# Metaprotein Expression Modeling for Label-Free Quantitative Proteomics

Joseph E. Lucas,[*,†,‡] J. Will Thompson,[†,‡] Laura G. Dubois,[†] Jeanette McCarthy,[†] Keyur Patel,[†] Hans Tillman,[†] Alex Thompson,[†] John McHutchison,[†] and M. Arthur Moseley[†]

*Duke University*

E-mail: joe@stat.duke.edu

**Abstract**

We describe a Bayesian hierarchical model for analyzing unbiased, label-free proteomics data which utilizes the covariance of peptide expression across samples as well as MS/MS-based identifications to group peptides – a strategy we call metaprotein expression modeling. Our metaprotein model takes into account the possibility of misidentifications, post-translational modifications and systematic differences between samples due to changes in instrument sensitivity or differences in total protein concentration. We describe the similarities and differences between this metaprotein model and protein level quantitation algorithms, then demonstrate the clinical/translational utility of the model for differentiating biological phenotypes in the context of a cohort of patients with Hepatitis C.

[*]To whom correspondence should be addressed
[†]Duke University
[‡]The first two authors contributed equally

# Introduction

The field of proteomics has made remarkable advances in analytical hardware and software which have provided increasingly sensitive and robust analyses on platforms capable of detecting low abundance proteins from complex mixtures, such as serum and cell lysates. The nanoscale liquid chromatography and mass spectrometry (LC/MS) proteomic technology, which has become the state-of-the-art for differential expression proteomic studies in most major laboratories around the world, typically uses a 'bottom-up' approach, where the samples are subjected to a total proteolytic digestion, and the peptide 'surrogates' of the protein are quantified and identified using tandem mass spectrometry. There are two general approaches used for bottom-up differential expression proteomics via LC-MS, those based on the use of stable isotope labeling or tagging of the peptides, and the so-called label-free methods (no labeling).[1–4] Advances in both approaches have occurred in recent years such that currently both relative and absolute quantitation of proteins is possible from complex mixtures by either labeled or label-free methodologies.[5]

Although a specific advantage of the labeling approaches is the ability to heavily fractionate the samples to "dig deeper" into the proteome while maintaining quantitative capabilities, extensive fractionation of the sample is often impractical in the context of a clinical study with tens or even hundreds of samples. The proteomics community has seen a significant increase in the use of the label free approach due to increased instrument stability and software sophistication, and it is emerging as the method of choice for larger clinically-based studies where use of the labeling strategies is impossible or less practical. In particular, an advantage of label-free strategies which measure area-under-the-curve (AUC) of the LC-MS peak is that any of a number of commercial or open-source software packages can be used to extract ion intensities from each individual analysis, and statistical analysis on the relative abundance of these ions can be performed even in the absence of a peptide identification. The ability to precisely and reproducibly quantify thousands of proteolytic peptides using the label-free method was demonstrated by Wang, et al and has been since employed and reproduced in a number of laboratories.[4,6]

Techniques for aggregating peptides into larger units generally revolve around protein identi-

fications. A variety of approaches exist to combine individual peak areas to generate a relative or absolute aggregate expression levels. Once peptides are assigned to their parent proteins, using an algorithm such as ProteinProphet, either the peptide frequency of observation ("spectral counting") or MS intensity is used to estimate protein abundance.[7] The spectral counting approaches have gained a fairly large degree of use in the community due to their ease of implementation, however they generally suffer from a limited dynamic range and they are insensitive to small changes in expression level due to the large number of species which have peptides observed only 1-3 times.[5] Label-free AUC approaches generally overcome these limitations by locating a peak in the retention-time and accurate-mass matrix using sophisticated software, and extracting the area under the LC-MS peak. An important characteristic of AUC label-free studies is that they need to be performed on high mass-accuracy instruments for the best results, which limits the application of this approach to more expensive QToF, FT-ICR, or Orbitrap instruments.

This current research aims to overcome several major oversimplifications that can occur during the aggregation of peptides into groups based on protein identifications. Error in protein-level quantitation can first occur due to incorrect peptide identifications. Even at a relatively low peptide false-discovery rate (i.e. 1%), the fraction of proteins detected that contain at least one false peptide identification is much higher because multiple peptides match back to the same protein. If a false-positive peptide is included in the protein level quantitation it can cause increased error in the protein-level quantitation. This can be partially overcome utilizing only the highest-quality or best-ionizing peptides for protein quantitation, however in current implementations of "Top 3" quantitation the individual peptide confidence is not utilized as an inclusion parameter.[8,9] A second type of error in protein quantitation occurs when many homologs share a common peptide. In this situation the protein grouping algorithm, such as ProteinProphet, makes an informed decision about which parent sequence a peptide belongs to and typically associates all of the peptide intensity to that parent sequence. This can deliver erroneous protein quantitation results when multiple homologs are present. A final challenge is with post-translationally modified or proteolytically processed peptides, which may show a biologically relevant and different expression pattern than the

proteotypic peptides. In these cases, they should not be grouped together with the other peptides for the purposes of modeling expression.

This is paper describes a statistical model which is designed to address limitations to current "state of the art" for assessing differential expression in mass spectrometry proteomics. For example, Daly et al.[10] describe a mixed effects model for estimation of protein level differential expression, however, this model treats each protein independently and therefore can not handle systematic correlation between proteins. Additionally, it relies on identifications to group features and is therefore restricted to the analysis of identified features and is subject to the introduction of noise due to incorrect identifications. Finally, it is designed around the detection of differential expression between experimental groups and therefore may not be appropriate for other types of studies. Another work is the software product Corra from Brusniak et al.,[11] which focuses on a computational framework rather than any particular statistical model. Corra is a wrapper for statistical procedures that target proteomic LC-MS data and we are examining the possibility of adapting the metaprotein model we describe in this paper for inclusion in the Corra framework. The work of Karpievitch et al.[12] describes a fixed effects model that is similar to that in[10] except that it models the proteins concurrently, but treats replicate measurements from the same subject as independent observations. As in the work of Daly et al.,[10] identifications are relied upon and are assumed to be correct and the model is designed to detect differential expression in the presence of known design vectors. Finally, all of these models rely on maximum likelihood estimation, the output of which are point estimates of model parameters without uncertainty. In fact, there is uncertainty about the accuracy of the model parameter estimates, and the approach we will outline addresses this shortcomming.

We present here a metaprotein classification approach which aims to minimize the effects of the previously mentioned types of error on a quantitative proteomic data set, by utilizing the expression profile of a peptide or group of peptides to assist its grouping with similarly-quantified peptides, as well as the traditional grouping by common "parent" protein sequence. In particular, our approach:

- Allows for the subtraction of large scale correlational structure between proteins that likely

arise from technical rather than biological variability (batch effects).

