UNIVERSITY OF CALIFORNIA SAN DIEGO

Penalty-Based Dynamic Programming for the Identification of Post-Translational Modifications in Peptide Mass Spectra

A dissertation submitted in partial satisfaction of

the requirements for the degree of

Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Laurence E. Bernstein

Committee in charge:

Professor Nuno Bandeira, Chair
Professor Steven Briggs, Co-Chair
Professor Vineet Bafna
Professor Peter Dorrestein
Professor Pavel Pevzner

The Dissertation of Laurence E. Bernstein is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California San Diego

2018

DEDICATION

To my father,

who instilled in me the love of math and science,

the life-long desire to ask the hard questions,

and the sense of humor to deal with it all.

EPIGRAPH

Computer science is to biology what calculus is to physics. It's the natural mathematical technique that best maps the character of the subject.

— Harold Morowitz

People think that computer science is the art of geniuses but the actual reality is the opposite, just many people doing things that build on each other, like a wall of mini stones.

— Donald Knuth

Research is what I'm doing when I don't know what I'm doing.

— Wernher von Braun

The good thing about science is that it's true whether or not you believe in it.

— Neil deGrasse Tyson

Nothing is foolproof. Fools are so ingenious.
— Unknown

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Dr. Pavel Pevzner for starting me on this path and Dr. Nuno Bandeira for helping me reach the end of it.

My thanks to the other members of my committee: Dr. Vineet Bafna, Dr. Steven Briggs, and Dr. Pieter Dorrestein for giving me support and sage advice along the journey.

My thanks to all the members of the bioinformatics group, particularly those that came before me and showed me the way: Dr. Stefano Bonissone, Dr. Natalie Castellana, Dr. Adrian Guthals, Dr. Sangtae Kim, and June Snedecor.

A special thanks to Dr. Nuno Bandeira, Dr. Adrian Guthals, Dr. Sangtae Kim, and Dr. Seungjin Na, upon whose work mine is largely based and without which mine would not be possible.

My thanks to Jeremy Carver and Julie Wertz, whose help developing and debugging the RaVen workflows was invaluable.

My thanks to my mother Judith, step-father Daniel, and my siblings: Robert, Arthur and Maya, who have always supported me in ways too numerous to mention.

Thanks to my good friends who helped me stay sane through it all.

And last, but the opposite of least, to my beautiful and brilliant wife Pia, whose sacrifices great and small have made this, and so many other things, possible.

Chapter 1 in full has been submitted for publication to *Proceedings of the National Academy of Sciences of the United States* under the title "Spectral networks identification of rare post-translational modifications and hypermodified regions in lens crystallin proteins" Bernstein, Laurence E., Julie Wertz, Seungin Na, and Bandeira, Nuno. 2018. The dissertation author is the primary author of this paper.

Chapter 2 in full has been submitted for publication to *Frontiers in Genetics* under the title "Discovery of post-translational modifications and proteomics diversity in colorectal cancer." Bernstein, Laurence E., and Bandeira, N. 2018. The dissertation author is the primary author of this paper

VITA

1998 Bachelor of Arts in Computer Science,
University of California San Diego

2018 Doctor of Philosophy in Bioinformatics and Systems Biology,
University of California San Diego

PUBLICATIONS

2018 Bernstein, Laurence E., Julie Wertz, Seungin Na, and Bandeira, Nuno. 2018. "Spectral Networks Identification of Rare Post-Translational Modifications and Hypermodified Regions in Lens Crystallin Proteins.", submitted to *Proceedings of the National Academy of Sciences of the United States*

2018 Bernstein, Laurence E., and Bandeira, N. 2018. "Discovery of post-translational modifications and proteomics diversity in colorectal cancer." submitted to *Frontiers in Genetics*

CONFERENCE PROCEEDINGS

2016 Bernstein, L., Na S. & Bandeira N. (2016). *Enabling Massive Blind Penalty Search using Multiple Enzyme Proteomics*, Proceedings of the 64th ASMS Annual Conference on Mass Spectrometry and Allied Topics, San Antonio, Texas, June 5-9, 2016

2015 Bernstein, L. & Bandeira N. (2015). *Peptide Variant Discovery in Lens Tissue using Penalty Based Spectral Alignment* , Proceedings of the 63rd ASMS Annual Conference on Mass Spectrometry and Allied Topics, St Louis, Missouri, May 31 – June 4, 2015

2014 Bernstein, L. & Bandeira N. (2014). *Blind Spectral Alignment with Adaptive Penalties*, Proceedings of the 62nd ASMS 62nd Annual Conference on Mass Spectrometry and Allied Topics, Baltimore, Maryland, June 5-9 2014

ABSTRACT OF THE DISSERTATION

Penalty-Based Dynamic Programming for the Identification of Post-Translational Modifications in Peptide Mass Spectra

by

Laurence E. Bernstein

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2018

Professor Nuno Bandeira, Chair

Professor Steven Briggs, Co-Chair

Tandem mass spectrometry (MS/MS) has long been the leading method of identifying peptides and proteins in complex biological samples and many algorithms have been created for this purpose. Many of the methods for searching MS/MS spectra against a database of known proteins must restrict the number of post-translational modifications (PTMs) that they can identify because the larger the number of PTMs being considered, the larger the search space,

which in turn increases both computational complexity and the potential for false matches. In addition these algorithms cannot discover new peptides or homologues or be used with species for which a protein database does not exist. Newer algorithms have been developed that perform "open" or "blind" searches capable of finding any possible modifications, however these methods increase the search space even further, often resulting in lower performance and the generation of many putative modification masses that must be sifted through manually to determine which are real.

To address the shortcomings of the existing methods, we created a new blind database search algorithm based on spectral networks. Our method uses a modification of the standard spectral tagging filtration techniques tailored for contig-consensus spectra generated from spectral networks, along with, the first of its kind, penalty-based, dynamic programming spectrum-database alignment algorithm that is able to accurately to identify both a priori specified modifications as well as novel PTMs. We then developed a workflow based on these new techniques that combines previous work in clustering, spectral alignment, spectral networks, and multi-spectral assembly. Because our new algorithm only identifies spectra that lie within the spectral networks, we created a workflow, called RaVen, that merged our method with MS-GF+ and combines the results from both methods resulting in a method with massive improvement in overall identification rates above existing methods while at the same time identifying many more rare modifications in samples. We also propose an improved way of measuring the accuracy of blind search algorithms: "peptide variants" which better meet captures the goals of blind search methods and does not rely on precise localization of modifications (which is very difficult to achieve for most algorithms).

CHAPTER 1 - Spectral networks identification of rare post-translational modifications and hypermodified regions in lens crystallin proteins

1.1. Abstract

Post-translational modifications and sequence cleavage events are key to protein-level regulation of biological processes and have been repeatedly implicated in disease phenotypes. Such events have potentially even greater relevance in the case of long-lived proteins in cellular contexts with little or no protein turnover, where long-term accumulation of protein damage can degrade protein function, as is the case with crystallin proteins in cataract lens. Building on high-throughput tandem mass spectrometry for the analysis of proteomics samples, various algorithms have been developed for the detection of unexpected modifications; however, current approaches still estimate the significance of detected modifications primarily by their frequency of occurrence on multiple peptides. This approach is generally effective for the detection of widespread artifacts and sample-handling modifications, but has limited ability to detect rarer modifications which are potentially more likely to have biological significance. Our proposed RaVen approach assesses the significance of detected modifications using spectral networks algorithms to match spectra of peptides with different modification states, thereby substantially increasing the detectability of modifications with strong signal even if they occur on only a single site in the whole proteome. Using this approach, RaVen not only detects nearly all modifications reported by previous algorithms, but further detects over 60 known and putatively novel rare modifications in lens crystallins spanning both chemical and enzymatic post-translational modifications, as well as several putative sequence polymorphisms. Furthermore,

by capitalizing on the detection and consensus interpretation of multiple modified variants of peptides with overlapping sequences, RaVen detects modifications on over two thirds of all crystallin peptides and reveals for the first time the occurrence of hypermodified protein regions covered by nearly three hundred distinct peptide sequence and modification variants. All data and search results were deposited in MassIVE (MSV000082143) and are available via ProteomeXchange with identifier PXD009167.

1.2. Introduction

Mass spectrometry (MS) has, for some time, been the leading method for high-throughput identification of proteins in complex biological samples and many algorithms have been developed for this task including: Sequest [1], Mascot [2] and MS-GF+ [3]. Such `restricted' search algorithms have made significant progress in the ability to identify proteins and peptides with at most a few common post-translational modifications (PTMs), but challenges still remain for sensitive identification of PTMs, especially when considering a large selection of possible PTMs . As a result, biological inquiries of proteomics data typically restrict the number of searched modifications to five or fewer possibilities, even though hundreds of possible variations are documented in Unimod [4]. All these algorithms must perform the same task of comparing spectra gathered by mass spectrometry experiments to a database of known proteins in order to find the best match. However the databases themselves have the limitation that they normally only contain unmodified protein sequences and not the many various forms of the protein that can possibly result from post-translational modifications. Because of this, these algorithms all suffer from two serious problems. First, as the number of PTMs being considered

increases, the computational complexity grows rapidly (and often exponentially). Second, as the search space grows, either by increase in the number of modifications allowed or increase in the size of the database, the potential for false matches rises significantly, resulting in far fewer identifications.

This problem becomes even more pronounced when we expand the search to unknown modifications. Such "open" searches expand the database by ten thousand times for a single modification or one million times for multiple modifications. As such, a range of algorithms such as MS-Alignment[5][6], OpenSea [7], TagRecon [8] and MODa [9] have been proposed for spectrum identification in the presence of unexpected modifications. More recently Chick et al. [10] extended Proteome Discoverer to allow open database searches by vastly increasing precursor ion tolerance. By using a large ion tolerance, the open search technique is able to identify unknown modification masses. Kong et al. improved on this approach with MSFragger [11] greatly increasing the speed and thereby making it more practical. However both MODa and the open search method generate large sets of putative modification masses. Most of these masses have very low occurrence rates, resulting in the necessity for manual examination to determine those that reflect true PTMs. An algorithm that can handle the increased virtual database size as well as allow for blind PTM searching is highly desirable for many current datasets, such as cancer or antibody sets where there are many differences from the nominal protein database, as well as for sets where the true protein database of interest is not known and only a homologous database is available. Our RaVen workflow (Figure 1.1) approaches the problem of open modification discovery from a different perspective. First, RaVen capitalizes on the principle that peptides with unexpected modifications are nearly always co-expressed with either unmodified or commonly-modified versions of the same peptides. As such, similar to

alignment of sequencing reads to detect related or overlapping sequences, RaVen builds on alignment between spectra [12] [13] to detect related variations of peptide sequences or post-translational modifications. Second, RaVen combines pairwise spectral alignments into multi-aligned protein `contigs' [14] and leverages the much-increased signal-to-noise ratio in contig-consensus sequences to derive de novo sequence tags used to filter the database search space, a novel multi-spectrum extension of the tag filtering approach initially proposed by Mann and Wilm [15] and later incorporated into multiple search tools [9][16][7][8]. Third, RaVen uses spectral alignment frequencies of putative modification masses to score hypothetical variations (e.g., PTMs, polymorphisms, etc.) only on sequences selected by high-scoring contig matches. RaVen thus utilizes spectral alignment to address two major challenges in the detection of unexpected modifications: a) it builds on correlated fragmentation of related peptides to construct contig spectra with longer sequence coverage and much higher signal-to-noise ratios and b) it considers unexpected modifications only at positions in the database where the sequence matches the high-quality de novo sequences derived from spectral contigs. Finally, RaVen's shifted focus towards peptide `variants' (differently-modified or polymorphic versions of a peptide sequence) and protein variant regions (PVRs) advances the expectations of open modification searches from the traditional detection of common (typically sample-handling) modifications, into a more detailed assessment of the occurrence of variations on specific peptides and proteins. By further combining this detection with the analysis of differential expression across multiple conditions, we show how even infrequent unexpected variations (which are typically lost in traditional frequency-based approaches) can be explained as known-but-unexpected modifications with substantial changes in expression across healthy and disease states.

1.3. Results

Lens proteins constitute an ideal type of sample for blind search algorithms for detection of post-translational modifications. As is well known from the literature [17], lens crystallins do not turn over but rather continuously accumulate damage and modifications over an individual's life span. Several of these modifications have been implicated in disease phenotypes such as opaque lens and impaired vision, commonly referred to as cataracts. In addition to the relevance to phenotypes of protein-related disease, lens tissue is also ideal for this study from a technical perspective because the relative simplicity of the lens proteome (>95% abundance concentrated in only <15 proteins) makes extensive characterization of proteome diversity much more accessible to mass spectrometry instruments (i.e., more scanning events available to detect biological diversity of the same proteins, rather than superficially characterizing many more proteins). Finally, the specific lens dataset analyzed here covers a range of ages (3 days to 93 years old) and phenotypes (healthy or cataracts), and has already been studied in various previous publications [9][18][19][20][21], thereby defining a comprehensive point of reference against which to compare new results, as well as establishing a high bar for new algorithms aiming to detect novelty in proteomics samples.

The performance of RaVen was first assessed in comparison with state-of-the-art approaches for i) restricted database search for unmodified or commonly-modified peptides (MS-GF+ [3]) and ii) blind modification search for peptides with unexpected modifications (MODa [9]). MS-GF+ has been shown to match or outperform a range of algorithms for peptide and protein identification over a wide variety of sample types and collection methods [3] and supports only user-specified modifications. Also, MODa is a leading advanced multi-blind

search algorithm whose performance remains competitive with recent approaches (MSFragger [11]) while at the same time supporting multi-blind searches for peptides with two or more unexpected modifications and having already been tested on the same Lens dataset reanalyzed here. Open search methods, such as MSFragger, are also capable of finding unexpected modifications; however according to the authors of MSFragger, modification masses discovered by their method and MODa are very similar and MODa recovered a greater number of PSMs when run in semi-tryptic mode (as was done here). As illustrated in Figure 1.2a, RaVen identified 25,242 clustered spectra whereas MS-GF+ identified 15,183 spectra and MODa identified only 11,556 spectra. This corresponds to gain of 66% and 118%, respectively. The overlap in identified spectra further reveal that RaVen results included 95% of all identifications made by MS-GF+ and 86% of all MODa identifications, as well as show that RaVen identifies 34% more spectra than MS-GF+ and MODa combined. In addition, RaVen identifies 9054 spectra that were not identified by either of the other two algorithms, which is greater than 5 times the number of MODa-only identifications and nearly 11 times the number MS-GF+-only identifications.

To better assess the quality and significance of the peptide identifications resulting from the spectrum identifications, we define a peptide `variant' as a tuple (P,m) where P is a peptide sequence in the sample and m is the sum of all modification masses incident on P for a specific spectrum identification. Intuitively, a peptide variant corresponds to one modification form of a unique peptide sequence with a set of modifications resulting in a unique sum of modification masses. Since the putative biological relevance of blind search identifications is primarily proportional to the number of unique modified peptide forms, we correspondingly compared the relevance of identification results by enforcing variant-level false discovery rate (FDR), which is

a stricter and more accurate way to enforce 1% FDR for blind search results than the traditional 1% FDR at the level of peptide-spectrum matches (PSMs). As shown in Figure 1.2, RaVen identified 78% more variants than MS-GF+, which is not surprising since restricted searches force users to determine the set of modifications independently of the sample and are thus typically configured to search only for common modifications. However, even though MODa is not subject to the same limitations as MS-GF+ and supports multi-blind identification of multiple unexpected modifications per peptide, RaVen also identified 75% more variants than MODa. Overall, RaVen identified over 85% of all variants found by MS-GF+ and 76% of variants found by MODa, while also yielding over 4.5 times as many RaVen-only unique variants as MODa-only and over 12 times as many as MS-GF+-only unique variants.

1.4. Discovery of modifications

The assessment of putative modification masses selected by blind modification searches is typically done by counting the number of PSMs identified with a specific mass offset on a specific amino acid, and these are typically reported as an offset frequency table [6] with one column per amino acid (plus columns for N/C-termini) and one row per mass offset with cell values reporting the number of PSMs identified for each pair (Table 1.1). While this is a common way to assess the significance of putative modification masses resulting from mass offsets in identifications from blind searches, offset frequency tables have disadvantages that limit their utility for the analysis of putative modification masses. For example, Table 1.2 shows that using this simple procedure to construct the offset frequency table results in hundreds of non-zero entries with nearly every mass offset being assigned to multiple putative sites. First, it

should be noted that the number of PSMs for a peptide variant is not a direct indicator of its correctness or statistical significance; while there may be indirect correlations (e.g., more PSMs per peptide indicate higher abundance and may thus correlate with better signal-to-noise ratios), it is also possible for incorrect identifications to have high PSM counts as long as the search tool is consistent in how it assigns identifications to multiple MS/MS spectrum acquisitions from the same LC/MS precursor. To avoid the effect of this abundance-induced bias and instead assess the frequency of mass offsets using distinct observations of unique peptide sequences, RaVen constructs offset frequency tables using unique variant counts instead of PSM counts. Second, blind search algorithms have no prior knowledge of which sites are valid or invalid for unexpected modifications and thus tend to assign mass offsets to spurious sites when the spectra are not informative enough to precisely localize the site of the mass offset (which is very frequently the case). As a result, offset frequency tables often have high counts for correct pairs of sites and offset masses but these are almost always accompanied by many other incorrect site localizations where the lack of spectrum fragments resulted in the same mass offsets also being localized to other amino acids near the correct sites [21]. RaVen addresses this problem by i) assigning partial credit to all amino acids in spectrum mass ranges where there are no localizing peaks and ii) iteratively selecting the highest-scoring sites and correspondingly removing support for neighboring sites from variants explained by previously selected site and mass offset pairs (see Supplemental Methods and Table 1.2). Using this refined offset frequency table and requiring a minimum of 10 variants to consider a mass offset as a putative modification (see Table 1.3), we found that all but one discovered modification masses were correct (the single exception was a +17 Da mass due to a 13C error on oxidation) and that RaVen identifications included over 80% of all distinct modifications previously reported using MODa searches [9], as

well as detecting additional modifications that were previously missed (homoserine lactone on methionine and a carbon artefact on tryptophan [22].

Detection of rare modifications is usually difficult in blind searches because their dependency on offset frequency tables biases the analysis towards high-frequency modifications, which mostly tend to be sample handling modifications of no biological relevance, as is illustrated in Table 1.3 by the detection of 17 artefactual modifications for at most 8 post-translational modifications (PTMs). But in contrast to this methods-induced bias towards high frequency (and mostly artefactual) modifications, there is actually no a priori requirement for the functional significance of biological modifications to correspond to high frequency of occurrence on many distinct peptide sequences. In fact, the opposite may be closer to the truth as precise regulation of protein function would more likely bias selection towards enzymes that modify only specific targets and only in the right cellular context. To address this challenge, RaVen uses a unique feature of spectral networks instead of offset frequency table for the detection of rare modifications (i.e., occurring on <10 distinct variants). Using the statistical significance of spectrum/spectrum alignments [12] between spectra of peptides differing by the modification mass (see Figure 1.3), RaVen detects a total of 112 putative modification masses. As illustrated in Figure 1.4, 15% of these are masses attributable to common modifications such as those in Table 1.3 and 33% are the masses attributable to rare modifications found in Table 1.4. The bulk of the remaining identified masses fall into two categories: masses that are off by one Dalton from true modification masses due to carbon isotopes (20%) and masses which are actually a combination of two modification masses which could not be separated due to missing peaks in the data (29%). The remaining two masses (2%) were manually confirmed to correspond to correct identifications but could not be attributed to any combination of known modifications –

these are therefore reported as “Undetermined” and could correspond to new modifications. In total 80% of all masses identified with our technique were verifiable as correct (either individual or combined modifications) while 20% were off by one dalton due to carbon isotopes. By contrast, the MODa offset frequency table contains nearly 1000 cells with non-zero counts and still does not capture all the modifications discovered by RaVen. Moreover, one would have to inspect every one of those cells in order to capture as many of our modifications as possible since our spectral alignment approach discovers multiple modifications with a single representative PSM. Using spectral networks for selection of modifications, RaVen further identifies an additional 68 unique modifications and polymorphisms Table 1.4, over twice as many as could be detected by relying solely on the offset frequency table. As expected, the set of rare modifications includes few artifacts (only 26% of all rare modifications) and is instead enriched for chemical derivatives and amino acid substitutions. Together with the common modifications detected by RaVen, these additional modifications include all those previously reported using MODa [9] as well as glycosylation, trioxidation, malonylation, maleimide, acetyladehyde and 18 different types of substitutions. Finally, RaVen detects four uncategorized modifications that are well supported by the data (see Table 1.5).

1.5. Detection of modified peptides and hypervariable protein regions

Beyond the traditional analysis of distinct types of modifications detected by blind search, RaVen’s analysis of peptide variants as highly-correlated spectra grouped into spectral networks (see Figure 1.5 for an example) further focuses the analysis of results on how the same peptides or protein regions may be modified with different types or combinations of

modifications. In particular, RaVen analysis revealed that the Lysine N-term on the βS-crystallin peptide “KPIDWGAASPAVQSFR“ has the most distinct modifications of any amino acid in the sample. Table 1.6 lists the 15 different modifications found at this position, including the common modifications methylation, acetylation and carboxyethyl as well as less-common modifications such as formylation, carbamidomethyl, carboxymethyl, pyridylacetyl and maleimide. Table 1.6 further shows that the αA-crystallin peptide ”QWHLEGSFPVLATEPPK” was found to be modified with 21 different types of modifications, the highest number of all peptides in our sample, including both common PTMS such as carbamylation, pyro-glutamate, phosphorylation and sodium adduct, as well as rare modifications like metylsufonethyl, ethoxyformyl and potassium adducts.

Since protein activity, interactions and subcellular localization can be regulated by both modifications and cleavage events, it is important to characterize protein-level diversity by aggregating all peptide variants with overlapping sequences. RaVen thus defines a Protein Variant Region (PVR) as a region where each peptide overlaps with at least one other peptide mapped to the same region by at least 50% of its amino acids. While most PVRs identified by RaVen are composed of <10 variants, RaVen’s analysis of the diversity of protein-level variation further revealed many hypervariable protein regions with dozens of peptide variants (Table 1.8 and Figure 1.7), including one region on CRYAA with an unprecedented 256 distinct variants (Figure 1.6). The dramatic variability of CRYAA is also shown in Figure 1.8, illustrating the 202 different modifications identified by RaVen in the hundreds of variants covering regions of this protein. While there appears to be some correlation between the coverage of protein regions and the number of modifications detected in the same regions, the most variant portions of the protein were not the amino acids with the highest coverage and the diversity of modifications

found on distinct amino acids appears to be more correlated with location on the protein than with amino acid type; for example specific Lysines and Arginines are observed with up to six distinct modifications even though the vast majority were mostly unmodified. Altogether, RaVen identified 106 hypervariable protein regions with 10+ variants on 23 proteins.

1.6. Cohorts

The RaVen workflow integrates identification of variants with visualization of expression across multiple samples or cohorts by mapping relative abundances of variants to pie charts on spectral networks nodes representing peptide variants. In addition, RaVen automatically reports the number of spectra assigned to each peptide variant in each user-defined group to facilitate downstream statistical analysis of differential expression across conditions. While the statistical and biological significance of these observations cannot be established with the small size of the available groups, these observations still suggest that rare modifications could potentially play a role in functional differences between the groups of patients in the original study (PMID: 15080731).

We divided the samples by age and cataracts vs. non-cataracts. By doing this we were able to explore cohort- specific trends that were present in the sample. Previous papers have shown broad correlations between lens samples and certain types of modifications such as increased deamidation in aging [23][24] [25] and increased oxidation in cataracts samples [26]. Given the low mass resolution, high prevalence of carbon isotopes, and possibility of artefactual sample-handling modifications in this data set, we did not seek to establish deamidation as a differentially expressed modification in the cataracts group, but rather focused on modifications

of higher mass and more reliable in vivo detection using mass spectrometry experiments. In particular, RaVen identified cysteine methylation as a common modification in lens, and its detected higher expression in βS crystalline is in agreement with the literature (Haines et al, 2008). However, even though methylation of Cys 185 in βA1 crystallin was reported to be only mildly elevated (~40%) in cataract tissue, our mass spectrometry detection of this event estimated an >8-fold increase in our sample, thus suggesting that this modification may be more prevalent than previously thought (Table 1.9). While our lens data was not specifically processed for detection of phosphorylated peptides, RaVen's identification of 42 phosphopeptides allowed us to compare their expression across groups and confirm that these are generally not more prevalent in cataracts[27]. One phosphopeptide was found to have higher spectral counts in cataracts (Table 1.9); however, its low spectral counts moderated the significance of this finding and would require additional experiments to establish their differential expression. Finally, RaVen's identification of high levels of αB-crystallin Met-68 oxidation are also in agreement with the literature [28] but, in contrast to other studies suggesting higher oxidation in cataracts[29], RaVen identifications generally indicated lower oxidation levels for old age and cataracts samples while finding a nearly six-fold overall increase of oxidation in the infant group (Table 1.9). For example, peptide HWNEWGAFQPQ(M,16)QSLR from CRBB1 is 4 times more abundant than the unmodified peptide in infant samples, but is two to four times less abundant than the unmodified peptide in old-age and cataracts groups, despite the protein itself being much more abundant in these groups (~3 times higher spectral counts than in the infant group). Table 1.9 further reports the 10 modifications with the highest relative expression for each cohort. Of particular note, RaVen detected an undetermined modification of 55 daltons on arginine with high spectral counts in old age; this modification was also previously reported

in[20], 2006 but its identity remains undetermined to date. To illustrate how RaVen spectral networks support the inspection and validation of differentially expressed unexpected modifications, Figure 1.9 shows the spectral network for the cataract-specific peptide VQ(D,14)DFVEIHGK. Spectral networks facilitate manual inspection of RaVen results in two major ways. First, the neighbors of the modified peptide in the spectral network show that its spectrum is very strongly correlated to those of multiple spectra of the corresponding unmodified peptide (e.g., Figure 1.10), thereby strongly supporting the identification of the +14 dalton mass offset on this peptide sequence (which can be explained as either Asp methylation or an Asp→Glu polymorphism). Second, the pie charts overlaid with the nodes show the spectral counts of each peptide variant in the various groups covered in our data. In particular, the solid grey circle for VQ(D,14)DFVEIHGK illustrates that this variant occurs only in the cataracts group and is the only variant of this peptide that is cataracts-specific.

1.7. Discussion

The spectral networks approach underlying RaVen's discovery of modifications advances blind modification searches in three main directions. First, RaVen uses spectral assembly to derive high-quality network-level (instead of spectrum-level) de novo sequence tags used to select database locations for sequence/spectrum alignment using sample-specific modification mass frequencies. Second, RaVen shows that correlations of peptide fragmentation patterns between differently-modified variants of the same peptides can support discovery of non-common (or rare) modifications in a manner that is much more sensitive than what is typically reported using offset frequency functions. While this approach is not suitable to detecting peptide variants that occur in only one modified form in the set of searched files (note that RaVen

supports joint analysis of files from many samples), we would note that a) RaVen shows that using correlated variants does result in the detection of many modifications and b) the false discovery rate is expected to be much higher in cases when peptides would occur in only one modified form, without corroborating information from related variants. Third, RaVen's aggregation of multiple variants of overlapping peptide sequences in the same spectral networks further focuses the analysis of the results on the diversity of proteome variant that can be detected on the same peptides or on the same protein variants regions. In fact, our analysis of lens proteins reveals an unprecedented level of diversity with over 100 regions being identified with 10+ distinct peptide variants and with hundreds of modified variants identified to a single region of Alpha Crystalin A. While it remains unclear whether some of this diversity may be potentially related to structural protein features (e.g., solvent accessibility or disordered regions), RaVen's ability to detect hypermodified regions motivates a quest for the determination of the functional significance of this immense proteomics diversity at the level of modifications and cleavage events.

1.8. Online Methods

The first stage of the RaVen workflow uses the AlignGF spectral alignment algorithm [12] combined with the MetaSPS spectral assembly algorithm [30] to construct spectral contigs. As previously described [31], MS-Cluster is used to group spectra obtained by repeated acquisition of the same peptide; each cluster is represented by a cluster-consensus spectrum and the resulting identifications are assigned to all corresponding cluster members. Cluster-consensus spectra are processed in two separate ways. First, RaVen uses MS-GF+ [3] searches of cluster-

consensus spectra to provide the peptide spectrum matches (PSMs) used by AlignGF to estimate the false discovery rate (FDR) of spectral alignment. Second, RaVen uses MetaSPS [30] to construct spectral contigs whereby cluster-consensus spectra are converted to Prefix Residue Mass (PRM) spectra (i.e., peak intensities are converted to likelihood scores and peak masses are converted to putative prefix/N-term cumulative amino acid masses [32] and AlignGF spectral pairs (filtered at 1% FDR) are used to assemble spectra into contigs and thereby derive contig-consensus PRM spectra.

