# The state of the human proteome in 2012 as viewed through PeptideAtlas

## Supporting Information Tables and Figures

See Excel Spreadsheet for Tables S1-S6.

**Table S1. Twenty datasets added to the Human PeptideAtlas in summer 2012.**

**Table S2. Catalog of the 20,243[i] entries in Swiss-Prot, with the following shown for each: physical properties contributing to observability; observations in Human PeptideAtlas; observations in Human Protein Atlas; transcript abundance; estimations of observability as in Table S6.**

**Table S3. PA-unseen proteins with highest transcript abundance.**

**Table S4. Human Protein Atlas tissue and cell type terms. Each term is listed with a count, out of 3717 total possible, of antibodies that are strongly and reliably reactive to proteins not observed in the Human PeptideAtlas. Types with the largest numbers of PA-unseen proteins include several from the digestive system (orange highlighting). Those with the largest proportions of PA-unseen proteins include skeletal muscle myocytes, liver hepatocytes, and kidney cells in glomeruli (green highlighting).**

**Table S5. 2061, or 16% of all PA-seen proteins, were observed in only one sample type in the Human PeptideAtlas. 36 of the total 53 different sample types have at least one protein specific to it.**

**Table S6. (Left) Criteria used to compute observability score for each protein and to label some proteins as likely unobservable or requiring special handling to be observed in Table S2. (Right) Tallies.**

[i] The number of entries in this table is one fewer than the number of entries in the 2012-05 release of Swiss-Prot because one of the Swiss-Prot entries, GV30H4, was inadvertently excluded from PeptideAtlas

.

**Figure S1: Perl script demonstrating 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.**

```perl
#!/usr/local/bin/perl

use strict;

my @proteins;

my $nProteins=20243;
my $nDetectableProteins=12629;
my $nUndetectableProteins=$nProteins-$nDetectableProteins;
my $nExperiments=3684;
#my $FDR = 0.02;
my $FDR_high = 0.05;
my $FDR_low = 0.005;

for (my $iExp=0; $iExp < $nExperiments; $iExp++) {
  my $nCorrect = 100 + int(rand(900));
  my $FDR = (5 + rand(45)) / 1000;
  my $nWrong = int($nCorrect * $FDR);
  #my $nWrong = 11;

  # Mark a random set of proteins as correctly detected
  my @tmpSelected;
  for (my $iCorrect=0; $iCorrect<$nCorrect; $iCorrect++) {
    my $done = 0;
    while (!$done) {
      my $rand = int(rand($nDetectableProteins));
      if (!$tmpSelected[$rand]) {
        $proteins[$rand]++;
        $tmpSelected[$rand]++;
        $done = 1;
      }
    }
  }

  # Mark a random set of proteins as incorrectly detected
  my @tmpIncorrectlySelected;
  for (my $iWrong=0; $iWrong<$nWrong; $iWrong++) {
    my $done = 0;
    while (!$done) {
      # A detectable protein can be incorrectly detected
      my $rand =  int(rand($nProteins));
      #my $rand = $nDetectableProteins + int(rand($nProteins-$nDetectableProteins));
      if (!$tmpIncorrectlySelected[$rand]) {
        $proteins[$rand]++;
        $tmpIncorrectlySelected[$rand]++;
        $done = 1;
      }
    }
  }

  #printf("Exp $iExp Ids $nCorrect proteins with %.1f\% FDR = $nWrong wrong\n",$FDR*100);


}

  my $nDetectedCorrect = 0;
  for (my $i=0; $i<$nDetectableProteins; $i++) {
    if ($proteins[$i]) {
      $nDetectedCorrect++;
    }
  }


  my $nDetectedIncorrect = 0;
  for (my $i=$nDetectableProteins; $i<$nProteins; $i++) {
    if ($proteins[$i]) {
      $nDetectedIncorrect++;
    }
  }

  printf("Of $nDetectableProteins detectable proteins, $nDetectedCorrect were seen (%.1f\%)\n",
        $nDetectedCorrect/$nDetectableProteins*100);

  printf("Of %d undetectable proteins, $nDetectedIncorrect were claimed to be seen (%.1f\%)\n",
        $nProteins-$nDetectableProteins,$nDetectedIncorrect/($nProteins-
$nDetectableProteins)*100);

  printf("Of all $nProteins proteins, %d were claimed to be seen (%.1f\%)\n",

$nDetectedCorrect+$nDetectedIncorrect,($nDetectedCorrect+$nDetectedIncorrect)/$nProteins*100);
```

**Figures S2 and S3, along with several zoom-in views for each, are found in accompanying graphics files.**

**Figure S2. GO Molecular Function terms highly enriched among PA-unseen proteins. Terms with enrichment at P-value <= $10^{-10}$ are shaded , and only the nodes and edges which connect each of these terms with the root of the tree are depicted.**

**Figure S3. GO Biological Process terms highly enriched among PA-unseen proteins. Terms with enrichment at P-value <= $10^{-10}$ are shaded , and only the nodes and edges which connect each of these terms with the root of the tree are depicted.**