- Appropriately models both identified and unidentified features of the LC-MS output
- Utilizes feature identifications from MS/MS spectra, but allows for the probability that some of those identifications will be incorrect
- Produces a full posterior distribution on the model parameters, which leads to the quantification of uncertainty in the results.
- Admits the possibility that sections of a protein will be post-translationally modified and therefore may not be representative of the expression pattern of the protein as a whole.
- Makes use of correlational structure across samples, which provides significant information about feature relationships that is unused in many other approaches.
- Can be used in the creation of predictive models based on multiple proteins, rather than just the enumeration of proteins associated with a particular outcome.

In the next section we will describe the Hepatitis C experimental data set on which we exemplify the use of our model, including a description of the cohort, as well as an explanation of the sample preparation, LC-MS operation and data preprocessing. Following that, in the “Statistical Methodology" section, we will describe the statistical model and some of its properties. Finally, we apply the model to a Hepatitis C study.

# Experimental

## Sample Selection

Chronic hepatitis C (HCV) patients, n = 87, were selected for proteomic analysis from the Duke Hepatology Database and Biorepository. Samples analyzed were all pre-treatment serum aliquots, but patients were classified based on their sustained response to PEGylated Interferon/Ribavirin

combination therapy, the standard of care for HCV. Patients were matched as well as possible on the basis of relevant clinical parameters, including viral genotype, sex, viral load, metavir fibrosis score, and race. Patients were divided between Genotype 1 HCV non-responders (n=42), Genotype 1 HCV responders (n=34), and Genotype 2/3 HCV responders (n=20).

## Sample Preparation

Serum samples were statistically randomized (3X) and the sample processing was performed blind to treatment group. Samples were thawed on ice and vortexed briefly. 50 uL of each serum samples was diluted 1:4 with Agilent immunodepletion Buffer A (5185-5990) and centrifuged through a 0.22 $\mu$m filter (5185-5990) at 10000 rpm. A 10 uL aliquot was removed for total protein assay (mini Bradford, Bio-Rad, Inc), and the samples were pipette into glass LC vials and stored in the autosampler at 4° C until analysis. The Agilent Multi Affinity Removal Column HU-14, 4.6 x 100 mm (5188-6558), was installed on an Agilent 1100 HPLC system, and conditioned with five 20$\mu$L plasma depletions before the first sample injection (normal plasma from Bioreclamation, Inc). HPLC-based immunodepletion and fraction collection was then performed as instructed by the manufacturer, except that only 20 ul serum equivalent (100 ul total) was injected on the column in order to ensure that the column had sufficient capacity to remove the target proteins across the sample cohort. The 1 mL unbound fraction, collected from 7 to 17 minutes during depletion, was desalted and buffer exchanged into 50 mM ammonium bicarbonate (EMD 1.01131.0500) at least 100x using a 10 kDa MWCO filter (Amicon 4, Millipore, Inc). The final sample was concentrated to approx 50uL, after which protein concentration was measured with a mini-Bradford assay. A 5 ug aliquot of each sample was run on a Novex SDS-PAGE gel (Invitrogen, Inc) as quality control measure. Approximately 20 ug of each sample was aliquoted for digestion, and sample concentrations were normalized to approximately 0.8 ug/uL with 50 mM ammonium bicarbonate (EMD). Following normalization, Rapigest SF (Waters, 186001861) was added to 0.1% w/v. Samples were then reduced, alkylated, and digested following a standard in-solution digestion protocol (http://www.genome.duke.edu/cores/proteomics/sample-preparation/). The samples were

reduced in 10 mM dithiothreitol (VWR, VW1506-02), alkylated in 20 mM iodoacetamide (Calbiochem 407710), and digested with 0.4 ug sequencing grade modified trypsin (Promega V5111). After digestion overnight, samples were acidified to 1% trifluoroacetic acid (Pierce, PI28904) and heated for 2 hrs at 60° C to remove Rapigest. Samples were then centrifuged at 15,000 rcf for 10 mins and the supernatant was pipetted into total recovery LC vials (Waters Corporation).

## LC-MS Operation

Each sample was analyzed by injecting approximately 1 ug of total digested protein onto a 75um x 250 mm BEH C18 column (Waters) and separated using a gradient of 5 to 40% acetonitrile with 0.1% formic acid, with a flow rate of 0.3uL/min, in 120 minutes on a nanoAcquity liquid chromatograph (Waters). Electrospray ionization was used to introduce the sample in real-time to a Q-Tof Premier mass spectrometer (Waters), collecting data in MSE mode with 0.9 second alternating scans between low CE (6V) and a high CE ramp (15V to 40V). Data collection in this fashion supplies sufficient sampling across the chromatographic elution of a peptide for accurate quantitation, while also allowing acquisition of data used for the qualitative identifications. Technical reproducibility was assessed by running a subset of the samples in triplicate (n=6) and also by analyzing a pooled sample at predefined intervals. In addition, a number of data-dependent LC-MS/MS analyses were performed using the same LC gradient and injection volumes; these runs provided column conditioning prior to quantitative analysis, and in some cases complementary peptide identifications.

## Preparation of data for analysis

To accomplish data alignment and feature quantitation across all biological samples, we used the Rosetta Elucidator$^{TM}$v3.3 software package (Rosetta Biosoftware) to import and align all MSE and data-dependent acquisition (DDA) raw data files. Database searches were performed against a forward/reverse Swissprot database (v 56.5) with human taxonomy, using ProteinLynx Global Server v2.4 (IdentityE algorithm, Waters Corporation) for MSE searches or Mascot v2.2 for DDA

data. Database searches are either performed externally and results imported (PLGS 2.4) or queued directly from within Elucidator (Mascot) to allow identification of many of the quantified features in the proteomic dataset. All database searches were performed with high mass accuracy on precursor and product ions (typically 20 ppm precursor and 0.04Da product ion tolerance), with fixed carbamidomethylation(Cys), variable oxidation(Met) and variable deamidation(Asn and Gln). Annotation of the peptides is accomplished at an estimated 1% FDR using the Elucidator implementation of PeptideProphet algorithm.[13] Visual scripting within Elucidator is utilized to extract feature intensities for those features which have quantitative values above the 1000 counts (approximately 10th percentile) in 50% of the samples. The final file for statistical analysis is made up of a matrix of intensities, with the rows corresponding to isotope groups and the columns to technical observations (LC-MS analysis). An isotope group is defined as all of the peaks associated with a single peptide at a specific charge state and retention time. This level of quantitation combines peaks from the same peptide that differ according to the number of carbon 13's incorporated, but does not combine the same peptide measured at different charge states. The intensity of an isotope group for a given sample is the total volume under the feature peaks associated with that isotope group. This is monotonically related to the concentration of that isotope group in the original sample, and it is these intensities that we work with.

## Statistical Methodology

In order to estimate metaprotein abundance, we build our model from pre-processed data (described in the previous section) with intensity estimates aggregated at the isotope group level. We introduce the term metaprotein here to differentiate this approach from those in which peptide identifications lead to a fixed assignment of a particular peptide to a protein. In our modeling approach, we allow the possibility that an isotope group will be incorrectly identified, or be correctly identified, but have a pattern of expression that is distinct from the bulk of peptides from the corresponding protein. In practice, this new grouping approach often leads to metaproteins which may be dominated by isotope groups from a particular protein, but which contain isotope groups from

other proteins as well.