While contig-consensus spectra were previously shown to have very high signal-to-noise ratios and to enable very high quality de novo peptide sequencing [30][33][34][35], RaVen uses these spectra to generate shorter de novo sequences (i.e., tags) that are more likely to be 100% correct and can thus be used to reduce the database search space of which sequences are matched to which spectra. Database filters based on de novo sequencing tags have been previously shown to be very effective [16][8][15] and Raven contig tags (see Supplementary Methods) were also found to efficiently filter the database from a maximum of 1 trillion possible contig-DB match locations down to 318,147 using contig tags of length 5 with up to two missing peaks (Figure 1.11 and Figure 1.12), corresponding to a reduction of approximately 3 million times in the database search space.

1.9. Blind modification searches

The second stage of the RaVen workflow uses a new dynamic programming algorithm to align assembled PRM spectra to the protein sequence locations determined by the contig-DB matches. RaVen contig tag filters effectively reduced the search space of each assembled PRM

spectrum by over 1 million times to an average of only 11-2 candidate locations per spectrum. RaVen alignment of assembled spectra to protein sequences extends previous spectral alignment approaches [9][6][5] in three major ways. First, penalties are applied to dynamic programming match scores to discourage matches that are less likely to be correct. As detailed in Supplemental Methods, discovered modifications are segregated into known, putative and unknown categories based on user input and sample-specific frequencies of occurrence of each modification mass. As shown in Figure 1.3, some detected mass offsets are over 1000x more frequent than others (e.g., sample-handling modifications such as Methionine oxidation) and thus should be preferred over alternatives using more exotic modifications on peptide-spectrum matches (PSMs) whose scores are otherwise similar. In addition, penalties are also used to reflect observation of enzymatic digestion rules and are automatically adjusted to the mean of each spectrum's per-peak likelihood scores. Second, RaVen uses pre-computed gap-alignment blocks to optimize the calculation of aggregate penalties over regions of the spectra with many missing or unmatched peaks (see Figure 1.13) with each block being reused an average of 49 times for the lens data analyzed here. Third, since spectra of shorter/longer peptides or of precursors with different charge states tend to have different score distributions [33][36], we use the generating function model [37] for calculation of per-spectrum spectral probabilities to normalize PSM scores and make them more comparable for estimation of false discovery rate (FDR). Finally, since spectra of unmodified peptides or with poor MS/MS fragmentation may not align into spectral networks (Figure 1.14), RaVen further considers MS-GF+ PSMs for spectra that remain unidentified after spectral-sequence alignments PSMs are filtered at 1% PSM-level FDR. While this step cannot contribute new PSMs with unexpected modifications, it is still useful for increasing spectrum

identification rates and overall protein sequence coverage, thereby supporting the identification of both proteins and overall peptide sequence diversity.

For comparison purposes, all tools were run on the lens tandem mass spectrometry data using parameters as similar as possible. The human lens data consists of 786,291 MS/MS spectra acquired on a Thermo LCQ Classic instrument from seven different samples: normal samples from 0 days old, 2 years old, 18 years old, 35, 70 and 93-year-old lens, and 70 and 93-year-old cataract lens (see Table 1.10). Accordingly, parent mass tolerance was set to 2.5 daltons, fragment mass tolerance was set to 0.4 daltons, trypsin digest was specified but peptides were allowed to have non-tryptic termini; results were gathered at 1% FDR (either PSM-level or variant-level, depending on which comparison is being reported). Five common known modifications were set for RaVen and MS-GF+: oxidation (M+15.994915), pyroglutamate on N-terminal Q (Q -17.026549), N-terminal carbamylation (+43.005814), N-term acetylation (+42.010565), and deamidation on N and Q (N,Q+0.984016). MODa was set to "multi-blind" mode to allow for discovery of multiple modifications on a single peptide. All searches used the reviewed subset of UniProt human proteins 20,199 proteins, also referred to as SwissProt) as downloaded on May 5th, 2016, concatenated with the sequences of commonly observed contaminant proteins [38].

1.10. Site selection using offset frequency tables

First, to avoid arbitrary localization of modifications when there are no supporting peaks in the spectrum, RaVen records the modification as being on one of several amino acids found in the mass range (i.e., gap) between the two closest peptide cleavage ions in the spectrum. Second,

an initial set of counts for the offset frequency table is computed, spreading the count for these "gap modifications" over all amino acids in the gap; for example, if the original annotation was (PEP,16), this would yield a ⅔ count for P and a ⅓ count for E. Third, RaVen selects the highest count modification site X for each modification mass M, reassigns M to X on all gaps containing X and eliminates the counts from the gap assignment. In the case of (PEP,16), if (P,16) had the highest counts then the modification would be localized to "P" (note that the selection of which P is modified is not relevant at this stage, as the immediate goal is only to infer which amino acids are more likely to be modified by which modification masses). Fourth, the procedure is repeated until all gaps have been eliminated. In this way all gap PTMs are reassigned to their most likely single amino acid annotation.

1.11. Supplemental – Consensus spectrum tagging

MetaSPS [30] assembly of cluster-consensus spectra results in networks of spectra from peptides with overlapping sequences (i.e., contigs), with each contig also represented by a contig-consensus spectrum which MetaSPS originally used for de novo protein sequencing [13] for an overview of spectral alignment and network algorithms). In difference from this, RaVen uses contig-consensus spectra to derive short amino acid sequence tags which are matched to the database to determine the set of locations used for alignment of spectra assembled into the corresponding contig (see Figure 1.1). The concept of using short de novo sequence tags to filter database search matches was originally proposed by Mann and Wilm [15] and has since been used in various related approaches [38][9][18][8][39]. RaVen extends this concept by using contig-derived sequence tags to reduce the overall number of database locations that are matched

to contig-consensus spectra. RaVen de novo sequence tags are constructed by connecting contig-consensus spectrum peaks whose masses differ by one amino acid mass; peaks whose masses differ by two amino acid masses (e.g., X and Z) may also be connected by mass gaps corresponding to the summed amino acid masses (e.g., mass(X)+mass(Z)) and representing all possible permutations of the two amino acids (e.g., "XZ" or "ZX"). As previously shown [30], the accuracy of de novo sequencing is much higher in contig-consensus spectra than in non-assembled spectra because noise peaks are mostly eliminated (or have proportionally reduced scores) and because likelihood scores from signal peaks are accumulated from all spectra assembled into each contig. These higher signal-to-noise ratios in contig-consensus spectra thus allow for the determination of longer sequence tags than the 3-mer sequence tags that are traditionally derived from individual spectra (see Figure 1.11). Since multiple different tags may be derived from each spectrum, it is common [9][16] to rank the tags by the combined score of its matching spectrum peaks and then to use the top N tags per spectrum to filter the database. As shown in Figure 1.11 the choice of tag length and the number of allowed gaps per tag, has a substantial effect on the number of times that a tag is matched to the SwissProt database (Figure 1.11a), as well as on the number of contigs with at least one correct match (i.e., the minimum necessary to retain the correct identification). While the filtration efficiency increases with longer tags (i.e., longer sequences have less matches to the database), missing spectrum peaks due to incomplete peptide fragmentation generally constrain the length of the longest correct tags that can be extracted from contig-consensus spectra. As such, we have found that the best compromise between sensitivity and filtration efficiency is with tags of length 5 with up to two gaps of two amino acid masses (i.e., 5-2 in Figure 1.11); the next best option would be to use tags of length 6 with up to two gaps of two amino acid masses (i.e., 6-2 in Figure 1.11) but the

16x gain in filtration efficiency reduces sensitivity by 9%. Using 50 tags per contig-consensus spectrum and with each tag yielding ≈20 matches to the SwissProt database, this filtration step results in contig-consensus spectra being matched to only ≈1,000 locations in the database instead of at ≈11 million possible locations (i.e., considering matches at every amino acid on every protein), thus corresponding to >10,000-fold decrease in the search space. Since we are primarily interested in identifying modifications for proteins whose identification is already otherwise likely, we obtain further reduction in the search space by considering only proteins for which MS-GF+ has found at least one PSM (MS-GF+ identifies over 99% of all proteins in our data regardless of their modifications).

For every location where a contig tag is matched to the database, we perform a tag-extension procedure where the score of a tag match is increased if the database sequence flanking the matched tag also matches the masses of peaks flanking the tag in the contig-consensus spectrum (see Figure 1.1b in the main text); the score of each tag match is thus increased by the summed score of all contig-consensus spectrum peaks matched to the flanking database sequence. Scores updated with the tag extension procedure are then used to remove contig tag matches whose score is not less than 80% of the maximum extended tag match score for the same contig. As shown in Figure 1.12a, lowering the threshold of acceptable tag match scores down to 75% or 70% of the top score yields additional identifications but the number of additional identifications is very low when weighed against the very large increase in the number of contig tag matches (see Figure 1.12), thus resulting in a poor tradeoff of increased sensitivity versus speed. Once the final set of contig tag matches is determined, assembled spectra are then aligned to the highly reduced set of database locations determined by the set of tag matches for the corresponding contig. As shown in Figure 1.12, the selected 80% threshold for contig tag

match scores results in only 536,927 matched database locations for 8348 contigs, thus resulting in each assembled spectrum being matched to an average of only 11-12 database locations, corresponding to a reduction of six orders of magnitude (i.e., $1.0x10^6$) in the number of locations where assembled spectra are aligned to the database.

1.12. Supplemental – Penalty Based Alignment

Spectrum/sequence alignment algorithms were first introduced in 2001 [5] and have since been adopted in various approaches [9][6][40] used to find high-scoring spectrum/sequence matches while allowing up to K unexpected modifications per match. Because allowing for multiple modifications per match tended to yield many false-positive matches, K was usually set to 1 (MODa[9] allows for multi-blind searches with K>1 but requires the interleaving occurrence of sequence tags between putative modifications). Rather than explicitly limiting the number of possible modifications per peptide, RaVen allows for any number and type of modifications but restricts their excessive usage by imposing score penalties that are inversely proportional to the sample-specific evidence in support of each putative modification mass.

To determine penalties for our alignment we use information from the spectral pairs that were identified during construction of the spectral networks. If there is a parent mass difference D (exceeding the parent mass tolerance) between two spectra matched by an AlignGF [9] spectral pair, then D supports the possibility that some modification (or other peptide transformation, such as a sequence extension) of mass D could have occurred within (or at either end of) the higher-mass peptide whose spectrum is being observed. By estimating the frequency of all such mass differences between all statistically significant AlignGF spectral pairs, we obtain

a histogram as shown in Figure 1.3, where high-count parent mass differences indicate the presence of putative modifications that are likely to be occurring in the sample (especially sample-handling modifications, as these are most likely to induce modifications on many peptides). While it is unavoidable that some false positive spectral pairs may be detected by the spectral matching process, we note that the corresponding "noise" introduced into this histogram is likely to distribute the detected parent mass differences across random mass difference bins, as there is no prior reason to expect false positive spectral pairs to induce biases towards specific parent mass differences. Furthermore, we impose 1% FDR on AlignGF spectral pairs and set a minimum per-mass frequency threshold for putative modifications of 0.5%, thus making it highly unlikely that false positive mass differences will substantially affect the frequencies histogram. Since spectral alignment cannot determine the directionality of modifications (i.e., a mass difference of D daltons could correspond to either a mass increase of +D or a mass decrease of –D on the modified peptide) and frequently cannot assign the precise amino acid localization of the modification (most spectra have incomplete fragmentation), the histogram provides evidence only towards the absolute value of the modification mass. The resulting histogram frequencies of the putative modifications are then used to define penalties used in the dynamic programming alignment algorithm used for scoring spectrum-sequence matches (described below).

Previous peptide-spectrum matching algorithms[3, 41–43] have established the utility of assessing the quality of matches using functions of the sum of the total intensity of PRM peaks that match the theoretical amino acid masses of a peptide from the database. Similarly, RaVen uses scored PRM spectra for spectrum-sequence matches, defining the base matching score as the sum of the PRM intensity scores, which is then modified by penalties derived from the

frequency of occurrence of the modification masses used to make the spectrum match the sequence. In order to properly combine penalties with the summed PRM score, the penalty values are scaled as multipliers of the average PRM intensity score in each spectrum, thereby making penalties automatically adjusted to each spectrum's peak scores and facilitating their intuitive interpretation as the "number of peak equivalents" required for the match to benefit from using a modification mass. More formally, we define peak equivalent (PE) as follows. Suppose we have two possible database matches, M1 and M2 for the same spectrum S, whose average peak score is A. Suppose also that annotations M1 and M2 have identical amino acid sequences, have nearly identical annotations, but M2 contains the modification *m*. If the modification *m* has a penalty of 1.0 peak equivalents, then M2 will have a higher match score than M1 if and only if M2 matches at least (1.0 * A) more total peak score than M1. Intuitively, using peak equivalents allows us to model penalties in a way that relates directly to each spectrum, in that adding a modification with a penalty of P into the alignment requires matching additional P (average) peaks to compensate for the penalty.

We divide all possible modifications into three categories: known, putative and unknown. We define known modifications as those that are specified a-priori. These are modifications of the type that are expected to appear often in most samples such as oxidation or deamidation. For our data we specified known modifications for oxidation of methionine, pyroglutamate of n-terminal glutamine, deamidation of asparagine and glutamine, and n-terminal acetylation and carbamylation. We assign modifications in this category a penalty of 0.01 peak equivalents. This penalty is small enough that adding such a modification will not greatly impact our final score but greater than zero so that adding a known modification still yields a lower score than no modification at all. The second category of putative modifications, are those for which there is

evidence of their presence in the sample from the histogram of spectral pairs as described above. These modifications may include modifications such as methylation, or sodium adducts which are given a penalty between a minimum value of 1.0 and a maximum value of 1.5; we map the frequency of mass differences of the spectral pairs to the peak equivalent range using a simple scaling procedure:

$$PutativePenalty = 1.5 - \frac{freq - minFreq}{maxFreq - minFreq} * 0.5$$

Where maxFreq is the frequency of the most common mass difference and is given the minimum penalty of 1.0 and minFreq is the least common (but still above a minimum threshold of 0.05% to exclude noise) is given the maximum penalty of 1.5. After excluding the know modification masses and masses corresponding to amino acids (which are likely just peptide extensions) we found 23 mass differences that were labeled as putative modifications in our data.

While the frequencies of mass differences in the spectral pairs histogram helps to reduce penalties for very common modifications, the detection of rare modifications is also emphasized by considering spectrum-specific putative penalties for mass differences of spectral pairs incident on each spectrum. These penalties are also scaled in the same manner as above from the minimum (1.0) to maximum (1.5) peak equivalents but for a given spectrum S and a putative modification of mass m, the scaling is based on the best AlignGF score (i.e., -log(AlignGF probability)) for a spectral pairs (S,S') where mass(S')-mass(S)=m and the normalization is based on the range of AlignGF scores over all pairs in the data.

$$SpectrumSpecificPenalty = 1.5 - \frac{BestPairScore - minAGFScore}{maxAGFScore - minAGFScore} * 0.5$$

All modifications not covered by the previous two categories are considered "unknown" modifications. These are modifications for which we have no significant prior evidence and are

the least likely to actually represent real modifications. We therefore assign these modifications a penalty of 1.5 times the maximum putative penalty.

With our penalty based approach, additional penalties from other sources of information can be added to the algorithm. In particular, we also apply a relatively small penalty based upon proper tryptic cleavage of the candidate database peptide sequences. Tryptic penalties are applied when the amino acid preceding the peptide is not K, nor R, when there is a K or R in the middle of the peptide or when the final amino acid is neither K nor R. In any of these cases a cumulative 0.5 peak equivalent penalty is applied for each "missed cleavage" location. These cleavage penalties are in addition to any other penalties from modifications. For future experiments, similar penalties can be created easily for other proteases.

Using these penalties our alignment score consists of the sum of matched peaks minus the penalty from any modifications or missed cleavages that were used to achieve the match. The recurrence relationship for PSM scoring is defined as follows:

$S[i][j]$	the PSM scoring matrix
$D[0..m]$	amino acid sequence of the database peptide
$T[0..m]$	masses of peaks of the theoretical database spectrum
$M[0..n]$	masses of the peaks of the peptide spectrum
$I[0..n]$	intensities of the peaks of the peptide spectrum (i.e., PRM scores)
$P[A][m]$	the penalty for a modification of mass *m* on amino acid string *A*
C[A]	Cleavage penalty for amino acid string *A*
SSP	Spectrum specific penalty

$\Delta M(a,b) = M[b] - M[a]$	Delta mass between peaks in the peptide spectrum
$\Delta T(c,d) = T[d] - T[c]$	Delta mass between peaks in the theoretical spectrum
$\Delta P(a,b,c,d)$ $= \Delta T(c,d) - \Delta M(a,b)$	Delta mass between a gap in the peptide spectrum and amino acid sequence in the database (i.e., the hypothesis for scan number k, d modification mass)

The Recurrence Relation is thus defined as follows

$$S[i][j] = I\,[0], where\ i\ =\ 0\ or\ j\ =\ 0$$
$$S[i][j] = I[j] + S[x][y] + C\big[D[x..i]\big] + min\left\{P\big[D[x..i]\big][\Delta P(x,i,y,j)], SSP\right\}\ \forall\ 0 \le x < i, 0 \le y < j$$

Notice that for our alignment we use unlimited "lookback" in both dimensions. This is so that we may match gaps where either there are noise peaks in the spectrum and we wish to ignore them in favor of the true peaks, or there are missing peaks in the spectrum (which is often the case for the contig-consensus spectra) and we wish to match multiple amino acids from the database to adjacent peaks in the spectrum. In theory this makes the algorithm $O(n^2m^2)$, where n is number of masses in the spectrum and m is the number of amino acids in the database sequence, and is potentially quite slow. However, in practice the lookback can be limited to any amount desired so we limit lookback to 1500 daltons in the spectrum and 8 amino acids in the database (additional details discussed below).

To reduce the runtime of the alignment algorithm we employ an optimization for all gap matches. A gap match is any match in the alignment that does not compare consecutive theoretical masses to consecutive spectral peaks. Any such match can be thought of as matching a "gap mass" G (the mass between the start and end peaks being considered), with the database string (the sequence of amino acids being matched). Because our algorithm uses the sum of the peak scores minus the sum of any penalties, we simply need to add the scores of the two bounding peaks, and then subtract any penalties required to match the mass G to the theoretical mass of the database string. Since the best score will always be the same for the same sequence of amino acids and the same mass, we can precompute the answers for each amino acid sequence and store these gap-alignment (GA) blocks for each mass. Additionally, the order of the amino acids does not matter since modifications appear on a single amino acid without respect to order (we do not consider context) and by definition a gap contains no interior spectrum peaks to

determine the localization of modifications. The only limiting factor is the amount of memory required to cache all such GA blocks. For our purposes, we precompute all 2-mers, 3-mers and 4-mers with 1 dalton bins. This results in less than 16000 GA blocks requiring only ≈100 megabytes of memory. For each bin we compute the lowest penalty that could be achieved at that mass for the amino acid sequence in question. Each GA block appears as in Figure 1.13. Once computed, the solution for any gap match up to four amino acids will then be a single lookup in the GA block corresponding to the amino acid sequence in question, i.e. - O(1). If we wish to compute the penalty for a gap of five to eight amino acids we may combine the values for two GA blocks, one that represents the initial 4-mer of string, and another that represents the remainder. For example, for a 7-mer, we combine the GA blocks for the beginning 4-mer and the trailing 3-mer. The answer can be found in O(n) time where n is the length of a single GA block (1500). This is done by finding the minimal value (penalty):

$$\min\left[GABlock1[m] + GABlock2[M-m]\right] \;\; \forall \;\; 0 \leq m \leq N$$

Where *m* is the mass of the bin from one block and M is the total mass of the spectral gap we are computing. This method can be used to combine more than two GA blocks for even longer amino acid gaps, but the runtimes increase quickly and hence this becomes undesirable for values above the precomputed size times two. Since the average amino acid mass is approximately 120 daltons, allowing our GA blocks to be 1500 daltons allows sufficient length for eight average amino acid masses combined with multiple large (over 100 dalton) modification masses. Also, this is a very large spectral gap to have no matching peaks; a gap this large is already indicative of a poor quality spectrum and as such, is likely to be misidentified

regardless. This was also further supported by our observation that considering larger gaps did not positively influence identifications in our data.

1.13. Supplemental - FDR

To determine which identifications are significant we used the standard False Discovery Rate (FDR) method [44] using the target database and a shuffled decoy database. We use a shuffled database rather than the common reversed database because we do not know the orientation of the peptide in contig-consensus spectra [30]. In addition, given that the computed alignment scores are highly dependent on the total intensity of peaks in each individual spectrum, we use a method similar to that used by MS-GF to compute p-values to derive a "p-value score" or "p-score" suitable for estimation of FDR. This p-score is computed by calculating the distribution of scores for all possible peptides using only the known modifications (putative and unknown modifications are not considered). Using the actual score of the alignment match (which includes the penalties for all modifications), we determine the chance of a random match (with known modifications) scoring higher than our penalty-scored match. Because we do not consider putative or unknown modifications in the null space of all possible matches, this is not a true p-value as computed in MS-GF but still works well as a spectrum-specific normalized identification score.

1.14. Supplemental – Variants

Since site localization is often unknown or uncertain for putative or unknown

modifications, we define the concept of a “peptide variant” or simply “variant”. All peptides which have the same amino acid sequence and whose total mass differs by no more than the parent mass tolerance are considered to be a single peptide variant. For example the two peptides: VARIANT and VARIA(N,1)T are considered the same variant if the parent mass tolerance was at least 1 dalton, while VARIANT and VARIA(N,1)TS would never be considered the same variant because they do not have the same unmodified peptide. The two peptides: (V,12)ARIA(N,1)T and VA(R,15)IANT would be considered the same variant if the parent mass tolerance was 2 daltons, since the unmodified peptides are the same and their total mass is within 2 daltons. Because the lens data we used is low resolution, we set a mass tolerance of 2.5 daltons for our analysis.

We propose that counting identifications of peptide variants is better than simply counting spectrum identifications or unique modifications for multiple reasons. First, if performance of an algorithm is measured solely in terms of additional identifications, just using spectral identifications does not necessarily reflect the amount of new information that could be of potential biological relevance (e.g., there is not much additional information in identifying more spectra to peptides that were already otherwise identified). Second, if two different identifications differ only with respect to their modifications, and those modifications fall within the parent mass tolerance it is often not possible to determine if the PTM identification is accurate without a manual inspection of the spectrum. Third, since localization of modifications is usually beyond the scope of blind search algorithms, it is likely that two peptides with similar total modification masses on the same underlying peptide sequence would actually represent the same modification, while just being mislabeled by the algorithm as differing PTM combinations. As such, identification of variants focuses the analysis of the results in the discovery of distinct

new modifications rather than on the discovery of new sites for the same modifications (which would require addressing other issues such as co-elution of similarly modified peptides).

Once we have obtained the list of FDR filtered identifications for the individual spectra using our new alignment we then reduce all the PSMs to one single PSM that is a representative of each variant. The PSM chosen is the one with the highest p-value (or p-score) of all PSMs in the variant group.

Since MS-GF+ identifications are used to estimate false discovery rates of AlignGF spectral pairs, we further reuse those identifications for spectra that are not directly identified by our alignment (e.g., spectra with no pairs in the dataset or poor spectra with no sequence tags). Since the p-values used to compute FDR by MS-GF+ and the p-scores used in our FDR computation are not comparable, we use search-specific PSM q-values as the PSM scores used for variant representatives. Finally, variant representatives are ranked by their corresponding variant q-scores and variant-level FDR is estimated as usual [44].

1.15. Supplemental – Detection of common modifications using variants-based offset frequency function

As described in the main text, traditional approaches to blind database search [9][11][45] traditionally use offset frequency functions based on PSM counts to detect frequent (amino acid, mass offset) pairs and thus cannot readily distinguish modifications occurring on many different sequences (and thus more likely to be real) from modifications identified on only a few sequences but with high spectral counts (as is often the case for very abundant peptides). RaVen makes this distinction by computing offset frequency functions not on PSMs but on counts of unique peptide variants with each detected mass offset (Table 1.2). Moreover, since the

localization of masses with blind searches is not known a priori, RaVen assigns partial credit to multiple amino acids where there are no distinguishing peaks and uses an iterative strategy to converge to the minimal set of amino acids that covers all identified peptide variants. For example: a peptide variant containing a detected mass of +12 on a gap (ABCC, +12) would not have any spectrum peaks supporting any of the four amino acids in the ABCC subsequence so RaVen would assign the frequency of observation of the mass offset in proportion to the number of times each amino acid is observed in the subsequence: (A,12) ¼, (B,12) ¼, (C,12) ½. Once the initial table of frequencies is constructed RaVen uses a greedy set cover algorithm to selected assignments of detected mass offsets to amino acids or N/C-termini. The method proceeds as follows:

1. For every cell in the offset frequency table, find the site with the highest counts.
    a. For every variant identification with a subsequence that contains that site and mass, assign the modification to the selected site. I.e., if mass +12 has the most counts on H, change (ABH, +12) to AB(H, +12).
2. Recompute all the frequencies in the offset frequency table using the updated variant identifications.
3. Repeat steps 1 and 2 until all gaps have been resolved.

Using this procedure RaVen's variant-based PTM table has greatly increased ability to localize modifications (see Figure 1.15).

1.16. Supplemental – Detection of rare modifications using spectral networks and alignment probabilities

While offset frequency functions have been very successful for the detection of sample-handling modifications or very abundant post-translational modifications, their foundations on the frequency of observation of a mass offset on many peptides intrinsically bias against the

detection of rare, but potentially biologically more relevant post-translational modifications. In fact, it is not unreasonable to expect that functional post-translational modifications would mostly occur on only very few sites where corresponding modification enzymes (e.g., kinases or acetylases) would have been evolutionarily refined to modify with a high degree of specificity and only under controlled regulatory contexts. To specifically support the detection of rare but highly-significant mass offsets, RaVen capitalizes on the significance of spectral alignments between the spectrum with the putative mass offset and all of its neighbors in the corresponding spectral networks. Intuitively, this supports the notion that a putative modification is significant if there are other spectra with highly correlated fragmentation patterns assigned to overlapping peptide sequences with zero or other different modifications – a type of event that may occur only once yet still result in very significant spectral alignment p-values [12].

For example, if a spectrum S is identified as having a modification of mass X and is also aligned to a spectrum S' where mass(S)-mass(S')=X then that supports the hypothesis that the modification mass X is not merely an alignment anomaly since the likelihood of the spectral pairing algorithm randomly paring two spectra that happen to have the same mass difference is very low (identifications are derived by aligning each spectrum to the database but spectral pairs are derived by aligning spectra to spectra). Also, since spectral alignments are filtered at a 1% FDR threshold and database identifications are also filtered at 1% FDR, the odds of such an event happening by random chance would be very low. The more neighbors in the network that support the existence of the modification and the higher their pair scoring, the less likely the modification is simply a random product of the alignment algorithm and more likely is a true discovery. The supported mod score is thus determined by summing the scores of all pairs of the spectrum that have a mass difference equal to the modification mass and that are in the same

peptide region. Using this method, from the thousands of modifications discovered, we can rapidly find modifications that may be present in only a handful of spectra, but are well supported by additional network data.

1.17. Supplemental – Detection of variants with chohort-specific fold changes in spectral counts

In order to detect cohort-specific fold changes in spectral counts, we compute the ratio of the number of PSM's containing the modification in the cohort of interest to the sum of those that occur in the control cohort. However, different cohorts may contain different total number of spectra or have different rates of identification so we need to find an appropriate baseline for comparison. As such, we select the most prevalent form of the peptide (often this is the unmodified form, but may be one with a very common modification) and compute the fold change as follows:

N_{cases} = Number of PSMs with the modification in the cohort of interest
$N_{control}$ = Total PSMs with the modification in the control cohort
MP_{cases} = Number of PSMs of the most prevalent form in the cohort of interest
$MP_{control}$ = Total PSMs of the most prevalent in the control cohort

$$FoldChange = \frac{\frac{\mathrm{Ncases}}{\mathrm{MPcases}}}{\frac{\mathrm{Ncontrol}}{\mathrm{MPcontrol}}}$$

To reduce exposure to highly-variable ratios resulting from low spectral counts, we limit our consideration to variants with at least 15 total identified spectra in all cohorts and with most prevalent forms that have at least 5 spectra in each cohort. Finally, because many of these

modifications are rare, we consider any cohorts with zero identifications to have had one identification.

1.18. Acknowledgements

This work was supported by the US National Institutes of Health grant 2 P41 GM103484-06A1 from the National Institute of General Medical Sciences; NB is an Alfred P. Sloan Research Fellow.

This chapter has been submitted for publication to *Proceedings of the National Academy of Sciences* under the title “Spectral networks identification of rare post-translational modifications and hypermodified regions in lens crystallin proteins” Bernstein, Laurence E., Julie Wertz, Seungin Na, and Bandeira, Nuno. 2018. The dissertation author is the primary author of this paper.

