Huttenhain, R.; Schiess, R.; Selevsek, N.; Aebersold, R.; Moritz, R. L. PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics* **2012**, *12* (8), 1170−5.
(56) Kusebauch, U. Manuscript in preparation.
(57) Deutsch, E. W.; Campbell, D. S.; Mendoza, L.; Sun, Z.; Farrah, T.; Kusebauch, U.; Chu, C.; Stevens, J.; Slagel, J.; Picotti, P.; Brusniak, M.-Y.; Lam, H.; Bletz, J.; Wang, G.; He, W.-w.; Hood, L.; Aebersold, R.; Moritz, R. L., SRMAtlas: Generating targeted proteomics transition atlases for complete proteomes; In preparation.
(58) Farrah, T.; Deutsch, E. W.; Aebersold, R. Using the Human Plasma PeptideAtlas to study human plasma proteins. *Methods Mol. Biol.* **2011**, *728*, 349−74.