Let $X$ be a $P \times N$-dimensional matrix consisting of measurements on $P$ isotope groups across $N$ samples. We utilize a modification of the latent factor model outlined previously in.[14–17]

$$X = \mu 1'_N + A\Lambda' + \varepsilon \tag{1}$$

The $P-$dimensional vector $\mu$ has elements $\mu_i$ representing the mean expression of isotope group $i$ and $1_N$ is a column vector of ones. The $N \times K$-dimensional matrix $\Lambda$ represents latent factors which will be learned from the data and $A$ is a $P \times K$-dimensional matrix of factor loadings with elements $a_{i,k}$. The random variable $\varepsilon$ is a $P \times N$ matrix of idiosyncratic noise.

Our goal is to estimate relative protein concentration from this model using the latent factors in $\Lambda$. Recall that we have identifications for some subset of the isotope groups. With this in mind, suppose we identify each column of $A$ and the corresponding column of $\Lambda$ with one identified protein. If we set $a_{i,k} = 1$ when isotope group $i$ is from a peptide identified as coming from protein $k$ and $a_{i,k} = 0$ otherwise, then our model is describing the expression pattern of each isotope group as a noisy approximation of the expression pattern of the protein, where the protein is known.

Retaining, for the time being, the idea of fixing $a_{i,k}$ in this way, we wish to handle the possibility of changing sensitivity and changing protein concentration from sample to sample. To account for this, we introduce an additional set of latent factors into equation 1.

$$X = \mu 1'_n + BH' + A\Lambda' + \varepsilon \tag{2}$$

We now introduce latent factors $H$ and factor loadings $B = (b_{i,j})$ where $j = 1 \cdots J$ which we use to account for systematic structure in the data that is sample specific. Because these features will span almost all peptides, we utilize a generic Gaussian prior for the elements of $B$.

$$b_{i,j} \sim N(m_0, v_0)$$

This distribution represents our belief that these effects span all isotope groups, but with varying effect sizes. This prior also minimizes identifiability issues between $B$, which is not sparse, and $A$ which is very sparse with somewhat informative priors.

We want to modify our prior on $A$ to allow for possible post-translational modifications and for misidentifications. With this in mind, we want to relax our strict assignment of zeros and ones in the loadings matrix $A$. Instead, our prior distribution for $a_{i,k}$ will reflect our level of certainty that we know which factor should represent the expression of this peptide. When we have an identification for peptide $i$ and have mapped that peptide to protein $k$, our prior distribution will reflect an increased certainty that $a_{i,k} \neq 0$.

We introduce a $p-$dimensional vector of latent variables $(z_i)$ which identifies the non-zero column of $A$ for each isotope group. When we have an identification that suggests that isotope group $i$ comes from protein $k$, our prior distribution for $z_i$ is

$$\begin{aligned} z_i &\sim Multinomial(1, q_i) \\ q_i &\sim Dir(a_0, \cdots, a_0, a_k, a_0, \cdots a_0) \end{aligned}$$

where $a_k$ is substantially larger than $a_0$ to reflect our prior belief that $z_i = k$. For peptides which do not have identifications, we utilize an unbiased prior $z_i \sim Dir(a_0)$. Because different peptides showing similar expression patterns may, nonetheless, show a different magnitude of expression of that pattern due to the relative sensitivity of the mass spectrometer for the peptide, we model each of the non-zero elements of $A$ independently, such that $a_{i,k} \sim N(m_a, v_a)$ when $z_i = k$ and $a_{i,k} = 0$ otherwise.

To complete the model specification, we assume a conjugate, row specific inverse gamma prior for the variance of $\varepsilon$. We also assume that the individual columns of $\Lambda$ arise from a uniform distribution on the $N$-dimensional sphere of radius $\sqrt{N}$. The model is fit via Markov chain Monte Carlo (MCMC) and the result of this fit is a set of draws from the posterior distribution of all of the model parameters. All prior distributions are conjugate, and therefore we may use Gibbs sampling

to update the model parameters at each step of the MCMC.

**Overlapping Peaks**

We have restricted our peptides to belong to just one metaprotein. As an alternative, one might allow more than one non-zero element in each row of $A$. This is equivalent to assuming that more than one metaprotein is responsible for the expression pattern seen in each peptide, which would be expected in cases where multiple isotope groups have highly overlapping peaks. Although the extent to which we see multiple isotope groups in a single peak is unclear, this type of structure can be accounted for with relaxed priors on $A$. If $a_{i,k}$ is an element of $A$, then a point mass mixture prior accomplishes our goals.

$$a_{i,k} \sim (1-q_k)\delta_0(a_{i,k}) + q_k N(a_{i,k}|m_0, v_0) \tag{3}$$

where $\delta_0$ is the distribution describing a point mass at 0. This distribution represents our prior belief that some, but not all, metaproteins will be required to describe the expression pattern of each isotope group. The normal distribution allows the magnitude of the effect to vary. For each of the metaproteins, we estimate the number of associated isotope groups by our prior distribution on the mixing probability $q_k$:

$$q_k \sim Be(\alpha_0, \gamma_0)$$

This approach allows for the restriction on isotope group association with just one metaprotein to be relaxed. However, as the resolution of mass spectrometry increases, and as fractionation in multiple dimensions (such as 2D chromatography and ion mobility) makes the distinction between polypeptides clearer, this modification to the model will become less and less important. Further, experience suggests that the vast majority of measured peaks are single species. Because of this, the addition of features to deal with overlapping peaks can introduce more noise than it removes, particularly when the number of samples in the experiment (and therefore the amount of

information available from the correlation structure) is limited.

# Results and discussion

To date, the algorithms and models for aggregating mass spectrometry features in larger groups than peptides have relied on protein identifications and do not use correlations across samples in any way. The model we have described makes use of these correlations to identify peptides that do indeed show consistent expression in addition to agreement in identification. We define a “dominant metaprotein" for a protein to be the metaprotein(s) that contains a majority of peptides from that protein. One of the features we have observed from studying posterior parameters from our model is that, in many cases, an identified isotope group (one with a protein label) does not follow the expression pattern of its corresponding dominant metaprotein. That is to say, for any given protein there is often a “consensus" expression pattern that many of the isotope groups from that protein follow, but that there is also a large minority of isotope groups which do not follow that expression pattern.

One of the strengths of our approach is the inherent modeling of uncertainty in all of the model parameters. Because we fit our model with MCMC, we obtain random draws from the posterior distribution of reasonable model parameters. Thus, rather than identifying fixed, specific sets of isotope groups which associate with each other, we obtain a list of possible groupings, all of which explain the data well.