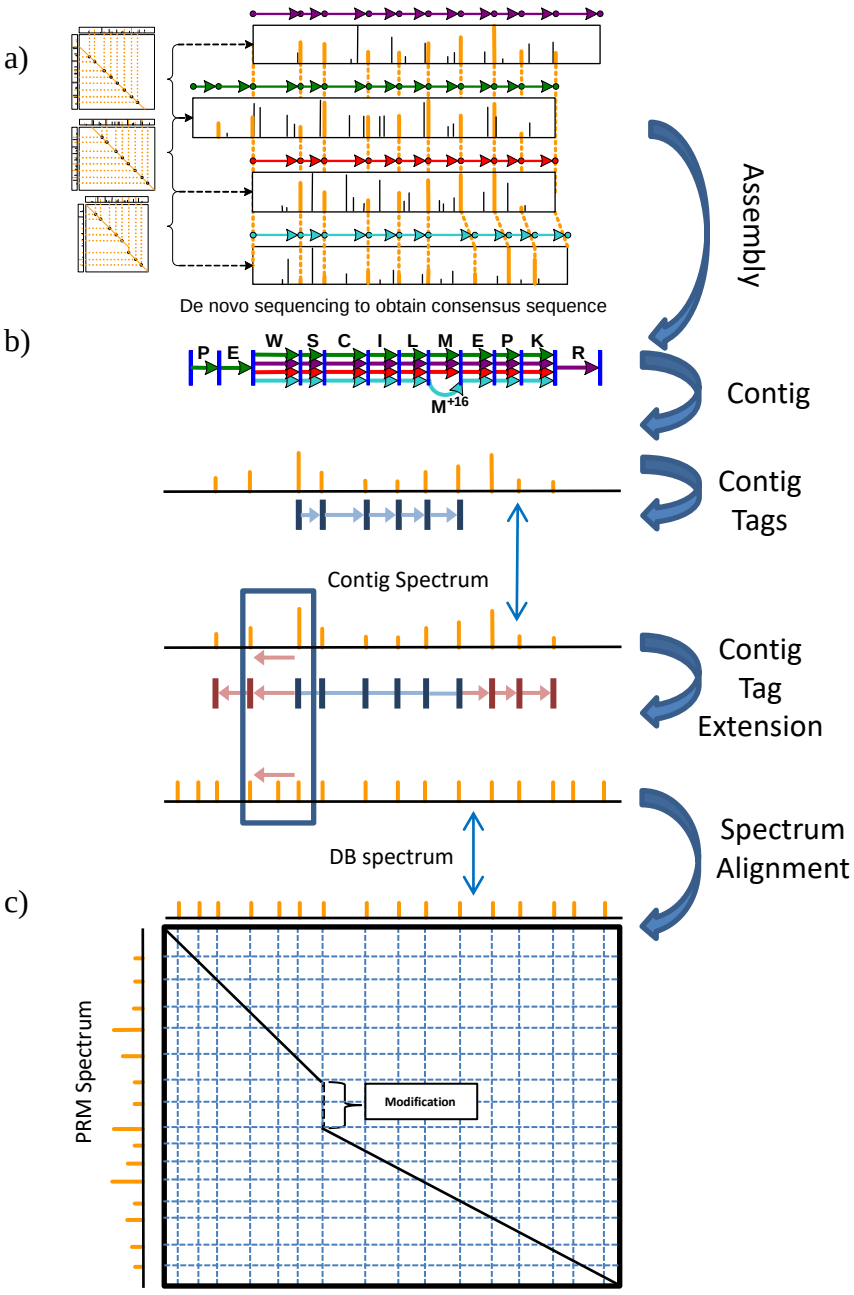


Figure 1.1: The RaVen Workflow
(a) Tandem mass spectra are aligned using AlignGF and assembled into spectrum contigs using MetaSPS; (b) De novo sequence tags are extracted from contig consensus spectra and matched to the database to select putative match positions; (c) contig-assembled spectra are aligned to tag-filtered database sequences using a dynamic programming algorithm parameterized with sample-specific modification frequency scores.

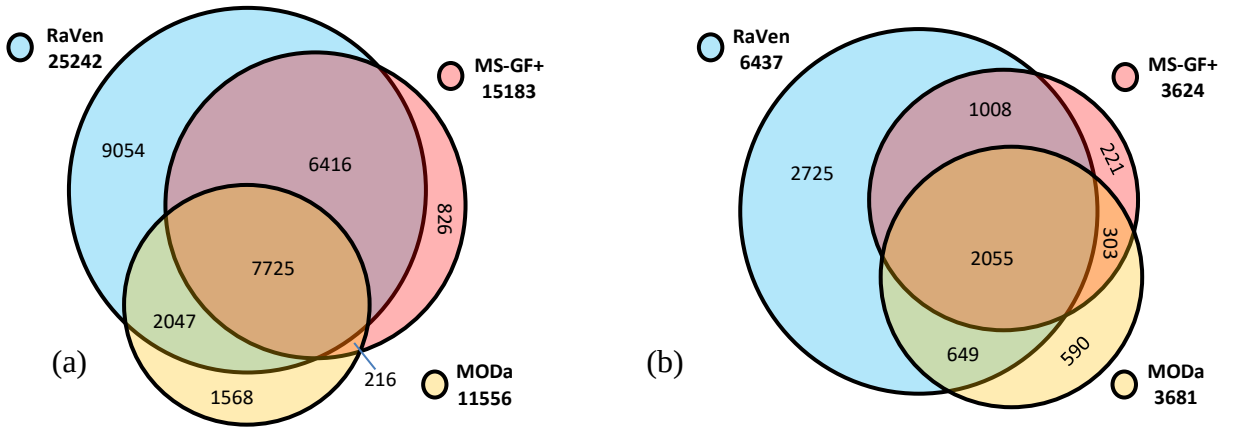


Figure 1.2: Comparison of identification rates of RaVen, MS-GF+ and MODa.
(a) shows the number of cluster identifications for each algorithm while (b) shows number of variants (i.e., unique modified peptide sequences, regardless of the localization of modifications). Totals for all identifications by each algorithm are shown next to the colored circles with the names of each algorithm.

Table 1.1: Upper portion of PTM Table from MODa showing all cells with >10 counts.
Green cells are known modifications (according to unimod.org) while red cells are unknown. Table exhibits very poor localization of the modification sites, particularly for those modifications with very high counts.

Mass	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	nTerm	cTrm	Total
1	43	0	43	82	33	61	31	30	19	72	22	347	48	232	16	59	32	47	13	26	70	3	1256
43	27	2	18	16	0	64	24	33	61	20	33	2	5	12	21	15	9	40	5	12	317	7	419
22	17	0	46	44	12	20	6	9	1	29	0	13	11	15	0	15	19	15	5	10	33	0	287
16	9	0	3	9	3	24	11	0	4	10	134	2	7	1	3	5	5	9	39	6	72	2	284
42	46	0	7	5	0	2	2	1	55	0	51	1	0	0	0	18	0	12	1	1	150	0	202
-17	0	0	8	5	3	0	4	1	6	7	1	17	7	101	11	2	4	8	4	1	91	2	190
14	8	122	1	0	1	1	19	0	13	2	2	1	0	0	1	2	1	3	2	0	3	2	179
28	7	4	4	4	1	4	4	10	41	9	1	5	5	4	1	46	17	3	2	0	29	3	172
2	8	0	6	11	2	6	1	2	3	7	7	32	6	15	5	15	5	9	1	7	13	1	148
58	1	36	0	2	0	6	0	1	56	1	17	0	2	0	1	2	2	5	0	0	37	2	132
-18	0	0	11	24	3	0	7	3	1	10	2	0	3	13	2	19	20	1	2	2	41	4	123
23	3	1	4	5	1	10	0	1	1	12	0	4	4	3	0	6	1	1	0	6	1	0	63
17	2	0	0	1	1	4	1	4	1	1	19	2	2	2	0	1	2	0	2	4	9	0	49
72	1	1	0	5	0	1	1	1	21	1	0	0	1	1	5	0	3	2	2	2	22	4	48
32	2	0	1	4	0	2	0	0	0	0	5	0	1	2	0	1	0	3	23	0	10	0	44
44	3	1	2	2	2	2	0	2	6	2	1	1	0	1	1	1	0	3	13	0	9	0	43
55	0	0	0	3	1	3	0	0	0	0	0	0	0	2	30	1	0	2	0	1	11	0	43
80	1	3	1	1	0	0	0	1	0	5	0	2	3	0	0	19	3	1	0	2	5	0	42
29	1	0	0	0	1	1	0	9	15	8	0	0	0	1	0	4	0	1	0	0	3	0	41
54	3	0	1	2	2	2	0	0	0	0	0	0	0	0	27	2	0	0	1	1	12	0	41
12	1	0	1	4	0	0	5	1	1	1	0	2	2	1	0	1	0	1	16	0	14	1	37
24	3	0	8	9	1	2	0	1	0	0	0	2	0	1	0	4	1	3	1	0	7	0	36
116	0	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	35
40	0	31	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	30	0	34
161	0	0	1	1	1	1	0	0	0	0	0	2	2	22	0	0	0	1	0	1	1	0	32
4	1	0	3	1	0	0	0	2	0	0	0	1	2	0	2	0	0	1	17	0	4	0	30
15	0	9	0	0	0	1	1	0	7	2	0	2	0	2	0	0	0	0	4	0	2	0	28
-48	0	0	1	2	0	0	0	0	1	1	20	0	0	0	0	0	0	0	0	0	5	0	25
-42	0	0	1	1	3	0	1	0	0	0	0	0	0	0	14	0	1	0	0	1	2	0	22
-1	2	0	0	6	0	0	0	3	1	1	0	0	1	5	0	0	2	0	0	0	15	1	21
59	1	3	3	0	3	1	0	0	0	2	2	0	1	1	1	0	1	0	1	0	6	0	20
-13	0	0	0	0	0	0	0	0	0	1	0	0	0	11	1	5	0	1	0	0	18	0	19
-43	0	0	3	2	6	0	1	0	2	0	0	0	0	1	0	0	2	0	0	1	0	0	18
26	2	0	0	1	0	0	2	1	2	1	0	0	0	1	2	1	1	4	0	0	14	0	18
73	0	9	0	0	0	1	0	1	0	0	0	0	0	0	1	6	0	0	0	0	1	0	18
156	4	0	1	2	2	0	0	1	0	1	1	0	1	0	0	0	2	1	0	2	8	5	18
-2	0	0	1	0	0	0	0	3	0	4	1	0	1	2	0	2	1	1	1	0	12	0	17
30	1	0	0	1	0	1	0	0	1	2	0	0	0	0	1	9	1	0	0	0	3	0	17
-30	0	0	1	1	0	0	0	0	0	0	1	0	0	1	0	4	8	0	0	0	2	0	16
64	0	3	2	0	0	0	0	0	3	0	1	0	0	0	0	5	2	0	0	0	6	0	16
48	0	7	0	0	2	0	0	0	0	1	1	0	1	0	0	0	1	2	0	0	2	0	15
-6	1	0	1	3	0	0	0	0	0	0	7	0	0	0	1	1	0	0	0	0	10	1	14
3	0	0	0	4	0	2	0	0	0	3	0	0	0	0	0	1	2	1	0	1	2	0	14
103	1	0	1	0	0	0	0	0	0	1	0	1	0	0	0	7	0	1	0	0	1	0	12
-14	0	0	0	0	0	0	0	4	0	0	1	1	0	1	3	1	0	0	0	0	7	0	11
56	0	1	0	1	0	0	1	0	0	0	3	0	0	1	0	3	0	0	1	0	6	0	11
101	0	8	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	1	0	11
105	0	10	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11

Table 1.2: Upper portion of the PTM Table using the RaVen method using variants with recursive localization by frequency showing all cells with >10 counts.
Highlighted cells are those with values of at least 10 counts. Green cells are known modifications (according to unimod.org) while red cells are unknown. Compare this smaller number of high count cells and large proportion of known modifications with the same table as created by MODa.

Mass	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	nTerm	cTrm	Total
1	0	0	0	0	1	0	0	0	0	0	0	361	0	256	0	2	0	0	0	0	13	1	633
43	2	4	3	2	0	0	1	6	21	1	8	0	0	1	1	0	4	7	0	0	330	1	389
16	1	1	0	0	4	0	2	1	4	0	223	0	4	0	0	0	0	0	21	0	17	2	278
22	13	1	37	53	3	24	4	3	2	8	0	13	6	22	0	26	15	12	2	3	3	0	248
42	1	2	0	0	0	0	0	0	25	0	0	0	0	0	0	1	0	0	0	0	173	1	202
-17	0	0	5	2	0	0	0	0	5	11	1	19	1	41	4	1	1	1	1	0	88	6	184
-18	0	0	16	12	4	2	0	0	0	3	1	1	2	8	1	20	23	1	1	0	20	2	113
14	0	86	2	2	0	3	6	0	3	1	1	0	0	0	0	4	1	0	1	0	3	1	112
57	0	42	0	0	0	2	1	0	3	0	1	0	0	1	0	0	1	0	0	0	43	1	93
28	2	1	2	1	0	0	1	1	9	2	1	0	0	0	0	38	15	2	0	0	12	1	86
80	0	0	0	1	1	1	0	0	3	0	0	2	1	0	0	32	8	1	0	2	2	2	54
17	0	0	0	1	0	0	0	0	1	0	22	0	2	0	1	5	0	0	1	2	0	5	39
58	0	0	0	0	0	0	0	0	29	0	3	0	0	2	0	0	0	0	0	0	5	0	38
15	0	8	3	4	0	3	0	0	9	0	0	0	1	0	0	2	0	3	1	0	2	1	36
3	6	0	1	3	1	1	0	0	0	8	0	2	2	7	0	1	1	0	0	0	3	1	35
4	3	0	0	4	0	0	0	2	0	1	0	0	0	1	0	5	0	0	18	0	1	0	34
72	0	5	0	0	0	0	0	0	13	2	0	0	0	1	3	0	0	0	6	1	4	0	34
38	3	0	9	6	2	5	0	0	0	0	0	0	2	0	0	0	1	3	0	1	1	0	32
44	1	0	2	0	0	3	0	2	7	2	0	1	0	0	0	1	0	0	9	1	3	1	31
23	1	0	1	1	6	1	0	1	0	3	0	0	0	6	0	7	0	0	0	1	0	3	30
24	0	0	7	3	0	0	0	0	0	4	0	1	0	4	0	3	0	2	1	0	3	1	28
32	0	1	0	0	0	0	0	0	5	0	1	1	0	0	0	1	1	5	13	0	1	1	28
26	2	0	0	0	0	2	7	0	1	5	0	0	0	3	0	3	2	0	0	0	5	1	27
29	3	2	1	1	0	0	0	0	5	0	0	1	0	2	0	9	0	0	0	0	1	4	27
12	0	0	2	1	1	0	3	0	0	0	0	0	0	1	0	2	1	0	12	0	2	0	24
54	0	0	0	0	0	0	1	0	0	0	0	0	0	0	17	0	0	0	0	0	5	0	22
-16	0	0	0	1	1	5	0	0	0	5	0	0	0	3	1	3	0	1	0	0	1	2	21
-48	0	0	0	0	0	1	0	0	0	0	17	0	0	0	0	0	0	0	0	0	1	0	19

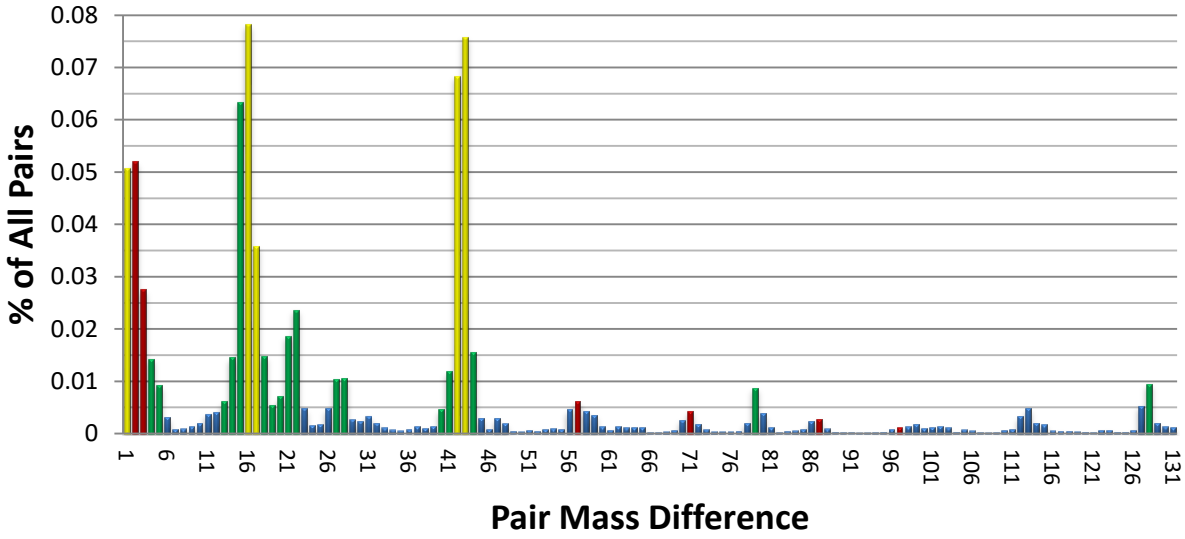


Figure 1.3: Histogram of pairwise deltas used to generate penalties for spectral alignment.

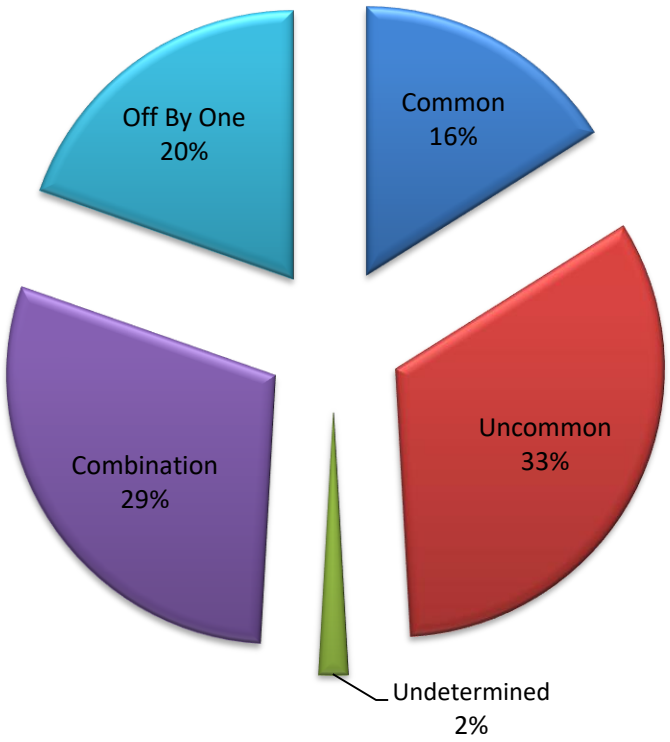


Figure 1.4: Pair Supported Modifications
Using Pair Supported Modification Calling (PSMC) rather than PSM frequency table easily identifies uncommon modifications. Of the modification masses identified 49% are verifiable as correct as is (either common or uncommon). Another 29% are correct in combination with other nearby modifications on the peptide or were combinations of more common modifications, and another 20% were simply off by one dalton due to the mass errors in the data. 2% of the modifications found were supported by the spectra but not readily identifiable as combinations of known modifications (possibly novel).

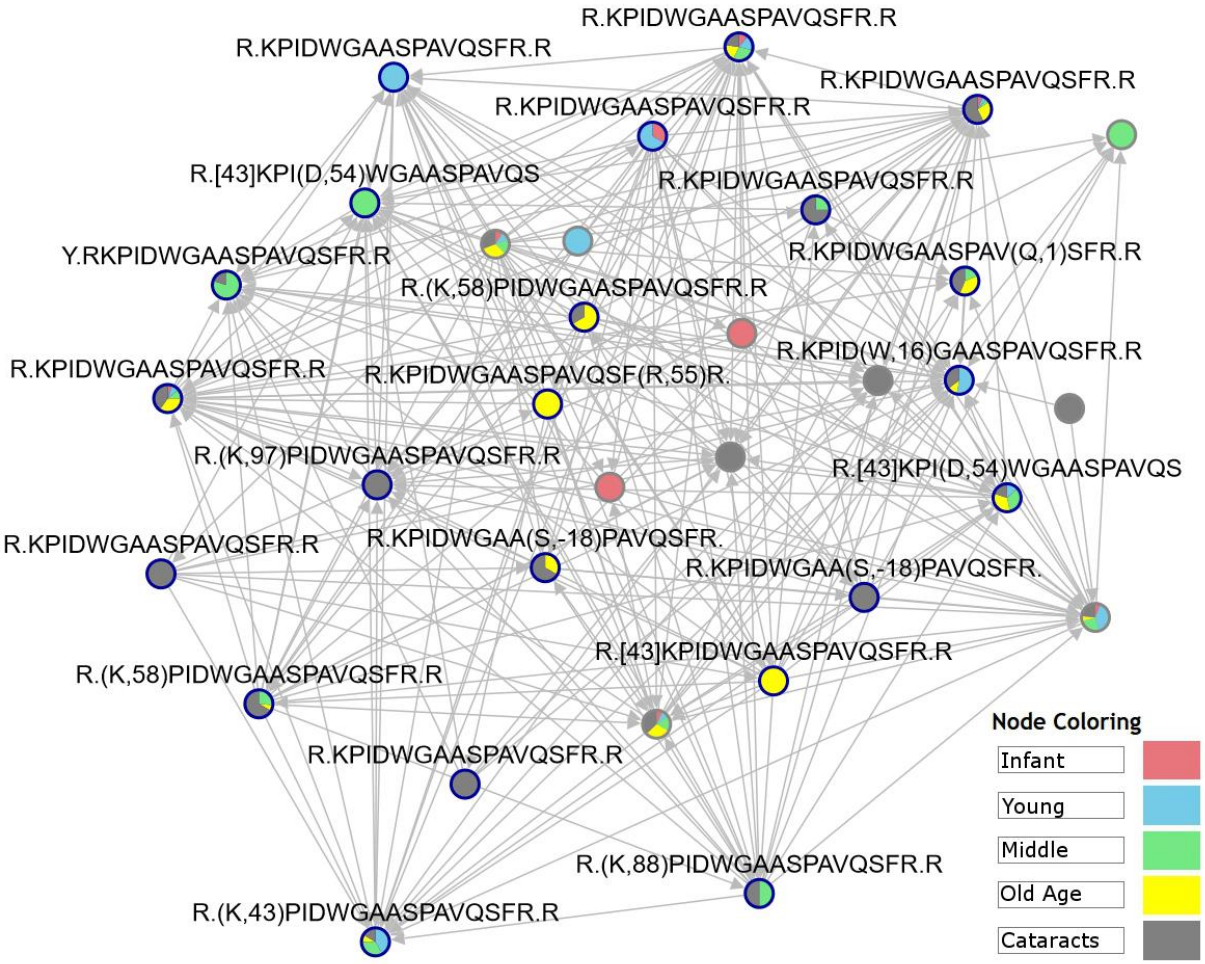


Figure 1.5: Snapshot of the network containing the KPIDWGAASPAVQSR peptide with many of the modifications present for this peptide along with pie charts showing the distribution of the peptides in the various cohorts.

Figure 1.6: All 256 manually verified discovered variants of the largest Protein Variant Region (PVR) identified in the region spanning amino acids 65 to 146 of the αA-crystallin protein (P02489).

Table 1.3: Common modifications with 10 or more variants in the RaVen offset frequency table.

Mass	Site	Classification	Description	#Variants
-48	M	Chemical Derivative	Homoserine lactone	14
-42	R	Artefact	Ornithine	10
-18	D	Chemical Derivative	Dehydration	16
	E	Artefact	Pyro-glu	12
	S	PTM	Dehydration	18
	T	PTM	Dehydration	23
-17	N	Chemical Derivative	N-Succinimide	19
	Q	Artefact	Pyro-glu	42
1	N	Artefact	Deamidation	339
	Q	Artefact	Deamidation	251
4	W	Chemical Derivative	Kynurenin	14
12	W	Artefact	Carbon	12
14	C	PTM	Methylation	84
16	M	Artefact	Oxidation	221
	W	Artefact	Oxidation	21
22	D	Artefact	Sodium adduct	36
	E	Artefact	Sodium adduct	52
	Q	Artefact	Sodium adduct	18
28	K	Artefact	Dimethylation	16
	N-term	Artefact	Formylation	12
	S	Artefact	Formylation	37
	T	Artefact	Formylation	14
32	W	PTM	Di-oxidation	14
42	K	PTM	Acetylation	24
	N-term	PTM	Acetylation	165
43	K	Multiple	Carbamylation	15
	N-term	Multiple	Carbamylation	341
54	R	Other	Methylglyoxal-derived hydroimidazolone	16
57	N-term	Artefact	Carboxyamidomethylation	44
58	K	Artefact	Carboxymethyl	26
72	K	PTM	Carboxyethyl	12
80	S	PTM	Phosphorylation	34

Table 1.4: Rare modifications observed with less than 10 variants in the offset frequency table but well supported by additional spectral pairs data.

Mass	Site	Classification	Description	#Vars
-41	R	Substitution	R→D	3
-34	F	Substitution	F→I/L	5
-32	M	Substitution	M-->V	7
-30	S	Substitution	S-->G	5
-28	N	Chemical Derivative	Pyrrolidone	1
-28	R	Substitution	R-->K	2
-27	K	Substitution	K→T	2
-22	H	Substitution	H-->N	2
-19	R	Substitution	R-->H	1
-16	L	Substitution	L-->P	5
	S	Chemical Derivative	Deoxy	4
-15	Q	Substitution	Q-->I/L	3
-14	I	Substitution	I-->V	5
-13	N	Substitution	N-->T	1
-1	C	Multiple	Deydro	1
10	S	Substitution	S-->P	4
12	E	Artefact	Carbon	1
	H	Artefact	Carbon	2
	T	Artefact	Carbon	3
14	D	PTM/SAAP	Methylation	2
	H	PTM	Methylation	6
	K	PTM	Methylation	2
	S	PTM	Methylation	5
16	F	Artefact	Oxidation	4
	H	Artefact	Oxidation	2
	K	PTM	Oxidation	3
17	E	Artefact	Ammonium	2
24	I	Substitution	I --> H	1
26	N-term	Other	Acetaldeyhde	4
	H	Other	Acetaldeyhde	4
27	S	Chemical Derivative	Ethyl amino	1
28	D	Artefact	Ethylation	1
30	V	Substitution	V-->E	1
34	H	Chemical Derivative	Chlorination	
38	D	Artefact	Cation:K	8
	E	Artefact	Cation:K	5
	K	Artefact	Cation:K	1

Mass	Site	Classification	Description	#Vars
40	N-term C	Artefact	Pyro-carbamidomethyl	7
41	N-term	Chem Derivative	Amidine	1
43	M	Artefact	Carbamylation	7
44	A	Substitution	A-->N	2
	K	Artefact	Carboxylation	8
	W	Artefact	Carboxylation	8
48	C	Chem Derivative	Trioxidation	6
53	E	Chem Derivative	Iron	1
57	K	Other	Carbamidomethyl	4
60	A	Substitution	A→M	1
64	C	PTM	Sulfur Dioxide	1
71	C	Artefact	Propionamide	2
72	H	Chem Derivative	Ethoxyformyl	2
	R	Multiple	Dihydroxyimidazolidine	2
73	I	Substitution	I-->W	1
80	T	PTM	Phosphorylation	8
86	C	Chem Derivative	Malonylation	1
87	C	Artefact	Acrylamide adduct	2
88	K	Other	Thioacyl	1
94	K	Other	Acrolein addition	1
97	K	Chem Derivative	Maleimide	5
101	C	PTM	HN2_Mustard	8
	K	PTM	HN2_Mustard	8
105	C	Chem Derivative	Pyridylethyl	5
106	K	Chem Derivative	methylsulfonylethyl	1
114	C	Chem Derivative	Dicarbamidomethyl	1
116	C	Chem Derivative	2-succinyl	8
119	K	Chem Derivative	Pyridylacetyl	1
120	C	Chem Derivative	Ethylsulfonylethyl	3
127	N-term	Chem Derivative	N-Succinimidyl-2-morpholine acetate	1
140	K	Chem Derivative	Maleimide	1
145	N-term	Chem Derivative	CAMthiopropanoyl	1
161	N	PTM	Hexosamine	2
162	C	Glycosylation	Hex	1
	K	Glycosylation	Hex	1
174	C	Chem Derivative	Thiadiazolydation of cysteine	1
209	C	Artefact	Carbamidomethylated DTT modification of cysteine	2

Table 1.5: Modification masses that are well supported by the data but have not been previously categorized in UniMod or in the literature.

Mass	Site	#Variants
25	D	1
88	K	6
115	K	2
118	Q	3

Table 1.6: Most modified single amino acid 'K' at position 159 in the βS-crystallin protein.