The posterior parameters of greatest interest will depend on the specific application, but often we will be most interested in the vector of factor memberships, $z$, which describes which peptides seem to group together most often. In data sets intended for the generation of predictive models in clinical/translational studies (as well as other types of studies), we will be interested in $\Lambda$. The columns of this matrix define our estimates of the expression patterns of the metaproteins across our samples, and can be treated as independent variables in any type of model that is appropriate for the study.

## Features of the Factor Model

One of the strengths of our approach is the ability to collect isotope groups based not only on identifications, but also on their coexpression across samples. Associated with each meta-protein is a vector of factor scores, representing the expression of the meta-protein as well as a collection of isotope groups which make up that meta-protein. Figure 1 shows a heatmap of all of the peptides from the dataset that are identified as belonging to the protein Apo E. Note that, while the majority of those peptides share a common expression pattern, three (labeled 45, 31 and 53) show a very different, conflicting pattern. Our meta-protein model automatically groups the co-expressing peptides into the same factor. While assigning the peptides with conflicting patterns to other meta-proteins that more closely match their expression patterns. We fit our model to all 109 proteins which have more than 1 identified peptide in the data set. Heatmaps similar to Figure 1 for each of these proteins are available in the supplementary material. Examination of these figures shows that the presence of peptides that show expression patterns significantly different from their corresponding dominant metaprotein is the rule, rather than the exception.

There are a few reasonable explanations for this. The most obvious possible explanation is that the poorly conforming peptides are those with the lowest overall intensity, and therefore subject to smaller signal to noise ratios. However, examination of heatmaps showing the exact same peptides, but now sorted by mean intensity across the samples rather than by meta-protein membership demonstrates that there is not a strong predominance of low intensity peptides among those that do not coexpress with the other peptides from the protein. All 109 of those heatmaps are available in the supplementary material. We tested, using a non-parametric Kruskal-Wallis test, for association between meta-protein membership and mean signal intensity for each of the metaproteins, and found that, of the 109 proteins tested, only 2 showed significant association (p-value<0.01, APOB and CERU).

Another possible explanation for the presence of poorly co-expression peptides within a single protein is misalignment between runblocks. The data set was analyzed in three runblocks, one of which occurred months after the original two, and it is often the case that there are large shifts

in retention time between runblocks, particularly when there are large time intervals between. We expect this to be a rare occurrence, however. For example, if the accuracy of the alignment algorithm is 99%, we would expect around 34 examples of this (based on the fact that there are 3398 identified peptides in the dataset being used). Under this condition, however, we would expect to see a peptide that coexpresses well in two run batches but is mismatched in the third, and this is not generally the case. We are able to find some examples that fit this pattern, however, it is almost always the case that when a peptide does not share an expression pattern with the bulk of peptides from the same protein in one runblock, it also does not share that pattern in the other run blocks.

A third explanation for peptides that are uncorrelated is mis-identification. However, by the same logic presented above, we expect around 34 such misidentifications, assuming that identification is correct 99% of the time. In fact, examining the list of peptides that do not belong to their dominant metaprotein, we see that more than half fall into this category (1640 out of 3398). This is true dispite our prior distribution assigning a 500x greater likelihood of a peptide belonging to its dominant meta-protein than to a different metaprotein.

The last explanation for lack of coherence among the peptides in a protein is post-translational modification. If proteins are extensively and dynamically modified after translation, then we should expect many of them to exhibit expression patterns that do not match the bulk of peptides from that same protein. Also, if peptide modification is a significant contributor to observed patterns of expression, then we also expect to find peptides that have targets for post-translational modification to be more likely to be found outside their dominant meta-protein. We examined the probability of peptides containing Glutamine and Asparagine, which are known sites of deamidation, to belong to their dominant metaproteins. Correcting for the number of peptides inside and outside their dominant metaproteins, we find that peptides containing Glutamine are approximately 1.2 times more likely to not follow the dominant expression pattern for any given protein, and that this is a statistically significant difference (p-value .0013, fisher's exact test). In addition, peptides with Asparagine are 1.22 times more likely to fail to coexpress with the dominant group of peptides

from a protein (p-value 0.0010). In addition to these two sites of post-translational modification, we examined the motifs NxT and NxS, which are known sites of N-linked glycosylation. For these two motifs, 25 of the 30 peptides which contain the "NxT" motif and 53 of the 60 peptides which contain the "NxS" motif follow expression patterns that are different from their dominant metaproteins (odds ratios 4.3 and 6.6 respectively, p-values 0.0013 and 2.1e-8 respectively). Additionally, both Serine and and Threonine are known to be sites of O-linked glycosylation as well as phosphorylation. We find that both Threonine and Serine are also more likely to show odd expression patterns (Threonine: odds ratio 1.2, p-value .002 and Serine odds ratio 1.3, p-value 8.1e-6). We tested Proline (odds ratio=1, p=.96) and Histidine (odds ratio=1.1, p=.40) as negative controls.

Thus, peptides with any of these post-translational modification motifs are significantly less likely to follow the dominant expression pattern for the protein from which they are derived. This suggests that post-translational modification is a pervasive feature of plasma proteomics, and that protein level quantitation is likely to either introduce errors by summing across uncorrelated parts of a protein (if quantitation is accomplished through summation across all associated peptides) or to miss critical post-translational modifications (if quantitation is accomplished by summation of only the top three peptides). The ability of our model to identify and properly treat those peptides that fail to follow the dominant expression pattern of the associated protein is critically important.

## Comparison with Protein Level Quantitation

While there are similarities between our meta-protein model and various techniques for protein level quantitation, the ability to group peptides based on co-expression across all samples, and therefore identify peptides that show evidence of post-translational modification is a critical difference. Nonetheless, it is interesting to compare our model to protein level quantitation algorithms. For this purpose, we will examine two such algorithms. The summation algorithm estimates protein level quantitation by summing total expression across all peptides from a protein. This algorithm automatically gives peptides with high intensity measurements a larger effect on the estimated protein level quantitation. A second algorithm, Top 3, estimates the protein level expression

as the mean of the three peptides with the highest intensity (average across the samples). These two algorithms typically give similar results, however, an examination of the Figure 1 shows that this treatment can potentially introduce noise from peptides that don't share the consensus expression pattern.

Of the 87 subjects in this study, we have available antibody assay mesurements of both Apo B and Apo E on 38. We compared the two protein level quantitation algorithms and our metaprotein model to the antibody assay "gold standard". Correlations are all generally high and examination of Figure 2 shows that the three techniques are generally in agreement, even on outliers. The top three isotope groups identified as coming from Apo B show a high level of correlation, and all of these peptides are members of the main Apo B metaprotein. Also, a large majority of the Apo B peptides show this same expression pattern and are assigned to the same metaprotein, thus agreement between the three methods on Apo B is not surprising.

However, examination of the top three isotope groups from Apo E (Figure 1) shows a different picture. The second most abundant isotope group from Apo E is in a different metaprotein because it shows a substantially different expression pattern from the bulk of the Apo E isotope groups. In addition, if we delete the two outliers from the data (they are outliers by all three quantitation methods), the correlations between the three top Apo E isotope groups and the antibody assay of Apo E activity are .59, .23 and .56 respectively (p-values of .0002, .17 and .0004 respectively). Thus, this second most abundant isotope group should be adding noise to the Top 3 protein level estimate of Apo E. Interestingly, the correlation between the Top 3 estimate and the antibody assay is .60. This is higher than any of the three separately, which suggests that the antibody assay is in fact measuring an aggregation of two different forms of Apo E.

## Metaprotein Expression in a Hepatitis C Cohort

We obtained pre-treatment serum samples from 87 patients with Hepatitis C who have a known response or non-response to the standard of care treatment with interferon and Ribavirin. Serum from the patients was measured with open platform LC/MS/MS as described in the Experimental

section. The overall goal of the study was to predict who among the study subjects will respond to therapy and who will not. We are also interested in estimating which proteins and peptides are potential markers of response, allowing for future, targeted assay development.

Analysis of this data set proceeds in two steps. First, the model described above is fit to the proteomic data without regard to the phenotype of interest. There were a total of 6,729 peptides in the data set with either positive identifications or with average expression levels greater than the mean. Of these 3620 had identifications. These were matched with 265 different proteins of which 111 had two or more associated, identified peptides.

### Technical Variation

This data set was collected in three run blocks. Two of these were consecutive and the third was run months later. There are significant batch effects present in comparing the first two to the third even after correcting for observed total protein. Even though we did not include explicit design vectors for batch in our regression, our inclusion of a factor matrix descrbing systematic effects ($\beta$ and $H$) allows this source of technical variation to be automatically modeled without foreknowledge. We find that the first row of $H$ perfectly distinguishes runblocks 1 and 2 from runblock 3, with values between .95 and 1.05 for the former and between -.95 and -1.05 for the latter.

### Prediction of Outcome

As with the computation of protein level expression, our 111 metaproteins may be used in any context as independent predictor variables. Additionally, we may assess the level of association between metaproteins and other biological phenotypes in either case. In general, we find that the correlation between our metaprotein model and protein level quantitation for estimation is high when there are a large number of peptides from the given protein. For example, one of the most abundant proteins in this data set is Apolipoprotein B (represented by 409 isotope groups), and correlation between estimated expression from our factor model and from summation of all Apo B identified peptides is 0.97. However, when there are fewer peptides available or if there are

many misidentifications or modifications we find evidence that our factor model gives improved estimation of protein level expression patterns. For example, pregnancy zone protein, which has only three associated isotope groups in the data set, is known to be overexpressed in women compared to men. While both show differential expression, the factor model gives a p-value of 4.5e-4 (statistically significant even after Bonferroni correction for multiple hypotheses) as compared to a p-value of .0014 for estimation by summation over identified peptides (not significant after multiple hypothesis correction).

Our first step in the analysis of the posterior distribution involves comparing the mean metaprotein expression patterns for all 111 metaproteins to the "response to therapy" phenotype. We find three such metaproteins to be significantly associated (ANOVA p-value $= 9.8e-5$) even after correction for multiple hypothesis testing. Protein level quantitation by summation yields only 2 and protein quantitation by top 3 yields zero. Furthermore, the relevant expression pattern is far clearer for the metaprotein analysis. Figure 3 shows the most predictive “protein" (ZA2G), as computed by the summation protein quantitation algorithm. Examination of the metaprotein with the most peptides from ZA2G, built from our model, shows that ZA2G is indeed identified as predictive by the metaprotein model as well. However, our metaprotein model identifies only those peptides from ZA2G that are highly correlated with each other, and it additionally identifies a number of other peptides from other proteins that share the same expression pattern across the samples (also shown in Figure 3). The result is better separation between responders and non-responders.

In addition to strong associations for three metaproteins, we find that there are a total of 13 metaproteins with p-values less than .01 (random association would dictate only 1). Thus there is clear evidence of the presence of blood-born markers of response to therapy in Hepatitis C.

**Identification of Candidate Peptides**

We would like to identify a set of candidate peptides for use in future targeted studies such as multiple reaction monitoring (MRM) or antibody studies. From the analysis of associations between averaged metaproteins and the phenotype, we are confident that there are markers of interest. How-

ever, analysis of summarized posterior parameter estimates is unsuited to the discovery of which markers are most relevant. In particular, as peptides are added and dropped from metaproteins, posterior summaries average the metaprotein makeup as well as the metaprotein expression pattern. However, it is unclear, and even unlikely, that the average metaprotein expression pattern would result from estimating the aggregate expression of the average peptides in that metaprotein. Thus, if we were to base our choice of candidate peptides on these averages, we should expect to need to recompute the metaprotein expression patterns.

We instead propose to obtain draws from the MCMC chain, and for each draw build a predictor and observe which peptides are included in that predictor. In this way, the values of $\Lambda$ are computed directly from the peptides included in the corresponding metaprotein. Additionally, by keeping track of which peptides are most often included in the predictors, we obtain a list of candidate biomarkers for future study.

Because we have 111 metaproteins but only 87 samples, direct regression is not possible in this context. We instead use variable selection with model averaging. Variable selection allows regression with a small subset of the total number of predictors, while model averaging allows us to properly account for uncertainty in which models are correct. In this context, model averaging has been shown to outperform the single best model for predictive accuracy on hold out data sets.[18] We use a publicly available implementation of variable selection and model averaging called Shotgun Stochastic Search.[19]

As mentioned above, our analysis consists of two steps. The first is the generation of a factor model to explain the variation seen in the peptide concentration data. This step is unsupervised; it does not take into account any phenotype data in any way. The variable selection and model averaging just described constitutes the second, supervised step. Results from this analysis will vary slightly at each step of the MCMC chain. This allows us to estimate both the accuracy of predictors generated by this model as well as the uncertainty in that accuracy. Figure 4 shows receiver operating curves (ROC) of the model for 200 steps of the MCMC, and compares this to the same ROC's from predicting randomly generated phenotypes with the same number of cases

and controls. We see that the accuracy is significantly better than chance for all 200 draws from the Markov chain.

The collection of most used isotope groups in this modeling approach are show in Figure 5. We show the locations of the associated peptides (those with identifications) in their respective proteins. We note that THBG has a region showing a statistically significant dearth of identified peptides in the list (Kolmogorov-Smirnov test). This potentially indicates alternatively spliced proteins or proteins with post-translational modifications which are associated with disease outcome. Verification of the statistical significance of the peptides for patient differention based on treatment response is underway with additional sample cohorts, using both open platform and MRM based quantitation.

# Conclusions

We have outlined a model for the accurate assessment of protein level expression data from open platform proteomic data. The model offers an approach to aggregation of mass spectrometry proteomics data that differs significantly from protein level quantitation, and that offers significant advantages over previous approaches. We have shown that it can be used to gain significant insights into translational studies, and have exemplified its use with a study of response to therapy in patients with Hepatitis C.