Most Modified Protein Position	Type
[42]KPIDWGAASPAVQSFR	Acetylation
(K,44)PIDWGAASPAV(Q,1)SFR	Carboxy
(K,58)PIDWGAASPAVQSFR	Carboxymethyl
(K,72)PIDWGASPAVQSFR	Carboxyethyl
(K,97)PIDWGAASPAVQSFR	Maleimide
(K,101)PIDWGAASPAVQSFR	HN2_mustard
(K,119)PIDWGAASPAVQSFR	Pyridylacetyl
[58](K,94)PIDWGAASPAVQSFR	Acrolein addition
R(K,14)PIDWGAASPAVQSFR	Methylation
R(K,28)PIDWGAASPAVQSFR	Formylation
R(K,57)PIDWGAASPAVQSFR	Carbamidomethyl
(K,127)PIDWGAASPAVQSFR	N-Succinimidyl-2-morpholine acetate
(K,140)PIDWGAASPAVQSFR	Malemide
(K,162)PIDWGAASPAVQSFR	Hexose glycosylation
(K,88)PIDWGAASPAVQSFR	Unknown

Table 1.7: Most modified peptide at position 236 (C-terminal) in the αA-crystallin protein.

Most Modified Peptide	Type
(Q,-17)WHLEGSFPVLATEPPK	Pyro-glu
(Q,-17)(W,4)HLEGSFPVLATEPPK	Pyro-glu & Kynurenin
(Q,-17)(W,12)HLEGSFPVLATEPPK	Pyro-glu & Carbon Adduct
(Q,-17)WHLEGSFPVLAT(E,22)PPK	Pyro-glu & Sodium Adduct
(Q,-17)WHLEG(S,28)FPVLATEPPK	Pyro-glu & Formylation
(Q,-17)(W,32)HLEGSFPVLATEPPK	Pyro-glu & Dioxidation
QWHLEGSFPVLAT(E,22)PPK	Sodium Adduct
(Q,-17)WHLEGSFPVLAT(E,38)PPK	Pyro-glu & Potassium Adduct
(Q,-17)(W,44)HLEGSFPVLATEPPK	Pyro-glu & Carboxy
Q(W,32)HLEGSFPVLATEPPK	Dioxidation
[43]QWHLEGSFPVLATEPPK	Carbamylation
QW(H,72)LEGSFPVLATEPPK	Ethoxyformyl
(Q,-17)W(H,72)LEGSFPVLATEPPK	Pyro-glu & Ethoxyformyl
(Q,-17)WHLEGSFPVLA(T,80)EPPK	Phosphorylation
[43]QW(H,26)LEGSFPVLATEPPK	Substitution H→Y
QWHLEGSFPVLATEPP(K,43)	Carbamylation
(Q,-17)WHLEGSFPVLAT(E,22)PP(K,43)	Sodium Adduct & Carbamylation
(Q,-17)WHLEGSFPVLATEPP(K,72)	Pyro-glu & Carboxyethyl
(Q,-17)WHLEGSFPVLAT(E,22)PP(K,72)	Sodium Adduct & Carboxyethyl
(Q,-17)WHLEGSFPVLATEPP(K,106)	Pyro-glu & Methylsulfonylethyl
(Q,-17)WHLEGSFPVLATEPP(K,114)	Pyro-glu & Double Carbamidomethylation

Table 1.8: Top 20 largest protein regions in terms of total variant count.

Protein	PVR Sequence	Count
CRYAA	SDRDKFVIFLDVKHFSPEDLTVKVQDDFVEIHGKHNERQDDHG-YISREFHRRYRLPSNVDQSALSCSLSADGMLTFCGPKIQ	294
CRBS	LSSCRAVHLPSGGQYKIQIFEKGDFSGQMYETTEDCPSIMEQFHMRE	102
CRBB1	RVGSVKVSSGTWVGYQYPGYRGYQYLLEPGDFRHWNEWGAFQPQMQSLRR	93
CRBB2	EKAGSVLVQAGPWVGYEQANCKGEQFVFEKGEYPRWDSWTSSR	92
CRBB1	TKGKGAPPAGTSPSPGTTLAPTTVPITSAKAAELPPGNYR	87
CRYAB	KYRIPADVDPLTITSSLSSDGVLTVNGPRKQVSGPER	86
CRBB1	LRDKQWHLEGSFPVLATEPPK	84
CRBA1	RMEFTSSCPNVSERSFDNVRSLKVESGAWIGYEHTSFCGQQFILERG	81
CRYAA	MDVTIQHPWFKRTLGPFYPSRLFDQFFGEGLFEYDLLPFLSSTISPYYR	80
CRBS	QYLLDKKEYRKPIDWGAASPAVQSFRR	76
CRBB2	RGLQYLLEKGDYKDSSDFGAPHPQVQSVRR	75
CRYAB	LRAPSWFDTGLSEMRLEKDRFSVNLDVKHFSPEELKVK	71
CRBS	RYDCDCDCADFHTYLSR	66
CRBB1	SDRLMSFRPIKMDAQEHKISLFEGANFKGNTIEIQGDDAPSLWVYGFSDRVGSVK	65
CRBS	CNSIKVEGGTWAVYERPNFAGYMYILPQGEYPEYQRW	63
CRYAB	KVKVLGDVIEVHGKHEERQDEHGFISREFHR	59
CRBA1	MTIFEKENFIGRQWEISDDYPSLQAMGWFNNEVGSMK	58
CRBB2	KKMEIIDDDVPSFHAHGYQEKVSSVR	54
CRBS	MSKTGTKITFYEDKNFQGRR	53
CRBB1	GFDRVRSIIVSAGPWVAFEQSNFRGEMFILEKGEYPR	51

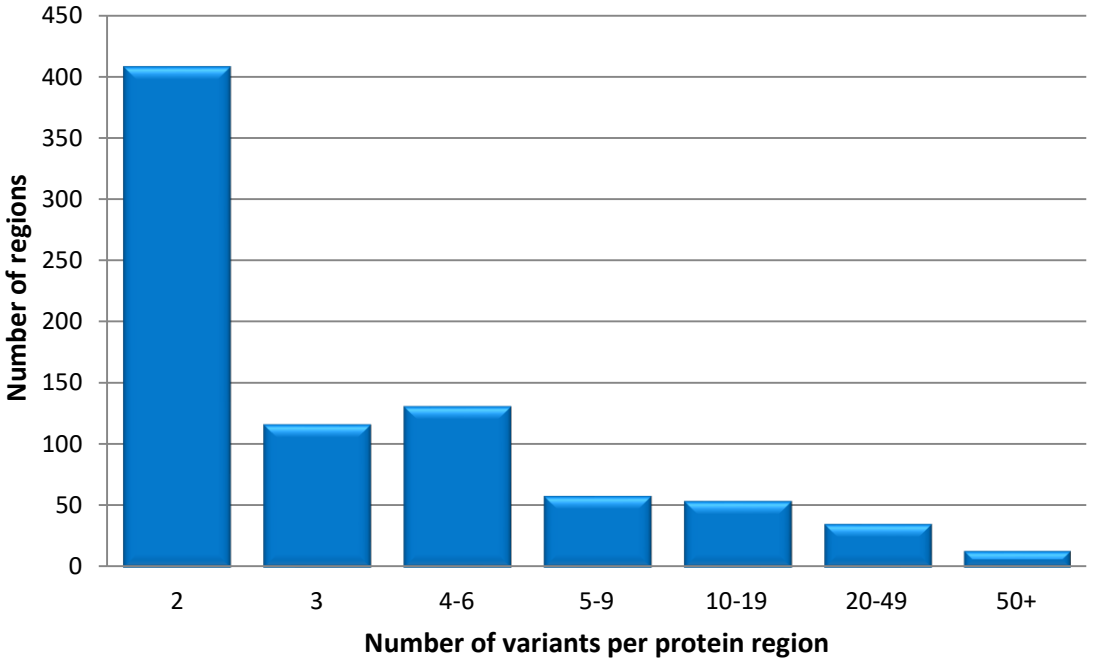


Figure 1.7: Histogram of number of variants in various protein regions throughout the sample.

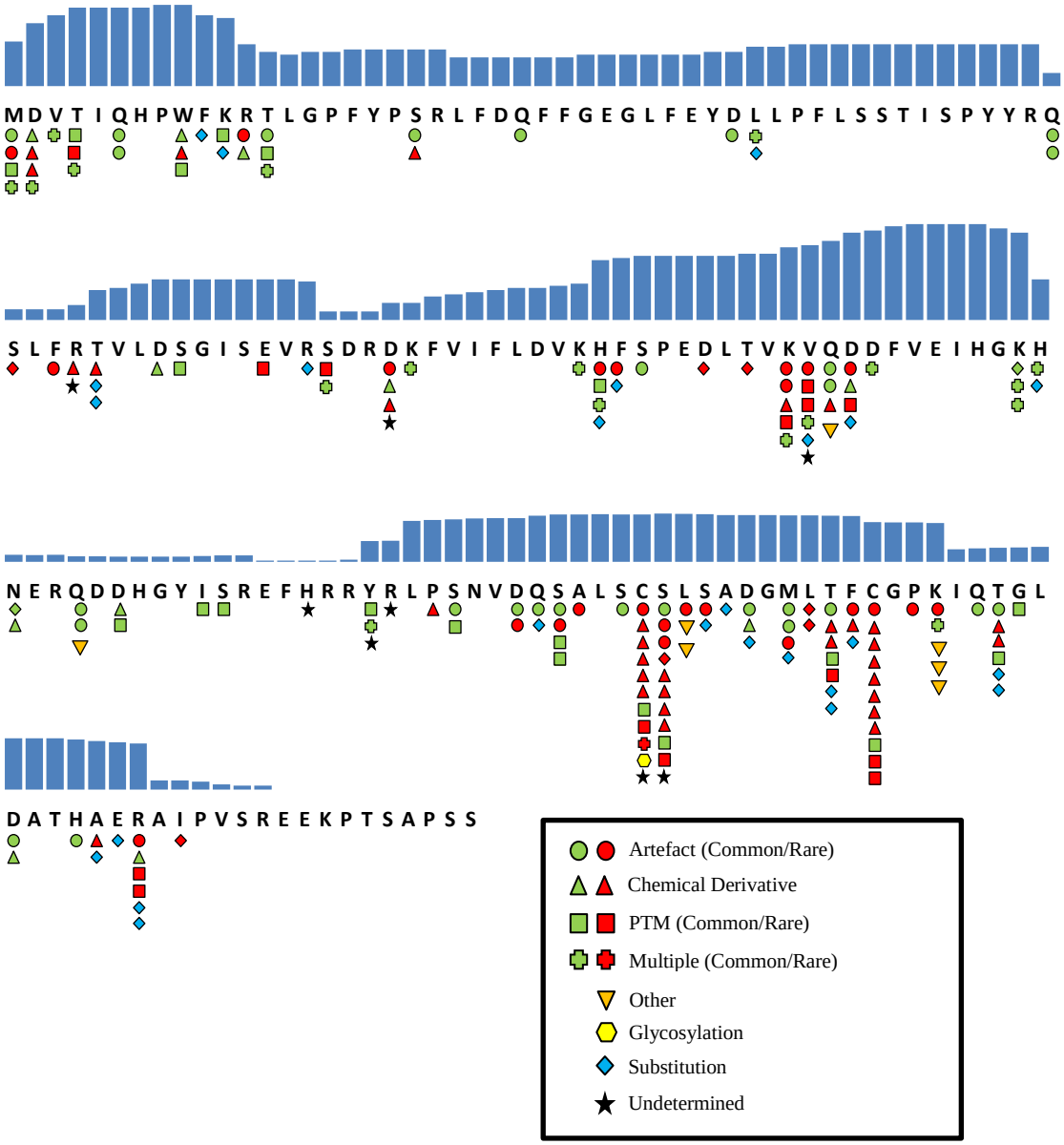


Figure 1.8: Diagram of variants and modifications of the CRYAA protein.
Above the sequence is the histogram of the number of variants that cover each amino acid, while below the sequence are symbols denoting what types of modifications were found at that location. Each symbol represents a unique modification found at that site. Even though the localization of these modifications may not be precise, the diagram still reveals the diversity of protein-level modifications, as well as the variation between areas of high and low modification rates.

Table 1.9: Top 10 modifications detected with the highest increase in spectral counts for the Cataract, Old Age, and Infant cohorts.

Peptide	Protein	Location	Classification	Cohort	Fold Incr
R.HWNEWDA(S,27)QPQLQSVR.R	CRBB3_HUMAN	180-195	Ethyl amino	Cataract	18.5
K.VQ(D,14)DFVEIHGK.H	CRYAA_HUMAN	89-99	Methyl	Cataract	16.3
M.[42]ASDHQTQAGKPQ(S,10)LNPK.I	CRBB2_HUMAN	2-18	Acetyl & Substitution S→P	Cataract	9.4
R.YRLPSNVDQSALSCSLSADGMLTF(C,120)GPK.I	CRYAA_HUMAN	118-145	Ethylsulfonylethyl	Cataract	7.4
R.SYETTTD(C,105)PNLQPYFSR.C	CRGC_HUMAN	16-32	Pyridylethyl	Cataract	5.4
R.GFQYVLE(C,14)DHHSGDYK.H	CRBA4_HUMAN	159-174	Methyl	Cataract	5.2
K.EY(R,55)KPIDWGAASPAVQSFR.R	CRBS_HUMAN	156-174	Undetermined	Cataract	4.7
R.HWNEWGAFQPQMQ(S,-30)LR.R	CRBB1_HUMAN	215-230	Substitution S→G	Cataract	4.2
C.SLSADGMLTF(C,14)GPK.I	CRYAA_HUMAN	132-145	Methyl	Cataract	4.2
D.(S,80)SDFGAPHPQVQSVR.R	CRBB2_HUMAN	174-188	Phospho	Cataract	3.9
K.(E,22)YR(K,44)PIDWGAASPAVQSFR.R	CRBS_HUMAN	156-174	Sodium & Carboxy	Old Age	12.35
R.GYQYILE(C,14)DHHGGDYK.H	CRBA1_HUMAN	178-193	Methyl	Old Age	8.7
M.(D,25)IAIHHPWIR.R	CRYAB_HUMAN	2-11	Undetermined	Old Age	8.2
R.T(N,-17)AMSGLVR.A	BFSP2_HUMAN	44-52	N-Succinimide	Old Age	6.4
R.G(E,22)YPSWDAWGG(N,1)TAYPAER.L	CRBA4_HUMAN	72-90	Sodium	Old Age	6.3
N.[42]PTPGSLGPWK.I	CRBA1_HUMAN	23-32	Acetyl	Old Age	4.7
K.VQDDFV(E,22)IHGK.H	CRYAA_HUMAN	89-99	Sodium	Old Age	4.4
R.GEMFILE(K,58)GEYPR.W	CRBB1_HUMAN	111-123	Carboxymethyl	Old Age	4.4
T.[26]TLAPTTVPITSAK.A	CRBB1_HUMAN	38-50	Acetaldeyhde	Old Age	4.2
K.GDFSGQMYETT(E,22)DCPSIME(Q,1)FHMR.E	CRBS_HUMAN	102-125	Sodium	Old Age	4.0
R.HWNE(W,16)GAFQPQ(M,16)QSLR.R	CRBB1_HUMAN	215-230	Oxidation x2	Infant	274.2
K.VLEGV(W,44)IFYELPNYR.G	CRBS_HUMAN	132-146	Carboxylation	Infant	45.4
R.SLHVLEGC(W,16)VLYELPNYR.G	CRGC_HUMAN	123-140	Oxidation	Infant	19.9
R.[28]YRLPSNVDQSALSCSLSADGMLTFCGPK.I	CRYAA_HUMAN	118-145	Formyl	Infant	19.3
_.[42](M,16)DIAIHHP(W,16)IR.R	CRYAB_HUMAN	1-11	Acetyl + Oxidation x2	Infant	19.2
_.[42](M,-48)DIAIHHPWIR.R	CRYAB_HUMAN	1-11	Acetyl & Homoserine lactone	Infant	13.5
R.HWNEWGAFQPQ(M,16)QSLR.R	CRBB1_HUMAN	215-230	Oxidation	Infant	12.4
K.[26]HFSPEDLTVK.V	CRYAA_HUMAN	79-88	Acetaldeyhde	Infant	11.7
K.GL(M,16)(M,16)ELSEDCPSIQDR.F	CRGC_HUMAN	100-115	Oxidation x2	Infant	11.5
K.[26]IQTGLDATHAER.A	CRYAA_HUMAN	146-157	Acetaldeyhde	Infant	9.4

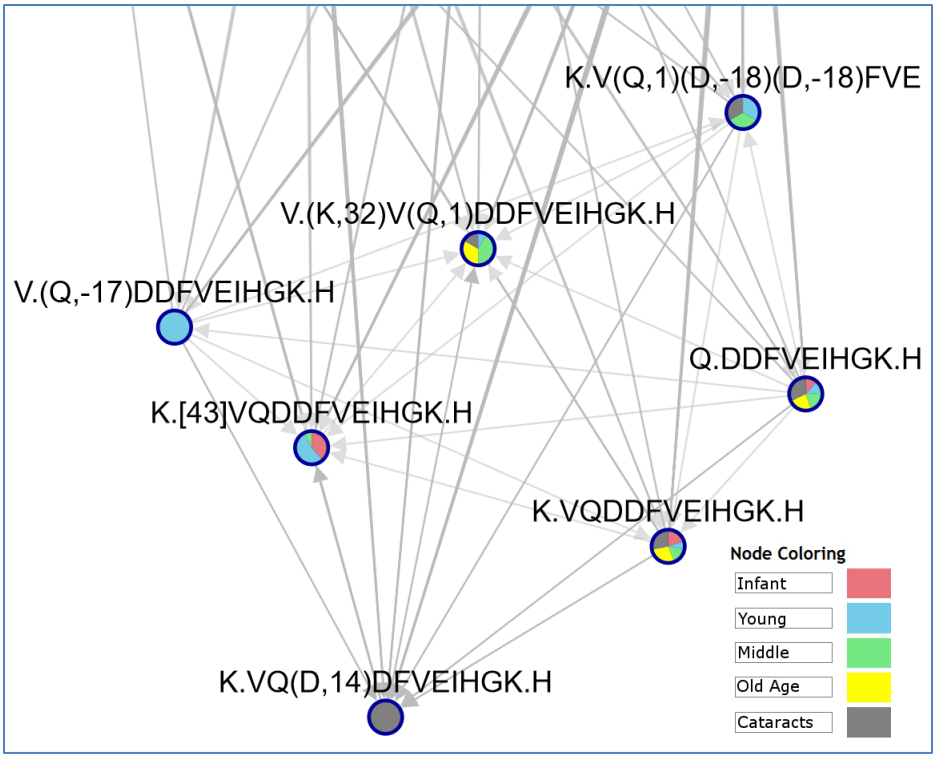


Figure 1.9: A section of the spectral network showing the VQ(D,14)DFVEIHGK peptide and representative neighbor peptides matched by spectral alignment, as well as per-node pie charts with cohort-specific relative abundance showing that this peptide is only present in the Cataracts cohort.

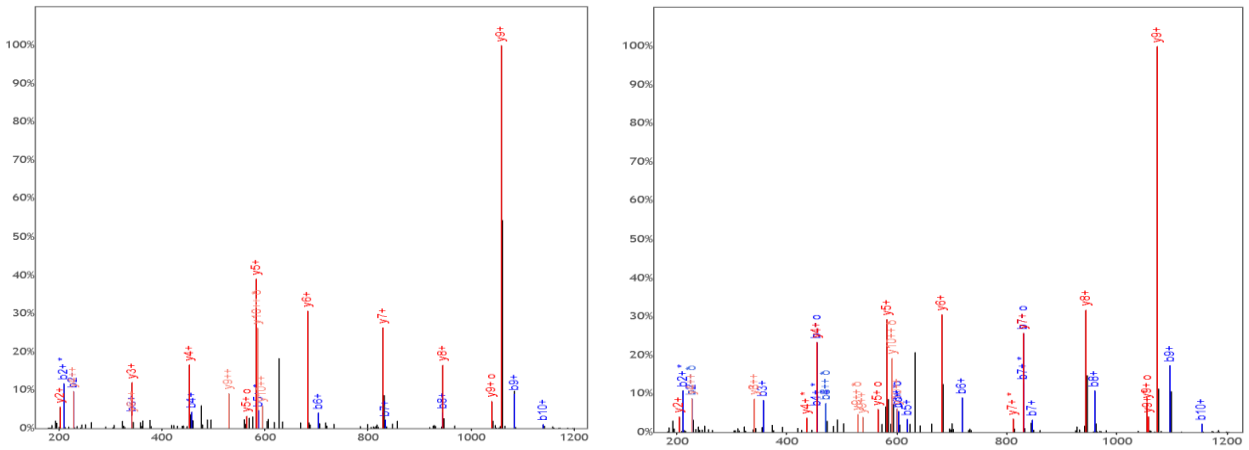


Figure 1.10: Spectra showing the match between modified and unmodified forms of the VQDFVEIHGK peptide.

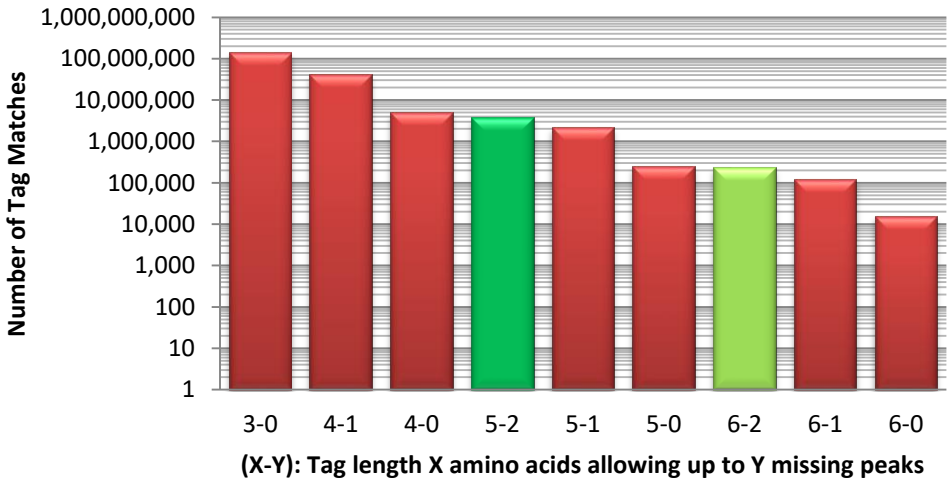
Table 1.10: Table of all lens data files used in analysis.

Filename	Number of Spectra	Condition
0d.mgf	13206	Any
2yo_insol.mgf	56620	Insoluble
2yo_sol.mgf	53945	Soluble
18yo_insol.mgf	67847	Insoluble
18yo_sol.mgf	56393	Soluble
32yo_sol.mgf	13304	Soluble
35yo_insol.mgf	60445	Insoluble
35yo_sol.mgf	52286	Soluble
70yo_insol.mgf	64112	Insoluble
70yo_sol.mgf	65263	Soluble
93yo_insol.mgf	13404	Insoluble
93yo_sol.mgf	17171	Soluble
70yo_cat_insol.mgf	59605	Insoluble (with Cataracts)
70yo_cat_sol.mgf	62325	Soluble (with Cataracts)
93yo_cat_insol.mgf	65283	Insoluble (with Cataracts)
93yo_cat_sol.mgf	65082	Soluble (with Cataracts)

(a)

Tag Length	Allowed Tag Gaps	Target Matches	Contigs with Matching Tags
3	**0**	143,634,098	2683
4	**1**	41,965,703	3524
4	**0**	5,055,268	2003
5	**2**	3,772,770	3828
5	**1**	2,154,127	3059
5	**0**	249,219	1485
6	**2**	235,791	3494
6	**1**	122,534	2384
6	**0**	15,596	930

(b)



(c)

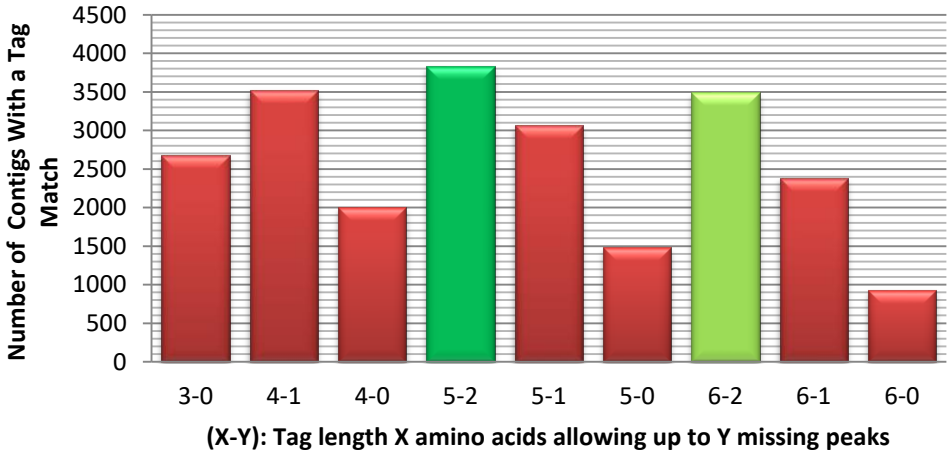
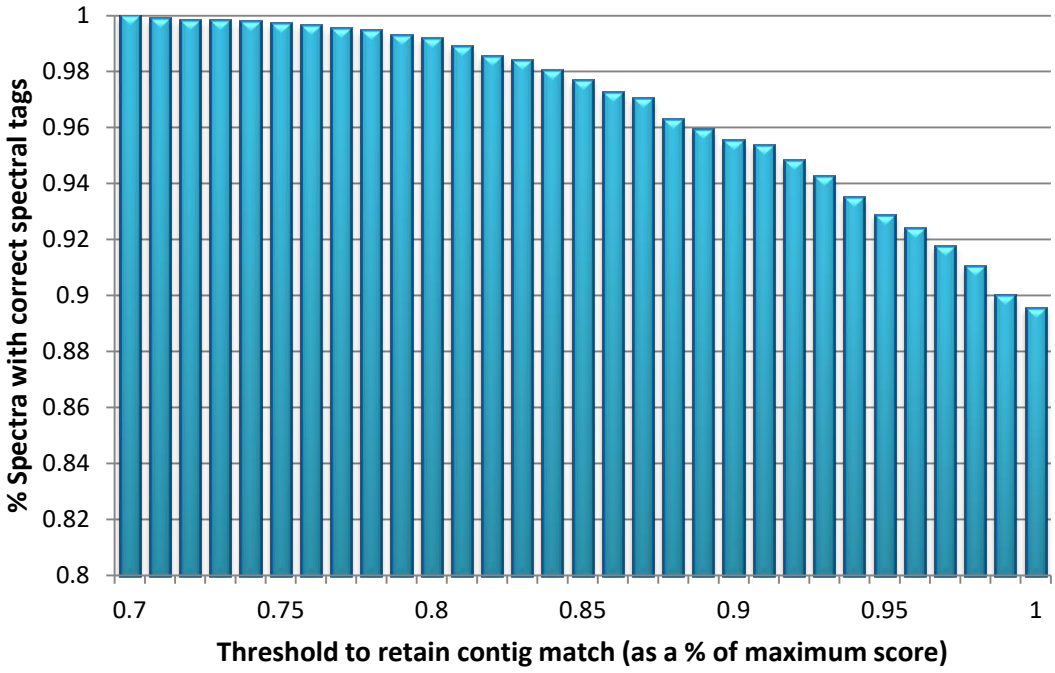


Figure 1.11: Number of contig tags generated and the number of contigs for which a tag can be found. Length 5 with 2 allowed missing peaks gives best performance while maintaing "reasonable" number of matches. One could also consider tags of 6-2 if willing to suffer a drop in performance in return for a 10x speedup in processing. However processing speed for contig alignment step is not really a large factor. (a) Table showing tag matches versus number of contigs with at least one match. (b) Number of tags rises rapidly with the use of short tags. (c) Number of correct tags does not increase with smaller tag sizes, but rather when using large tag sizes while allowing missing peaks.