## Acknowledgement

Supported in part by Duke University's CTSA grant 1 UL1 RR024128-01 from NCRR/NIH. Supported in part by a gift from David H. Murdock. We gratefully acknowledge Waters Corporation and Rosetta Biosoftware, Inc for hardware and software support for the data presented in this manuscript.

## Supporting Information Available

Supplementary data on protein and peptide identifications has been uploaded according to the MI-APE standard to the Proteomics Identification Database (http://www.ebi.ac.uk/pride/), accession numbers xxx (in progress). Additionally, Matlab software for fitting the model described above is available at: xxx (in progress).

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

## References


(1) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D.; Steen, H.; Pandey, A.; Mann, M. *Molecular and Cellular Proteomics* **2002**, *1*, 376–386.

(2) Wiese, S.; Reidegeld, K.; Meyer, H.; Warscheid, B. *Proteomics* **2007**, *7*, 340–350.

(3) Gygi, S.; Rist, B.; Gerber, S.; Turecek, F.; Gelb, M.; Aebersold, R. *Nat Biotechnol* **1999**, *17*, 994–999.

(4) Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T.; Hill, L.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. *Analytical Chemsistry* **2003**, *75*, 4818–4826.

(5) Kito, K.; Ito, T. *Curr Genomics* **2008**, *9*, 263–274.

(6) Wang, G.; Wu, W.; Zeng, W.; Chou, C.-L.; Shen, R.-F. *Journal of Proteome Research* **2006**, *5*, 1214–1243.

(7) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal Chem* **2003**, *75*, 4646–4658.

(8) Silva, J. C.; Denny, R.; Dorschel, C.; Gorenstein, M. V.; Li, G.-Z.; Richardson, K.; Wall, D.; Geromanos, S. J. *Molecular and Cellular Proteomics* **2006**, *5*, 598–607.

(9) Silva, J. C.; Gorenstein, G.-Z., M. V.; Li; Vissers, J. P. C.; Geromanos, S. J. *Molecular and Cellular Proteomics* **2006**, *5*, 144–156.

(10) Daly, D. S.; Anderson, K. K.; Panisko, E. A.; Purvine, S. O.; Fang, R.; Monroe, M. E.; Baker, S. E. *Journal of Proteome Research* **2008**, *7*, 1209–1217.

(11) Brusniak, M.-Y.; Bodenmiller, B.; Campbell, D.; Cooke, K.; Eddes, J.; Garbutt, A.; Lau, H.; Letarte, S.; mueller, L. N.; Sharma, V.; Vitek, O.; Zhang, N.; Aebersold, R.; Watts, J. D. *BMC Bioinformatics* **2008**, *9*.

(12) Karpievitch, Y.; Stanley, J.; Taverner, T.; Huang, J.; Adkins, J. N.; Ansong, C.; Heffron, F.; Metz, T. O.; Qian, W.-J.; Yoon, H.; Smith, R. D.; Dabney, A. R. *Bioinformatics* **2009**, *25*, 2028–2034.

(13) Keller, A.; Nesvizhskii, A.; Kolker, E.; Aebersold, R. *Anal Chem* **2002**, *75*, 4646–4658.

(14) Carvalho, C.; Chang, J.; Lucas, J.; Nevins, J.; Wang, Q.-L.; West, M. *Journal of the American Statistical Association* **2008**, *103*, 1438–1456.

(15) Lucas, J.; Carvalho, C.; Wang, Q.; Bild, A.; Nevins, J.; West, M. In *Bayesian Inference for Gene Expression and Proteomics*; Vannucci, M., Do, K.-A., Müller, P., Eds.; Cambridge University Press, 2006; pp 155–176.

(16) Lucas, J.; Carvalho, C.; Chen, J.-Y.; Chi, J.-T.; West, M. *PLoS One*

(17) Lucas, J.; Carvalho, C.; West, M. *Statistical Applications in Genetics and Molecular Biology*

(18) Raftery, A.; Madigan, D.; Hoeting, J. *Journal of the American Statistical Association* **1997**, *92*, 191–197.

(19) Hans, C.; Dobra, A.; West, M. *Journal of the American Statistical Association* **2007**, *102*, 507–516.

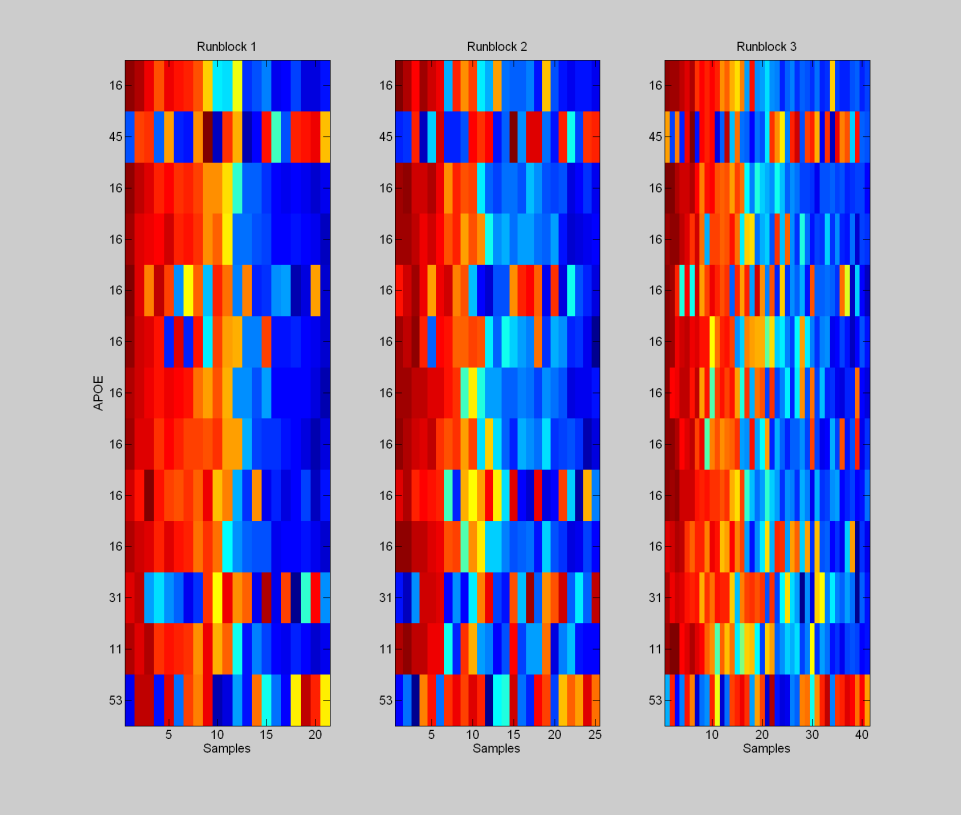


Figure 1: A heatmap of all of the isotope groups in the data set that are identified as originating from the protein Apoliprotein E. The labeling on the y-axis indicates which metaprotein that peptide was assigned to. The peptides are ordered from top to bottom from highest to lowest mean intensity across the samples. Note that, while the majority of those peptides share a common expression pattern (those assigned to metaprotein 16), three show a very different, conflicting pattern. Our meta-protein model automatically groups the co-expressing peptides into the same factor. While assigning the peptides with conflicting patterns to meta-proteins that more closely match their expression patterns.

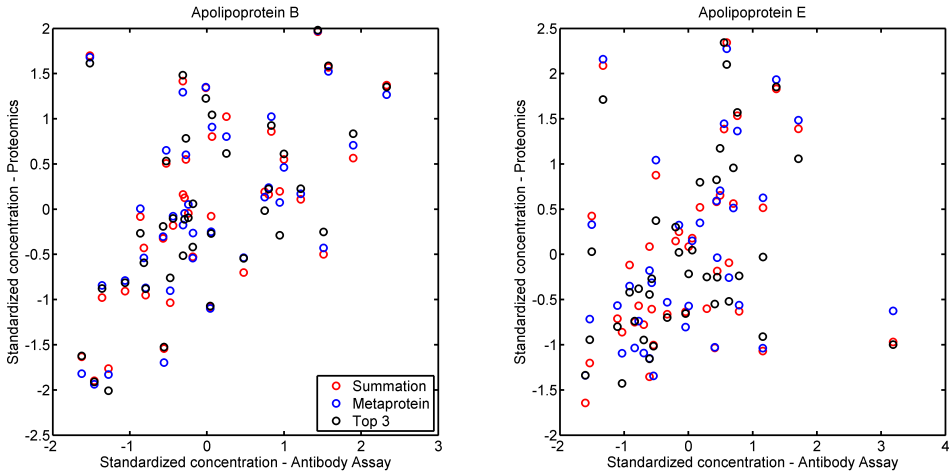


Figure 2: A comparison between the antibody assay estimate of the protein level expression patterns for Apo B (panel A) and Apo E (panel B) (x-axes in both figures), and the LC-MS/MS estimates using the summation, top 3 and metaprotein methods for protein level expression. We note that the three algorithms are in agreement to a high degree. Correlations for Apo B are .55, .54 and .56 for Metaprotein, Summation and Top 3 models respectively and they are .31, .28 and .32 for respectively for Apo E. Note that there is an outlier in each of the top left and bottom right of the Apo E figure, and without these two outliers, correlations are just under .6 for all three methods.

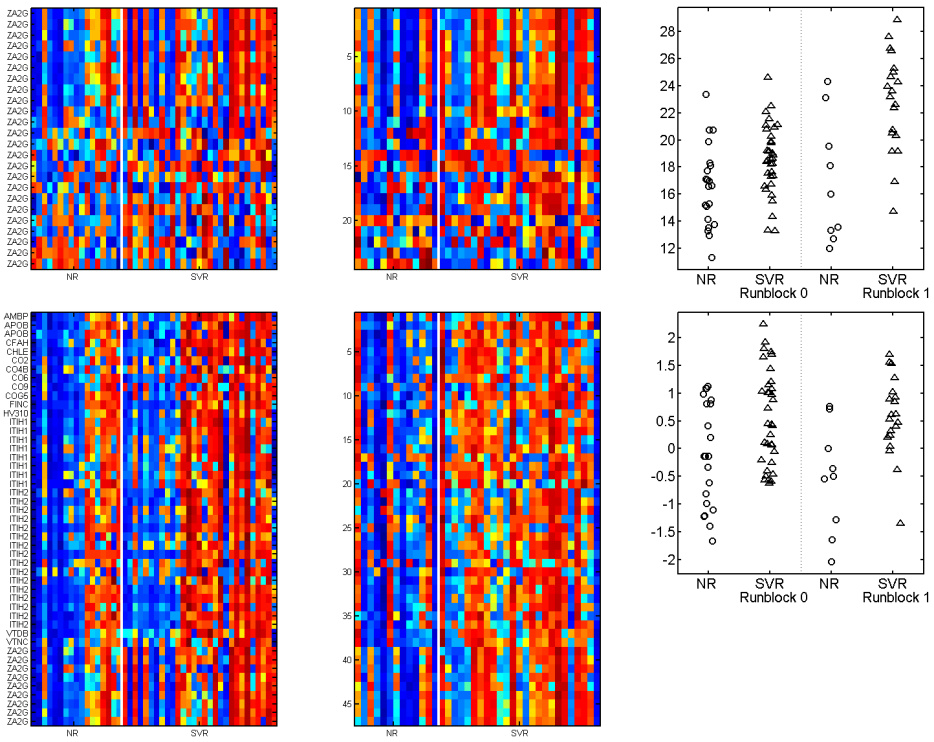


Figure 3: Comparison of metaprotein and summation approaches to aggregation of large numbers of isotope groups. Panels A (runblock 1) and B (runblock 2) show all of the isotope groups identified as coming from the protein ZA2G. This is one of the two proteins significantly associated with the response to therapy phenotype in patients with Hepatitis C. Samples are on the x-axis and isotope groups are on the y-axis. Panels D and E show the same type of heatmap, but now for the metaprotein containing the largest numbers of isotope groups from ZA2G. Notice that there are a number of isotope groups from other proteins that are highly correlated with the included ZA2G isotope groups and are therefore included in this metaprotein. Also, as can be seen in A and B, correlation of individual isotope groups from the ZA2G protein is high for about half of the isotope groups, but quite poor for many of the others. Panel C shows the predictive performance of the summation algorithm of "protein" level quantitation for ZA2G, split by runblock, and panel F shows the same for the metaprotein. Note that the metaprotein shows better performance and that the performance of the metaprotein is more consistent across the two separate runblocks. Performance of the top 3 algorithm for protein level quantitation is not shown because it is not statistically significantly associated with response to therapy.