(a)



(b)

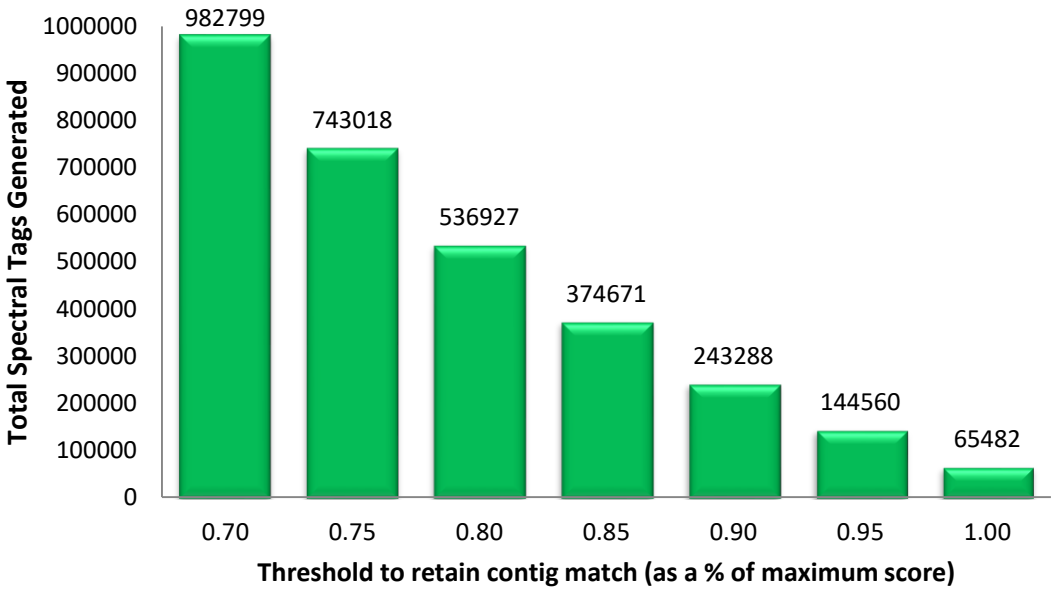


Figure 1.12: Choice for percentage of contig matches kept to form spectral tags for 46,396 spectra in 8348 contigs with at least one correct match.
(a) Ability to find correct PSM drops off slowly before 80% retention threshold for contig tags. At an 80% threshold more than 99% of all spectra tags are still retained. (b) The total number of spectral tags falls rapidly as retention threshold becomes stricter.

Figure 1.13: Pre-computed Gap-Alignment Block (GAB)
An example of a pre-computed gap-alignment block for the amino acid string "MN". Each bin represents a gap mass or mass difference between two peaks in the spectrum of interest, so bin 144 is a (rounded) mass of 144 daltons. A mass of this size corresponds to the sum of masses for Methionine and Asparagine with no modifications., therefore the penalty for this match is 0. All other bins are given values that correspond to the lowest penalty explanation for an annotation with that mass. The bin at 145 corresponds to a +1 modification which is explained by the known modification of Deamidation on Asparagine and therefore has a penalty of K, the penalty for one known mod. Similarly the mass bin at 160 corresponds to the known +16 modification of Oxidation of Methionine and also receives the value K. The bin at 161 corresponds to a modification of +17, or simultaneous modifications of Oxidation of Methionine and Deamidation of Asparagine, and therefore has a penalty of 2 known modifications or 2*K. The bin at 162, corresponds to a modification of +18, which for the purposes of this example is a putative modification discovered in the pair data, and therefore receives the penalty value (P_{18}) for a modification of +18 daltons computed from the frequency of a +18 shift in the pairs data. The bin at 163 corresponds to a +19 modification which is best explained by a putative modification of +18 on M combined with a Deamidation of Asparagine and receives a penalty for both or: P_{18} + K. However, the bin at 164 has no better explanation than an unknown modification of +20 daltons (two unknowns that added to +20 would be even worse) and therefore receives the penalty for an unknown modification: U as do all other mass bins that cannot be explained by sums of known and/or putative modifications. Once computed this gap-alignment block can be used to look up a match between the database string "MN" and any gap between two peaks in the spectrum that has mass between 114 and 1499 daltons in O(1) time. Note: In no gap of less than 114 daltons could actually be considered since that is the minimum mass of two amino acids (2 Glycines).

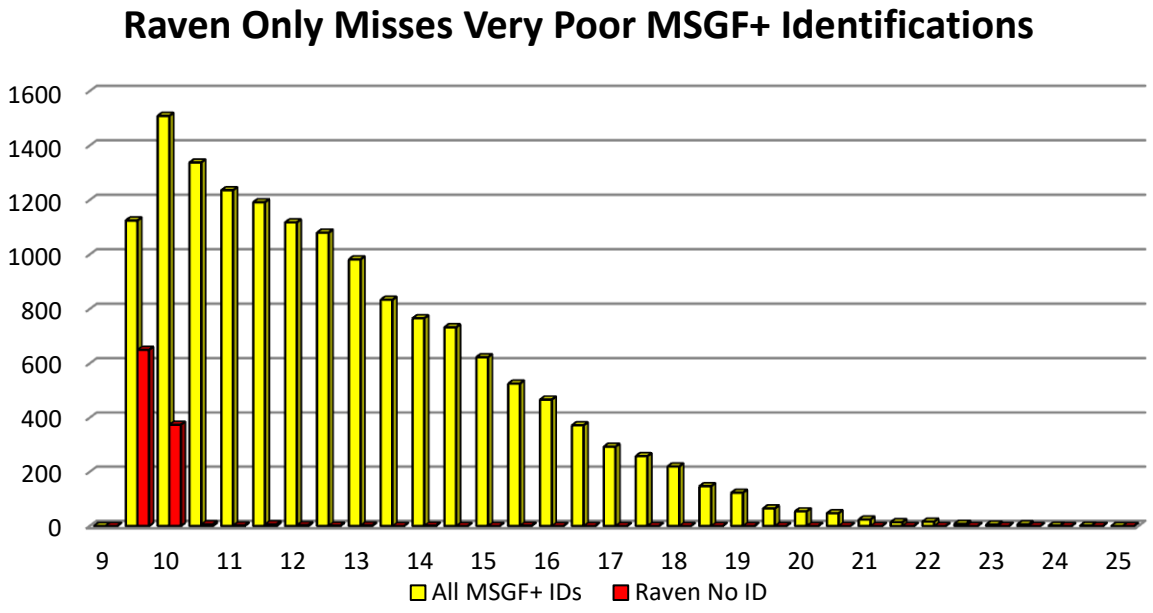


Figure 1.14: Raven Only Misses Very Poor MSGF+ Identifications
Histogram of the –log(p-values) of all the identifications made by MSGF+ versus only those that Raven failed to identify. The values for the MSGF+ identifications range from 9.5 to 24.5, however for the 1046 PSMs which Raven failed to identify, 98% had a value between 9.5 and 10.0. Raven only misses MSGF+ identifications which have a low probability of being correct.

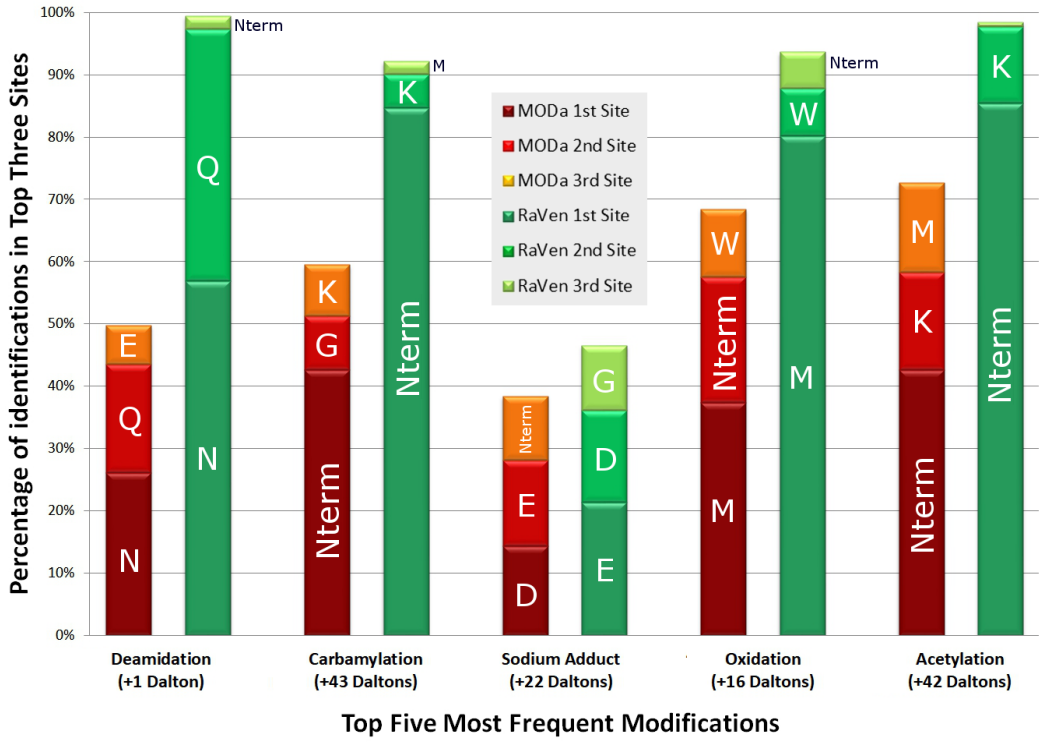


Figure 1.15: Percentage of identifications localized to the top three sites with the highest counts in the offset frequency table for the five most frequent modifications shows that RaVen greatly improves localization over MODa.

CHAPTER 2 - Discovery of post-translational modifications and proteomics diversity in colorectal cancer

2.1. Abstract

High throughput proteomics analysis of tumor tissue allows for the observation of translated gene products as well as post-translational modifications which could be related to altered cancer biology. Recognizing this potential, the NIH Clinical Proteomic Tumor Analysis Consortium (CPTAC) program has supported the acquisition and study of reference collections of tumor proteomics mass spectrometry data. While these collections have analyzed changes in protein expression and translation of genomic events, limited attention has been given to the detection and potential role of post-translational modifications in tumor tissues. Using a newly-developed blind database search approach (RaVen) we reanalyze the CPTAC colorectal cancer data (MSV000079852) and reveal dozens of previously undetected post-translational modifications in over 15,000 modified peptide variants, including some found to be potentially differentially expressed between healthy and tumor samples. In addition to detecting both known and novel modifications, as well as rare modifications and amino acid polymorphisms, our analysis further reveals the occurrence of hyper-modified protein regions in colorectal cancer including a single region in hemoglobin subunit beta (HBB) covered by over 70 distinct peptide variants. The novel detection of such proteomics diversity illustrates how blind modification searches can reveal novel proteomics events of potential biological significance, including the detection of previously-undetected modification variants mapped to functional protein regions.

2.2. Introduction

Deeper understanding of the molecular basis of cancer requires large-scale analysis of not only genomes but also of altered cancer proteomes [46], [47], [48], [49], [50], [51], [52], [53]. To address this need, in 2006 the National Cancer Institute (NCI) at the National Institutes of Health (NIH) started the Clinical Proteomic Tumor Analysis Consortium (CPTAC) and held their first annual meeting in 2008 [54]. Since that time, NCI has continued this effort with follow-on funding in 2011 and established 8 lead centers and many collaborating groups around the world. One result of this effort is the CPTAC Data Portal, which is a repository of all 6.3 TB of proteomic data collected by Proteome Characterization Centers (PCCs) in the consortium [55]. However, to date, the primary focus of studies analyzing this data has been towards the discovery of genetic variants in cancerous tissue [56], [57], [58] or have sought to quantify protein abundance to examine differences in protein expression between normal and cancerous tissue [49], [50], [52], [53]. However, while post-translational modifications and sequence cleavage events are key to protein-level regulation of biological processes and have been repeatedly implicated in disease phenotypes, the analysis of these events has not been a primary focus of CPTAC studies of colorectal cancer proteomics data, potentially leaving a wealth of proteomics information undiscovered within this community resource dataset.

One of the largest studies conducted by Zhang et al. [46] in 2014, performed a proteomic analysis using three different analysis tools: Pepitome [59], Myrimatch [60], and MS-GF+ [36]. Pepitome is a spectral library matching algorithm and therefore can only match previously annotated spectra and cannot make discoveries of any new PTMs. Myrimatch and MS-GF+ are both database matching algorithms that must use a list of a-priori specified PTMs, which in this case were set to allow only three modifications: oxidation of methionine, n-terminal acetylation

and n-terminal pyroglutamine. In 2010, Dasari et al, developed a new search method called TagRecon and tested it on yeast whole cell lysate samples from the CPTAC data. They compared their results with those from Inspect, X!Tandem and Myrimatch, but all these database engines were configured to allow only the same three modifications: oxidation of methionine, n-terminal acetylation and n-terminal pyroglutamine. In 2016, Rudnick et al created the Common Data Analysis Platform (CDAP) [61] which used only MS-GF+ for the identification of peptides in all data sets. In addition, because the standard database search techniques such as Myrimatch, MS-GF+ and others like Sequest [1] or Mascot [2] cannot identify variant peptides not in the database the authors in Zhang et al. were forced to create a customized database using matched RNA sequences from RNA-seq data.

Our recently created RaVen methodology (Bernstein, L., Wertz, J., Na, S., and Bandeira, N., 2018) is a blind database search approach allowing for unexpected and previously unknown modifications to be discovered in complex biological samples. Because such an approach can yield modifications of any mass, it is also capable of detecting somatic protein variants as amino acids with mass modifications. This allows RaVen to detect a wide range of proteogenomic events that may be linked to disease phenotypes. Using 12 case and 12 control samples from the CPTAC colorectal data set we show how RaVen can reveal protein regions of high variability, previously undetected amino acid substitutions, and post-translational modification events with high differential expression in the sample. We demonstrate that using a novel technique such as RaVen to reanalyze existing community databases has great potential in expanding our knowledge of disease and cancer proteome diversity.

2.3. Material and Methods

Our RaVen (Bernstein, L., Wertz, J., Na, S., and Bandeira, N., 2018) workflow builds on a variety of spectral clustering, alignment and filtering algorithms to reduce the size of the input space and allow a deep investigation of the spectra of greatest interest (those with rare or unknown modifications). Input spectra are first clustered using the MS-Cluster software [31], to reduce the size of the input spectra data. Each cluster is represented by a cluster-consensus spectrum and the resulting identifications are assigned to all corresponding cluster members.

This reduced set of clustered spectra then becomes the input to two parallel processes. First, the clustered spectra are input to the MS-GF+ [3] algorithm and the identifications from M-SGF+ are used in multiple phases of the follow-on processing. Second, the clustered spectra are input to the PepNovo [62] scoring method. PepNovo uses a likelihood ratio model to determine whether the peaks observed in the mass spectrum are more likely to have been produced under this fragmentation model than under a model that treats peaks as random events. PepNovo replaces the peak intensities with likelihood values enhancing the contrast of B and Y ion pairs while reducing other (presumably noise) peaks creating prefix-residue mass (PRM) spectra [63]. The PRM spectra along with the MS-GF+ identifications are used as inputs to the AlignGF algorithm [12] which performs pairwise alignments. The resulting pairs are used for creation of spectral networks [64] as well as identification of putative modification masses. The spectral networks are then input to the MetaSPS [30] algorithm which assembles cluster-consensus spectra from peptides with overlapping sequences (i.e., contigs) and derives short de novo sequences (tags) to match to the database and identify unmodified contigs. It is these contigs that are first identified by RaVen.

To reduce the overall number of database locations that are required to be matched to the contig-consensus spectra, RaVen uses de novo sequence tags similar to that originally proposed by Mann and Wilm [15] and used in various related approaches [38] [9] [18] [8] [39]. These tags are constructed by connecting contig-consensus spectrum peaks whose masses differ by the mass of one or two amino acids. Accuracy of de novo sequencing has been shown [30] to be much higher in these contig-consensus spectra than in non-assembled individual spectra because noise peaks are mostly eliminated and likelihood scores from peaks are combined from the multiple assembled spectra. At all locations where a contig tag is matched to the database, RaVen performs a tag-extension procedure and keeps only the best matches. Using this final set of extended tags, the assembled spectra are aligned to the highly reduced set of database locations with a penalty based spectrum/sequence alignment algorithm (Bernstein, L., Wertz, J., Na, S., and Bandeira, N., 2018). The penalties are divided into three categories: known modifications, putative modifications and unknown modifications. Known modification penalties are fixed at a very small value and used when introducing a modification of a-priori known type (such as M+16 for oxidation of methionine). Putative modifications are discovered using the information obtained from the spectral pairs output of the AlignGF algorithm creating a histogram of the parent mass differences to capture possible modification masses that occur frequently in the sample. These frequently occurring masses are indicative of possible modifications in the peptides and from them we derive putative modifications and set their penalty proportional to their probability of occurrence in the sample pairs. Unknown penalties are set to a large (greater than the largest putative penalty) value and used for all modifications that are in neither the known or putative categories. Using this method, RaVen finds alignments for all spectra in all contigs and then uses a standard False Discovery Rate (FDR) method [44] with the target

database and a shuffled decoy database to determine the set of identifications. It then removes any spectra identified by RaVen from the list of MS-GF+ identifications, combines the remaining MS-GF+ identifications with all RaVen identifications, and thresholds the new combined list at 1% FDR to arrive at the final set of peptide-spectrum identifications. Since site localization is often difficult to achieve in blind searches, RaVen uses the concept of a peptide `variant' to better assess the quality and significance of the peptide identifications. For any given spectrum identification S, RaVen defines the variant for S as the pair (P,m) where P is the peptide sequence and m is the sum of all modification masses on P. RaVen then enforces variant-level FDR (with variant scores set to the best score of all spectra identified to the same variant), which is stricter and more accurate than the traditional 1% FDR at the level of peptide-spectrum matches and avoids the problems of false localization in blind search results.

Putative modification masses in blind modification searches are typically reported as an offset frequency table [6] with one column per amino acid (plus columns for N/C-termini) and one row per mass offset with cell values reporting the number of PSMs identified for each pair. This approach, while widely used, has severe disadvantages in that finding biologically relevant modifications within the table is hampered by the sheer size of the table and large number of non-zero entries. Blind searches can also spread the same modification over multiple cells in the table due to the combination of smaller masses into a single modification mass and errors due to mass accuracy, therefore correctly identified modifications that occur infrequently can be difficult to pick out even when an algorithm correctly identifies them. Modification masses resulting in large numbers of cell counts are typically not biologically relevant and instead are more likely the result of sample handling modifications while rare modifications are almost always the results of biologically relevant processes. This means that finding modifications of

interest can require evaluating individual identifications by hand. RaVen's use of variants alleviates these issues by reporting counts of variants rather than counts of PSMs. RaVen also uses a localization procedure that assigns partial credit to all amino acids in the variant where the modification location cannot be determined due to lack of spectrum peaks. RaVen then iterates through the table selecting the site with the globally highest counts for all variants containing that amino acid and mass combination and removing counts for neighboring sites. This procedure is repeated until all modification masses have been assigned a single amino acid site. Furthermore RaVen can verify the mass modifications against information from pair alignments obtained during the AlignGF stage of our workflow and use information from spectral network neighbors to provide independent verification of the correctness of assigned modification masses.

2.4. Results

Using RaVen's localized, variant-based offset frequency table we compiled a list of the twenty most frequently occurring modifications in the data (Table 2.11). As expected, this list is topped by the commonly occurring modifications that restricted algorithms normally include in their searches, and that were used in previous examinations of the CPTAC data [46], [61]: i.e. -oxidation, deamidation, pyro-glu, and acetylation. Also comprising nearly 50% of the list are other common sample handling artefacts such as carbamylation, formylation, carbamidomethyl, and sodium adduct. As we anticipated, these modifications tell us little or nothing of biological relevance, and instead we must look to the rarer modifications that go undetected by most standard algorithms.

In Table 2.12 we see the list of top twenty rare modifications being observed with ten or fewer identifications. Contrary to what we see in Table 1, there are only four artefactual

modifications amongst these rare modifications: oxidation (on two different amino acids), carbamidomethyl, and dicarbamidomethyl. Eight of the remaining modifications are identified as amino acid substitutions, which is not surprising given the nature of the data set. To confirm these rare observations RaVen uses spectral networks to compare the identified spectra to other similar spectra.

One such example amongst the rare modifications is a -30 dalton modification on methionine, which corresponds to a methionine to threonine substitution. This substitution is the result of a single nucleotide polymorphism (SNP) of AUG to ACG and RaVen detects this change on multiple proteins (HBB, H2B1K, and FABPL) with very high confidence. Along with the direct spectral evidence in support of this identification (see Figure 2.16), RaVen also finds corroborating evidence for this identification in the spectral networks where the modified variant is aligned to a spectrum of an unmodified version of the same peptide that has nearly identical peak structure (see Figure 2.17). Although we find no substitutions of this type reported on any of these three proteins in UniProt, there is a reported methionine to lysine (AUG to AAG SNP) unstable hemoglobin variant reported [65] at that location lending credence to the possibility that other such SNPs may exist at that location.

Similarly, the rarest modification in the list, n-succinimidyl-2-morpholine acetate (SMA) is detected on a mere 5 peptides; however, RaVen shows strong support for this modification in the networks. In Figure 2.18 we see that not only is the SMA identification of the individual spectra very strongly supported by the matching peaks, but additionally the network shows a very similar spectral structure in a variant with the much more common +26 acetylaldehyde modification on the n-term.

RaVen also performs analysis of protein regions, some of which contain unusually high diversity in terms of peptide cleavages and modification masses that may point to specific regions of biological interest on certain proteins. To accomplish this analysis, RaVen identifies a Protein Variant Region (PVR) as a region where each peptide overlaps with at least one other peptide by at least 50% of its amino acids. Most proteins (~90%) contain 20 or less identified peptide variants but RaVen identifies over 70 proteins with more than 100 variants. Not surprisingly, as shown in Figure 2.19, the top 10 proteins with the most variants contain many of the longest proteins; however, the number of discovered variants is not proportional to protein length, rather the protein with the highest variant density (Table 2.13) is the short hemoglobin protein sp|P68871|HBB_HUMAN. Within this protein there is a wide range of coverage over the entire length of the protein (Figure 2.20). While some amino acids have no peptides which include their position, others have over 70 variants in the protein variant region.

The region of highest variability occurs on hemoglobin protein HBB from position 66 to position 85. Within this region of 20 amino acids we identify 71 verifiable unique variants (see Figure 2.21), including 12 different cleavage variants, 10 variants with no modification, and 61 variants with at least one modification. In addition, RaVen detects 25 variants with modifications that correspond to amino acid substitutions. These represent 16 unique substitutions since some variants contain duplicate substitutions in combination with additional modifications. Seven of these substitutions have previously been reported in UniProt; however, the other 9 substitutions are novel discoveries by RaVen. One of the detected substitutions corresponds to a +16 dalton modification near the n-term and might be attributed to the more common oxidation on lysine; however, we report these modifications as valine to aspartic acid substitutions (a substitution reported in UniProt at this protein position) based on information from multiple identifications

which reveals that on three of the four variants with this modification, n-terminal lysine oxidation is not possible due to the presence of n-terminal formylation or carboxylation.

While we do not presume that all discovered modifications are localized precisely (this is the reason we advocate using peptide variants to measure performance of blind searches), RaVen brings together multiple sources of information to help make a final determination on the most likely annotation. For example, we report the variant $K^{-1}V^{-28}$LGAFSDGLAHLDNLK that has both a -1 dalton modification and a -28 dalton modification on consecutive amino acids which generally is an unlikely annotation, particularly in light of the fact that there is not an intervening peak in the spectrum to support splitting the -29 mass into two modifications. While it is not possible to prove that RaVen's interpretation is the proper one, RaVen has multiple lines of evidence to support this choice. First, the spectral network shows a good correlation to a variant with only the -28 dalton modification. Second, there are numerous instances of the -28 modification on valine (valine to alanine substitution) and the -1 modification on lysine (lysine to allysine substitution) individually throughout the data. Third, we only detect the presence of the -29 mass modification on the n-term lysine when followed by a valine.

While proteogenomics algorithms have been developed to search mass spectrometry proteomics data against new genomic or transcriptomic sequences [66], there has been limited analysis of the potential of blind modification searches to either dispute or confirm the correctness of proteogenomics identifications – a comparison that is especially relevant when claiming amino acid polymorphisms or short sequence extensions (e.g., novel splicing events or exon boundaries) whose induced mass offsets might also correspond to modification masses. To illustrate this potential, we performed a detailed comparison of the RaVen identifications and the official CPTAC proteogenomics results for patient TCGA-AA-3518-01A-11, which consists of

65,723 peptide spectrum matches (PSMs) released by the CPTAC data portal for this analysis. While the high number of spectrum identifications initially suggests the potential for detection of many novel translated sequence variations, our remapping of the reported peptide sequences to the UniProt reference human proteome reveals that only 0.06% (36 PSMs) actually cover sequences that are not already reported in UniProt – a significantly lower level of detection of proteomics diversity than what we find through RaVen blind modification search. But since proteogenomics and blind modification searches consider distinct search spaces for spectrum identification, parallel analyses of the same data also provide the opportunity to contrast results from both searches to either challenge or increase confidence in the detection of novel proteomics events. Illustrating this concept, we focus on one specific polymorphism event detected in patient TCGA-AA-3518-01A-11, identified to SALFAQINQGESITHALK by the proteogenomics search and matched by RaVen's detection of a -16 Da modification in its identification SS^{-16}LFAQINQGESITHALK (Alanine is 16 Da lighter than Serine). In addition to explaining the detected delta mass as a translated polymorphism supported by mRNA data, this match further helps explain the +10 Da modification discovered by RaVen on a different variant of the same peptide sequence. While this +10 Da offset could potentially be interpreted as a polymorphism of S to P, combining information from the proteogenomics identification and RaVen's blind search results suggests a different, less surprising conclusion. In RaVen's spectral networks, the -16 variant forms a pair with the +10 variant (Figure 2.22). This spectral network evidence combined with the widespread detection of Acetaldehyde (+26 Da) on peptide N-termini (50 detected variants in the same dataset) yields the much less surprising interpretation of the exact same +10 Da offset as $^{+26}$SALFAQINQGESITHALK which combines the N-terminal offset of +26 Da with the -16 Da from the S to A sequence change. Complementary to this

example where proteogenomics helps interpret blind search results, the latter can also help eliminate false positives from proteogenomics searches – typical examples include reinterpreting N-term extensions by Glycine as the much more common N-term Carbamidomethylation or realizing that certain polymorphisms (e.g., Alanine to Serine) can often be explained by common modifications (e.g., oxidation) on the same or on nearby amino acids (e.g., oxidation on a Methionine adjacent to a putatively mutated Alanine). As such, it would be highly recommended to conduct both types of searches for samples where proteogenomics is expected to be needed for the detection of translation products from mutated genomes.

Comparing cases and controls is a traditional approach to finding differential expression of biological events that may be related to the phenotypes separating patients (or other samples) into groups. In the same spirit, the CPTAC data acquisition for colorectal cancer tumors was also complemented by acquisition of proteomics mass spectrometry data from healthy patients, even though those were all acquired at a later stage and possibly on different mass spectrometry instruments, thereby potentially compromising the power of the dataset to determine the association between the groups and changes in expression of proteins or peptides (due to potentially large batch effects). Nevertheless, we used spectral counts to conduct a preliminary analysis of the differential expression of modified peptides between the tumor and control groups using 12 cases and 12 controls of age and gender-matched colorectal samples. Since spectral counts are typically low for modified peptides (i.e., under 10 spectra per modified peptide variant), we opted to consider ratios of change in expression of at least 2-fold. As such, the minimum number of spectra to obtain a consistent 2-fold variation between groups would be to have spectral counts of 1 for each observation of a peptide in each patient in one group (12 counts per group) and spectral counts of 2 for each observation of the same peptide in each

patient in the other group (24 counts per group), altogether requiring 36 spectra across 24 patients. To allow some margin for missing data, which is typically the case for low spectral count, we considered only 16,534 variants with at least 30 total spectra observed across all groups. Of these variants, there are over 1,500 unmodified peptides with a fold change of over 4, which are most likely due to changes in protein expression as previously reported by [46]. However, our primary interest is to consider the changes in the level of post-translational modifications rather than changes in overall protein expression, hence we focused our analysis on the changes in fractions of peptide observations that are observed with a modification versus all other states of the exact same peptide sequence (either unmodified or with other modifications). These changes in fractions of modified variants were thus calculated using the following ratio:

V_{case} = Number of spectra of the variant of interest in the case cohort
$V_{control}$ = Number of spectra of the variant of interest in the control cohort
T_{case} = Total number of spectra of all variants of the peptide in the case cohort
$T_{control}$ = Total number of spectra of all variants of the peptide in the control cohort

$$FoldChange = \log_2 \frac{\left(\frac{\mathrm{Vcase}}{\mathrm{Tcase}}\right)}{\left(\frac{\mathrm{Vcontrol}}{\mathrm{Tcontrol}}\right)}$$

We then perform a t-test using all 12 case and control cohorts, where each observation is the per-patient ratio:

$$PerPatientRatio = \log_2 \left(\frac{\mathrm{Vcase}}{\mathrm{Tcase}}\right)$$

Of the 2,213 variants with a modification greater than or equal to 4 daltons (we chose 4 daltons as a minimum modification mass to minimize confusion with C^{13} parent mass errors), we observe 73 variants with a fold change of at least 4, of which 60 pass a t-test threshold of p-value

<0.01. Among these 60 variants we discover the peptide PVSSAASVYAGAGGSGSR on K1C18_HUMAN (P05783) with a variant containing phosphorylation on the serine at position 34 - a phosphorylation that has been shown to regulate interaction with YWHAE [67] and proposed as a marker of progression of human liver disease [68]. But while our data does not detect a significant variation of phosphorylation at serine 34 (Table 2.14), we do detect over 4-fold change in another variant of this peptide with oxidation of tyrosine at position 36 (a new modification site not listed in UniProt or PhosphoSitePlus), thus suggesting that oxidation could also potentially play a role in P05783 interactions. We further note that this oxidized tyrosine variant was detected with more spectral counts than any other oxidized tyrosine in the sample; out of 18,593 peptides containing the amino acid tyrosine, only 73 (less than 0.5%) were detected to be oxidized, thus strongly suggesting that this is not an artefactual sample handling modification.