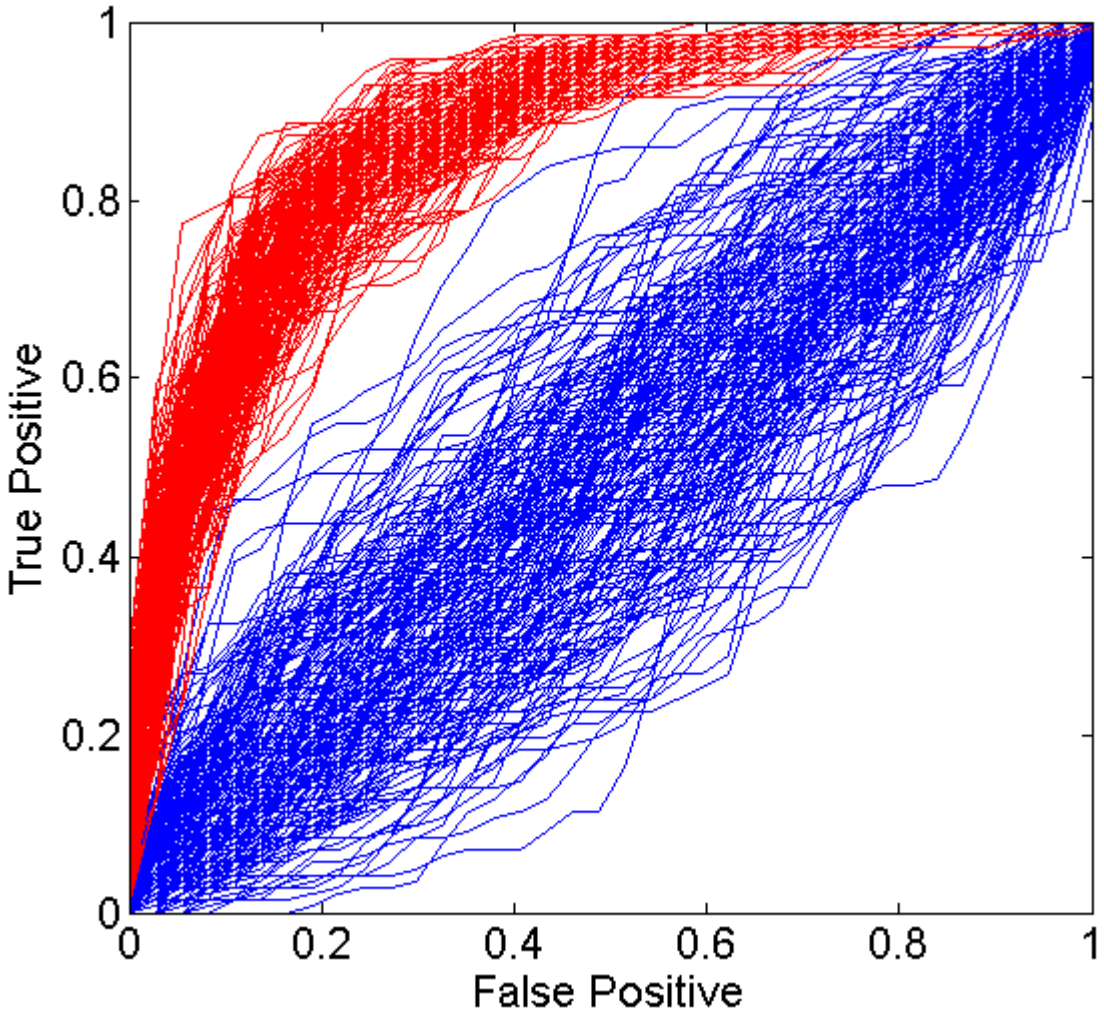


Figure 4: Receiver operating curves generated from 200 draws from the MCMC chain (red) compared to ROC's generated from predictors of randomly generated outcomes (blue). The random predictors were generated from a mean zero multivariate normal distribution using a covariance matrix generated from the factor matrix $\Lambda$ for the corresponding MCMC draw. The outcome variable (response/non-response) was then permuted before the 'random predictor' model was fit.

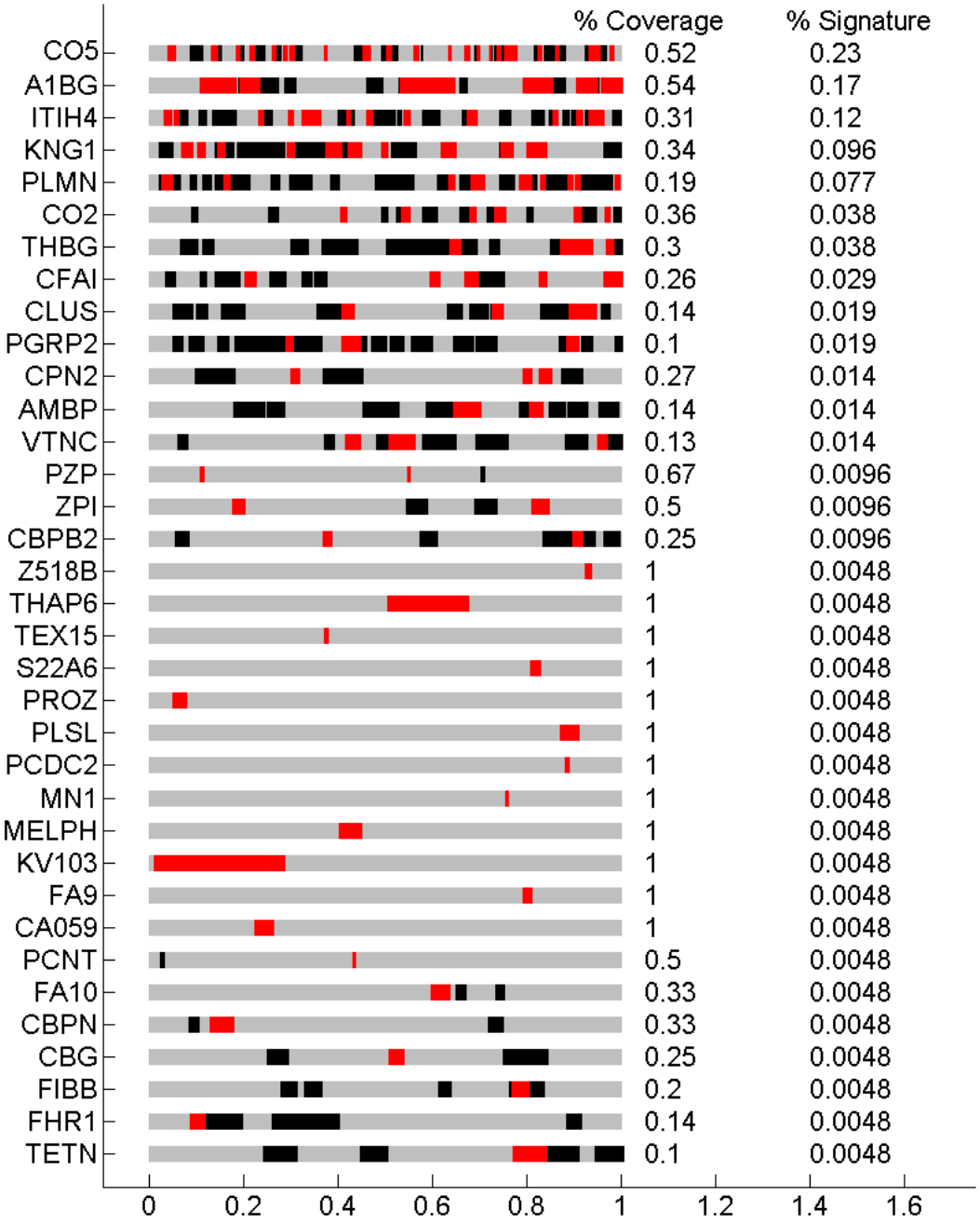


Figure 5: Shows the locations of the 600 most used peptides which were also identified. Proteins with fewer than 1% of their peptides represented in the figure were filtered out. We note that THBG has a region showing a statistically significant dearth of identified peptides in the list (Kolmogorov-Smirnov).