2.5. Discussion

While several proteomics studies of CPTAC data have been conducted since the inception of the program in 2006, most of these have used focused on quantifying protein abundance and some on the analysis of translated genetic variants, typically using restricted search algorithms such as MS-GF+ and Myrimatch. The results presented here expand on this approach by using the RaVen blind database search method to detect far greater diversity than previously reported for post-translational modifications, sites and modified peptide variants. RaVen's blind search technique reveals both previously undetected modifications, as well as new polymorphisms, thereby confirming the premise of this special issue that there is a substantial

amount of proteomics “dark matter” that is almost always undetected in proteomics experiments. We thus expect that blind search re-analyses of public datasets such as demonstrated here will reveal far more proteomics diversity in health and disease than have been reported to date, thereby providing the possibility of detection of novel linkages between post-translational modifications and their impact on disease phenotypes.

2.6. Conflict of Interest

NB was a co-founder, had an equity interest and received income from Digital Proteomics, LLC through 2017. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. Digital Proteomics was not involved in the research presented here.

2.7. Author Contributions

Conceptualization, L.B., N.B.; Methodology, L.B., N.B.; Software, L.B.; Validation, L.B., N.B.; Formal Analysis L.B., N.B.; Writing, L.B., N.B.; Supervision, N.B.;

2.8. Funding

This work was supported by the US National Institutes of Health grant 2 P41 GM103484-06A1 from the National Institute of General Medical Sciences; NB is an Alfred P. Sloan Research Fellow.

2.9. Acknowledgements

Chapter 2 in full has been submitted for publication to *Frontiers in Genetics* under the title “Discovery of post-translational modifications and proteomics diversity in colorectal cancer.” Bernstein, Laurence E., and Bandeira, N. 2018. The dissertation author is the primary author of this paper

Table 2.11: The top 20 most common modifications are nearly all artefacts and sample handling modifications.

Top 20 Common Modifications				
Mass	**Site**	**Classification**	**Description**	**#Peptides**
16	M	Artefact	Oxidation	1656
1	N	Multiple	Deamidation	1563
-17	Q	Artefact	Pyro-glu	650
42	N-term	Post-translational	Acetylation	303
26	N-term	Other	Acetaldehyde	150
57	N-term	Artefact	Carbamidomethyl	142
-48	M	Chemical Derivative	Homoserine lactone	99
43	N-term	Multiple	Carbamylation	98
43	M	Artefact	Carbamylation	98
28	N-term	Artefact	Formylation	86
16	P	Post-translational	Oxidation	56
1	Q	Multiple	Deamidation	50
209	C	Artefact	Carbamidomethyl DTT	34
40	C	Artefact	Pyro-carbamidomethyl	30
-18	T	Post-translational	Dehydration	30
32	W	Chemical Derivative	Dioxidation	26
22	E	Artefact	Sodium Adduct	23
-18	E	Artefact	Pyro-glu	22
22	D	Artefact	Sodium Adduct	21
16	Y	Post-translational	Oxidation	18

Table 2.12: The top 20 rare modifications contain many substitutions and possibly biologically relevant modifications.

Top 20 Rare Modifications				
Mass	**Site**	**Classification**	**Description**	**#Peptides**
16	W	Artefact	Oxidation	10
14	N-term	Chemical derivative	Methyl	10
50	L	AA substitution	Leu/Ile->Tyr substitution	10
-34	C	Chemical derivative	Dehydroalanine	10
12	N-term	Chemical derivative	Thiazolidine	9
100	N-term	Post-translational	Succinyl	9
14	K	Post-translational	Methyl	9
57	K	Artefact	Carbamidomethyl	9
-30	M	AA substitution	M->T/1	8
50	I	AA substitution	I->Y/2	8
16	F	Artefact	Oxidation	7
114	N-term	Artefact	Dicarbamidomethyl	7
-28	V	AA substitution	V->A/1	6
4	W	Chemical derivative	Trp->Kynurenin	6
50	S	AA substitution	S->H/2	6
14	H	Post-translational	Methyl	6
30	G	AA substitution	G->S/1	6
57	E	AA substitution	E->W/2	6
-14	A	AA substitution	A->G/1	6
127	N-term	Chemical derivative	N-Succinimidyl-2-morpholine acetate (SMA)	5

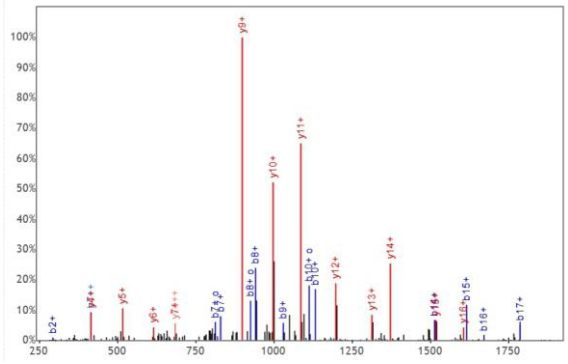
Table 2.13: Top 10 proteins with the highest variant density.
"Modified coverage" is the total number of amino acids in modified peptide variants with sequences mapping to the corresponding protein; "Modified density" is modified coverage divided by protein length.

Proteins with highest AA variant coverage density			
Protein	Length	Modified coverage	Modified density
sp|P68871|HBB_HUMAN	147	5007	34.06122
sp|P60709|ACTB_HUMAN	375	6482	17.28533
sp|P01834|IGKC_HUMAN	106	1647	15.53774
sp|P02768|ALBU_HUMAN	609	7484	12.289
sp|P62805|H4_HUMAN	103	1112	10.79612
sp|P07148|FABPL_HUMAN	127	1347	10.6063
sp|P04406|G3P_HUMAN	335	3185	9.507463
sp|O60814|H2B1K_HUMAN	126	1124	8.920635
sp|Q8IUE6|H2A2B_HUMAN	130	986	7.584615
sp|P06702|S10A9_HUMAN	114	831	7.289474

Table 2.14: Variants of the PVSSAASVYAGAGGSGR peptide on protein K1C18.
The unmodified, and functional phosphorylated forms show no significant fold change between cases and controls; however, the novel oxidized and doubly-oxidized forms near the functional phosphor-serine residue were detected with a very large 4-fold change.

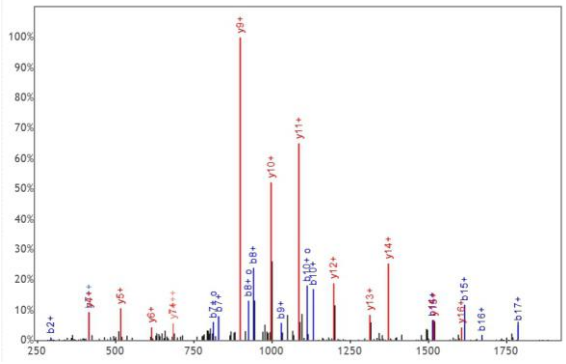
Variant		Controls	Cases	Fold Change
PVSSAASVYAGAGGSGSR	Unmodified	568	491	0.9
PVSSAA(S,80)VYAGAGGSGSR	Phosphorylation	26	18	0.7
PVSSAASV(Y,16)AGAGGSGSR	Oxidation	20	90	4.5
PVSSAASV(Y,32)AGAGGSGSR	Di-oxidation	7	40	5.7
[28]PVSSAASVYAGAGGSGSR	Formylation	20	12	1.7
[57]PVSSAASVYAGAGGSGSR	Carbamidomethyl	33	41	1.2
[12]PVSSAASVYAGAGGSGSR	Thiazolidine	35	36	1.0

(a)



FFESFGFLSTPDAV(M,-30)GNPK

(b)



YQLQSQENFEAF(M,-30)K

Figure 2.16: Two spectra showing detection of the M,-30 modification (M → T substitution).
Both spectra show strong peak correlation with good localization of the modification by matches to the surrounding peaks. (a) Identification of the modification on the protein sp|P68871|HBB_HUMAN (b) Detection of the modification on the protein sp|P07148|FABPL_HUMAN.

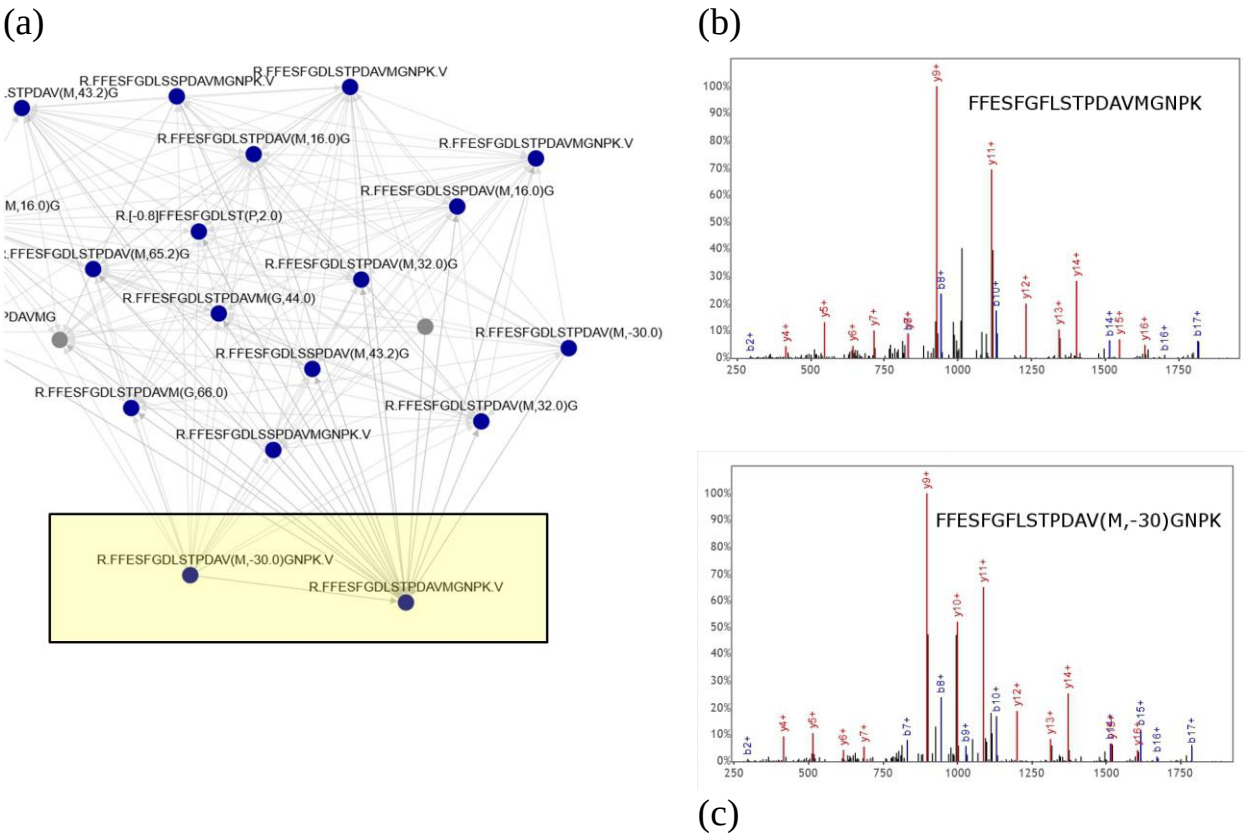


Figure 2.17: Spectral network for the FFESFGFLSPDAVMGNPK peptide on protein sp|P68871|HBB_HUMAN. (a) Network showing the pair of spectra with the unmodified and modified versions of the peptide. (b) Spectrum of the unmodified version of the peptide.
(c) Spectrum of the M,-30 modified (M → T substitution) version showing very high correlation with the peaks of the unmodified version in (b).

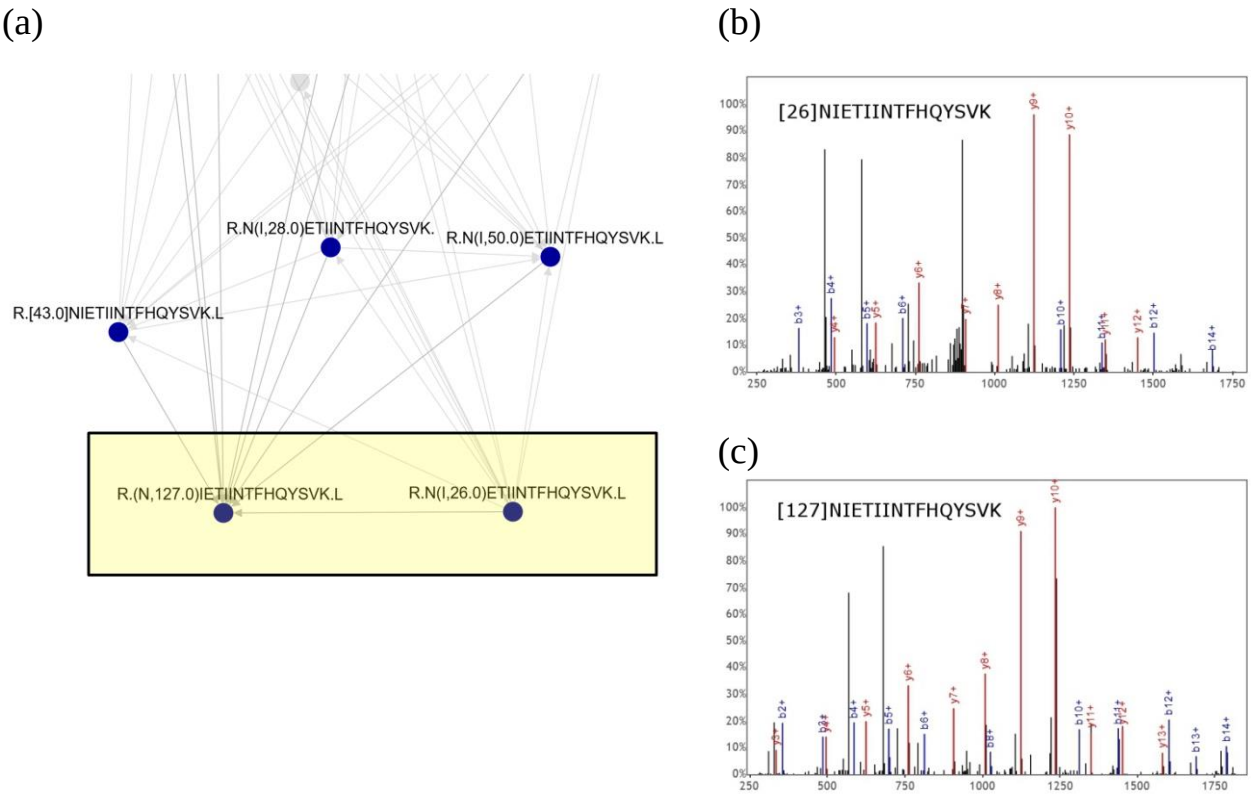


Figure 2.18: A portion of the spectral network for the NIETIINTFHQTSVK peptide on protein sp|P06702|S10A9_HUMAN.
(a) Network showing the pair of spectra with the n-terminal +127 (SMA) and n-terminal +26 (acetaldehyde) versions of the peptide. (b) Spectrum of the +26 version of the peptide. (c) Spectrum of the +127 version of the peptide showing very high correlation with the peaks of the +26 version in (b).

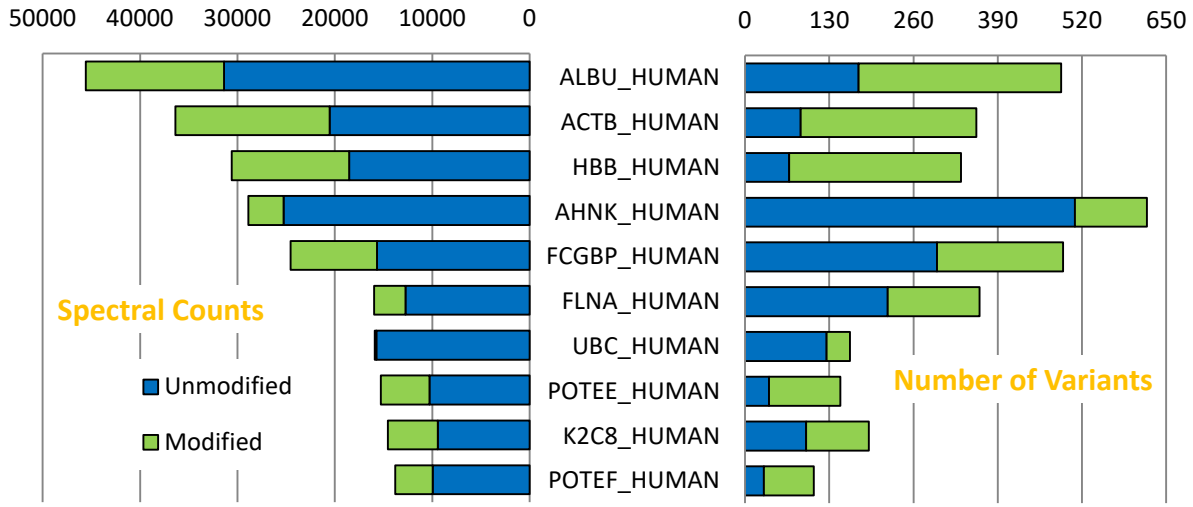


Figure 2.19: Top 10 proteins with the highest abundance (by spectral counts).
From the comparison it can be seen that there is not a strong connection between spectral abundance of proteins in the sample and the number of peptide variants detected per protein.

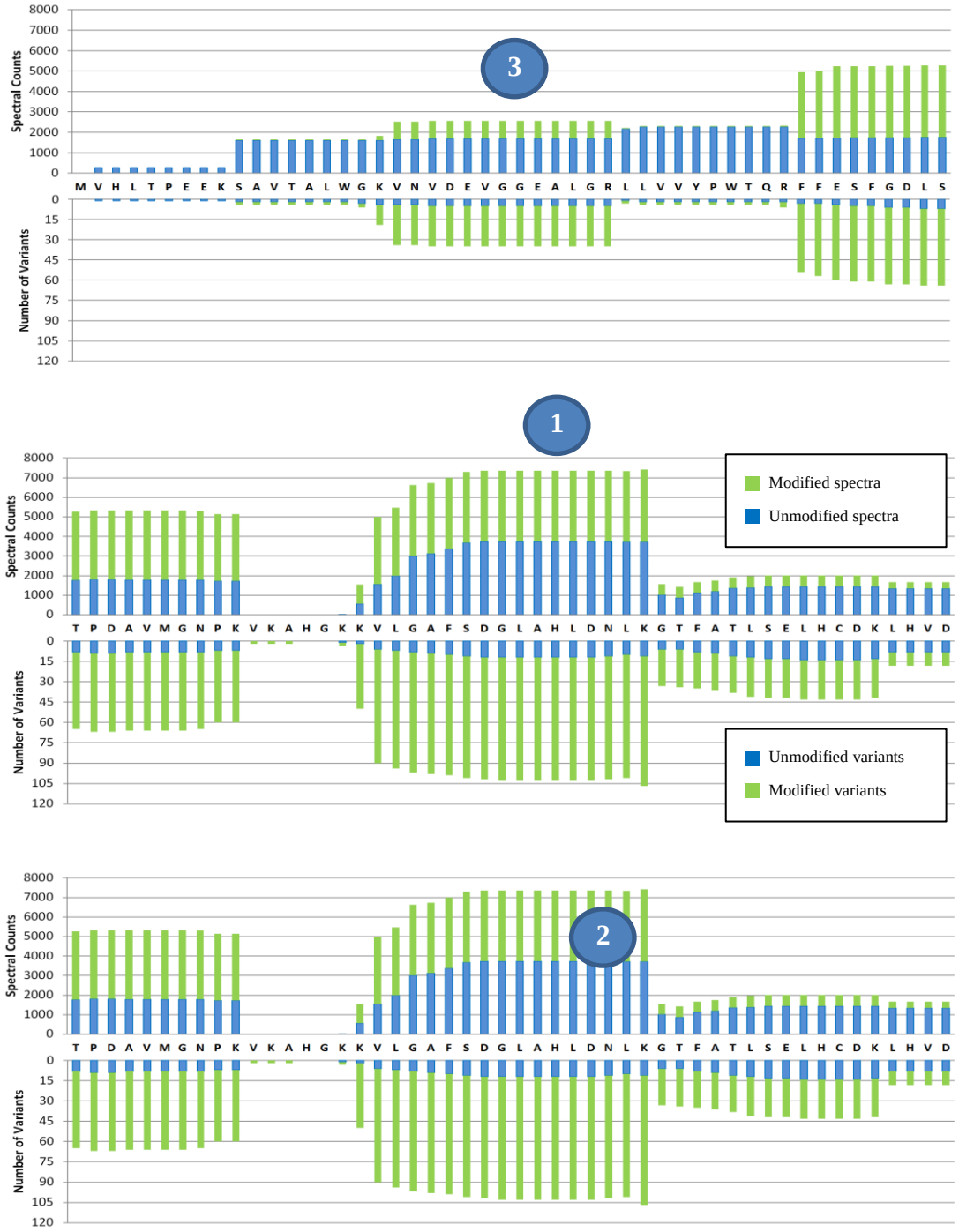


Figure 2.20: Histogram of variant coverage for the protein HBB (P68871).
Unmodified spectra and variants are shown in blue, while modified versions are shown in green. We can see two types of regions that are well covered. (1) High coverage regions with over 4500 spectra per amino acid. These regions also have over 50 variants per amino acid that are nearly all modified. (2) Medium coverage regions with under 3000 spectra, which are either well conserved with less than 5 modified variants per amino acid, or (3) very diverse sections with dozens of modified variants per amino acid. While higher coverage (spectral counts) does facilitate the detection of more peptide variants, the observation of areas of both low and high modification diversity within these regions clearly illustrates that peptide abundance is not the only determinant of the observed protein diversity.

Figure 2.21: 71 manually verified discovered variants of the largest Protein Variant Region (PVR) identified in the region spanning amino acids 66 to 85 of the HBB protein (P68871).
Modifications in blue are modifications which reflect substitutions while modifications in green are substitutions that have previously been discovered and listed in Uniprot.

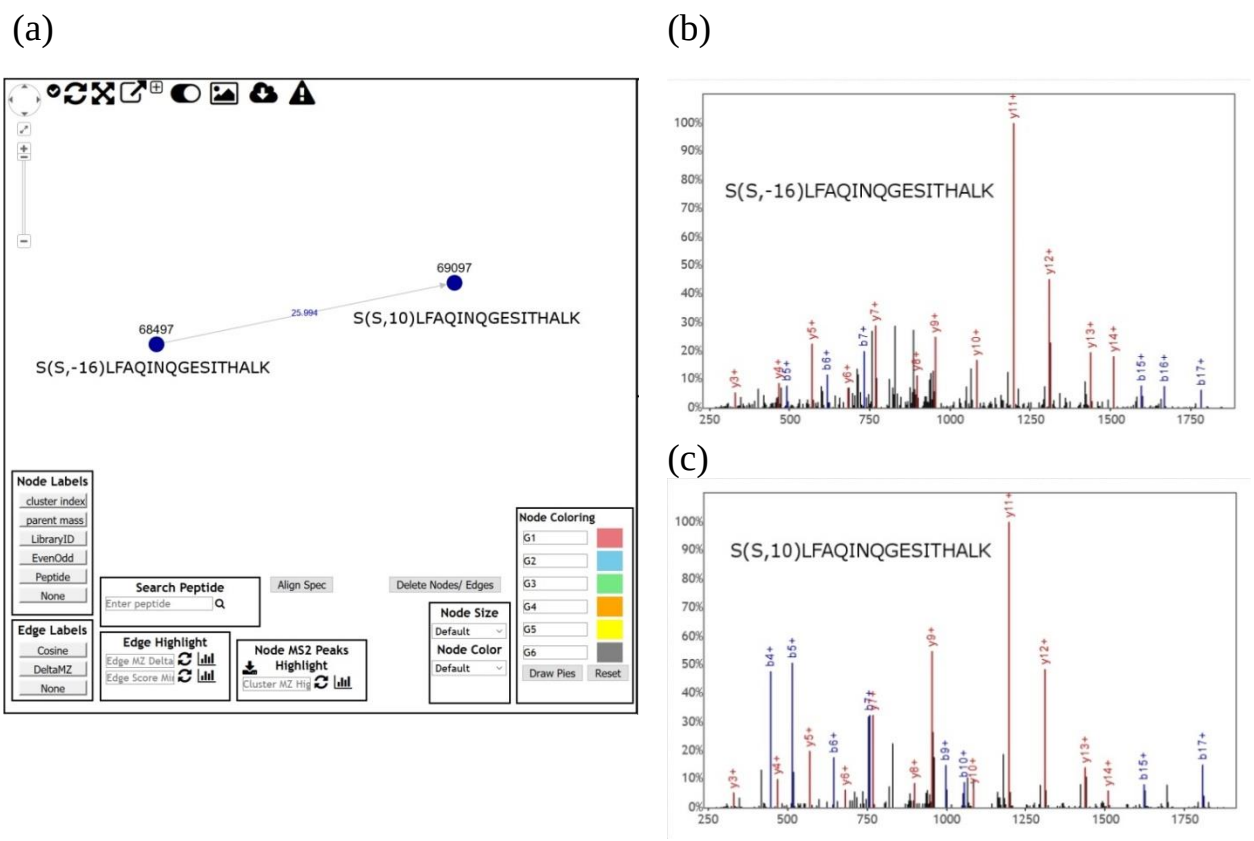


Figure 2.22: Spectral network for SSLFAQINQGESITHALK peptide on protein K1C18.
(a) Network showing the pair of spectra with a 26 dalton mass difference. (b) Spectra with S to A polymorphism. (c) Spectra with +10 mass modification corresponding to S to A polymorphism and acetyladehyde +26.

CHAPTER 3 - Enabling massive blind database search using multiple enzyme proteomics

3.1. Introduction

Traditional methods of database search in mass spectrometry, such as Sequest [1], Mascot [2] and MS-GF+ [3] face significant challenges with increases in the size of the database being searched. Additionally, these methods allow only a pre-specified list of known post-translational modifications (PTMs) to be discovered in the data. Newer, so-called "blind" or "open" search methods allow for the discovery of unexpected modifications but often suffer from either reduced overall identification ability or large numbers of false-positive identifications which require manual post-processing in order to determine the true novel discoveries. Both these blind and non-blind methods suffer severe degradation in terms of speed and identification rates as database search size grows, while restricted methods also degrade with increased size of their allowed modification lists.

In 1994, Mann et al. [15] first proposed the use of small de novo tags in order to pre-filter the database and reduce the overall search space, and many search methods have incorporated this strategy [7][8] [9][16]. While it is possible to increase the size of the tags to handle large databases, this poses additional challenges in that standard mass spectra contain large amounts of noise and missing spectral peaks which greatly reduce the likelihood of finding longer length tags and thus results in reduced sensitivity and lowered numbers of identifications. Furthermore as database sizes increase the number of matches to a short (length 3) tag becomes very large,

particularly because protein databases are not random, but contain many homologs resulting in far more matches than would be encountered randomly for some tags.

Most existing algorithms use raw MS/MS spectra for identification; however, the RaVen workflow includes multiple steps that process the spectra making the prohibition of longer tags no longer valid. First, RaVen uses MS-Cluster to form cluster-consensus spectra obtained from repeated acquisition of the same peptide. Second, RaVen invokes MetaSPS [30] to construct spectral contigs converts cluster-consensus spectra to Prefix Residue Mass (PRM) spectra [32]. These cluster-consensus PRM spectra have far fewer noise peaks resulting in a much-increased signal-to-noise ratio [30]. Since RaVen uses these spectra for tagging, the previously held belief that long tags cannot work no longer applies, instead we have already shown [69] that longer tags with gaps produce superior identification results and faster identification times. The MetaSPS algorithm; however, has also succeeded in generating even longer contigs when the samples are processed using multiple enzymes [33]. This is due to the increased overlap in peptides when multiple enzymes are used due to their differing cleavage locations (Figure 2). RaVen is able to capitalize on this and use even longer tags than in single enzyme experiments resulting in greater gains in processing speed, while at the same time not just maintaining the performance that has previously shown to exceed other algorithms, but actually increasing performance as tag length grows due to the reduction in the virtual database search space.

3.2. Methods

To perform the analysis we used the RaVen software exactly as described in Bernstein, L., Wertz, J., Na, S., and Bandeira, N., 2018 [69] with the sole exception being the change in tagging parameters. A brief description of the salient tagging features of RaVen is given here.

Rather than use spectral tagging, RaVen uses contig-derived sequence tags to reduce the database locations that are matched to contig-consensus spectra. Standard de novo sequence tags are constructed by connecting spectrum peaks whose masses differ by one amino acid mass (Figure 3.1a); however, RaVen extends this method to allow peaks whose masses differ by two amino acid masses (e.g., X and Z) and that may also be connected by mass gaps corresponding to the summed amino acid masses (e.g., mass(X)+mass(Z)) and all possible permutations thereof (e.g., "XZ" or "ZX"). As Guhtals et al showed [30], the accuracy of de novo sequencing with these contig-consensus spectra are much higher than in non-assembled spectra since most noise peaks are eliminated or have reduced intensity while signal peaks have larger scores due to accumulation of signal from multiple assembled spectra. Since multiple tags may be found in each spectrum (Figure 3.1a), it is common [9][16] to rank each tag by summing the scores of all spectrum peaks which match the tag and then to match the top N scoring tags per spectrum to the database. Obviously, filtration efficiency increases with longer tags (i.e., longer sequences have less matches to the database); however, missing spectrum peaks constrain the length of the longest correct tags that can be extracted from any given contig-consensus spectra. We therefore allow gaps, or missing peaks, when matching the de novo tags as shown in Figure 3.1b.

Previously we have found that using tags of length 5 with up to two gaps of two amino acids, and taking the top 50 tags, resulted in the best filtration vs. identification performance [69]. However, given the much longer contig-consensus spectra generated as the result of

assembly with peptides from multiple enzyme digestion we used length 5 as a starting point and tested with tags from length 6 to length 9. Although we also tested the use of various numbers of allowed gaps, we discovered that 2 gaps remained the best choice. We also expanded our list from 50 to 200 top tags for filtration, because with tags of such length and high filtration rate, we did not want to filter any matching tags and during testing it was discovered that using either top 100 or top 200 tag lists produced nearly identical filtration results.

3.3. Results

To test our long tagging approach we used the six protein test sample (Table 3.1) used by Guthals, A., Clauser, K. R., and Bandeira, N. [30] and included samples from seven different enzymatic digestions (see Table 3.2). We assessed the performance of our RaVen approach against state-of-the-art approaches for both restricted and unrestricted database searches. For the restricted algorithm we chose MS-GF+ [3] because it has been shown to exceed the performance of many other existing algorithms over a variety of sample types and collection methods. For the blind approach, we chose MODa [9] which is one of the leading multi-blind search algorithms.

To examine the effect on increased database size, we ran tests using four databases of increasing size from 4 megabytes to 287 megabytes (see Table 3.3). We use the standard False Discovery Rate (FDR) method [44] using the target database and a shuffled decoy database. The smallest database we tested against consisted of the six known proteins of the sample plus the entire yeast proteome, plus a list of common contaminants resulting in a database of 6,645 target proteins (~ 4 million amino acids). Because larger databases are required for proper sampling of the space of random peptides [70], we did not test with any smaller database (for instance only the 6 sample peptides). For our larger databases we added the entire Drosophila proteome to the

first database to create a database with a total of 34,828 proteins (~21 million amino acids). We then added the entire human proteome to the six sample proteins, contaminants, plus Yeast and Drosophila database resulting in a database with 120,725 proteins (~64.5 million amino acids). For our final database we searched against the entire Uniprot Reference Protein Set consisting of 585,133 proteins (~264.7 million amino acids).

In previous experiments we have shown that when using tags with contig-consensus sequences the best performance is obtained when using tags of length five with allowed gaps of up to two peaks [69]. However, these searches were conducted with contig-consensus spectra created from a single enzyme digestion. While perfect digestion with a single enzyme provides no overlap of the resulting peptides (Figure 3.2a) only by relying on missed cleavages (Figure 3.2b) can overlapping peptides be achieved. However, different digestion enzymes cleave proteins in different locations (Figure 3.2c), so by combining the digestion of multiple peptides, excellent overlapping can be achieved (Figure 3.2d) which leads to increased contig length as shown by Guthals et al [30]. Using these long contig-consensus spectra we tested even longer length tags (from 6 to 9) than we previously thought possible. As can be seen in Figure 3.3, at 1% spectrum level FDR we see increased performance on all database searches at as the tag lengths increase from 5 to 6 and then 7, whereas after that point the performance drops precipitously (at tag lengths 8 and 9).

While peptide-spectrum (PSM) identifications at a fixed FDR are a common measure of performance for database search algorithms, we instead use a measurement of discovered peptide variants where a variant corresponds to a unique peptide sequence with a unique sum of modification masses [69]. Our reason for using this measure instead, is that the biological relevance of identifications in the context of blind searches is in the discovery of uniquely

modified peptide forms. We therefore impose a 1% peptide variant level FDR which is stricter than the traditional PSM level FDR. Using this measure we compared the results of our search with results from MODa and MS-GF+. As can be seen in Figure 3.5, RaVen outperforms MODa blind search by nearly 30% and MS-GF+ by 70% on the largest database set.

One curious effect that can be seen in the figure is that the performance of MODa and MS-GF+ appears to rise with the largest database. This is a highly suspicious result, and merited further investigation, upon which we discovered the reason for this increase is that the full Uniprot Reference database contains a great number of homologues to the proteins and contaminants in the sample. Although we eliminate double identical counting peptides appearing on multiple proteins, any search algorithm that includes possible modifications and particularly blind searches which are able to match any modification mass may identify homologues in the database by detecting a single amino acid change as a modification mass. For example: a peptide containing a phenylalanine (mass 147 daltons) along with an oxidation (mass +16 daltons) might be matched to a peptide from a homologous protein with a tyrosine (163 daltons) in the same location. To get a more accurate picture of the true identifications we removed all target proteins from the sample, common contaminants and all their homologues in all databases. We then counted the number of remaining (false positive) identifications found by each algorithm. As can be seen in Figure 3.7, at smaller database sizes, both RaVen and MS-GF+ maintain approximately a 1% FDR with the small databases, while MODa has ~2% FDR rate. When the database becomes very large, the false positive rate for all algorithms rises significantly. RaVen yields 3.94% false identifications with the full Uniprot reference, while MODa and MS-GF+ have over twice the rate of RaVen (8.22% and 8.74% respectively). There remains the possibility; however, that there were unexpected contaminants within the sample, so we further

examined identifications without modifications that were agreed upon by all three algorithms both MODa and MS-GF+. Inspecting this list of identifications reveals that these additional discoveries are also possible homologues of proteins in the sample, which were not taken into account in the first pass of eliminating homologues. Using this additional exclusion list, all the algorithms adhere closely to the estimated 1% FDR expected for small database sizes; however, once again when tested against the Uniprot reference, all the algorithms show an increase in the number of false positives. Again, RaVen shows the smallest adverse effect, remaining under 2% false identifications on the full Uniprot reference, while MODa increased to nearly 4% and MS-GF+ to nearly 6%.

The speed gains from using longer tags are, of course, directly related to the length of the tag used. In our earlier work [69] we noted that tags of length 5 with 2 allowed gaps yielded a >10,000-fold decrease in the search space for the contig-consensus spectrum. A similar analysis for the largest database yields the results seen in Table 3.4. From this table we can see that while our 6 protein sample only yielded 620 total contigs, the 5-2 tags yielded ~750 tags per contig-consensus spectrum, a figure similar to that seen in our previous research. These contig-consensus spectra matches are elongated through a tag-extension procedure and the tag match scores are increased by the score of any spectral peaks that match the flanking sequences in the database. Any extended tag matches that score below 80% of the best match are then thrown out and the remaining locations are used to align the assembled spectra. As we can see in Table 3.4, there is a linear relationship between tag length and resulting database matches and therefore overall filtration efficiency. Using tags of length 7 we filter to approximately 50 database locations for each assembled spectrum, resulting in over a $5x10e^{-6}$ reduction in the search space and search times of approximately 20 minutes for the largest database. We can compare these

figures to what would be required using the standard tags of length 3 that would match, on average, ~33,000 locations in the database for every tag on every spectrum, or 6.6 x10e^{-6} tags per spectrum if using the top 200 tags, resulting in only a 40-fold decrease in search space.

3.4. Discussion

By taking advantage of spectral networks and contig assembly of spectra, our RaVen method can leverage the highly increased signal-to-noise ratio of assembled spectra and long tags to achieve very high database filtration rates even in standard MS/MS experiments. If multiple enzymes are used in the digestion of the sample, we have shown that tag lengths can be increased even further due to the increased overlap of the resultant peptides and longer contig-consensus spectra created by the assembly process. Guthals et al. also showed that peptide overlap was improved through the use of MS/MS spectra triplets from experiments run with multiple fragmentation methods: i.e. - electron-transfer dissociation (ETD), collision-induced dissociation (CID) and higher-energy collision-induced dissociation (HCD) [30] and this would be true for the RaVen identification method as well. It should be noted, that our RaVen method relies on the construction of a spectral network which is completed using the alignment of pairs of spectra. This is an O(N^2) matching process, which for very large data sets can take a substantial amount of time; however, our use of AlignGF for matching spectral pairs partially mitigates this issue since AlignGF also uses a tagging-based approach to reduce the total number of spectral alignments required [12]. Furthermore, we are less concerned about overall identification time, but rather increased detection of rare peptide variants even when searching against very large databases.

3.5. Acknowledgements

The material in this chapter was presented in large part at the 2016 Annual Meeting for the American Society of Mass Spectrometry, but has been substantially enhanced and refined.

Table 3.1: Six proteins in the test sample.

sp|P00433|PER1A_ARMRU	Peroxidase
sp|P00974|BPT1_BOVIN	Pancreatic trypsin inhibitor
sp|P07288|KLK3_HUMAN	Prostate Specific Antigen
sp|P0A6F5|CH60_ECOLI	Chaperonin E. Coli
sp|P41160|LEP_MOUSE	Leptin
sp|P68082|MYG_HORSE	Myoglobin

Table 3.2: Seven enzymes used in the sample preparation.

Trypsin
Chymotrypsin
Pepsin
Glu-C
Lys-C
Arg-C
AspN

Table 3.3: Four databases used for all database searches.

Database	Number of proteins	Number of amino acids (millions)
6-mix proteins with Yeast	6,645	3.96
6-mix proteins with Yeast and Drosophila	34,828	21.66
6-mix proteins with Yeast and Drosophila and Human	120,725	64.50
All Uniprot Reference Protein Set	585,133	264.70

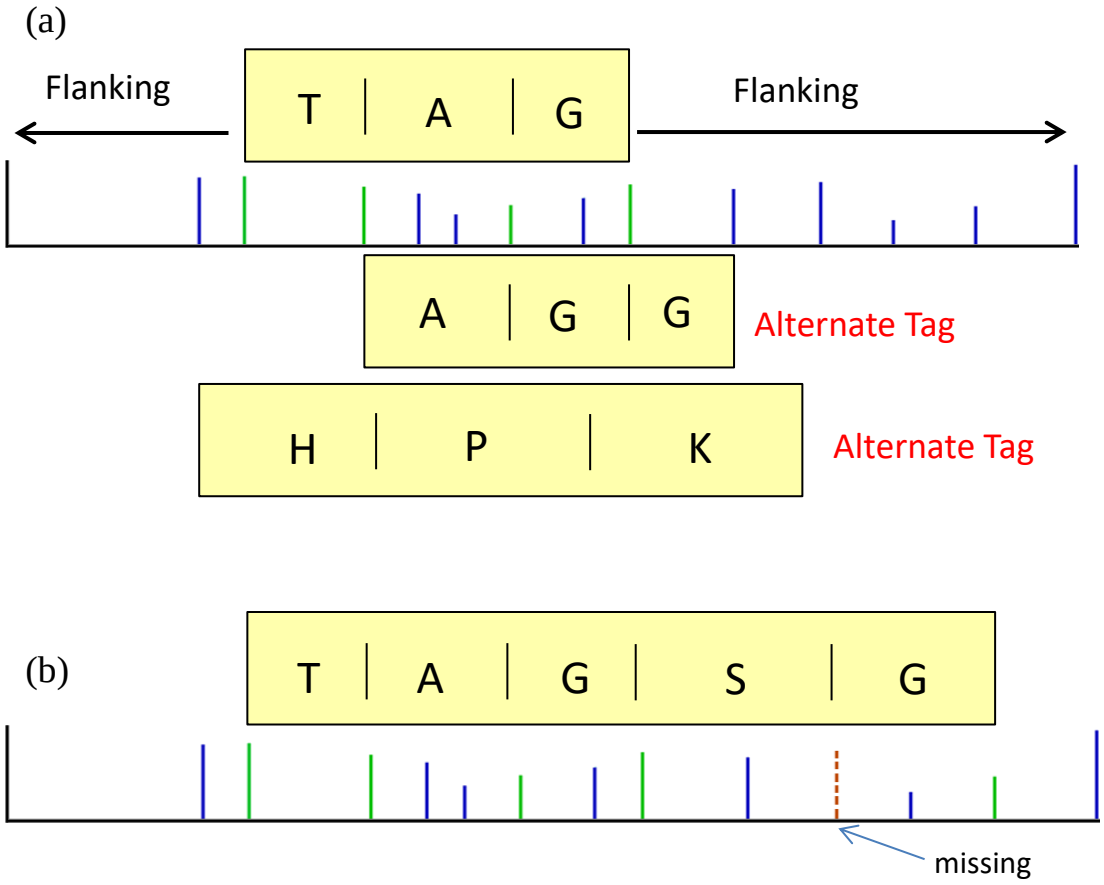


Figure 3.1: Denovo spectrum tagging. (a) Tags of standard length 3. Each spectrum may have multiple different tags of equal length that match different peaks in the spectrum. (b) Tag of length 5 with one missing peak.

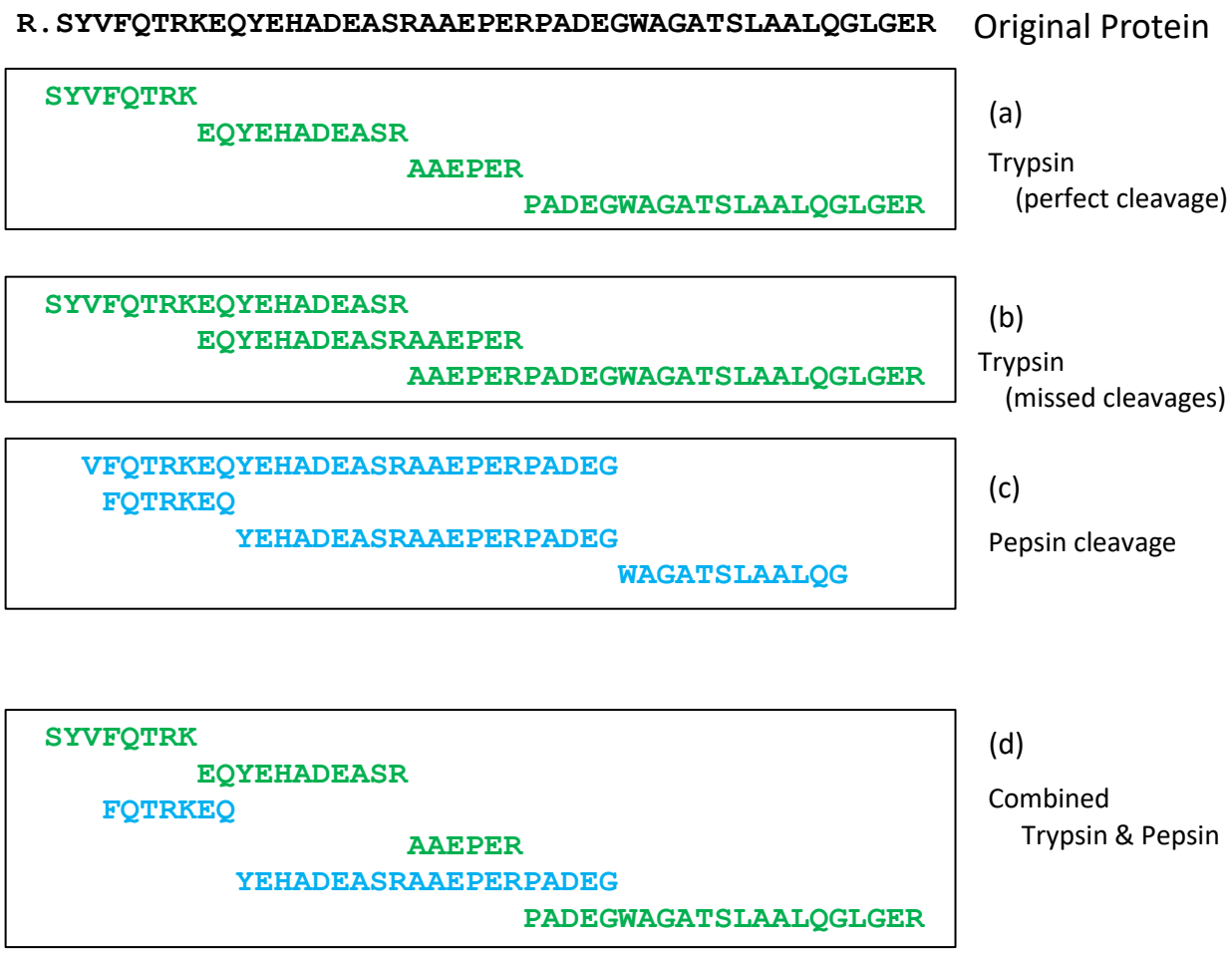


Figure 3.2: Enzymatic cleavage using multiple enzymes. (a) Perfect cleavage using trypsin results in no overlap of resultant peptides. (b) Only missed cleavages result in overlapping peptide. (c) Perfect cleavage with Pepsin results in no overlap of peptides. (d) Combined peptides from multipleenzymes results in long contigs.

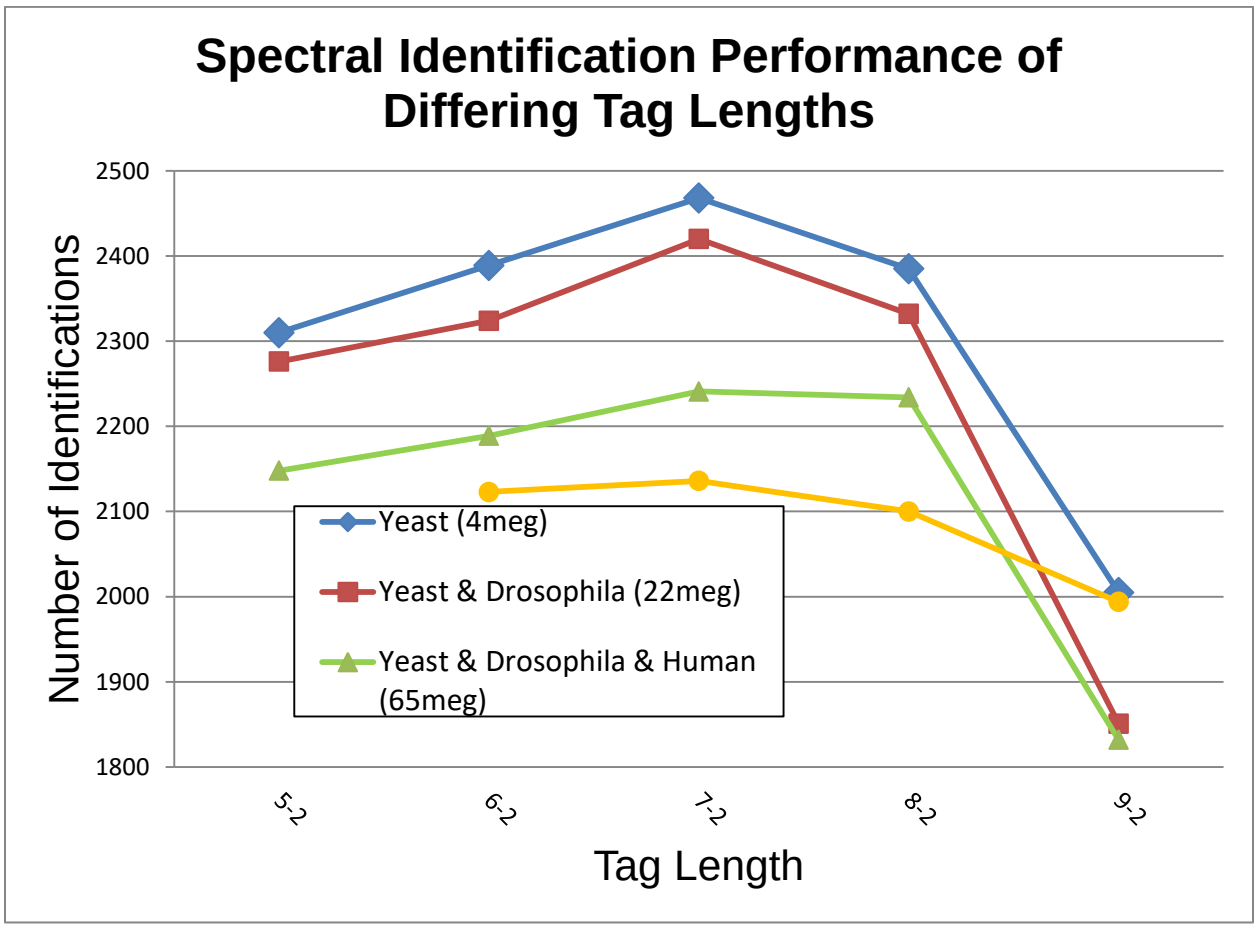


Figure 3.3: Graph of Tag Length versus spectral identification performance of RaVen Performance increases with an increase of tag length from 5 to 7 but falls off rapidly thereafter.

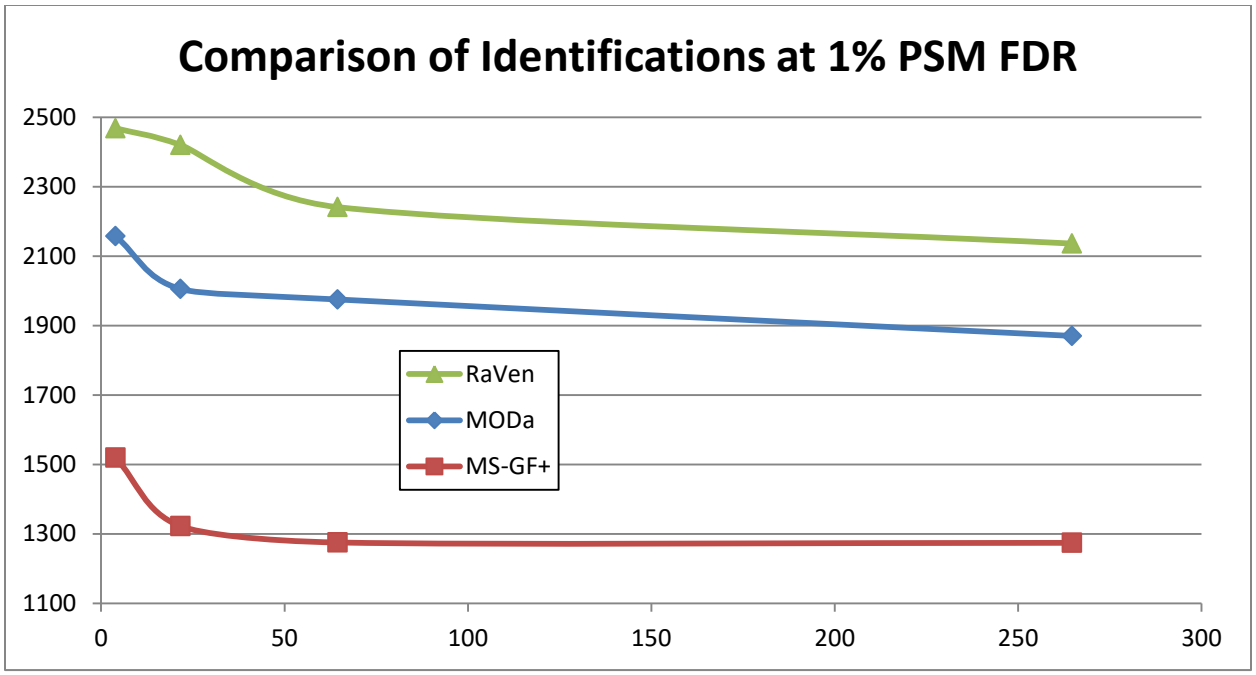


Figure 3.4: Comparison of the number of spectral identifications by all three algorithms at 1% PSM FDR shows RaVen performs significantly better than either MODa or MS-GF+.
Both Raven and MODa show a small decline in the number of identified spectra with very large database size.

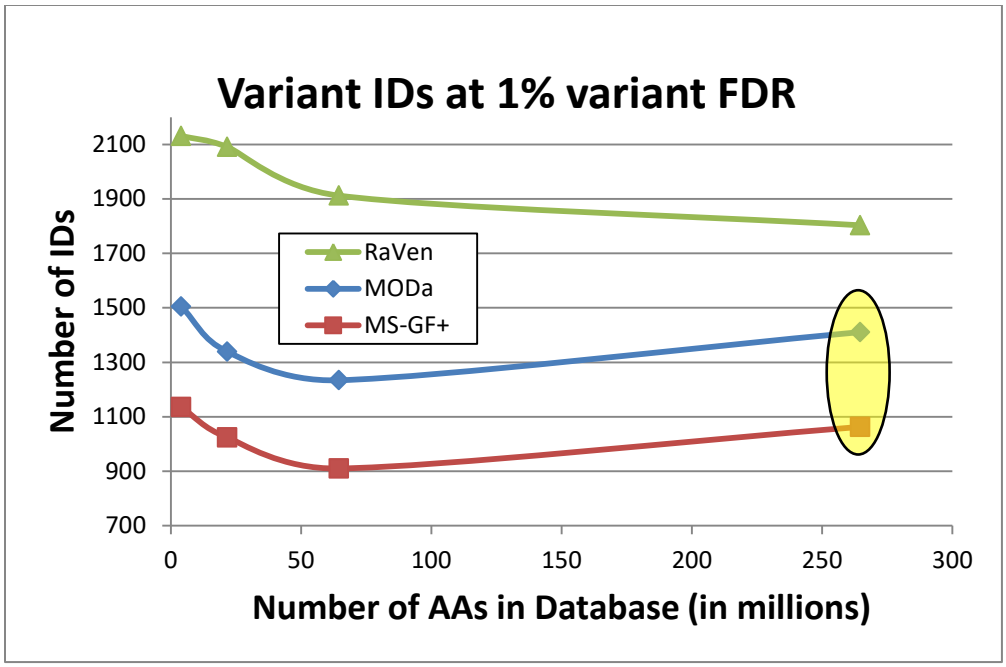


Figure 3.5: Comparison of peptide variant identifications at 1% FDR for all three search methods (RaVen, MODa and MS-GF+) shows RaVen performance far exceeds both other methods. At large database size MODa and MS-GF+ show an unusual rise in identifications, which appears to be the result of the large number of homologues in the Uniprot database (see text).

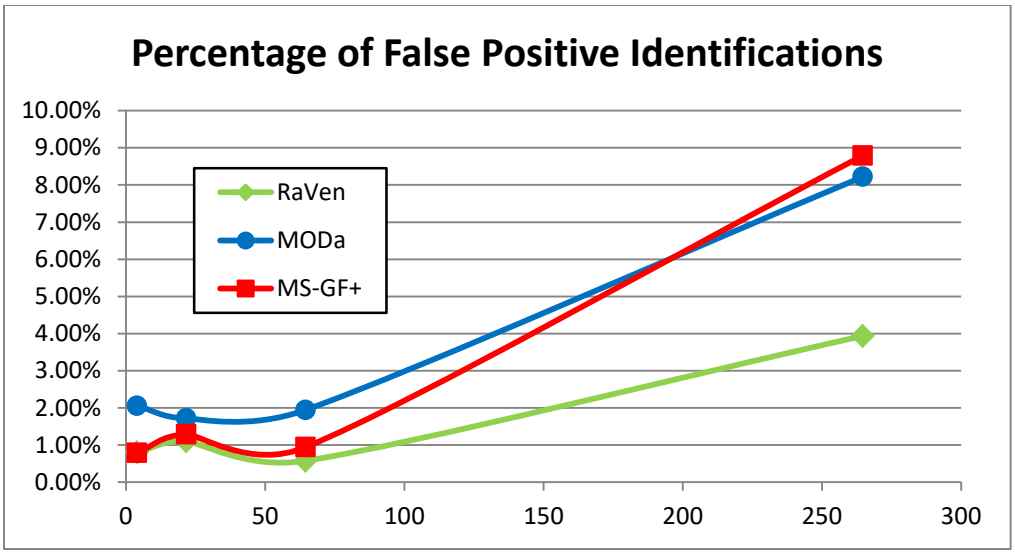


Figure 3.6: Percentage of false positive identifications by all search algorithms rises dramatically as database size increases.
Both RaVen and MS-GF+ maintain approximately a 1% FDR with the small database, while MODa has ~2% FDR rate. When the database becomes very large, MODa and MS-GF+ have over twice the rate of false positives compared to RaVen.

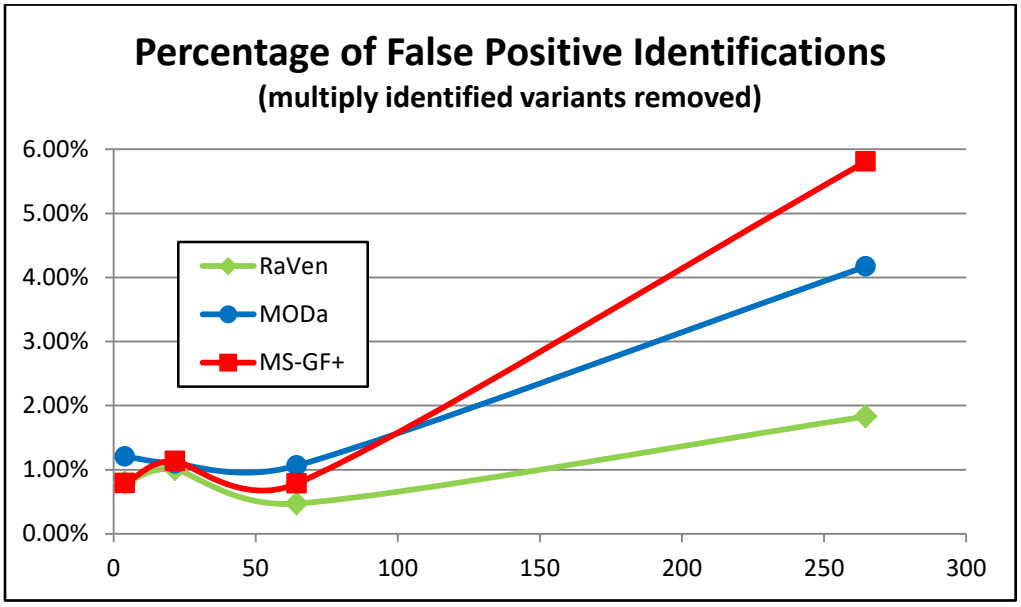


Figure 3.7: Percentage of false positive identifications by all search algorithms when variants identified by all methods are removed.
All methods maintain approximately ~1% FDR with the small databases; however, when the database becomes very large, RaVen false positives increase slightly to 1.8% which MODa and MS-GF+ have 4.2% and 5.8% false positive rates respectively.

Table 3.4: Number of tag matches and resulting filtration rates for tags of varying lengths.

Tag Length	Contig-Consensus Matches	Matches per Contig (620 contigs)	Spectral Matches	Matches per spectrum (2628 spectra)	Filtration Efficiency
5	45997	781.17	184,458	70.19	3.77E+06
6	30484	270.11	142,676	54.29	4.88E+06
7	23800	160.23	124,102	47.22	5.61E+06
8	17580	108.40	99,387	37.82	7.00E+06
9	13594	74.34	81,141	30.88	8.57E+06

REFERENCES

1. Eng, J. K., McCormack, A. L., and Yates, J. R. "**An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database.**" *Journal of the American Society for Mass Spectrometry* 5, no. 11 (1994): 976–89. doi:10.1016/1044-0305(94)80016-2, Available at http://www.ncbi.nlm.nih.gov/pubmed/24226387

2. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. "**Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data.**" *Electrophoresis* 20, no. 18 (1999): 3551–67. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2, Available at http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1522-2683(19991201)20:18%3C3551::AID-ELPS3551%3E3.0.CO;2-2/abstract

3. Kim, S. and Pevzner, P. A. "**MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics.**" *Nature communications* 5, (2014): 5277. doi:10.1038/ncomms6277, Available at http://www.nature.com/ncomms/2014/141031/ncomms6277/abs/ncomms6277.html

4. Creasy, D. M. and Cottrell, J. S. "**Unimod: Protein Modifications for Mass Spectrometry.**" *Proteomics* 4, no. 6 (2004): 1534–6. doi:10.1002/pmic.200300744, Available at http://doi.wiley.com/10.1002/pmic.200300744

5. Pevzner, P. A., Mulyukov, Z., Dancik, V., and Tang, C. L. "**Efficiency of Database Search for Identification of Mutated and Modified Proteins via Mass Spectrometry.**" *Genome research* 11, no. 2 (2001): 290–9. doi:10.1101/gr.154101, Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=544186&tool=pmcentrez&rendertype=abstract

6. Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P. A. "**Identification of Post-Translational Modifications by Blind Search of Mass Spectra.**" *Nature biotechnology* 23, no. 12 (2005): 1562–7. doi:10.1038/nbt1168, Available at http://www.nature.com/doifinder/10.1038/nbt1168

7. Searle, B. C., Dasari, S., Wilmarth, P. A., Turner, M., Reddy, A. P., David, L. L., and Nagalla, S. R. "**Identification of Protein Modifications Using MS/MS de Novo Sequencing and the OpenSea Alignment Algorithm.**" *Journal of proteome research* 4, no. 2 (2005): 546–54. doi:10.1021/pr049781j, Available at http://pubs.acs.org/doi/abs/10.1021/pr049781j

8. Dasari, S., Chambers, M. C., Slebos, R. J., Zimmerman, L. J., Ham, A.-J. L., and Tabb, D. L. "**TagRecon: High-Throughput Mutation Identification through Sequence Tagging**" *Journal of Proteome Research* 9, no. 4 (2010): 1716–1726. doi:10.1021/pr900850m, Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2859315/

9. Na, S., Bandeira, N., and Paek, E. "**Fast Multi-Blind Modification Search through Tandem Mass Spectrometry.**" *Molecular & cellular proteomics : MCP* 11, no. 4 (2012): M111.010199. doi:10.1074/mcp.M111.010199, Available at http://www.mcponline.org/content/11/4/M111.010199.short

10. Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L., and Gygi, S. P. "**A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun Proteomics as Modified Peptides**" *Nature Biotechnology* 33, no. 7 (2015): 743–749. doi:10.1038/nbt.3267, Available at http://www.ncbi.nlm.nih.gov/pubmed/26076430

11. Kong, A. T., Leprevost, F. V, Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. "**MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry-Based Proteomics.**" *Nature methods* 14, no. 5 (2017): 513–520. doi:10.1038/nmeth.4256, Available at http://www.nature.com/doifinder/10.1038/nmeth.4256

12. Na, S., Payne, S. H., and Bandeira, N. "**Multi-Species Identification of Polymorphic Peptide Variants via Propagation in Spectral Networks.**" *Molecular & cellular proteomics : MCP* 15, no. 11 (2016): 3501–3512. doi:10.1074/mcp.O116.060913, Available at http://www.ncbi.nlm.nih.gov/pubmed/27609420

13. Guthals, A., Watrous, J. D., Dorrestein, P. C., and Bandeira, N. "**The Spectral Networks Paradigm in High Throughput Mass Spectrometry.**" *Molecular bioSystems* 8, no. 10 (2012): 2535–44. doi:10.1039/c2mb25085c, Available at http://xlink.rsc.org/?DOI=c2mb25085c

14. Bandeira, N., Tang, H., Bafna, V., and Pevzner, P. "**Shotgun Protein Sequencing by Tandem Mass Spectra Assembly.**" *Analytical chemistry* 76, no. 24 (2004): 7221–33. doi:10.1021/ac0489162, Available at http://dx.doi.org/10.1021/ac0489162

15. Mann, M. and Wilm, M. "**Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags**" *Analytical Chemistry* 66, no. 24 (1994): 4390–4399. doi:10.1021/ac00096a002, Available at http://dx.doi.org/10.1021/ac00096a002

16. Frank, A., Tanner, S., Bafna, V., and Pevzner, P. "**Peptide Sequence Tags for Fast Database Search in Mass-Spectrometry.**" *Journal of proteome research* 4, no. 4 (2005): 1287–95. doi:10.1021/pr050011x, Available at http://pubs.acs.org/doi/abs/10.1021/pr050011x

17. Toyama, B. H. and Hetzer, M. W. "**Protein Homeostasis: Live Long, Won't Prosper.**" *Nature reviews. Molecular cell biology* 14, no. 1 (2013): 55–61. doi:10.1038/nrm3496, Available at http://www.ncbi.nlm.nih.gov/pubmed/23258296

18. Searle, B. C., Dasari, S., Turner, M., Reddy, A. P., Choi, D., Wilmarth, P. A., McCormack, A. L., David, L. L., and Nagalla, S. R. "**High-Throughput Identification of Proteins and Unanticipated Sequence Modifications Using a Mass-Based Alignment Algorithm for MS/MS de Novo Sequencing Results.**" *Analytical chemistry* 76, no. 8 (2004): 2220–30.

doi:10.1021/ac035258x, Available at http://pubs.acs.org/doi/abs/10.1021/ac035258x

19. Bandeira, N., Tsur, D., Frank, A., and Pevzner, P. A. "**Protein Identification by Spectral Networks Analysis.**" *Proceedings of the National Academy of Sciences of the United States of America* 104, no. 15 (2007): 6140–5. doi:10.1073/pnas.0701130104, Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1851064&tool=pmcentrez&rendertype=abstract

20. Wilmarth, P. A., Tanner, S., Dasari, S., Nagalla, S. R., Riviere, M. A., Bafna, V., Pevzner, P. A., and David, L. L. "**Age-Related Changes in Human Crystallins Determined from Comparative Analysis of Post-Translational Modifications in Young and Aged Lens: Does Deamidation Contribute to Crystallin Insolubility?**" *Journal of proteome research* 5, no. 10 (2006): 2554–66. doi:10.1021/pr050473a, Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2536618&tool=pmcentrez&rendertype=abstract

21. Tanner, S., Payne, S. H., Dasari, S., Shen, Z., Wilmarth, P. A., David, L. L., Loomis, W. F., Briggs, S. P., and Bafna, V. "**Accurate Annotation of Peptide Modifications through Unrestrictive Database Search**" *Journal of Proteome Research* 7, no. 1 (2008): 170–181. doi:10.1021/pr070444v, Available at http://pubs.acs.org/doi/abs/10.1021/pr070444v

22. Savitski, M. M., Nielsen, M. L., and Zubarev, R. A. "**ModifiComb, a New Proteomic Tool for Mapping Substoichiometric Post-Translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures.**" *Molecular & cellular proteomics : MCP* 5, no. 5 (2006): 935–48. doi:10.1074/mcp.T500034-MCP200, Available at http://www.mcponline.org/lookup/doi/10.1074/mcp.T500034-MCP200

23. Hains, P. G. and Truscott*, R. J. W. "**Post-Translational Modifications in the Nuclear Region of Young, Aged, and Cataract Human Lenses**" (2007): doi:10.1021/PR070138H, Available at http://pubs.acs.org/doi/abs/10.1021/pr070138h

24. Lampi, K. J., Wilmarth, P. A., Murray, M. R., and David, L. L. "**Lens β-Crystallins: The Role of Deamidation and Related Modifications in Aging and Cataract**" *Progress in Biophysics and Molecular Biology* 115, no. 1 (2014): 21–31. doi:10.1016/j.pbiomolbio.2014.02.004, Available at http://www.ncbi.nlm.nih.gov/pubmed/24613629

25. Takemoto, L. "**Deamidation of Asn-143 of Gamma S Crystallin from Protein Aggregates of the Human Lens.**" *Current eye research* 22, no. 2 (2001): 148–53. Available at http://www.ncbi.nlm.nih.gov/pubmed/11402392

26. Xing, K.-Y. and Lou, M. F. "**Effect of Age on the Thioltransferase (Glutaredoxin) and Thioredoxin Systems in the Human Lens**" *Investigative Opthalmology & Visual Science* 51,

no. 12 (2010): 6598. doi:10.1167/iovs.10-5672, Available at http://www.ncbi.nlm.nih.gov/pubmed/20610843

27. Bloemendal, H., Jong, W. de, Jaenicke, R., Lubsen, N. H., Slingsby, C., and Tardieu, A. "**Ageing and Vision: Structure, Stability and Function of Lens Crystallins**" *Progress in Biophysics and Molecular Biology* 86, no. 3 (2004): 407–485. doi:10.1016/j.pbiomolbio.2003.11.012, Available at http://www.ncbi.nlm.nih.gov/pubmed/15302206

28. Yanshole, L. V, Cherepanov, I. V, Snytnikova, O. A., Yanshole, V. V, Sagdeev, R. Z., and Tsentalovich, Y. P. "**Cataract-Specific Posttranslational Modifications and Changes in the Composition of Urea-Soluble Protein Fraction from the Rat Lens.**" *Molecular vision* 19, (2013): 2196–208. Available at http://www.ncbi.nlm.nih.gov/pubmed/24227915

29. Truscott, R. J. W. "**Age-Related Nuclear Cataract?Oxidation Is the Key**" *Experimental Eye Research* 80, no. 5 (2005): 709–725. doi:10.1016/j.exer.2004.12.007, Available at http://www.ncbi.nlm.nih.gov/pubmed/15862178

30. Guthals, A., Clauser, K. R., and Bandeira, N. "**Shotgun Protein Sequencing with Meta-Contig Assembly.**" *Molecular & cellular proteomics : MCP* 11, no. 10 (2012): 1084–96. doi:10.1074/mcp.M111.015768, Available at http://www.mcponline.org/content/11/10/1084.full.pdf+html

31. Frank, A. M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., and Pevzner, P. A. "**Clustering Millions of Tandem Mass Spectra.**" *Journal of proteome research* 7, no. 1 (2008): 113–22. doi:10.1021/pr070361e, Available at http://dx.doi.org/10.1021/pr070361e

32. Frank, A. M. "**A Ranking-Based Scoring Function for Peptide-Spectrum Matches.**" *Journal of proteome research* 8, no. 5 (2009): 2241–52. doi:10.1021/pr800678b, Available at http://pubs.acs.org/doi/abs/10.1021/pr800678b

33. Guthals, A., Clauser, K. R., Frank, A. M., and Bandeira, N. "**Sequencing-Grade de Novo Analysis of MS/MS Triplets (CID/HCD/ETD) from Overlapping Peptides.**" *Journal of proteome research* 12, no. 6 (2013): 2846–57. doi:10.1021/pr400173d, Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4591044/

34. Guthals, A., Gan, Y., Murray, L., Chen, Y., Stinson, J., Nakamura, G., Lill, J. R., Sandoval, W., and Bandeira, N. "**De Novo MS/MS Sequencing of Native Human Antibodies.**" *Journal of proteome research* 16, no. 1 (2017): 45–54. doi:10.1021/acs.jproteome.6b00608, Available at http://pubs.acs.org/doi/abs/10.1021/acs.jproteome.6b00608

35. Trevisan-Silva, D., Bednaski, A. V, Fischer, J. S. G., Veiga, S. S., Bandeira, N., Guthals, A., Marchini, F. K., Leprevost, F. V, Barbosa, V. C., Senff-Ribeiro, A., and Carvalho, P. C. "**A Multi-Protease, Multi-Dissociation, Bottom-up-to-Top-down Proteomic View of the**

Loxosceles Intermedia Venom." *Scientific data* 4, (2017): 170090. doi:10.1038/sdata.2017.90, Available at http://www.nature.com/articles/sdata201790

36. Kim, S., Mischerikow, N., Bandeira, N., Navarro, J. D., Wich, L., Mohammed, S., Heck, A. J. R., and Pevzner, P. A. "**The Generating Function of CID, ETD, and CID/ETD Pairs of Tandem Mass Spectra: Applications to Database Search.**" *Molecular & cellular proteomics : MCP* 9, no. 12 (2010): 2840–52. doi:10.1074/mcp.M110.003731, Available at http://www.mcponline.org/content/9/12/2840.full.pdf+html

37. Kim, S., Gupta, N., and Pevzner, P. A. "**Spectral Probabilities and Generating Functions of Tandem Mass Spectra: A Strike against Decoy Databases.**" *Journal of proteome research* 7, no. 8 (2008): 3354–63. doi:10.1021/pr8001244, Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2689316&tool=pmcentrez&rendertype=abstract

38. Tanner, S., Shu, H., Frank, A., Wang, L.-C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. "**InsPecT: Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra.**" *Analytical chemistry* 77, no. 14 (2005): 4626–39. doi:10.1021/ac050102d, Available at http://dx.doi.org/10.1021/ac050102d

39. Han, Y., Ma, B., and Zhang, K. "**SPIDER: Software for Protein Identification from Sequence Tags with De Novo Sequencing Error**" *Journal of Bioinformatics and Computational Biology* 03, no. 03 (2005): 697–716. doi:10.1142/S0219720005001247, Available at http://www.worldscientific.com/doi/abs/10.1142/s0219720005001247

40. Bandeira, N., Pham, V., Pevzner, P., Arnott, D., and Lill, J. R. "**Automated de Novo Protein Sequencing of Monoclonal Antibodies.**" *Nature biotechnology* 26, no. 12 (2008): 1336–8. doi:10.1038/nbt1208-1336, Available at http://dx.doi.org/10.1038/nbt1208-1336

41. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. "**InsPecT: Identi Cation of Posttranslationally Modi Ed Peptides from Tandem Mass Spectra**" *Anal Chem* 77, (2005): 4626–4639.

42. Kim, S., Gupta, N., and Pevzner, P. A. "**Spectral Probabilities and Generating Functions of Tandem Mass Spectra: A Strike against Decoy Databases**" *Journal of proteome research* 7, no. 8 (2008): 3354–3363.

43. Bandeira, N., Pham, V. C., Pevzner, P. A., D., A., and J.R., L. "**Automated de Novo Protein Sequencing of Monoclonal Antibodies**" *Nature Biotechnology* 26, (2008): 1336–1338.

44. Elias, J. E. and Gygi, S. P. "**Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry.**" *Nature methods* 4, no. 3 (2007): 207–14. doi:10.1038/nmeth1019, Available at http://dx.doi.org/10.1038/nmeth1019

45. Liu, X., Sirotkin, Y., Shen, Y., Anderson, G., Tsai, Y. S., Ting, Y. S., Goodlett, D. R., Smith,

R. D., Bafna, V., and Pevzner, P. A. "**Protein Identification Using Top-Down.**" *Molecular & cellular proteomics : MCP* 11, no. 6 (2012): M111.008524. doi:10.1074/mcp.M111.008524, Available at http://www.ncbi.nlm.nih.gov/pubmed/22027200

46. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J. C., Carr, S. A., Tabb, D. L., Coffey, R. J., Slebos, R. J. C., Liebler, D. C., and CPTAC, the N. "**Proteogenomic Characterization of Human Colon and Rectal Cancer**" *Nature* 513, no. 7518 (2014): 382–387. doi:10.1038/nature13438, Available at http://www.nature.com/articles/nature13438

47. Slebos, R. J. C., Wang, X., Wang, X., Zhang, B., Tabb, D. L., and Liebler, D. C. "**Proteomic Analysis of Colon and Rectal Carcinoma Using Standard and Customized Databases**" *Scientific Data* 2, (2015): 150022. doi:10.1038/sdata.2015.22, Available at http://www.nature.com/articles/sdata201522

48. Tabb, D. L., Wang, X., Carr, S. A., Clauser, K. R., Mertins, P., Chambers, M. C., Holman, J. D., Wang, J., Zhang, B., Zimmerman, L. J., Chen, X., Gunawardena, H. P., Davies, S. R., Ellis, M. J. C., Li, S., Townsend, R. R., Boja, E. S., Ketchum, K. A., Kinsinger, C. R., Mesri, M., Rodriguez, H., Liu, T., Kim, S., McDermott, J. E., Payne, S. H., Petyuk, V. A., Rodland, K. D., Smith, R. D., Yang, F., Chan, D. W., Zhang, B., Zhang, H., Zhang, Z., Zhou, J.-Y., and Liebler, D. C. "**Reproducibility of Differential Proteomic Technologies in CPTAC Fractionated Xenografts.**" *Journal of proteome research* 15, no. 3 (2016): 691–706. doi:10.1021/acs.jproteome.5b00859, Available at http://www.ncbi.nlm.nih.gov/pubmed/26653538

49. Song, E., Gao, Y., Wu, C., Shi, T., Nie, S., Fillmore, T. L., Schepmoes, A. A., Gritsenko, M. A., Qian, W.-J., Smith, R. D., Rodland, K. D., and Liu, T. "**Targeted Proteomic Assays for Quantitation of Proteins Identified by Proteogenomic Analysis of Ovarian Cancer.**" *Scientific data* 4, (2017): 170091. doi:10.1038/sdata.2017.91, Available at http://www.ncbi.nlm.nih.gov/pubmed/28722704

50. Xu, Z., Wu, C., Xie, F., Slysz, G. W., Tolic, N., Monroe, M. E., Petyuk, V. A., Payne, S. H., Fujimoto, G. M., Moore, R. J., Fillmore, T. L., Schepmoes, A. A., Levine, D. A., Townsend, R. R., Davies, S. R., Li, S., Ellis, M., Boja, E., Rivers, R., Rodriguez, H., Rodland, K. D., Liu, T., and Smith, R. D. "**Comprehensive Quantitative Analysis of Ovarian and Breast Cancer Tumor Peptidomes.**" *Journal of proteome research* 14, no. 1 (2015): 422–33. doi:10.1021/pr500840w, Available at http://www.ncbi.nlm.nih.gov/pubmed/25350482

51. Coscia, F., Watters, K. M., Curtis, M., Eckert, M. A., Chiang, C. Y., Tyanova, S., Montag, A., Lastra, R. R., Lengyel, E., and Mann, M. "**Integrative Proteomic Profiling of Ovarian Cancer Cell Lines Reveals Precursor Cell Associated Proteins and Functional Status.**"

Nature communications 7, (2016): 12645. doi:10.1038/ncomms12645, Available at http://www.ncbi.nlm.nih.gov/pubmed/27561551

52. Mertins, P., Mani, D. R., Ruggles, K. V, Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J. T., Gatza, M. L., Wilkerson, M., Perou, C. M., Yellapantula, V., Huang, K., Lin, C., McLellan, M. D., Yan, P., Davies, S. R., Townsend, R. R., Skates, S. J., Wang, J., Zhang, B., Kinsinger, C. R., Mesri, M., Rodriguez, H., Ding, L., Paulovich, A. G., Fenyö, D., Ellis, M. J., Carr, S. A., and NCI CPTAC, the N. "**Proteogenomics Connects Somatic Mutations to Signalling in Breast Cancer.**" *Nature* 534, no. 7605 (2016): 55–62. doi:10.1038/nature18003, Available at http://www.ncbi.nlm.nih.gov/pubmed/27251275

53. Huang, K.-L., Li, S., Mertins, P., Cao, S., Gunawardena, H. P., Ruggles, K. V, Mani, D. R., Clauser, K. R., Tanioka, M., Usary, J., Kavuri, S. M., Xie, L., Yoon, C., Qiao, J. W., Wrobel, J., Wyczalkowski, M. A., Erdmann-Gilmore, P., Snider, J. E., Hoog, J., Singh, P., Niu, B., Guo, Z., Sun, S. Q., Sanati, S., Kawaler, E., Wang, X., Scott, A., Ye, K., McLellan, M. D., Wendl, M. C., Malovannaya, A., Held, J. M., Gillette, M. A., Fenyö, D., Kinsinger, C. R., Mesri, M., Rodriguez, H., Davies, S. R., Perou, C. M., Ma, C., Reid Townsend, R., Chen, X., Carr, S. A., Ellis, M. J., and Ding, L. "**Proteogenomic Integration Reveals Therapeutic Targets in Breast Cancer Xenografts.**" *Nature communications* 8, (2017): 14864. doi:10.1038/ncomms14864, Available at http://www.ncbi.nlm.nih.gov/pubmed/28348404

54. Tao, F. "**1st NCI Annual Meeting on Clinical Proteomic Technologies for Cancer**" *Expert Review of Proteomics* 5, no. 1 (2008): 17–20. doi:10.1586/14789450.5.1.17, Available at http://www.ncbi.nlm.nih.gov/pubmed/18282119

55. Edwards, N. J., Oberti, M., Thangudu, R. R., Cai, S., McGarvey, P. B., Jacob, S., Madhavan, S., and Ketchum, K. A. "**The CPTAC Data Portal: A Resource for Cancer Proteomics Research**" *Journal of Proteome Research* 14, no. 6 (2015): 2707–2713. doi:10.1021/pr501254j, Available at http://www.ncbi.nlm.nih.gov/pubmed/25873244

56. Lin, Y.-Y., Gawronski, A., Hach, F., Li, S., Numanagić, I., Sarrafi, I., Mishra, S., McPherson, A., Collins, C. C., Radovich, M., Tang, H., and Sahinalp, S. C. "**Computational Identification of Micro-Structural Variations and Their Proteogenomic Consequences in Cancer**" *Bioinformatics* 34, no. 10 (2018): 1672–1681. doi:10.1093/bioinformatics/btx807, Available at http://www.ncbi.nlm.nih.gov/pubmed/29267878

57. Vasaikar, S. V, Straub, P., Wang, J., and Zhang, B. "**LinkedOmics: Analyzing Multi-Omics Data within and across 32 Cancer Types.**" *Nucleic acids research* 46, no. D1 (2018): D956–D963. doi:10.1093/nar/gkx1090, Available at http://www.ncbi.nlm.nih.gov/pubmed/29136207

58. Cha, S. W., Bonissone, S., Na, S., Pevzner, P. A., and Bafna, V. "**The Antibody Repertoire**

of Colorectal Cancer" *Molecular & Cellular Proteomics* 16, no. 12 (2017): 2111–2124. doi:10.1074/mcp.RA117.000397, Available at http://www.ncbi.nlm.nih.gov/pubmed/29046389

59. Dasari, S., Chambers, M. C., Martinez, M. A., Carpenter, K. L., Ham, A.-J. L., Vega-Montoto, L. J., and Tabb, D. L. "**Pepitome: Evaluating Improved Spectral Library Search for Identification Complementarity and Quality Assessment**" *Journal of Proteome Research* 11, no. 3 (2012): 1686–1695. doi:10.1021/pr200874e, Available at http://pubs.acs.org/doi/10.1021/pr200874e

60. Tabb, D. L., Fernando, C. G., and Chambers, M. C. "**MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis**" *Journal of Proteome Research* 6, no. 2 (2007): 654–661. doi:10.1021/pr0604054, Available at http://pubs.acs.org/doi/abs/10.1021/pr0604054

61. Rudnick, P. A., Markey, S. P., Roth, J., Mirokhin, Y., Yan, X., Tchekhovskoi, D. V, Edwards, N. J., Thangudu, R. R., Ketchum, K. A., Kinsinger, C. R., Mesri, M., Rodriguez, H., and Stein, S. E. "**A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline.**" *Journal of proteome research* 15, no. 3 (2016): 1023–32. doi:10.1021/acs.jproteome.5b01091, Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5117628/

62. Frank, A. and Pevzner, P. "**PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling**" *Analytical Chemistry* 77, no. 4 (2005): 964–973. doi:10.1021/ac048788h, Available at http://dx.doi.org/10.1021/ac048788h

63. Dancík, V., Addona, T. A., Clauser, K. R., Vath, J. E., and Pevzner, P. A. "**De Novo Peptide Sequencing via Tandem Mass Spectrometry.**" *Journal of computational biology : a journal of computational molecular cell biology* 6, no. 3–4 (1999): 327–42. doi:10.1089/106652799318300, Available at http://www.ncbi.nlm.nih.gov/pubmed/10582570

64. Bandeira, N. "**Spectral Networks: A New Approach to de Novo Discovery of Protein Sequences and Posttranslational Modifications.**" *BioTechniques* 42, no. 6 (2007): 687, 689, 691 passim. Available at http://www.ncbi.nlm.nih.gov/pubmed/17612289

65. Sciarratta, G. V and Ivaldi, G. "**Hb Matera [Beta 55(D6)Met----Lys]: A New Unstable Hemoglobin Variant in an Italian Family.**" *Hemoglobin* 14, no. 1 (1990): 79–85. Available at http://www.ncbi.nlm.nih.gov/pubmed/2384314

66. Woo, S., Cha, S. W., Bonissone, S., Na, S., Tabb, D. L., Pevzner, P. A., and Bafna, V. "**Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer.**" *Journal of proteome research* 14, no. 9 (2015): 3555–67. doi:10.1021/acs.jproteome.5b00264, Available at http://www.ncbi.nlm.nih.gov/pubmed/26139413

67. Ku, N. O., Liao, J., and Omary, M. B. "**Phosphorylation of Human Keratin 18 Serine 33 Regulates Binding to 14-3-3 Proteins.**" *The EMBO journal* 17, no. 7 (1998): 1892–906. doi:10.1093/emboj/17.7.1892, Available at http://www.ncbi.nlm.nih.gov/pubmed/9524113

68. Toivola, D. M., Ku, N.-O., Resurreccion, E. Z., Nelson, D. R., Wright, T. L., and Omary, M. B. "**Keratin 8 and 18 Hyperphosphorylation Is a Marker of Progression of Human Liver Disease**" *Hepatology* 40, no. 2 (2004): 459–466. doi:10.1002/hep.20277, Available at http://www.ncbi.nlm.nih.gov/pubmed/15368451

69. Bernstein, L., Wertz, J., Na, S., and Bandeira, N. "**Spectral Networks Identification of Rare Post-Translational Modifications and Hypermodified Regions in Lens Crystallin Proteins**" (2018):

70. Gupta, N., Bandeira, N., Keich, U., and Pevzner, P. A. "**Target-Decoy Approach and False Discovery Rate: When Things May Go Wrong.**" *Journal of the American Society for Mass Spectrometry* 22, no. 7 (2011): 1111–20. doi:10.1007/s13361-011-0139-3, Available at http://www.ncbi.nlm.nih.gov/pubmed/21953092